Des ontologies pour indexer un site Web

E. DESMONTILS, C. JACQUIN

IRIN, Université de Nantes
2, rue de la Houssinière, BP92208
F-44322 NANTES Cedex 3, France
{desmontils,jacquin}@irin.univ-nantes.fr

Résumé

Cet article présente une nouvelle approche d'indexation d'un site Web s'appuyant sur l'exploitation d'ontologies et sur des techniques de traitement automatique des langues. L'objectif général est de construire un index structuré d'un site Web. La structure est donnée par une ontologie orientée terminologie spécifique au domaine visé par le site. Afin d'exploiter des ontologies dans un contexte linguistique, les représentants des concepts des ontologies sont donc désambiguïsés à l'aide d'un thesaurus. Le processus d'indexation s'appuie d'abord sur des techniques d'extraction de termes qui prennent en compte le marquage HTML pour déterminer l'importance de ces termes dans les pages (un coefficient nommé fréquence pondérée est calculé). Ensuite, l'utilisation d'un thesaurus permet de passer au niveau conceptuel. Chaque concept candidat est évalué en déterminant son niveau de représentativité de la page, c'est-à-dire l'importance du concept relativement aux autres concepts de la page. Ensuite, l'index structuré proprement dit est construit en attachant, à chaque concept de l'ontologie, les pages du site dans lesquelles ces concepts sont trouvés. Finalement, un certain nombre d'indicateurs permettent d'évaluer l'indexation du site par rapport à l'ontologie proposée.

Mots-clés : Indexation par le contenu, Indexation Structurée, TALN, Ontologies, Thesaurus.

1 INTRODUCTION

Rechercher de l'information sur Internet signifie accéder à des sources d'information qui sont hétérogènes, distribuées et qui sont de moins en moins stables : certaines apparaissent, disparaissent ou sont mises à jour. Dans ce contexte une question se pose : comment rechercher de l'information qui soit la plus pertinente possible pour une requête donnée ? De nombreux moteurs de recherche nous aident dans cette tâche difficile. La plupart d'entre eux s'appuient sur une base de donnée, généralement centralisée, et utilisent de simples mots clés pour accéder à l'information. Dans ce type de système, le taux de rappel[1] est satisfai-

[1] le rapport entre le nombre de pages effectivement récupérées concernant un domaine et le nombre total de pages concernant ce domaine.

sant. Par contre, le taux de précision[2] est mauvais. De plus, du fait de la dimension gigantesque du Web, les indexations ne sont pas faites très régulièrement, donc les informations centralisées ne sont pas toujours à jour.

Certains travaux dans la communauté multi-agent (Yuwono & Lee, 1996; Ashish & Knoblock, 1997; Cazalens *et al.*, 2000) montrent qu'un agent localisé sur le site Web peut grandement améliorer le processus de recherche d'information. Dans ce contexte, l'agent a la connaissance du contenu des pages qu'abrite le site et est donc plus apte à répondre efficacement et précisément à des requêtes. Le travail présenté dans cet article s'inscrit dans ce contexte : Comment déterminer le contenu des pages d'un site Web et représenter cette connaissance ? Nous nous intéressons donc à construire l'index structuré d'un site. Comme dans certains travaux de désambiguïsation de requête (Gonzalo *et al.*, 1998) ou d'annotation de pages Web (Luke *et al.*, 1996; Fensel *et al.*, 1998), l'indexation est relative à une connaissance représentée par une ontologie. Contrairement aux outils classiques d'indexation, nous ne nous basons pas sur les mots-clés mais sur les concepts qu'ils représentent.

Dans la suite de cet article, nous présenterons d'abord le processus général d'indexation structurée (section 2), puis, après avoir exposé les caractéristiques des ontologies utilisées (section 3), nous indiquerons comment est évaluée la représentativité d'un concept dans une page (section 4) puis comment est évalué le processus d'indexation (section 5).

2 PROCESSUS GÉNÉRAL

L'objectif du processus est donc de construire un index structuré des pages d'un site Web en fonction d'un domaine de connaissance. La structure est donnée par une ontologie de ce domaine. Le processus d'indexation comporte les phases suivantes (figure 1) :

1. D'abord, pour chacune des pages est constitué un index à plat des termes, avec leur fréquence respective (nombre d'occurrences) pondérée par les marqueurs HTML qui leurs sont relatifs.
2. Ensuite, un thesaurus permet de déterminer tous les concepts candidats associés aux termes précédemment acquis. Dans notre expérimentation, nous utilisons le thesaurus WordNet (Miller, 1990).
3. Pour chaque concept candidat, le calcul d'un coefficient de représentativité permet d'évaluer sa représentativité dans la page étudiée. Ce calcul s'appuie sur la fréquence pondérée et sur une mesure de similarité entre concepts. Cette mesure permet aussi de déterminer en contexte le sens le plus probable d'un terme. Ainsi, un concept sera d'autant plus important dans une page qu'il sera fortement en relation avec d'autres concepts de cette page. Cette évaluation permet de relativiser la fréquence pondérée. Elle accentue

[2] le rapport entre le nombre de pages effectivement récupérées concernant un domaine et le nombre total de pages récupérées.

l'importance des concepts fortement en relation avec les autres et diminue celle des concepts plus ou moins isolés (même s'ils ont une fréquence importante).

