Enrichissement de la sémantique d'une ontologie de domaine par utilisation de règles d'association comme contraintes d'attributs

Oumar Sy

Section d'informatique

U.F.R de Sciences Appliquées et de Technologie
Université Gaston Berger de Saint-Louis
B.P. 234 Saint-Louis
SENEGAL

oumar.sy@ugb.edu.sn

RÉSUMÉ: L'ontologie est, au sens large, une connaissance qualitative. Les règles d'association sont également des connaissances qualitatives extraites de bases de données. Dans ce papier, nous montrons comment la sémantique d'une ontologie de domaine peut être enrichie par des règles d'association (RA) extraites à partir de la base de données de son domaine d'application. L'approche proposée consiste à utiliser les connaissances découvertes comme contraintes d'attributs dans les schémas de l'ontologie de domaine et de la base de données sous-jacente. Notre approche repose sur deux principes : *(i)* le schéma de la base de données et l'ontologie de domaine sont stockés dans le même catalogue, et *(ii)* un ensemble de RA est préalablement extrait. L'ontologie et les données ainsi enrichies en sémantique conduisent à une réduction considérable de l'espace de recherche et la qualité des règles lors des nouvelles tâches d'extraction. Ce qui permet d'améliorer les performances et les capacités des systèmes d'information (SI) en termes de tâches de gestion et/ou d'aide à la prise de décision.

MOTS-CLÉS : ontologie de domaine, règle d'association, règle de gestion, contrainte d'attribut, schéma de base de données, modélisation conceptuelle.

ABSTRACT: In this paper, we show how the semantic of domain ontology can be enriched with associations rules (AR), extracted from the database of its application domain. The proposed approach consists on using the discovered knowledge as attributes' constraints in the database and ontology scheme. Our proposal is based upon two principles: *(i)* the database scheme and domain ontology are stored in a single catalog and, *(ii)* a set of AR is already extracted. As a consequence, the enriched data and ontology lead to reduce the search space and to improve the quality of the AR in future mining tasks. This also allows to improve the IS' capabilities in terms of business tasks and/or decision making.

KEYWORDS: domain ontology, association rule, business rule, attribute constraint, database scheme, conceptual modeling.

1. Introduction

L'ontologie est, au sens large, une connaissance qualitative. Les règles d'association sont également des connaissances qualitatives extraites des bases de données. Par ailleurs, les attentes des utilisateurs ne se limitent plus seulement à des informations alphanumériques mais aussi à des connaissances et des synthèses élaborées. C'est précisément dans ce sens que se réalise la tâche de data mining (DM) d'extraction de connaissances sous forme de règles d'association qui sont des modèles typiques de relations entre les données [1]. Cette capacité du DM à découvrir des connaissances prouve qu'un modèle n'est pas une représentation naïve du réel, mais une interprétation liée aux opérations basées sur la connaissance [7]. Par conséquent, l'information contenue dans les bases de données des SI doit être suffisamment pertinente et cohérente.

Par ailleurs, le but d'une ontologie est, à partir de l'observation du monde réel, d'identifier les instances pertinentes pour ensuite codifier les relations entre elles [3]. L'ontologie doit, à ce titre, satisfaire le besoin en informations des acteurs du domaine, en étant compatible avec les objets et les traitements-métiers. De plus, chaque SI comporte très souvent une ontologie non explicite [5] et tout modèle de données présume implicitement ou explicitement une forme de modélisation de la signification des données [8]. Présentement, trois approches abordent le problème de l'intégration des ontologies au KDD [4]. Il s'agit des approches Onto4KDD, KDD4Onto et Onto4KDD4Onto.

Les travaux basés sur Onto4KDD concernent l'utilisation des ontologies pour améliorer le processus de l'ECD [2][10][14]. KDD4Onto explore des techniques de fouille pour l'acquisition automatique ou semi-automatique de connaissances à partir de données [3]. La dernière approche est une combinaison des deux premières approches. Le double objectif visé [4] a été d'améliorer le processus d'extraction de connaissances par les ontologies et de réutiliser les connaissances découvertes pour en créer de nouvelles ou de réévaluer celles existantes dans l'ontologie de domaine.

La connexion entre les ontologies et le KDD s'avère ainsi un challenge prometteur. Et notre démarche est similaire à celle développée dans [4]. Cependant, aucune de ces approches n'a abordé le problème tel que nous le déclinons dans ce papier.

