Extraction des connaissances à partir du Web pour la recherche des images géoréférencées

Houda BOUAMOR[1]

LIMSI-CNRS et Université Paris-Sud 11

B.P. 133, 91403 Orsay Cedex

houda. bouamor@limsi.fr

RÉSUMÉ. Les bases de données géoréférencées connaissent un rôle croissant dans une grande variété de domaines d'application. La création manuelle de ces bases de données est cependant une opération coûteuse. Cela a suscité un intérêt pour l'automatisation de leur construction, par exemple, par l'exploitation des informations géographiques présentes sur le Web. Dans ce travail, nous présentons une nouvelle approche automatique pour la construction d'une base de données géoréférencées multilingues et à large échelle en se basant principalement sur l'encyclopédie collaborative Wikipédia pour identifier les noms géographiques, catégoriser ces noms, trouver leurs coordonnées géographiques et les classer selon une estimation de leur pertinence. La base de connaissances obtenue a été intégrée dans ThemExplorer, une application de recherche d'images géoréférencées.

ABSTRACT. Geolocalized databases are becoming necessary in a wide variety of application domains. The manual creation of such databases is an expensive operation which stimulated the interest for the automation of their construction, by mining geographic information from the Web. In this article, we present and evaluate a new automated approach for creating a geographical database. Our technique is based on Wikipedia to identify geographical names, categorize them, find their geographical coordinates and rank them. Then this database has been integrated in ThemExplorer, an application for geographic image retrieval.

MOTS-CLÉS : Bases de données géoréférencées, Wikipédia, extraction d'information, fouille de données, ThemExplorer.

KEYWORDS: Geographic databases, Wikipedia, Information extraction, data mining, ThemExplorer

[1] Ce travail a été réalisé dans le cadre d'un stage de Master 2 au CEA/LIST/ LIC2M

1. Introduction

La structure minimale d'un thésaurus géographique est défini par Hill [4] de la façon suivante : chaque entité de la base doit être renseignée avec les informations suivantes : son nom (***Musée du Louvre***), sa localisation géographique représentée par les coordonnées (***48.862 N, 2.336 E***) et sa classe parent regroupant cette entité avec d'autres éléments de même nature (***Musée***). Les travaux de Hill ont permis de définir un schéma standard pour les bases de données géographiques et de mettre en place un système de classification hiérarchique des catégories géographiques. Or, la construction manuelle de bases de données géographiques, telles que Alexandria [4] ou Geonames[2] s'avère un processus long et fastidieux ce qui motive donc l'intérêt pour des travaux visant à automatiser ce processus.

Dans cet article, nous décrivons une méthode permettant de créer automatiquement une base de connaissances multilingue et à large échelle pour le domaine géographique en exploitant l'encyclopédie Wikipédia, une source d'informations riche en contenu semi-structuré qui a été utilisée dans plusieurs travaux de recherche (par exemple [2,5]). Un des objectifs de ce travail est l'amélioration de ThemExplorer [7], application de recherche d'images géoréférencées du CEA LIST, par l'intégration des données acquises.

Cet article est structuré de la façon suivante : dans la section 2, nous donnons un aperçu des travaux liés à notre approche. Dans la section 3, nous exposons notre méthode de création automatique du thésaurus géographique et la section 4 présente les différentes évaluations effectuées afin de valider notre approche.

2. État de l'art et approche suivie

Il existe un fort besoin d'accès à des informations géographiques, tel que montré par l'étude de Sanderson et Han [9] qui révèle que jusqu'à 37% des requêtes soumises à un moteur de recherche concernent des informations géographiques. Aujourd'hui, on dispose de plusieurs grandes sources d'informations géographiques créées manuellement telles que Alexandria et Geonames. Geonames contient plus de 8 millions de noms correspondant à environ 6,5 millions de lieux. Le nom, le type et les coordonnées sont **renseignés** pour chaque entité. Chaque entité appartient à une classe géographique bien déterminée ce qui offre la possibilité de mettre en place une navigation thématique. Les principales limites de Geonames concernent la variation de la couverture des différentes régions du monde et l'impossibilité d'afficher les résultats par ordre de pertinence. Or, comme souligné par Toriani et *al.* [10], ce classement est essentiel pour l'exploitation efficace des thésaurus géographiques en recherche d'information.

