Découverte de motifs fréquents guidée par une ontologie

Yaya TRAORE[1,2], Cheikh Talibouya DIOP[2], Sadouanouan MALO[3], Moussa LO[2], Stanislas OUARO[1]

[1]Université de Ouagadougou (Burkina Faso)(yaytra@yahoo.fr , ouaro@yahoo.fr)

[2]Université Gaston Berger de Saint Louis (Sénégal) (cheikh-talibouya.diop@ugb.edu.sn , moussa.Lo@ugb.edu.sn)

[3]Université Polytechnique de Bobo Dioulasso (Burkina Faso) (sadouanouan@yahoo.fr)

RÉSUMÉ. L'extraction des motifs fréquents en fouille de données génère une quantité énorme de motifs fréquents et requiert par conséquent 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 utiles qui intègre les connaissances décrites par l'expert et représentées dans une ontologie associée aux données. L'approche utilise l'ontologie pour bénéficier de plus d'informations structurées afin d'éliminer certains motifs fréquents de l'analyse. Les expérimentations réalisées avec notre approche donnent des résultats satisfaisants.

ABSTRACT. The frequent pattern mining generates a huge amount of patterns and therefore requires the establishment of an effective post-treatment to target the most useful. This paper proposes an approach to discover the useful frequent patterns that integrates knowledge described by the expert and represented in an ontology associated with the data. The approach uses the ontology for benefit from more structured information to remove some frequent patterns of the analysis. The experiments realized with our approach give satisfactory results.

MOTS-CLÉS: Ontologie, Motifs fréquents, fouille de données

KEYWORDS: Ontology, Frequent pattern, data mining

1. 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 extraits 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 préliminaires et l'énoncé du problème qui seront utiles dans l'article. La section 3 présente les travaux liés à notre approche. Nous développons notre approche dans la section 4 puis nous terminons par une conclusion et des perspectives.

2. Préliminaires et position du problème

Dans cette section, nous définissons les différentes notions puis nous introduisons 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.

2.1. 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'attributs, et R est une relation binaire entre les individus et les attributs. Un couple $(i, a) \in R$ dénote le fait que l'individu $i \in$

I contient l'attribut $a \in A$. 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 associés à un attribut par : $g : A \rightarrow I$ tel que pour $a \in A, g(a) = \{i / i \in I \text{ et } (i, a) \in R\}$.

2.2. Motif fréquent

Un motif (ou itemset) est un sous ensemble d'attributs. Le support d'un motif est la proportion d'individus associés à ce sous ensemble de motif. Un motif est fréquent si son support est supérieur à un seuil minimal fixé *minsup*. Formellement cela se définit comme suit : Soit $A1 \subseteq A$ un motif. Notons $Supp(A1)$ son support :

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

Dans cette formule : $g(A1) = \cap_{a \in A1} g(a)$

et $|g(A1)|$ donne le nombre d'individus associé aux attributs $a \in A1$ et $|I|$ donne le nombre total des individus. Un motif $A1$ est fréquent si son support est supérieur au seuil de support minimal *minsup* : $Supp(A1) >= minsup$.

2.3. Ontologie

Les ontologies sont utilisées de manière générale pour formaliser les connaissances d'un domaine. Elle est définie par Gruber [10] comme "une spécification explicite d'une conceptualisation". Cette définition a ensuite été complétée par [16] qui définissent une ontologie comme une "spécification formelle et explicite d'une conceptualisation partagée". En s'appuyant sur ces définitions, et en s'inspirant des travaux de [8], nous considérons une ontologie illustrée par la figure 1, comme un ensemble structuré des concepts pertinents d'un domaine. Chaque concept est dénoté par un ensemble de termes consensuels qui n'est pas propre à un individu mais accepté par une communauté d'utilisateurs afin d'en garantir une utilisation plus large et partagée des connaissances du domaine. Ainsi nous définissons formellement l'ontologie par $O = (S, L)$ où S, L représentent respectivement la structure conceptuelle, et la structure lexicale :

- La structure conceptuelle est définie par $S = (C, R, H, \sigma_R)$ où :
  - C, R sont des ensembles disjoints contenant les concepts et les relations associatives,
  - $H \sqsubseteq C \times C$ est une taxonomie de concepts. $(c_1, c_2) \in H$ signifie que le concept c_1 est subsumé par c_2. Notons H^{C_1} une sous partie de H qui désigne l'ensemble des concepts subsumés par C_1. $C_{11} \in H^{C_1}$ signifie que :
    - $\exists\, C_{1i} \in C$ tel que C_{11} est subsumé par C_{1i} et $C_{1i} \in H^{C_1}$,

- ou que C_{11} est subsumé par C_1.

