# Comparaison de dissimilarités pour l'analyse de l'usage d'un site web

Fabrice Rossi*, Francisco De Carvalho**, Yves Lechevallier*, Alzennyr Da Silva*,**

*Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex – France
**Centro de Informatica - CIn/UFPE
Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil

**Résumé.** L'obtention d'une classification des pages d'un site web en fonction des navigations extraites des fichiers "logs" du serveur peut s'avérer très utile pour évaluer l'adéquation entre la structure du site et l'attente des utilisateurs. On construit une telle typologie en s'appuyant une mesure de dissimilarité entre les pages, définie à partir des navigations. Le choix de la mesure la plus appropriée à l'analyse du site est donc fondamental. Dans cet article, nous présentons un site de petite taille dont les pages sont classées en catégories sémantiques par un expert. Nous confrontons ce classement aux partitions obtenues à partir de diverses dissimilarités afin d'en étudier les avantages et inconvénients.

## 1 Introduction

La conception, la réalisation et la maintenance d'un site web volumineux sont des tâches difficiles, en particulier quand le site est écrit par plusieurs rédacteurs. Pour améliorer le site, il est alors important d'analyser les comportements de ses utilisateurs, afin de découvrir notamment les incohérences entre sa structure *a priori* et les schémas d'utilisation dominants. Les utilisateurs contournent en effet souvent les limitations du site en navigant (parfois laborieusement) entre les parties qui les intéressent, alors que celles-ci ne sont pas directement liées aux yeux des concepteurs. A l'opposée, certains liens sont très peu utilisés et ne font qu'encombrer la structure hyper textuelle du site.

Une méthode d'analyse dirigée par l'usage consiste à réaliser une classification du contenu du site à partir des navigations enregistrées dans les logs du serveur. Les classes ainsi obtenues sont constituées de pages qui ont tendance à être visitées ensembles. Elles traduisent donc les préférences des utilisateurs. La principale difficulté de cette approche réside dans la nature des observations (les navigations). Comme celles-ci sont de taille variable, on peut en déduire de nombreuses mesures de dissimilarité entre les pages visitées, selon qu'on tient compte de la durée de la visite, du nombre de fois que la page est vue, etc. Dans le contexte de la classification, il est alors difficile de choisir *a priori* quelle mesure de dissimilarité est la plus adaptée à l'analyse du site.

Dans cet article, nous étudions un site web peu volumineux (91 pages), très bien structuré, et au contenu sémantique bien défini. Grâce à cet exemple de référence, nous comparons différentes dissimilarités afin de mesurer leur aptitude à révéler ce contenu sémantique.