An Approach to Convert Unprocessed Weblogs to Database Table

Kiruthika M, Dipa Dixit, Pranay Suresh, Rishi M

Department of Computer Engineering, Fr. CRIT, Vashi, Navi Mumbai

Abstract

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools to find the desired information and to track and analyze their usage patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge. Web mining is the area which deals with mining data from web. Web Mining includes mining Web linkage structures, web contents and web access patterns. Web usage mining is nothing but mining Weblog records to find useful patterns which may help in extracting information about users/customers. In this paper, we discuss about a simplest approach of how the logs from a Webserver can be retrieved and can be viewed. Also, how these weblogs can be converted into database table .Now any required information can be mined for further analysis by querying the database table?

1. Introduction

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior. The general pattern tracking analyzes the weblogs to understand access patterns and trends. This analysis can shed light on better structure and grouping of resource providers. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling advertisements.

2. Tools for collection and analysis of Weblogs

2.1 Weblog Expert and Log Viewer 1.1

Weblog Expert automates the entire process of analyzing log files by scheduling and downloading raw or compressed log files and it allows processing Weblogs in Apache or IIS formats. Weblog Expert takes care of routine tasks associated with downloading. Built-in scheduler makes it possible to set the log analyzer to download the logs periodically, automatically retrieving the logs from a Web server via HTTP and FTP protocols. Recognizing raw, ZIP and GZ compressed Weblogs. The log analyzer can process logs in Apache and IIS formats. This gives the logs in the text format and this can be viewed.
Log Viewer 1.1 reads and displays any type of web server log in a convenient tabular format. It enables date filters, regular expressions search, logs monitoring, reports generation and more. It can read very large log files and enables indexing and very quick searches of IP's and URL's. The Log Reader helps to read extremely large files without consuming memory. Once the log files are parsed by the log viewer, the user can create aggregation reports; define log monitoring that can send alerts via email, SNMP, JMS messages and more.

2.2 Virtual PC and WinProxy

To collect weblogs in your PC, another tool namely virtual PC combined with Win Proxy can be used. The following steps can be followed to build the setup. Initially, a set up for Virtual PC can be activated which is supported by Windows XP. Then, WinProxy software can be loaded on to the Virtual PC. WinProxy software will act as a server and our own PC as a client. Request for URL given by user through client will get recorded in form of weblogs by WinProxy server. Thus, Weblog records are collected on server and these can be used for further analysis.

3. An approach to convert unprocessed weblogs to a database table.

Algorithm:
Log into mySQL command line client.
Create a mySQL database.
Create a table into the database with column definitions and primary key.
To import the log file,
Upload the log file into the home directory, so that the location of logfile is '/logfile.txt'.
Load this infile on the LOCAL file system into mySQL table by treating spacebar character as column delimiter and '\n' character as line delimiter.

4. Unprocessed weblogs

Below we have given a sample of 25 records out of 400 from a weblog file. This weblog file with the name “log.txt” has been taken as input.

#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer)
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=323-21-603&q=3&y=2002&id=671 200 156 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21) http://www.cs.depaul.edu/courses/syllabilist.asp depaul.edu/courses/syllabilist.asp
2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/default.asp - 200 16 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/news/news.asp?theid=573
2002-04-01 00:00:29 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) -
2002-04-01 00:00:29 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/default.asp - 200 94 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:30 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 62 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) -
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 63 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:32 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /resources/ug_scholarships.asp section=advising 200 15 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/advising/
2002-04-01 00:00:34 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-96-302&q=3&y=2002&id=576 200 171 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:00:35 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabisearch.asp - 200 125 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/courses/
2002-04-01 00:00:36 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/default.asp - 200 31 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:40 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/nsf_scholarships.asp - 200 625 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/resources/ug_scholarships.asp?section=advising

2002-04-01 00:00:44 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/2002/gradse2002.asp - 200 171 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/programs/
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /resources/gae_guide.asp |-|0|404_object_not_found 302 0 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+98) http://google.yahoo.com/bin/query?p=%22semantic+object+model%22+differences+%22entity-relationship+model%22+database+design&hc=0&hs=0
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /shared/404.asp 404;http://www.cs.depaul.edu/resources/gae_guide.asp 200 31 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+98) http://google.yahoo.com/bin/query?p=%22semantic+object+model%22+differences+%22entity-relationship+model%22+database+design&hc=0&hs=0
2002-04-01 00:01:07 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/courses.asp depcode=96&deptmne=se&courseid=550 200 140 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/programs/2002/gradse2002.asp
2002-04-01 00:01:09 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 719 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/
2002-04-01 00:01:14 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /people/facultyinfo.asp id=210 200 672 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/programs/courses.asp?depcode=96&deptmne=se&courseid=550
2002-04-01 00:01:15 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 post /cti/advising/display.asp - 200 1625 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+aol+7.0;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/cti/advising/display.asp?subpage=form&edit=yes&course=336-21-901&q=3&y=2002&id=322
2002-04-01 00:01:20 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-96-303&q=3&y=2002&id=576 200 172 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:01:31 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=313-94-601&q=3&y=2002&id=618 200 109 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:01:36 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 719 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/

5. Implementation

The above algorithm was implemented using MySQL and conversion was done. Output information was verified with input weblog file and found correct. By querying the database, required attributes can also be selected i.e., Attributes which are of interest for extracting information can be targeted and selected. The above approach was tested with a weblog file of approximately 400 records.

A sample of unprocessed weblogs from the input weblog file (log.txt) is already shown.

