DATA ENGINEERING : MODELING AND INTEGRATION ISSUES

Par Madame Ana Carolina SALGADO Discipline : INFORMATIQUE

Ce rapport présente mes principaux résultats en trois axes de recherche depuis 1989 : les Bases de Données Géographiques, l'Intégration de Données et la prise en compte de la Sémantique dans les Systèmes Pair-a-pair (P2P). Une base de données géographique est dédiée à la représentation, au stockage et à la récupération d'informations référencées dans l'espace. Les techniques traditionnelles de modélisation n'étaient pas adéquates pour le traitement de ces types de données. La difficulté vient du fait que la plupart des données sont validées en termes de leur localisation dans l'espace, du temps et de leur disponibilité. Dans ce contexte, notre contribution a été la proposition d'un modèle de données géographiques orienté-objet, MGeo+, et son langage de requête, LinGeo. Nous avons aussi travaillé sur l'analyse des méthodes d'accès spatiales et sur la proposition d'un langage de requêtes visuel et son interface utilisateur. Les systèmes d'intégration de données sont des outils qui offrent un accès uniforme à des sources de données distribuées et hétérogènes. Cela est accompli en identifiant les hétérogénéités et en fournissant une vue unifiée sur les diverses sources. Les utilisateurs envoient leurs requêtes sur cette vue intégrée sans perdre du temps à naviguer sur le Web. Nous travaillons sur la spécification et l'implémentation d'un système d'intégration de données et, en particulier, sur les aspects d'évolution du

schéma de médiation et de la qualité des schémas. Les schémas et les instances des sources de données hétérogènes, dynamiques et distribuées contiennent rarement des descriptions sémantiques explicites qui puissent être utilisées pour dériver le sens des éléments du schéma (entité, attributs et associations). L'information sémantique implicite doit être extraite pour clarifier la signification des éléments du schéma. Pour permettre cela, une ontologie du domaine fournira les informations des associations sémantiques entre les termes du vocabulaire partagé par les sources. Cependant, l'information sémantique a un rapport avec la compréhension des gens et est une tâche dépendante du contexte et qui nécessite une connaissance spécifique du domaine. Le concept de contexte peut être employé pour améliorer la prise de décision afin de résoudre l'hétérogénéité sémantique des processus d'intégration de données une fois qu'il aide à la compréhension sémantique du schéma des sources et de leurs contenus. Nous présentons notre proposition d'un modèle de contextes, d'un gestionnaire de contextes indépendant du domaine, d'une ontologie d'informations contextuelles pour l'intégration de données et d'une approche pour la prise en compte des aspects sémantiques dans les systèmes pair-a-pair (P2P).

Abstract : This report includes the main results in three research areas we have been working on since 1989: Geographical Databases, Data Integration and Semantic Issues in PDMS (Peer Data Management Systems). A Geographical Database is a collection of inter-related and geo-referenced data. By definition, it is a database directed to the representation, storage and access to the information, which is spatially referenced. Traditional techniques of data modeling were not adequate for the treatment of geographical data. The difficulty consists of the fact that most of these data are validated in terms of its spatial localization, time, and the reliability of the collection. In this context, our contribution was the proposal of an object-oriented geographic data model MGeo+ and its query language LinGeo. We also have worked on spatial access methods' analysis and on the proposal of a visual query language for geographical data along with its user interface. The data integration systems are tools that offer a uniform access to distributed and heterogeneous Web data sources. This is done by resolving the heterogeneities and giving to the disparate sources an uniform view. Users submit queries over the integrated view without having to spend a lot of time in searching and browsing the Web. We have been working on the specification and implementation of a data integration system mainly interested in the evolution of the mediation schema, query reformulation and quality issues. Schemas and instances drawn from heterogeneous, dynamic and distributed data sources rarely contain explicit semantic descriptions which could be used to derive the meaning or purpose of schema elements (e.g. entity, attribute and relationship). Implicit semantic information needs to be extracted in order to

clarify the meaning of the schema elements. To achieve this, an ontology of a given knowledge domain will provide the information regarding semantic relations among the vocabulary terms shared by the data sources. Semantic interpretation, however, regards people's understanding and it is a context-dependent task which requires a specific understanding of the shared domain knowledge. Context may be employed as a way to improve decision-making over heterogeneity reconciliation in data integration processes since it helps to understand the data schema semantics as well as the data content semantics. We present our proposal to a context-oriented model and a domain-independent context manager, a contextual ontology to data integration and a semantic-based approach to peers' organization in a PDMS.

INFORMATIONS COMPLÉMENTAIRES

Aris M. OUKSEL,Professeur des Universités, à l'Université de l'Illinois, Chicago, Etats-Unis - Rapporteur**Stefano SPACCAPIETRA**, Professeur, à l'Ecole Polytechnique Fédérale de Lausanne, Suisse - Rapporteur**Patrick VALDURIEZ,**Directeur de Recherche, à l'Institut National de Recherche en Informatique et Automatique, Nantes - Rapporteur**Zohra BELLAHSENE,**Professeur des Universités, à l'Université de Montpellier - Examinateur **Patrick BREZILLON,**Directeur de Recherche, au Centre National de la Recherche Scientifique, Université de Paris 6 - Examinateur **Mokrane BOUZEGHOUB,**Professeur des Universités à l'Université de Versailles Saint Quentin en Yvelines - Tuteur