Indexation sémantique de documents sur le Web: application aux ressources humaines

E. DESMONTILS, C. JACQUIN, E MORIN

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

Résumé

Dans cet article nous présentons nos travaux en cours relatifs à la gestion des ressources humaines sur le Web. Nous nous intéressons plus particulièrement à l'extraction et à la structuration du contenu des Curriculum Vitae dans le but de faciliter le processus de recherche et d'offre d'emploi sur le Web.

Mots-clés : Indexation sémantique, ontologies linguistiques, ressources humaines

1 INTRODUCTION

Les technologies de l'Internet ont introduit de nouvelles pratiques dans la gestion des ressources humaines. Les personnes en recherche d'emploi peuvent déposer leur Curriculum Vitae (CV) via le Web sur des serveurs dédiés ou l'envoyer directement à une entreprise par le biais du courrier électronique. En sens inverse une entreprise peut aussi diffuser ses offres d'emploi via le Web [1]. Ces dernières années, le Web est devenu dans ce domaine, un vecteur de diffusion de l'information très utilisé. Pourtant, ces grandes masses de données sont souvent mal exploitées car les techniques disponibles de gestion de CVs sont limitées face à l'afflux d'information à traiter. Dans cet article, nous montrons une approche liée au Web sémantique qui permet d'extraire et de structurer la connaissance issue des CVs. Cette approche est liée à des travaux qui ont été ménés dans le cadre de l'indexation par le contenu de documents issus du Web (Desmontils & Jacquin, 2002).

2 LES TRAVAUX EXISTANTS

2.1 Le système BONOM

Le système BONOM (Cazalens & Lamarre, 2001) (Cazalens *et al.*, 2002) est un système multi-agent développé à l'IRIN pour la recherche d'informations et de connaissances distribuées sur le Web. Il propose :

- une organisation d'agents selon une hiérarchie de domaines informationnels (thèmes). Les agents sont structurés en groupes relevant chacun d'un domaine. Le système dispose principalement de trois types d'agent: des agents sites, des agents intermédiaires et des agents utilisateurs ;
- des mécanismes d'analyse de sites (situés sur les agents sites) qui permettent d'indexer le contenu d'un site Web en fonction d'ontologie(s) représentative(s) des domaines couverts par le site ;
- des protocoles d'interaction permettant, à travers une recherche, d'atteindre les sites les plus pertinents. Il est à noter que ce sont les agents sites qui s'inscrivent auprès d'agents succeptibles de leur envoyer les requêtes auxquelles ils peuvent répondre.

Dans le cadre du système BONOM, nous avons étudié et développé des fonctionnalités au niveau des agents site qui disposent de mécanismes semi-automatiques qui indexent leurs pages et structurent la connaissance qui leur est propre à l'aide d'ontologie (Desmontils & Jacquin, 2002).

1. Selon le sondage IPSOS Liaisons sociales/France Télécom du 5 février 2002 réalisé auprès de 300 entreprises de plus de 100 salariés, certaines grandes entreprises reçoivent jusqu'à 80 000 CVs électroniques par an. 41% des compagnies trouvent ces candidatures "trop nombreuses et pas assez ciblées" et 32% qu'elles ne sont pas "de bonne qualité" !

2.2 Indexation semi-automatique de documents Web par le contenu

Le processus d'indexation semi-automatique des documents s'appuie sur des techniques issues du traitement automatique des langues et de l'ingénierie des connaissances (Desmontils & Jacquin, 2002). Dans le cadre de ces recherches, une suite d'outils a été développée qui permet d'effectuer une indexation structurée d'un site Web selon une ontologie donnée. Ces recherches ont mené à la détermination des processus suivants (figure 1) :

