Extraction de concepts guidée par le contexte

Lobna Karoui, Marie-Aude Aufaure

Ecole Supérieure d'Electricité
Plateau de Moulon 3 rue Joliot Curie
91192 Gif-sur-Yvette cedex, France
{Lobna.karoui, Marie-Aude.Aufaure}@supelec.fr
http://www.supelec.fr/ecole/si/pers.html

Résumé. Les ontologies constituent la brique supportant les échanges et le partage des informations en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Dans cet article, nous présentons une méthode d'extraction de concepts ontologiques utilisant un algorithme de clustering non supervisé et guidé par le contexte à partir de pages Web. Notre méthode est basée sur une approche unifiée intégrant des dimensions complémentaires pour l'acquisition de connaissances conceptuelles. En particulier, nous exploitons les caractéristiques structurelles des documents HTML afin de localiser et de définir un contexte approprié pour chaque terme en respectant ses différentes positions dans le corpus. Notre définition contextuelle permet de sélectionner les co-occurrents sémantiquement proches et de définir une mesure de pondération appropriée pour chaque couple de termes. Notre méthode se base sur une évaluation interactive et incrémentale de la qualité des clusters par l'utilisateur. Nous l'avons expérimentée sur un corpus du domaine portant sur le tourisme. Les premiers résultats obtenus montrent bien que la prise en compte du contexte des termes guidant le clustering améliore considérablement la pertinence des concepts extraits

1 Introduction

Les ontologies constituent la brique supportant les échanges, le partage et la recherche d'information en étendant l'interopérabilité syntaxique du web en une interopérabilité sémantique. Elles permettent de représenter un ensemble de concepts formellement définis, acceptés par une communauté d'utilisateurs. Selon les domaines et les besoins applicatifs, les ontologies seront plus ou moins riches, allant de simples métadonnées, à des taxonomies jusqu'à de véritables bases de connaissances. Elles servent de squelette de structuration sémantique des données et représentent une valeur ajoutée pour leur manipulation, leur traitement et leur interrogation. La réalisation du web sémantique dépend de la construction des ontologies et de leur déploiement. Le problème majeur concerne les coûts de cette construction pour un grand nombre d'ontologies de domaines et d'applications. La classification automatique est un angle d'attaque permettant de déterminer des regroupements des données en classes et d'extraire des concepts du domaine et leurs relations ; ce qui constitue un point crucial pour cette tâche lourde et complexe que représente la construction d'ontologies

Dans cet article, nous présentons une approche unifiée d'extraction de connaissances à partir de documents HTML en vue de construire une ontologie du domaine. En effet, l'abondance et l'importance de pages HTML comme une riche source d'information sont un fait

indéniable avec l'expansion continue du Web. Notre approche se base sur une architecture composée principalement d'une étape de prétraitement, une étape de traitement et une étape de formalisation et d'évaluation. L'étape de prétraitement utilise différents modules pour la constitution, la représentation, le traitement et l'analyse du corpus. L'étape de traitement exploite des dimensions complémentaires telles que les différentes analyses et la représentation relationnelle issues de l'étape précédente et en particulier les caractéristiques structurelles du corpus afin d'améliorer la pertinence des concepts ontologiques extraits. Le processus d'extraction des concepts ontologiques est détaillé dans cet article ; il a comme objectif principal l'amélioration du processus de sélection des co-occurrents sémantiquement proches pour chaque terme et la pondération des couples de termes. Pour atteindre cet objectif, la structure des documents HTML est exploitée, en particulier les relations entre les balises HTML, pour définir un contexte approprié pour chaque terme en respectant ses différentes positions dans le corpus. Notre définition contextuelle, déduite des différentes analyses de l'étape de prétraitement, est représentée par une hiérarchie contextuelle permettant de découvrir les différentes associations d'un terme avec les autres termes et de raffiner sa pondération ainsi que sa similarité avec ses co-occurrents. Afin d'obtenir des classes de termes, nous avons défini une méthode de décomposition guidée par le contexte et utilisant un algorithme de clustering non supervisé à savoir les cartes de kohonen. Notre approche se base sur une évaluation interactive et incrémentale de la qualité des clusters par l'utilisateur. Nous l'avons expérimentée sur un corpus du domaine portant sur le tourisme. Les premiers résultats obtenus montrent que la prise en compte du contexte des termes guidant le clustering améliore considérablement la pertinence des concepts extraits.

Dans la section suivante nous présentons quelques travaux similaires. Dans la section 3, nous exposons notre approche et nous détaillons le module d'extraction de concepts ontologiques. La section 4 présente nos expériences, les résultats obtenus ainsi que leur évaluation. Dans la section 5, nous concluons sur le travail présenté et nous présentons nos perspectives par rapport à l'architecture générale présentée.

