Approche de découverte de nouvelles catégories dans un wiki sémantique basée sur les motifs fréquents

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 : Dans cet article, nous proposons une approche de découverte de nouvelles catégories potentiellement utiles dans un wiki sémantique. Les pages du wiki sont sémantiquement annotées et des tags (mots clés) peuvent être associés librement à celles-ci. Les pages sont créées par les utilisateurs autorisés à partager des informations sur le wiki. Les catégories permettent d'organiser les liens entre les pages dans le wiki. Elles sont créées par les experts. Notre contribution dans ce papier consiste à extraire parmi les tags qui sont associés librement aux pages, les motifs fréquents de tags qui sont identifiés comme de nouvelles catégories utiles qui guideront l'expert dans la création ou la modification de catégories dans le wiki. Nous utilisons l'ontologie associée au wiki pour bénéficier de plus d'informations structurées afin de sélectionner les tags de la fouille dans le prétraitement et d'éliminer certains motifs de tags de l'analyse dans la phase de fouille.

Mots-clés : Wiki sémantique, Ontologie, Motifs fréquents

1 Introduction

Nos travaux rentrent dans le cadre du projet[1] « Mise en place d'une plateforme web social et sémantique pour le partage de connaissances des communautés ouest-africaines» qui répond à un besoin de disposer d'un cadre de partage de connaissances sur les communautés ouest africaines en s'appuyant sur les technologies du web social et sémantique. Il s'agit de s'appuyer sur les méthodes de l'ingénierie de la connaissance et, en particulier sur les technologies du Web sémantique pour proposer des solutions de partage de connaissances à nos communautés. C'est ainsi qu'en vue de la mise en place d'une plateforme web social et sémantique, un wiki sémantique, est développé autour du moteur Semantic MediaWiki (Krötzsch et al., 2006). Un état de l'art sur les wikis sémantiques est disponible dans (Buffa et al., 2007) et (Meilender, 2013). La plateforme est organisée autour (i) des pages de catégories qui servent à l'organisation des informations dans le wiki ;(ii) des pages de propriétés qui servent à préciser les liens sémantiques qu'il y a entre les informations du wiki ; (iii) des pages (dites normales) qui sont les informations qu'on veut présenter sur le wiki. Les pages de catégories et de propriétés sont créés par l'expert du domaine. Les pages normales sont créées

1. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/ED/pdf/SenegalGastonBergerFR.pdf

par les utilisateurs autorisés à partager des connaissances sur le wiki. La plateforme propose à tous ses utilisateurs un champ (Figure 1) qui permet d'associer des tags (mots clés) à chaque page normale à la création. La quantité immense de tags stockés sur les pages cache plusieurs connaissances qu'il faut extraire pour réorganiser les liens entre les pages. Cela suppose alors une maintenance de cette plateforme par les experts du domaine. Cette maintenance consiste en l'annotation de nouvelles ressources, l'organisation des liens entre les ressources existantes. Pour guider l'expert dans cette maintenance, nous proposons dans ce papier une approche de découverte de nouvelles catégories utiles à partir des tags stockés sur les pages normales du wiki.

