Un processus d'acquisition d'information pour les besoins de l'enrichissement des BDG

Khaoula Mahmoudi*
Sami Faïz ** ***
* Laboratoire URISA -Unité de Recherche en Imagerie Satellitaire et ses Applications
Ecole Supérieur des communications de Tunis (SUPCOM)
khaoula.mahmoudi@insat.rnu.tn
** Institut National des Sciences Appliquées et de Technologie (INSAT)
*** Laboratoire de Télédétection et Systèmes d'Informations à Références Spatiales (LTSIRS)
sami.faiz@insat.rnu.tn

Résumé. Les données constituent l'élément central d'un Système d'Information Géographiques (SIG) et leur coût est souvent élevé en raison de l'investissement substantiel qui permet leur production. Cependant, ces données sont souvent restreintes à un service ou pour une catégorie d'utilisateurs. Ce qui a fait ressortir la nécessité de proposer des moyens d'enrichissement en informations pertinentes pour un nombre plus important d'utilisateurs. Nous présentons dans ce papier notre approche d'enrichissement de données qui se déroule selon trois étapes : une identification de segments et de thèmes associés, une délégation et enfin, un filtrage textuel. Un processus de raffinement est également offert. Notre approche globale a été intégrée à un SIG. Son évaluation a été accomplie montrant ainsi sa performance.

1 Introduction

Les données dans un SIG (Faïz, 1999), sont souvent recueillies pour les besoins propres d'une institution, voire d'un service. Face à cette réalité, il devient judicieux de déployer de nouvelles sources pour répondre aux besoins d'un nombre plus important d'utilisateurs. Ceci est qualifié d'enrichissement de bases de données géographiques (BDG). C'est dans ce contexte que s'inscrit notre approche (Mahmoudi et Faïz, 2006_a, Mahmoudi et Faïz, 2006_b, Faïz et Mahmoudi, 2005). Cette dernière utilise la technique de résumé de documents multiples (Barzilay et McKeown, 2005) permettant d'extraire l'information pertinente sous une forme abrégée. Pour assurer l'extraction dans des temps raisonnables et conformément au paradigme multi-agents (Ferber, 1999), nous adoptons trois classes d'agents: agent *interface*, agent *géographique* et agent *tâche*. L'interaction entre les agents est achevée par envoi de messages. L'enrichissement est réalisé en trois phases : une identification de segments et de thèmes, une délégation et enfin, un filtrage textuel. S'ajoute à ces étapes de base, une approche, exercée à la demande, pour un raffinement du processus.

La section 2 présente, certains travaux d'enrichissement des BDG dans les SIG ainsi que notre approche pour cet enrichissement. La section 3 est dédiée à la mise en œuvre et l'évaluation de notre système.

2 l'enrichissement des BDG

2.1 Travaux relatifs à l'enrichissement des BDG

L'*enrichissement* des SIG permet l'acquisition d'informations supplémentaires indispensables pour une bonne prise de décision. On parle d'enrichissement spatial et d'enrichissement sémantique. Concernant l'aspect spatial et dans le cadre du processus de généralisation (Plazanet, 1996), par exemple, l'enrichissement procure les BDG d'informations en terme de structure des formes, des connaissances se rapportant à l'ordre des opérations et aux algorithmes appropriés. Un autre flot de travaux se rapporte à l'aspect sémantique (dit aussi factuel ou descriptif) des BDG. Dans cette catégorie, nous pouvons citer Metacarta (MetaCarta, 2005) et GeoNode (Hyland et al., 1999).

Le projet Metacarta accompli l'*enrichissement* par son produit *Geographic Text Search* (GTS). GTS permet de relier des documents textuels à des entités géographiques localisées sur la carte venant enrichir les données de la BDG. MetaCarta GTS est offert comme une extension au système d'information géographique *ArcGIS*.

GeoNode (*Geographic News On Demand Environment*) exploite la technique d'extraction d'informations pour aboutir à l'enrichissement, via le système *Alembic*. GeoNode permet d'extraire les entités nommées et les évènements associés qui seront visualisées d'une manière geospatiale. Le SIG *ArcView* supporte GeoNode.

Ce qui marque notre approche est qu'à l'opposition des travaux existants, elle va au-delà de la simple localisation de l'information, pour permettre la synthèse de ces informations. De même, à l'opposé des travaux existants dont la source de données aurait été déjà pré-établies sous forme de librairies (données traitées et classées) par exemple, notre approche préconise la génération en temps réels (information toujours récente et mise à jour) des documents requis peu importe le type d'information réclamée (nous ne pouvons jamais prédire tous les besoins des utilisateurs). Egalement, et outre l'aspect sémantique, nous exploitons la composante spatiale des BDG afin d'améliorer les résultats de l'enrichissement.