2 Travaux similaires

Nombreuses sont les recherches qui ont utilisé la structure HTML. Par exemple, Buyukkoten (Buyukkoten et al, 2001) discute une méthode d'extraction du contenu des documents HTML en transformant une page Web en une hiérarchie d'unités textuelles sémantiques. Ces unités sont définies en analysant les caractéristiques syntactiques d'un document HTML comme le texte contenu dans les balises (<p>, <frame>, etc.). Egalement, Kiyota et Kurohashi (kiyota et kurohashi, 2001) présentent un extracteur de phrases et un générateur de résumés basé sur une analyse syntactique, la méthode du tf.idf de Salton et la structure HTML. Ils considèrent que les mots-clés appartenant aux titres et aux sous-titres sont plus importants que les mots-clés qui apparaissent dans les listes et les tables c'est pourquoi ils leur associent un poids plus important. Il existe aussi (Cai et al, 2004) qui ont comme but d'améliorer la pertinence de la recherche documentaire dans le Web.

Egalement, il existe différents travaux qui ont développé des méthodologies ou des techniques d'automatisation pour la construction d'ontologies. Dans (Faure et al, 1998), les auteurs présentent ASIUM, un système d'apprentissage à partir de textes techniques. Des clusters de base sont formés par des termes apparaissant avec le même verbe et avec le même rôle syntaxique ou la même préposition fournie par un analyseur syntaxique. Les auteurs

ayant développé une méthode de clustering basée sur ces idées obtiennent en sortie une ontologie avec des relations taxonomiques. Le système WebOntEx (Hahn et Elmasri, 00) a pour objectif d'extraire semi-automatiquement des ontologies en analysant les pages web appartenant au même domaine. L'extraction des connaissances est basée sur les balises HTML (<b>, <h1>), les balises de lemmatisation (verbe, nom) et les balises conceptuelles (entité, attribut) en utilisant WorldNet et la méthode de programmation logique inductive. L'un des problèmes majeur de cette méthode est la lourde tâche manuelle à réaliser au cours de laquelle l'utilisateur devra développer le cœur de l'ontologie et extraire les patterns génériques à partir des pages Web. OntoMiner (Davulcu et al, 98) analyse des ensembles de sites web d'un domaine spécifique et génère une taxonomie de concepts particuliers ainsi que leurs instances. Cet outil utilise les régularités HTML des documents pour générer une structure hiérarchique codé en XML. Maedche et Staab (2001) proposent un environnement d'apprentissage d'ontologies (Text-To-Onto) basé sur une architecture générale de découverte de structures conceptuelles à partir de différentes sources (XML, DTD, schéma de BD, etc.). Cet environnement possède une librairie de méthodes d'apprentissage et des outils linguistiques pour extraire des concepts, leurs relations taxonomiques et non taxonomiques. DODDLE II (Sugiura, 2004) est un environnement de développement d'ontologies permettant d'extraire des relations taxonomiques en utilisant à la fois les termes du domaine et WorldNet. Pour découvrir des relations non taxonomiques, les auteurs retrouvent les collocations d'un ensemble de 4 termes d'où la notion de collocation du domaine. Pour obtenir les relations non taxonomiques, les auteurs sélectionnent les paires de concepts dont le produit de leurs vecteurs a dépassé un seuil fixé par l'utilisateur et utilisent également les règles d'associations afin d'extraire ces relations. Cette méthode a été testée sur un ensemble réduit de contrats internationaux de vente de marchandises. Une extension de cet environnement, appelée DODDLE-R (Sugiura, 2004), génère des structures de langage naturel, utilise ces structures et les relations non taxonomiques de DODDLE II pour construire un modèle RDF et se base sur les relations taxonomiques afin de déduire les classes de la hiérarchie résultante. SYNDIKATE (Hahn et Romacker, 2001) est un système pour l'acquisition automatique de connaissances à partir de textes allemands. L'approche d'apprentissage de concepts résultant de procédures de compréhension de textes se base sur deux sources : (1) les connaissances à priori (lexique et ontologie) (2) les structures syntaxiques dans lesquelles des objets lexicaux inconnus apparaissent. Ce système extrait également des relations non taxonomiques à partir de l'interprétation sémantique du texte. ASIUM est un système qui dépend de la structure des documents analysés. Quant à Text-to-Onto, l'une de ses faiblesses est l'existence de bruit. En effet, certains termes, n'ayant pas de relations sémantiques et existants dans la même classe, peuvent nuire au processus d'interprétation de la classe et induire une perturbation au sein du groupe de mots. Pour DODDLE, il nécessite l'existence du vocabulaire du corpus à étudier dans WorldNet à défaut l'opération de matching est impossible à réaliser. Cette dépendance à une connaissance à priori, pareille pour WebOntoEx, tout en étant un atout pour extraire une connaissance plus pertinente, représente une limite. SYNDIKATE dépend à la fois d'une ontologie générique du domaine et d'un lexique sémantique. Concernant la notion de contexte, ces systèmes focalisent leur analyse sur le contexte syntaxique d'un terme à l'exception de DODDLE qui considère que le contexte d'un groupe de 4 mots est l'ensemble de 4 mots qui le précède directement. Pour le système OntoMiner, les auteurs se basent sur les régularités existantes dans les documents HTML alors qu'en réalité la majorité des documents Web manquent de régularités HTML.

