École nationale
de la statistique
et de l'analyse
de l'information

Yann BUSNEL
Tel. : +33 (0)2 99 05 33 30
Fax : +33 (0)2 99 05 32 05
yann.busnel@ensai.fr

IRESI — INTRODUCTION AUX RESEAUX INFORMATIQUE
L3RI — SUJET DE PROJET 2015

Octobre 2015

Présentation du projet

Votre projet consiste à programmer et tester analytiquement un algorithme présenté pendant le 3e cours et de l'implémenter dans le langage qui vous convient. Vous pourrez retrouver ce cours à l'adresse suivante :
http://www.ensai.fr/files/_media/images/Enseignants%20chercheurs%20-%20doctorants/ybusnel%20-%20images/iresi/CM3_Data_streams-1.pdf

L'article de recherche correspondant est également disponible au téléchargement, à l'adresse suivante :
http://www.ensai.fr/files/_media/images/Enseignants%20chercheurs%20-%20doctorants/ybusnel%20-%20images/iresi/Codeviance_paper.pdf

Vous devrez donc implémenter cet algorithme, ainsi que la manière de lui donner des paires de flux en entrée, pour en sortir la valeur de codeviance approchée.

Vous devez ensuite l'exécuter sur chaque paire possible des traces réelles décrites ci-dessous, que vous retrouverez aux adresses suivantes :
http://www.ensai.fr/files/_media/images/Enseignants%20chercheurs%20-%20doctorants/ybusnel%20-%20images/iresi/traces.zip

Ces traces sont relativement anciennes mais elles ont le bon goût d'être relativement "petites" (entre 28 000 et 750 000 requêtes). Elles sont tirées de Internet Traffic Archive (http://ita.ee.lbl.gov/html/traces.html). Attention, il faudra nécessairement faire un travail d'extraction des informations pertinentes. Ici, nous nous intéressons, en fonction de la trace, soit aux adresses IP présentes soit aux noms des fichiers demandés, servant à chaque fois d'identifiant des items.

Etant donné que vous possédez la trace, vous pouvez donc l'analyser statiquement dans un

premier temps. Vous calculerez ainsi le résultat exact de codeviance que doit rendre votre algorithme (il sera bon d'extraire au préalable d'autres statistiques telles que le nombre d'items distincts, la distribution de fréquences des items, *etc*.).

Cet algorithme étant fondés sur l'aléatoire, vous devez itérer chaque simulation plusieurs fois et en extraire la moyenne et l'écart type de vos résultats.

S'il vous reste du temps, vous pourrez pousser l'analyse numérique en ajoutant un générateur de flux d'entrée, prenant en paramètre une distribution de probabilité, une taille de flux et un nombre d'élément distincts maximum, puis générant un flux correspondant. Ces derniers pourront ainsi être donnés en entrée de votre algorithme, permettant d'enrichir l'analyse.

Vous devez rendre pour le **17 novembre 2015** un rapport qui reprendra votre analyse du problème, les difficultés rencontrées lors du développement de votre simulateur ad-hoc et les résultats expérimentaux tirés de ces simulations. La soutenance de ce rapport aura lieu le jeudi **19 novembre 2015, à 10h15**, en Salle i53.

Nous aurons une séance de TP le **5 novembre 2015**, laquelle vous permettra de me poser les questions qui vous tarauderont ainsi que de résoudre des blocages potentiels.

Bon travail !

Description des traces réelles

EPA-HTTP

Description

This trace contains a day's worth of all HTTP requests to the EPA WWW server located at Research Triangle Park, NC.

Format

The logs are an ASCII file with one line per request, with the following columns:

1. **host** making the request. A hostname when possible, otherwise the Internet address if the name could not be looked up.
2. **date** in the format "[DD:HH:MM:SS]", where **DD** is either "29" or "30" for August 29 or August 30, respectively, and **HH:MM:SS** is the time of day using a 24-hour clock. Times are EDT (four hours behind GMT).
3. **request** given in quotes.
4. **HTTP reply code**.
5. **bytes in the reply**.

Measurement

The logs were collected from 23:53:25 EDT on Tuesday, August 29 1995 through 23:53:07 on Wednesday, August 30 1995, a total of 24 hours. There were 47,748 total requests, 46,014 **GET** requests, 1,622 **POST**requests, 107 **HEAD** requests, and 6 invalid requests. Timestamps have one-second precision. The WWW server software used was not recorded.

Privacy

The logs fully preserve the originating host and HTTP request. Please do not however attempt any analysis beyond general traffic patterns.

Acknowledgements

The logs were collected by Laura Bottomley (*laurab@ee.duke.edu*) of Duke University. Please include a corresponding acknowledgement in publications analyzing the logs.