La constitution de bases de données géographiques à partir de corpus non structurés est illustrée par les travaux de Rattenbury et *al.* [8]. Leur approche vise à

[2] http://www.geonames.org/

séparer les noms géographiques des autres informations associées aux photographies géoréférencées de la base Flickr. Leur évaluation mesure une précision de 82% et un rappel de 50%. Buyukkoten et *al.* [3] proposent eux une méthode permettant l'identification et l'exploitation des informations géographiques à partir des sites Web afin de permettre aux moteurs de recherche d'exploiter ces informations à des fins de classement.

Le projet Gazetiki [6] est une base d'informations géographiques construite automatiquement par combinaison d'informations extraites de sources hétérogènes : Wikipédia, Panoramio et AllTheWeb. Son but est d'enrichir et compléter Geonames en utilisant un modèle du domaine permettant une intégration facilitée des deux ressources.

Notre travail se situe dans la continuité de celui de Gazetiki, consiste également à enrichir les informations de Geonames en ajoutant des données géographiques ainsi que des catégories relatives aux entités contenues dans Wikipédia. Alors que Gazetiki ne porte que sur l'anglais, notre projet à une visée multilingue et couvre six langues. Le processus d'extraction des coordonnées et de la valeur de pertinence des entités est similaire à celui de Gazetiki. L'identification des entités et leur catégorisation se fait de manière différente : la première se base sur la reconnaissance des articles géoréférencés de Wikipédia, la seconde sur l'extraction de plusieurs catégories candidates à partir des différentes parties d'un article, puis d'un processus de validation de la catégorie finale comme Ahern et *al.*[1], nous extrayons des fragments de pages à partir de Wikipédia afin de les catégoriser, mais nous exploitons également d'autres sources du Web. La limitation à un domaine précis nous permet de réaliser des analyses plus spécifiques et plus complexes des documents que dans d'autres travaux [2,8].

3. Extraction des informations à partir de Wikipédia et construction de la base des connaissances

Les articles de Wikipédia sont constitués de textes écrits en langage naturel, et comportent d'autres types d'informations structurées : les **Infobox***,* les informations sur les catégories, les coordonnées géographiques, et des liens vers les pages écrites en d'autres langues. Nous avons téléchargé les versions archivées de Wikipédia en six langues : *français, anglais, italien, espagnol, allemand* et *néerlandais,* afin de les utiliser pour construire notre gazetteer multilingue final. Notre travail consiste donc à extraire les éléments du tuple décrivant une entité géographique : *nom, localisation, catégorie, valeur de pertinence*.

La valeur de pertinence associée à chaque objet permet de classer les entités les plus populaires pour les présenter en priorité à l'utilisateur, ce qui est utile pour les régions à forte densité de monuments comme ***Paris*, *New York*...** Cette valeur a été calculée de la même façon que dans Gazetiki [6] .

3.1. Sélection des articles géoréférencés de Wikipédia

L'identification des articles géoréférencés repose sur la présence du couple {latitude, longitude}. Nous remarquons que le format des coordonnées géographiques n'est pas homogène notamment parce qu'elles sont souvent introduites par des non géographes. Le tableau 3.1 présente le nombre d'articles géoréférencés par langue.

	Anglais	**Français**	**Italien**	**Espagnol**	**Allemand**	**Néerlandais**
Nombre d'articles géoréférencés	242 142	76 477	88 513	45 534	96 405	122 915

Tableau 3.1: Nombre d'articles géoréférencés dans Wikipédia (Avril 2008) pour chacune des langues

Par ailleurs, certains articles ne sont écrits que dans une seule langue. Pour l'anglais, par exemple, il existe **107 611** articles parmi les **242142** qui ne sont pas traduits, **53438** ont une traduction en français, **48464** en italien, **28031** en espagnol, **56429** en allemand et **81305** en néerlandais.