3 Architecture du système de découverte de connaissances

Ayant comme tâche la recherche, la découverte et la structuration des connaissances conceptuelles à partir des pages Web en vue de construire une ontologie de domaine, nous avons définit une architecture unifiée de découverte de connaissances pour le web sémantique. Les composants de notre système assistent l'ingénieur de connaissances lors de la constitution et du traitement du corpus, la représentation des données, l'analyse du corpus, la découverte et la structuration des connaissances en produisant une ontologie de domaine ou en la raffinant. Notre but est d'aider l'utilisateur lors de la construction de l'ontologie grâce aux mécanismes interactifs de notre système. En exploitant les fonctionnalités de notre système, l'utilisateur pourra intervenir lors de l'étape de prétraitement (supprimer des mots ou des documents, enrichir le corpus, etc.) et l'étape de traitement (donner des noms aux classes de mots, etc.). En plus, l'expert du domaine pourra évaluer, soit individuellement ou en collaboration avec d'autres experts, les classes de mots, l'ontologie, etc. Le but de notre système est de produire des ontologies de domaines qui pourront être intégrées dans des systèmes de recherche d'information afin de spécifier, restreindre et confirmer la requête en utilisant les concepts de l'ontologie au lieu des mots clés. Ainsi, nous définissons cet environnement de travail pour la découverte des concepts en exploitant profondément les caractéristiques des documents. Dans les sections suivantes, nous présentons brièvement les différents modules conçus dans l'architecture.

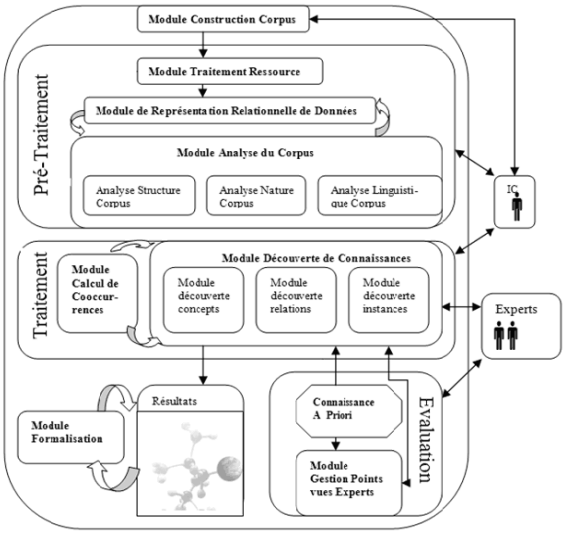

FIG. 1 – *Architecture d'un système de découverte de connaissances pour le Web Sémantique*

3.1 Modules de constitution, de traitement, de représentation et d'analyses du corpus

Dans cette section, nous expliquons les différents modules inclus dans la phase de prétraitement du corpus. Ces modules forment les premières fonctionnalités de notre système.

Modules de constitution et de traitement du corpus. Notre source d'information est un ensemble de documents HTML portant sur le même domaine. Ce choix se justifie par l'énorme quantité de pages HTML disponibles sur le Web. Le document est composé d'un contenu textuel riche en balises de structure, de liens entre les documents et de balises permettant d'associer différents styles typographiques aux termes. L'ensemble des pages HTML est nettoyé permettant ainsi de conserver que le texte associé aux balises que nous avons définies et catégorisées comme balises clés (<Keywords>, <TITRE_URL>, etc.), balises bloc (<h1>, <table>) et balises en-lige(<b>, <i>). Par exemple les termes qui suivent un lien hypertexte sont suffisamment représentatifs du contenu pointé par ce lien. Ils sont associés à la balise <TITLE_URL> (Exemple : <a href="http://www....> Activité sportives à Grenoble ==> <TITRE_URL> Activité sportives à Grenoble). Ce module permet aussi de corriger le codage des caractères spéciaux (é --> é).

Module de représentation des données. L'objectif de ce module est de représenter l'intégralité du corpus traité à l'aide d'une table relationnelle. Les termes du corpus sont les tuples de cette relation et leurs propriétés (type grammatical, son lemme, etc.) constituent les attributs. Cette représentation permet tout type d'interrogation SQL de la source d'informations par l'utilisateur. Par exemple, l'utilisateur pourra, à partir d'un concept type « hébergement », revenir à son contexte source (par exemple les documents dans lesquels il a été cité ainsi avoir une idée sur ses co-occurrents comme hôtel, gîte, etc.).

