De l'importance du pré-traitement des données pour l'utilisation de l'inférence grammaticale en *Web Usage Mining*

Thierry Murgue

Eurise – Université Jean Monnet
23 rue du docteur Paul Michelon
42023 Saint-Étienne Cedex 2
thierry.murgue@univ-st-etienne.fr

Résumé. Le *Web Usage Mining* est un processus d'extraction de connaissance qui permet la détection d'un type de comportement usager sur un site internet. Cette tâche relève de l'extraction de connaissances à partir de données : plusieurs étapes sont nécessaires à la réalisation du processus complet. Les données brutes, utilisées et souvent incomplètes correspondent aux requêtes enregistrées par un serveur. Le pré-traitement nécessaire de ses données brutes pour les rendre exploitables se situe en amont du processus et est donc très important. Nous voulons travailler sur des modèles structurés, issus de l'inférence grammaticale. Nous détaillons un ensemble de techniques de traitement des données brutes et l'évaluons sur des données artificielles. Nous proposons, enfin, des expérimentations mettant en évidence l'affectation des algorithmes classiques d'inférence grammaticale par la mauvaise qualité des logs bruts.

1 Introduction

Le *Web Usage Mining* a été introduit pour la première fois en 1997 (Cooley et al. 1997). Dans cet environnement, la tâche est d'extraire de manière automatique la façon dont les utilisateurs naviguent sur un site web. Depuis 1995, Catledge et Pitkow ont étudié la manière de catégoriser les comportements utilisateurs sur un site web (Catledge 1995). Le processus d'extraction de connaissance – pré-traitement, fouille, interprétation – est basé sur la disponibilité de données fiables : divers travaux on été menés sur la façon de traiter les données récupérables depuis un site web (Cooley et al. 1999, Pitkow 1997, Chevalier et al. 2003). Une grande majorité de chercheurs utilisent de manière systématique les informations contenues dans les enregistrements du serveur (fichiers de logs), mais ces données, sous forme brute, ne sont pas complètes : un pré-traitement est donc nécessaire. L'étape suivante du Web Usage Mining consiste à apprendre des modèles de comportement utilisateurs depuis ces données. Ainsi, ces dernières années, de nombreuses méthodes de traitement (Tanassa et al. 2004) et d'apprentissage ont été utilisées dans ce domaine : recherche de séquences fréquentes (Frias-Martinez 2002, Gery 2003), travaux sur l'utilisation de modèles structurés de type chaîne de Markov ou modèle de Markov caché (Hmm) (Pitkow 1999, Bidel et al. 2003). Certains chercheurs ont notamment travaillés sur des modèles grammaticaux : certains (Borges 1999) en utilisant des *n-grams*, d'autres (Karampatziakis et al. 2004) en étudiant le comportement