Extraction des informations d'entreprises

A. Fotsoh Tawofaing[*,**] **— A. Le Parc Lacayrelle**[**] **— C. Sallaberry**[**] **— T. Moal**[*]

[*] *Laboratoire LUIPPA, BP-1155, 64013 PAU Université Cedex, France*
[**] *Cogniteev, 2 Rue Doyen Georges Brus, 33600 Pessac, France*

RÉSUMÉ. *Ce papier propose un service de recherche d'entreprises sur le web. Le besoin d'information est exprimé par des critères thématiques (activités et métiers de l'entreprise, produits commercialisés, ...) ainsi que des critères spatiaux (lieu). Le système comprend un premier module de filtrage de sites web d'entreprises. Un second module analyse ces sites afin d'en extraire automatiquement toutes les informations contextuelles (activités, métiers, produits, contacts, adresses, ...). Notre proposition se distingue des services similaires que l'on peut trouver sur le web par (i) la richesse des informations thématiques (catégorisation détaillée des activités, des produits, des métiers) ; (ii) la provenance web des informations ; (iii) la combinaison de critères de recherche thématique, spatial et plein texte. Un prototype est développé pour mettre en œuvre notre proposition.*

ABSTRACT. *Searching information about local businesses is not a trivial problem to address. Most of existing services are supplied with manually recorded data. Based on the observation that more and more businesses are presented on the web, we propose in this paper a new approach, which consists to extract companies targeted information (addresses, activities, jobs, products, emails, fax) from websites, to supply a local businesses search service. The retrieval information module combines thematic, spatial and full-text criteria. A prototype of this service is implemented to experiment our proposal.*

MOTS-CLÉS : *Analyse du web, Extraction d'informations spatiales, Extraction d'informations thématiques.*

KEYWORDS: *Web mining, Spatial information extraction, Thematic information extraction.*

1. Introduction

De plus en plus d'entreprises sont présentes sur le web et publient de nombreuses informations relatives à leurs activités, leurs métiers, leurs produits et leurs coordonnées (adresses, numéros de téléphone, numéros de fax, adresses mail). Partant de ce constat, le projet Cognisearch Business vise à exploiter ces données pour offrir un service de recherche d'entreprises capable de répondre à des besoins d'information du type « charpente en chêne au sud de Poyartin ». Ce type de recherche comporte une dimension thématique (charpente en chêne) et une dimension spatiale (au sud de Poyartin).

Plusieurs solutions sur le web proposent des services permettant de répondre à ce type de recherche. Nous pouvons les classer en trois catégories principales : (i) les fournisseurs de données, comme Factual[1] ou Axciom[2], qui collectent et commercialisent les informations d'entreprises. Les informations alimentant ces services proviennent généralement des «données ouvertes», du crawl du web ou même de plateformes partenaires ; (ii) les annuaires, qui sont des bases de données d'informations d'entreprises consultables en ligne. Dans cette catégorie, nous pouvons citer des services comme Google Maps, Google My Business[3], Pages Jaunes, ou même société.com (annuaire d'entreprises françaises déclarées au registre du commerce) ; (iii) les réseaux sociaux, qui sont beaucoup plus orientés partage d'informations et d'appréciations portant sur les commerces, les places et les événements dans une région. C'est le cas des services comme Yelp[4], Foursquare[5], Facebook Places[6]. Dans les catégories (ii) et (iii) les données proviennent, en général, de contributions (salariés et utilisateurs) et de fournisseurs de données. Cependant, ces solutions (i) ne permettent pas de prendre en compte toutes les relations topologiques dans la dimension spatiale ; (ii) peuvent mal interpréter la dimension thématique dans certains cas ; (iii) s'appuient en grande partie sur des données saisies manuellement. (Ahlers, 2013) propose un système d'analyse de pages web pour enrichir la base de données entreprises des Pages Jaunes. L'approche proposée consiste à croiser les données des Pages jaunes, avec celles de l'annuaire DMOZ[7] pour identifier les pages web associées à une entreprise et en extraire des informations (adresses, numéros de téléphone, adresses mail, données commerciales, données fiscales). Cependant, la proportion d'entreprises existant dans une région et qui sont référencées sur DMOZ reste faible pour le cas de la France (par exemple, on a 2580 entrées DMOZ au total pour la région d'Aquitaine alors qu'il y a plus de 250000 entreprises déclarées au registre du commerce), d'où l'enrichissement limité de la base de données.

