Annotation Sémantique De Pages Web

BENYAHIA Kadda[1], LEHIRECHE Ahmed[1], LATRECHE Abdelkrim[1]

[1] Laboratoire EEDIS, Université Djillali Liabes de Sidi Bel Abbes, ALGERIE
benyahiaka@gmail.com

Abstract. L'annotation d'une page web constitue l'outil qui permet d'associer une sémantique au contenu de la page. Enrichir le partage d'information, améliorer les échanges et augmenter l'interopérabilité sur le web sont les principaux objectifs. En effet, avec la grande masse de données gérées à travers le monde et surtout avec l'avènement du web, l'annotation manuelle de ces pages est impossible. Dans cet article nous nous intéressons à l'annotation semi-automatique de page web, nous présentons un système d'annotation sémantique de pages web basé sur l'utilisation d'une ontologie. Notre approche consiste à relier les mots clés représentant la page à annoter aux concepts de l'ontologie pour aider l'auteur à réaliser l'annotation. Les mots qui entrent dans la composition de l'annotation sont déterminés à partir d'une analyse mixte : le calcul du degré de similarité et le calcul de la fréquence.

Keywords: Annotation, Web Sémantique, Ontologie, Degré de Similarité, Calcul de la Fréquence, RDF.

1 Introduction

Le Web Sémantique essai de répondre à la nécessité d'accéder seulement à l'information directement utilisable. Cette problématique est née du fait que les moteurs de recherches n'utilisent que le mot pour découvrir l'objet de la requête d'un usagé. Une solution serait l'ajout d'une couche sémantique. L'objet de la recherche devient alors un contenu i .e un sens. Cette solution n'est possible qu'a la condition que chaque document soit doté d'une couche sémantique. L'annotation en est l'une de ces solutions.

Annoter c'est accompagner un texte de notes, de remarques, des explications, de commentaires pour aider leurs lecteurs à le comprendre. Actuellement et avec ce grand volume d'information, il est difficile d'annoter manuellement des millions de ressources mises à la disposition des utilisateurs.

L'indexation d'un texte [8], consiste à repérer dans son contenu certains mots ou expressions particulièrement significatifs (Appelés termes d'indexation) dans un contexte donné, et à créer un lien entre ces termes et le texte d'origine. Il existe trois types d'annotation, manuelle : lorsque le document est analysé par un spécialiste du domaine ou un documentaliste, automatique : lorsque cette tâche est réalisée complètement par la machine, et semi automatique lorsque une partie se faite automatiquement et l'intervention du spécialiste est nécessaire pour l'autre partie.

Selon Salton [9], l'indexation manuelle peut conduire à deux indexations différentes d'une même page.
L'indexation sémantique prend en compte la sémantique des mots, Desmontils [3] a indexé une page avec des mots clés attachés à une ontologie. Yan Bodain [13] a proposé un outil d'annotation KATIA qui permet d'annoter une page web, en sélectionnant une région de texte et en choisissant l'élément de l'ontologie correspondant dans l'arbre hyperbolique. Baz [2] a présenté un modèle d'annotation qui construit un noyau sémantique pour chaque document avec les concepts et leur proximité. L'annotation des documents en utilisant des ontologies de domaine est pratiquée dans le domaine biopuces [5], le domaine médical, Lylia [6] a utilisé la technique de propagation des annotations sur les documents en utilisant une ontologie, Amardeilh [1] a présenté un outil d'annotation Ontopop qui est basé sur la combinaison des outils d'extraction d'information (EI) avec les outils de représentation des connaissances du Web service.
La plupart de ses systèmes utilisent le poids sémantique des mots clés de la page dans leurs démarches d'annotation.
Dans cet article nous présentons un système d'annotation basé sur :1) l'extraction des mots selon leurs poids sémantiques et leurs valeurs statistiques dans la page (2), ces mots sont associés aux concepts de l'ontologie.
Le reste de cet article est organisé comme suit : la Section 2 introduit notre démarche d'annotation de pages web. Nous présentons l'expérimentation dans la Section 3, et nous finirons par une conclusion et des perspectives dans la Section 4.

