Approche Incrémentale d'Extraction et d'Alignement Sémantique Guidée par une Ontologie

Annotation Sémantique

Mouhamadou Thiam[*,***] — Nacéra Bennacer[**] — Nathalie Pernelle[*] — Moussa Lo[***]

[*] LRI, Université Paris-Sud 11 INRIA Saclay Ile de France
{Nathalie.Pernelle, Mouhamadou.thiam}@lri.fr

[**] SUPELEC, 3 rue joliot- curie F-91192 Gif-sur-Yvette cedex, France
Nacera.Bennacer@supelec.fr

[***] LANI, Université Gaston Berger UFR S.A.T, B.P 234 Saint-Louis Sénégal
Lo_Moussa@yahoo.fr

RÉSUMÉ. *SHIRI*[1] est un système d'intégration de documents semi-structurés guidé par une ontologie. Il s'appuie sur une approche automatique, non supervisée et guidée par une ontologie pour extraire, aligner et annoter sémantiquement des nœuds balisés de documents. Il utilise RDF/OWL pour la représentation des ressources et SPARQL pour les interroger. Le papier se focalise sur l'algorithme *Extract-Align* qui exploite des patterns pour extraire des termes et des entités nommées. Les résultats des expériences sur un corpus de documents HTML sont très prometteurs et montrent comment le comportement incrémental de l'algorithme permet de peupler l'ontologie, de réduire l'accès à des ressources externes et d'augmenter le nombre de termes alignés directement avec l'ontologie.

ABSTRACT. *SHIRI* is an ontology-based system for integration of semi-structured documents related to a specific domain. It relies on an automatic, unsupervised and ontology-driven approach to extract, align and annotate semantically tagged document nodes. *SHIRI* uses RDF/OWL languages for resources representation and SPARQL for their querying. The paper focuses on *Extract-Align* algorithm which exploits terms patterns and named entity ones. We experiment and validate the algorithm on a HTML corpus related to call for papers in computer science. The obtained results are very promising and show how the incremental behaviour of the algorithm enriches the ontology, reduces the access to extern resources as the Web and increases the number of terms aligned directly with the ontology.

MOTS-CLÉS : Extraction d'information, Annotation sémantique, Alignement, Ontologies, Documents semi-structurés, OWL, RDF/RDFS

KEYWORDS : Information Extraction, Semantic Annotation, Alignment, Ontology, Semi-structured documents, OWL, RDF/RDFS

1. Système Hybride d'Intégration et de Recherche d'Information, Digiteo labs project (LRI, SUPELEC)

1. Introduction

Les informations disponibles sur le Web sont en général au format HTML, et sont donc plus ou moins bien structurés syntaxiquement. La nécessité d'automatiser le traitement de ces informations, leur exploitation par des applications et leur partage justifie l'intérêt que porte la recherche sur le Web sémantique. A cause de l'absence de sémantique, l'interrogation de ces ressources est généralement basée sur des mots clés. Ce qui n'est pas satisfaisant parce que ne garantissant pas la pertinence des réponses qui sont faites de documents entiers. L'annotation de ressources Web avec des métadonnées sémantiques devrait permettre une meilleure interprétation de leur contenu. La sémantique des métadonnées est définie dans une ontologie de domaine à travers des concepts du domaine et de leurs relations. L'annotation manuelle est une tâche longue et fastidieuse qui ne sied pas à l'échelle du Web. L'automatisation des techniques d'annotation est un facteur clé pour le Web du futur et son passage à l'échelle.

De nombreux travaux appartenant aux domaines de recherches complémentaires, tels l'intelligence artificielle, l'ingénierie des connaissances et la linguistique ont exploré la question de l'annotation de ces documents. Certaines activités sont basées sur des approches supervisées ou sur l'hypothèse de l'existence de modèles de structure soit dans les documents traités comme dans [7, 8, 10], soit dans le texte comme dans [3, 13]. En général, ces hypothèses sont incompatibles avec l'hétérogénéité et le grand nombre de documents disponibles sur le Web. Actuellement, une information peut apparaître dans différents types de structure en fonction de la forme du document.