1. http ://www.factual.com/
2. http ://www.acxiom.fr/
3. http ://www.google.com/business/
4. http ://www.yelp.fr/
5. http ://fr.foursquare.com/
6. http ://fr-fr.facebook.com/places/
7. http ://www.dmoz.org/

Notre objectif est de construire un service de recherche d'information d'entreprises qui s'appuie sur des données publiées sur le web et qui combine des critères spatiaux (prenant en compte différentes relations topologiques) et des critères thématiques avec la recherche plein texte. Notre proposition se distingue des services existants par son indépendance vis à vis des données enregistrées manuellement ; en effet, notre service est alimenté uniquement par des données publiques et des informations extraites du web. Ce service se veut également plus précis dans l'interprétation de la zone spatiale couverte par les besoins d'informations en prenant en compte les différentes relations topographiques (Vaid *et al.*, 2005) exprimées dans les requêtes. A la différence de (Ahlers, 2013), les sites web d'entreprises à partir desquels se fait l'extraction d'information sont identifiés par une heuristique à partir des données du registre du commerce. Cette démarche permet d'avoir un corpus plus exhaustif que celui construit à partir de l'annuaire DMOZ. Par ailleurs, au delà des données de localisation (adresses), nous extrayons également les métiers, produits et activités d'entreprises.

La suite de l'article est structurée comme suit. La section 2 présente l'architecture générale de notre service. La section 3 explique le processus de constitution de notre corpus de sites web. La section 4 détaille l'annotation de ces sites, tandis que la section 5 présente l'indexation des données. Le prototype développé est décrit en section 6. La section 7 montre une première évaluation. Enfin, la section 8 conclut l'article et présente les perspectives envisagées.

2. Architecture du service Cognisearch Businness

Nous proposons un service de recherche d'informations d'entreprises qui s'appuie uniquement sur des données publiées sur le Web. Nous avons défini un modèle de représentation des entités entreprise qui est composé de deux parties. La première correspond aux données d'immatriculation (de base) des entreprises (nom officiel, numéro SIRET, ...). La seconde partie est composée des données extraites des sites web des entreprises qui sont relatives à ses activités, produits et métiers. L'architecture de ce service est décrite figure 1. Une première étape consiste à identifier sur le Web les pages publiées par les entreprises (figure 1.1). Une fois le corpus de pages web constitué, une étape d'annotation (figure 1.3) permet d'en extraire automatiquement les informations relatives aux activités, aux produits, aux métiers ainsi qu'aux coordonnées de chaque entreprise (adresses, numéros de téléphones, emails, fax). Les informations annotées viennent compléter les données d'immatriculation des entreprises afin de constituer un premier index d'informations d'entreprises. En parallèle, le texte des sites web est indexé afin de permettre la recherche plein-texte (figure 1.6).