2 L'approche

La tâche de notre système consiste à prendre en entrée une page web et fournir en sortie le même contenu enrichi par des annotations sémantiques basées sur des représentations de la connaissance plus ou moins formelles.
Afin de réaliser cette tâche, nous nous appuyons sur le contenu qui se traduit par les mots clés qui représentent le mieux cette page.
Les différentes étapes de l'approche sont schématisées dans la figure 1.

2.1 L'analyse linguistique

Consiste à extraire les termes composants la page web. Le traitement linguistique représente le document à annoter par un ensemble de termes simples et importants. Cette extraction est le résultat d'un nettoyage de la page et de la segmentation du texte.

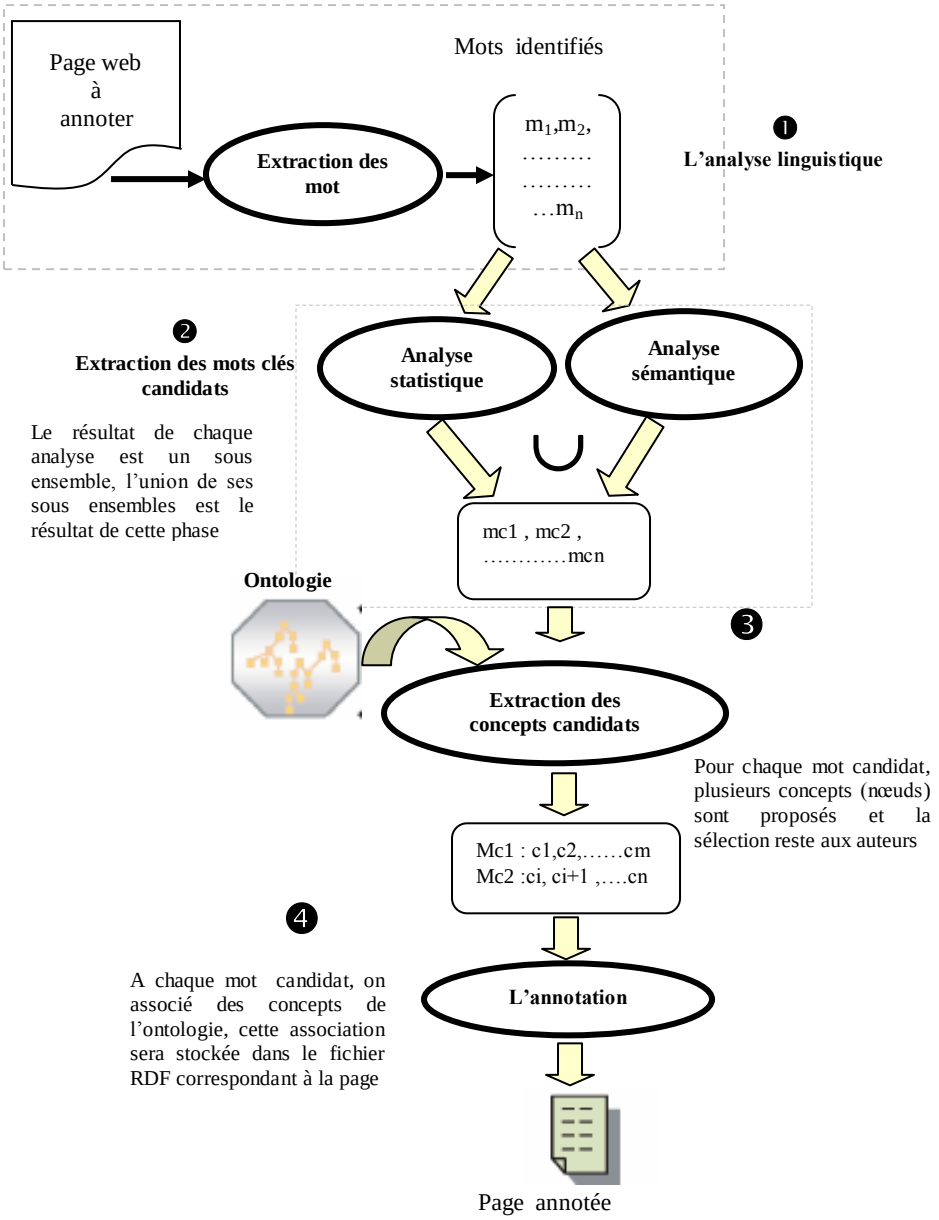

Fig. 1. Schéma synoptique de l'approche proposée

2.2 Sélection des mots clés candidats

Vise à déterminer l'ensemble des mots clés qui représente mieux la page web, cet ensemble est l'union de deux sous ensembles résultats de l'analyse sémantique et l'analyse statistique

2.2.1 L'analyse sémantique

Cette analyse consiste à Déterminer le poids d'un mot dans la page web, ce poids sémantique est calculé en se basant sur la mesure de similarité.
Des mesures de proximité sémantique ont été proposées dans la littérature (une douzaine) utilisant des structures de réseaux sémantiques ou hiérarchiques :

- Mesure basée sur le chemin (path based measures) entre les deux concepts à comparer telles que définies par Rada, Leacock ou Jiang en 1997.
- Mesure basée sur la notion de contenu d'information (Information Content ou IC) telle que celle définie par Wu et Palmer [12] et Resnik.
- Mesure basée sur une combinaison du chemin et du contenu d'information par D. Lin en 98.
