Motifs fréquents pour améliorer la catégorisation dans un wiki sémantique

Yaya Traore[1,2], **Cheikh Talibouya Diop**[1,] **Fatou Kamara-Sangare**[1], **Sadouanouan Malo**[3], **Moussa Lo**[1], **Stanislas Ouaro**[2]

1. *Université Gaston Berger de Saint -Louis, Saint-Louis, BP 234, Sénégal*
   {cheikh-talibouya.diop, fatou.kamara, moussa.lo}@ugb.edu.sn
2. *Université* Ouaga *1 Pr Joseph Ki-Zerbo, Ouagadougou*
   BP 7021, Burkina Faso
   {yaytra, ouaro}@yahoo.fr
3. *Université Polytechnique de Bobo Dioulasso, Bobo-Dioulasso,*
   BP 1091, Burkina Faso
   sadouanouan@yahoo.fr

RÉSUMÉ. Les wikis sémantiques sont des sites collaboratifs qui permettent aux utilisateurs d'une communauté de créer et de partager des connaissances. Les pages du wiki sont sémantiquement annotées et des tags (mots-clés) peuvent être associés librement à celles-ci. Les catégories permettent d'organiser les liens entre les pages dans le wiki. Elles sont créées par les experts. Cette manière d'utiliser le wiki permet de stocker une quantité énorme de données sémantiques sur les pages du wiki. Ces données sémantiques peuvent être exploitées et des nouvelles connaissances peuvent être extraites pour améliorer l'organisation du contenu du wiki. Dans ce papier, nous proposons une approche qui permet d'extraire parmi les tags des motifs fréquents de tags utiles pour guider la découverte de nouvelles catégories et améliorer la catégorisation du contenu du wiki. Nous utilisons l'ontologie associée au wiki pour bénéficier de plus d'informations structurées pour guider l'expert dans la création de ces nouvelles catégories dans le wiki. Les pages taggées par ces nouvelles catégories sont automatiquement catégorisées dans le wiki. L'originalité de l'approche consiste à construire un contexte d'extraction à partir du contenu du wiki et d'extraire des motifs fréquents de tags utiles pour améliorer la catégorisation du wiki. Les expérimentations réalisées sur un wiki sémantique dont les pages sont annotées montrent que la méthode permet d'améliorer la catégorisation du contenu du wiki et la recherche sémantique par catégorie.

ABSTRACT. Semantic wikis allow collaboration between users of a community for creating and sharing knowledge. The wiki pages are semantically annotated and tags (keywords) can be freely associated with them. The categories organize links between pages in the wiki and are created by the experts. In this way, data stored in semantic wikis can be extracted and the resulting knowledge pattern can be reused in order to improve the wiki organization. In this

paper, we propose a method which allows to extract frequent patterns useful of tags in the wiki for guiding new categories discovery and improving categorization. We use an external ontology of the wiki for guiding the experts to create these new categories in the wiki. Pages annotated by these new categories are categorized. The originality of our approach is to build from the content of wiki an extraction context and to extract knowledge units from this content. This will allow to improve the hierarchy category and improve the semantic research by categories. The experiments on a semantic wiki annotated show the substantial results.

MOTS-CLÉS : Wiki sémantique, motifs fréquents, ontologie, catégorisation.

KEYWORDS: Semantic wiki, frequent pattern, ontology, categorization.

DOI:10.3166/ISI.21.5-6.83-106 © 2016 Lavoisier

1. Introduction

Les wikis sémantiques sont des wikis qui utilisent les technologies du web sémantique pour permettre d'annoter les pages du wiki (Krötzsch *et al.*, 2007a). Cette annotation consiste à créer des catégories pour regrouper les pages, créer des propriétés qui servent à préciser les liens sémantiques entre les pages. Egalement les utilisateurs peuvent annoter les pages en associant un ensemble de tags avec celles-ci. Dans ce travail, nous utilisons un wiki sémantique développé autour du moteur Semantic MediaWiki (Krötzsch *et al.*, 2006). Ce wiki permet aux communautés de produire en collaboration une quantité immense d'annotation que les utilisateurs peuvent facilement exploiter grâce à des liens sémantiques et une base de connaissances formalisées. Dans la tendance du web sémantique, en profitant de la représentation des connaissances avec les ontologies, ces annotations dans les wikis sémantiques sont représentées au sein du schéma RDF et peuvent être interrogées en utilisant le langage SPARQL.

Les données sémantiques produites par les pages tagguées sont rarement utilisées pour réorganiser le contenu du wiki sémantique. Cependant, les tags stockés dans le wiki sémantique peuvent être extraits et les nouvelles connaissances qui en résultent peuvent être réutilisées dans le but d'améliorer la catégorisation du wiki. Cette catégorisation consiste à créer de nouvelles catégories ou sous-catégories parmi les tags. Les pages tagguées par ces nouvelles catégories sont automatiquement classées. Les techniques de découverte de connaissances sont des candidats pour extraire ces nouvelles connaissances cachées dans le contenu du wiki. Le but de ce travail est d'utiliser une technique de découverte de connaissances basée sur la fouille de motifs fréquents pour extraire ces nouvelles connaissances stockées parmi les tags et qui ne sont pas des catégories existantes. Les motifs fréquents de tags ainsi extraits sont utilisés pour améliorer la catégorisation dans le wiki.

Le reste de l'article est organisé comme suit : la section 2 présente les préliminaires et l'énoncé du problème. La section 3 présente les travaux liés à notre approche. Dans la section 4, nous développons l'approche proposée pour améliorer la catégorisation dans le wiki sémantique. Enfin, nous présentons les résultats

expérimentaux de notre approche dans la section 5. Nous terminons par une conclusion et des perspectives dans la section 6.

2. Préliminaires et position du problème

2.1. Terminologie du wiki sémantique

Le wiki est essentiellement défini par un ensemble de catégories, de pages et de propriétés. Chaque page du wiki est classée dans une ou plusieurs catégories et est associée à des tags. Une catégorie définit une thématique. Elle est destinée à regrouper les pages sur des sujets similaires dans le wiki. Les catégories sont organisées dans une hiérarchie. Dans notre contexte, nous utilisons une ontologie externe associée au wiki dont les concepts sont utilisés par les experts pour créer de nouvelles catégories dans le wiki. Un tag décrit une caractéristique particulière d'une page. Un tag peut décrire une catégorie dans le wiki et dans ce cas, le tag et la catégorie sont identiques. Les tags sont fournis par les utilisateurs et sont stockés sur les pages du wiki. Les experts peuvent utiliser ces tags pour réorganiser les liens entre les pages du wiki. Par exemple, la figure 1a montre un exemple du contenu d'un wiki sémantique. La partie (1) présente la hiérarchie des catégories, la partie (2) présente les pages créées et la partie (3) présente les tags associés aux pages.

Chaque page du wiki est considérée comme un élément d'une ontologie (Krötzsch *et al.*, 2007b). (Krötzsch *et al.*, 2007b) expliquent comment les catégories, les propriétés et les pages d'un wiki peuvent être utilisées respectivement comme concepts, propriétés et instances d'une ontologie. Selon le schéma défini par MediaWiki[1](SMW), une catégorie est définie comme une *owl:Class* et une page est définie comme une instance définie par SWIVT ontologie (Krötzsch *et al.*, 2012), qui fournit une base pour l'interprétation des données sémantiques de SMW. Par exemple, dans SWIVT ontologie (Krötzsch *et al.*, 2012), la page est définie comme une *swivt:page* et un tag est défini comme une *propriété:Tag*. Ainsi, les données sémantiques de SMW peuvent être stockées dans un triplestore RDF. On peut utiliser des requêtes SPARQL sur ce triplestore pour obtenir, par exemple, la liste des catégories, sous-catégories, pages et tags.

