CPM

Analyse de logs

Mars 2013

Alex Pavy

Structure du code

- Vues :

Filtres : choix filtres

Session : affichage statistiques sessions, export

- Modèle

Analyseur : calculs statistiques

LogsReader : Lecture / Ecriture

RegexExtractor : matching sur strings

Visit, Session : Stockage données

- Mediateur : Appel des méthodes

Filtrages

- Par type de fichier de requête : htm, jpg, …
- Par code de retour : 200, 404, …
- Par présence de Bot : googlebot, crawler, …

Création nouveau fichier de base

Sessions

- Visite : adresse après le « GET » + date
- Session : ip + id + userAgent + liste visites

Calcul sessions :

- Par temps limite :

Liste visites triée par dates

- Par referer :

Session contient aussi le string du referer

Statistiques

- Nombre de sessions

- Adresses ip qui ont le plus de sessions :

Liste adresses ip, nombre de session triée

- Requêtes plus fréquentes:

Liste requête, nombre de fois qu'elle est effectuée triée

Export Weka

- Liste toutes requêtes uniques

Sessions sous forme de vecteurs :

1 si requête présente

0 sinon

Optimisation

- Filtrage : une lecture, fichier en entier
- Création des sessions : une lecture , fichier en entier
- Statistiques : un parcourt des sessions
- Export : un parcourt des sessions

Exemple

bourges.txt

Analog

- Successful requests: 41,536
  Average successful requests per day: 5,935
  Successful requests for pages: 10,084
  Average successful requests for pages per day: 1,441
  Failed requests: 445
  Redirected requests: 1
  Distinct files requested: 353
  Distinct hosts served: 1,152
  Data transferred: 103.63 megabytes
  Average data transferred per day: 14.81 megabytes

Analog

- Reqs search term
- 461souris
- 374chauve
- 64chauves
- 34bourges
- 33les
- 21bats
- 21museum
- 21de
- 16la
- 16naturelle

Analog

- Reqs %bytes extension
- 10095 55.52% .jpg [JPEG graphics]
- 7580 21.00% .htm [Hypertext Markup Language]
- 17051 10.81% .gif [GIF graphics]
- 627 5.44% .class [Java class files]
- 1403 3.24% .html [Hypertext Markup Language]
- 1101 1.54% [directories]
- 26 0.90% .pdf [Adobe Portable Document Format]
- 2180 0.82% .js [JavaScript code]

Filtrage

- File bourges.txt filtered into bourges_FRF_jpg.txt

- Number of lines after filter: 8108

Génération de session

- % Sessions générées pour bourges_FRF_jpg.txt
- %
- % Il y a 1122 sessions

- Données des sessions exportées dans bourges_FRF_jpg.arff

Statistiques

- % Pages les plus visitées:
- % 206 : /robots.txt
- % 67 : /favicon.ico
- % 20 : /anglais/index_gb.htm
- % 20 : /teletypeBeanInfo.class
- % 9 : /actu/nouv%20espec/nouv%20espec.html
- % 9 : /actu/nouv%20espec/
- % 5 : /actu/photo/photo.html
- % 3 : /actu/aout/aout.html

Statistiques

- % Adresses ip ayant le plus de sessions :
- % 59 : zener.grpleg.net
- % 54 : aclermont-ferrand-101-1-2-56.w193-252.abo.wanadoo.fr
- % 51 : dyn-83-154-139-120.ppp.tiscali.fr
- % 50 : f02v-4-139.d3.club-internet.fr
- % 40 : arouen-102-2-1-71.w81-249.abo.wanadoo.fr
- % 40 : lns-p19-10-82-65-178-29.adsl.proxad.net
- % 39 : mix-bayonne-105-3-229.w193-249.abo.wanadoo.fr
- % 39 : mix-montpellier-203-1-22.w193-249.abo.wanadoo.fr
- %

Visites

- s:/jpeg/english.jpg d:Sun Sep 05 06:57:53 CEST 2004 | s:/actu/madagascar/ico_roussettes.jpg d:Sun Sep 05 06:57:54 CEST 2004
- s:/actu/veste/ico_veste.jpg d:Sun Sep 05 06:57:54 CEST 2004 | s:/actu/girouette/ico_faitage.jpg d:Sun Sep 05 06:57:54 CEST 2004
- s:/actu/nouv_espec/ico_microsc.jpg d:Sun Sep 05 06:57:54 CEST 2004
- s:/actu/veste/veste.jpg d:Sun Sep 05 06:58:24 CEST 2004
- s:/actu/veste/veste.jpg d:Sun Sep 05 06:59:50 CEST 2004
- s:/actu/veste/veste.jpg d:Sun Sep 05 07:00:48 CEST 2004

Fichier Weka

- @attribute user_ip string
- @attribute user_id string
- @attribute userAgent string
- @attribute /jpeg/vespbecht.jpg {0,1}
- @attribute /jpeg/english.jpg {0,1}
- …
- @data
- c-24-1-196-124.client.comcast.net, -, Mozilla/4.0 (compatible…), 1, 0, 0, 0, 0, 0, 0, 0, …, 0
- …

Avec export seulement int

- 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0
- 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Clustering - Weka

bourges_FRF_jpg2 :

- 0 1009 (90%)
- 1 113 (10%)

bourges_FRS_200 : Clustered instances: 1428

- 0 616 (84%)
- 1 117 (16%)
- 2 3 (0%)