- Mesure basée sur l'algorithme de Lesk que Patwardhan, Banerjee et Pederson en 2003 adapté à WordNet.

Nous avons utilisé la mesure de **Wu-Palmer,** cette mesure a l'avantage d'être simple à implémenter et d'avoir d'aussi de bonnes performances que les autres mesures de similarité selon D. Lin. Son principe est le suivant :
Dans un domaine de concepts, la similarité est définie par rapport à la distance qui sépare deux concepts dans la hiérarchie et également par leur position par rapport à la racine. La similarité entre C1 et C2 est :

$$Consim(C1, C2) = \frac{2 * depth(C)}{depth_C(C1) + depth_C(C2)} \quad (1)$$

Où C est le PPG de C1 et C2 (en nombre d'arcs), depth (C) est le nombre d'arcs qui sépare C de la racine et depthc (Ci) avec i le nombre d'arcs qui séparent Ci de la racine en passant par C.
Dans cette phase , un mot sera acceptée si et seulement s'il est fortement en relation avec d'autres mots de cette page. Cette décision dépend du choix d'un seuil défini par l'utilisateur Ce résultat est un ensemble **poids_sem.**

2.2.2 L'analyse statistique

Déterminer l'importance d'un terme dans une page web, dans cette analyse nous avons utilisé la technique de pondération des termes car elle permet d'affecter aux termes d'un document, un poids pour traduire son importance dans le document, donc son degré d'informativité. Dans cette technique on s'intéresse à la pondération locale qui mesure la représentativité locale d'un terme. La fonction utilisée est la fonction

normalisée qui permet de réduire les différences entre les valeurs associées aux termes du document. Elle est donnée par la formule suivante :

$$0.5 + 0.5 \frac{tf_{ij}}{\max_{t \in Dj}(tf_{ij})} \quad (2)$$

Où $\max_{t_i \in D_j}(tf_{ij})$ est la plus grande valeur de tf_{ij} des termes du document Dj.