La figure 1b présente une ontologie externe associée au wiki (a). Les catégories *BP*, *CC* sont respectivement les concepts *« Biological Process »* et *« Cellular component »* de l'ontologie. Les sous-catégories *A*, *B* représentent respectivement les sous concepts *Apoptsis*, *Immunity* de l'ontologie. Les tags *A*, *B*, *C*, *D*, *E*, *F*, *G*, *Ca* sont des sous-concepts des concepts *« Biological Process »* et *« Cellular component » et Ligand.*

Formellement le wiki sémantique est défini par $\mathrm{WS} = (\mathrm{C}, \subseteq, \mathrm{P}, \mathrm{T}, R_C, R_T)$:

– C représente l'ensemble des catégories ;

1. http://semanticweb.org/wiki/SMW

– ⊆ représente la relation de « sous-catégorie » entre les catégories de C ;

– P représente l'ensemble des pages ;

– T représente l'ensemble des tags ;

– R_C représente une relation binaire entre C et P. Un couple $(p, c) \in R_C$ dénote le fait que la page $p \in P$ est catégorisée par la catégorie $c \in C$,

– R_T représente une relation binaire entre P et T. Un couple $(p, t) \in R_T$ dénote le fait que la page $p \in P$ est taguée par le tag $t \in T$.

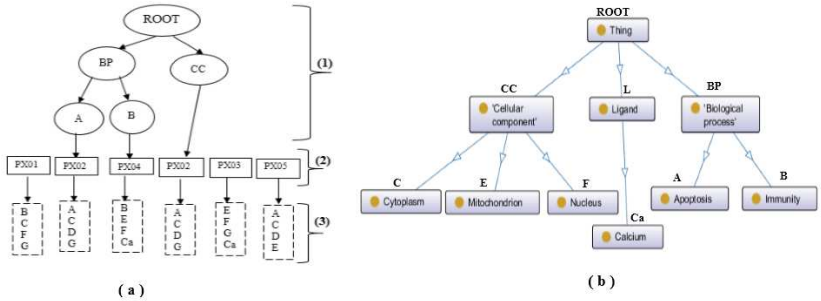

Figure 1. (a) Exemple du contenu d'un wiki : 1) la hiérarchie des catégories 2) les pages du wiki ; 3) les tags associés aux pages. (b) Ontologie externe associée au wiki (a)

2.2. Motifs fréquents

2.2.1. Contexte d'extraction

Un contexte d'extraction est un triplet $CE = (P, T, R_T)$ où P représente l'ensemble fini des pages du wiki sémantique, T l'ensemble fini des tags, R_T une relation binaire entre T et P tel que R_T (p, t) = 1 si la page p ϵ P est tagguée par t ϵ T sinon 0. Nous définissons la fonction g qui permet d'avoir l'ensemble des pages associées à un tag comme suit : $g : T \rightarrow P$ tel que pour tout $t \epsilon T, g(t) = \{p / p \in P$ et R_T (p, t) = 1$\}$. Dans la suite nous représentons un contexte d'extraction par un tableau pages/tags comme illustré par le tableau 1.

Tableau 1. Contexte d'extraction

Tags / Pages	t_1	…	t_i	…	t_n
p_1	$R_T(p_1, t_1)$	…	$R_T(p_1, t_i)$	…	$R_T(p_1, t_n)$
…	…	…	…	…	…
p_i	$R_T(p_i, t_1)$	…	$R_T(p_i, t_i)$	…	$R_T(p_i, t_n)$
…	…	…		…	…
p_n	$R_T(p_n, t_1)$	…	$R_T(p_n, t_i)$	…	$R_T(p_n, t_n)$

2.2.2. Motifs fréquents de tags

Un motif de tags est un sous ensemble de tags. Le support d'un motif de tags est la proportion de pages annotées par ce sous ensemble de motif. Soit $T_1 \subseteq T$ un motif de tags. Le support du motif T_1 est obtenu par la formule suivante :

$$Support\ (T_1) = \frac{|\ g(T_1)|}{|P|} \quad (1)$$

Dans la formule (1) : $g\ (T_1) = \cap_{t \in T_1}\ g(t)$ et $|\ g(T_1)|$ donne le nombre de pages tagguées par T_1 dans le wiki et $|\ P\ |$ donne le nombre total des pages du wiki. Soit *minsup* le seuil de support minimal. Un motif T_1 est fréquent si son support est supérieur au seuil de support minimal *minsup* fixé *: Support* (T_1)*>=minsup*.

Motifs fréquents de tags utiles. Soit C l'ensemble des catégories du wiki et f un motif fréquent de tags : f est un motif fréquent de tags utile dans le wiki si et seulement si : $\forall t\ \epsilon\ f \implies t$ n'est pas un élément de C. Les motifs fréquents de tags utiles sont les nouvelles connaissances fournies par les utilisateurs et stockées sur les pages. L'extraction de ces nouvelles connaissances permettra aux experts d'identifier de nouvelles catégories parmi ces tags stockés sur les pages et ainsi réorganiser le contenu du wiki.

2.3. Position du problème

Dans un wiki, toutes les catégories sont créées manuellement par les experts du wiki et tous les tags sont fournis par les utilisateurs du wiki. Le fait que le nombre de pages ne cesse de croître et que de nouveaux tags sont fournis par les utilisateurs est un défi majeur pour la gestion des catégories et la catégorisation des pages. L'énoncé du problème pour ce travail est le suivant : Comment améliorer la catégorisation du wiki en utilisant les tags incorporés dans les pages par les utilisateurs afin d'améliorer par exemple, la recherche sémantique par catégorie ?

Pour améliorer la navigation et la recherche sémantique dans le wiki, la hiérarchie des catégories doit être mise à jour périodiquement. Ainsi, nous proposons une approche qui utilise les motifs fréquents de tags utiles pour guider la création de nouvelles catégories et qui permet de catégoriser les pages de ces catégories. Dans ce qui suit, à la section 4, nous décrivons la méthodologie qui utilise les motifs fréquents de tags utiles pour améliorer la catégorisation du wiki.

3. Travaux liés

Les wikis sémantiques sont des outils qui produisent des données structurées et non structurées. Pour réduire les données non structurées et améliorer la structuration des données du wiki, (Boyer *et al.,* 2010) proposent un système qui suggère des annotations automatiquement aux utilisateurs. Cette méthode permet de

contrôler la création des pages mais elle ne donne pas la possibilité à l'utilisateur d'associer des tags à une page pour la décrire selon sa compréhension.

(Chernov *et al.*, 2006) proposent une méthode qui extrait des informations sémantiques de Wikipédia en analysant les liens entre les catégories. Les informations sémantiques obtenues sont utilisées pour améliorer les capacités de recherche dans Wikipédia et fournir aux contributeurs des suggestions significatives pour éditer les pages de Wikipédia. Cette méthode améliore la recherche par catégorie en définissant les liens sémantiques entre catégories mais elle ne permet pas de créer de nouvelles catégories et de catégoriser les pages dans le wiki.

(Shi *et al.*, 2011) proposent une méthode qui guide la réingénierie du wiki sémantique. Cette méthode est basée sur l'analyse des concepts formels (ACF) et son extension l'analyse relationnelle des concepts (ARC). La méthode permet de définir de nouvelles catégories et les relations entre les catégories. L'approche est intéressante mais elle ne montre pas de manière explicite comment une nouvelle catégorie est insérée dans le wiki et comment les pages sont catégorisées après cette insertion.