4. Parmi tous les concepts retenus à la phase précédente, un filtre est réalisé via l'ontologie et la représentativité des concepts. Il permet de sélectionner les concepts présents dans l'ontologie et dont la représentativité dépasse un certain seuil. Ceci permet ensuite de construire l'index structuré en associant la page concernée aux concepts de l'ontologie qu'elle contient.
5. Finalement, un certain nombre de mesures sont évaluées afin de caractériser l'indexation, c'est-à-dire de déterminer l'adéquation entre le site et l'ontologie. Ces mesures prennent en compte notamment le nombre de pages sélectionnées par l'ontologie, le nombre de concepts de l'ontologie présents dans les pages et la représentativité moyenne de ces concepts.

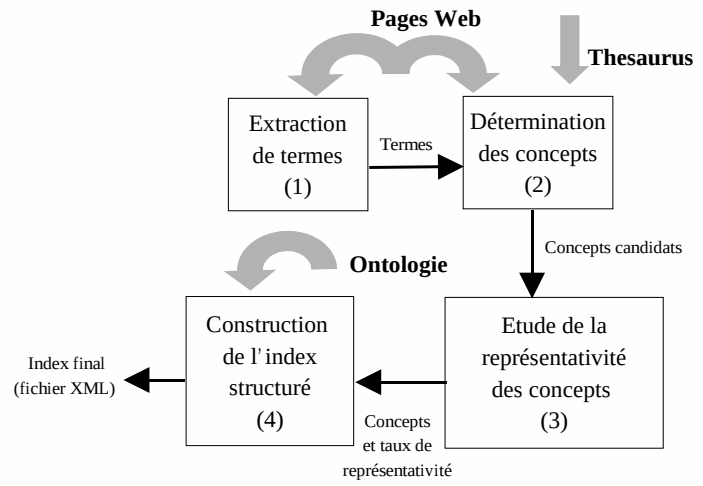

FIG. 1 – Le processus d'indexation

Dans notre approche, l'index est construit indépendamment des pages du site et est stocké sous un format XML (W3C, 1998). En cela, elle diffère notablement des approches basées sur l'annotation de pages Web telle qu'on peut la trouver dans SHOE (Luke *et al.*, 1996), WebKB (Martin & Eklund, 1999) ou KA2 (Fensel *et al.*, 1998). Dans tous les cas, l'objectif est de faire intervenir des informations sémantiques pour faciliter la recherche de l'information. Cependant, dans ces approches, des annotations sont disposées au sein du code HTML à l'aide de meta-marqueurs. Le concepteur des pages indique les connaissances manipulées au fur et à mesure de leur apparition. Le problème est que toute modification ou nouvelle génération de pages demande de recommencer tout ou partie du travail. Par contre, la précision du traitement est extrêmement fine.

De plus, les méthodes basées sur l'annotation sont totalement manuelles. Aussi, elles sont très coûteuses en temps et ne peuvent être mises en oeuvre que par des spécialistes (Heflin *et al.*, 1999). Au contraire, notre processus est semi-automatique et permet plus facilement d'avoir un point de vue global sur le site. Il permet aussi d'indexer des sites dont le code source des pages ne peut être modifié. Nous ne le considérons pas comme totalement automatique, car des ajustements peuvent être effectués par l'utilisateur en fin de processus. La contrepartie à cette automatisation est, bien évidemment, une moins bonne précision du traitement. Enfin, par rapport à l'approche par annotation, l'approche par indexation facilite la recherche d'information, car elle permet d'accéder directement aux pages concernant un concept. Par contre, l'approche par annotation demande un parcours de toutes les pages du site pour retrouver cette même information.

Le projet qui se rapproche le plus de notre étude est le projet CHIMERE (Segret *et al.*, 2000). Ce projet s'attache à extraire des informations de formulaires à partir d'une ontologie du domaine et de traitements linguistiques. Ces formulaires se composent de zones de saisie et de données textuelles. Mais les données textuelles sont très réduites. Les techniques linguistiques employées dans ce contexte ne peuvent s'adapter à notre cas d'étude qui traite des pages Web comportant principalement de grandes zones de données textuelles, souvent peu structurées.

Avant de présenter quelques résultats, nous allons étudier deux éléments importants : l'ontologie de référence et surtout la méthode d'évaluation des concepts.