3.2. Identification

Dans Wikipédia, chaque article a un nom unique qui est son titre, celui-ci comporte le nom de l'entité (ex : ***Château de Versailles***), accompagné dans certains cas du lieu où elle est située (ex : ***Manta (Ecuador)***) et dans d'autres, de sa catégorie (*ex :**Luzon (eiland)***). Le processus d'extraction du nom de l'entité passe par deux étapes. Dans la première, on extrait le nom dans la langue principale (celle dans laquelle l'article est écrit) en analysant la partie titre. Dans la seconde, on analyse les traductions, de cet article, afin d'extraire le nom de cette entité dans les autres langues. Le résultat est une entrée de la base de données comportant les noms des entités géographiques dans les six langues. La base de données ainsi créée sera enrichie par d'autres informations telles que la localisation de chacune des entités.

3.3. Localisation

Généralement, les coordonnées géographiques sont représentées par le couple (***latitude, longitude***) qui précise la position spatiale de l'objet. Mais dans Wikipédia, ce couple n'est presque jamais renseigné explicitement ou ne figure pas dans un format bien défini. Ce travail consiste à extraire toutes les coordonnées géographiques des entités représentées et à les convertir dans un format standard. Pour cela, nous avons extrait manuellement 31 motifs, parmi eux 7 sont communs aux six langues, un exemple de ces motifs est le suivant : *"Coor dms" : coordonnées en degré, minutes et secondes et direction N|S et E|O.*

3.4. Catégorisation

A partir de Wikipédia, nous avons établi un dictionnaire des catégories géographiques en anglais ainsi que leurs traductions dans les cinq autres langues. Pour un nombre limité de concepts, il n'existe pas de traduction, donc nous avons saisi la traduction manuellement. Ce dictionnaire nous permet de reconnaître la catégorie à extraire et d'enrichir la liste des catégories de *ThemExplorer*.

Notre méthode de catégorisation est apparentée à celle de Popescu et *al.* [6], mais nous utilisons plusieurs parties de la structure de l'article pour extraire des classes parents candidates et nous mettons en place une procédure de vote.

Les noms des objets géographiques contiennent souvent une référence explicite à leur catégorie, par exemple ***Tour* Eiffel**, **Golden Gate *Bridge***. Nous affectons temporairement cette catégorie à l'entité. Néanmoins, cela produit des erreurs pour des termes comme ***Cathedral of Learning*** qui n'est pas une cathédrale mais un gratte-ciel. De plus, cette méthode est inefficace pour les noms qui n'incluent aucune référence à leur classe, comme ***London Eye*** ou ***Parthenon.***

On analyse, ensuite, le contenu de la première phrase décrivant l'objet géographique. Celle-ci est habituellement une définition contenant une référence explicite à la catégorie cherchée. Par exemple, pour ***Notre Dame de Paris***, la première phrase est : ***Notre Dame de Paris is a Gothic cathedral on the eastern half of the Ile de la Cité.*** L'attribution de la catégorie est faite en deux étapes, nous cherchons la première apparition du verbe ***to be*** et retenons la partie à droite du verbe : ***a Gothic cathedral on the eastern half of …*** Puis, toutes les éléments du dictionnaire sont comparés au contenu de cette partie. Nous retenons comme deuxième catégorie temporaire celle qui apparaît la première.

Puis, on extrait les catégories prédéfinies par Wikipédia, situées en fin d'article et on les retient comme autres catégories temporaires.

Une procédure de vote est finalement mise en place pour choisir la catégorie la plus pertinente parmi toutes les catégories temporaires.

3.5. Extraction des entités en se basant sur le vocabulaire géographique

Il existe dans Wikipédia des articles définissant des entités géographiques mais qui ne sont pas géoréférencées. Pour découvrir ces entités, nous utilisons notre dictionnaire comme base pour la recherche des articles non géoréférencés. Cette méthode nous permet d'identifier le nom et la catégorie de chaque entité, mais pour les intégrer dans la base finale, il faut trouver leurs coordonnées géographiques.