La reconnaissance d'entités nommées a pour but de localiser et de classer les éléments dans le texte dans des catégories prédéfinies telles les noms de personnes, d'organisations, de lieux, les dates, etc. Certains systèmes non supervisés de reconnaissance d'entités nommées sont basés sur des ressources lexicales [9], ou sur les ressources lexicales construites grâce à des données disponibles sur le Web [13, 2]. Certaines approches utilisent le Web comme un corpus sur lequel ils appliquent des patterns afin de trouver des labels pour annoter une entité nommée d'une ressource [3]. Du fait que l'accès à des ressources externes nécessite beaucoup de temps, elle n'est appliquée que lorsque les autres stratégies échouent.

L'automatisation de l'annotation de documents hétérogènes peut aussi se fonder sur des termes qui décrivent les concepts et qui ne sont pas des entités nommées. Les techniques différentes d'extraction peuvent être classées en techniques linguistiques, statistiques ou hybrides [15, 14].

Une fois un terme ou une entité nommée est extrait, il est à comparer avec l'ensemble des termes qui appartiennent à l'ontologie (labels de concept ou entités nommées). Des mesures de similarité qui peuvent être utilisées pour estimer une similarité sémantique entre des termes ou entre des entités nommées ont été largement étudiées dans [6].

SHIRI [1] (cf. figure 1) peut être présenté comme un système d'intégration de documents semi-structurés basé sur une ontologie d'un domaine donné. Le but du système est de permettre aux utilisateurs d'accéder aux parties pertinentes de documents HTML comme réponses à leurs requêtes. SHIRI utilise des langages standardisés du W3C tels RDF, OWL pour la représentation des ressources et SPARQL pour leur interrogation. Le système repose sur une approche automatique, non supervisée et basée sur une ontologie pour l'extraction, l'alignement et l'annotation sémantique des nœuds de documents balisés. L'extraction de termes candidats à aligner avec l'ontologie s'appuie sur un ensemble de patterns d'entités nommées et de termes. Il procède de manière incrémentale