3 LES ONTOLOGIES

3.1 Définition d'une ontologie

Les ontologies sont devenues ces dernières années un domaine à part entière de la recherche en ingénierie des connaissances (conféré le nombre de conférences qui y font référence). Ce terme est issu du domaine de la philosophie et a comme signification originelle : «*explications systématiques de l'existence*». Les chercheurs en ingénierie des connaissances ont, sans fondamentalement en changer le sens, donner d'autres définitions de ce concept, mieux adaptées à leur souci de recherche. En parcourant la littérature, on peut s'apercevoir qu'il y a foison de définitions complémentaires qui sont fortement liées au point de vue de l'auteur et à l'utilisation qu'il fait des ontologies (Gruber, 1993; Guarino, 1997). Certains ont un point de vue plus formel et s'attachent à travailler sur un modèle abstrait des ontologies, d'autres au contraire, utilisent des ontologies dans des applications bien précises et en ont une vue plus pragmatique. En ce qui nous concerne, nous prendrons comme définition, celle issue de (Gomez-Perez, 1999) : «*les ontologies fournissent le vocabulaire commun d'un domaine et définissent de façon plus ou moins formelle le sens des termes et les relations entre ces derniers*». Ceci nous autorisera à pouvoir appeler ontologies, des hiérarchies de concepts plus ou moins formalisées. Par exemple, la figure 2 propose un extrait de l'ontologie sur les universités américaines du projet SHOE (Luke *et al.*, 1996).

```
<?xml version="1.0" encoding="ISO-8859-1"
      standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="2.1"
          description="...">
  <def-category name="Department"
                isa="EducationOrganization"
                short="university department"/>
  <def-category name="Program"
                isa="EducationOrganization"
                short="program"/>
  <def-category name="ResearchGroup"
                isa="EducationOrganization"
                short="research group"/>
  <def-category name="University"
                isa="EducationOrganization"
                short="university"/>
  <def-category name="Activity"
                isa="SHOEEntity"
                short="activity"/>
  <def-category name="Work"
                isa="Activity"
                short="work"/>
  <def-category name="Course"
                isa="Work"
                short="teaching course"/>
  ...
</ontology>
```

FIG. 2 – Extrait de l'ontologie SHOE sur les universités

3.2 Construction d'une ontologie orientée terminologie

Habituellement, les concepts des ontologies sont représentés par un terme linguistique (une étiquette). Cependant, dans notre contexte, ce terme est à la fois ambigu (représente potentiellement plusieurs concepts) et pas forcément unique (existence de synonymes). Dans le cadre de traitements utilisant des textes écrits en langage naturel, il convient donc de déterminer l'ensemble des synonymes (étiquettes candidates) permettant de définir de manière unique un concept. Ce type de traitement se retrouve de façon manuelle dans OntoSeek (Borgo *et al.*, 1997; Guarino *et al.*, 1999) ou de manière semi-automatique dans Mikrokosmos (O'Hara *et al.*, 1998).

Dans notre contexte, une ontologie est donc une hiérarchie de concepts chacun représenté par un terme (une étiquette) et un ensemble de synonymes de ce terme. Certains concepts sont reliés par la relation spécifique/générique, la relation partie/tout... Nous appelons ce type d'ontologie une ontologie orientée terminologie.

Notons que nos ontologies ne reflètent pas tous les aspects inhérents aux ontologies formelles (Gomez-Perez, 1999).

Nous proposons donc un processus permettant de déterminer toutes les étiquettes possibles d'un concept. Ce processus, se basant sur un thesaurus, utilise un certain nombre d'heuristiques similaires à celles proposées par le projet Mikrokosmos. Le principe général de ces heuristiques est de faire une correspondance entre les chemins suivant la relation "est un" ("isa") dans l'ontologie et les chemins d'hyperonymes dans le thesaurus. Selon le degré de correspondance, une confiance plus ou moins grande est donnée à tel ou tel ensemble de synonymes. Notons que des expérimentations réalisées prenant en compte la relation de composition n'a pas permis d'améliorer notablement les résultats.

L'utilisateur termine éventuellement la désambiguïsation des étiquettes à la main. En effet, le processus ne permet pas toujours de sélectionner de façon certaine le bon ensemble de synonymes. Les définitions des ensembles de synonymes potentiels sont présentées afin d'aider au choix final. La figure 3 propose un extrait de l'ontologie orientée terminologie sur les universités américaines construite à partir de celle du projet SHOE (figure 2). Cependant, le processus donne des résultats assez satisfaisants puisqu'il choisit le bon sens pour près de 75% des étiquettes associées aux concepts pour l'ontologie des universités de SHOE (Luke *et al.*, 1996) et de 95% lorsque certaines modifications ont été apportées (certaines contradictions par rapport au thesaurus utilisé ont été levées). Bien sûr, ce processus de désambiguïsation dépend du thesaurus utilisé (dans notre cas Wordnet).

4 REPRÉSENTATIVITÉ D'UN CONCEPT D'UNE PAGE

L'autre élément important de notre processus d'indexation est l'évaluation de l'importance d'un concept dans une page HTML. Il y a deux étapes essentielles : (1) l'extraction des termes de la page avec leur poids respectif et (2) la détermination des concepts candidats et le calcul de la représentativité des concepts dans la page.

4.1 Extraction des termes

Les termes présents dans la page sont extraits en utilisant des techniques classiques de traitement automatique des langues qui ont été enrichies. Tout d'abord, les marqueurs HTML sont supprimés. Le texte est ensuite découpé en phrases indépendantes et étiqueté en utilisant l'étiqueteur de Brill (Brill, 1995). Chaque mot des pages est alors annoté par sa catégorie grammaticale (nom, adjectif, verbe...). Après lemmatisation, les termes linguistiquement bien formés sont extraits en fonction de patrons morpho-syntaxiques (Daille, 1994), c'est-à-dire les termes de la forme "nom", "nom nom", "nom of nom", "adjectif nom"...

