Du texte à la base de données géographiques

Nesrine Hassini, Khaoula Mahmoudi, Sami Faiz

Laboratoire de Télédétection et Systèmes d'Information
à Références Spatiales, ENIT BP 37 le Belvédère 1002 Tunis, TUNISIE
hassnes@yahoo.fr, kamahmoudi@yahoo.fr, sami.faiz@insat.rnu.tn

Résumé. Avec la prolifération des données géographiques, il y a un fort besoin de concevoir des outils automatiques pour l'exploitation des connaissances géographiques incarnées dans les documents textuels. C'est dans ce contexte, que nous proposons une approche permettant de générer une base de données géographiques (BDG) à partir de textes. Notre approche s'articule autour de deux grandes phases : la génération du schéma de la BDG et la détermination des données qui serviront au remplissage de cette base. L'implémentation de notre approche a donné naissance à un outil que nous avons baptisé GDB Generator et que nous avons intégré dans le SIG : OpenJUMP.

1 Introduction

L'explosion récente des technologies mobiles et des données géo-référencées a fait émerger un nouveau type de données, qualifiées de géographiques. De ce fait, une prolifération des systèmes d'informations géographiques (SIG) a vu le jour pour assurer une meilleure exploitation de ces informations. En fait, pour atteindre cet objective on doit disposer d'un ensemble d'informations complet et cohérent. Des moyens d'enrichissement ont été proposés pour aboutir à l'enrichissement des BDG à des coûts réduits. Dans ce contexte, on trouve à titre d'exemple MetaCarta (www.metacarta.com) et PIV (Lesbegueries et al., 2006) qui opèrent l'enrichissement en liant les documents textuels aux entités géographiques correspondantes. Ces travaux, ne proposent pas (ou peu) de moyens pour gérer les contenus textuels des documents. L'outil SDET (Mahmoudi et Faïz, 2010) est un autre moyen d'enrichissement qui cherche à exploiter le contenu textuel pour extraire l'essentiel sous forme d'un résumé. Le présent travail s'inscrit dans le contexte d'enrichissement des SIG tout en proposant une vue structurée des informations sous format de BDG. Concrètement, notre approche s'articule autour de deux grandes phases : une correspondance du texte en schéma conceptuel et une génération de la BDG à partir du schéma suite à son remplissage.

2 Notre approche

Le processus général de génération d'une structure sous forme de BDG à partir de texte vise un enrichissement de la BDG initialement incarnée dans le SIG. L'idée est de chercher un ensemble d'attributs pouvant compléter les données existantes. Notre approche se décompose

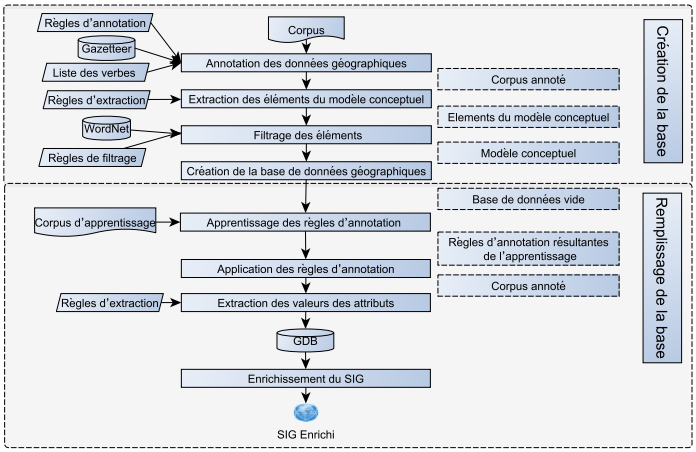

FIG. 1 – *Processus Général de notre approche d'enrichissement.*

en deux grandes phases (cf. figure 1). La première phase de notre processus d'enrichissement consiste à la génération de la structure de la BDG à partir de textes. Cette phase est déclenchée par une analyse morphosyntaxique résultant en un texte taggé. Le texte prétraité subit par la suite une annotation permettant de repérer les phases cibles. Inspirée du modèle proposé dans (Gaio et Nguyen, 2011), nous avons pu identifier différentes formes syntaxiques de phrases renfermant des données géographiques. En adoptant la même représentation symbolique, nous avons dégagé le modèle présenté par la figure 2.