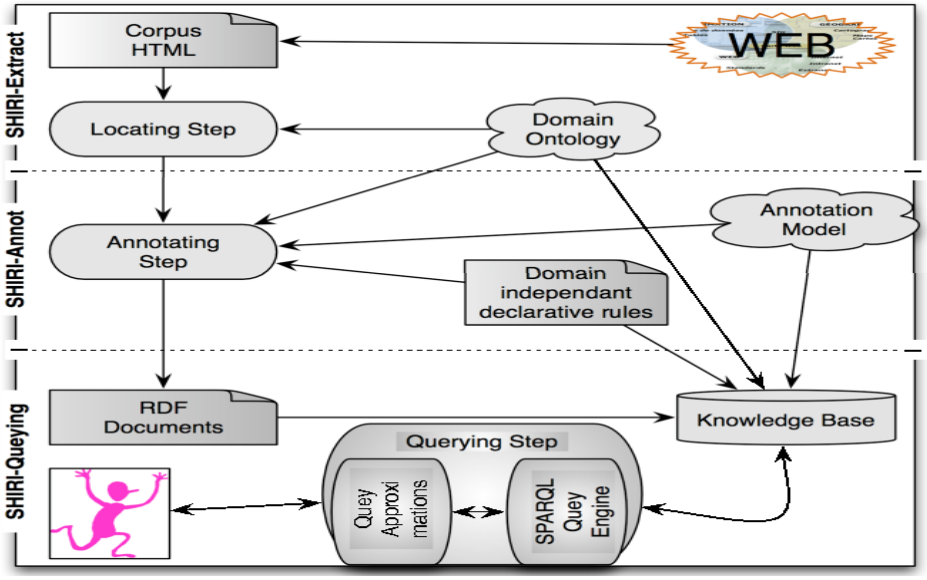

Figure 1. *Architecture Globale de SHIRI*

en peuplant l'ontologie avec les termes décrivant des instances de concepts appartenant au domaine. Ce qui permet ainsi de réduire l'accès aux ressources externes telles le Web. Les annotations des termes sont associées aux éléments balisés du document HTML (désormais appelés unité structurelle) [1]. En effet, à l'exception des entités nommées, les instances de concepts sont souvent noyées dans le texte et ne sont pas aisément dissociables. Même les techniques avancées de traitement du langage naturel, souvent adaptées à des corpus très spécifiques, ne réussissent pas à les extraire tous avec précision.

Dans cet article nous nous concentrons sur l'algorithme d'extraction et d'alignement *Extract-Align*. Nous l'avons expérimenté et validé sur un corpus de documents HTML qui concernent des appels à communication à des conférences informatiques et les résultats obtenus sont très prometteurs. Ces résultats montrent comment le comportement incrémental de l'algorithme *Extract-Align* permet d'enrichir l'ontologie et comment le nombre de termes et entités nommées alignés directement avec l'ontologie augmente. Dans la section 2, nous détaillons l'approche d'extraction et d'alignement. Dans la section 3, nous présentons les résultats des expérimentations. Dans la section 4, nous concluons et donnons quelques perspectives.

2. Approche d'extraction et d'alignement sémantique incrémental

Dans cette section, nous mettons l'accent sur l'extraction des termes candidats et leur alignement avec l'ontologie. L'approche d'extraction consiste à appliquer un ensemble de patterns sur les documents en entrée pour obtenir des termes candidats. Elle distingue deux types de patterns : des patterns pour les entités nommées et des patterns pour les termes. Les termes candidats sont à aligner avec les concepts de l'ontologie du domaine.

Cet alignement est effectué soit directement avec l'ontologie soit indirectement en utilisant le Web. L'ontologie est ensuite peuplée avec les termes candidats alignés qui sont exploités dans les prochains alignements.

2.1. Description de l'ontologie

Soit $\mathcal{O}(\mathcal{C}, \mathcal{R}, \preceq, \mathcal{S}, \mathcal{A}, \mathcal{L}_{\mathcal{EX}})$ l'ontologie de domaine où $\mathcal{C}$ est l'ensemble des concepts, $\mathcal{R}$ est l'ensemble des relations entre les concepts, $\preceq$ désigne la relation de subsomption entre les concepts et entre les relations. $\mathcal{S}$ définit le domaine et le codomaine de chaque relation et $\mathcal{A}$ est un ensemble d'axiomes et de règles définis sur les concepts et les relations.

$\mathcal{L}_{\mathcal{EX}}(\mathcal{L}, \mathcal{T}, prefLabel, altLabel, hasTerm, hastermNe)$ définit l'ensemble $\mathcal{L}$ des labels de concept et l'ensemble $\mathcal{T}$ de termes et entités nommées décrivant les concepts du domaine. Chaque concept $c \in \mathcal{C}$ est lié à un label préféré via la propriété $prefLabel$ [1] et les autres labels via $altLabel$[1] appartenant à $\mathcal{L}$. Chaque concept $c \in \mathcal{C}$ est relié à des termes par la propriété $hasTerm$ et à des entités nommées par $hastermNe$. Termes et entités nommées appartiennent à $\mathcal{T}$. Nous supposons que les ensembles $\mathcal{L}$ et $\mathcal{T}$ sont respectivement initialisés par un ensemble de labels et un ensemble de termes choisis par l'expert du domaine. Par exemple les labels et les termes sélectionnés pour le concept *Topic* du domaine informatique sont les suivants :
prefLabel(c, "Topic"), altLabel(c, "field"), altLabel(c, "area"), altLabel(c, "theme"),
hasTerm(c, "communications protocol"), hasTerm(c, "data encryption"),
hasTerm(c, "information"), hasTerm(c, "object-oriented programming language")

L'ensemble des termes $\mathcal{T}$ est enrichi par des termes candidats extraits au fur et à mesure que les documents sont traités. étant donné que cet enrichissement est automatique, certains termes ne sont pas pertinents, c'est pourquoi nous les distinguons des labels. L'expert peut valider l'ontologie et décider que certains termes deviennent des labels.