- $\sigma_R : R \rightarrow C \times C$ 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 défini par : $L = (L_C, L_R, F_C, F_R)$ 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 : $\forall l \in L_R, F_R(l) = \{r / r \in C\}$. Ces fonctions permettent d'accéder respectivement aux concepts et relations désignés par un label.

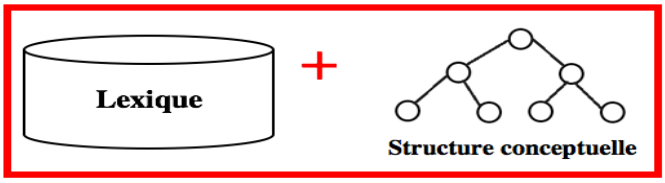


Figure 1. Ontologie

2.4. Position du problème

Soit CE un contexte d'extraction, F l'ensemble des motifs, O une ontologie de domaine et C, L respectivement l'ensemble des concepts et labels des concepts de O. Soit $C_i \in C$ un concept de C et H^{C_i} l'ensemble des concepts subsumé par le concept C_i.

Définition 1 : Soit t un terme, σ un seuil de similarité et *DSim* une distance sémantique (nous utilisons la distance de JaroWinkler [17] qui est une métrique bien adaptée pour la similarité entre deux chaînes de caractères courtes). Le terme t est sémantiquement proche d'un label associé au concept C_i ($t \equiv C_i$) si et seulement si : $\exists L_i \sqsubseteq L$ avec $F_C(L_i) = C_i$ tel que *DSim(t, L_i) >= σ*. Ainsi un terme t est sémantiquement proche du label d'un concept de O s'il existe un concept $C_i \in C$ tel que $t \equiv C_i$.

Définition 2 : Deux concepts $C_{i1} \in C$ et $C_{i2} \in C$ sont de la même hiérarchie de concepts de C_i ($C_i \in C$) si et seulement si $C_{i1} \in H^{C_i}$ et $C_{i2} \in H^{C_i}$.

Définition 3 : Un motif f de F n'est pas utile si les éléments qui le constituent sont sémantiquement proches du label d'un concept de l'ontologie ou si ces éléments sont de la même hiérarchie de concept. Formellement cela se définit comme suit : soit A, B, E, F, G, P des attributs ou items :

- si f est un « 1-itemset » (f est constitué d'un seul attribut) (par exemple $f = A$) alors f est un motif utile si et seulement si f n'est pas un élément de C,
- si f est un « N-itemset » (f est constitué de N attributs avec N >=2) (par exemple $f = ABP$) alors f est un motif utile si et seulement si :

- $\forall ai \in f$, s'il existe $C_i \in C / ai \in H^{C_i}$ alors il existe au moins un $aj \in f$ (avec $aj \not\equiv ai$) tel que $aj \notin H^{C_i}$,
- ou $\forall ai \in f$ et $\forall H^{C_i}$ une hiérarchie de concepts quelconque de C alors $ai \notin H^{C_i}$.

Par exemple considérons $C = \{C_1, C_2, A, B, E, F, G\}$ l'ensemble des concepts d'une ontologie et H^{C_1}={A,B} l'ensemble des concepts subsumés par le concept C_1 et H^{C_2}={E,F,G} l'ensemble des concepts subsumés par le concept C_2.
Considérons $F = \{f_1, f_2, f_3, f_4\}$ l'ensembles des motifs tel que : f_1=A, f_2=P, f_3=AB, f_4=AG.
Les motifs utiles de F sont $= \{f_2, f_4\}$:

- f_2 est un motif utile car $P \not\subseteq C$,
- f_4 est un motif utile car $A \in H^{C_1}$, $G \in H^{C_2}$ dont A et G ne sont pas de la même hiérarchie de concepts,
- f_1 n'est pas un motif utile car A est un sous concept de C_1 dont $A \in C$
- f_3 n'est pas un motif utile car A et B sont de la même hiérarchie des concepts de C_1 : $A \in H^{C_1}$ et $B \in H^{C_1}$.

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.

3 Travaux existants

Ce travail est la suite de nos travaux [18] où nous nous intéressons à la fouille de motifs fréquents de tags potentiellement utiles pour guider la découverte de catégories et sous catégories dans un wiki sémantique. De nombreux algorithmes[1] Apriori [1], Eclat [19], dEclat [21], PrePost [5], PrePost+ [6], Charm [20], Close [14], [15], FPMax [9], [3] ont été proposés pour la découverte des motifs fréquents. Parmi ces algorithmes nous considérons trois approches. 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 basée sur l'extraction des motifs fermés fréquents, inspirée de [14] et [15] 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. Enfin la troisième approche est basée sur l'extraction des motifs fréquents maximaux [3], [9] dont tous les sur-ensembles sont non fréquents et tous

[1] http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php

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. Comme dans la première approche, les algorithmes basés sur cette méthode doivent calculer les supports de tous les motifs fréquents depuis la base de données.