Module d'analyse du corpus. L'objectif de ce module est d'accomplir différentes analyses afin d'évaluer, d'enrichir et de caractériser notre corpus. La structure est étudiée par rapport aux unités structurelles telles qu'un paragraphe, une phrase et aux diverses balises (balise bloc, balise clefs, balises en-ligne). L'analyse de la nature du corpus est réalisée grâce à deux méthodes complémentaires. La première méthode est l'analyse factorielle de correspondance (AFC) (Benzecri, 1973). Le but de cette analyse est d'examiner si la distribution des termes se fait autour d'un même point ou s'il existe un ensemble de nuages séparés (des classes thématiques). La seconde méthode est la mesure du TF*IDF (Joachims, 1997) permettant de représenter chaque document par un vecteur de fréquences de mots qu'ils contiennent. Nous l'appliquons pour vérifier la répartition de la terminologie du domaine entre les documents du corpus. Le corpus est analysé sur le plan linguistique grâce à une analyse morphologique et syntaxique. L'analyse morphologique permet la reconnaissance d'un mot et sa structure interne. Elle est faite par le biais de l'outil Tree Tagger (Schmid, 1994) afin de retrouver le type grammatical de chaque terme et son lemme. L'analyse Syntaxique permet de définir les structures grammaticales du corpus (groupe nominal, groupe verbal, etc.). Elle se fait en utilisant l'outil Syntex (Frérot et al, 2003). Ces informations sont utiles pour découvrir le les patrons candidats pour raffiner le contexte structurel à un niveau linguistique approprié.

3.2 Module de découverte des connaissances

Dans cette section, nous nous focalisons sur la découverte des « concepts ontologiques » et la manière dont ils sont extraits en utilisant la notion de contexte.

Un « Contexte » dans (Brézillon, 1999) est défini comme : « Ce qui contraint une résolution de problème sans y intervenir explicitement ». Nous définissons le contexte comme un ensemble de circonstances (situations) qui entourent l'objet d'étude et reflète son environnement concret. Il fournit un support pour l'activité d'apprentissage et pour l'interprétation

sémantique. Dans notre cas, l'objet étudié est le terme, l'activité d'apprentissage est le clustering, l'interprétation sémantique est l'opération d'évaluation et de labellisation des classes de termes et la définition du contexte est déduite des analyses structurelles du corpus. Notre contexte est représenté par une hiérarchie par rapport à laquelle il est instancié en respectant les différentes positions d'un terme dans les balises HTML (contexte structurel). Le contexte structurel est basé sur l'existence ou non de relations entre les balises html. Dans les sections suivantes, nous expliquons et expérimentons ce contexte structurel.

Module de calcul des occurrences. Ce module est chargé d'un processus de sélection des cooccurrents sémantiquement proches et de pondération des mots. Généralement, si deux mots apparaissent ensemble dans une expression ou un document, nous leur attribuons l'un ou l'autre la valeur 1 (présent) ou une autre valeur entre 0 et 1 (fréquence d'apparition). Dans les deux cas, le contexte pourrait être approprié à quelques mots, mais pas à tous les mots d'un corpus. Par exemple, soient les trois phrases: "les Etats-Unis possèdent divers possibilités de logement. Dans la région Nord, ils offrent des hôtels et des résidences. Il existe également diverses activités de loisir." Si nous fixons le contexte à une phrase, nous constatons que 'logement' et 'hôtel', qui appartient au même concept, ont la valeur 0 parce qu'ils n'appartiennent pas à la même phrase. Si nous fixons le contexte à une fenêtre dont la taille est 14 mots, nous constatons que 'résidence' et 'loisir' ont la valeur 1 cependant ils n'appartiennent pas au même concept (la résidence appartient au concept 'logement' et 'loisir' est un autre concept). Pour ces raisons, nous avons raffiné le contexte afin de prendre davantage en compte la position du mot. La question qui se pose maintenant est : « comment attribuer une pondération à un terme reflétant son importance dans le domaine et mesurant la pertinence de ses co-occurrents dans la relation les liant sémantiquement» ?

L'existence d'une relation structurelle entre les éléments HTML peut révéler une relation sémantique implicite entre les termes associés. Le fait d'instancier le contexte par rapport au lien structurel permet de mieux cerner et révéler les concepts relatifs aux termes apparaissant dans, par exemple, les balises <h1> <p> ; <caption> <td> (titre d'un tableau cellule d'un tableau) ; <TITLE_URL> (titre d'un lien hypertexte) les titres d'une partie d'un document ; <TITLE_URL> les titres du document référencés ; etc.

Nous distinguons deux types de lien structurel : un lien physique qui dépend de la structure du document HTML (entre la balise <h1> et la balise <p> associée) et un lien logique qui n'est pas visuel puisque les éléments ne sont pas nécessairement consécutifs (entre <TITLE_URL> et les titres du document référencé par exemple). Pour caractériser les liens entre les balises, nous avons défini deux notions : « hiérarchie contextuelle » (H.C.) basée sur les balises HTML et « cooccurrence par liaison ».