2.2. Algorithme *Extract-Align*

L'approche d'extraction et d'alignement de *SHIRI* procède d'une manière incrémentale. Chaque appel à *Extract-Align* permet de traiter un sous-ensemble de termes. Plus précisément, à chaque invocation, l'algorithme est appliqué à un sous-ensemble de termes provenant de documents appartenant au même domaine, à l'ontologie $\mathcal{O}$ de ce domaine, à un ensemble de patterns $\mathcal{P}$ et à un ensemble $Processed$ contenant les termes traités dans les étapes précédentes. L'algorithme distingue deux types de patterns : les patterns syntaxiques d'entités nommées et les patterns syntaxiques de termes. Ces deux types de patterns sont utilisés pour extraire un ensemble de termes candidats noté $\mathcal{I}$ (voir l'exemple du tableau 1). Chaque terme $t \in \mathcal{I}$ est identifié par la séquence des mots numérotés en fonction de leur ordre d'apparition dans le document. Ces termes doivent être alignés avec l'ensemble des labels $\mathcal{L}$ et l'ensemble des termes $\mathcal{T}$ appartenant à l'ontologie $\mathcal{O}$.

A chaque étape, l'algorithme tente d'aligner directement les termes de $\mathcal{I}$ avec l'ontologie, autrement en utilisant le web. Chaque étape enrichit l'ensemble $\mathcal{T}$ des termes du domaine et des entités nommées. Donc le nombre d'appels au Web diminue avec le traitement des documents. C'est aussi la raison pour laquelle l'ensemble des termes traités mais non alignés sont gardés dans $Processed$.

1. Propriétés définies dans SKOS : Simple Knowledge Organization System

La fonction $alignTerm(t)$ est appliquée à chaque $t \in \mathcal{I}$ et retourne un jeu de concepts $\mathcal{C}_\sqcup \subset \mathcal{C}$ si elle réussit. Alors, t est ajouté à $\mathcal{T}$ et relié à tous les concepts $c \in \mathcal{C}_\sqcup$ par la relation $hasTerm$ ou $hastermNe$ selon le type du pattern correspondant (voir l'exemple ci-dessous). Invoquée, la fonction $alignTerm(t)$ utilise des mesures de similarité qui sont appropriées pour comparer deux entités nommées ou deux termes.

Les termes non alignés sont soumis au Web comme dans l'approche CPankow [3] : des patterns lexico-syntaxiques de Hearst pour l'hyponymie [5] sont utilisés pour construire des requêtes contenant le terme non aligné t. Ces requêtes sont soumises à un moteur de recherche afin de trouver un ensemble de labels candidats noté $\mathcal{L}_\sqcup$. Pour chaque $l \in \mathcal{L}_\sqcup$, la fonction $webAlign(l)$ est appelée et retourne un ensemble de concepts $\mathcal{C}_\sqcup \subset \mathcal{C}$. Si $webAlign(l)$ réussit, alors l et t sont ajoutés à $\mathcal{T}$. t est lié à chaque $c \in \mathcal{C}_\sqcup$ par la relation $hasTerm$ ou par la relation $hastermNe$ selon le type du pattern utilisé. l est reliée à chaque $c \in \mathcal{C}_\sqcup$ par la relation $hasTerm$. Comme l est extrait automatiquement, il est considéré comme un terme.