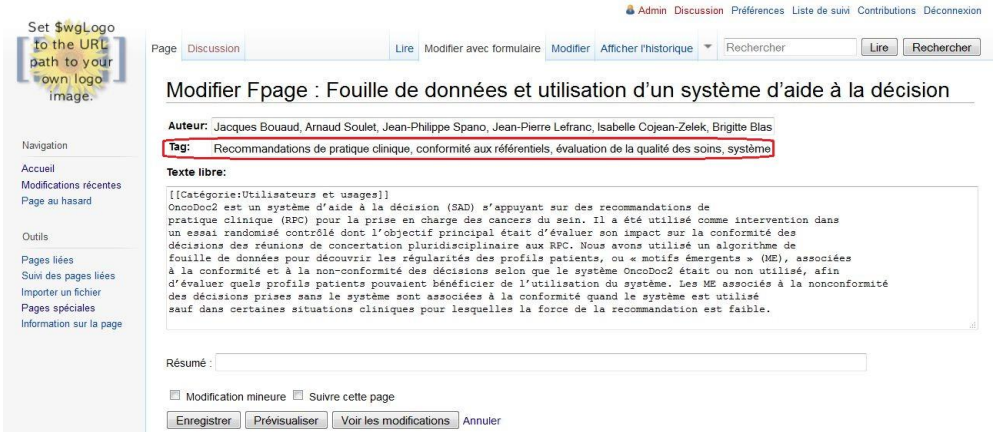

FIGURE 1 – Exemple de page tagguée.

Nous exportons l'ensemble des pages du wiki en RDF pour créer une base de connaissances du wiki. Cette base de connaissances est utilisée dans le processus de fouille pour bénéficier de plus d'informations structurées. La Figure 2 donne un extrait de la base de connaissances d'un wiki avec l'éditeur Protégé. Les concepts représentent les catégories du wiki, les relations représentent les propriétés du wiki et les instances représentent les pages du wiki. Les tags stockés sur les pages du wiki sont représentés par les valeurs des propriétés associées aux instances. SWIVT ontologie **(**Krözsch et al., 2012**)** fournit une base pour l'interprétation de données sémantiques exportées par Semantic MediaWiki.

Dans la suite de cet article, nous présentons à la section 2 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.

2 Définitions et notations

Dans cette section, nous définissons les différentes notions utilisées dans le reste de l'article.

Contexte d'extraction : Un contexte d'extraction est un triplet $CE = (P,T,R)$ où P représente l'ensemble fini des pages du wiki sémantique, T l'ensemble fini des tags, R une relation binaire entre T et P tel que $R(p, t) = 1$ 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 → P tel que pour t ϵ T, g (t) = {p/ p ϵ P}.

Motif fréquent 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. Un motif est fréquent si son support est supérieur à un seuil fixé σ. Soit $T_1 \subseteq T$ un motif de tags. Notons $Supp(T_1)$ (1) son support : T_1 est fréquent si $Supp(T_1) >= \sigma$.

$$Supp(T_1) = \frac{|\, g(T1)|}{|P|} \qquad (1)$$

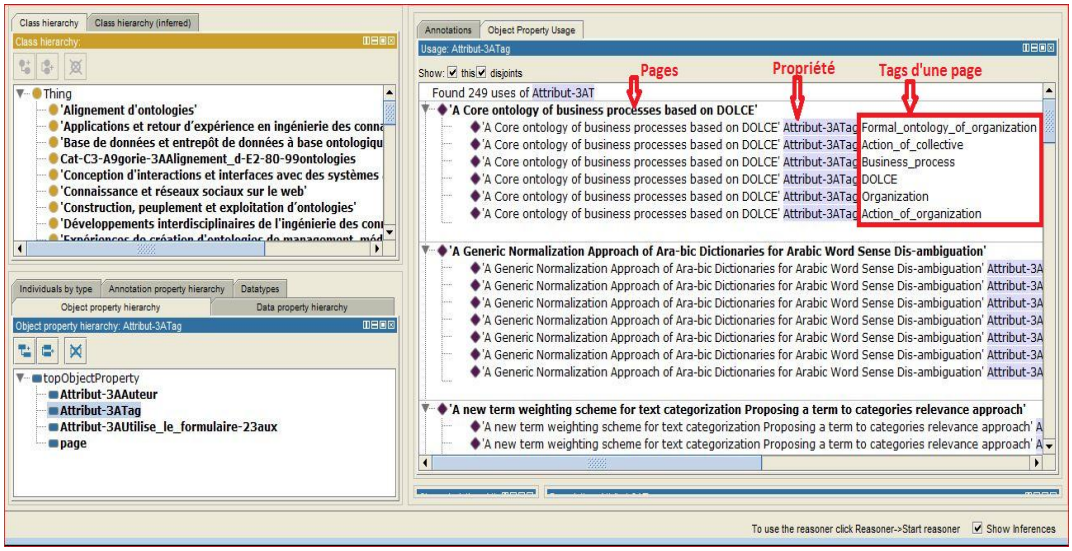

FIGURE 2 – Extrait de la base de connaissances du wiki.

