Découverte de motifs fréquents guidée par une ontologie

Yaya TRAORE[1,3], Sadouanouan MALO[2], Cheikh Talibouya DIOP[3], Moussa LO[3], OUARO Stanislas[1]
[1] Université de Ouagadougou , *Ouagadougou – BP 7021, Burkina Faso*
{yaytra,ouaro}@yahoo.fr
[2] Université Polytechnique de Bobo Dioulasso, *Bobo-Dioulasso – BP 1091, Burkina Faso*
sadouanouan@yahoo.fr
[3] Université Gaston Berger de Saint -Louis, *Saint-Louis – BP 234, Sénégal*
{cheikh-talibouya.diop,moussa.lo}@ugb.edu.sn

Résumé : La phase d'extraction des motifs fréquents en fouille de données génère une quantité énorme de motifs fréquents et requiert la mise en place d'un post-traitement efficace afin de cibler les motifs fréquents les plus utiles. Cet article propose une approche de découverte de motifs fréquents qui intègre les connaissances décrites par l'expert et représentées dans une ontologie associée aux données. Nous utilisons l'ontologie pour bénéficier de plus d'informations structurées afin d'éliminer certains motifs fréquents de l'analyse.

Mots-clés : Ontologie, Motifs fréquents, fouille de données

I. INTRODUCTION

La recherche de motifs fréquents est un domaine important de la fouille de données et de la découverte de connaissances dans les bases de données. A l'origine de ce domaine se trouvent les travaux d'Agrawal [1] sur la découverte de motifs fréquents. Le problème de la découverte de motifs fréquents consiste, étant donné un contexte d'extraction défini par un ensemble d'objets décrits par la liste de leurs attributs et un seuil de support minimal *minsup*, à découvrir tous les motifs qui apparaissent plus de *minsup* fois dans la base. L'extraction des motifs fréquents dans [1] consiste à parcourir itérativement par niveaux l'ensemble des motifs. Durant chaque itération ou niveau k, un ensemble de motifs candidats est créé en joignant les motifs fréquents découverts durant l'itération précédente k-1 ; les supports de ces motifs sont calculés et les motifs non fréquents sont supprimés. Dans cette approche le problème de recherche de motifs fréquents est exponentiel par rapport au nombre d'attributs. En effet, s'il y a n attributs, il y a 2^n motifs possibles et dans le pire des cas ils sont tous fréquents. Ainsi la quantité énorme de motifs fréquents extrait requiert la mise en place d'un post-traitement efficace afin de cibler les motifs fréquents les plus utiles.

L'objectif de cet article est d'utiliser une ontologie de domaine comme un support d'élagage sémantique pour éliminer des candidats du calcul des motifs fréquents avec l'algorithme [1]. La phase d'élagage sémantique permettra une réduction qualitative du nombre de motifs fréquents et de garder les plus utiles.

Le reste de l'article est organisé comme suit : à la section 2 nous présentons les définitions et notations qui seront utiles dans l'article. La section 3 présente les travaux liés à notre approche. Nous développons notre approche dans la section 4. Nous terminons par une conclusion et des perspectives.

II. DEFINTIONS ET POSITION DU PROBLEME

Dans cette section, nous définirons les différentes notions utilisées dans le papier puis nous introduirons les notations que nous utiliserons dans la suite du document. Enfin nous allons situer le problème étudié par rapport à ceux du domaine rencontrés dans la littérature.

1. Contexte d'extraction

Un contexte d'extraction est un triplet *CE= (I, A, R)* dans lequel *I* et *A* sont respectivement des ensembles finis d'individus et d'attribut, et *R* est une relation binaire entre les individus et les attributs. Un couple (i, a) ∈ *R* dénote le fait que l'individu i ∈ *I* contient l'attribut a ∈ *A*. Dans notre cas, le contexte d'extraction représente le jeu de données de la fouille.

Nous définissons la fonction g qui permet d'avoir l'ensemble des individus associées à un attribut par : g : A → I tel que pour a ϵ A, g (a) = {i/ i ϵ I}.

Exemple : Le tableau 1 ci-dessous (Tableau 1) illustre par exemple un contexte d'extraction dont I= {1, 2, 3, 4, 5} est un ensemble d'individus et A={ Ontologie, Ontologie de domaine, Mise à jour d'ontologie, Motifs fréquents} est un ensemble d'attributs associés à ces individus.