En outre, les candidats termes de $\mathcal{I}$ sont traités de façon progressive de la plus longue à la plus courte. Nous supposons que le terme est plus précis et plus significatif que les sous-termes qu'il contient. Par exemple "*distributed databases*" est plus précis que "databases". Mais pour un terme comme "*Interoperability of data on the Semantic Web*", l'alignement échouera très probablement. Nous notons un terme de longueur k se situant à la position i dans le document comme une séquence de k mots : $t_i^k = w_i w_{i+1} ... w_{i+k-1}$, où w_{i+j} désigne le mot à la position $i + j$, $j \in [0, k-1]$. Nous notons $\mathcal{I}^k = \{t_i^k, i \in [1, N]\}$ l'ensemble des termes extraits de longueur k. k varie de len (longueur maximale des termes en nombre de mots) à 1.

A l'itération k, l'algorithme traite les termes de $\mathcal{I}^k$ et $\mathcal{I} = \bigcup_{i=1}^{k} \mathcal{I}^k$. Nous disons que le terme $t_{i_2}^{k_2}$ est inclus dans $t_{i_1}^{k_1}$ si $k_2 < k_1$ et $i_2 \in [i_1, i_1 + k_1 - 1]$. Lorsque le système aligne un terme $x \in \mathcal{I}^k$ alors $\forall y \in \bigcup_{i=1}^{k-1} \mathcal{I}^i$ inclus dans x, y est supprimée de $\mathcal{I}$.

Exemple : Considérons le texte dans le tableau 1 et les deux patterns $P_t^1 = JN$ et $P_t^2 = N$, où J désigne un adjectif et N un nom, les termes extraits sont les suivants : $\mathcal{I}^1 = \{Areas_{71}, databases_{76}, intelligence_{79}, workshop_{81}, databases_{86}, Intelligence_{87}, areas_{88}\}$ et $\mathcal{I}^2 = \{distributed_{75}\, databases_{76}, artificial_{78}\, intelligence_{79}\}$. Les termes "$distributed_{75}\, databases_{76}$" et "$artificial_{78}\, intelligence_{79}$" sont alignés avec le concept "$Topic$". Ainsi, nous supprimons les candidats termes "$databases_{76}$" et "$Intelligence_{79}$" de $\mathcal{I}^1$.

Nous obtenons trois types de sorties résultant de l'invocation de l'algorithme *Extract-Align* : (1) des triplets RDF qui enrichissent l'ontologie avec les termes ou les entités nommées décrivant les concepts en utilisant les relations $hasTerm$ et $hasTermNe$, (2) des triplets RDF référençant des unités structurelles des documents, les concepts que ces unités contient en utilisant la relation $containInstanceOf$ et les valeurs des termes ou entités nommées correspondant en utilisant la relation $hasValueInstance$ et (3) l'ensemble de tous les termes traités.

3. Validation de l'Algorithme Extract-Align

Soit $\mathcal{O}$ l'ontologie du domaine des *"appel à participation à des conférences en informatique" ("call for papers" en anglais)*. Les entités nommées sont constituées par les *événements* (i.e. *conférences*, *workshops*), les *personnes*, leurs *affiliations* (i.e. *équipe*, *laboratoire* et/ou *université*), et les *lieux* (*université*, *ville* ou *pays*) où ont lieu les événements. Chaque concept est décrit par un label préféré et un ensemble de labels alternatifs.

Texte Original	Termes Candidats Extraits
... Areas$_{71}$ of$_{72}$ interest$_{73}$ are$_{74}$ distributed$_{75}$ databases$_{76}$ and$_{77}$ artificial$_{78}$ intelligence$_{79}$. The$_{80}$ workshop$_{81}$ SEMMA$_{82}$ focuses$_{83}$ also$_{84}$ on$_{85}$ databases$_{86}$. Intelligence$_{87}$ areas$_{88}$ ...	... [Areas$_{71}$] of$_{72}$ interest$_{73}$ are$_{74}$ [distributed$_{75}$ [databases$_{76}$]] and$_{77}$ [artificial$_{78}$ [intelligence$_{79}$]]. The$_{80}$ [workshop$_{81}$] SEMMA$_{82}$ focuses$_{83}$ also$_{84}$ on$_{85}$ [databases$_{86}$]. [Intelligence$_{87}$] [areas$_{88}$] ...