Le problème posé est, étant donné un ensemble de règles d'association découvertes, d'enrichir la sémantique d'une ontologie de domaine existante ou en cours de développement en utilisant les connaissances obtenues comme contraintes d'attributs généralement exprimées au travers des règles de gestion des SI.

Dans ce but, nous admettons que : ***(i)*** le schéma de la base de données et l'ontologie de domaine sont stockés dans le même catalogue et ***(ii)*** un ensemble de règles d'association est préalablement extrait.

La suite du papier est organisée comme suit : Les sections 2 et 3 présentent, respectivement, les concepts de règle d'association et d'ontologie du domaine. La section 4 examine la similarité entre les deux types de connaissances contenues dans les règles d'association et dans l'ontologie de domaine. Dans la section 5, nous étudions les contraintes d'attributs. Dans la section 6 est présenté un exemple d'application. La section 7 conclut le présent papier et dégage des perspectives.

2. Règles d'association

Les règles d'association sont obtenues par la tâche de data mining (DM) du processus d'extraction de connaissances à partir de bases de données (ECD ou *KDD acronyme de K*nowledge *D*iscovery in *D*atabases) [17]. Traditionnellement, les algorithmes d'extraction de règles d'association étaient appliqués aux transactions du panier de la ménagère où une transaction est un ensemble d'items. Dans ce contexte, une transaction est alors décrite par l'identifiant de la transaction et un ensemble d'items.

Plus formellement, étant donné $I = \{i_1, i_2, \ldots, i_m\}$ l'ensemble des items ; soit D l'ensemble des transactions où chaque transaction T est un ensemble d'items avec $T \subseteq I$. Une règle d'association est une structure de la forme (1).

$$X \Rightarrow Y \text{ où } X \subset I, Y \subset I, \text{ et } X \cap Y = \varnothing \qquad \textbf{(1)}$$

Dans le cas d'une table relationnelle, une transaction est un tuple, et les items des attributs de la relation [1] [9]. Les règles d'association extraites constituent une source d'informations utiles et pertinentes en ce sens qu'elles sont transformables en connaissances permettant d'orienter le déroulement d'un processus, la production de nouvelles informations ou la prise de décisions [7]. En effet, les connaissances découvertes peuvent être utilisées en vue d'expliquer le comportement actuel (modèles descriptifs) ou de prédire des résultats futurs (modèles prédictifs). C'est dans ce cadre que nous avons exploité la sémantique des bases de données relationnelles pour extraire des règles d'association pour la prédiction de valeurs manquantes [9].

Dans ce papier, nous considérons ce type de règles d'association quantitatives calculées sur la base des mesures de qualité, dites objectives, de support des itemsets (n-uplets) et de confiance des règles.

3. Les ontologies de domaine

De son origine philosophique au 17ème siècle à sa première utilisation dans le domaine des technologies de l'information et des sciences en 1967 [6], le terme "Ontologie" a suscité beaucoup d'intérêt. Développé d'abord en Intelligence Artificielle (IA) pour faciliter le partage et la réutilisation de connaissances, le concept d'ontologie a été l'objet de nombreux travaux à partir des années 90 dans divers domaines d'applications [16]. L'ontologie se distingue des autres systèmes de représentation de connaissances car elle permet de modéliser les connaissances de manière explicite et formelle en concepts et relations entre les concepts. La définition consensuelle de l'ontologie largement utilisée dans la littérature a été proposée par [18] pour qui une ontologie est *«une spécification explicite et formelle d'une conceptualisation partagée»*.

Une ontologie de domaine est ainsi nommée parce qu'elle correspond à la conceptualisation spécifique d'un domaine d'application particulier.

Plus formellement, une ontologie est décrite par une *dimension intensionnelle* et une *dimension extensionnelle*.

La dimension intensionnelle, la *T-Box*, décrit les concepts terminologiques du domaine et les relations entre les concepts. La T-Box comprend des relations unaires (description de termes conceptuels) qui modélisent les concepts de classes, et des relations binaires (relations sémantiques entre concepts). Une relation spéciale appelée *relation de subsumption,* et notée $\leq_s$, permet de modéliser les hiérarchies conceptuelles entre les catégories de classes. La T-Box est définie comme un ensemble d'*axiomes* et de *contraintes* qui constituent la structure sémantique de l'ontologie de domaine.