	Ontologie	Ontologie de domaine	Mise à jour d'ontologie	Motifs fréquents
1	1	0	1	1
2	1	0	0	1
3	0	0	0	0
4	1	0	0	0
5	0	1	1	1

Tableau 1. Contexte d'extraction

2. Motif fréquent

Un motif est un sous ensemble d'attributs. Le support d'un motif est la proportion d'individus associées à ce sous ensemble de motif. Un motif est fréquent si son support est supérieur à un seuil minimal fixé *minsup*. Soit $A_1 \subseteq A$ un motif. Notons $Supp(A_1)$ son support :

$$Supp(A_1) = \frac{|g(A1)|}{|I|}$$

A_1 est un motif fréquent si $Supp(A_1) >= minsup$.

3. Position du Problème

Soit CE un contexte d'extraction, F l'ensemble des motifs fréquents et O une ontologie de domaine. Un motif fréquent f de F est un motif fréquent utile si les éléments qui le constituent ne sont pas sémantiquement proches d'un concept ou si ces éléments ne sont pas de la même hiérarchie de concept.

Le problème qui nous intéresse est, étant donné un contexte d'extraction CE et un seuil de support minimal *minsup*, d'extraire tous les motifs fréquents utiles.

4. Ontologie

Selon [6], une ontologie est "une liste explicite et organisée de tous les termes, relations et objets qui constituent le schéma de représentation d'un domaine". Une ontologie est donc composée d'une structure conceptuelle et d'une structure lexicale comme l'illustre la figure 1.

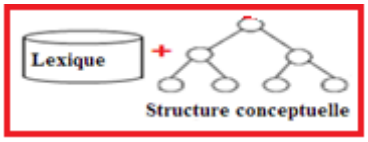

Figure 1- Ontologie

- La structure conceptuelle est formellement définie par S = (C, R, $\leq_C$, σ_R) où C, R sont des ensembles disjoints contenant les concepts et les relations associatives, $\leq_C$ définit la hiérarchie de concepts et $\sigma_R : R \rightarrow C \times C$ est la signature d'une relation entre concept.
- Le lexique contient tous les labels qui sont associés aux concepts et relation de la composante conceptuelle de l'ontologie. Il est formellement défini comme suit : L= (L_C,L_R,F_C F_R) où L_C et L_R sont des ensembles disjoints des labels des concepts, et des relations. F_C est une fonction définie sur l'ensemble des concepts par : $\forall$ l $\in$ L_C, F_C (l)={c / c $\in$ C} et F_R est une fonction définie sur l'ensemble des relations par : $\forall$l $\in$ L_R, F_R (l)={r/ r $\in$ R}. Ces fonctions permettent d'accéder respectivement aux concepts, relations désignés par un label.

L'utilisation de l'ontologie permet de bénéficier de plus d'informations structurées dans un domaine.

III. TRAVAUX EXISTANTS

Trois grandes approches ont été proposées pour l'extraction des motifs fréquents :

- La première approche, inspirée de [1], consiste à parcourir itérativement l'ensemble des motifs par niveaux. Durant chaque itération ou niveau k, un ensemble de motifs candidats est créé en joignant les motifs fréquents découverts durant l'itération précédente k-1 ; les supports de ces motifs sont calculés et les motifs non fréquents sont supprimés. Toutefois, tous les algorithmes qui utilisent cette approche doivent déterminer le support de tous les motifs fréquents.
- La seconde approche est basée sur l'extraction des motifs fréquents maximaux dont tous les sur-ensembles sont non fréquents et tous les sous-ensembles sont fréquents. Les algorithmes utilisant cette approche combinent un parcours par niveaux du bas vers le haut et un parcours du haut vers le bas de l'ensemble des motifs. Lorsque les motifs fréquents maximaux sont découverts, tous les motifs fréquents sont dérivés de ces derniers et un ultime balayage de la base de données est réalisé afin de calculer leur support. L'algorithme le plus efficace basé sur cette approche est l'algorithme [3]. Comme dans le premier cas, les algorithmes basés sur cette méthode doivent calculer les supports de tous les motifs fréquents depuis la base de données.
- La troisième approche basée sur l'extraction des motifs fermés fréquents, inspirée de [10], utilise la fermeture de la connexion de Galois. Les motifs fermés fréquents (et leurs supports) sont extraits de la base de données en réalisant un parcours par niveaux. Tous les motifs fréquents et leur support peuvent donc être déduits des motifs fermés fréquents avec leur support, sans accéder à la base de données.