Restrictions

The trace may be freely redistributed.

SDSC-HTTP (1 et 2)

Description

This trace contains a day's worth of all HTTP requests to the SDSC WWW server located at the San Diego Supercomputer Center in San Diego, California.

Format

The logs are an ASCII file with one line per request, with the following columns:

1. **host** making the request. Hosts are identified as **N+H:**, where **N** is a number identifying the hosts's network, and **H** is a number identifying the host within that network.
2. **timestamp** in the format "DAY MON DD HH:MM:SS YYYY", where **DAY** is the day of the week, **MON** is the name of the month, **DD** is the day of the month, **HH:MM:SS** is the time of day using a 24-hour clock, and **YYYY** is the year. Times are Pacific Daylight Time.
3. **filename** of the requested item. Empty for error conditions, non-empty for normal transactions. If non-empty it always starts with '/'.
4. **operation** performed by the server.
5. Successful operations always start with "Sent" and are terminated with a ':'. These include: "cache search", "range search", "search results", "WAIS search", "grep search", "text", "range", "binary", "cache", "exec binary", "exec", "HTTP/1.0 header", "CGI output", "CGI output [nph]", and "CGI redirection".
6. Failed operations do not start with "Sent" (they still are terminated with ':'). The **filename** associated with them is always empty. There are many failure modes and we do not list them here.
7. **remainder** of the transaction log. This is either the HTTP request, or information associated with a failed operation.

Routines to parse the logs are available from *webmaster@sdsc.edu*.

Measurement

The logs were collected from 00:00:00 PDT through 23:59:41 PDT on Tuesday, August 22 1995, a total of 24 hours. There were 28,338 requests and no known losses. Timestamps have 1 second resolution. Note that the server used was the GN server, not the more common CERN or NCSA server.

Privacy

The sites making requests to the server have had their addresses renumbered to preserve privacy. This was done by first splitting the address into a network number and a host number, and then renumbering each. Networks were numbered starting from 1 for the first network encountered in the trace. Hosts were numbered starting from 1 for the first host encountered in the trace for each network.

Acknowledgements

The logs were collected by Joshua Polterock (*webmaster@sdsc.edu*), Hans-Werner Braun (*hwb@sdsc.edu*), and K Claffy (*kc@sdsc.edu*), all of the San Diego Supercomputer Center. Please include a corresponding acknowledgement in publications analyzing the logs.

Restrictions

The trace may be freely redistributed

Calgary-HTTP

Description

This trace contains approximately one year's worth of all HTTP requests to the University of Calgary's Department of Computer Science WWW server located at Calgary, Alberta, Canada.

Format

The logs are an ASCII file with one line per request, with the following columns:

1. **host** making the request. Hosts are identified as either **local** or **remote** where **local** is a host from the University of Calgary, and **remote** is a host from outside of the University of Calgary domain.
2. **timestamp** in the format "DAY MON DD HH:MM:SS YYYY", where **DAY** is the day of the week, **MON** is the name of the month, **DD** is the day of the month, **HH:MM:SS** is the time of day using a 24-hour clock, and **YYYY** is the year. The timezone is -0700 between 30/Oct/1994:01:30:57 and 02/Apr/1995:03:03:26. For all other requests, the timezone is -0600.
3. **filename** of the requested item. Paths have been removed. Modified filenames consist of two parts: **num.type** , where **num** is a unique integer identifier, and **type** is the extension of the requested file.
4. **HTTP reply code**.
5. **bytes in the reply**.

Measurement

The logs were collected from October 24, 1994 through October 11, 1995, a total of 353 days. There were 726,739 requests. Timestamps have 1 second resolution.

Privacy

The sites making requests to the server have had their addresses removed to preserve privacy. Paths have been removed. Files were numbered from 1 for the first file encountered in the trace. Files retain the original file extension, so that the type of file can be determined.

Acknowledgements

The logs were collected by Robert Fridman of the University of Calgary Department of Computer Science, and contributed by Martin Arlitt (*mfa126@cs.usask.ca*) and Carey Williamson (*carey@cs.usask.ca*) of the University of Saskatchewan.

Publications

This is one of six data sets analyzed in an upcoming paper by
M. Arlitt and C. Williamson, entitled ``Web Server Workload Characterization: The Search for Invariants'', to appear in the proceedings of the *1996 ACM SIGMETRICS Conference on the Measurement and Modeling of Computer Systems*, Philadelphia, PA, May 23-26, 1996. An *extended version* of this paper is available on-line; see also the *DISCUS home page* and the group's *publications*.

Restrictions

The trace may be freely redistributed.