Pour chaque terme ainsi sélectionné, nous calculons ensuite sa fréquence pondérée. Cette fréquence est déterminée en prenant en compte la fréquence du terme et surtout les marqueurs HTML qui interviennent sur chacune de ses occurrences.

```
<?xml version="1.0" encoding="ISO-8859-1"
      standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="3.0">
  <def-category name="Course" short="teaching course"
                isa="Work">
    <sense name="Course" no="1" origin="WN"
           definition="..." convenience="1.0">
      <synset>class#4,course of instruction#1,
              course of study#2,course#1</synset>
    </sense>
  </def-category>
  <def-category name="Department"
                short="university department"
                isa="EducationOrganization">...
  </def-category>
  <def-category name="University" short="university"
                isa="EducationOrganization">
    <sense name="University" no="3" origin="WN"
           definition="..." convenience="1.0">
      <synset>university#3</synset></sense>
  </def-category>
  <def-category name="Program" short="program"
                isa="Information">
    <sense name="Program" no="4" origin="WN"
           definition="..." convenience="1.0">
      <synset>course of study#1,curriculum#1,program#4,
              syllabus#1</synset></sense>
  </def-category>
  <def-category name="ResearchGroup"
                short="research group"
                isa="EducationOrganization">
    <sense name="ResearchGroup" no="0" origin="TECH"
           definition="" convenience="1.0">
      <synset>research group#0</synset></sense>
  </def-category>
  <def-category name="Activity" short="activity"
                isa="HumanActivity">...
  </def-category>
  <def-category name="Work" short="work"
                isa="Activity">...
  </def-category>...
</ontology>
```

FIG. 3 – Extrait de l'ontologie orientée terminologie sur les universités

Il est à noter que la fréquence seule n'est pas un critère prépondérant. En effet, nous travaillons sur des pages qui sont de dimension assez restreinte. L'influence du marqueur dépend de son rôle dans la page. Par exemple, le marqueur "TITLE" donnera une importance considérable au terme (*10) alors que le marqueur "B" (pour bold) possède une influence bien moindre (*2). Le tableau 1 donne le poids des marqueurs les plus importants. Ces poids ont été déterminés de manière expérimentale (Gamet, 1998).

Description du marqueur HTML	**Marqueur HTML**	**Poids**
Titre du document	<TITLE></TITLE>	10
Mot clé	<meta name="keywords" content=...>	9
Hyper-lien	<A HREF=...></A>	8
Font 7	<FONT SIZE=7></FONT>	5
Font +4	<FONT SIZE="+4"></FONT>	5
Font 6	<FONT SIZE=6></FONT>	4
Font +3	<FONT SIZE="+3"></FONT>	4
Font +2	<FONT SIZE="+2"></FONT>	3
Font 5	<FONT SIZE=5></FONT>	3
Titre niveau 1	<H1></H1>	3
Titre niveau 2	<H2></H2>	3
Titre d'image	<IMG ... ALT="...">	2
Marqueur Big	<BIG></BIG>	2
Souligné	<U></U>	2
Italique	<I></I>	2
Gras	<B></B>	2
...	...	...

TAB. 1 – Les principaux marqueurs HTML et leur poids

La fréquence pondérée de chacun des termes est calculée relativement à celle ayant la plus forte valeur (normalisée). Ce seront donc des valeurs prises dans l'intervalle $[0, 1]$. Soit $F(T_i)$ la fréquence pondérée du terme T_i (i dans $1..n$), soient $\{t_{i,j}\}$ les p occurrences du terme T_i dans la page et $\{M_{i,j}\}$ les poids des p marqueurs HTML portant sur les termes $t_{i,j}$ (notons la présence d'un marqueur vide par défaut dont le poids est 1), les équations 1 et 2 présentent le calcul de $F(T_i)$.

$$F(T_i) = \frac{P(T_i)}{max_{k=1..n}(P(T_k))} \quad (1)$$

$$P(T_i) = \sum_{j=1}^{p}(M_{i,j}) \quad (2)$$

Le tableau 2 donne un extrait de l'index à plat ainsi obtenu pour la page HTML "http ://www.cs/washington.edu/news". Les termes sont ordonnés selon leur fréquence pondérée décroissante.

Termes	Fréquence pondérée
uw	1.00
cse	0.59
uw cse	0.45
computer	0.41
university	0.37
seattle	0.30
article	0.30
science	0.26
research	0.24
professor	0.24
computer science	0.18
...	...
university of washington	0.16
program	0.15
...	...
news	0.12
...	...
information	0.09
...	...
message	0.01
...	...