Parmi ces trois approches, nous nous intéressons à l'extraction de motifs fréquents avec la première approche [1] (**Algorithme 1).** Toutefois cette approche n'utilise pas les connaissances du domaine dans la phase d'élagage des motifs candidats. Ainsi nous proposons d'utiliser une ontologie comme une contrainte d'élagage de motifs candidats avant le calcul du support des motifs fréquents.

Algorithme 1 : Algorithme de découverte de motifs fréquents

Entrée :CE:contexte d'extraction (Base de transaction),minsup:seuil minimum de support
Sortie : F:motifs fréquents
Début
1. L_1=ensemble des 1-itemsets fréquents
2. K=2
3. Tant que($L_{K-1} \neq \emptyset$) faire
4. **// Phase de génération des candidats**
5. C_K = ensemble des K-itemsets C tels que : C = F1 $\cup$ F2 où F1 et F2 sont éléments de L_{k-1} et F1$\cap$F2 comporte (K-2) éléments
6. **//Phase d'élagage**
7. Supprimer de C_K tout candidat C tel qu'il existe un sous-ensemble de C de (K-1) éléments non présent dans L_{K-1}
8. **// Phase d'évaluation des candidats**
9. Calculer le support de chaque candidat C dans C_K
10. L_K ={Cϵ C_K / support(C) >= minsup}
11. K=K+1
12. Fintanque
13. Retourner F=$\cup L_K$

Fin

Un certain nombre de travaux utilisant les ontologies dans le processus d'extraction de connaissances à partir des données (ECD) existent. [5] utilisent l'ontologie pendant la phase de prétraitement, [4] utilisent l'ontologie dans le prétraitement et le post-traitement, [10] l'utilisent dans le post-traitement pour réduire la quantité de règles extraites à partir des schémas de règles. Dans ces approches la disponibilité d'un expert du domaine est nécessaire pour valider les correspondances entre les concepts de l'ontologie et les sous-ensembles d'enregistrement de la base de données, ce qui n'est pas toujours possible. Dans notre cas, il s'agit d'utiliser l'ontologie comme une contrainte d'élagage. Ainsi en s'inspirant de [2] qui utilise deux types de contraintes, définies sur la base des conditions exprimées dans l'ontologie (les contraintes d'abstraction qui sont utilisées pour définir la généralisation de certains items et les contraintes d'élagage qui sont utilisées pour exclure des items de l'analyse), nous utilisons l'ontologie comme une contrainte d'élagage sémantique pour enlever les motifs candidats dont les éléments existent dans la même hiérarchie de concepts de l'ontologie ou sont sémantiquement proches d'un concept de l'ontologie.

IV. DESCRIPTION DE L'APPROCHE

1. Description de l'approche

L'approche proposée dans cet article consiste à mettre en place un système utilisant une ontologie de domaine comme une contrainte d'élagage pour enlever certains motifs candidats du calcul de motifs fréquents afin de réduire le nombre de motifs fréquents qui sera extrait par l'algorithme Apriori [1]. Notre démarche qui se décompose en deux phases illustrée par la figure 2, introduit une phase d'élagage sémantique des motifs candidats. L'élagage sémantique (phase 1) élimine du calcul des motifs fréquents tout candidat dont les éléments sont :

- sémantiquement proches d'un concept de l'ontologie,
- dans la même hiérarchie de concepts de l'ontologie.

La phase (2) calcul le support des motifs candidats qui n'ont pas été éliminé dans la phase d'élagage sémantique.