La dimension factuelle, *A-Box,* est un ensemble d'*assertions* qui doivent satisfaire les contraintes définies sur les axiomes de la T-Box. La A-Box est une extension (instance) de la T-Box. A ce titre, elle décrit un état du réel.

Soit Σ= {*SEMENCE, CEREALE, RIZ, BLE, MAIS, LEGUMINEUSE, SOJA, ARACHIDE*} l'ensemble des concepts de l'univers du discours et les relations taxonomiques suivantes: $\leq_\sigma$*(CEREALE,RIZ),* $\leq_s$*(CEREALE,BLE),* $\leq_s$*(CEREALE, MAIS),* $\leq_s$ (*LEGUMINEUSE, SOJA),* $\leq_s$*(LEGUMINEUSE, ARACHIDE)*.

Etant donné ce domaine d'application et admettant que *SEMENCE* est la racine de l'ontologie, la figure 1 montre les deux taxonomies actuelles de cette ontologie de domaine que nous appellerons *Semences*.

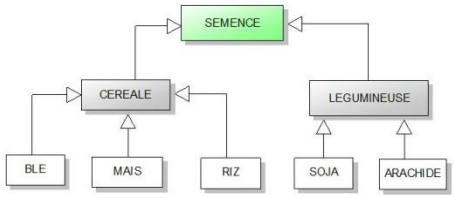

Figure 1 *Exemples de taxonomies du domaine Semences*

Nous admettrons ici une ontologie terminologique contenant des termes avec leurs définitions, des propriétés et des contraintes sur ces propriétés. De plus, nous considérons que les composantes de l'ontologie, la *T-Box* et la *A-Box*, sont conçues dans l'esprit d'*OntoViews CM* [19]. A titre d'exemple, une ***semence***[1] est une *«graine sélectionnée pour être semée»* [11]. Le concept *Semence* est décrit au minimum par quatre propriétés : *PURETE VARIETALE, PURETE SPECIFIQUE, FACULTE GERMINATIVE.* Soit par exemple la déclaration (2) représentant ce concept dans la T-Box.

Semence (SID, S_ESPECE, VAR, GEN, PV, PS, FGERM) ***(2)***

Dans la A-Box correspondante, on pourra enregistrer une définition pour chaque terme telle que par exemple une définition explicitant la propriété *PV* comme *«taux de graines s'écartant de la plante modèle de la variété. Pour les semences certifiées la pureté variétale est de l'ordre de 99,7%»*.

Cependant, les applications nécessitent des données alphanumériques pour effectuer les traitements et répondre aux requêtes. Par exemple, la base de données stockera un tuple sous la forme ***(3).***

<1,'RIZ', 'ORIZA SATIVA', 'BASE', 94.7, 91.2, 83> ***(3)***

Ce tuple renseigne qu'une semence de riz de variété '*ORIZA SATIVA',* de niveau '*BASE'* avec une pureté variétale de *94.7%,* une pureté spécifique de *91.2%* et une faculté germinative égale à *83%* est enregistrée. De plus, des données concernant les acteurs (opérateurs privés semenciers, multiplicateurs de semences) et des données connexes (parcelles exploitées) peuvent être stockées dans la base de données. Ces données sont contrôlées par des règles de gestion ou règles-métiers du type (4) ci-après.

«L'agriculteur multiplicateur devra avoir un contrat de multiplication en bonne et due forme avec un établissement semencier agrée» ***(4)***

[1] «Semence vient, au travers du latin, du grec sperma = semence, germe. Ce terme a également donné le terme sperme. Par analogie entre l'homme et l'agriculteur d'une part, la femme et la terre de l'autre, il a pris en agriculture le sens de graine que l'on plante en vue d'une récolte. De même que l'homme était supposé "ensemencer" la femme pour qu'elle porte un enfant, l'agriculteur ensemençait la terre pour qu'elle porte une récolte» [WikipédiA]

Intuitivement, la règle-métier **(4)** suppose un ensemble connu d'établissements semenciers agréés. La base de données est construite en créant une table pour chaque concept de classe défini dans l'ontologie [19]. Autrement dit, si aucune optimisation n'est effectuée, la base de données comportera au moins autant de tables que de concepts dans l'univers Σ.