Une hiérarchie contextuelle est une hiérarchie de balises. Elle illustre les relations possibles dans les documents HTML et entre eux. La cooccurrence est définie par le fait d'avoir deux mots dans le même contexte (paragraphe, texte, etc.). Dans notre étude, le contexte est variable et il est déduit de la hiérarchie contextuelle. En respectant la structure de H.C, nous établissons des liaisons entre les termes si :

- Les termes sont encadrés par la même balise bloc (TAB. 1 : Exemple 1). Dans ce cas, on parle de *cooccurrence par voisinage* et le contexte est fixé à la balise en soit (<H1>).
- Les termes sont encadrés par des balises qui à leur tour sont reliées par un lien physique ou logique schématisé dans la hiérarchie contextuelle. Dans ce deuxième cas

(TAB. 1 : Exemple 2), nous parlons de *cooccurrence par liaison* (balises non consécutives) et le contexte est l'association des deux balises (<title> + <h1>).

Exemple 1	Exemple 2
<H1> événement maritime </H1>	<TITLE> Catégories de logements et d'établissements d'hébergement </TITLE> <KEYWORDS> *** </KEYWORDS> <HYPERLINK> *** <TITLE_URL> *** <H1> Résidences de tourisme </H1> <P> un établissement touristique ayant certaines caractéristiques communes avec un hôtel…… </P>

TAB. 1 – *Exemples de contextes d'utilisation*

La cooccurrence par voisinage permet de retrouver les cooccurrents d'un mot dans un seul contexte (phrase, balise, etc.) alors que la cooccurrence par liaison est une cooccurrence pour laquelle les cooccurrents d'un mot dépendent à la fois de la position du mot dans un contexte et de la relation de ce contexte avec les autres contextes existants. Ainsi, le contexte est générique et sera instancié selon l'appartenance du terme à une balise (prendra des valeurs différentes par exemple dans un cas où le contexte est une balise comme <B> et dans un autre cas où le contexte est l'association de <H1> et <Title>, etc.). Dans l'exemple 2 (TAB. 1), si nous considérons le terme « logement », en respectant la liaison logique existante entre <TITLE> et <H1> (figurant dans H.C), nous trouvons les co-occurrents de « logement » dans la réunion des deux balises bien qu'elles soient éloignées. Ce second type de liaison (logique) est sémantique puisqu'un titre de document aura une relation avec les sous titres du même document. Si nous considérons le terme « résidences », nous retrouvons ses co-occurrents dans l'association des deux balises <h1> et <p> qui sont reliées par un lien physique conformément à H.C. Ces deux balises représentent le contexte instancié pour le terme « résidences» en respectant son appartenance à <h1>. Si ce même terme existe dans une autre balise, le contexte sera différent et sera une nouvelle instance.
Dans les cas ou nous ne retrouvons ni un lien logique ni un lien physique entre deux balises, nous considérons la balise seule en tant qu'unité contextuelle et nous appliquons la cooccurrence par voisinage dans la même balise html. Dans l'exemple 1 (TAB. 1), nous avons comme co-occurrent de « événement » le terme « maritime » dans la balise <h1>. Cette balise représente le contexte du terme « événement ».

L'application du contexte générique en relation avec la structure html et les liens sémantiques existants entre les balises permet de représenter l'adaptabilité d'un terme dans le corpus et modélise un contexte dynamique. Ainsi, notre modèle contextuel respecte la position d'un terme afin de pouvoir prendre en considération diverses situations dans lesquelles le terme a été cité. Le calcul de pondération d'un terme par rapport à son co-occurrent prendra en compte les différents contextes (instanciés grâce à H.C) dans lesquels le mot apparait.

La pondération d'un terme est calculée en utilisant l'indice d'équivalence (Michelet, 1988) qui permet d'évaluer la force de lien entre deux termes.

$$E_{ij} = C^2_{ij} / (C_i \times C_j) \qquad (1)$$

C_i : occurrence du terme i / C_{ij}: cooccurrences des deux termes i et j

Module d'extraction des concepts. Afin d'obtenir des concepts ontologiques d'une façon incrémentale, nous initions le processus avec les termes appartenant aux balises clefs et aux titres. Ces termes sont ceux que le concepteur du site a choisi comme mots clefs, glossaire, etc. afin de présenter un résumé de l'information sémantique du domaine. Pour vérifier

l'impact de la définition contextuelle structurelle sur l'extraction des concepts du domaine, nous présentons nos premières expérimentations dans lesquelles la mesure de similarité est la distance euclidienne et le contexte général est limité aux deux premiers niveaux de notre modèle contextuel (balises clefs, titres, sous titres). Cette méthode devra envisager l'expérimentation d'autres mesures de similarités et devra intégrer progressivement tous les éléments html ainsi que les termes afin d'obtenir un ensemble de concepts qui couvrent le domaine. Les groupes de termes sont obtenus en appliquant une méthode de clustering non supervisée. Parmi les méthodes que nous avons testées à savoir Kmeans et les cartes de Kohonen, nous avons choisit de présenter les résultats de la deuxième méthode puisqu'ils sont légèrement meilleurs. Cette méthode (Kohonen, 2001) représente un terme par un vecteur de termes retrouvés en fonction du contexte.