- Un processus qui extrait un ensemble de concepts candidats issus de pages Web, appelé index à plat. Il permet d'extraire les termes bien formés issus des pages d'un site à l'aide d'analyses linguistiques (à l'aide de patrons morpho-syntaxiques) et de calculer un coefficient relatif à leur importance dans la page (celui-ci est fonction de la fréquence des termes et du poids des marqueurs HTML associés). Et d'autre part, il permet aussi de construire les concepts candidats représentatifs du contenu des pages à l'aide du thesaurus WordNet (Miller, 1990) et d'une mesure de similarité sémantique (qui prend en compte le contexte des concepts dans les pages).
- Un processus concernant la désambiguïsation des labels des concepts de l'ontologie. Les pages Web sont des documents faiblement structurés et écrits en langage naturel. Pour faire le lien entre ces documents et les ontologies, un processus de désambiguïsation des labels associés aux concepts des ontologies a été déterminé. Ce processus s'appuie sur des heuristiques exploitant les relations de généralisation et de spécialisation présentes dans l'ontologie et les relations d'hyperonymie et d'hyponymie présentes dans une ontologie linguistique spécialisée construite à partir d'une ontologie issue du projet SHOE (Heflin *et al.*, 1999) et du thesaurus WordNet (nous avons travaillé sur les universités américaines).
- Un processus d'appariement des concepts candidats issus des documents et des concepts de l'ontologie. À partir de l'ontologie linguistique, certains concepts candidats de l'index à plat sont retenus et les documents correspondants (les pages Web dans notre cadre d'étude) sont associés aux concepts correspondant de l'ontologie. Ceci permet de construire un index structuré des documents. La structure est donnée par l'ontologie. De nombreuses expérimentations et la mise au point de mesures d'évaluation spécifiques ont été effectuées.

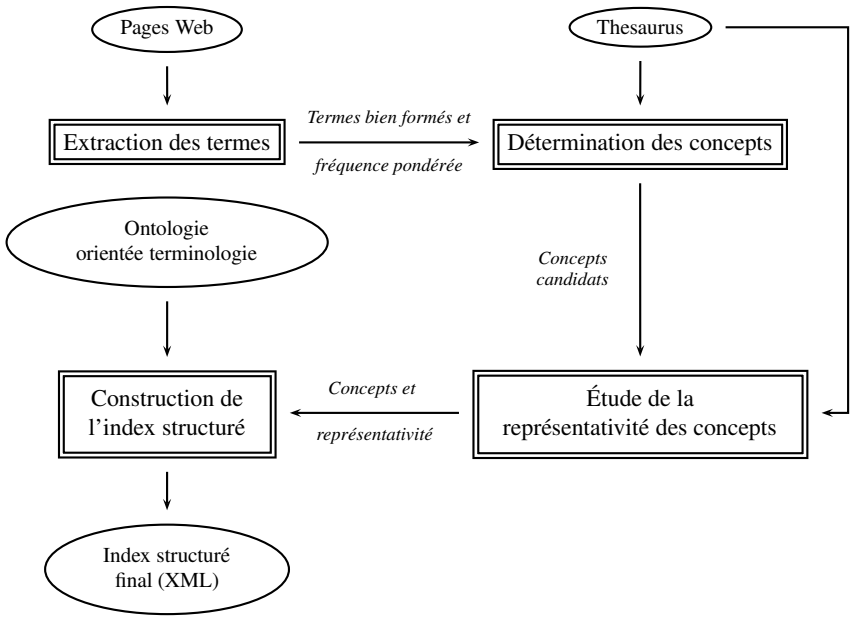

FIG. 1 – *Le processus d'indexation*

3 TRAVAUX EN COURS: APPLICATION AUX RECRUTEMENTS

Dans le cadre d'un contrat [2], nous étudions l'adaptation du système BONOM à la recherche de profils dans un ensemble de sites de CVs libres.

3.1 Adaptation du système multi-agent

Dans le système BONOM, les agents sont organisés relativement à une hiérarchie de domaine. Dans le cadre du recrutement, leur structuration s'effectue relativement à un domaine d'activité. Si on prend l'exemple des domaines de l'informatique, on aura un domaine relatif aux bases de données, un domaine relatif aux réseaux... Ceci signifie que les requêtes, émises par les agents utilisateurs, seront véhiculées via cette hiérarchie vers les agents sites qui se sont inscrits pour recevoir des requêtes de ces domaines spécifiques. À chaque domaine spécifique est associé une ontologie relative aux emplois/métiers liés à ce domaine. Cette ontologie est partagée entre les agents intermédiaires de ce domaine, les agents sites et les agents personnels qui les contactent. Selon les CVs qu'ils ont à mettre en valeur, les agents sites s'adressent à des nœuds de la hiérarchie du domaine d'activité.