TAB. 2 – Termes extraits et leur fréquence pondérée (résultat trié par la fréquence pondérée) provenant de http ://www.cs.washington.edu/news/

4.2 Détermination et évaluation des concepts

A ce stade du processus, nous disposons d'un index à plat de la page traitée et à chaque terme de cet index est associée sa fréquence pondérée. L'objectif suivant est de déterminer les concepts candidats et leur importance relative dans la page. Les concepts candidats sont construits en utilisant un thesaurus. Pour chaque terme, le thesaurus fournit les concepts potentiels sous la forme d'une liste de synonymes (supposée unique pour un concept donné). Ensuite, pour chaque concept candidat, son degré de représentativité dans la page est calculé en fonction de la fréquence pondérée et de la mesure de similarité par rapport aux autres concepts de la page (similarité cumulée). Nous commencerons par définir la mesure de similarité entre deux concepts qui permet d'évaluer la proximité sémantique de deux concepts d'une page relativement au thesaurus en utilisant la relation d'hyperonymie.

Dans notre contexte, nous utilisons la mesure de similarité définie par (Wu & Palmer, 1994). Cette mesure prend en compte le concept le plus proche qui subsume deux concepts (ce qui les rapproche) tout en normalisant ensuite le calcul par ce qui les différencie. La mesure est donnée par l'équation 3 où C est le concept le plus proche qui subsume C_1 et C_2 (en nombre d'arc), $depth(C)$ est le nombre

d'arc qui sépare C de la racine de la hiérarchie, et $depth_c(c_i)$ avec i élément de $\{1, 2\}$ est le nombre d'arc qui sépare c_i de la racine de la hiérarchie en passant par C.

$$sim(c_1, c_2) = \frac{2 * depth(c)}{depth_c(c_1) + depth_c(c_2)} \quad (3)$$

Cette mesure est moins performante que la mesure de Resnik (Resnik, 1999) mais plus que celle, plus traditionnelle, nommée "edge counting". Cette dernière calcule le nombre minimal d'arc qui sépare deux concepts en passant par le concept le plus proche qui les subsume. Dans notre contexte la mesure de Resnik (Resnik, 1999), n'est pas facilement applicable, car celle-ci demande des évaluations de fréquences sémantiques de concepts sur corpus que nous n'avons pas à notre disposition (ceci demande un étiquetage sémantique manuelle de gros corpus).

Pour évaluer l'importance relative d'un concept dans la page, il est nécessaire de connaître la similarité cumulée de celui-ci avec les autres concepts présents dans la page. Cette mesure, notée $\widehat{sim}$, associée à un concept dans une page, est la somme de toutes les mesures de similarité calculées entre lui et les différents concepts candidats inclus dans la page étudiée. Dans cette formule, pour des facilités d'écriture, nous avons unifié un concept avec le synset (ensemble de synonymes) qui le représente dans WordNet. La mesure utilisée est donnée dans la formule 4 où un terme T_k a l_k synsets associés et il y a m termes différents dans la page étudiée.

$$\widehat{sim}(synset_i(T_k)) = \sum_{j \in [1,k-1] \cup [k+1,m]} \sum_{l=1}^{l_j} sim(synset_i(T_k), synset_l(T_j)) \quad (4)$$

Dans ce calcul, pour un concept c, tous les calculs de similarités avec les autres concepts candidats ne sont pas pris en compte afin de discriminer les données. Pour cela, un seuil est appliqué.

Finalement, la représentativité du concept dans la page est calculée par une combinaison linéaire de la fréquence pondérée du terme associé à ce concept et de sa similarité cumulée (préalablement normalisée) avec les autres concepts candidats de la page. Ce coefficient est primordial pour la qualification des réponses aux requêtes futures. L'équation 5 illustre ce calcul. Les valeurs déterminées expérimentalement pour α et β sont respectivement $2/3$ et $1/3$.

$$representativite(synset_i(T_k)) = \frac{\alpha * F(T_k) + \beta * \widehat{sim}(synset_i(T_k))}{\alpha + \beta} \quad (5)$$

Le tableau 3 illustre l'effet de ce coefficient de représentativité sur l'ordre des concepts.

Concept	**Fréquence pondérée**	**Représentativité**
uw#0	1.0	0.67
award#2, accolade#1, honor#1, honour#2, laurels#1	0.20	0.47
computer#1, data processor#1, electronic computer#1, information processing system#1	0.41	0.45
university#2	0.37	0.39
information#1, info#1	0.09	0.39
cse#0	0.59	0.39
course of study#1, program#4, curriculum#1, syllabus#1	0.15	0.35
calculator#1, reckoner#1, figurer#1, estimator#1, computer#2	0.41	0.34
article#3, clause#2	0.30	0.34
news#1, intelligence#4, tidings#1, word#3	0.12	0.34
voice#6	0.01	0.34
voice#2, vocalization#1	0.01	0.34
message#2, content#2, subject matter#1, substance#6	0.01	0.34
language#1, linguistic communication#1	0.01	0.34
submission#1, entry#4	0.01	0.33
subject#1, topic#1, theme#1	0.01	0.33
...	...	...

