Prise en compte des « Points de Vue » pour l'annotation d'un processus d'Extraction de Connaissances à partir de Données

Hicham Behja(*),(**),(***), Brigitte Trousse(**), Abdelaziz Marzak(***)

(*) ENSAM Meknes Marjane 2 ; B.P. 4024
Beni Mhammed Meknes Maroc
(**) INRIA Sophia Antipolis Projet AxIS,
BP 93, 06902 Sophia Antipolis, France
(***){Nom.Prenom}@sophia.inria.fr
(***) Faculté des Sciences Ben M'sik de Casablanca
Avenue Driss Harti B.P 6621 Casablanca, Maroc
marzak@hotmail.com

Résumé. Dans cet article on propose une nouvelle approche qui rend explicite la notion de point de vue dans une analyse multivues issue d'un processus d'Extraction de Connaissances à partir de Données (ECD). Par point de vue, nous entendons la vision particulière d'un analyste lors de son processus ECD, vision référant à un corps de connaissances qui lui est spécifique. On cherche, d'une part, à faciliter la réutilisabilité et l'adaptabilité du processus, et d'autre part à garder une trace des points de vues sous-jacents aux analyses faites. Le processus d'ECD sera vu comme un processus de génération et de transformation de vues qui seront annotées par des métadonnées pour garder la sémantique de la connaissance extraite. Un positionnement de notre approche vis-à-vis des travaux méthodologiques du processus d'ECD sera donné. Des éléments de modélisation du processus ECD basé sur les points de vue seront décrits au niveau ontologique. Enfin, on illustrera notre approche sur l'analyse des usages d'un site web à partir des fichiers log, selon le point de vue fiabilité.

1 Introduction

Le processus d'ECD est un processus itératif et interactif, constitué principalement de trois grandes étapes: prétraitement, fouille de données et postraitement (Fayyad *et al.,* 1996(a)). Il se présente comme un processus complexe tant au niveau des techniques et méthodes qu'au niveau des données manipulées (Gancarski et Trousse, 2004).

Dans le cadre de nos recherches sur l'analyse des usages d'un système d'information (Tanasa et Trousse, 2004), nous nous intéressons au processus d'ECD appliqué aux données Web. Ces données peuvent être hétérogènes (textes, hypertexte, images, vidéo, etc.), incohérentes, évolutives, incomplètes mais en général, on peut dire qu'elles sont bien structurées dans le sens où elles respectent un format bien connu par les analystes d'ECD (comme le format CLF «Common Log File »). Mais cette structure reste relative et liée au point de vue de l'analyste d'ECD. Les mêmes données peuvent être vues mal structurées du point de vue de l'analyste des comportements des utilisateurs d'un système d'information (SI) basé sur le Web. En effet ces données sont relativement pauvres pour répondre aux objectifs de l'analyste en termes de comportement utilisateur. Cette approche nécessite de plus amples informations sur les utilisateurs qui se présentent comme les acteurs centraux d'une analyse des comportements d'un SI basé sur le Web. Ces objectifs passent,