Génération d'un extrait textuel à partir de bases de données

Ghada Landoulsi, Khaoula Mahmoudi, Sami Faiz

Laboratoire de Télédétection et Système d'Information à Référence Spatiale (LTSIRS)
École National d'Ingénieurs de Tunis, B. P, 37, 1002 Tunis-Belvédère, Tunisie.
ghadalandoulsi@yahoo.fr, khaoula.mahmoudimapa@laposte.net,
sami.faiz@insat.rnu.tn

Résumé. Dans ce papier, nous présentons une approche dédiée à la transformation d'une base de données en un extrait textuel. L'idée sous-jacente à notre proposition est d'apporter plus de sémantique aux données de la base. Cet objectif est atteint moyennant l'utilisation des ontologies comme ressources sémantiques. Notre approche prend comme input un ensemble de bases de données et associe à chacune une ontologie. Une ontologie globale est générée, à partir de laquelle des règles d'association sont proposées pour mieux expliciter sa sémantique. Enfin, la génération d'un extrait textuel prend lieu.

1 Introduction

De nos jours, les volumes de données textuelles gérées et échangées ne cessent d'augmenter. Ceci est dû au fait qu'un texte est plus riche en sémantique que tout autre support informationnel. Toutefois, pour les systèmes de gestion de bases de données, les informations gérées se présentent principalement sous format structuré. Bien que cette structuration offre un confort au niveau de l'exploitation de données, elle présente une limite quant à la représentation des connaissances d'une manière explicite. Généralement, les données textuelles sont plus significatives que les schémas conceptuels des bases de données, dans la mesure où elles permettent une description sémantique des relations entre les éléments d'un domaine. Pour profiter de l'abondance des données structurées et en même temps expliciter les sémantiques qui y sont incarnées, nous proposons une approche visant la génération des textes à partir de bases de données. Cette approche s'articule autour de trois grandes phases : (i) une phase de prétraitement de données pour la génération d'ontologies à partir des bases de données, (ii) une phase de génération de règles d'association pour mieux expliciter la sémantique de l'ontologie globale et (iii) une phase de génération d'un extrait textuel. Dans le reste de cet article la section 2 détaille les trois étapes de base de notre approche. La section 3 est consacrée à la mise en œuvre de notre approche.

2 Approche proposée

Notre approche pour la génération d'un extrait textuel est basée essentiellement sur les ontologies comme ressources sémantiques. En fait, ce choix est justifié par la capacité des

ontologies d'offrir une modélisation sémantique des concepts et des relations associées. Étant donné que nous visons la production d'un extrait textuel, une phase de génération de règles d'association s'avère indispensable pour guider le passage de l'ontologie au texte. Concrètement, cette approche comporte principalement trois grandes phases (voir figure 1).

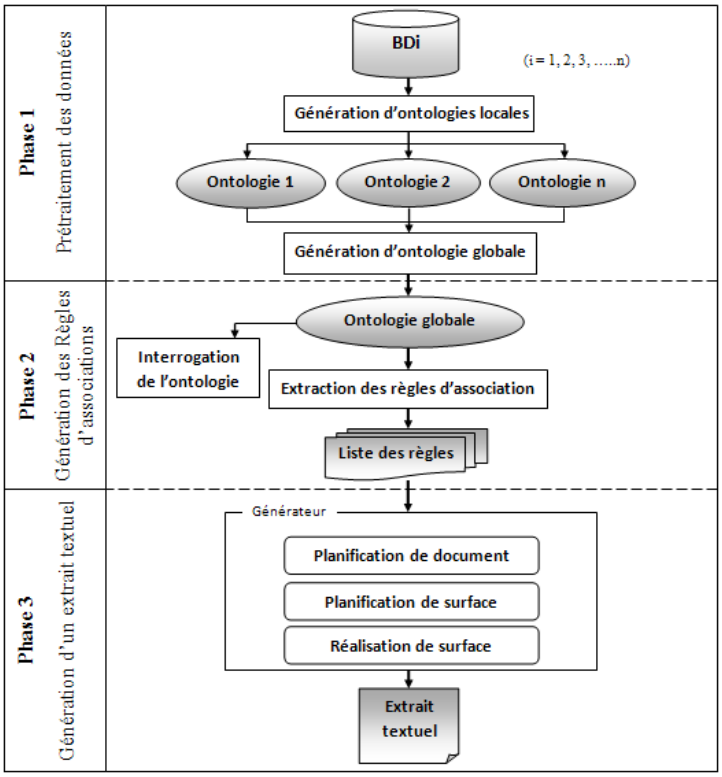

FIG. 1 – *Architecture générale de l'approche proposée*

2.1 Phase de prétraitement des données

Cette phase vise l'association d'une ontologie globale aux bases de données alimentées comme inputs à notre approche. Ceci étant accompli en trois étapes. La première dite de conceptualisation, assure la transformation de chaque base de données fournie en entrée en une ontologie qu'on qualifiera de locale. Une phase d'ontologisation, pour aboutir une ontologie globale est accomplie par la suite pour faire face à l'hétérogénéité des ontologies locales générées. Il s'agit de procéder à un mapping pour découvrir la correspondance sémantique entre les éléments dans différentes ontologies locales. Pour ce faire, nous avons utilisé l'algorithme FOAM (Framework for Ontology Alignment and Mapping) (Ehrig, 2007). L'idée sous-jacente à ce dernier est de procéder à un calcul de similarité de certaines paires d'entités (E_1, E_2) des ontologies à intégrer. Une fois, l'ontologie globale créée, une étape d'opérationnalisation prend lieu. L'objectif étant d'évaluer la consistance de l'ontologie.