Dans la section suivante, nous présentons quelques aspects des contraintes d'intégrité rencontrées dans les bases de données relationnelles.

4. Les contraintes d'attributs

La conception des schémas de bases de données et des ontologies consiste en une description de la structure des données et la spécification d'un certain nombre de contraintes. Les contraintes existantes peuvent être classées dans deux catégories principales : les *contraintes structurelles* ou *implicites* et les *contraintes syntaxiques* dites *explicites*.

Les contraintes *implicites* (unicité de clé, clés étrangères et références) sont inhérentes au modèle relationnel et ne sont pas en considération dans ce papier.

Les contraintes explicites sont des contraintes sémantiques qui dépendent du domaine d'application et sont généralement prises en compte par un langage de spécification de contraintes. On y distingue les *contraintes de colonnes*, les contraintes de table, les contraintes d'assertion, les *contraintes de domaines* et les *déclencheurs*.

Une contrainte de colonne ou de domaine est appliquée à un *attribut* de relation. Elle garantit la cohérence et la signification des valeurs de l'attribut dans la base de données relationnelle. De plus, la définition d'un attribut sur un type de données induit implicitement un domaine des valeurs possibles pour l'attribut considéré. Une restriction peut alors être réalisée sur une plage du domaine pour contraindre les valeurs pouvant apparaître pour l'attribut. Par exemple, pour le niveau (génération) d'une variété de semence les valeurs possibles sont prises dans l'ensemble {'G0', 'G1', 'G2', 'G3', 'BASE', 'R1', 'R2'} qui est identique à {'PREBASE' , 'BASE', 'R1', 'R2'} avec PREBASE $\in$ { 'G0', 'G1', 'G2', 'G3'}.

5. Des règles d'association aux contraintes d'attributs

Soit les deux règles d'association [9] (5.1) et (5.2) sur le domaine *Semences :*

$$<VAR = \text{'JAYA'} \Rightarrow FGERM \in [90, 95] > \quad \textbf{(5.1)}$$

$$<VAR = \text{'JAYA'} \wedge GEN = \text{'R1'} \Rightarrow FGERM \in [92, 95]> \quad \textbf{(5.2)}$$

Intuitivement, chacune de ces règles est une conceptualisation sur le domaine d'application. La règle (5.1) spécifie que la faculté germinative est comprise entre 90 et 95 tandis que la deuxième, plus spécifique, renseigne que la faculté germinative est comprise entre 92 et 95 si la variété de riz est 'JAYA' et la génération de semence 'R1'.

Plus généralement, les valeurs contenues dans les conséquents (parties droites) des règles spécifient les valeurs pouvant être prises par la classe d'objets dont les attributs constituent les antécédents des règles (parties gauches).

Une analyse fine de la structure et de la sémantique des règles d'association et des contraintes d'intégrité, nous permet de conclure que : ***(i)*** les règles d'association de type (5.1) et (5.2) sont équivalentes à des *contraintes de domaines* pour les attributs numériques continus ou discrets ; et plus spécifiquement ***(ii)*** les règles de type (5.2) sont assimilables à l'implémentation de *déclencheurs*.

6. Exemple d'application

Le domaine d'application considéré est la certification de semences. D'après la règle-métier (4) énoncée dans la section 3, un agriculteur multiplicateur «*détenant un contrat de multiplication en bonne et due forme avec un établissement semencier agréé»* peut accéder aux connaissances (tel que le pedigree d'une espèce) et aux données connexes de la base. La figure 2 donne un aperçu sur la connexion entre l'ontologie de domaine et le schéma de la base de données.

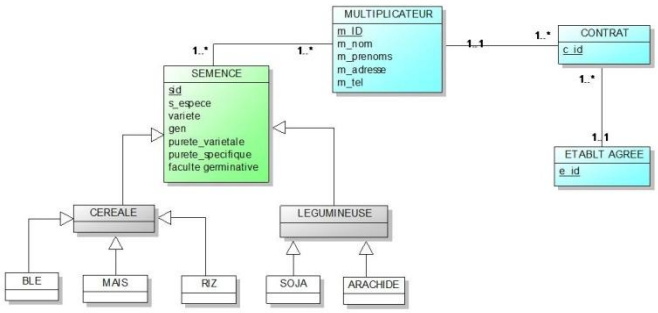

Figure 2 *Aperçu du domaine d'application*

Dans [9], les règles calculées sont de la forme (6).