2.2 Notre approche pour l'enrichissement des BDG

Notre processus d'enrichissement émane d'un besoin informationnel réclamé par les utilisateurs des SIG. Un utilisateur soumettant une requête à la BDG, mais, se trouvant en situation d'insatisfaction, peut lancer notre processus d'enrichissement. En fait, les documents relatifs à une ou plusieurs entités géographiques sont confiés à l'agent *interface* qui va les distribuer entre les agents *tâche*. Chaque agent *tâche* procède à la segmentation de son document en parties thématiquement cohérentes. Notre proposition est une adaptation de l'algorithme de TextTiling (Hearst, 1997). Le texte est initialement découpé en blocs de taille fixe. En fait, il arrive que lors du découpage, deux phrases hétérogènes soient classées sous le même bloc, ainsi, lors du calcul de la similarité, nous obtenons forcément des valeurs élevées faisant preuve de l'homogénéité des deux blocs. Pour résoudre le problème de localisation non précise des frontières, nous intervenons lors du découpage initial en appliquant la procédure C99 (Choi, 2000), reconnue pour être la plus appropriée pour les textes courts. Cette procédure remplace la similarité entre phrases par le rang dans un contexte local. Un score de cohésion (métrique du cosinus) est attribué aux blocs adjacents. Les similarités sont présentées sous forme de graphe et aplanie. Les vallées observées sur le

graphe correspondent à des scores faibles indiquant une potentielle rupture thématique.

Les segments précédemment détectés, subissent une annotation (étiquetage) par les agents *tâche* moyennant l'attribution de thème. Le mot le plus fréquent est assigné comme le thème du texte en cas de distribution hétérogène des fréquences des mots. Pour une distribution homogène, nous partons de l'idée que le texte est un ensemble de termes contribuant à développer un thème donné. Pour déterminer ce thème, nous exploitations les relations sémantiques du thesaurus WordNet (Miller, 1990). Pour expliciter notre propos nous considérons ce texte: *"The state run Tunisian Radio and Television Establishment (ERTT) operates two national TV channels and several radio networks. Until November 2003 the state had a monopoly on radio broadcasting."* Suite à l'investigation du WordNet, nous présentons les relations les plus intéressantes: *Hypernymy (television) = broadcasting, Hypernymy (TV) = broadcasting, Hypernymy (radio) = broadcasting, Synonym (TV) = television, Topic-domain (network) = broadcasting*...Nous recalculons la fréquence des mots qui est le nombre de liens sémantiques. Ainsi, nous pouvons déduire que *broadcasting* est le mot le plus tranchant considéré comme mot central relevant le thème général du texte. A l'issue de cette phase, les thèmes sont expédiés à l'agent *interface*.

Les thèmes décelés sont assignés aux agents *tâche* par une délégation entreprise par l'agent *interface* (cas d'une seule entité géographique) ou un des agents *géographique* (en cas d'une multitude d'entités). Cette délégation consiste à distribuer les différents thèmes entre les agents *tâche*, et ce tenant compte d'un coût qui reflète le coût de communication (en terme de messages échangés) et la charge du travail (en terme de volume textuel à prendre en charge). Dorénavant, chaque délégué détient à sa disposition des *documents générés* qui sont l'ensemble des segments textuels relatifs aux thèmes qui lui sont affecté.

Par la suite, chaque délégué entame une phase de filtrage visant la condensation dans un format de résumés de ses *documents générés* pour ne retenir que l'essentiel de l'information (Mann et Thompson, 1988). Ainsi, pour une phrase formée de deux unités lexicales et reliée par une relation d'exemplification (par la présence de *cues ; for instance*, *for example*), nous ne retenons que la partie (appelée nucléus) qui ne contient pas ce *cue*, obligatoire pour la compréhension du texte. L'autre contenant le *cue* est subsidiaire (appelée satellite). En cas d'absence de *cues*, nous calculons la similarité entre les unités et nous ne retenons que les unités les plus dissimilaires.

A l'issue des étapes susmentionnées, il se trouve que parfois l'utilisateur reste insatisfait par les résultats de l'enrichissement. Ceci peut provenir d'une des raisons suivantes : (i) la recherche de l'information n'a pas ciblé les documents les plus pertinents dans le web, (ii) l'utilisateur est submergé par une masse colossale de documents qui ne sont pas tous pertinents par rapport à l'entité en question, (iii) L'information visée n'est pas disponible dans les documents du corpus. L'idée est d'examiner le voisinage de l'entité sujette de la recherche pour dégager les entités avec les quelles, elle détient des relations spatiales (Mahmoudi et Faïz, 2006_b). Notre thèse est argumentée par la première loi de la géographie décrivant la nature des systèmes géographiques dans lesquels *"everything is related to everything else, but near things are more related than distant things"* (Bin et Itzhak, 2006). Dans ce cas, l'utilisateur peut lancer un processus de raffinement que nous avons intégré au processus de base. Il exploite les relations spatiales (adjacence, superpositions…) afin de mieux décrire les entités sujettes de la recherche et ainsi mieux cibler les informations pertinentes.