4 Présentation et évaluation des résultats

Afin d'évaluer l'intérêt du modèle contextuel présenté dans la section 3.2, nous avons appliqué deux définitions de contextes sur le même corpus. Le premier contexte est un contexte statique permettant d'encadrer un mot dans une fenêtre d'une taille précise. Nous cherchons les co-occurrents d'un mot dans un espace de 10 mots. Cette définition de contexte considère que tous les mots possèdent la même importance. Le second contexte se base sur notre hiérarchie contextuelle, appliquée uniquement sur les mots appartenant aux balises clefs définies (section 3.2). En appliquant la méthode de clustering dans les deux cas, nous expérimentons différentes alternatives de nombre de classes allant de 20 à 400. Nous présentons le cas établi sur 306 classes (résultats plus significatifs).

Dans un processus de clustering, la qualité d'une classe est généralement basée sur l'homogénéité ou la compacité. Dans (Vazirgiannis et al, 2003), des critères d'évaluation statistique de l'apprentissage non supervisé sont définis. Cependant, les applications liées à l'extraction de connaissance et à la construction d'ontologie ne peuvent pas appliquer ces standards définis pour d'autres applications. En effet, l'homogénéité de la classe n'implique pas que les mots lui appartenant sont sémantiquement proches ou que le label associé satisfait l'expert du domaine. Concernant la découverte de connaissance, l'évaluation reste un challenge. Dans (Holsapple et Joshi, 2005), les auteurs ont proposé une méthode d'évaluation basée sur une ontologie construite manuellement. Dans (Navigli et al, 2004), les auteurs proposent une évaluation qualitative par les experts de domaine qui répondent à un questionnaire dans lequel ils évaluent la qualité des concepts découverts. Dans d'autres travaux, l'évaluation et le processus de labellisation sont basés sur un thesaurus. Mais le thesaurus ne couvre pas forcément tous les aspects spécifiques d'un domaine. Dans notre cas, certains termes de nos classes n'apparaissent pas dans le Thesaurus de OMT (Organisation Mondiale du Tourisme). Ainsi, nous avons proposé une évaluation manuelle par des experts du domaine. Nous présentons les résultats à ces deux experts. D'abord et individuellement, chacun d'entre eux évalue et labélise manuellement les classes de mots ce qui revient à lui associer un concept appartenant au domaine et relatif à son contenu. Ensuite, ils travaillent ensemble pour discuter des résultats de leurs propositions de labels et nous fournissent une évaluation unique sur laquelle ils se sont mis d'accord. Pour évaluer et présenter les résultats de l'expertise, nous avons défini six critères : la distribution des termes, la pondération de paires de termes, la similarité de paires de termes, les concepts extraits, l'interprétation sémantique et le degré de généralité des concepts extraits. Concernant la distribution des termes, avec le contexte stati-

que, nous obtenons 74 classes parmi lesquelles il existe une classe contenant 55% des termes. Alors que pour le contexte basé sur la structure HTML, seuls 13% des termes initiaux sont regroupés dans la même classe.

Pondération et Similarité de paires de termes. La pondération est calculée entre deux termes, le terme à classer et son co-occurrent dans le contexte. Notre contexte structurel permet d'obtenir de meilleurs résultats que le contexte basé sur une fenêtre. Par exemple, le terme « hébergement » et son co-occurrent sémantique « établissement » sont retrouvés avec une pondération plus importante avec notre hiérarchie contextuelle. Nous remarquons également que « loisir » et « oie », qui sont des mots n'ayant aucune relation sémantique, possèdent une faible pondération dans notre contexte en comparaison avec le premier contexte. En ce qui concerne la similarité entre termes, nous remarquons que la similarité calculée dans notre contexte est meilleure que celle obtenue avec le premier contexte. Par exemple, nous retrouvons les termes « archipel » et « île » qui forment le groupe nominal « archipel d'îles » avec une faible similarité dans notre contexte par rapport au premier contexte. Alors que dans ce dernier, nous retrouvons des mots n'ayant pas de relations sémantique comme « bicyclette » et « cuisine » et qui ont une similarité plus faible que celle retrouvé avec le second contexte (FIG.2).