Parmi ces trois approches, nous nous intéressons à l'extraction de motifs fréquents avec la première approche [1] (illustrer par l'*Algorithme Apriori).* Toutefois, aucune des approches de fouille n'intègre de manière explicite les connaissances du domaine dans l'algorithme. L'expert est obligé de refaire une phase de post-traitement pour retenir les motifs les plus utiles. Vu la quantité de motifs qui peut être générée, cette tâche n'est pas facile. Ainsi l'intégration des connaissances du domaine est un moyen pour réduire la charge de travail. Nous proposons d'utiliser une ontologie comme une contrainte d'élagage de motifs candidats avant le calcul du support des motifs fréquents.

Un certain nombre de travaux utilisant les ontologies dans le processus d'extraction de connaissances à partir des données (ECD) existent. [7] utilise l'ontologie pendant la phase de prétraitement, [4] utilise l'ontologie dans le prétraitement et le post-traitement, [13] l'utilise 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. [22] propose l'intégration des connaissances de domaine dans le processus d'extraction des connaissances (ECD) multi-vues. Cette approche est intéressante mais ce processus est mené par plusieurs experts avec différents points de vue. Dans notre cas, il s'agit d'utiliser l'ontologie comme un support d'élagage. Ainsi en s'inspirant des travaux de [23] qui exploitent les connaissances disponibles des experts dans l'extraction des règles d'association, des travaux de [11] et [12] qui montrent qu'une ontologie peut être utilisée pour améliorer la recherche d'information et 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 un support d'élagage sémantique pour enlever les motifs candidats dont les éléments existent dans la même hiérarchie de concepts ou sont sémantiquement proches d'un concept de l'ontologie. L'ontologie est utilisée dans la phase de fouille des motifs fréquents. Cela permet de calculer uniquement le support des candidats retenus après l'élagage sémantique au lieu de calculer le support de tous les candidats. Dans la section suivante nous illustrons notre approche avec l'algorithme Apriori [1].

4 Approche de découverte de motifs fréquents

4.1. Description de l'approche

L'approche proposée dans cet article consiste à mettre en place un système utilisant une ontologie de domaine comme un support 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 illustrées 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.

L'approche est illustrée par *l'algorithme 2*.

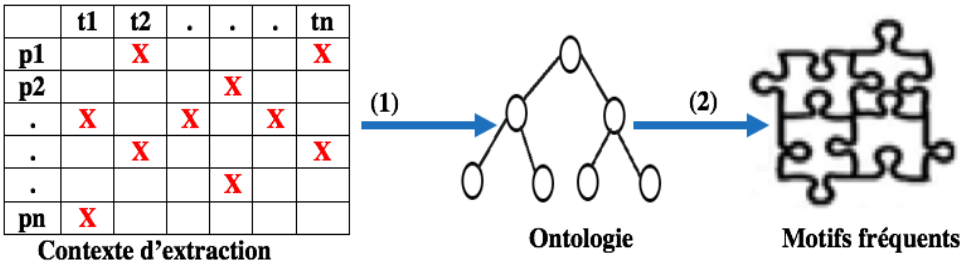


Figure 2. Approche de découverte de motifs fréquents

4.2. Algorithme de l'approche

L'algorithme (*Algorithme 2*) qui illustre notre approche est une adaptation de l'algorithme de [1] (*Algorithme Apriori*). 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).

L'originalité de l'approche consiste à intégrer de manière explicite les éléments d'une ontologie de domaine (ligne N°9 et ligne N°15 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 utiles qu'on ne peut pas déduire en raisonnant avec une ontologie de domaine.

Algorithme 1 : Algorithme Apriori : 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 ∪ 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'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=∪L_K

Fin

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 utiles
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 qui n'est pas un motif utile**
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. **Supprimer de L_1 tout motif I tel qu'il existe un concept C_i de O tel que I≡C_i**
16. Retourner F=∪L_K

Fin

4.3. Illustration de l'approche

Le Tableau 1 ci-dessous illustre par exemple un contexte d'extraction dont I= {PX01, PX02, PX03, PX04, PX05} est un ensemble d'individus et A={Apoptosis(A), Immunity(B),Cytolysis(C),Membrane(D),Mitochondrion(E),Nucleus(F), Cytoplasm(G), 3D-Structure (H), Calcium (Ca)} est un ensemble d'attributs associés à ces individus. La figure 3 illustre un extrait d'une ontologie du domaine.