3. Filtrage de sites web

Les données d'immatriculation, telles que renseignées au registre du commerce sont issues automatiquement de la ressource societe.com, et constituent les données de

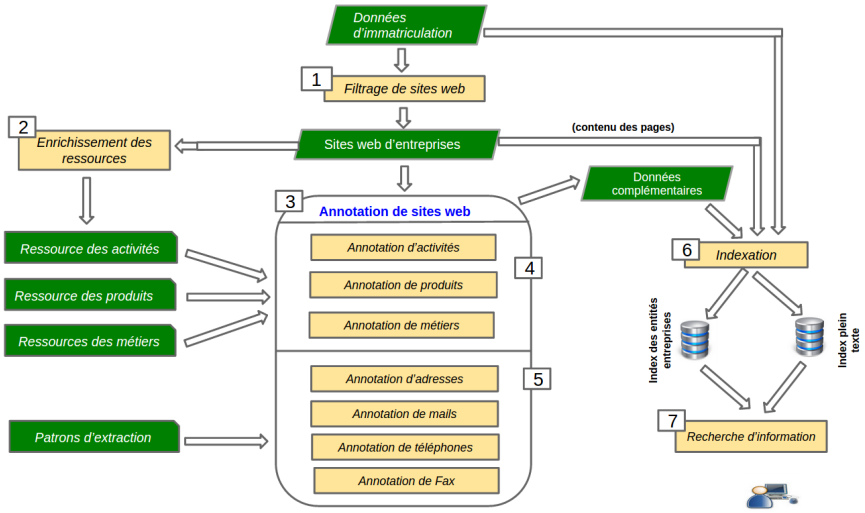

Figure 1. *Architecture de Cognisearch Business*

base d'une entité entreprise. Un traitement basé sur une heuristique combine ces données pour filtrer le web et retrouver, s'il existe, le site web associé à chaque entreprise (figure 1.1). Nous avons testé plusieurs combinaisons sur 300 compagnies françaises. Les annuaires d'entreprises et les réseaux professionnels ont été répertoriés dans une liste (appelée *liste noire*) afin d'être exclus lors de la recherche. L'algorithme retenu est le suivant. Pour chaque entreprise du registre du commerce, nous commençons par interroger le web en combinant son nom commercial (si il existe) avec sa localisation. Nous conservons, parmi les trois sites les plus pertinents, le premier qui n'est pas dans la liste noire. Si tous les sites sont dans la liste noire, nous recommençons le même principe mais en combinant cette fois le nom officiel de l'entreprise avec sa localisation. Si de nouveau, les trois premiers sites trouvés sont dans la liste noire, nous en concluons que l'entreprise n'a pas de sites web. L'ensemble des sites web ainsi obtenu, constitue le corpus d'entrée pour le processus d'annotation (figure 1.3).

4. Annotation des sites web

L'étape d'annotation (figure 1.3) permet d'extraire automatiquement des informations thématiques et spatiales relatives aux différentes entreprises. En ce qui concerne les informations spatiales, très peu de sites web utilisent les microformats ou des balises particulières pour les définir. Les adresses peuvent être situées n'importe où dans la page et les différentes informations qui les composent ne sont pas écrites forcément dans le même ordre.

4.1. *Extraction d'informations thématiques*

Afin d'annoter automatiquement les métiers, les produits et les activités contenues dans les sites web, nous avons construit, par transformation de modèles, trois ressources de type connaissance (sous forme d'ontologies au format OWL) décrivant, respectivement, les activités, les produits et les métiers d'entreprises. Ces ressources ont été construites à partir des deux hiérarchies d'organisation des activités et des produits définies par l'INSEE[8] (NAF[9] et CPF[10]) et de la hiérarchie définie par Pôle Emploi qui organise les métiers et emplois en catégories socio-professionnelles (ROME[11]). Le choix de ces ressources se justifie notamment par le niveau de finesse dans la description des différentes catégories.

L'ontologie relative aux activités ainsi construite est pauvre en vocabulaire. Nous avons de ce fait, mis en œuvre un processus pour son enrichissement de façon semi-automatique (figure 1.2). Ce processus utilise l'apprentissage pour regrouper les expressions des sites web qui sont communes à une catégorie précise d'activité. Pour cela, chaque classe de la hiérarchie des activités est associée à un ensemble de sites web d'entreprises. Un algorithme de clustering permet de mettre en évidence les expressions communes à tous les sites web d'une classe. Ces expressions constituent un vocabulaire potentiel pour enrichir la ressource. Une phase de validation par un expert intervient par la suite pour sélectionner les expressions pertinentes. L'algorithme de clustering que nous avons utilisé dans notre proposition est Latent Dirichlet Allocation (LDA) illustré dans (Blei *et al.*, 2003).