Nouvelle catégorie : Dans notre contexte, une nouvelle catégorie est un motif fréquent de tags utilisés sur les pages du wiki et qui n'est pas dans la liste des catégories existantes de celui-ci. Soit C l'ensemble des catégories du wiki et f un motif fréquent de tag, f est une nouvelle catégorie si et seulement si elle vérifie la propriété suivante : $\forall t \in f \implies t \notin C$.

Distance sémantique de (Cilibrasi et al., 2006) : Pendant la fouille, certains tags peuvent ne pas être retenus à l'étape de prétraitement alors qu'ils peuvent avoir des corrélations avec les tags du domaine de la fouille. Nous utilisons la distance sémantique de (Cilibrasi et al., 2006) pour les sélectionner. Dans notre étude nous adaptons la distance sémantique proposée par (Cilibrasi et al., 2006) **(**notée DCV**)** entre deux termes (tags) t_1 *et* t_2 ainsi qu'il suit:

$$\boldsymbol{DCV}(t_1, t_2) = \frac{\max\{\log(fr(t_1)), \log(fr(t_2))\} - \log(fr(t_1, t_2))}{\log(M) - \min\{\log(fr(t_1)), \log(fr(t_2))\}} \qquad (2)$$

M désigne le nombre de page du wiki, $fr(t_1)$ la fréquence de t_1, $fr(t_2)$ la fréquence de t_2 et $fr(t_1, t_2)$ la fréquence de t_1 *et* t_2. Cette mesure désigne une mesure de la proximité sémantique entre t_1 *et* t_2 et varie entre 0 et 1. 0 indique que t_1 est « sémantiquement » proche de t_2 et 1 le contraire. Deux tags t_1 *et* t_2 cachent donc une relation si $\boldsymbol{DCV}(t_1, t_2)$=0.

3 Travaux existants

Un certain nombre de travaux utilisant les ontologies dans le processus d'extraction de connaissances à partir des données (ECD) existent. (Euler et al., 2004) utilisent l'ontologie pendant la phase de prétraitement, (Brisson et al., 2006) utilisent l'ontologie dans le prétraitement et le post-traitement, (Marinica et al., 2010) 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'enregistrements de la base de données, ce qui n'est pas toujours possible. Dans notre étude, nous utilisons un wiki sémantique qui stocke une quantité immense de tags sur les pages. Dans ce contexte, les travaux de (Tobias et al., 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. Notre objectif est de fouiller dans les tags stockés sur les pages pour identifier de nouvelles catégories. Nous nous inspirons des travaux de (Yaya et al.,2014) sur

l'apport de l'ontologie dans le prétraitement de l'extraction des connaissances à partir d'un wiki sémantique, des travaux de (Christian et al., 2010) sur l'extraction des catégories, propriétés pour créer une ontologie et l'enrichir au fur et à mesure à partir d'un wiki, les travaux de (Jian et al., 2006), (Julien et al., 2010), (Fleischhacker et al., 2012) sur la fouille de données utilisant une base de données RDF. Sur la base de ces travaux, nous proposons d'utiliser la base de connaissances obtenue à partir du wiki sémantique pour bénéficier de plus d'informations structurées pour sélectionner automatiquement les tags de la fouille sur la base d'une proximité sémantique avec l'objectif de la fouille. Certains tags *a priori* rejetés peuvent avoir des corrélations avec les tags retenus que nous détectons grâce à la distance sémantique de (Cilibrasi et al., 2006). Pour construire le contexte d'extraction qui sera le point d'entrée de l'étape de la fouille, nous utilisons la relation entre les pages et les tags. Dans la phase de fouille en s'inspirant de (Antunes, 2007) nous utilisons la structure conceptuelle de l'ontologie comme une condition d'élagage pour enlever certains motifs de tags de l'analyse.