$$\langle A_{i1} = v_1 \wedge Ai_2 = v_2 \wedge \ldots \wedge A_{ik} = v_k \Rightarrow A_{i0} \in E \rangle \qquad \textbf{(6)}$$

Dans (6), A_{i0} est l'attribut prédit et E est soit un intervalle soit un ensemble de valeurs selon que l'attribut A_{i0} est de type continu ou discret. Les règles ont été calculées avec un seuil de support de 14% et une confiance de 100%. Pour simplifier, nous allons considérer les quatre règles ((7.1) à (7.4)) invoquant la variété 'JAYA'.

$$\langle \text{VAR} = \text{'JAYA'} \Rightarrow \text{FGERM} \in [90,95] \rangle \quad \textbf{(7.1)}$$

$$\langle \text{VAR} = \text{'JAYA'} \wedge \text{GEN} = \text{'R1'} \Rightarrow \text{FGERM} \in [92,95] \rangle \quad \textbf{(7.2)}$$

$$\langle \text{VAR} = \text{'JAYA'} \wedge \text{PS} = \text{'MOYENNE'} \Rightarrow \text{FGERM} \in [92,95] \rangle \quad \textbf{(7.3)}$$

$$\langle \text{VAR='JAYA'} \wedge \text{GEN='R1'} \wedge \text{PS='MOYENNE'} \Rightarrow \text{FGERM} \in [93,95] \rangle \quad \textbf{(7.4)}$$

Sur la base de (7.1), et connaissant uniquement le nom de l'espèce ('JAYA'), un expert du domaine est en mesure d'affirmer que la faculté germinative est comprise entre 90 et 95. Si en plus, il sait que la génération est 'R1' ou que la pureté est 'MOYENNE', les règles (7.2) et (7.3) lui autorisent de restreindre le domaine sur [92, 95]. La règle (7.4) est la plus spécifique.

En résumé, les règles découvertes offrent la possibilité de créer trois domaines permettant la restriction de l'attribut *FGERM*. De plus, ces connaissances découvertes garantissent des élagages importants lors de nouvelles tâches d'extraction par les algorithmes d'extraction.

7. Conclusion et perspectives

Dans ce papier, nous avons mis en exergue la pertinence de la réutilisation des connaissances découvertes dans les règles d'association comme contraintes d'attributs pour enrichir l'ontologie de domaine. De fait, plus qu'une simple hiérarchie de concepts du domaine, l'ontologie inclut un ensemble de contraintes sur ces concepts.

Le premier intérêt de l'approche est une meilleure description des schémas. En effet, le manque de représentation conceptuelle des connaissances imposerait la collaboration des experts du domaine pour la construction de l'ontologie à partir des bases de données existantes ou inversement. Ainsi, contrairement à [4] où la construction de l'ontologie est la dernière étape, nous avons proposé dans [19] une approche d'intégration de l'ontologie du domaine au cycle de vie des bases de données.

Un second avantage est la réduction de l'espace de recherche lors de nouvelles tâches d'extraction car il a été montré que seulement 20% des données effectivement stockées dans la base de données sont concernées par 80% des requêtes utilisateurs [13] [15]. De plus, il est rare que les attributs et les tuples d'une table soient tous utilisés dans les requêtes.

Par ailleurs, aucun des travaux [2] [3] [4] [10] [14] ayant abordé le rapprochement entre le DM et les ontologies n'a évoqué l'enrichissement sémantique par les contraintes d'intégrité dans les schémas.

En perspective, nous envisageons de mettre en place une interface d'administration de l'ontologie de domaine et corollairement le schéma de la base de données sous-jacente dans l'esprit de [19]. Comme préalable, l'objectif est de développer prochainement une ontologie de la certification des semences. Nous exploiterons à ce titre la structure de la base de données dénommée «GECCSEM»[2]

8. Bibliographie

[1] ANDREA, B., FURLETTI, B., GROSSI, V., ROMEI, A. «Pushing constraints in association rules mining: An Ontology-Based Approach», *IADIS International Conference WWW/Internet 2007*

[2] CLAUDIA, M. AND FABRICE, G., «IMPROVING POST-MINING OF ASSOCIATION RULES WITH ONTOLOGIES», *the XIII International Conference «Applied Stochastic Models and Data Analysis», (ASMDA-2009), Vilnius, LITHUANIA, June 30-July 3, 2009.*

[3] COSTELLO, J. C, DAN, S., JEFF, G., MEHMET, D., *«DATA-DRIVEN ONTOLOGIES», Pacific Symposium on Biocomputing* 14:15-26, 2009.