	A	B	C	D	E	F	G	H	Ca
PX01	1	0	0	0	0	1	1	0	0
PX02	1	1	1	0	0	0	1	0	0
PX03	0	1	1	0	1	1	0	0	1
PX04	1	1	1	0	1	0	0	1	0
PX05	0	1	1	1	1	1	0	0	0

Tableau 1. Contexte d'extraction

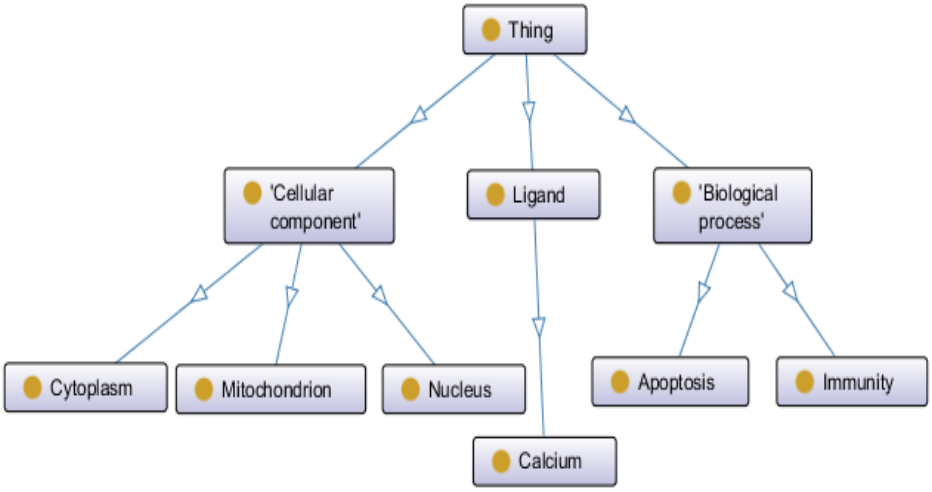
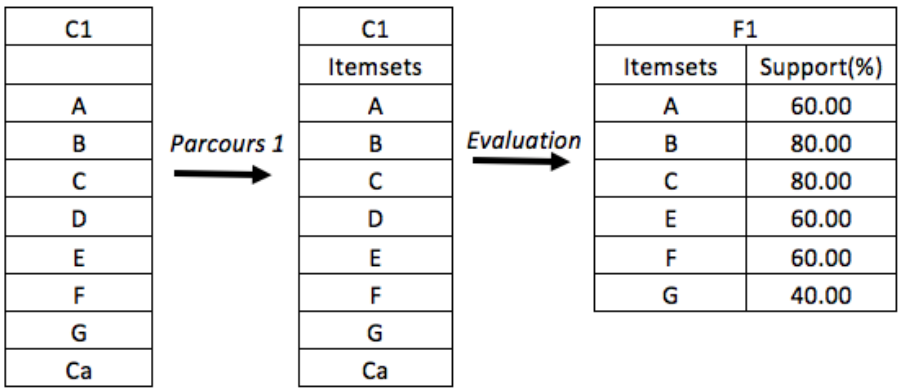


Figure 3. Extrait d'une ontologie de domaine

Dans cet exemple nous montrons le processus de découverte de motifs fréquents guidée par l'ontologie de la figure 3 sur le contexte d'extraction du Tableau 1 avec un support *minsup*=40%.
Déroulement de l'approche :

- **Calcul de L_1 : Candidats : [A, B, C, D, E, F, G, H, Ca]**

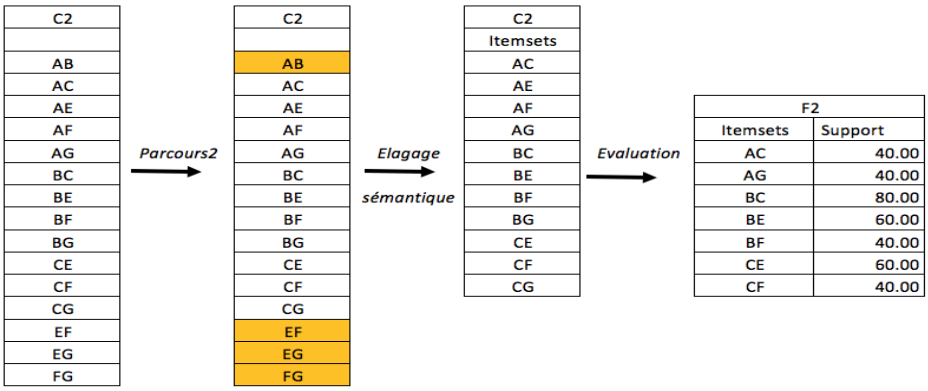
A la première itération, chaque item (ou attribut) de I est un 1-itemset (sous ensemble d'attribut ou motif) de C1. Un premier parcours de CE permet de trouver le support de chaque 1-itemset. Tous les 1-itemsets fréquents, i.e. de support supérieur ou égal à 40% seront gardés dans F1.