4 Description de l'approche

Notre approche (Figure 3) se situe dans le cadre global de l'extraction de connaissances à partir de données (ECD**)** (Fayyad et al., 1996) et se déroule en deux grandes étapes ci-dessous expliquées :

- Etape 1: Construction du contexte d'extraction (Algorithme 1) : (1) exportation de toutes les pages du wiki sémantique en RDF; (2) définition et expression de l'objectif de fouille par un mot clé ;(3) sélection des tags proches de l'objectif de la fouille en utilisant l'ontologie et ceux corrélés grâce à la distance DCV. (4) l'ensemble des tags du domaine et ceux corrélés forment les tags sélectionnés pour le CE; (5) construction du contexte d'extraction : sélection des pages de chaque tag et définition de la relation.
- Etape 2: Algorithme de découverte (Algorithme 2) : cette étape est basée sur l'algorithme Apriori (Agrawal ,1994**) :** (6) sélection de motifs de tags candidats et élagage en utilisant la structure conceptuelle de l'ontologie du wiki pour enlever des motifs de tags candidats qui sont sémantiquement proches des catégories existantes dans le wiki. (7) calcul des motifs fréquents de tags qui sont identifiés comme de nouvelles catégories utiles.

L'innovation de notre approche est la sélection non supervisée des tags du contexte d'extraction sans l'aide de l'expert et l'introduction d'une contrainte d'élagage basée sur la structure conceptuelle de l'ontologie associée au wiki dans algorithme Apriori (Agrawal ,1994**)**. A notre connaissance c'est la première fois que notre approche est utilisée dans la phase de prétraitement et de fouille de l'ECD à partir d'un wiki sémantique.

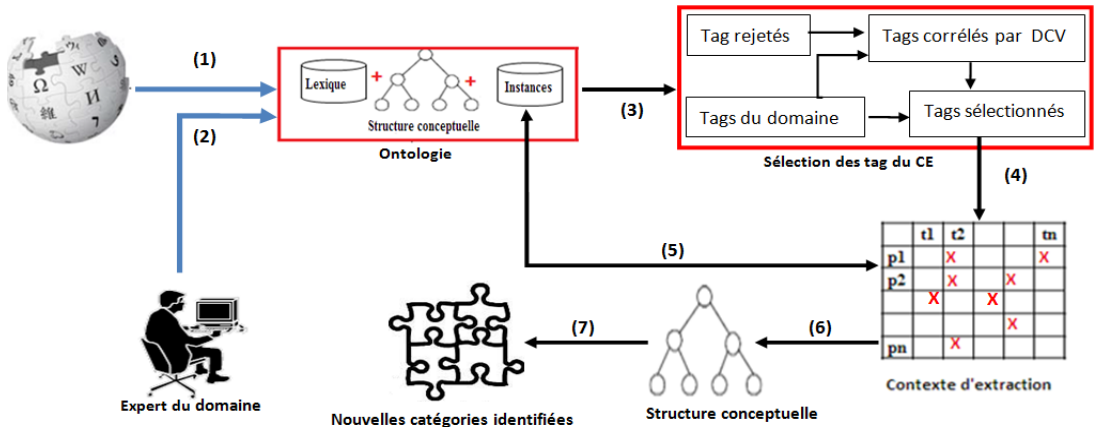

FIGURE 3 – Approche de découverte de nouvelles catégories dans un wiki sémantique.