3 Mise en œuvre et évaluation de notre approche

Notre approche pour l'enrichissement des BDG a été mise en œuvre en utilisant le

langage Java ce qui a permis une implantation d'un système distribué transposant les concepts du système multi-agents. Cette mise en œuvre a donné naissance à un outil que nous avons baptisé *Semantic Data Enrichment Tool* ou SDET (Mahmoudi et Faiz, 2007, Mahmoudi et Faiz, 2006_c). SDET offre un ensemble de fonctionnalités visant l'enrichissement des données initialement stockées dans la BDG allant de la création de *documents générés* relatifs aux thèmes décelés du corpus textuel jusqu'à la condensation des idées essentielles incarnées dedans sous forme de résumés. En plus, SDET offre le moyen d'exploiter les relations spatiales que les entités géographiques y maintiennent pour un éventuel raffinement de la recherche. Ce système à été intégré au SIG OpenSource : OpenJump.

Pour l'évaluation des résumés générés, nous avons utilisé l'outil ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) adopté par la campagne d'évaluation de systèmes de résumé DUC (Document understanding conference). ROUGE calcule le taux de chevauchement entre les résumés produits par un système de résumé et les résumés produits par des humains (appelé modèle ou référence) en utilisant des n-grammes. Formellement, ce score s'exprime par la formule suivante :

$$\text{ROUGE-N} = \frac{\sum_{s \in \{références\}} \sum_{n-gram \in s} count_{match}(n-gram)}{\sum_{s \in \{références\}} \sum_{n-gram \in s} count(n-gram)}$$

ROUGE-N, dénote un score basé sur le nombre de n-grammes (1<=n<= 4) communs entre les résumés automatiques et les résumés modèles. $Count_{match}$(n-gram) est le nombre de n-grammes partagés entre le résumé produit par le système et le résumé modèle et Count (n-gram) est le nombre de n-grammes dans le résumé modèle.

Rouge-2, Rouge-SU-4 et BE se sont imposés. ROUGE-2 calcule le nombre de paires de mots successifs (nombre de *bigrammes*) en commun entre les résumés automatiques et les résumés modèles. Rouge-SU-4 correspond au rappel en *"bigrammes à trous"* (skip units) de taille maximum 4. Pour, BE (*Basic Elements*) elle a est définie comme unité sémantique minimale qui consiste en deux éléments et une relation (entités-relation) entre ces éléments. Par exemple, à partir de la phrase *"two Libyans were indicted for the Lockerbie bombing in 1991"* nous détectons la BE suivante : indicted|libyans|obj, tel que *obj* est la relation objet entre *indicted* et *libyans*. La similarité entre le résumé référence et le résumé système en terme de BEs, permet de juger que le résumé système est un bon résumé.

Comme pour la conférence DUC, nous avons défini deux types de références basses, appelées aussi *baseline*. Ces résumés sont créés automatiquement en fonction des règles suivantes : une *baseline* sélectionne les 150 premiers mots du document le plus récent, une autre *baseline* sélectionne la première phrase dans les (1, 2,…, n) documents de l'ensemble à résumer trié par ordre chronologique jusqu'à atteindre 150 mots. Ces règles découlent du fait que d'après des études de différents documents textuels, l'importance du commencement du document par rapport à sa fin a été approuvée. Ainsi, les premières lignes du texte couvrent en général une partie importante de l'essentiel du texte. Pour notre évaluation, nous avons généré des résumés automatiques avec les systèmes suivants : le système MEAD (Radev et al., 2003) et les méthodes *baselines* (*baseline1 et baseline2*). Nous avons utilisé l'outil ROUGE pour comparer les résumés obtenus avec notre système SDET, MEAD, *baseline1 et baseline2* avec les résumés humains comme le montre le Tableau 1.

Le score le plus élevé étant le meilleur, il indique le système le plus performant. SDET est classé au premier rang avec les meilleures notes d'évaluation. Les deux *baselines* donnent les résultats les plus faibles. Notons que si l'hypothèse de base qui stipule que les premières

lignes du texte couvrent en général une partie importante de l'essentiel, cela reste insuffisant dans le cas de résumé multi-documents qui nécessite une prise en charge particulière des similarités et des différences informationnelles à travers tous les documents.