Tableau 1. *Exemples de termes candidats extraits*

Par exemple, *scientist* et *people* sont des labels alternatifs du concept *Person*. L'expert a également exploité WordNet et a sélectionné un ensemble de 353 termes du domaine tels que {*Communications protocol, data encryption, information, ...*} qui sont liées au concept *Topic* par la propriété *hasTerm*. Le corpus que nous avons collecté est composé de 691 documents HTML (250542 mots après pré-traitement).

Les entités nommées sont extraites automatiquement de la collection de documents par une technique spécialisée, Senellart [2], qui exploite DBLP (Digital Bibliography and Library Project) pour identifier avec précision les noms de personnes et les dates. Nous utilisons également l'ensemble des patterns syntaxiques de C-Pankow [3] pour extraire d'autres instances d'entités nommées. Quant aux termes, ils sont extraits selon les patterns définis dans Arppe [4].

Pour extraire les labels Web des termes candidats, un ensemble de requêtes est construit avec chaque terme ou entité nommée. Les requêtes, comme dans l'approche C-Pankow [3], sont construites en utilisant les patterns de Hearst, de la copule et de la définition (phrase nominale introduite par le déterminant The). Les labels Web sont considérés comme bons lorsque la valeur de la mesure de similarité entre le document contenant l'entité nommée et celui d'où est extrait le label est supérieure à 0.2. Pour l'alignement des termes candidats ou des labels Web avec les concepts de $\mathcal{O}$, nous utilisons l'outil Taxomap [12]. Son objectif est de découvrir des correspondances entre les concepts de deux taxonomies. Il effectue un alignement orienté d'une ontologie source vers une ontologie cible. L'outil s'appuie sur des techniques terminologiques et structurelles appliquées de façon séquentielle. Nous exploitons uniquement les stratégies terminologiques de Taxomap c'est à dire celles utilisant des mesures de similarité basées sur la syntaxe appliquées aux labels des concepts et des termes (inclusion de terme, similitude par n-gramme, ...). Pour l'alignement des entités nommées, la mesure de similarité considérée entre les termes est l'égalité stricte.

Par exemple, dans nos expériences, le terme "*reinforcement learning*" est directement aligné avec le concept $Topic$ car ce dernier comportait d'ejà du terme $learning$. Le terme "*World Wide Web*" n'a pas été aligné directement avec $\mathcal{O}$. Un des labels candidats obtenus avec le Web est "*information resources*". Taxomap l'aligne avec le terme $information$ relié au concept $Topic$. Les termes "*Reinforcement learning*", "*information ressources*", "*World Wide Web*" sont alors ajoutés à $\mathcal{T}$ et reliés au concept $Topic$ par la propriété $hasTerm$.

Le tableau 2 montre les résultats que nous avons obtenus avec les patterns d'entités nommées. Nous présentons les mesures de précision et de rappel pour les entités nommées qui sont alignés soit directement, soit indirectement en utilisant des labels provenant du Web. Puisque la granularité de l'annotation de *Shiri-Annot* est l'unité structurelle, nous considérons qu'une entité nommée incomplète est correcte car le nom complet apparaît dans l'unité structurelle d'où elle est extraite. Par exemple, le terme "*International Confe-*

Concepts	Patterns d'Entités Nommées					Patterns de Termes		
	Alignés Avec $\mathcal{O}$	Utilisant le Web	Précision	Précision avec NE incomplètes	Rappel	Alignés Avec $\mathcal{O}$	Précision	Rappel
Affiliation	0	1317	84.18%	86.07%	70.83%	165	96.97%	91.95%
Location	0	1097	98.53%	99.02%	91.86%	143	80.42%	78.77%
Person	745	362	89.47%	90.85%	79.5%	206	63.59%	59.01%
Event	0	741	64.35%	86.13%	84.47%	80	65.00%	65.00%
Date	456	0	97.58%	97.58%	74.17%	-	-	-
Topic	-	-	-	-	-	276	65.58%	59.34%