(Rosenfeld *et al.*, 2010) proposent une approche qui permet de détecter les problèmes d'incohérence et de redondance dans les wikis sémantiques et assiste les utilisateurs dans la maintenance du wiki. Cette maintenance consiste par exemple à fusionner deux catégories en une seule, définir des relations de sous-catégories, renommer une catégorie existante, modifier les annotations incorrectes des ressources. Comme (Shi *et al.*, 2011), cette approche fait de la réingénierie du wiki. Elle est intéressante mais ne présente pas de manière explicite comment une catégorie est ajoutée dans le wiki et comment une page est classée après une maintenance. Aussi cette approche ne prend pas en compte les tags fournis par les utilisateurs et stockés sur les pages.

En s'inspirant des travaux de (Shi *et al.*, 2011 ; Rosenfeld *et al.*, 2010 ; Gil *et al.*, 2013) et de nos travaux précédents (Traoré *et al.*, 2015) sur la découverte de nouvelles catégories dans un wiki sémantique, nous proposons une approche pour améliorer la catégorisation basée sur les motifs fréquents de tags utiles.

De très nombreux algorithmes ont été développés pour résoudre le problème de recherche de motifs fréquents (Agrawal et Srikant, 1994 ; Pasquier *et al.*, 1998 ; 1999 ; Zaki, 2000 ; Zaki *et al.*, 2002 ; 2003 ; Grahne et Zhu, 2003 ; Deng *et al.*, 2012 ; Deng et Lv., 2015). Un ensemble d'algorithme de fouille de motifs fréquents implémentés en Java est disponible dans (Fournier-Viger, 2016). Les travaux de (Tobias et Andreas, 2011), proposent une extension de Semantic Mediawiki pour extraire des motifs fréquents de nuages de tags à partir d'une propriété, mais cette extension ne permet pas de détecter de nouvelles catégories. Comme (Tobias et Andreas, 2011), nous utilisons l'algorithme Apriori (Agrawal et Srikant, 1994). Nous insérons dans cet algorithme une phase d'élagage sémantique qui supprime les motifs de tags dont les éléments correspondent à une catégorie ou une sous-

catégorie existante dans le wiki. Ainsi tous les motifs fréquents de tags utiles sont extraits et sont analysés pour identifier des nouvelles catégories.

Pour insérer les nouvelles catégories trouvées, nous utilisons une ontologie externe. Cette ontologie décrit les connaissances du domaine du wiki. Toutes les catégories ou sous-catégories du wiki existent comme concepts ou sous-concepts dans cette ontologie. Plusieurs travaux dans la littérature ont montré l'intérêt de réutiliser une ontologie pour contrôler le vocabulaire d'un domaine. Parmi ces travaux nous nous intéressons aux travaux de (Gennari *et al*., 1994), qui explique la réutilisation d'une ontologie et les travaux de (Buffa *et al*., 2007 ; 2008 ; Meilender, 2013) qui montrent comment une ontologie peut être utilisée avec un wiki sémantique. Dans ce cas, il est nécessaire de définir une correspondance entre les éléments du wiki et de ceux de l'ontologie. En s'inspirant des travaux de (Hernandez, 2006) sur l'utilisation d'une ontologie de domaine pour la modélisation du contexte en recherche d'information, des travaux de (Marinica *et al.*, 2008) sur la fouille de règles d'association guidée par des ontologies et des schémas de règles, des travaux de (Schönberg *et al.*, 2010) et des travaux (Filipiak et Ławrynowicz, 2014) qui génèrent automatiquement à partir d'une ontologie le contenu de Semantic Mediawiki, nous proposons d'utiliser l'ontologie externe pour identifier la place de chaque nouvelle catégorie. Si la place est détectée dans l'ontologie nous l'insérons dans le wiki en utilisant la correspondance entre les concepts de l'ontologie et les catégories dans le wiki. Ainsi, nous proposons un algorithme qui effectue cette tâche et qui catégorise les pages après insertion d'une catégorie dans le wiki.

4. Approche et méthodologie proposée

Dans cette section, nous donnons des détails sur la méthodologie (figure 2) utilisée pour améliorer la catégorisation du wiki. La méthode est composée de trois étapes :

– Etape 1 : cette étape sélectionne l'ensemble des tags du contenu du wiki pour construire un contexte d'extraction. Après nous utilisons l'algorithme basé sur Apriori (Agrawal et Srikant, 1994) où nous avons inséré une phase d'élagage sémantique qui supprime les motifs de tags dont les éléments correspondent à une catégorie ou une sous-catégorie existante. Les motifs fréquents obtenus sont les motifs fréquents de tags utiles.

– Etape 2 : Les motifs fréquents de tags utiles permettent de générer des règles d'association pertinentes dont les conséquents sont identifiés comme les nouvelles catégories. Nous proposons un algorithme qui insère la nouvelle catégorie ou sous-catégorie dans la hiérarchie des catégories existantes. Cet algorithme utilise une ontologie externe du wiki qui contient toutes les catégories ou sous-catégories correspondantes du wiki.

– Etape 3 : la dernière étape consiste à classer les pages dans les catégories ou sous-catégories nouvellement définies dans le wiki et supprimer les liens de sur-

catégorisation. La sur-catégorisation est le fait de classer une même page dans une catégorie et ses sous-catégories.

Le reste de la section est organisé comme suit. Tout d'abord, nous expliquons comment les données sont récupérées et comment nous construisons le contexte d'extraction. Ensuite, nous présentons l'algorithme basé sur Apriori, conformément à l'approche pour extraire les motifs fréquents de tags utiles. Enfin, nous proposons un algorithme pour la nouvelle catégorisation dans le wiki.

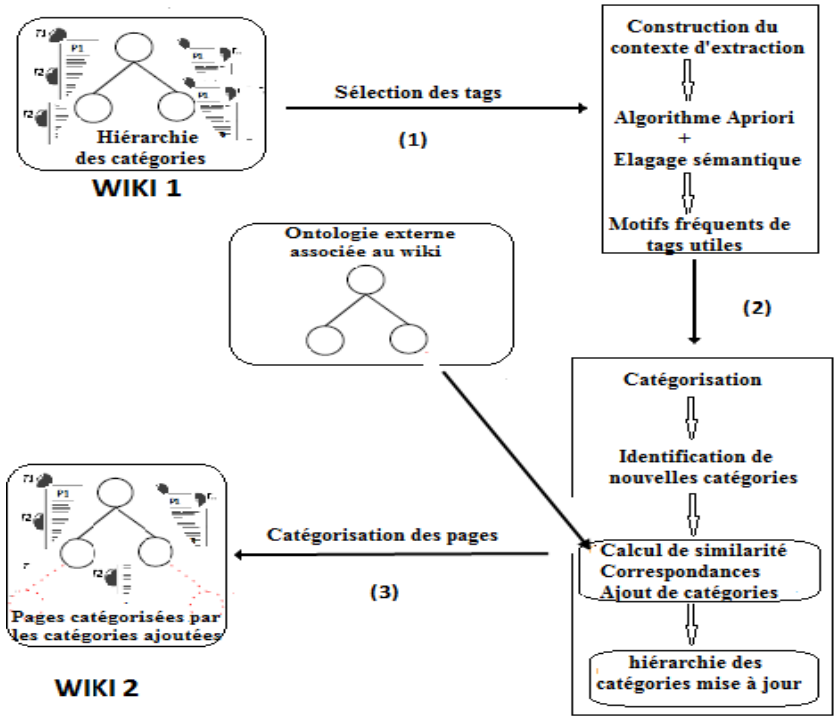

Figure 2. Méthodologie pour améliorer l'organisation du wiki guidée par les motifs fréquents de tags utiles

4.1. Construction du contexte d'extraction

Dans un premier temps, nous exportons l'ensemble des pages du wiki en RDF pour créer une base de connaissances du wiki. Soit KB cette base de connaissances. SWIVT ontologie (Krötzsch *et al.*, 2012) fournit une base pour l'interprétation de données sémantiques exportées par Semantic MediaWiki. Après l'obtention de la

base de connaissances en RDF, nous la chargeons dans virtuoso[2] et nous sélectionnons l'ensemble des catégories (*owl:Class*), l'ensemble des pages (*swivt:page*) et l'ensemble des tags *(property:Tag*) avec des requêtes SPARQL sur KB.