L'extraction d'informations relatives aux métiers, aux produits et aux activités d'entreprises est réalisée en utilisant les ontologies construites. Chaque syntagme nominal, trouvé dans le texte de la page web, est annoté de l'identifiant de la classe associée dans l'ontologie correspondante.

En ce qui concerne les emails, les numéros de téléphone et numéros de fax, des patrons sont utilisées pour leur extraction. Ces patrons ont été mis au point à partir de l'observation d'un échantillon de sites web. Par exemple, le patron permettant d'extraire les emails est le suivant :
$Email \rightarrow Login(\text{"@"}|\text{"}at\text{"})NomDomaine(\text{"."}|\text{"}(dot)\text{"})ExtensionDomaine$

4.2. *Extraction d'informations spatiales*

Deux approches principales peuvent être utilisées pour l'annotation automatique des adresses : une première utilise des patrons d'extraction ((Borges *et al.*, 2007), (Blohm, 2011), (Ahlers et Boll, 2008)) et une deuxième s'appuie sur les techniques d'apprentissage ((Loos et Biemann, 2008), (Taghva *et al.*, 2005)). Nous avons choisi

8. http ://www.insee.fr/fr/
9. Nomenclature Française des Activités
10. Classification des Produits Française
11. Répertoire Opérationnel des Métiers et des Emplois

Noms de champs	Exemples	Noms de champs	Exemples
CA : Complément Adresse	Résidence Rigaud	CP : Code Postal	64000
INV : Introducteur Nom Voie	Avenue	C : Commune	Pau
BP : Boite Postale	BP 1167	NV : Numéro Voie	10 ter
CS : Course Spéciale	CS 2587	D : Département	Gers
NC : Numéro Courrier	CEDEX 01	P : Pays	France
NVo : Nom Voie	Avenue de l'université		

Tableau 1. *Les différents champs pouvant composer une adresse*

une approche basée sur les patrons d'extraction. Ces approches exploitent en général des gazetiers répertoriant les noms de toutes les voies, le noms des villes, ... Dans le contexte français, une ressource complète contenant tous les noms de voie n'est pas disponible en accès libre. Une des difficultés est donc d'identifier le nom de la voie dans une adresse. En effet, il peut y avoir un ou plusieurs compléments qui peuvent être positionnés avant et/ou après le nom de la voie. Le tableau 1 répertorie les différentes informations que l'on peut trouver dans une adresse française.

L'observation d'un échantillon de 160 sites web d'entreprises a permis de définir des patrons, dont voici un extrait :

$$Adresse \rightarrow CA?\ \ ((BP\ \ CS)\,|\,(CS\ \ BP)\,|\,BP\,|\,CS)?\ \ NV?\ \ NVo\ \ CA?$$
$$((BP\ \ CS)\,|\,(CS\ \ BP)\,|\,BP|\,CS)?\ \ ((CP\,C)\,|\,(C\,CP))\ \ NC?\ \ D?\ \ P?$$
$$Adresse \rightarrow CA?\ \ ((BP\ \ CS)\,|\,(CS\ \ BP)\,|\,BP\,|\,CS)?\ \ ((CP\,C)\,|\,(C\,CP))$$
$$NC?\ \ D?\ \ P?$$
$$Adresse \rightarrow CA?\ \ ((BP\ \ CS)\,|\,(CS\ \ BP)\,|\,BP\,|\,CS)?\ \ NV?\ \ NVo\ \ CA?$$
$$((BP\ \ CS)\,|\,(CS\ \ BP)\,|\,BP|\,CS)?\ \ CP\ \ NC?\ \ D?\ \ P?$$