5.1 Steps for transferring text file into database in MySQL

Enter password: *********
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 27

```
Server version: 5.0.41-community-nt MySQL
Community Edition (GPL)
Type 'help;' or '\h' for help. Type '\c' to clear the
buffer.

mysql> use weblogdb;
Database changed
mysql> show tables;
+--------------------+
| Tables_in_weblogdb |
+--------------------+
| weblogtb1          |
+--------------------+
1 row in set (0.46 sec)

mysql> load data local infile '/log.txt'
    -> into table weblogtb1
    -> fields terminated by ' '
    -> lines terminated by '\n'
    ->

(date,time,cip,csuname,ssitename,scompname,sip,
sport,csmethod,csuristm,csuriquery,scstatus,timeta
ken,csversion,cshost,csUagent,csref);

Query OK, 432 rows affected, 17 warnings (0.58
sec)

Records: 432   Deleted: 0  Skipped: 0  Warnings:
17
```

Now all the records are entered into database from where retrieval can be done.

6. Results

The results are shown in two parts:
Database table showing all the attributes is shown in Table no:1.

This table shows all the fields and their content in a tabular format with respect to the input weblog file which was in text format. (ie;) the whole information about weblogs are now transferred in a readable format/table/database.
Retrieval of relevant attributes are shown in Table no:2.
This retrieval helps in further analysis of weblogs to extract any information related to the user/session etc.

7. Conclusion

The approach discussed in this paper can be utilized for applications where very few attributes of weblogs are of interest. In such cases, pre-processing of weblogs can be avoided. This decision is totally dependent on what type of application pattern or knowledge a person is looking for from weblogs. Eventhough few steps of preprocessing are avoided, this approach can help in extracting enough information about users/sessions.

8. References

[1] Margaret H Dunham, "*Data mining introductory and advanced topics*" 5th Ed.
[2] Jiawei Han and Micheline Kamber, "*Data mining: concepts and techniques*".
[3] "*Mining access patterns efficiently from weblogs*" by Jian Pei,Jiawi, HanBehzad Mortazavi-asl and Hua Zhu Simon Fraser University,Canada.
[4] "*Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm*" by Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan Mohamad Mohsin.
[5] "*Research on Path Completion Technique in Web Usage Mining*" by Yan LIa,b, Boqin FENGa, Qinjiao MAOa aSchool of Electronics and Information Engineering, Xi' an Jiaotong University, Shaanxi, China bSchool of Computer Science and Engineering, Xi' an University of Technology, Shaanxi, China.
[6] "*Applying Web Usage Mining Techniques to Discover Potential Browsing Problems of Users*" by I-Hsien Ting, Chris Kimble,Daniel Kudenko Department of Computer Science, the University of York, Seventh IEEE International Conference on Advanced Learning Technologies,2007

Table No1:

Date	Time	cip	csUname	s-sitename	s-compname	sip	s-port	cs-method
2002-04-01	12/30/1899 12:0...	1cust62.tnt40.chi...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	ac9781e5.ipt.aol....	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	alpha1.csd.uwm....	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	12-250-96-248.cli...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	alpha1.csd.uwm....	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	ac9781e5.ipt.aol....	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	chf-il11-202.rasse...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	12-250-96-248.cli...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	ac9781e5.ipt.aol....	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	66-79-37-44.coas...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	66-79-37-44.coas...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	chf-il11-202.rasse...	-	w3svc3	bach	bach.cs.depaul.e...	80	get
2002-04-01	12/30/1899 12:0...	w010.z06422106...	-	w3svc3	bach	bach.cs.depaul.e...	80	get

cs-uri-stm	cs-uri-query	sc-status	Time-taken	cs-ver	cs-host	csUagent	csRef
/courses/syllabu...	course=323-21-6...	200	156	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/advising/default...	-	200	16	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/default.asp	-	302	0	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	-
/courses/default....	-	200	94	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/default.asp	-	302	0	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	-
/news/default.asp	-	200	62	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	-
/news/default.asp	-	200	63	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	-
/resources/ug_s...	section=advising	200	15	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/courses/syllabu...	course=468-96-3...	200	171	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/courses/syllabis...	-	200	125	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/programs/defaul...	-	200	31	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/advising/nsf_sc...	-	200	625	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/programs/2002/...	-	200	171	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/resources/gae_...	l-l0l404_object_n...	302	0	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://google.yah...
/shared/404.asp	404;http://www....	200	31	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://google.yah...
/programs/cours...	depcode=96&de...	200	140	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/courses/syllabili...	-	200	719	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/people/facultyin...	id=210	200	672	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...
/cti/advising/dis...	-	200	1625	http/1.1	www.cs.depaul.e...	mozilla/4.0+(com...	http://www.cs.d...

Table No 2:

Date	Time	cip	cs-uri-stm	csRef
2002-04-01	12/30/1899 12:0...	1cust62.tnt40.chi...	/courses/syllabu...	http://www.cs.d...
2002-04-01	12/30/1899 12:0...	ac9781e5.ipt.aol....	/advising/default...	http://www.cs.d...
2002-04-01	12/30/1899 12:0...	alpha1.csd.uwm....	/default.asp	-
2002-04-01	12/30/1899 12:0...	12-250-96-248.cli...	/courses/default....	http://www.cs.d...
2002-04-01	12/30/1899 12:0...	w010.z06422106...	/default.asp	-
2002-04-01	12/30/1899 12:0...	alpha1.csd.uwm....	/news/default.asp	-
2002-04-01	12/30/1899 12:0...	w010.z06422106...	/news/default.asp	-
2002-04-01	12/30/1899 12:0...	ac9781e5.ipt.aol....	/resources/ug_s...	http://www.cs.d...
2002-04-01	12/30/1899 12:0...	chf-il11-202.rasse...	/courses/syllabu...	http://www.cs.d...