Nous proposons *l'Algorithme 1* pour construire le contexte d'extraction. Dans cet algorithme, nous avons en entrée la base de connaissances du wiki en RDF. L'algorithme utilise des requêtes SPARQL pour sélectionner l'ensemble des tags (ligne N° 2) et l'ensemble des pages (ligne N° 4). A la ligne N° 7, l'algorithme utilise la fonction g (voir la section 2.2.1 pour la définition) pour obtenir toutes les pages associées à un tag puis construit le contexte d'extraction.

Algorithme 1. Construction du contexte d'extraction

Entrée :
KB= Base de connaissances du wiki en RDF
Sortie :
CE : Contexte d'extraction
Début
1: *// Extraction des tags par des requête sparql*
2: *T=ensemble des tags extraits de KB avec sparql*
3: *// Extraction des pages par des requête sparql*
4: *P=ensemble des pages extraites de KB avec sparql*
5: *// Constuction du contexte d'extraction CE*
6: **Pour chaque** *page* $p \in P$ **faire**
7: **Pour chaque** *tag* $t \in T$ **faire**
8: **Si** $(p \in g(t))$ **alors**
9: *CE(p, t)=1*
10: **Sinon**
11: *CE(p, t)=0*
12: **FinSi**
13: **Finpour**
14: **Finpour**
15: *Retourner CE*
Fin

Exemple : En déroulant l'algorithme 1 sur le wiki (figure 1a), on construit le contexte d'extraction ci-après (tableau 2).

2. http://virtuoso.openlinksw.com/

Tableau 2. Exemple de contexte d'extraction du wiki (figure 1a)

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

4.2. Découverte de motifs fréquents de tags utiles

De nombreux algorithmes[3] ont été proposés pour la découverte des motifs fréquents. Dans cet article, nous nous intéressons à l'algorithme Apriori (Agrawal et Srikant, 1994). L'extraction des motifs fréquents dans Apriori (Agrawal et Srikant, 1994) consiste à parcourir itérativement par niveau 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.

Nous introduisons dans cet algorithme une phase d'élagage sémantique (figure 3) des motifs candidats. L'élagage sémantique (phase (1)) élimine du calcul des motifs fréquents tout candidat dont les éléments correspondent à une catégorie ou une sous-catégorie existante dans le wiki. La phase (2) calcule le support des motifs candidats qui n'ont pas été éliminés lors de la phase d'élagage sémantique. Les motifs obtenus sont des motifs fréquents de tags utiles.

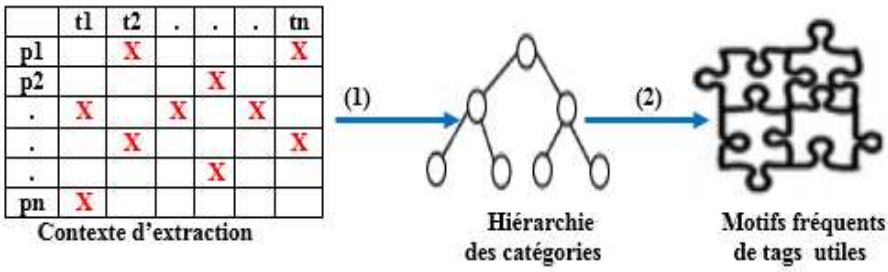

Figure 3. Approche de découverte de motifs fréquents de tags utiles dans le wiki

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

Nous utilisons *l'Algorithme 2* pour effectuer cette tâche. Nous insérons à la ligne N° 9, l'étape d'élagage sémantique qui supprime les motifs de tags qui ne sont pas utiles, par exemple, si $\boldsymbol{f}$ est un motif de tags candidat alors $\boldsymbol{f}$ est un motif utile ssi : $\forall \boldsymbol{t} \in \boldsymbol{f} \Rightarrow \boldsymbol{t}$ ne correspond pas à une classe ou une sous-classe de KB. Ainsi, les motifs fréquents de tags retournés à la ligne N°15, sont les motifs fréquents de tags utiles dans le wiki.

Exemple : Pour un seuil minimal *minsup=40 %*, les motifs fréquents de tags utiles extraits à partir du contexte d'extraction (tableau 2) utilisant l'Algorithme 2 sont :

1) motifs fréquents de tags utiles : {C,D,E,F,G,Ca};

2) motifs fréquents de tags utiles : {C D, C G, E F, E Ca, F G, F Ca} ;

3) motifs fréquents de tags utiles : { E F Ca}.

Algorithme 2. Découverte de motifs fréquents de tags utiles

Entrée :
CE : contexte d'extraction, KB : Base de connaissances du wiki,
minsup : seuil minimal de support

Sortie :
F : motifs fréquents de tags 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 :* $C = F1 \cup F2$ *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'élagage sémantique***
9: ***Supprimer de*** C_K ***tout candidat C qui n'est pas un motif de tags utile***
10: *// Phase d'évaluation des candidats*
11: *Calculer le support de chaque candidat C dans* C_K
12: $L_K = \{C \in C_K / Support(C) >= minsup\}$
13: $K=K+1$
14: Fintanque
15: *Retourner* $F=\cup L_K$

Fin

4.3. Catégorisation

L'objectif de ce travail consiste à utiliser les motifs fréquents de tags utiles pour créer de nouvelles catégories et à classer les pages dans ces nouvelles catégories. Tout d'abord, nous utilisons les motifs fréquents de tags utiles pour générer les règles d'associations. Une règle d'association est une expression d'implication de la forme $T_i ==> T_j$, où $T_i \subseteq T$, $T_j \subseteq T$ et $T_i \cap T_j = \emptyset$. Dans cette règle T_i est appelé l'antécédent et T_j le conséquent.

Nous analysons ces règles d'association pour sélectionner les règles pertinentes et importantes. De nombreuses mesures (Vaillant, 2006) ont été proposées pour mesurer la qualité des règles d'association en plus de la mesure de confiance (formule 2) utilisée pour sélectionner ces règles. Parmi ces mesures, nous choisissons la mesure lift (formule 3) (Brin *et al.*, 1997) pour mesurer la qualité des règles. La mesure lift mesure l'amélioration apportée par la règle d'association par rapport au contexte d'extraction où T_i et T_j seraient indépendants. Une grande valeur du lift (valeur de lift supérieure à 1) traduit une forte corrélation entre l'antécédent et le conséquent de la règle. Alors nous utilisons cette mesure pour sélectionner les règles importantes.

$$Confiance\ (T_i ==>\ T_j) = \frac{Support(T_i \cup T_j)}{Support(T_i)} \quad (2)$$

$$lift\ (T_i ==>\ T_j) = \frac{Confiance\ (T_i ==> T_j)}{Support(T_j)} \quad (3)$$

Soit *minlift* le seuil minimum de la mesure lift. Les règles pertinentes et importantes sont celles qui ont une valeur de lift supérieur *à minlift* >1. Enfin, les règles d'association pertinentes et importantes obtenues sont utilisées pour définir la règle de catégorisation. Cette tâche est exécutée par la procédure *NouvellesCategories (Algorithme 3.1).*

Règle de Catégorisation. Soit une règle $r_i : T_i ==> T_j$ une règle, $p \epsilon P$ une page du wiki, $minconf$ le seuil minimum de confiance et *minlift* le seuil minimum de la mesure lift. Nous définissons la règle de catégorisation comme suit :

– r_i est une règle d'association *: Confiance (* r_i *)* $> minconf$;

– si ($|T_j| = 1\ et\ lift(r_i) > minlift$) alors T_j est une nouvelle catégorie ;

– $si\ p\ est\ tagguée\ \ par\ le\ tag\ T_j\ alors\ p\ est\ catégorisée\ par\ T_j$.