Système	Rouge-2	Rouge-SU4	BE
SDET	0,47908	0,30106	0,23048
MEAD	0,35194	0,20117	0,10580
Baseline1	0,20200	0,12132	0,09851
Baseline2	0,17240	0,08104	0,05141

TAB. 1 – *Les résultats de notre évaluation avec le package ROUGE/BE.*

4 Conclusion

La disponibilité des informations pertinentes répondants aux besoins de la quasi totalité des utilisateurs SIG est devenue un objectif visé par les concepteurs de tels systèmes. En effet, les informations manipulées sont fournies par des BDG qui deviennent très vite insuffisantes et ne répondent plus à l'ensemble des besoins des utilisateurs. Face à cette situation, nous avons proposé et mis en œuvre un processus d'enrichissement pour les BDG. Il s'agit d'une identification de segments et de leurs thèmes, une délégation puis un filtrage. Une phase supplémentaire peut être invoquée en cas de résultats insatisfaisants, il s'agit d'un raffinement visant un ciblage plus pertinent des informations requises. La mise en œuvre du processus d'enrichissement a donné lieu à un outil que nous avons baptisé SDET. Les résultats générés par notre système ont été évalués en utilisant le package ROUGE. Les résultats de cette évaluation sont très convaincants faisant preuve de l'intérêt de notre démarche pour l'enrichissement des BDG.

Références

Barzilay, R., K. McKeown (2005). *Sentence Fusion for Multidocument News Summarization*. Computational Linguistics, 31(3), 297-328.

Bin, J., O. Itzhak (2006). *Spatial Topology and its Structural Analysis based on the Concept of Simplicial Complex*. 9th AGILE Conference on Geographic Information Science, Visegrád, Hungary, 204–212.

Choi, F. (2000). *Advances in domain independent linear text segmentation*. In : NAACL'00.

Faïz, S. (1999). *Systèmes d'Informations Géographiques : Information Qualité et Data mining*. Editions C.L.E, 362 p.

Faïz, S., K. Mahmoudi (2005). *Semantic Enrichement of Geohraphical Databases*. Encyclopedia of database tehnologies and applications. Editors : Rivero L., Doorn J. & Ferraggine V. Editions Idea Group, Etats-Unis, septembre, 587-592.

Ferber, J. (1999). *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. 1st edn. Addison-Wesley Professional.

Hearst, M.A. (1997). *TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages*. Computational linguistics, vol. 23, No. 1, 33-46.

Hyland, R., C. Clifton, R. Holland (1999). *GeoNODE: Visualizing News in Geospatial Environments*. In Proceedings of the Federal Data Mining Symposium and Exhibition '99, AFCEA, Washington D.C.

Lin, C-Y. (2004). *ROUGE*: *A Package for Automatic Evaluation of Summaries*. Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL, Barcelona, Spain, 74-81. http://www.isi.edu/~cyl/ROUGE/

Mahmoudi, K., S. Faïz (2006_a). Une approche distribuée pour l'extraction de connaissances : Application à l'enrichissement de l'aspect factuel des BDG. *Revue des Nouvelles Technologies de l'Information, Editions Cépaduès, Janvier, 107-118.*

Mahmoudi, K., S. Faïz (2006_b). *L'apport de l'information spatiale pour l'enrichissement des bases de données.* INFORSID'2006, Hammamet, Tunisie, juin, 323-338.

Mahmoudi, K., S. Faïz (2007). *L'outil SDET pour le complètement des données descriptives liées aux bases de données géographiques*. Actes 7èmes Journées d'Extraction et Gestion des Connaissances (EGC'07), Session Demo, Namur Belgique, Janvier, 179-180.

Mahmoudi, K., S. Faïz (2006_c). *SDET: A Semantic Data Enrichment Tool Application to Geographical Databases.* International Conference On Signal-Image Technology & Internet–Based Systems (SITIS'2006), IEEE, ACM, Hammamet, Tunisie, Décembre, 88-97.

Mann, W. C., S.A. Thompson (1988). Rhetorical structure theory: Toward a functional theory of text organization. *An Interdisciplinary Journal for the Study of Text 8(2):243-281.*

Miller, G. (1990). Wordnet: An on-line lexical database. *International Journal of Lexicography (special issue), 3(4):235-312.*

MetaCarta (2005): GTS versus MetaCarta GeoTagger. MetaCarta, Inc.

Plazanet, C. (1996). *Enrichissement des bases de données géographiques: analyse de la géométrie des objets linéaires pour la généralisation cartographique (application au routes)*. PhD thesis, Université de Marne-la-Vallée.

Radev D., J. Qi H. Otterbacher, et D. Tam (2003). *Mead reducs: Michigan at DUC 2003*. In DUC03, Edmonton, Alberta, Canada: ACL, 160–167.

Summary

Data is essential to the Geographic Information Systems (GIS). in many situations we have to add complementary data to the geographic database (GDB) to support decision makers. To this end we have proposed along this paper a data enrichment process to add supplementary data to the already stored data. This process is accomplished through three steps: a segmentation and theme identification, delegation and text filtering. Besides, a refinement of this overall process can be eventually performed. The implementation of our approach was reported along with the evaluation of the generated results.