4. Résultats et Évaluations

4.1. Résultats

Nous obtenons au final une base de connaissances multilingue contenant environ **700 000** entités : *noms, catégories, localisation et valeur de pertinence.*

Nous avons comparé nos résultats avec ceux de Geonames, celui-ci intègre des références à des articles Wikipédia, sans effectuer une analyse de leur contenu autre que l'extraction des coordonnées géographiques. Comme le montre la figure 5.1, notre approche assure une meilleure couverture dans l'extraction d'articles pertinents pour le domaine géographique dans la plupart des langues. La différence observée s'explique par le traitement de versions différentes de Wikipédia mais aussi par la grande diversité de motifs traités dans notre travail.

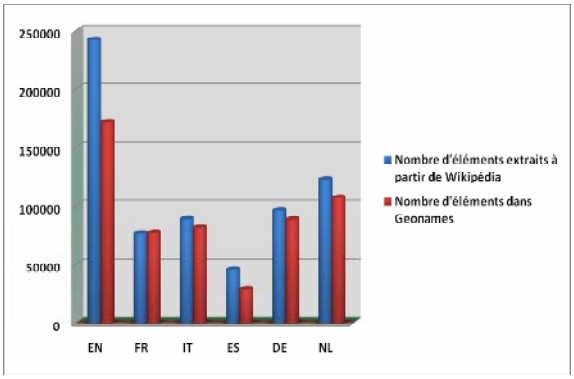

Figure 4.1- Nombre d'entités géographiques extraites de Wikipédia par notre approche et par Geonames.

4.2. Analyse de l'extraction d'articles géoréférencés

Dans Wikipédia, nous avons été confrontés à la polysémie de certains mots utilisés comme déclencheurs pour identifier les articles géoréférencés (ex : ***l*ongitud** *:* en espagnol désigne la longitude, mais aussi la longueur dans d'autres contextes) ; les faux positifs ont été éliminés car ils ne comprennent aucun des critères que nous avons définis. Parmi les résultats obtenus, on a observé la présence d'articles décrivant des lieux tels que ***Byrgius (crater)*** qui est un cratère de la lune. Les coordonnées de ce cratère sont des coordonnées sélénographiques renseignées dans un Infobox comportant ses caractéristiques.

4.3. Évaluation de la catégorisation

La qualité de la catégorisation a été évaluée semi-automatiquement. Nous avons choisi un échantillon de **1200** entités dans les six langues (200 par langue). Les résultats (tableau 4.2) montrent que la catégorisation réalisée est correcte dans plus de **95%** des cas, ce qui représente un taux de succès très satisfaisant pour une méthode complètement automatique. On observe une faible variabilité de la qualité de la catégorisation avec le changement de la langue (valeur minimale de 94% et maximale de 98%).

	EN	FR	DE	IT	ES	NL
Erreurs (/200)	12	4	8	12	10	6
Précision	94%	98%	96%	94%	95%	97%

Tableau 4.2- Résultats de l'évaluation de la catégorisation

4.4. Analyse des erreurs de catégorisation

Les erreurs sont causées principalement par des définitions complexes. Par exemple, le verbe ***to be*** est parfois suivi par une référence à la position géographique de l'objet et non par sa classe parent : **X est situé à l'est de Y et est un Z**. Dans ce cas, au lieu d'extraire Z, il est possible de trouver un élément du vocabulaire géographique dans Y qui sera extrait par notre algorithme. Nous avons aussi remarqué quelques erreurs de catégorisation en raison de particularités linguistiques des catégories en l'absence de certains termes. Par exemple, la langue allemande est une langue agglutinante, donc les catégories sont plus difficiles à extraire.