La règle de catégorisation permet d'identifier toutes les nouvelles catégories (*Algorithme 3.1)* à partir des règles d'association qui satisfont la condition : $|T_j| = 1\ et\ lift(r_i) > minlift$. En effet si : $|T_j| = 1$ alors T_j est constituée d'un seul tag. La valeur de $lift(r_i) > minlift$ traduit une forte corrélation entre T_i et T_j. Cela traduit que toutes les pages tagguées par T_i sont associées à T_j. On peut donc définir le tag T_j comme une nouvelle catégorie et toutes les pages tagguées par T_i sont catégorisées par T_j : ($\forall\ p\ \in g(T_i)$ ou $\forall\ p\ \in g(T_j)$) alors $R_C(p, T_j) = 1$

Ainsi, tous les conséquents des règles d'association pertinentes et importantes sont identifiés comme des nouvelles catégories qu'il faut ajouter dans le wiki.

4.3.1. Mise à jour de la hiérarchie des catégories

4.3.1.1. Identification des nouvelles catégories

Nous proposons la procédure *NouvelleCategories* (*Algorithme 3.1.*) pour identifier les nouvelles catégories. Cette procédure utilise en entrée les motifs fréquents de tags utiles et pour chaque motif, la procédure *ap_genrules* (ligne N° 8) est utilisée pour générer les règles d'association dont la confiance est supérieure au seuil minimal de confiance $minconf$. Pour chaque règle générée, la procédure détermine la taille du conséquent de la règle R_{cons} et la valeur de la mesure lift. Si $|R_{cons}| = 1$ et lift(R) $> minlift$ alors le conséquent R_{cons} de cette règle est ajouté à la liste des nouvelles catégories (ligne N° 14).

Algorithme 3.1. Nouvellecategories : permet identifier les nouvelles catégories

Entrée :
F: Motifs fréquents de tags utiles, $minconf$: seuil minimal de confiance
$minlift$: seuil minimal de lift
Sortie :
CR : Liste des nouvelles catégories
Début
1: *K=2*
2: *$CR=\emptyset$*
3: *// R = une règle d'association*
4: *//f_k est un k-motif fréquent de tags utiles*
5: Pour chaque $f_k \in F$ **faire**
6: *// H_1 =Candidat à une règle d'association*
7: *H_1={t/ t ∈ f_k} // 1-tag consequents de règle.*
8: R=ap_genrules(f_k, H_1, minconf)
9: *//R_{cons}=consequent de la règle d'association R*
10: *R_{cons} = Consequent(R)*
11: *//vlift= valeur de la mesure lift de la règle*
12: *vlift=Confiance(R)/Support(R_{cons})*
13: **Si** *($|R_{cons}| == 1$ et $R_{cons} \notin CR$) et (vlift > minlift)* **alors**
14: $CR = CR \cup \{R_{cons}\}$
15: FinSi
16: Fin pour
17: *Retourner CR*
Fin.

4.3.1.2. Ajout d'une nouvelle catégorie dans le wiki

Pour ajouter ces nouvelles catégories trouvées dans la hiérarchie des catégories, nous utilisons une ontologie externe qui définit les connaissances du domaine du wiki. Cette ontologie permettra de faire la mise à jour de la hiérarchie des catégories en évaluant la similarité sémantique entre chaque nouvelle catégorie et l'ensemble

des concepts de l'ontologie. La mesure de similarité sémantique utilisée est celle de JaroWankler (Winkler, 1999) qui est une métrique bien adaptée pour la similarité entre deux chaînes de caractères courtes. En particulier, la fonction *SimilaryJaroWankler* calcule la similarité entre un concept et la nouvelle catégorie. Elle évalue pour un concept, la similarité entre chaque terme qui le désigne et la nouvelle catégorie. Si un terme similaire est trouvé alors le concept est similaire à la nouvelle catégorie. La procédure *MajHierarchie (Algorithme 3.2.)* est utilisée pour exécuter cette tâche. Cette procédure reçoit en entrée la base de connaissances du wiki, l'ontologie externe, une nouvelle catégorie et un seuil de similarité. A la ligne N° 8, la procédure utilise la fonction *SimilaryJaroWankler* pour calculer la similarité entre chaque concept et la nouvelle catégorie. Si le concept similaire C_i est trouvé alors la procédure récupère son parent dans C_h. La procédure cherche alors la correspondance du concept C_h avec les catégories dans le wiki. A la ligne N° 18, si la catégorie Cat du wiki correspond à C_h, la catégorie R_{cons} est créée comme une sous-catégorie de Cat. Autrement C_h est créée comme une sous-catégorie du ROOT et R_{cons} une sous-catégorie de C_h. Au cas où la nouvelle catégorie R_{cons} n'est pas trouvée dans l'ontologie externe alors elle est ajoutée au ROOT comme une sous-catégorie (ligne N° 26).

Algorithme 3.2. MajHierarchie : permet d'ajouter une nouvelle catégorie dans le wiki

Entrée :
R_{cons} : nouvelle categorie, KB : Base de connaissances du wiki, O : Ontologie externe
θ : seuil de similarité
Sortie :
Hiérarchie des catégories mise à jour
Début
1: *//C=ensemble des concepts de O*
2: *K=Faux // K=Vrai signifie que R_{cons} est une sous-catégorie*
3: *X= θ // X est initialisé à la valeur du seuil de similarité*
4: **Pour chaque** *concepts $C_i \in C$* **faire**
5: *SC=Ensembles des sous concepts de C_i*
6: *SC=SC ∪ {C_i} // Prendre en compte le concept C_i*
7: **Pour chaque concept** *$sc_i \in SC$* **faire**
8: *sim=SimilaryJaroWankler(sci, R_{cons})*
9: **Si** *sim>= X* **alors**
10: *X =sim*
11: *K=Vrai*
12: *C_h=C_i //C_h = concept parent de R_{cons}*
13: **FinSI**
14: **FinPour**
15: **FinPour**
16: **Si** *(K== $Vrai$)* **alors**
17: *Cat= Catégorie correspondante à C_h dans le wiki*
18: **Si** *Cat existe* **alors**
19: *Ajouter R_{cons} comme sous-catégories de Cat*

20: *Supprimer les liens de sur-categorisation avec Cat pour toutes les pages p où* $R_C(p,Cat)$=1 and $R_T(p,R_{cons})$=1
21: **Sinon**
22: *Ajouter* C_h *comme sous-catégorie du ROOT*
23: *Ajouter* R_{cons} *comme sous-catégorie de* C_h
24: **FinSi**
25: **Sinon**
26: *Ajouter* R_{cons} *comme sous-catégorie du ROOT*
27: **FinSi**
Fin

4.3.2. Catégorisation des pages tagguées

La catégorisation de pages consiste à classer les pages dans une ou plusieurs catégories. Nous proposons l'*Algorithme 3.3* pour cette tâche de catégorisation. Cet algorithme reçoit en entrée les motifs fréquents de tags utiles, la base de connaissances du wiki, l'ontologie externe O et les valeurs de seuils minimums de confiance *minconf*, de lift *minlift* et de similarité. Pour chaque motif fréquent de tags utiles obtenus par l'*Algorithme 2*. La procédure *Nouvellecategories (Algorithme 3.1)* est utilisée (ligne N° 2). Pour chaque nouvelle catégorie identifiée, la procédure *MajHierarchie* (Algorithme 3.2) est utilisée (ligne N° 5) pour ajouter la nouvelle catégorie dans le wiki. Enfin, les pages associées à elle sont catégorisées (ligne N° 9).

