

# Mesure d'audience sur Internet par populations de fourmis artificielles

Nicolas Labroche

UPMC, LIP6, Pole IA, 8 rue du Capitaine Scott  
75015 Paris  
nicolas.labroche@lip6.fr  
<http://lofti.lip6.fr>

**Résumé.** Nous présentons dans ce travail un outil pour la mesure d'audience sur Internet, reposant sur l'extraction de profils de navigation représentatifs de l'activité des internautes sur les sites. Ces profils sont obtenus par l'application d'un algorithme de classification non supervisée – inspiré du système de reconnaissance chimique des fourmis – sur des sessions de navigations construites à partir des fichiers log du site étudié. Cet algorithme de classification a été associé à une représentation multimodale des sessions utilisateurs permettant d'employer l'ensemble des informations à disposition dans les fichiers log (impacts sur les pages, heure de connexion, durée, séquence des pages, ...), ainsi qu'à une mesure de similarité adaptée pour créer les profils de chacun des clusters obtenus. Il reste cependant d'autres modalités (basées sur le contenu des documents accédés) qui pourraient améliorer la capacité de l'outil à donner du sens aux profils découverts.

## 1 Introduction

La mesure d'audience sur Internet s'attache à extraire et à donner du sens aux navigations des internautes. Ses champs d'applications sont nombreux : personnalisation automatique des sites Web en fonction des pages accédées (Mobasher et al., 1999) ou encore recommandation dynamique de pages aux internautes en fonction de leur navigation passée (Yan et al., 1996). Dans ce cadre, la représentation des navigations des internautes est cruciale car elle détermine le type de recherche d'informations qui pourra être conduit par la suite.

Généralement, les informations de navigation des internautes sur un site Internet sont extraites d'un fichier log, présent sur le serveur Web, qui recense, pour simplifier et de manière idéale, l'ensemble des demandes de pages du site de la part des internautes. Ces requêtes clientes sont ensuite triées, filtrées et regroupées en sessions qui constituent pour chaque internaute, l'ensemble des informations issues de leur navigation sur un site à un moment donné (Cooley et al., 1999).

Plusieurs représentations des sessions ont été utilisées dans la littérature. Par exemple, le système WebMiner (Cooley et al., 1999) utilise un vecteur de transactions qui indique pour chacune des pages du site si elle a été accédée au moins une fois durant la session de façon à extraire des règles d'association. Dans Masseglia et al. (1999), les auteurs conservent les dates d'accès à chacune des pages pour extraire des règles séquentielles. D'autres représentations ont été utilisées comme la durée de visite ou le nombre d'impacts par page