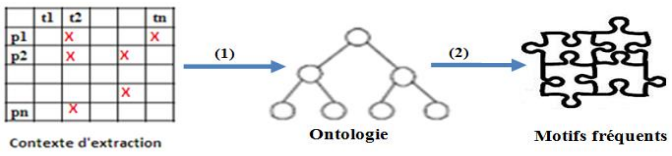

Figure 2- Approche de découverte de motifs fréquents

2. Algorithme de l'approche

L'algorithme (Algorithme 2) de notre approche est une adaptation de l'algorithme de [1] (illustré par l'algorithme 1). Elle intègre après la génération des candidats (ligne N°5) et la phase d'élagage (ligne N°7), une phase d'élagage sémantique (ligne N° 9) qui élimine chaque candidat dont les éléments vérifient les conditions en (1).

Notre approche utilise l'algorithme 2 ci-dessous :

Algorithme 2 : Algorithme de découverte de motifs fréquents guidée par une ontologie

Entrée : CE:contexte d'extraction (Base de transaction),O:l'ontologie de domaine ,minsup:seuil minimum de support
Sortie : F:motifs fréquents
Début
1. L_1=ensemble des 1-itemsets fréquents
2. K=2
3. Tant que($L_{K-1} \neq \emptyset$) faire
4. **// Phase de génération des candidats**
5. C_K = ensemble des K-itemsets C tels que : C = F1 ∪ F2 où F1 et F2 sont éléments de L_{k-1} et F1∩F2 comporte (K-2) éléments
6. **//Phase d'élagage**
7. Supprimer de C_K tout candidat C tel qu'il existe un sous-ensemble de C de (K-1) éléments non présent dans L_{K-1}
8. **//Phase d'élagage sémantique**
9. Supprimer de C_K tout candidat C tel que les éléments de C sont sémantiquement proches d'un concept de l'ontologie O ou sont dans la même hiérarchie de concepts de l'ontologie O
10. **// Phase d'évaluation des candidats**
11. Calculer le support de chaque candidat C dans C_K
12. L_K ={C ϵ C_K / support(C)>= minsup}
13. K=K+1
14. Fintanque
15. Retourner F=∪L_K

Fin

L'originalité de l'approche consiste à intégrer de manière explicite les éléments d'une ontologie de domaine (ligne N°9 de l'algorithme 2) pour élaguer sémantiquement certains motifs candidats dans la découverte des motifs fréquents. L'apport de l'ontologie dans l'approche est d'abord sa terminologie, son expressivité et la puissance de son raisonneur qui permet de bénéficier de plus d'informations structurées afin d'élaguer sémantiquement certains motifs candidats dans le calcul des motifs fréquents. Le résultat donne des motifs fréquents qu'on ne peut pas déduire en raisonnant avec une ontologie de domaine.

V. Validation de l'approche

Pour valider notre approche, nous allons utiliser le contexte d'extraction du tableau 1 et une ontologie (figure 3) sur des thématiques de recherches en informatique. La figure 4 montre une démarche expérimentale de la découverte de motifs fréquents utilisant l'algorithme 2 (Apriori avec notre approche) et l'algorithme 1 (Apriori sans notre approche).

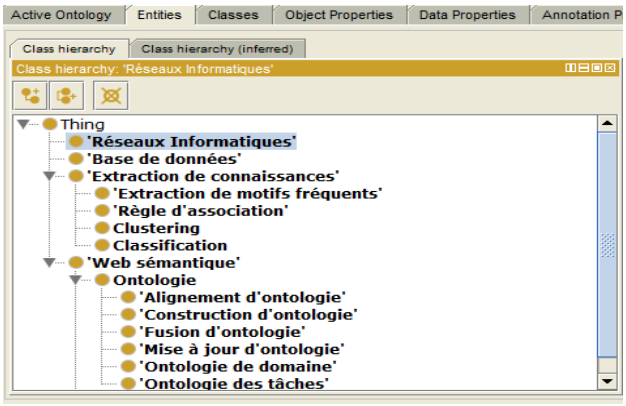

Figure 3 – Extrait de l'ontologie des thématiques

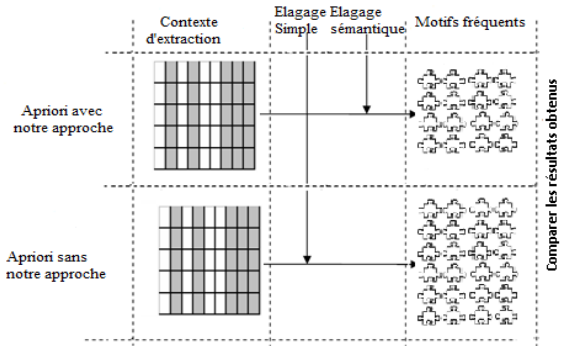

Figure 4- La démarche d'expérimentation

Le tableau 2 donne les résultats des motifs fréquents extraits avec l'algorithme 1 sur le contexte d'extraction avec *minsup*=20%.