Algorithme 3.3. Algorithme de catégorisation dans le wiki

Entrée :
F: Motifs fréquents de tags utiles, O*: Ontologie externe associée au wiki,*
θ*: seuil de similarité, minconf : seuil minimum de confiance, KB : Base de connaissances du wiki, minlift: seuil minimum de lift*
Sortie :
Hiérarchie des catégories mise à jour, Pages classées dans les nouvelles catégories
Début
1: *// CR =liste des nouvelles catégories*
2: ***CR=Nouvellecategories (F, minconf, minlift)***
3: *//Ajout des nouvelles catégories dans le wiki*
4: **Pour chaque** *nouvelle catégorie* $R_{cons} \in CR$ **faire**
5: ***MajHierarchy(*** R_{cons} ***,KB, O, θ)***
6: *// g(*R_{cons}*)= ensemble fini des pages associées à* R_{cons}
7: *// Catégorisation des pages tagguées par* R_{cons}
8: **Pour chaque** *page p* $\in g(R_{cons})$ **faire**
9: Classer p dans R_{cons} // $R_C(p, R_{cons})$ =1
10: **Finpour**
11: **FinPour**
Fin

Exemple: Soit l'ontologie de la figure 1b. En déroulant l'algorithme 3.3 sur les motifs fréquents de tags utiles extraits par l'algorithme 2 avec *minconf*=40%, *minlift* = 1,5 et θ = 0,83 on a les résultats suivants :

Les règles d'association générées sont :

– **R1:** D == > **C**; **R2:** Ca == > **E**; **R3:** Ca == > **F** ;

– **R4:** Ca E == > **F**; **R5:** F E == >**Ca**.

En déroulant la procédure *Nouvellecategories* on obtient la liste des nouvelles catégories suivantes : *C, E, F, Ca*

En déroulant la procédure *MajHierarchie* sur chacune des nouvelles catégories avec l'ontologie de la figure 1b, les nouvelles catégories *C, E, F* sont identifiées et placées dans le wiki sous la catégorie *CC.* Pour la nouvelle catégorie *Ca* la procédure a identifié une nouvelle hiérarchie qui commence avec le concept *L*. La figure 4b montre que la catégorisation du wiki (a) a été améliorée après le déroulement de notre méthode. En effectuant une recherche sémantique avec la catégorie CC dans le wiki (a) on trouve une seule page. En effectuant la même recherche dans le wiki (b) quatre pages de plus sont restituées. Ainsi avec ces nouvelles catégories ajoutées, le wiki (b) devient plus intéressant que le wiki (a).

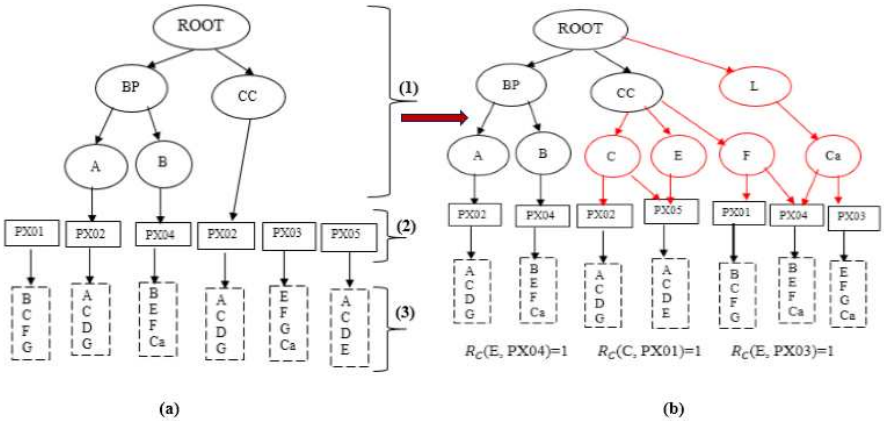

Figure 4. (a) wiki avant notre méthode, (b) Wiki obtenu après notre méthode

5. Expérimentation

Dans cette section nous décrivons les résultats expérimentaux obtenus en appliquant notre approche sur un wiki construit avec les données téléchargées de la base de données biologique Uniprot (www.uniprot.org). Cette base de données contient des données sur les protéines. Ces protéines sont étiquetées avec des mots-clés qui sont classés dans 10 catégories : *« Biological process », « Cellular component », « Coding sequence diversity », « Developmental stage », Disease,*

Domain, Ligand, « Molecular function », « Post-translational modification », « Technical term ».

Dans cette base de données, pour chaque protéine nous nous intéressons à deux colonnes : l'entrée des protéines dans la base de données et les mots-clés.

Ainsi nous avons constitué un ensemble de *129 311 pages* sur les protéines (provenant de l'organisme Thalina). Ces pages sont annotées par *478 mots-clés*. Le reste de la section, est organisé comme suit : dans un premier temps, nous décrivons le wiki sémantique construit à partir de ces données. Ensuite, nous présentons le protocole expérimental et nous terminons par les résultats obtenus après l'application de notre méthode.

5.1. Construction du wiki sémantique

A partir des données téléchargées de la base données de Uniprot, nous construisons un wiki sémantique (tableau 3) avec 10 catégories et 300 sous-catégories parmi les 478 mots-clés (ou tags) associés aux pages. Sur les *129 311 pages, 99 904 sont catégorisées* (tableau 2) *et 29 407 sont non catégorisées*.

Les catégories *« Cellular component »* et *Ligand* sont des catégories inutilisées dans le wiki. Dans la suite du document, nous parlerons de wiki 1 pour désigner ce wiki et de wiki 2 pour désigner le wiki obtenu après l'application de notre méthode.

Tableau 3. Contenu du wiki sémantique (wiki 1) construit

Catégories	Sous catégories	Nombre de pages
Biological process	207	9 930
Cellular component	0	0
Coding sequence diversity	5	2 941
Developmental stage	0	1 146
Disease	0	1 426
Domain	0	37 480
Ligand	0	0
Molecular function	88	14 468
Post-translational modification	0	994
Technical term	0	31 519

L'ontologie externe associée à ce wiki est également téléchargée sur Uniprot (www.uniprot.org). Il s'agit de l'ontologie UniProtKB (figure 5) qui décrit la structure hiérarchique des mots-clés associés aux pages des protéines. Elle est constituée de 10 concepts et 1 182 sous concepts qui représentent les mots clés utilisés.