L_1= {A, B, C, E, F, G}

- **Calcul de L_2 : Candidats : [AB, AC, AE, AF, AG, BC, BE, BF, BG, CE, CF, CG, EF, EG, FG]**

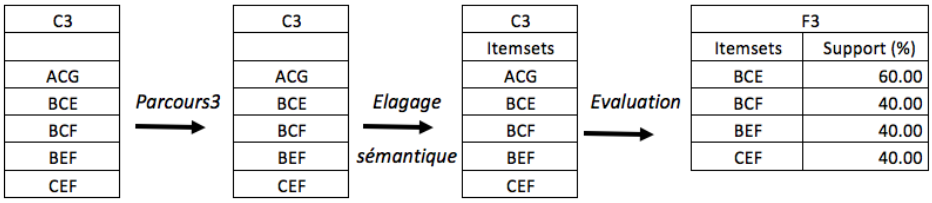
Afin de découvrir les 2-itemsets fréquents, on effectue dans la seconde itération une jointure de F1 ⋈ F1 pour trouver l'ensemble C2 des candidats de taille 2. Après le calcul des candidats un élagage sémantique est effectué pour enlever les candidats qui sont dans une même hiérarchie de concept dans l'ontologie. Ainsi les candidats AB et EF sont de la même hiérarchie de concept dans l'ontologie. Donc ils sont élagués dans la phase d'élagage sémantique. On évalue ensuite le support des candidats et les 2-itemsets fréquents, i.e. de support supérieur ou égal à 40% seront gardés dans F2.



L_2= {AC, AG, BC, BE, BF, CE, CF}

- **Calcul de L_3 : Candidats : [ACG, BCE, BCF, BEF, CEF]**

Pour découvrir les 3-itemsets fréquents, on effectue dans la troisième itération une jointure de F2 ⋈ F2 pour trouver l'ensemble C3 des candidats de taille 3. Après le calcul des candidats un élagage sémantique est effectué pour enlever les candidats qui sont dans une même hiérarchie de concept dans l'ontologie. Il n'y a pas de concept ou des sous concepts de la même hiérarchie parmi les candidats. Donc on évalue le support des candidats et les 3-itemsets fréquents, i.e. de support supérieur ou égal à 40% seront gardés dans F3.



L_3= {BCE, BCF, BEF, CEF}

- **Calcul de L_4 : Candidats : [BCEF]**

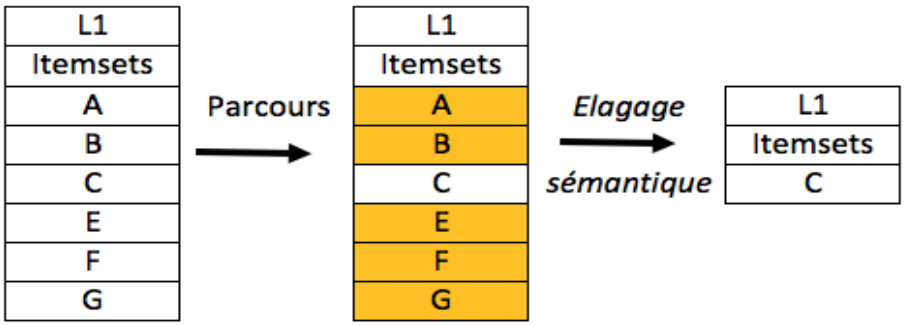
Pour découvrir les 4-itemsets fréquents, on effectue dans la quatrième itération une jointure de F3 ⋈ F3 pour trouver l'ensemble C4 des candidats de taille 4. Après le calcul des candidats un élagage sémantique est effectué pour enlever les candidats qui sont dans une même hiérarchie de concept dans l'ontologie. On a un seul candidat et ses éléments ne sont pas des concepts de la même hiérarchie. Donc on évalue le support du candidat qui est égale à 40%.

L_4= {BCEF}

On n'a plus de combinaison possible avec les éléments de L_4. Donc on arrête le calcul et on élague les concepts de l'ontologie dans **L_1**.

- **Elagage des concepts de l'ontologie dans L_1= {A, B, C, E, F,G}**

Afin d'enlever les connaissances du domaine dans la découverte, on effectue un dernier parcours sur les 1-itemsets fréquents dans L_1 pour élaguer sémantiquement les concepts ou sous concepts de l'ontologie qui sont des 1-itemsets fréquents gardés dans F1. Les concepts ou sous concepts A, B, E, F sont élagués.



L_1= {C}

L'ensembles des motifs découverts sont :

- Frequent 1-itemsets: [C]
- Frequent 2-itemsets: [AC, AG, BC, BE, BF, CE, CF]
- Frequent 3-itemsets: [BCE, BCF, BEF, CEF]
- Frequent 4-itemsets: [BCEF]

Le nombre de motifs extraits après application de l'algorithme avec l'élagage sémantique est égal à 13. Sans l'élagage sémantique le nombre de motifs extraits avec l'algorithme Apriori *(Algorithme Apriori)* est 21. La lecture des résultats montre que l'utilisation de l'ontologie a réduit le nombre de motifs fréquents extraits de 21 à 13 motifs fréquents.