TAB. 3 – Les concepts extraits après calcul du coefficient de représentativité (résultats triés sur la représentativité) provenant de http ://www.cs.washington.edu/news/

Si on regarde le coefficient de représentativité, on s'aperçoit que des concepts sont remontés dans la liste (par rapport à la table 2) : par exemple news#1 (fréquence pondérée 0.12, représentativité 0.34) ou information#1 (fréquence pondérée 0.09, représentativité 0.39). Ceci est un bon résultat pour une page qui est la page des "news" du département informatique de l'université de washington. Si l'on regarde un peu plus en aval dans le fichier résultat, on s'aperçoit que les concepts news#4 et news#2 ont un coefficient de représentativité de 0.31, donc pas très différent de celui de news#1 qui est de 0.34. Si on regarde plus en détail dans le thesaurus on s'aperçoit que ces trois sens dans le thesaurus, ont exactement les mêmes concepts qui les subsument. On voit ici un problème inhérent à Wordnet qui est sa trop forte granularité. Il a été développé par des linguistes et n'est pas toujours très discriminant dans les applications de traitement informatique des langues (O'Hara *et al.*, 1998).

5 APPARIEMENTS DES CONCEPTS DE L'ONTOLOGIE ET DES CONCEPTS CANDIDATS

A ce stade du processus nous disposons, d'une part, d'une ontologie dont les étiquettes associées aux concepts ont été désambiguïsées, et, d'autre part, pour chaque page, d'une liste des concepts extraits et leur coefficient de représentativité. Le stade suivant consiste à apparier les concepts de l'ontologie à ceux extraits des pages. Chaque concept de l'ontologie est recherché dans l'ensemble des concepts extraits d'une page. Si ce concept est trouvé et si le coefficient de représentativité dépasse un certain seuil alors l'URL de la page et son coefficient relatif (représentativité) sont ajoutés dans l'ontologie.

Afin d'évaluer l'adéquation entre l'ontologie et le site Web cinq coefficients sont calculés. Lorsque plusieurs ontologies sont susceptibles d'être associées au site, la meilleure peut être choisie eu égard à cette évaluation. Les quatre premiers coefficients (normalisés entre 0 et 1) définissent :

- la proportion de concepts présents directement dans les pages (Degré d'Indexation Direct ou DID) ;
- la proportion de concepts présents indirectement dans les pages (Degré d'Indexation Indirecte ou DII) qui est calculée en tenant compte de la relation générique/spécifique et du DID ;
- la proportion de pages concernées par l'ontologie (Degré de Couverture de l'Ontologie ou DCO),
- la représentativité moyenne des concepts sélectionnés (RMC).

Actuellement, les coefficients (DID, DII, DCO et RMC) sont évalués pour différents seuils concernant le coefficient de représentativité (de 0 et 1 par pas de 0,05). Ensuite, pour chacun de ces coefficients est calculée sa moyenne pondérée. Par exemple, l'équation 6 présente le calcul de la moyenne pondérée du degré d'indexation direct ($\overline{DID}$).

$$\overline{DID} = \sum_{s=0}^{1} (s * DID_s) \tag{6}$$

Cette pondération permet de donner plus de poids aux concepts les plus représentatifs des pages. Une ontologie représentative d'un site possède des coefficients proches de 1. Notons toutefois que cette évaluation dépend aussi du thesaurus utilisé puisqu'elle dépend des relations entre concepts. Finalement, l'évaluation globale de l'indexation (le DAOS : Degré d'Adéquation Ontologie Site) est une combinaison linéaire de ces moyennes pondérées. Pour l'instant, les coefficients sont évalués de manière expérimentale. L'évaluation actuelle est donnée par l'équation 7 où s est un site et o une ontologie. Après avoir analyser manuellement un échantillon représentatif des résultats d'indexation, pour les différents seuils testés, l'expérience montre qu'une valeur de 0.3 pour le coefficient de représentativité donne de bons résultats. En dessous de ce seuil, trop de concepts peu représentatifs du contenu sont conservés. Pour ce seuil, la discrimination des

concepts est relativement efficace (sachant qu'elle est d'autant plus efficace que les pages sont de plus grande dimension).

$$DAOS_{s,o} = \frac{\overline{DII}_{s,o}}{2} + \overline{DID}_{s,o} + 2 * \overline{DCO}_{s,o} + 2 * \overline{RMC}_{s,o} \qquad (7)$$

La figure 4 présente les résultats de l'indexation d'un site[3] de 1315 pages HTML. Le symbole "***" indique le seuil effectif de l'indexation. La figure 5 présente un extrait de l'index structuré pour ce seuil. Ce site concerne le département "Computer Science" de l'université de Washington. Il a été choisi parce qu'à priori intéressant par rapport à l'ontologie. Cependant, le degré d'adéquation de ce site par rapport à l'ontologie (DAOS) n'est pas très élevé (51). Ceci peut s'expliquer par le fait que l'ontologie utilisée (c'est une adaptation de celle du projet SHOE qui comporte 79 concepts), n'est pas exhaustive du point de vue du domaine qu'elle devrait recouvrir. Par exemple le site étudié comporte de nombreuses pages personnelles qui sont rarement indexées via l'ontologie.