Dans le premier patron, le nom de voie, le code postal et la commune sont obligatoires, et peuvent être complétés par d'autres informations («10 Rue du Maréchal Foch, 49000 Angers», «Résidence des Aubiers, 3e Étage, 14 ter Rue de la République, 64000 Pau, France»). Cette forme est la plus fréquente dans l'échantillon (75% des adresses). Pour le deuxième patron, le code postal et la commune sont obligatoires et le nom de voie est absent («Résidence Rigaud 33350 Mouliets-et-Villemartin»). Le patron 3 permet entre autre d'identifier le cas où on a un code postal et un nom de voie sans commune («10 Place de la République, F-33600»). Cette dernière forme est assez rare (moins de 4% des adresses de l'échantillon).

5. Indexation

Les annotations sont extraites des pages web pour la construction des entités finales à indexer. Ces annotations sont rajoutées aux données d'immatriculation de chaque entreprise ainsi que l'adresse du site web correspondant. Les coordonnées géolocalisées

correspondant à chaque adresse extraite sont calculées avant l'ajout. L'entité finale est stockée dans un index (figure 1.6). De plus, une opération parallèle consiste à extraire le contenu textuel des pages de chaque site web d'entreprise et à l'indexer.

Les index ainsi construits sont utilisés pour répondre à des besoins d'information qui supportent des critères d'interrogation multidimensionnels et exploitent les caractéristiques spatiales, thématiques et plein texte contenues dans les index (figure 1.7).

6. Prototype

Un premier prototype mettant en oeuvre notre approche a été développé. Il traite des données relatives aux entreprises de la région Aquitaine pour laquelle 254 000 entreprises ont été identifiées. Parmi ces entreprises, nous nous sommes intéressés uniquement à celles traitant de 6 domaines d'activités : commerce, construction, hébergement & restauration, enseignement, information & communication et activités scientifiques & techniques. Ceci a réduit la liste initiale à 115 000 entreprises. Le module de filtrage de sites web a permis d'identifier un site pour 22 000 d'entre elles. Le corpus d'analyse relatif à ces 22 000 entreprises est constitué de 550 000 pages web. Il a été constitué en utilisant l'outil Nutch[12] de la fondation Apache.

Le module d'annotation des sites web utilise la plateforme GATE[13] couplée avec le framework HADOOP[14] pour gérer la volumétrie et la scalabilité du processus. Cette étape a permis d'annoter 30 000 adresses, 44 000 labels d'activités, 12 500 labels de produits et 28 000 labels de métiers. Les entités entreprises, construites à partir des annotations extraites, ont été indexées sous Elasticsearch[15], ainsi que le texte des pages web. Les deux index ainsi construits (index des entités entreprises et index plein texte) ont une taille globale de l'ordre de 3 GB.

7. Évaluation

Nous avons fait une première évaluation du processus d'annotation pour les adresses et les activités. Etant donné qu'il n'existe pas de campagnes d'évaluation pour ce type d'entité, nous nous sommes inspirés du processus utilisé dans TREC (Voorhees et Harman, 2005). Les résultats sont résumés dans le tableau 2. Le rappel et la précision sont calculés pour chaque page, les valeurs globales de ces métriques étant calculées en faisant la moyenne arithmétique. La F_1-mesure est évaluée en faisant la moyenne harmonique des métriques globales obtenues pour l'ensemble de l'échantillon.