N°	Motifs fréquents	Support
1	Ontologie de domaine, Mise à jour d'ontologie, Motifs fréquents	20%
2	Ontologie de domaine, Mise à jour d'ontologie	20%
3	Ontologie de domaine, Motifs fréquents	40%
4	Mise à jour d'ontologie, Motifs fréquents	40%
5	Ontologie, Mise à jour d'ontologie, Motifs fréquents	20%
6	Ontologie, Mise à jour d'ontologie	20%
7	Ontologie, Motifs fréquents	60%

Tableau 2 - Motifs fréquents extrait sans notre approche

Le tableau 3 donne les résultats des motifs fréquents extraits avec l'algorithme 2 sur le Contexte d'extraction avec *minsup*=20%. Les motifs fréquents N°2 et N°6 du tableau 2 sont enlevés.

N°	Motifs fréquents	Support
1	Ontologie de domaine, Mise à jour d'ontologie, Motif fréquent	20%
2	Ontologie de domaine, Motif fréquent	40%
3	Mise à jour d'ontologie, Motif fréquent	40%
4	Mise à jour d'ontologie, Ontologie, Motif fréquent	20%
5	Ontologie, Motif fréquent	60%

Tableau 3 - Motifs fréquents extrait avec notre approche

La lecture des résultats (Tableau 2 et 3) montre que l'utilisation de l'ontologie à réduit le nombre de motifs fréquents extraits de 7 à 5 motifs fréquents.
Aussi on peut observer que les motifs N°2 et N° 6 du tableau 2 enlevés dans le tableau 3 sont de la même hiérarchie de concepts de l'ontologie (Figure 3) et peuvent être extrait avec l'ontologie.

VI. Conclusion

Cet article a présenté une approche de découverte de motifs fréquents basée sur l'algorithme Apriori [1] en intégrant une phase d'élagage sémantique. Cette phase d'élagage utilise une ontologie comme un support d'élagage pour supprimer les motifs candidats du calcul. De nombreuses perspectives s'offrent à la suite de nos travaux. La première d'entre elles est d'évaluer notre approche avec un volume important de données afin d'analyser plus en détail l'impact de notre proposition. Les autres perspectives seront consacrées au développement des techniques exploitant les résultats de l'algorithme 2 pour extraire des règles d'association.

REFERENCES

[1] Agrawal R. and Srikant R. Fast algorithms for mining association rules in larges databases,Proc. VLDB conf.,pp 478-499,September 1994
[2] Antunes, C. ONTO4AR : A Framework for Mining Association Rules. In Proceedings of the International Workshop on Constraint-Based Mining and Learning (CMILE ″UPKDD), Warsaw, Poland, pp.37– 48.
[3] Bayardo R. J., « Efficiently Mining Long Patterns from Databases », *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, juin 1998, p. 85-93.
[4] 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.
[5] 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
[6] Gennari J. H., S. W. Tu, T. E. Rothenfluh et M. A. Musen., Mapping domains to methods in support of reuse.

International Journal of Human-Compute Studies,,41:399–424, 1994

[7] Hernandez Nathalie (2006). Ontologie de domaine pour la modélisation du contexte en recherche d'information, Thèse de Doctorat, Université Paul Sabatier de Toulouse

[8] Hernandez Nathalie, Hubert Gilles, Mothe Josiane, Ralalason Bachelin,(2008). RI et Ontologies – Etat de l'art 2008, Rapport interne, N° IRIT/RR-2008-14-FR , Juillet 2008

[9] 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.

[10] N. Pasquier Y. Bastide R. Taouil et L. Lakhal. (1998) . Pruning closed itemset lattices for association rules. In Actes des 14 journées Bases de Données Avancées (BDA'98), pages 177-196.