[4] GOTTGTROY, P. «Ontology Driven Knowledge Discovery Process: a proposal to integrate Ontology Engineering and KDD», *11th Pacific-Asia Conference on Information Systems, 2007.*

[5] GUARINO N. Formal Ontology and Information Systems, *in N. Guarino (Ed.) Formal Ontology in Information Systems*, Amsterdam, Netherlands: IOS Press, 1998.

[6] GUIZZARDI, G. On Ontology, ontologies, Conceptualizations, Modeling Languages, and (Meta)Models, *Databases and Information Systems IV*, O. Vasilecas et al. (Eds.), IOS Press, 2007.

[7] HERNANDEZ, N., Ontologie de domaine pour la modélisation du contexte en recherche d'information, Thèse, Université Paul Sabatier (Toulouse), 2006.

[2]GECCSEM est l'acronyme de *«Gestion du Contrôle et de la certification des Semences»*. Nous avons conçu cette base de données dans le cadre d'un projet financé par l'Union Européenne et piloté par l'UGP-STABEX (Unité de Gestion de Projets -Stabilisation des Exportations) pour le compte de la DA-DISEM en 2008.

[8] HIRSCHHEIM R. A., H.-K.KLEIN, AND LYYTINEN, K., Information Systems Development and Data Modeling: Conceptual and Philosophical Foundations. Cambridge; New York: Cambridge University Press, 1995.

[9] JAMY, S.I., TAO-YUAN, J., DOMINIQUE, L., GEORGES, L., SY, O., «Extraction de règles d'association pour la prédiction de valeurs manquantes», *Revue Africaine de la Recherche en Informatique et Mathématique Appliquée ARIMA,* vol. Spécial CARI04, pp. 103–124, 2004.

[10] KERDPRASOP, N. AND KERDPRASOP, K. Moving Data Mining Tools toward a Business Intelligence System, World Academy of Science, Engineering and Technology, 2007.

[11] LACHARME, M. «La production de semences certifiées Règles à suivre à l'exploitation», Mémento Technique de Riziculture», Juin 2001, http://crrmc.ilemi.net/IMG/pdf/10_Production_des_semences_certifiees.pdf

[12] MAN, L., XIAO-YONG D., SHAN, W., LEARNING ONTOLOGY FROM RELATIONAL DATABASE*, Proceedings of the Fourth International Conference on Machine Learning and Cybernetics*, Guangzhou, 18-21 August, 2005

[13] NANCY, D., ESPINASSE, B. *Ingénierie des Systèmes d'Information : Merise deuxième génération,* SYBEX, 1998.

[14] PINTO, F., MANUEL, F. S., MARQUES, A. «Ontology based Data Mining – A contribution to Business Intelligence», *Proceedings of the 10th WSEAS Int. Conference on MATHEMATICS and COMPUTERS in BUSINESS and ECONOMICS*

[15] RICHARD, D. Une méthodologie de conception physique de base de données pour le système expert SECSI, Thèse de l'Université Paris VI, 1989

[16] SHU-HSIEN, L., JEN-LUNG, C., TZE-YUAN, H.: «Ontology-based data mining approach implemented for sport marketing », Expert Systems with Applications 36 (2009) 11045–11056, journal homepage: www.elsevier.com/locate/eswa

[17] SOULET, A. Un cadre générique de découverte de motifs sous contraintes fondées sur des primitives, THESE, Université de Caen / Basse-Normandie, 2006

[18] STUDER R., BENJAMINS R., FENSEL D. Knowledge Engineering: Principles and Methods*, Data Knowledge Engineering*, 1998.

[19] SY, O., DUARTE, D., LO, M. "Integrating Ontologies in Database Scheme: Ontology-Based Views Conceptual Modeling*, the 6th International Conference on Signal-Image Technology & Internet-Based Systems SITIS*, 15-18 December, Kuala Lumpur (Malaysia), 2010.