12. http ://nutch.apache.org/
13. http ://gate.ac.uk/
14. http ://hadoop.apache.org/
15. http ://www.elastic.co/

	Pertinentes	Annotées	Pertinentes et annotées	Précision	Rappel	F_1-mesure
Adresses	309	286	251	0,80	0,81	0,80
Activités	1131	1119	631	0,45	0,52	0,49

Tableau 2. *Évaluation du processus d'extraction d'information*

En ce qui concerne les adresses, notre échantillon est constitué de 240 pages web contenant au moins une adresse. 309 adresses ont été annotées par un expert, tandis que notre annotateur en a trouvé 286. Parmi ces 286, seules 251 sont jugées pertinentes, 13 étant incomplétement annotées. Ces résultats s'expliquent, d'une part, par le fait que nous avons considéré, dans nos patrons d'extraction, que toutes les adresses avaient obligatoirement un code postal, ce qui n'est pas le cas dans notre échantillon. D'autre part, certaines villes sont mal orthographiées.

En ce qui concerne les activités, notre échantillon est constitué de 100 pages web relatives à des entreprises travaillant dans le domaine de la toiture (charpente, couverture, zinguerie, ...). 1131 activités ont été annotées par un expert tandis que notre annotateur en a trouvé 1119. Parmi ces 1119, seules 631 sont pertinentes. Ces résultats s'expliquent par un défaut de vocabulaire dans l'ontologie des activités en dépit de son enrichissement. Par exemple, l'expert va annoter «charpente traditionnelle en chêne» alors que l'ontologie contient seulement le label «charpente en chêne».

8. Conclusion

Nous avons présenté dans cet article la première étape de la construction du service de recherche géo-localisé d'informations portant sur les entreprises. Cette étape est centrée sur l'extraction des informations sur le web et leur restructuration dans des index. Ce processus s'appuie sur plusieurs problématiques de recherche différentes, qui deviennent complémentaires dans le processus de construction des entités entreprises. Il s'agit notamment de l'apprentissage qui est utilisé pour l'enrichissement des ressources de type connaissance. De plus, le processus d'annotation de texte combine les approches basées sur les patrons d'extraction et celles exploitant les bases de connaissances. Il permet d'extraire les adresses des sites malgré les différences de formes que l'on peut trouver. Un premier prototype mettant en oeuvre notre approche a été développé. Il montre la faisabilité et l'intérêt de l'approche.

Une deuxième étape consistera à exploiter ces index pour répondre à des besoins d'information combinant les critères spatiaux, thématiques et plein texte. Une évaluation du service avec un jeu de requêtes représentatif sera également menée au terme de cette dernière étape.

9. Bibliographie

Ahlers D., « Business entity retrieval and data provision for yellow pages by local search », *Workshop of Integrating IR technologies for Professional Search*, 2013.

Ahlers D., Boll S., « Retrieving Address-based Locations from the Web », *International Workshop on Geographic Information Retrieval (GIR)*, ACM, New York, NY, USA, p. 27-34, 2008.

Blei D. M., Ng A. Y., Jordan M. I., « Latent dirichlet allocation », *Journal of machine Learning research*, vol. 3, p. 993-1022, 2003.

Blohm S., *Large-scale pattern-based information extraction from the world wide web*, KIT Scientific Publishing, 2011.

Borges K. A. V., Laender A. H. F., Medeiros C. B., Davis Jr. C. A., « Discovering Geographic Locations in Web Pages Using Urban Addresses (GIR) », *International Workshop on Geographical Information Retrieval (GIR)*, ACM, p. 31-36, 2007.

Loos B., Biemann C., « Supporting web-based address extraction with unsupervised tagging », *Data Analysis, Machine Learning and Applications*, p. 577-584, 2008.

Taghva K., Coombs J. S., Pereda R., Nartker T. A., « Address extraction using hidden markov models », *International Symposium on Electronic Imaging Science and Technology (IST/SPIE)*, p. 119-126, 2005.

Vaid S., Jones C. B., Joho H., Sanderson M., « Spatio-textual Indexing for Geographical Search on the Web », *International Conference on Advances in Spatial and Temporal Databases*, SSTD'05, Berlin, Heidelberg, p. 218-235, 2005.

Voorhees E. M., Harman D. K., *TREC : Experiment and Evaluation in Information Retrieval*, The MIT Press, 2005.