2.2 Phase de génération de règles d'association

L'ontologie créée lors de la phase précédente est riche sémantiquement par rapport aux bases sources. Néanmoins, sa sémantique n'est pas exprimée d'une manière explicite. En fait, le passage à une vue textuelle est tributaire d'une interprétation claire de l'ontologie en question. C'est l'objectif de la phase en cours. Une fois l'ontologie globale sollicitée moyennant une requête utilisateur, les données sont injectées comme inputs de la phase de génération de règles. Lors de cette phase nous avons adopté l'algorithme Apriori (Agrawal et Srikant, 1994) pour déterminer les règles. Une fois les règles produites, ils seront considérés comme inputs pour déclencher la phase de génération de l'extrait textuel.

2.3 Phase de génération d'un extrait textuel

Cette phase est assurée en suivant trois étapes : une planification de document, une planification de surface et enfin une réalisation de surface. Quant à la planification de document elle vise la détermination de la forme finale de l'extrait textuel à générer. Plus explicitement, l'ensemble de règles subit une sélection basée sur un calcul de support et de confiance pour maintenir un sous ensemble de règles jugées pertinentes. Une fois le contenu déterminé, il doit être structuré. La structuration du contenu consiste à analyser les règles d'association sélectionnées afin d'extraire les éléments d'information qui les constituent. Une fois les règles analysées, une étape d'organisation est nécessaire. Pour assurer cette structuration, nous proposons l'utilisation des schémas de McKeown (1985). Ces derniers présentent un ensemble de spécifications sur la méthode d'organisation des éléments d'information. Une fois le contenu de notre extrait textuel est déterminé et structuré, il convient de convertir les éléments constituant les règles d'association en termes lexicaux grâce à deux fonctions qui sont la lexicalisation et l'agrégation. La dernière étape de cette phase est consacrée pour la réalisation de surface. Elle consiste en fait à traduire la représentation conceptuelle en extrait textuel compréhensible.

3 Implémentation

Pour la mise en œuvre de notre approche un ensemble de bases de données géographiques du domaine de l'agriculture a été fourni comme input à notre système. Ces bases de données ont été téléchargées à partir d'un système mondial d'information sur l'eau et l'agriculture de la FAO (Food and Agriculture Organization) AQUASTAT [1]. Chacune de ces bases doit être convertie en une ontologie locale, via l'activation du plugin DataMaster (Nyulas et al., 2007) dédié à la conceptualisation. Pour ce faire, nous avons opté pour l'éditeur d'ontologies open-source Protégé-OWL [2]. L'intégration des ontologies locales est réalisée grâce au plugin Prompt (Noy et Musen, 2001) en appliquant l'algorithme FOAM. Pour l'extraction des règles d'association, l'interrogation de notre ontologie est indispensable. Le résultat de la requête est enregistré sous format CSV. Ce fichier est exploité pour la génération des règles d'association. Ces dernières sont exploitées moyennant l'utilisation du langage prolog [3]. Notre objectif étant de générer un extrait textuel (voir figure 2).

1. http ://www.fao.org.
2. http ://protege.stanford.edu
3. http ://www.swi-prolog.org

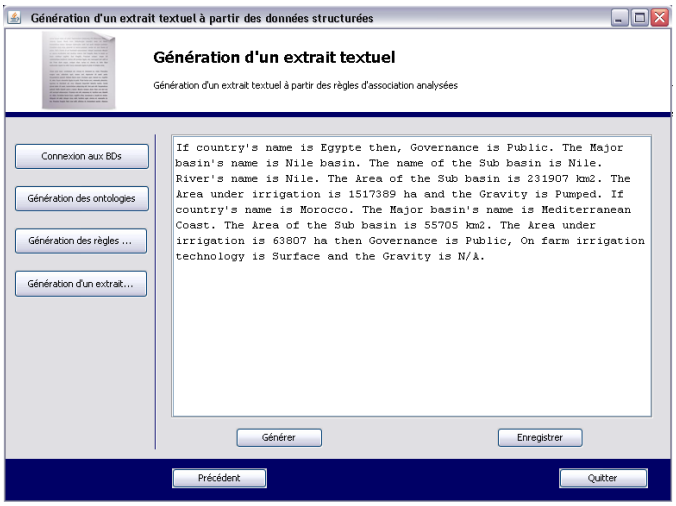

FIG. 2 – *Générateur d'un extrait textuel*

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, Santiago de Chile, Chile*, pp. 487–499.

Ehrig, M. (2007). *Ontology Alignment: Bridging the Semantic Gap*, Volume 4 of *Semantic Web and Beyond*. New York: Springer.

McKeown, K. (1985). *Text Generation : Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge, England.

Noy, N. et M. Musen (2001). Anchor-prompt: Using non-local context for semantic matching. In *Proceedings of the Workshop on Ontologies and Information Sharing at the 17th International Joint Conference on Artificial Intelligence (IJCAI), Seattle USA*, pp. 63–70.

Nyulas, C., M. O'Connor, et S. Tu (2007). Datamaster a plug-in for importing schemas and data from relational databases into protégé. In *Proceedings of the 10th International Protégé Conference, Budapest, Hungary, July 15-18*.

Summary

The backbone of an information system is a database storing the data to manage. The main feature of a database is the structure it offers to make handling data a straightforward task. Although, the abstraction adopted while setting up a database makes it poor semantically. In this paper we propose an approach to allow transforming a database to a textual view to exhibit its incarnated semantics. To achieve this objective we exploit the ontologies. The idea underlying this choice is that an ontology allows highlighting the semantic relationships among entities.