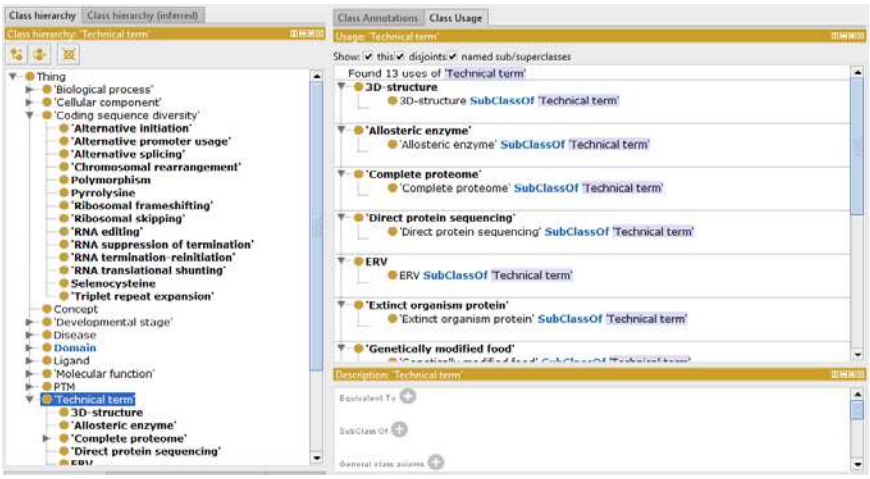

Figure 5. Ontologie UniProtKB téléchargée sur Uniprot (www.uniprot.org)

5.2. Protocole expérimental

L'objectif de l'expérimentation est de montrer que le wiki 2 obtenu après l'application de notre méthode est plus intéressant que le wiki 1. Ainsi pour évaluer l'apport de la méthode, nous vérifions dans un premier temps si les nouvelles catégories identifiées sont des tags importants dans le wiki. Ensuite, nous évaluons l'apport de la méthode en effectuant une recherche sémantique par catégorie sur les deux wikis.

Dans qui ce suit, nous nous intéressons aux catégories *« Cellular component »*, *Domain*, *Ligand*, *« Post-translational modification »*, *« Technical term »* pour évaluer l'approche. Nous définissons les tags importants par catégorie dans les données téléchargées et les pages attendues dans une recherche sémantique par catégorie.

Tags importants. Pour définir les tags importants dans le wiki, nous utilisons la mesure de la fréquence inverse (IDF). En effet *l'IDF* mesure l'importance d'un terme dans une collection de données. Nous l'appliquons dans notre contexte pour mesurer l'importance d'un tag. Cette mesure est définie comme suit :

$$\forall t \in T,\ IDF(t) = \log\left(\frac{|P|}{|g(t)|}\right) \quad (4)$$

Dans la formule (4) : $|P|$ =nombre total de pages, $|g(t)|$=nombre de pages associées au tag t.

L'IDF classe les tags par degré d'importance en considérant le tag qui a la plus petite valeur comme le plus important. Ainsi nous définissons les tags importants dans le wiki comme ceux dont la valeur de *l'IDF(t)* est inférieure à un seuil d'importance β fixé : $IDF(t) < \beta$. En appliquant la formule 4 sur le jeu de données et en fixant de manière empirique un seuil de degré d'importance à 5, le nombre de tags importants est défini dans le tableau 4.

Pages attendues dans une recherche sémantique par catégorie. Une page est attendue dans une recherche sémantique avec une catégorie si elle est associée à un mot clé de cette catégorie dans l'ensemble des données téléchargées. Le nombre de pages attendues par catégorie est donné dans le tableau 4.

Métrique d'évaluation. Pour évaluer la capacité de notre approche à trouver des résultats pertinents, nous utilisons la mesure du rappel définie dans l'équation (5). Soit X l'ensemble des réponses trouvées avec notre approche et Y l'ensemble des bonnes réponses. Le rappel est défini comme suit :

$$Rappel = \frac{|X \cap Y|}{|Y|} \quad (5)$$

Tableau 4. Nombre de tags importants par catégorie et pages attendues dans une recherche sémantique par catégorie

Catégories	Nombre de tags importants	Nombre de pages attendues
Cellular component	10	77 600
Domain	8	76 916
Ligand	13	60 409
Post-translational modification	4	19 070
Technical term	3	89 561

5.3. Résultats et discussions

Les résultats expérimentaux dans le tableau 5 montrent que la méthode permet d'identifier de nouvelles catégories parmi les tags avec un taux de rappel de 57,89 %. Ce score du rappel montre la capacité de la méthode à identifier les nouvelles catégories parmi les tags importants stockés sur les pages.

Tableau 5. Nombre de tags importants et les nouvelles catégories identifiées par notre méthode avec minsup=1 %, minconf=50 % et minlift=1.5

Catégories	Nombre de tags importants	Nouvelles catégories	Rappel (%)
Cellular component	10	6	60,00
Domain	8	7	87,50
Ligand	13	7	53,87
Post-translational modification	4	2	50,00
Technical term	3	0	0

Le tableau 6 donne les statistiques des deux wikis. Le nombre de sous-catégories du wiki 2 a augmenté de 22 sous-catégories par rapport au wiki 1 et le nombre de pages catégorisées a augmenté de 14 412 pages.

La figure 6 montre que le nombre de pages non catégorisées a diminué. Parmi les *29 407 pages* non catégorisées, *5 801 pages* ont été classées dans les sous-catégories de *« Cellular component »* et *8 611 pages* classées dans les sous-catégories de *Ligand*.

Tableau 6. Statistiques sur le wiki 1 et le wiki2

	Wiki 1	Wiki 2
Catégories	10	10
Sous-catégories	300	322
Catégories inutilisées	2	0
Pages catégorisées	99 904	114 316
Pages non catégorisées	29 407	14 995

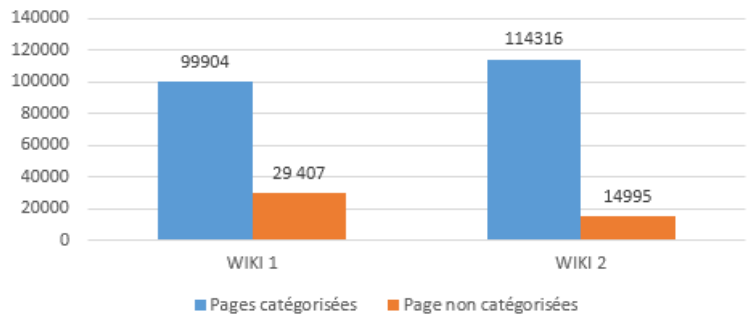

Figure 6. Histogramme des pages catégorisées et non catégorisées dans le wiki 1 et le wiki 2

Pour évaluer la contribution de la méthode dans une recherche sémantique par catégorie dans le wiki, nous utilisons les catégories définies dans le tableau 4. Le tableau 7 donne la valeur du rappel de la recherche sémantique avec chacune de ses catégories. La valeur du rappel est 0 dans le wiki 1 avec les catégories *« Cellular component »* et *Ligand*. Ce résultat est normal puisque dans le wiki 1 *« Cellular component »* et *Ligand* n'ont pas de pages. Ce résultat est amélioré dans le wiki 2 respectivement de 60,77 % et 44,52 % grâce aux nouvelles sous-catégories qui ont été définies pour *« Cellular component »* et *Ligand*. La recherche sémantique avec la catégorie *« Technical term »*donne le même résultat dans le wiki 1 et dans le wiki 2 car aucune sous-catégorie n'est identifiée pour cette catégorie. Dans le wiki 2, la recherche sémantique avec les catégories *Domain* et *« Post-translational modification »* est améliorée respectivement de 43,12 % et 37,39 %. Ces résultats étaient considérés comme du silence dans le wiki 1. Ainsi, notre approche améliore les résultats d'une recherche sémantique par catégorie dans le wiki 2 comparés aux résultats avec les mêmes catégories dans le wiki 1.