5 Expérimentation

5.1. Protocole expérimental

Nos expérimentations ont été réalisées en utilisant une base de données biologiques Uniprot (www.uniprot.org) qui contient des protéines de toutes les espèces. Ces protéines sont étiquetées avec des mots-clés qui peuvent être utilisés pour récupérer des sous-ensembles particuliers de protéines. Une ontologie « Keywords » disponible également sur Uniprot, décrit les connaissances associées aux mots clés. Cette ontologie contient 10 concepts et 1182 sous concepts.

Nous avons ainsi constitué un jeu de données de 129311 lignes de protéines. Ces données ont été téléchargées à partir de Uniprot : les mots clés sont les attributs du contexte d'extraction. L'ontologie « Keywords » (figure 4) représente l'ontologie de domaine associées aux données du contexte d'extraction.

La figure 5 illustre la démarche expérimentale qui consiste à évaluer la découverte de motifs fréquents utilisant l'algorithme 2 (Apriori avec notre approche) et l'algorithme

Apriori (Apriori sans notre approche) afin de vérifier la pertinence de notre proposition. Nous avons fait varier la valeur de *minsup* entre 5% et 20%. Pour valider les résultats expérimentaux nous utilisons également notre approche avec l'algorithme Close [14] pour vérifier si elle diminue également le nombre de motifs.

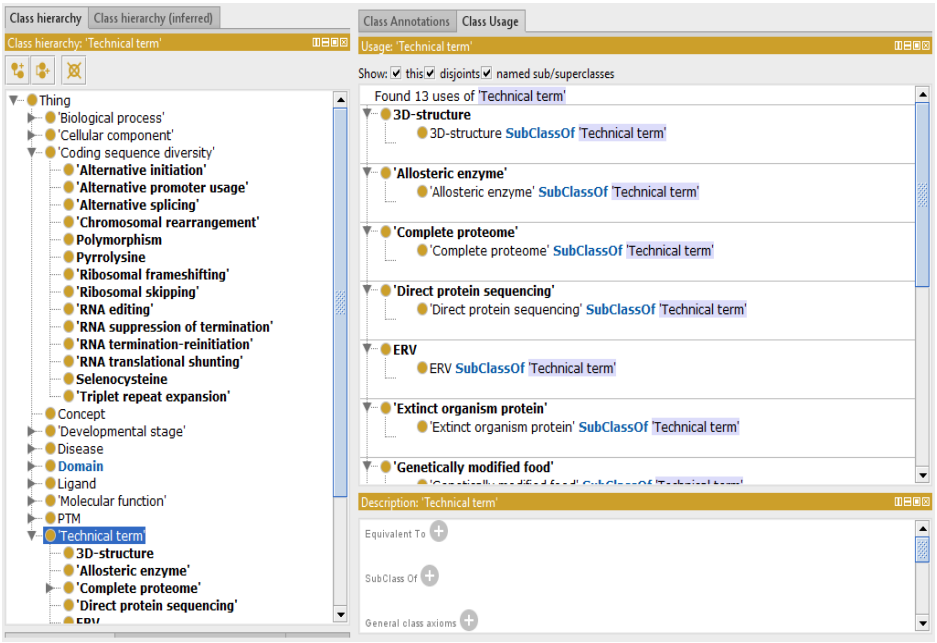


Figure 4. Ontologie de domaine des mots clés des protéines sur Uniprot(www.uniprot.org)