Le résultat de cette étape est un ensemble de mots nommé degré_signif
Les mots clés candidats = poids_sem $\cup$ degré_signif

2.3 L'extraction des concepts candidats

Dans cette étape on utilise une Ontologie de domaine, nous avons fait un passage des mots clés candidats à l'ontologie pour définir les concepts correspondants.
A chaque passage d'un terme à l'ontologie, un ensemble de concepts sera présenté aux auteurs pour choisir les concepts à utiliser dans l'étape de l'annotation.
Cette étape est semi-automatique, la recherche et la proposition se fait par notre système et le choix des concepts les plus significatifs reste aux auteurs. L'automatisation de cette tâche fait l'objet de plusieurs recherches

2.4 L'annotation

C'est la dernière phase, elle consiste à associer à chaque mot clé des concepts de l'ontologie (nœuds).
Après la proposition des concepts candidats, et le choix effectué par l'auteur dans l'étape précédente, une association entre ces mots et ces concepts élus sera stockée dans un fichier RDF correspond à la page.

3 Expérimentations et résultats

Nous allons montrer, en utilisant un ensemble de pages web l'intérêt de la démarche que nous avons proposé pour l'annotation semi-automatique des pages web.
Pour cela nous utilisons 21 pages annotées généralement par des auteurs, notre démarche consiste à comparer l'annotation obtenue par notre approche qui utilise une analyse sémantique et une analyse statistique pour la sélection des mots clés candidats avec celle obtenue par l'utilisation de la technique de calcul de similarité uniquement dans l'étape de l'extraction des mots clés.
Dans cette étape d'évaluation on a utilisé différentes Ontologie selon le domaine de la page utilisée pour l'évaluation. nous avons utilisé quatre autre ontologie pour l'évaluation, La figure 2 présente l'ontologie du domaine « Recherche » un extrait de PROTEGE2.0..

Afin de représenter le résultat, nous avons défini un indice de qualité d'annotation :

$$Iqa = \frac{Ac}{Ae} \in [0,1] \quad (3)$$

– Ac : nombre d'annotations correctes par page ;
– Ae : nombre d'annotations par page.

Le tableau 3.1 et la figure3 représentent les résultats de la comparaison pour les 21 pages évaluées et la figure 4 présente un extrait du fichier RDF de l'annotation résultat

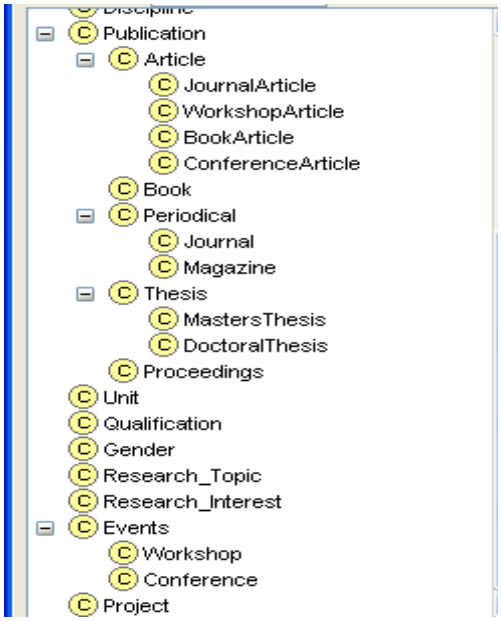

Fig. 2. Extrait de l'ontologie du domaine « Recherche ».

Table 1. Les résultats de la comparaison

Méthode	Iqa
A- calcul de similarité	0.62
B- calcul de similarité + calcul de fréquence	0.71