Tableau 7. Apport de la recherche sémantique par catégorie dans le wiki 1 et dans le wiki 2

Catégories	Rappel de la recherche sémantique dans le wiki 1 (%)	Rappel de la recherche sémantique dans le wiki 2 (%)
Cellular component	0	60,77
Domain	48,73	91,84
Ligand	0	44,52
Post-translational modification	5,21	42,61
Technical term	35,19	35,19

6. Conclusion

Dans ce papier, nous avons proposé une approche basée sur les motifs fréquents qui permet d'organiser le contenu d'un wiki sémantique. Notre méthode aide à identifier de nouvelles catégories potentielles et à définir les relations entre ces catégories et les pages. En utilisant une ontologie externe associée au wiki, nous proposons un algorithme qui insère ces nouvelles catégories dans la hiérarchie des catégories existantes. Nos expérimentations montrent que la méthode proposée permet de définir des catégories parmi les tags stockés sur les pages et diminue le nombre de pages non catégorisées dans le wiki sémantique.

Bien que les expérimentations réalisées donnent des résultats satisfaisants, le fait d'utiliser une ontologie développée spécifiquement pour la création des catégories

dans le wiki sémantique, réduit la portabilité de notre approche vers d'autres types de problèmes. Ainsi de nombreuses perspectives s'offrent à la suite de nos travaux. La première d'entre elle, est de faire appel à des ressources linguistiques pour résoudre les problèmes linguistiques relatifs aux annotations sémantiques telles que les synonymes, les ambiguïtés, les co-références, etc. La deuxième perspective est d'une part, de valider l'approche, au-delà des expérimentations, par des experts pour avoir des commentaires sur la qualité perçue de l'approche et d'autre part, d'étendre les expérimentations de notre approche sur d'autres wikis sémantiques afin d'analyser plus en détail l'impact de notre proposition. Les autres perspectives seront consacrées au développement d'une méthodologie de mise à jour de l'ontologie associée au wiki et d'une méthode qui permet de prédire les catégories d'une nouvelle page créée et tagguée par un ensemble de tags dans le wiki.

Remerciements

Ce travail a été effectué pendant notre séjour de recherche (2016) à l'UGB. Ce séjour a été 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 financé ce séjour de recherche.

Bibliographie

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

Boyer A., Brun A., Skaf-Molli H. (2010). Human Computer Collaboration to Improve Annotations in Semantic Wikis, *6th Conference on Web Information Systems and Technologies (Webist 2010)*, April, Valencia, Spain, p. 8.

Brin S., Motwani R. et Silverstein C. (1997). Beyond market baskets: Generalizing association rules to correlations, In Proc. of the ACMSIGMOD Conference, Tucson, Arizona, p. 265-276.

Buffa M., Gandon F., Ereteo G. (2007). Wiki et web sémantique, In *F. Trichet (Ed.), IC'2007 : 18e Journées Francophones d'Ingénierie des connaissances*.

Buffa M., Gandon F., Ereteo G., Sander P., Faro C. (2008). SweetWiki: A semantic wiki, *Web Semantics: Science, Services and Agents on the World Wide Web* Vol. 6, n° 1, February, p. 84-97, Semantic Web and Web 2.0.

Chernov S., Iofciu T., Nejdl W. and Zhou X. (2006). Extracting semantic relationships between wikipedia categories, *In 1st International Workshop SemWiki2006 – From Wiki to Semantics, co-located with the ESWC 2006*, Budva.

Deng Z. H., Lv S.-L. (2015). PrePost$^+$: An efficient N-lists-based algorithm for mining frequent itemsets via Children–Parent Equivalence pruning. *Expert Systems with Applications*, vol. 42, n° 13, 1 August, p. 5424-5432.

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, vol. 55, n° 9, p. 2008-2030.

Filipiak D., Ławrynowicz A. (2014). Generating semantic media Wiki content from domain ontologies, *SWCS'14 Proceedings of the Third International Conference on Semantic Web Collaborative Spaces*, vol. 1275, Germany ©2014, p. 68-76.

Fournier-Viger P. (2016). *SPMF An Open-Source Data Mining Library http://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php*

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

Gil Y., Knight A., Zhang K., Zhang L., Sethi R. (2013). An Initial Analysis of Semantic Wikis, *In Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI)*.

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

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

Krötzsch M., Vrandecic D., Völkel M. (2006). Semantic Mediawiki, *ISWC 2006:5th International Semantic Web Conference*, Athens,Ga, USA, November 5-9.

Krötzsch M., Schaffert S., Vrandecic D. (2012). *Swivt ontology specification*. http://semantic-mediawiki.org/swivt/.

Krötzsch M., Schaffert S., Vrandecic D. (2007a). Reasoning in semantic wikis, *In Reasoning web 2007*, vol. 4636 of Lecture Notes in Computer Science, Springer, p. 310-329.

Krötzsch M., Vrandecic D. Kolkel M., Haller H., Studer R. (2007b). Semantic wikipedia, *J. Web Sem*, p. 251-261.

Marinica C., Guillet F., Briand H. (2008). Vers la fouille de règles d'association guidée par des ontologies et des schémas de règles, *Atelier Qualité des Données et des Connaissances, EGC'08.*

Meilender T. (2013). *Un wiki sémantique pour la gestion des connaissances décisionnelles – Application à la cancérologie*, Thèse de Doctorat Université de Lorraine.

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

Pasquier N., Bastide Y., Taouil R., Lakhal L. (1999). *Efficient Mining of Association Rules using Closed Itemset Lattices. Information Systems*, Elsevier Science, vol. 24, n° 1, p. 25-46.

Rosenfeld M., Fernández A., Díaz A. (2010). Semantic Wiki Refactoring. A Strategy to Assist Semantic Wiki Evolution, *In Proceedings of the Fifth Workshop on Semantic Wikis (SemWiki 2010), co-located with 7th European Semantic Web Conference, ESWC.*

Schönberg C., Pree H., Freitag B. (2010). Rich ontology Extraction and Wikipedia Expansion Using Language Resources, *Proc. of the 11th int. Conf. on Web-Age Information Management, Jiuzhaigou,China, LNCS*, vol. 6184.

Shi L., Toussaint Y., Napoli A., Blansché A. (2011). Mining for Reengineering: An Application to Semantic Wikis Using Formal and Relational Concept Analysis, *in The Semanic Web: Research and Applications, 8th Extended Semantic Web Conference* Proceedings, *ESWC'11*, Heraklion, Crete, Greece, May 29-June 2, p. 421-435.

Tobias B., Andreas F. (2011). *FrequentPattern TagCloud, Semantic MediaWiki Extension*, Documentation, University of Heidelberg.

Traoré Y., Malo S., Diop C. T., Lo M., Ouaro S. (2015). Approche de découverte de nouvelles catégories dans un wiki sémantique basée sur les motifs fréquents, *IC'15*, June 2015, Rennes, France. collection AFIA.

Vaillant B. (2006) *Mesurer la qualité des règles d'association : Etudes formelles et expérimentales*, Thèse de doctorat École nationale supérieure des Télécommunications de Bretagne.

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

Zaki M. J. (2000). Scalable algorithms for association mining. *IEEE TKDE Journal*, vol. 12, n° 3, p. 372-390.

Zaki M. J., Hsiao C.-J. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining, *In 2nd {SIAM} International Conference on Data Mining.*

Zaki M. J., Gouda K. (2003). Fast vertical mining using diffsets, *In: SIGKDD*, p. 326-335.