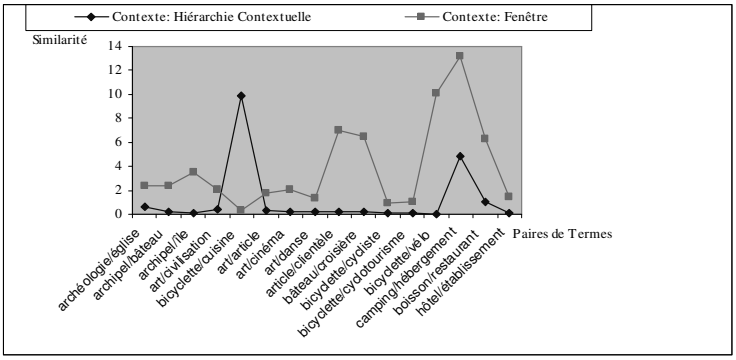

FIG. 2 – *Similarités entre termes avec deux définitions de contextes*

Interprétation sémantique. En évaluant les classes, les experts du domaine remarquent qu'il existe trois types de classes : les classes *acceptables*, les classes *incorrectes* et les classes *inconnues*. Une classe acceptable est une classe que l'expert est capable de labelliser et dans laquelle les termes appartenant au même groupe sont proches sémantiquement. Une classe incorrecte est une classe qui présente l'un des deux cas suivants : soit elle contient certains termes qui n'ont pas de relations avec le concept extrait de cette classe, soit elle contient plusieurs concepts clairement identifiés par l'expert. Une classe inconnue est une classe dont les termes n'ont aucune relation sémantique ; l'expert ne peut pas en donner une interprétation sémantique. Dans notre expérimentation, nous obtenons avec notre définition de contexte plus de classes acceptables (53.2%) et moins de classes inconnues (20.51%) et légèrement plus de classes incorrectes (26.28%) en comparaison avec le premier contexte pour lequel nous obtenons respectivement (40.54%), (33.78%) et (25.67%). A titre d'exemple, comme classe acceptable, nous avons une classe constituée de : {essentiel, relaxation, remise, santé, verdure} (relatif au tourisme de santé) et comme classe inconnue,

« berceau » et « cœur ». Un exemple de classe incorrecte, est la classe {archéologie, bière, boisson, cidre, ethnologie, expérience, peuple}.

Concepts extraits. Avec notre définition contextuelle, l'expert a labellisé 79.48% des classes en comparaison avec 66.21% pour le premier contexte. Nous prenons en compte uniquement les classes acceptables dans les deux cas et nous calculons la précision. Dans notre étude, « la précision est le ratio des termes pertinents ayant entre eux une importante similarité sémantique par rapport à l'ensemble des termes d'une classe donnée». Comme résultats, nous obtenons respectivement 81.36% et 86.18% pour le premier contexte et le second.

Degré de généralité des concepts extraits. Un autre élément qui détermine la qualité d'un concept extrait est son degré de généralité. Dans une ressource ontologique, il est important d'avoir un ensemble important de concepts généraux qui résument le domaine et formant les concepts les plus génériques de l'ontologie. Ainsi, nous focalisons notre intérêt sur les concepts extraits des classes acceptables. Nous établissons une évaluation manuelle en nous basant sur le thésaurus de l'Organisation Mondiale du Tourisme (OMT). Ce thésaurus contient des termes génériques représentant les concepts clés du domaine. Respectivement pour le premier et le second contexte, nous obtenons 60% et 78.31% de concept généraux. Par exemple comme concepts généraux, nous avons « tourisme religieux », « tourisme de santé », etc.

Discussion. Dans cette section, nous avons montré que notre définition de contexte guidant le clustering permet d'obtenir de meilleurs résultats (concepts extraits) sur divers points de vue par rapport au premier contexte (fenêtre). Notre contexte générique se base sur une hiérarchie contextuelle fondée sur l'existence de liens entre les balises HTML. Bien entendu, pour que notre approche reste fonctionnelle le corpus traité doit comporter un minimum de structure dans les documents. Cependant, l'analyse de structure réalisée dans l'étape de prétraitement permet d'adapter la définition du contexte en tenant compte des balises les plus utilisées. Egalement, l'absence de certaines balises n'affecte pas le fonctionnement de notre méthode. Par exemple, si nous ne disposons pas de balises sous titres (<h1>), notre méthode reste réalisable puisqu'elle est incrémentale et nous permet de chercher les co-occurrents d'un mot dans les autres niveaux inférieurs de H.C (les balises <p>, <td>, etc.). En plus, en cas d'absence de balise type <title> qui contient les premiers mots générant nos premiers concepts, nous avons d'autres informations provenant des balises définies. Concernant la notion de corpus, l'étape de prétraitement nous permet d'explorer tout type de corpus qu'il soit constitué par l'ingénieur de connaissance ou imposé (un benchmark) à condition qu'il soit relatif à un seul domaine. Cependant, nous avons obtenus des classes incorrectes et inconnues, certes moins qu'avec le contexte fenêtre mais qui devront être corrigées dans les recherches à venir.

5 Conclusion et perspectives

La construction d'ontologie est une tâche difficile et lourde au regard de la diversité langagière du Web. Nous avons donc proposé une architecture d'un système interactif allant de la construction d'un corpus de pages web relatif à un domaine jusqu'à la formalisation des connaissances stockées dans une ontologie en passant par une phase de prétraitement des données, une phase de découverte de connaissances incrémentale permettant l'intervention

de l'utilisateur . Ce papier décrit le processus d'extraction de concepts ontologiques guidée par le contexte. Ce contexte est modélisé par une hiérarchie contextuelle qui représente un contexte structurel dynamique. Nous avons également présenté les expérimentations faites sur les premiers niveaux de notre hiérarchie. Les résultats obtenus ont montré l'importance de la définition du contexte pour améliorer la sélection des co-occurrents sémantiquement proches, la pondération des termes, et par conséquent la pertinence des concepts ontologiques extraits. Dans les travaux à venir, nous envisageons de corriger les classes incorrectes et inconnues afin d'enrichir l'ensemble des concepts extraits. Puis, nous allons poursuivre nos expérimentations concernant les autres niveaux de la hiérarchie contextuelle et nous allons définir un contexte linguistique et un contexte documentaire puis les combiner avec le contexte structurel afin de tenir compte des cas où les documents HTML sont pauvres en structures, d'améliorer la finesse de décomposition des clusters et de construire une hiérarchie de clusters. Convaincus que l'évaluation et la labellisation sont deux tâches indissociables, nous les intégrons de façon incrémentale dans notre processus d'extraction. Nous envisageons de poursuivre notre réflexion par rapport à la découverte des relations.