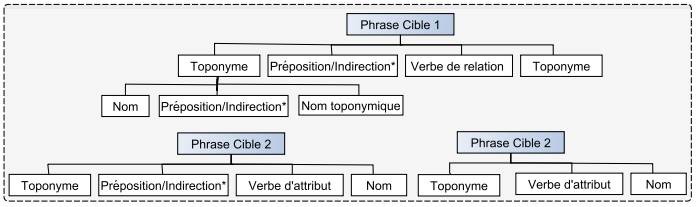

FIG. 2 – *Modèles des phrases à détecter dans le texte (avec * signifie que le composant correspondant peut être présent zéro ou plusieurs fois). Les autres composants doivent apparaître au moins une fois.*

Les phrases annotées seront injectées comme entrée de la phase d'extraction des concepts géographiques. Pour pouvoir dégager les éléments de base du modèle classes/relations nous nous sommes appuyés sur un ensemble de règles. Ces dernières sont dégagées empiriquement suite à l'étude de corpus textuels se rapportant aux données géographiques. Le tableau 1 décrit ces règles pour chaque élément du modèle.Une fois les données extraites, une étape de filtrage s'avère nécessaire pour la suppression des redondances et des noms propres au niveau des classes et des attributs. La tâche de suppression des redondances nécessite la détermination des synonymes de chaque élément du modèle. Ceci nous a amené à exploiter le thésaurus

WordNet (Miller et al., 1990) pour accomplir cette tâche. Pour ce faire, nous avons appliqué un ensemble de règles illustrées par le tableau 1.

Règles d'extraction des éléments du modèle conceptuel		
Identification des classes	*Identification des attributs*	*Identification des relations*
Si (Phrase Cible 1)ou(Phrase Cible 2) Alors Si "∃" N Alors Ajouter singulier (N) à LCC Sinon // ∄ N Ajouter NT à LCC	Si (Phrase Cible 2) Alors Pour toute[(C=N)ou(C=NT)] faire Ajouter Nom à la LA de C	Si (Phrase Cible 1) Alors Si[(∃(N1,N2))et(N1∈LCC) ET(N2∈LCC] Alors Pour tout (binôme (C1,C2)=(N1,N2)) faire Ajouter à LR de (C1) la relation (VR,C2) Et à LR de (C2) la relation (VR, C1)
Règles de filtrage des éléments du modèle conceptuel		
Suppression des noms propres	*Suppression des noms des classes*	
Si (Phrase Cible 3) Alors Si NT ∈ LCC Alors Pour toute C=NT faire Remplacer C par N	Pour toutes C1 et C2 faire Si C1=C2 ou C1 ∈ Synonymes (C2) Alors Copier LA de (C2) à LA(C1) Copier LR de (C2) à LR(C1) Supprimer C2 de la LCC	

TAB. 1 – *Règles d'extraction et de filtrage des éléments du modèle conceptuel (Avec LCC : liste classe candidate, CC : classe candidate (LA : liste attributs, LR : liste relations), R : relation (VR : verbe relation, CR : classe relation), NT : nom toponymique, N : nom de toponyme).*

La deuxième phase de notre approche consiste à fournir un corpus d'apprentissage annoté préalablement pour pouvoir dégager les règles d'annotation. Pour accomplir l'apprentissage nous avons adopté l'algorithme de SVM à marges inégales (Li et Shawe-Taylor, 2003). Par annotation nous entendons le repérage des données qui vont servir comme valeurs des attributs dont nous disposons. Les règles dégagées vont être appliquées au texte annoté à la première phase. Ceci résulte en un remplissage de la BDG préétablie. La base résultante et la BDG incarnée au SIG vont subir un appariement pour pouvoir ajouter tout attribut non présent dans la base initiale.