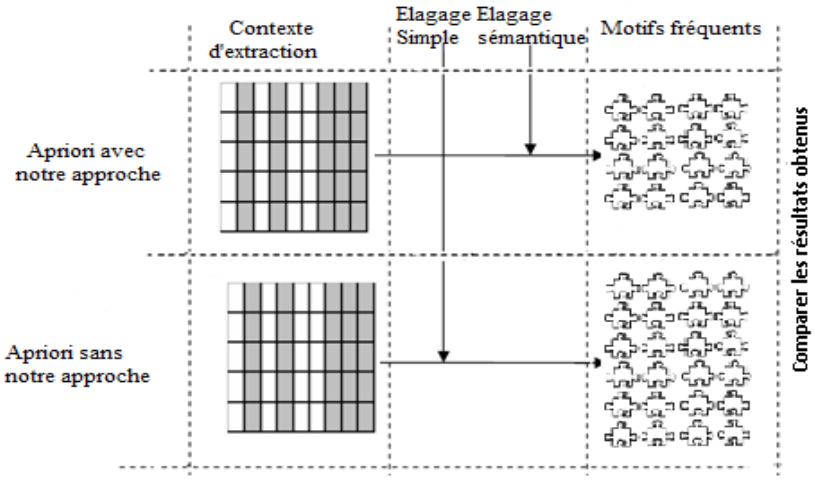


Figure 5. La démarche d'expérimentation

5.2. Résultats et discussions

Les figures 6 et 7 montrent les résultats expérimentaux obtenus par l'application de notre approche avec les algorithmes de fouille comme *Apriori* et *Close*. Sur la figure 6 nous observons que le nombre de motifs extraits en intégrant l'élagage sémantique dans l'algorithme Apriori est toujours inférieur au nombre de motifs extraits avec Apriori. La baisse du nombre de motifs extraits est liée à l'utilisation de l'ontologie comme un filtre pour élaguer de manière automatique les motifs qui peuvent être obtenus en raisonnant avec l'ontologie. Cette baisse est également fonction du minimum de support *(minsup*) utilisé. Plus le minimum de support *(minsup)* est élevé alors le nombre de motifs diminue et tend à se rapprocher. Cela s'explique par le fait que durant chaque itération ou niveau k, les motifs fréquents découverts sont des motifs fréquents du niveau k-1.

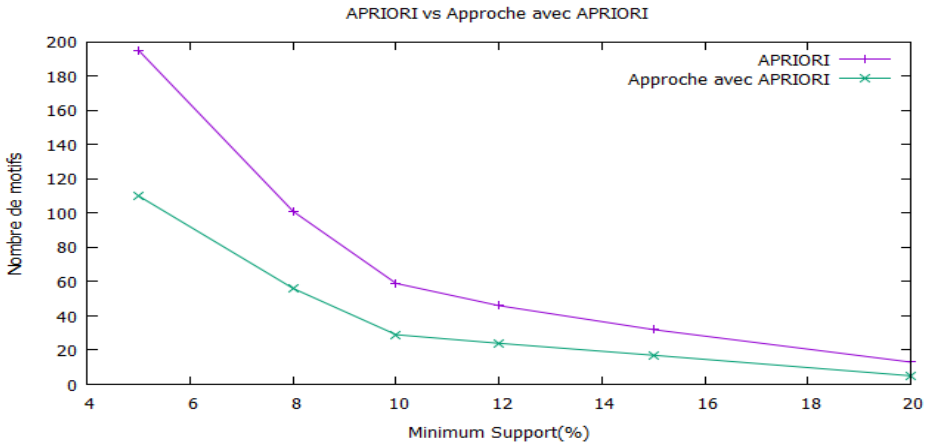


Figure 6. Nombre de motifs extraits avec l'algorithme APRIORI et notre Approche avec APRIORI

Nous observons avec la figure 7 que le nombre de motifs diminue également en intégrant l'élagage sémantique dans l'algorithme Close. Ainsi nous pouvons dire qu'en intégrant l'ontologie dans la découverte de motifs fréquents, seuls les motifs qu'on ne peut pas déduire en raisonnant avec l'ontologie sont conservés et les autres qui sont des concepts ou sous concepts de l'ontologie sont enlevés.

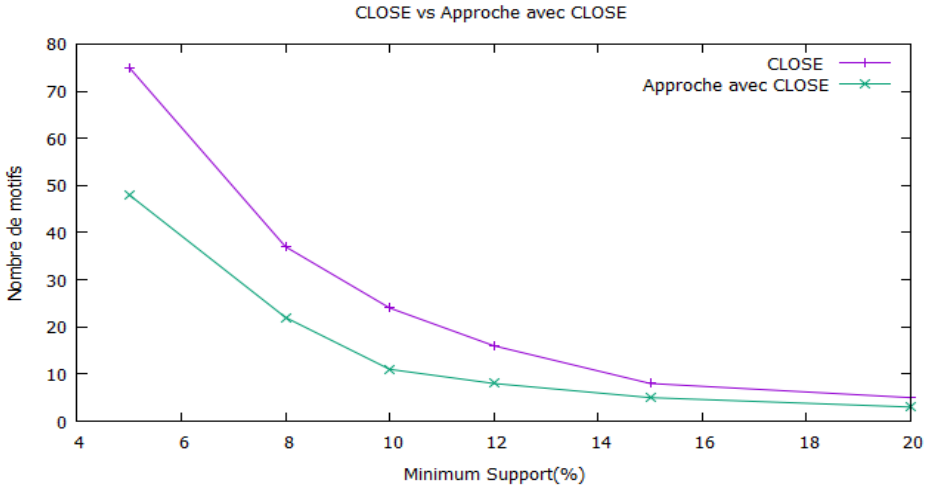


Figure 7. Nombre de motifs extraits avec l'algorithme CLOSE et notre Approche avec CLOSE

Le résultat obtenu va améliorer le post-traitement qui consiste à enlever certains motifs et à converser les plus utiles. Par exemple, on peut observer sur la figure 6 que le nombre de motifs fréquents passe de 195 motifs à 110 motifs fréquents extraits avec *minusup*=5%. L'approche a enlevé 85 motifs qui peuvent être obtenus en utilisant l'ontologie. Si l'expert devrait mettre une seconde par exemple, pour étudier l'utilité d'un motif alors cela correspond à une diminution de la charge de travail de l'expert de 85 secondes.