6 CONCLUSION

Dans cet article, nous avons présenté un processus semi-automatique qui permet d'indexer un site Web par le contenu. Ce processus permet de construire un index structuré à partir d'une ontologie orientée terminologie et d'un site Web. Après la construction de l'index à plat où à chaque terme est associée sa fréquence pondérée, nous cherchons à déterminer les concepts que ces termes sont susceptibles de représenter. Pour chaque concept obtenu est calculée sa représentativité dans la page dans laquelle il a été trouvé. Finalement, les concepts les plus représentatifs des pages sont sélectionnés et, parmi ceux-ci, ceux appartenant à l'ontologie sont conservés. L'index structuré final est organisé selon l'ontologie. A chaque concept est associé un ensemble de pages du site où il se trouve.

Ce processus comporte un certain nombre d'avantages sur les méthodes classiques d'indexation (seulement basées sur la recherche des mots dans les pages) et même sur les méthodes d'annotation :

1. les pages retournées (répondant à la requête) comportent les concepts demandés et pas seulement les mêmes termes ;
2. ces concepts sont représentatifs des thèmes traités dans la page ;
3. les termes présents dans la page ne sont pas forcément ceux de la requête mais peuvent être des synonymes ;
4. les pages peuvent comporter non seulement les concepts demandés mais aussi des concepts plus spécifiques ;
5. l'importance des concepts n'est pas seulement dépendant de leur fréquence mais des marqueurs HTML qui leurs sont associés et surtout des autres concepts de la page avec lesquels il existe une relation sémantique particulière

[3] "http ://www.cs.washington.edu/"

Seuil	DCO(%)	RMC(%)	DID(%)	DII(%)	
0.0	91.25	69.2	74.68	81.01	
0.05	90.42	68.65	74.68	81.01	
0.1	87.91	67.93	72.15	78.48	
0.15	83.19	66.66	69.62	75.95	
0.2	78.63	65.58	68.35	75.95	
0.25	71.18	65.0	65.82	73.42	
0.3	63.73	64.74	60.76	70.89	***
0.35	53.46	67.27	54.43	67.09	
0.4	50.04	67.34	51.9	64.56	
0.45	43.57	70.12	48.1	63.29	
0.5	42.43	69.95	48.1	63.29	
0.55	32.85	72.74	37.97	56.96	
0.6	30.19	72.67	37.97	56.96	
0.65	26.92	72.91	37.97	56.96	
0.7	22.97	73.71	36.71	56.96	
0.75	21.29	73.77	36.71	56.96	
0.8	18.78	74.48	35.44	55.7	
0.85	16.73	74.02	32.91	54.43	
0.9	15.67	73.45	31.65	53.16	
0.95	14.68	71.9	31.65	53.16	
1.0	13.99	71.05	31.65	53.16	
Moy.	46.19	70.15	49.49	64.26	
MP.	30.0	71.61	40.32	58.57	

DAOS:51.48

FIG. 4 – Exemple de résultats de l'analyse d'une indexation

Le processus d'indexation que nous présentons peut être utile pour la recherche d'information mais aussi pour évaluer l'adéquation d'un site par rapport à un domaine afin de le faire référencer dans ce domaine par des moteurs de recherche hiérarchiques classiques (Yahou !, Excite...).