Références

Benzecri, J.-P. . *L'analyse des correspondances*. Dunod, 1973.

Brézillon, P.:*Context in problem solving: A survey*, The Knowledge Engineering Review, Volume: 14, Issue: 1,Pages: 1-34, 1999.

Buyukkokten, O., Garcia-Molina, H. and Paepcke, A.: *Accordion summarization for end-game browsing on PDAs and Cellular Phones*. In Proc. Of Conference on Human Factors in Computing Systems (CHI01), 2001.

Cai, D., Yu, S., Wen, J. and Ma, W. : *Block-based web search*. Proceedings of the 27 th annual international ACM SIGIR conference on research and development in information retrieval, pages 456-463, 2004.

Compton, P. and Jansen R : *knowledge in context: a strategy for expert system maintenance*. In : J.Siekmann (Ed): Lecture Notes in Artificial Intelligence, Subseries in computer sciences, Vol.406, 1988.

Davulcu, H., Vadrevu, S. and Nagarajan, S. : OntoMiner: Boostrapping ontologies from overlapping domain specific web sites. In AAAI'98/IAAI'98 Proceedings of the 15th National Conference on Artificial Intelligence, 1998.

Faure, D., C. Nedellec and C. Rouveirol (1998). *Acquisition of semantic knowledge uing machine learning methods : the system ASIUM*. Technical report number ICS-TR-88-16, Laboratoire de recherche en informatique, inference and learning group, Paris-sud.

Frérot, C., D. Bourigault and C. Fabre (2003). *Marier apprentissage endogène et ressources exogènes dans un analyseur syntaxique de corpus*. Le cas du rattachement verbal à distance de la préposition « de », in Revue t.a.l., 44-3.

Hahn, U. and M. Romacker (2001). *The SYNDIKATE Text Knowledge Base Generator*. Proceedings of the 1st International Conference on Human Language Technology Research, San Diego, USA.

Han, H. and Elmasri, R.: "Architecture of WebOntEx: A system for automatic extraction of ontologies from the Web". Submitted to WCM2000.

Holsapple, C. and Joshi, K.D.: *A collaborative approach to ontology design*. Communications of ACM, 45(2): 42-47, 2005.

Joachims, T. . *A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization*. In Proc. 14th International Conference on Machine Learning, Morgan Kaufmann, pages 143-151, 1997.

Kiyota, Y. and Kurohashi, S.: *Automatic summarization of Japanese sentences and its application to a WWW KWIC index*. Proceedings of the 2001 Symposium on applications and the internet, page 120, 2001.

Kohonen, T. *Self organizing Maps*. Eds Springer, 2001.

Michelet, B. . *L'analyse des associations*. Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris, 1988.

Meadche, A. and S. Staab (2001). *Ontology learning for the semantic Web*. IEEE journal on Intelligent Systems, Vol. 16, No. 2, 72-79, 2001.

Navigli, R., Velardi, P., Cucchiarelli, A. and Neri, F.: *Quantitative and qualitative evaluation of the ontolearn ontology learning system*. In Proc. Of ECAI-2004 Workshop on ontology learning and population, Valencia, Spain, Aug.2004.

Schmid, H.. *Probabilistic Part-of-Speech Tagging Using Decision Trees*. IMS-CL, Institut Für maschinelle Sprachverarbeitung, Universität Stuttgart, Germany, 1994.

Sugiura, N, N., Izumi and T. Yamaguchi. *A supprt environment for domain ontology development with general ontologies and text*. IEEE Computational Intelligence Bulletin, February 2004, Vol.3 No.1, 2004.

Vazirgiannis, M., Halkidi, M. and Gunopoulos, D.: *uncertaintly handling and quality assessmen in data mining*. Springer, 2003.

Summary

Ontologies provide a common layer that plays a major role in information exchange and share support. In this paper we present an integrated framework involving complementary dimensions to drive the (semi) automatic acquisition conceptual knowledge process from HTML Web pages. Our approach takes advantage from structural HTML document features and the word location to identify the appropriate term context. Our context definition improves the word weighting, the selection of the semantically closer cooccurrents and the relevant extracted ontological concepts. We use an unsupervised clustering method for term groups' generation. Notice that the chosen clustering method relies on a user incremental quality evaluation process. In this paper, we summarize the most significant results obtained by applying our method on a corpus dedicated to the tourism domain. The first results show how the definition of an appropriate context improves the relevance of the extracted concepts.