3.2 Adaptation du système d'indexation sémantique

3.2.1 Extraction d'information à partir de CVs

Par rapport aux documents généraux contenus dans un site Web, les CVs ont des spécificités propres. Il s'agit de traiter des documents très courts (le plus souvent ne dépassant pas une page) et syntaxiquement pauvres (par exemple, absence de sujet et de verbe dans les phrases). En revanche, un CV apparaît comme un document relativement bien structuré au sein duquel il est facile pour le lecteur d'identifier les données signalétiques, la formation, les expériences professionnelles, les loisirs du scripteur sans pour autant être un spécialiste du domaine. Nous avons donc affaire à un objet visuel dont les propriétés visuelles sont directement exploitées par le lecteur lors de sa compréhension du CV (Virbel, 1985) (Pascual, 1991) (Péry-Woodley & Rebeyrolle, 1998). Ces caractéristiques dispositionnelles représentent des caractéristiques identitaires et structurelles importantes dans le CV et s'ajoutent aux caractéristiques lexicales, syntaxiques et typographiques traditionnelles exploitées en extraction d'information.

Dans un premier temps, nous exploitons les caractéristiques dispositionnelles d'un CV pour identifier les différentes zones informationnelles. Cette identification repose, d'une part, sur le repérage des titres des zones informationnelles (à l'aide d'une liste de mots clés), et d'autre part, sur le formatage propre à chaque CV de ces titres (pour un CV donné, les titres ont presque toujours le même style). Suite à l'identification de ces différentes zones, nous les traitons individuellement en exploitant des grammaires locales.

Données signalétiques

L'identification des données signalétiques comme l'identité, l'adresse, le téléphone, l'âge, la situation familiale, l'adresse électronique du scripteur repose sur une analyse de surface qui met en œuvre, entre autre, les notions d'évidences internes (par exemple, une adresse électronique devra comporter le symbole @) et externes (par exemple, l'adresse sera précédée par un code postal) définies par (McDonald, 1998). À cet effet, le système de reconnaissance d'entités nommées *Nemesis* développé par (Fourour, 2002) sera adapté à cette tâche.

Formations

Les informations de cette zone sont le plus souvent organisées sous la forme d'une énumération chronologique, inversée ou non, du parcours scolaire. Chaque item de l'énumération renvoie à une étape scolaire de ce parcours. À ce niveau la principale difficulté est d'identifier la nature de la formation réalisée. Nous devrons déterminer si des connaissances génériques sur les formations sont suffisantes (par exemple, le nom, l'abréviation des diplômes) ou si nous devons exploiter des ressources caractéristiques du domaine d'étude (ontologie des formations). Au sein de cette zone, nous pouvons aussi exploiter la régularité syntaxique de la construction des items (par exemple : si le premier élement donne la date, le nom du diplôme puis son lieu d'obtention, les autres items auront la même construction). D'autres questions devront aussi être

2. avec la fondation «Recherche & Emploi» patronnée par la Fondation de France et la société de travail intérim Védiorbis

précisées : Est-ce qu'il s'agit d'une formation validée ou non? Y a-t-il un classement, une mention relative à l'obtention de cette formation?

Expériences professionnelles

L'expérience professionnelle est une donnée précieuse pour appréhender le profil et le niveau d'expérience du scripteur. À ce niveau, il faut déterminer, le nom de l'entreprise et sa localisation géographique, la durée et le type de la mission exercée par le scripteur. Ici encore, l'énumération est régulièrement exploitée ; chaque item renvoyant à une expérience professionnelle particulière. À cette étape, notre principale difficulté provient de la construction du sens de cette expérience à partir de l'identification de ses composants (nom de l'entreprise, localisation géographique, type de mission). Le repérage des noms de lieu et d'entreprise sera réalisé par le système *Nemesis* (Fourour, 2002). L'identification et la mission reposera sur le repérage de mots clés (par exemple: stage, projet, réalisation, gestion) et sur l'extraction de termes. Les mots clés servant d'ancre pour repérer la mission, alors que le candidat terme permettra de la caractériser.

Compétences

Les informations disséminées à l'intérieur de cette zone sont relativement difficiles à identifier dans la mesure où elles sont le plus souvent caractéristiques d'un domaine d'étude (par exemple, en informatique, on s'intéressera aux connaissances du scripteur sur les systèmes d'exploitation, les environnements logiciel, les méthodes d'analyse). À ce niveau, l'usage de données spécifiques au domaine est nécessaire. Dans le cadre de ce travail, nous envisageons d'exploiter une ontologie spécifique du domaine. Le niveau de maîtrise d'une langue est aussi une donnée relative aux compétences du scripteur. L'extraction de cette information engendre moins de problèmes, car elle est peu caractéristique du domaine et sa définition tend à se normaliser.