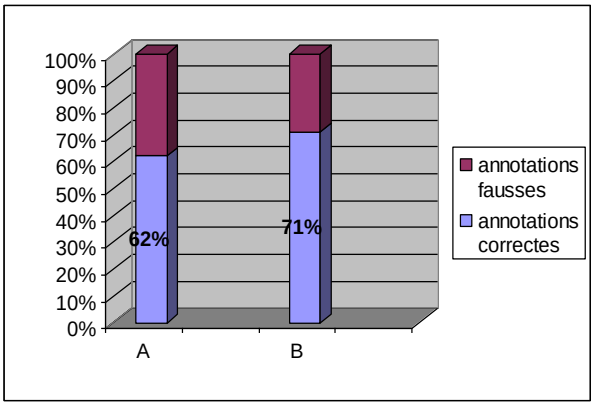

Fig. 3. Représentation Des Résultats.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:Inst="D:/annotation/Ontologies/Institute#">
<rdf:Description rdf:about="file:/D:/Annotation/Exemples/univ.htm">
<rdf:type rdf:resource=" D:/annotation/Ontologies/Institute# "/>
<Inst :Discipline> Mathematics </Inst : Discipline >
<Inst : Discipline > Computer_science </Inst : Discipline >
<Inst : Discipline > Commerce </Inst : Discipline >
<Inst :Research_Topic> Semantic_Web </Inst : Research_Topic >
<rdf:Description>
```

Fig. 4. Un extrait d'annotation Rdf

4 Conclusion Et Perspectives

Dans cet article nous nous sommes intéressés à l'annotation semi-automatique d'une page web. Nous proposons une démarche basée sur un calcul sémantique et un calcul statistique dans l'étape d'extraction des mots candidats de la page à annoter qui seront ensuite relié à une ontologie de domaine par l'intervention de l'auteur de la page.

Les résultats des expérimentations donnent 71% d'annotations correctes. Ces résultats sont très encourageants. La démarche que nous proposons offre des résultats d'annotation très intéressants tout en satisfaisant le critère du passage à l'échelle qui est un point très crucial dans un contexte ou la masse de données est très importante.

Comme perspectives, nous projetons d'appliquer notre méthode sur un plus grand nombre de page Web et d'une complexité plus élevée afin de faire une étude comparative effective.
Nous travaillons sur l'intégration des connaissances de l'utilisateur dans le processus d'annotation et l'exploitation de l'annotation dans les systèmes de recherche d'informations.

Références

1. Amardeilh.F, Web Sémantique et Informatique Linguistique: propositions méthodologiques et réalisation d'une plateforme logicielle, PhD thesis, institut de recherche en informatique de toulouse 2007
2. Baziz M, Indexation conceptuelle guidée par ontologie pour la recherche d'information. PhD thesis, Institut de recherche en informatique de Toulouse, université Paul Sabatier, 2005.
3. Desmontils.E , Jacquin.C, and Morin.E. Indexation sémantique de documents sur le web : application aux ressources humaines. In Proceedings of Journées de l'AS-CNRS Web sémantique, Octobre 2002.
4. Doumi N. et Lehireche.A, Une ontologie pour le lexique arabe, in proceeding du 2ème congrès international de "l'ingénierie de la langue arabe et de l'ingénierie de la langue", CRSTDLA, UA, 2005.
5. Khelif.K and Dieng-Kuntz. R, Annotations sémantiques pour le domaine des biopuces. In Proceedings of 15èmes journées francophones d'ingénierie des connaissances, 2004.
6. Lylia.A, Annotation de documents par le contexte de citation basée sur une ontologie. 2006
7. Roberston.S.E et Walker.S, On relevance weights with little relevance information. In proceeding of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 16-24. ACM press, 1997.
8. Salton.G , A comparaison between manual and automatic indexing methods. In Proceedings of Journal of American documentation, 1971.
9. Salton.G , Another look at automatic text-retrieval systems. Commun. ACM, 29(7) : 648–656, 1986.
10. Toumouh.A, Lehireche.A, Widdows.D, Malki.M, Adapting Word Net to the Medical Domain using Lexicosyntactic Patterns in the Ohsumed Corpus: 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06), , Dubai/Sharjah, UAE. 2006.
11. Widdows.D, Toumouh.A, Dorow.B, Lehireche.A, Ongoing Developments in Automatically Adapting Lexical Resources to the Biomedical Domain, International Conference on Language Resources And Evaluation, Italy, LREC 2006.
12. Wu.z et .Palmer.M, verb semantic and lexical selection, proceedings of the 32nd annual meeting of the associations for computational linguistics. Pages133-138, 1994
13. Yan Bodain, Logiciel d'annotation pour la conception de cours sur le Web sémantique, IHM 2006