Projet LOG - LP STID Menton 2004/2005

Analyse des fichiers logs des sites Web de l'INRIA

Sergiu CHELCEA
Doru TANASA
Brigitte TROUSSE

Dans ce document vous trouverez la description :

- des termes utilisés dans le processus de Web Usage Mining (WUM),
- des données utilisées, de leur prétraitement et du résultat final de prétraitement (fichier *csv*)

1. Description des termes du WUM

Dans le cadre du processus de WUM on utilise des termes tels que ***utilisateur, serveur Web, site Web, page Web, ressource Web, etc.*** Afin de mieux se comprendre, avant de commencer, voyons la signification exacte de chacun de ces termes :

Ressource Web - Toute ressource (image, fichier html, etc.) accessible via un protocole HTTP.

Serveur Web - Un serveur qui donne accès à des ressources Web.

Page Web - Ensemble des informations, consistant en une ou plusieurs ressources Web, identifiées par un seul URI. Exemple : un fichier HTML, un fichier image et un applet Java accessibles par un seul URI constituent une page Web.

Présentation de page (Page View) - Le fait d'afficher une page Web dans l'environnement visuel client à un moment précis dans le temps.

Navigateur Web (Browser) - Logiciel de type client chargé d'afficher des pages pour l'utilisateur et de faire des requêtes HTTP au serveur Web (ex. Internet Explorer, Netscape Navigator, Opera, etc.). Le navigateur Web, sa version ainsi que le système d'exploitation employé par l'utilisateur sont notés dans le champ "User Agent" (champ 5) du fichier log.

Utilisateur - Personne qui accède à des pages Web situées sur un ou plusieurs serveurs Web, en utilisant un navigateur (browser).

Session utilisateur - Un ensemble délimité des clicks utilisateurs sur un ou plusieurs serveur Web.

Navigation ou visite – sous-ensemble d'une session utilisateur. La distance en temps entre toutes 2 requêtes consécutives est inférieure à un seuil prédéfini (en général le seuil est de 30 minutes).

2. Données

Les requêtes des utilisateurs sont stockées dans les fichiers log Web du serveur Web. Ces fichiers log Web peuvent atteindre plusieurs centaines de Megaoctets par jour pour un serveur Web de taille moyenne (le serveur Web de l'INRIA collecte plus de 50 MO de logs Web par jour).

Les données du projet ***LOG*** représentent un mois de logs Web et elles sont structurées en sessions et navigations. Nous fournissons aussi des informations sur les utilisateurs du site (unité de recherche, projet ou service) lorsqu'ils viennent d'une machine INRIA.

Les données utilisées dans le prétraitement sont les deux fichiers log provenant de deux serveurs Web de l'INRIA, pour la période du 1er au 31 octobre 2004 :

1. http://www.inria.fr/ (site national, le siège)
2. http://www-sop.inria.fr/ (unité de recherche de Sophia Antipolis)

A. Nettoyage des données

Dans le processus de nettoyage des données nous avons éliminé les requêtes :

1. Concernant un fichier de type : .jpg, .gif, .png, .ico, .css, .class, .jsc, etc.
2. Provenant des serveurs d'indexation Altavista de l'INRIA (User Agent égal à "AltaVista Intranet V2.0")
3. Provenant des robots d'indexations connus (liste prédéfinie) ou identifiés (requêtes pour « robots.txt »)

B. Codage du nom de machine

Par raison de sécurité nous avons choisi de rendre anonymes ces données. Voici le processus par lequel nous avons accompli cela :

machine.organisation.pays (ex. "manon.unice.fr", "bot.google.com") devient

ID.example.com.pays (ex. "123.example.com.fr", "456.example.com.com")

Les machines d'INRIA (ne sont pas données dans le fichier csv actuel) :
Pour Sophia nous avons gardé aussi un identificateur pour le projet ou service auquel la machine appartient

Par exemple :
"gentiane.inria.fr"
devient
"789.example.com.99.projet.sophia.inria.fr" (99 est l'ID du projet axis)

"tempete.inria.fr"
devient
"111.example.com.88.service.sophia.inria.fr" (tempete apartient au service Semir et l'ID du service Semir est 88)

Pour les quatre autres sites de l'INRIA (Rocquencourt, Rhône-Alpes, Lorraine, Rennes) on a respectivement :
"123.example.com.rocquencourt.inria.fr"
"123.example.com.inrialpes.inria.fr"
"123.example.com.loria.inria.fr"
"123.example.com.irisa.inria.fr"

C. Groupement des données en sessions et navigations

1. Nous avons groupé les requêtes provenant d'un même IP avec un même User Agent en leur attribuant un même "ID_Session"
2. En suite, nous avons divisé une session" (définie comme l'ensemble des requête ayant le même ID_Session) dans plusieurs navigations. Une nouvelle navigation commence lors de la première requête d'une session ou lorsque la distance (en temps) entre deux requêtes consécutives (d'une même session) dépasse 30 minutes. Toutes les requêtes d'une navigation reçoivent le même ID_Navigation.

Dans ces conditions une **navigation** est constituée par l'ensemble des requêtes ayant le même couple (ID_Session, ID_Navigation).

D. Description des données prétraitées (fichier *csv*)

Le fichier *csv* obtenu après le prétraitement contient 12 champs séparés par ¨;¨ et décrits ci dessous (11 champs utilisables en effet) :

1. ID_Navigation - un identificateur pour la navigation de la session
2. ID_Session - un identificateur pour la session
3. IP - l'IP ou le nom de la machine, codé (voir B.)
4. Login - le login de l'utilisateur (non utilisé, à ignorer)
5. User_Agent + Version + OS :
   - le navigateur de l'utilisateur du fichier log +
   - la version du navigateur +
   - l'OS de l'utilisateur (Windows, Linux, etc.)
6. Date_req - la date de la requête (Jour Mois Année)
7. Time_req - le temps de la requête (heure:min:sec)
8. Duration - le temps passé sur la page
9. Bytes - dimension du fichier demandé
10. Status - le statut de la requête (voir les "status codes")
11. URL - l'adresse URL du fichier
12. Referer - l'adresse de la page de provenance de la requête