4.5. Évaluation de l'extraction d'entités non géoréférencées

Nous avons extrait un grand nombre de nouvelles entités décrites dans des articles de Wikipédia non géoréférencés. En nous basant uniquement sur le champ « catégorie » de l'Infobox, nous obtenons **12 117** (41%) nouvelles entités avec leur bonne catégorie. Nous avons ensuite analysé la première phrase des articles ainsi que la partie relative à la catégorisation pour trouver d'autres entités qui possèdent des informations liées à leur catégorie géographique. On a évalué cette méthode d'extraction en choisissant au hasard **200** entités. L'extraction est correcte à **92.5%**. Les erreurs proviennent surtout des articles de catégorisation qui ont des titres comportant le terme ***category*** (ex : ***category : cities in Ecuador)***. De plus, nous avons remarqué que quelques articles sont relatifs à des personnes. Leur existence s'explique par l'apparition du mot ***states*** dans la définition de l'entité ou dans la partie de catégorisation.

5. Conclusion et Perspectives

Le travail qu'on a présenté dans cet article a pour objectif de construire une base de connaissances géographiques multilingue et à large échelle. Nous avons réussi à ajouter plus de 700 000 entités à l'application existante de recherche d'images géolocalisées sur Internet ThemExplorer. Par ailleurs, et les évaluations montrent des résultats encourageants. En nous inspirant des travaux de Popescu et *al.* [6], nous avons extrait pour chaque entité : son nom, sa catégorie géographique, ses coordonnées et une mesure de pertinence et ce à partir de Wikipédia.

Une première perspective de ce travail est d'appliquer la méthode d'extraction d'articles non géoréférencés et de catégorisation pour les autres langues. Une fois les noms des entités et leurs catégories extraits, on essaiera de découvrir leurs coordonnées géographiques à partir de Panoramio et de calculer leur mesure de pertinence pour les ajouter à la base finale. Une deuxième perspective concerne

l'évaluation d'une catégorisation multilingue des entités. Les résultats obtenus pour une catégorisation monolingue montrent une précision de plus de **95%**, mais nous sommes convaincus qu'il est possible de les améliorer en utilisant plusieurs langues. Une troisième perspective consiste en l'application de ces méthodes sur d'autres sources d'information, comme Flickr qui contient un très grand nombre d'images géoréférencées.

Références

[1] S. Ahern, M. Naaman, R. Nair, J. Yang. 2007. WorldExplorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. *In Proc. of JCDL (Vancouver, Canada, June 2007).*

[2] S. Auer, C. Bizer, G. kobilarov, J. Lehmann, R. Cyganiak et Z. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. *In Proc.of ISWC 2007 (Busan, Korea, November 2007)*

[3] O. Buyukkoten, J. Cho, H. Garcia-Molina.1999.Exploiting Geographical Location Information of Web Pages. *In WebDB'99, (1999).*

[4] L. L. Hill, J. Frew, et Q. Zheng. 1999. "*Geographic Names: The Implementation of a Gazetteer in Georeferenced Digital Library*". In *CNRI D-Lib Magazine (January 1999).*

[5] J. Kazama et K. Torisawa .2007. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. *In Proc.of the 2007 Joint Conference on Empirical Methods in Natural Language Processing et Computational Natural Language Learning, p. 698-707, (Prague, June 2007).*

[6] A. Popescu, G. Grefenstette et P.A. Moëllic.2008. Gazetiki: Automatic Creation of a Geographical Gazetteer, *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries.*

[7] A. Popescu, P.A. Moëllic et I. Kanellos.2008.ThemExplorer: Finding and Browsing Geo-Referenced Images, *International Workshop on Content-Based Multimedia Indexing, 2008. CBMI 2008.*

[8] T. Rattenbury, N. Good et M. Naaman. 2007. Towards Automatic Extraction of Event and Place Semantics from Flickr Tags. *In Proc.of SIGIR 2007 (Amsterdam, The Netherlands, July 2002).*

[9] M. Sanderson, Y.Han.2007. Search Words and Geography. *In GIR'2007, November 9, 2007, Lisbon, Portugal.*

[10] C.Toriani, S.Battle et S.Cayzer. 2006. Sharing, Discovering and Browsing Geotagged Pictures on the Web. *3rd Italian Semantic Web Workshop on Semantic Web Applications and Perspectives.*