Tableau 2. *Résultats des Patterns d'Entités Nommées et de Termes*

rence" est incomplet, mais l'unité structurelle la contenant contient le nom complet qui est "*International Conference on Web Services*". Approximativement, le Web a permis d'obtenir 74% des entités nommées alignées. Toutes les affiliations, les événements et les lieux sont retrouvés grâce au Web. En outre, le tableau montre que, en prenant en compte les entités nommées incomplètes, la précision augmente surtout pour les événements. Ces entités nommées sont souvent partiellement extraites en raison de leur longueur et de leur complexité. Le tableau 2 montre aussi que, pour les patterns de termes, le concept $Topic$ est enrichi à 78% de termes. D'autres entités nommées ont été également trouvées avec une bonne précision et un bon rappel par exemple les affiliations qui sont souvent décrites avec des termes complexes. Le tableau 3 montre que beaucoup de termes se répètent plusieurs fois. Ceci est dû au fait que les documents sont du même domaine. Cependant, comme l'algorithme traite les termes du plus long au plus court, la plupart d'entre eux ne sont pas traités car ils sont inclus ou équivalents à des termes traités (cf tableau 3).

Longueur	1	2	3	4	5	7	**Total**
Termes Extraits	101430	32712	17912	5704	966	104	158828
Termes Extraits (différents)	14413	15806	10020	3797	602	48	44680

Tableau 3. *Nombre de Termes Extraits par Taille*

La figure 2 montre : (1) L'évolution du nombre de termes extraits par rapport au nombre de documents (par dix) (2) l'évolution du nombre d'appels Web en fonction du nombre de documents (par dix). Ces résultats montrent que le nombre d'invocations du Web diminue avec le nombre de documents traités. Cela s'explique par le comportement incrémental de l'algorithme *Extract-Align* : (1) plus l'ontologie est peuplée par de nouveaux termes, plus un candidat terme est susceptible d'être aligné directement et (2) tous les termes passés au Web et dont les tentatives d'alignement échouent sont stockés dans les données traitées (Processed).

4. Conclusions et perspectives

Dans ce papier, nous avons présenté une approche automatique, non suppervisée et guidée par une ontologie pour l'extraction, l'alignement et l'annotation sémantique d'éléments balisés de documents. L'algorithme *Extract-Align* procède de façon incrémentale et peuple l'ontologie avec des termes décrivant des instances de concepts du domaine et réduit ainsi progressivement le besoin d'accéder à des ressources externes telles le Web.

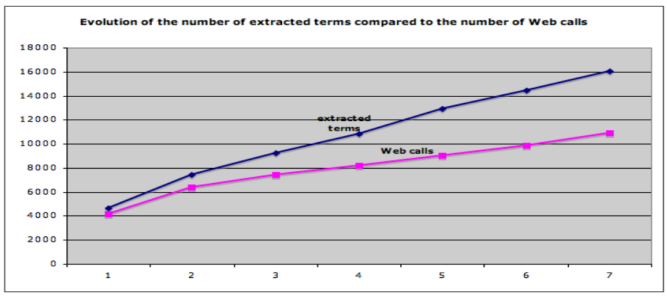

Figure 2. *Evolution des nombres de termes extraits et des appels Web sur le nombre de documents (par 10)*

Nous avons expérimenté et validé notre approche sur un corpus de documents HTML concernant les appels à communication à des conférences en informatique et les résultats obtenus sont prometteurs. Ces résultats montrent comment l'ontologie est enrichie et comment le nombre de termes (ou entités nommées) alignés directement avec l'ontologie augmente. L'ontologie construite peut être validée par un expert du domaine afin de sélectionner parmi les termes ceux qui sont à supprimer et ceux qui deviennent des labels de concept.