6. Conclusion

Cet article a présenté une approche de découverte de motifs fréquents en intégrant une phase d'élagage sémantique dans le processus de découverte. Cette phase d'élagage utilise une ontologie comme un support d'élagage pour enlever certains motifs candidats du calcul. Notre approche aide l'expert dans la découverte de motifs fréquents utiles en enlevant les motifs qui peuvent être retrouvés en utilisant l'ontologie de domaine associée aux données de la fouille. Les expérimentations réalisées montrent que l'utilisation de l'ontologie comme un support d'élagage diminue le nombre de motifs extraits. De nombreuses perspectives s'offrent à la suite de nos travaux. La première d'entre elles est d'évaluer notre approche avec une base de connaissances comme DBpedia 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.

Remerciements :

Ce travail est fait pendant notre séjour de recherche (2016) à l'UGB et partiellement financé par le Bureau Afrique de l'Ouest de l'Agence Universitaire de la Francophonie (AUF) dans le cadre du programme « Horizons Francophones » et par le Centre d'Excellence en Mathématiques, informatique et TIC (CEA-MITIC) de l'UGB. Nous remercions l'AUF et le CEA-MITIC pour avoir contribué à la réalisation de ce travail.

7. Bibliographie

[1] Agrawal R. and Srikant R. (1994). Fast algorithms for mining association rules in larges databases, *Proc. VLDB conf.,pp 478-499,September 1994*

[2] Antunes, C. (2007). 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., (1998). « 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] Deng, Z. H., Wang, Z. H., & Jiang, J. J. (2012). A new algorithm for fast mining frequent itemsets using n-lists. *Science China Information Sciences,* September 2012, Volume 55, issue 9, pp 2008–2030

[6] Deng, Z. H. (2015). PrePost^{+}: An efficient N-lists-based algorithm for mining frequent itemsets via Children–Parent Equivalence pruning. *Expert Systems with Applications*, Volume 42,Issue 13, 1 August 2015, Pages 5424–5432

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

[8] Gennari J. H., S. W. Tu, T. E. Rothenfluh et M. A. Musen (1994). Mapping domains to methods in support of reuse. *International Journal of Human-Compute Studies*,41:399–424, 1994

[9] G Grahne, J Zhu (2003). High performance mining of maximal frequent itemsets - *6th International Workshop on High Performance Data*

[10] Gruber, T.R., 1993a. Toward Principles for the Design of Ontologies Used for Knowledge Sharing, in: International Journal of Human-Computer Studies - *Special issue: the role of formal ontology in the information technology*, Volume 43 Issue 5-6, PP. 907 - 928

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

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

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

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

[15] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L. (1999). Efficient Mining of Association Rules using Closed Itemset Lattices. Information Systems, Elsevier Science, 24(1), pages 25-46

[16] Studer, R., Benjamins, V.R., Fensel, D., 1998. Knowledge Engineering: Principles and Methods. Data Knowl Eng 25, 161–197. doi:10.1016/S0169-023X(97)00056-6

[17] Winkler W. E., (1999) The state of record linkage and current research problems, *Statistics of Income Division, Internal Revenue Service Publication R99/04*,1999

[18] Yaya Traoré, Sadouanouan Malo, Cheikh Talibouya Diop, Moussa Lo, Stanislas Ouaro (2014). Extraction des connaissances dans un wiki sémantique : apport des ontologies dans le prétraitement,*5th Journées Francophones sur les Ontologies (JFO), pp.127-138,14-16 Nov. 2014,* Hammamet, Tunisie

[19] Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE TKDE Journal, 12(3), 372–390*

[20] Zaki, M. J. and Ching-Jui Hsiao. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. *In 2nd {SIAM} International Conference on Data Mining. Apr 2002*

[21] Zaki, M. J. & Gouda, K. (2003). Fast vertical mining using diffsets. *In: SIGKDD, pp. 326–335*

[22] Zemmouri El Moukhtar (2013). Représentation et gestion des connaissances dans un processus d'Extraction de Connaissances à partir de Données multi-points de vue, Thèse de doctorat, Ecole Nationale Supérieur d'Arts et Métiers – Meknès

[23] Z. Nazeri and E. Bloedorn. (2004). Exploiting available domain knowledge to improve mining aviation safety and network security data. In P. Buitelaar, J. Franke, M. Grobelnik, G. Paass, and V. Svatek, editors, *Proceedings of the Workshop on Knowledge Discovery and Ontologies at ECML/PKDD'04,* Pisa, Italy, September 2004