Traitement et exploration du fichier Log du Serveur Web, pour l'extraction des connaissances : Web Usage Mining

Mostafa Hanoune*, Faouzia Benabbou*
* Université Hassan II- Mohammedia, Faculté des sciences Ben M'Sik, Laboratoire TIM (Technologies de l'information et Modélisation), Casablanca, Maroc
m_hanoune@yahoo.fr, hgfbenabbou@menara.ma.

Résumé. Le but dans ce travail consiste à concevoir et réaliser un Outil Logiciel, en utilisant les concepts du Web Usage Mining pour offrir aux web masters l'ensemble des connaissances, y inclut les statistiques sur leurs sites, afin de prendre les décisions adéquates. Il s'agit en fait, d'extraire de l'information à partir du fichier log du serveur Web, hébergeant le site Web, et de prendre les décisions pour découvrir les habitudes des internautes, et de répondre à leurs besoins en adaptant le contenu, la forme et l'agencement des pages web.

1 Introduction

L'activité sur le Web et les données résultantes ont connu une croissance très rapide, vu la croissance exponentielle du nombre des documents mis en ligne.

D'après des statiques sur des sites spécialisés, le nombre des utilisateurs d'Internet dans le monde a dépassé le milliard (1 022 863 307), au mois de mars 2006[1], et le nombre de sites Web a atteint 74,4 millions au mois d'Octobre 2005[2]. Ces données, en particulier celles relatives à l'usage du Web, sont traitées dans le Web Usage Mining (WUM). Dans cet article, nous décrivons les fonctionnalités majeures du logiciel que nous avons conçu et réalisé, et qui permet l'analyse des fichiers Logs afin de comprendre le comportement des internautes sur un site Web (Site de l'université Hassan II- Mohammedia www.univh2m.ac.ma Casablanca, Maroc).

2 Proposition

L'apport de ce travail réside principalement dans les points suivants :

1. Connaissances sur les visiteurs :
   (a) Le pourcentage des visiteurs par semaine par mois et par an
   (b) Avoir une visibilité internationale : d'où proviennent nos visiteurs ?
2. Connaissances sur les pages :
   (a) Les pages les plus et les moins consultées (pages populaires et pages impopulaires)
   (b) Les combinaisons des pages consultées
   (c) Savoir quels sont les liens qui nous référencent le mieux
3. Connaissances sur les navigateurs et les OS
   (a) Le pourcentage des navigateurs les plus utilisés

[1] http://www.internetworldstats.com/stats.htm
[2] http://www.netcraft.com

(b) Le pourcentage des systèmes d'exploitations les plus utilisés

Le travail est subdivisé en trois parties distinctes :

- La première partie présente la conception, par la méthode UML, de la solution mise en place,
- La deuxième partie correspond aux différentes étapes du prétraitement et nettoyage du fichier Log,
- La dernière est consacrée à l'exploration et l'analyse de fichier Log.

Références

Charrad, M., M. Ben Ahmed et Y. Lechevallier (2005). Web Usage Mining: WWW pages classification from log files. *In Proceeding of International Conference on Machine Intelligence*, Tozeur, Tunisia, 5-7 Novembre.

Cooley, R., B. Mobasher, et J. Srivastava (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems*.

Malika Charrad, Mohamed Ben Ahmed, Yves Lechevallier (2005), Extraction des connaissances à partir des fichiers Logs,*Actes de l'atelier Fouille du Web des 6èmes journées francophones «Extraction et Gestion des Connaissances »*.

Pierrakos, D., G. Paliouras, C. Papatheodorou, et C.D. Spyropoulos (2003). Web Usage Mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction*, 13:311-372.

Srivastava, J., R. Cooley, M. Deshpande et P.-N. Tan (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKKD Explorations.*

Tanasa, D. et B. Trousse (2003), Le prétraitement des fichiers Logs Web dans le Web Usage Mining Multi-sites. *In Journées Francophones de la Toile*.

Hanene Azzag, Christiane Guinot Gilles Venturini (2005), Classification hiérarchique et visualisation de pages Web, *Actes de l'atelier Fouille du Web des 6èmes journées francophones « Extraction et Gestion des Connaissances »*.

Khalid Benabdeslem, Younès Bennani (2005),Classification et visualisation des données d'usages d'Internet, *Actes de l'atelier Fouille du Web des 6èmes journées francophones « Extraction et Gestion des Connaissances »*.

Thomas Guyet, Catherine Garbay, Michel Dojat, (2006), Algorithme d'apprentissage de scénarios à partir de séries symboliques temporelles, Fouille de données temporelles, Atelier à EGC 2006, 17 janvier 2006

Summary

The goal of this work consists in designing and producing a Software Tool, by using the concepts of the Web Usage Mining, to offer to the Webmasters complete knowledge in order to make appropriate decisions. It works as follows: it extracts information from log files and it makes decisions to discover the behaviours of the users. It adapts to the users behaviour by adapting the contents and general aspects of the Web pages.