Une perspective à court terme est l'exploitation du modèle d'annotation pour la reformulation des requêtes en vue de les adapter aux différents niveaux de précision des annotations. Une autre perspective à plus long terme est d'étudier comment pondérer les triplets représentant les annotations. Nous prévoyons également d'appliquer notre approche à d'autres domaines comme le commerce électronique.

5. Bibliographie

[1] THIAM M. AND PERNELLE, N. AND BENNACER, N., « Contextual and Metadata-based Approach for the Semantic Annotation of Heterogeneous Documents », *Proceedings of the First International Workshop on Semantic Metadata Management and Applications, SeMMA 2008*, n° 346, 2008.

[2] SENELLART P., « Understanding the Hidden Web », *PHD Thesis, University of Paris 11*, n° XVI-146 p., 2007

[3] CIMIANO P., HANDSCHUH S., STAAB S., « Gimme'The Context : Context Driven Automatic Semantic Annotation With C-PANKOW. », *Proceedings of the 14th international conference on World Wide Web*, n° 1-59593-046-9, 2005.

[4] ANTTI ARPPE, « Term Extraction from unrestricted Text », *the Nordic Conference on Computational Linguistics*, vol. NoDaLiDa, n° 1995.

[5] HEARST, M. MARTI, A., « Automatic acquisition of hyponyms from large text corpora », *Proceedings of the 14th International conference on Computational linguistics, pages 539–545 1992*.

[6] WILLIAM W. COHEN, PRADEEP RAVIKUMAR, AND STEPHEN E. FIENBERG., « A comparison of string distance metrics for name-matching tasks », *In IIWeb 2003, pages 73-78*.

[7] CRESCENZI V., MECCA G., MERIALDO P., « RoadRunner : Towards Automatic Data Extraction from Large Web Sites », *Very Large Data Bases Conference 2001*.

[8] DAVULCU H., VADREVU S. AND NAGARAJAN S., « OntoMiner : Automated Metadata and instance Mining from News Websites », *The International Journal of Web and Grid Services (IJWGS)*, n° 1, No. 2, pp. 196-221, Inderscience Publishers, 2005.

[9] BORISLAV P., ATANAS K., ANGEL K., DIMITAR M., DAMYAN O., MIROSLAV G., « KIM - Semantic Annotation Platform », *Journal of Natural Language Engineering*, vol. 10, n° issue 3-4, Cambridge University Press, pages 375-392, 2004.

[10] BAUMGARTNER R. AND FLESCA S. AND GOTTLOB G, « Visual Web Information Extraction with Lixto », *The VLDB Journal, pages 119-128, 2001 citeseer.ist.psu.edu/baumgartner01visual.html*

[11] SODERLAND S., « Learning information extraction rules for semi-structured and free text », *Machine Learning, 34(1-3) :233-272, 1999.*

[12] FAYCAL HAMDI, HAIFA ZARGAYOUNA, BRIGITTE SAFAR, CHANTAL REYNAUD, « TaxoMap in the OAEI 2008 alignment contest, Ontology Alignment Evaluation Initiative (OAEI) 2008 Campaign », *Int. Workshop on Ontology Matching, 2008.*

[13] JÉRÔME EUZENAT AND PAVEL SHVAIKO, « Ontology matching », *Ontology matching-Springer-Verlag, 2007*, n° isbn 3-540-49611-4.

[14] ETZIONI, M. CAFARELLA, D. DOWNEY, S. KOK, A. POPESCU, T. SHAKED, S. SODERLAND, D. WELD, AND A. YATES, « Unsupervised named-entity extraction from the web : An experimental study », *Artificial Intelligence*, n° 165(1) :91-134, 2005.

[15] R. NAVIGLI, P. VELARDI, « Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites », *Computational Linguistics*, n° 30(2), MIT Press, 2004, pp. 151-179.

[16] DROUIN, PATRICK, « Term extraction using non-technical corpora as a point of leverage », *In Terminology* n° vol. 9, no 1, p. 99-117.2003