Pour l'heure, nous sommes en train d'indexer d'autres sites toujours relatifs au monde des universités américaines (du fait de l'ontologie dont nous disposons) afin de comparer les résultats à ceux obtenus sur le site de l'université de Washington. Mais, afin d'améliorer les résultats d'indexation, il faut améliorer la couverture de l'ontologie par rapport à notre domaine d'étude. De plus l'utilisation d'un thesaurus moins généraliste, donc plus ciblé sur un domaine, permettrait aussi d'améliorer les résultats. En ce qui concerne la détermination des concepts en contexte, nous travaillons sur l'apport de la prise en compte d'autres relations que la relation générique/spécifique. Nous avons aussi développé une mesure relative à la relation partie/tout, mais elle demande à être évaluée par expérimentation.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no"?>
<!DOCTYPE ontology SYSTEM "http://.../onto.dtd">
<ontology id="university-ont" version="3.0" description="">
  <def-category name="University" short="university"
                isa="EducationOrganization">
    <sense name="University" no="3" origin="wn" convenience="1.0">
      <synset>university#3</synset>
      <page name="http://www.cs.washington.edu/news/recent/latest10.html"
            frequence="0.4" representativity="0.85"/>
      <page name="http://www.cs.washington.edu/news/"
            frequence="0.37" representativity="0.85"/>
      <page name="http://www.cs.washington.edu/commercialization/policyprop.html"
            frequence="0.3" representativity="0.75"/>
      <page name="http://www.cs.washington.edu/homes/ghulten/"
            frequence="0.38" representativity="0.85"/>
      <page name="http://www.cs.washington.edu/news/chan.html"
            frequence="0.86" representativity="0.85"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/"
            frequence="1.0" representativity="0.0"/>
      <page name="http://www.cs.washington.edu/homes/tiwary/"
            frequence="0.5" representativity="0.55"/>...
    </sense>
  </def-category>
  <def-category name="Department" short="university department"
                isa="EducationOrganization">
    <sense name="Department" no="1" origin="wn" convenience="1.0">
      <synset>department#1,section#11</synset>
      <page name="http://www.cs.washington.edu/education/courses/590m/"
            frequence="0.4" representativity="0.65"/>
      <page name="http://www.cs.washington.edu/leadership/"
            frequence="0.5" representativity="0.35"/>
      <page name="http://www.cs.washington.edu/homes/lazowska/chair/summer.support.html"
            frequence="0.67" representativity="0.75"/>
      <page name="http://www.cs.washington.edu/education/courses/590b/"
            frequence="0.4" representativity="0.65"/>
      <page name="http://www.cs.washington.edu/info/public/"
            frequence="1.0" representativity="0.75"/>
      <page name="http://www.cs.washington.edu/education/courses/590zpl/"
            frequence="0.4" representativity="0.65"/>
      <page name="http://www.cs.washington.edu/education/courses/510/"
            frequence="0.4" representativity="0.65"/>
      <page name="http://www.cs.washington.edu/education/courses/490ap/"
            frequence="0.33" representativity="0.65"/>
      <page name="http://www.cs.washington.edu/homes/carlson/"
            frequence="0.33" representativity="0.35"/>
      <page name="http://www.cs.washington.edu/"
            frequence="0.33" representativity="0.65"/>...
    </sense>
  </def-category>...
</ontology>
```

FIG. 5 – Extrait de l'ontologie SHOE sur les universités

RÉFÉRENCES

ASHISH N. & KNOBLOCK C. A. (1997). Semi-automatic generation internet information sources. In *2nd IFCIS Conference on Cooperative Information Systems (CoopIS)*.

BORGO S., GUARINO N., MASOLO C. & VETERE G. (1997). Using a large linguistic ontology for internet-based retrieval of object-oriented components. In *SEKE*.

BRILL E. (1995). Transformation-based error-driven learning and natural language processing : a case study in part-of-speech tagging. *Computational linguistics*, **21**.

CAZALENS S., DESMONTILS E., JACQUIN C. & LAMARRE P. (2000). A web site indexing process for an internet information retrieval agent system. In I. C. S. PRESS, Ed., *International Conference on Web information Systems Engineering (WISE'2000)*, p. 245–249, Hong-Kong.

DAILLE B. (1994). *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Paris 7.

FENSEL D., DECKER S., ERDMANN M. & STUDER R. (1998). Ontobroker : Or how to enable intelligent access to the www. In *the 11th Banff Knowledge Acquisition for Knowledge-Based System Workshop (KAW'98)*, Banff, Canada.

GAMET J. (1998). *indexation de pages web*. rapport de dea informatique, université de Nantes.

GOMEZ-PEREZ A. (1999). Développements récents en matière de conception, de maintenance et d'utilisation des ontologies. In REVUE TERMINOLOGIES NOUVELLES, Ed., *colloque Terminologie et intelligence artificielle de Nantes*, p. 9–20.

GONZALO J., VERDEJO F., CHUGUR I. & CIGARRAN J. (1998). Indexing with wordnet synsets can improve text retrieval. In *Coling-ACL'98*.

GRUBER T. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition*, **5**, 199–220.

GUARINO N. (1997). Some organizing principles for a unified top-level ontology. In *Spring Symposium series on ontological engineering*, p. 57–63, Stanford.

GUARINO N., MASOLO C. & VETERE G. (1999). Ontoseek : Content-based access to the web. *IEEE Intelligent Systems and Their Applications*, **14**(3), 70–80.

HEFLIN J., HENDLER J. & LUKE S. (1999). Applying ontology to the web : A case study. In *International Work-Conference on Artificial and Natural Neural Networks (IWANN)*.

LUKE S., SPECTOR L. & RAGER D. (1996). Ontology-based knowledge discovery on the world-wide-web. In *the workshop on internet-based information system, AAAI'96*, Portland, Oregon.

MARTIN P. & EKLUND P. (1999). Embedding knowledge in web documents. In *the 8th International World Wide Web Conference (WWW8)*, Toronto, Canada.

MILLER G. A. (1990). Wordnet : an online lexical database. *International Journal of Lexicography*, **3**(4), 235–312.

O'HARA T., MAHESH K. & NIREMBURG S. (1998). Lexical acquisition with wordnet and mihrokosmos ontology. In *Coling-ACL'98*.

RESNIK P. (1999). Semantic similarity in a taxonomy : an information based measure and its application to problems of ambiguity in natural language. *Journal of artificial intelligence research*, **11**.

SEGRET M.-S., POMPIDOR P. & HÉRIN D. (2000). Extraction et intégration d'informations semi-structurées dans les pages web - projet chimère. In *IC'2000*, p. 277–288.

W3C (1998). Extensible markup language (xml) 1.0. w3c recommendation, reference : Rec-xml-19980210.

WU Z. & PALMER M. (1994). Verb semantics and lexical selection. In *the 32nd annual meeting of the association for computational linguistics*, Las Cruses, New Mexico.

YUWONO B. & LEE D. L. (1996). Wise : A world wide web resource database system. *IEEE Transactions on Knowledge and Data Engineering*, **8**(4).