3 Implémentation

L'approche détaillée tout au long de ce papier a fait l'objet d'une implémentation qui a pu aboutir à l'outil GDB Generator. Etant donné que le but est d'enrichir les SIG, cet outil se présente comme un Plugin développé en Java et intégré dans le système OpenJump (http ://www. openjump. org). Si l'utilisateur a affaire à des informations non stockées dans la base initiale, il peut lancer notre outil. En backoffice l'outil s'appuie sur un corpus de textes relatifs à l'entité en question. En utilisant l'environnement d'édition visuel intégré à GATE (http ://www.gate.ac.uk), notre outil débute par l'annotation des éléments géographiques dans les textes. Ceci étant accompli en définissant des règles d'annotation en langage de grammaire JAPE (http ://gate.ac.uk/sale/tao/splitch8.htm).

Une fois extraits et filtrés, ces éléments forment le modèle conceptuel de la base à créer en adoptant le SGBD postgresql (www.postgresql.org). La détection des données qui serviront à remplir la base a été accomplie sous le système Gate en intégrant deux sessions de Machine-Learning pour appliquer l'apprentissage au corpus préalablement annoté et l'application des règles d'annotation générées à notre corpus. La base créée servira à enrichir la base existante par des nouveaux attributs selon les besoins de l'utilisateur(cf. figure 3).

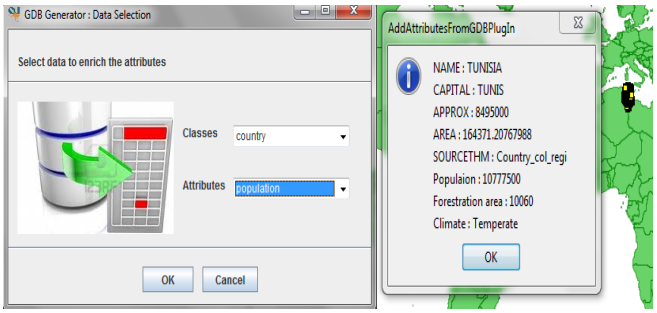

FIG. 3 – *La liste des attributs enrichie.*

4 Conclusion

Nous avons présenté dans cet article une approche pour automatiser la création et le remplissage d'une BDG pour enrichir les SIG. La mise en oeuvre de notre approche a donné naissance à l'outil baptisé GDB Generator. Ce dernier a été intégré au SIG OpenJUMP. Comme perspectives, nous proposons d'enrichir nos règles d'annotation en prévoyant plus de formes syntaxiques. La prise en compte de la composante spatio-temporelle est prévue aussi comme extension de notre approche.

Références

Gaio, M. et V. Nguyen (2011). Utilisation de la relation " verbe-preposition-toponyme" pour un inventaire lexical automatique. *30e colloque international sur le lexique et la grammaire, Nicosie : Chypre*, 22–33.

Lesbegueries, J., M. Gaio, et P. Loustau (2006). Geographical information access for non-structured data. *In proceedings of the 2006 ACM symposium on Applied computing*, 83–89.

Li, Y. et J. Shawe-Taylor (2003). The svm with uneven margins and chinese document catego-rization. *17th Pacific Asia Conference on Language, Information and Computation (PACLIC17), Singapore*.

Mahmoudi, K. et S. Faïz (2010). Towards geographic databases enrichment. *Lecture Notes in Computer Science (LNCS), Vol. 6177/2010, Springer-Verlag, Berlin*, 216–223.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, et K. J. Miller (1990). Introduction to wordnet: An on-line lexical database*. *International journal of lexicography 3*(4), 235–244.

Summary

In this paper we present an approach to generate a geographic database from texts. The implementation of the proposed approach has given born to the GDB Generator tool. The latter was integrated as a new facility to the OpenJump GIS.