Algorithme 1 : Algorithme de construction du CE

Entrée : M=mots clés de la fouille, BC=Base de connaissances
Sortie : CE: contexte d'extraction
Début

```
1.   /* Extraction de tags ,de pages par de requêtes sparql */
2.    T=ensemble des tags extraits de BC
3.    P=ensemble des pages extraites de BC
4.    T_O=ensemble des tags de BC similaire au mots clés M
5.    T_R= T \ T_O // T_R =Tags rejetés (non proches du domaine M)
6.    T_S = T_O // T_s=Tags du contexte d'extraction CE
7.    // application de la distance sémantique de Cilibrasi
8.   Pour chaque tag t_r ϵ T_R faire
9.       Pour chaque tag t_o ϵ T_O faire
10.         Si (DCV(t_r,,t_o ) = 0 ) alors
11.          T_S = T_S ∪ t_r
12.         Finsi
13.      Finpour
14.   Finpour
15.   /*  Construction du contexte d'extraction CE  */
16.   Pour chaque page p de P faire
17.       Pour chaque tag t de Ts faire
18.            Si (p ϵ g(t)) alors
19.              CE(p, t)=1
20.            Sinon
21.              CE(p, t)=0
22.            FinSi
23.       Finpour
24.   Finpour
25.   Retourner CE
```

Fin

Algorithme 2 : Algorithme de découverte de nouvelles catégories

Entrée :CE:contexte d'extraction (Base de transaction),S:structure conceptuelle de l'ontologie,minsup:seuil minimum de support
Sortie : F:motifs fréquents de tags
Début

```
1. L_1=ensemble des 1-itemsets fréquents
2. K=2
3. Tant que( L_K-1 ≠ ∅ ) 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 qu'il existe un élément de C qui
   est sémantiquement proche d'un concept de la structure conceptuelle S
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

5 Conclusion

Cet article a présenté une approche de découverte de nouvelles catégories utiles dans un wiki sémantique en utilisant la base de connaissances à base ontologique du wiki dans le processus. De nombreuses perspectives s'offrent à la suite de nos travaux. La première d'entre elles est d'évaluer notre approche sur un wiki sémantique avec un volume important d'annotations 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 réorganiser les pages du wiki.

Références

AGRAWAL R. AND SRIKANT. R.(1994) Fast algorithms for mining association rules in large databases, *Proc. VLDB conf., pp 478-499*, September 1994.

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.

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.

BUFFA M., GANDON F., ERETEO G. (2007). Wiki et web sémantique. *In F. Trichet (Ed.),* IC'2007 : 18eJournées Francophones d'Ingénierie des connaissances.

CHRISTIAN SCHÖNBERG, HELMUTH PREE,BURKHARD FREITAG (2010). Rich ontology Extraction and Wikipedia Expansion Using Language Resources, Proc. of the 11th int.Conf. on Web-Age Information Management ,Jiuzhaigou,China, LNCS,volume 6184.

CILIBRASI R., VITANYI P. (2006). Similarity of Objects and the Meaning of Words. In Proceedings of the Third international conference on Theory and Applications of Models of Computatio TAMC'06, Beijing, China, pages 21-45.

DBPEDIA EN FRANÇAIS. (2014). Dernière consultation : Décembre 2014. http://fr.dbpedia.org/

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.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996a). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), pp. 27-34.

FLEISCHHACKER D. VÖLKER J., AND STUCKENSCHMIDT H. (2012). Mining rdf data for property axioms. In OTM Conferences (2), pages 718-735.

JIANG T. ET TAN A. (2006). Mining rdf metadata for generalized association rules: knowledge discovery in the semantic web era. In WWW, pages 951-952.

JULIEN RABATEL, SANDRA BRINGAY AND PASCAL PONCELET. (2010). Contextual sequential pattern mining. In (ICDMW), ICDM, pages 981-988. IEEE.

KRÖZSCH M. ET VRANDECIC D. (2012). Swivt ontology specification. dernière consultation Janvier 2015, http://semantic-mediawiki.org/swivt/

KRÖTZSCH M., VRANDECIC D., VÖLKEL M. (2006).Semantic Mediawiki. ISWC 2006:5th International Semantic Web Conference,Athens,Ga, USA, November 5-9.

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.

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

TOBIAS BECK, ANDREAS FAY. (2011). FrequentPattern TagCloud, Semantic MediaWiki Extension, Documentation, University of Heidelberg.

YAYA TRAORE, 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.