3.2.2 Indexation des CVs

Construction des ontologies

Nous avons déterminé les types d'ontologies dont nous aurons besoin dans ce contexte d'applications : ontologies métiers, ontologies compétences, ontologies formations. Le formalisme à utiliser pour représenter ces connaissances reste à préciser. Il devra pouvoir permettre de représenter les connaissances du domaine mais aussi permettre de prendre en compte les différentes formes linguistiques sous lesquelles peut apparaitre un concept (les synonymes). En effet, c'est ce point précis qui permettra de passer du niveau linguistique (passage des formes linguistiques dans les CVs au niveau des concepts de l'ontologie). le processus de désambiguisation des étiquettes liées aux concepts des ontologies développé dans le laboratoire et qui permet de construire une ontologie linguistique sera mis en œuvre sur les ontologies définie dans ce nouveau cadre d'application. La définition précise des différents concepts, relations... propres aux ontologies sera réalisée à l'aide de l'analyse des CVs menée à la phase précédente mais aussi à l'aide de ressources externes comme le ROME (Répertoire Organisationnelles des Métiers et Emplois) et les documents du CIGREF (Club Informatique des Grandes Entreprises Françaises).

Instanciation des ontologies

Après la phase d'extraction d'informations par des processus issus du domaine du TALN (Traitement Automatique des Langues) et de génération de concepts candidats, le travail consistera à adapter à notre cas d'étude, les processus d'appariement de concepts dévéloppés au laboratoire pour relier les concepts issus des CVs aux concepts des ontologies. Le résultat final consistera en des ontologies dont certains concepts seront reliés aux concepts issus des CVs.

4 CONCLUSION

À l'heure actuelle, nous avons étudié les informations qui peuvent être extraites automatiquement des CVs à l'aide de techniques issues du TALN. Nous avons commencé la construction des ontologies en nous restreignant à un domaine d'étude qui est celui des emplois/métiers en informatique. Bien que ce projet débute, nous pouvons déjà en tirer un certain nombre d'enseignements. Après analyse d'un corpus de CVs, on

constate une certaine régularité structurelle. Par conséquent, contrairement aux documents Web généraux que nous traitions dans le cadre du projet BONOM, les CVs sont des documents relativement structurés. Cette structure (zone concernant la formation, l'expérience...) peut être exploitée efficacement pour catégoriser l'information de manière plus fine. Par rapport au projet BONOM, où nous disposions d'une ontologie linguistique générale (WordNet) pour extraire les concepts et une ontologie spécialisée pour réaliser l'indexation, ici l'utilisation d'une ontologie linguistique générale semble moins primordiale. En effet, les termes issus des CVs sont peu ambigus, car d'une part l'agent site qui indexe connait le domaine dans lequel un CV s'insére (c'est un domaine que le site couvre et il sait de quel domaine relève un CV) et d'autre part la structuration forte des CVs aide aussi à lever des ambiguités. Donc, des ontologies spécialisées relatives aux formations, aux emplois/métiers, aux compétences peuvent suffire du fait du domaine d'étude.

RÉFÉRENCES

CAZALENS S., DESMONTILS E., JACQUIN C. & LAMARRE P. (2002). Sources d'informations et de connaissances : de la gestion locale à la recherche distribuée. *Revue l'Objet*, **8(4)**.

CAZALENS S. & LAMARRE P. (2001). An organization of internet agents based on a hierarchy of information domains. In Y. DEMAZEAU & F. J. GARIJO, Eds., *Proceedings MAAMAW*.

DESMONTILS E. & JACQUIN C. (2002). *Indexing a Web Site with a Terminology Oriented Ontology*, In *The Emerging Semantic Web*, p. 181–197. IOS Press, i.f. cruz and s. decker and j. euzenat and d. l. mcguinness edition.

FOUROUR N. (2002). Nemesis: un système de reconnaissance incrémentielle des entités nommées pour le français. In *TALN'02*, p. 255–264, Nancy, France.

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)*.

MCDONALD D. (1998). Internal and external evidence in the identification and semantic categorization of proper names. In B. B. . J. PUSTEJOVSKY, Ed., *Corpus Processing for Lexical Acquisition , Language, Speech and Communications*: MIT Press.

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

PASCUAL E. (1991). *Représentation de l'architecture textuelle et génération de texte*. PhD thesis, Univ. De Toulouse.

PÉRY-WOODLEY M. P. & REBEYROLLE J. (1998). Domain and genre in sublanguage text: definitional microtexts in three corpora. In R. C. . A. T. A. RUBIO, N. GALLARDO, Ed., *First International Conference on Language Resources and Evaluation*, p. 987–992.

VIRBEL J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. In *Cahiers de grammaire*, p. 5–72: cahiers de grammaire.