Indirect Positive and Negative Association Rules in Web Usage Mining

Dhaval Patel
Department of Computer Engineering,
Dharamsinh Desai University
Nadiad, Gujarat, India

Malay Bhatt
Department of Computer Engineering,
Dharamsinh Desai University
Nadiad, Gujarat, India

ABSTRACT

One of the purposes of Web usage mining is to find out interesting user association rules from web server logs. It has become vital for personalization, effective web site management, business and support services, creating adaptive web sites, and so on. In the web domain, items correspond to pages and transactions to user sessions. Indirect associations, type of infrequent pattern provide useful insight into the data. The concept of indirect association is to indirectly connect two rarely co-occurred pages via a third page called transitive page. Mining positive and negative association rules in web usage data become a hot spot. So, these all information leads find new approach for discovering efficient rules for web. Indirect positive and Negative Association Rules are discussed here which can be used for Web Recommendation, personalization etc. Mining indirect positive and negative association rules for the web is explored very little so far in the research work. The presented Proposed Approach of algorithm is to extract the positive and negative association rules from web session log file and then discover the indirect positive and negative association rules from it.

Keywords

Association Rules, Indirect Association Rules, Positive and Negative Association Rules, Preprocessing, Web Usage Mining

1. INTRODUCTION

Internet is one of the fastest and necessary areas of data and information gathering. But different types of data have to be managed and organized for accessing efficiently. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web [1,4]. Web mining contains a wide range of applications that aim at discovering and extracting hidden information in data stored on the Web. There are different important purposes of Web mining like, it is to provide a mechanism to make the data access more efficiently and adequately [1,5]. The information about the activities of users is stored in log files which are derived and discovered by the web mining.

Web data mining is separated into three fields: Web Content Mining, Web Structure Mining, and Web Usage Mining [4]. Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs. The purpose of WUM is to reveal the knowledge hidden in the log files of a web server to improve web based applications [4, 5]. Web log file is a simple plain text file which record information about each user from web server. Presentation of log files data in three different formats. Out of them NCSA log format file [11] is used here.

2. RELATED WORK:

PriyankaPatilet al. [3] focused on data preprocessing stages and log file description. It presents algorithm for field extraction and data cleaning. C.p. sumathiet al. [7] describes about the preprocessing and its different various steps to preprocess the web server log file. They explained the format of web server log file and after it techniques which are used to preprocess it. MohdHelmyAbd Wahab et al. [6] presents the pre-processing techniques and one Algorithm for Reading Server Logs and other for transfer to database using Active server pages and dll on IIS Web Server Logs.

XiufengPiao et al. [2] proposed the algorithm which based on the correlation and dual confidence, can mine the positive and negative association rules. After pruning rules with dual confidence result shows that given algorithm can reduce the scale of meaningless association rules, and mine interesting negative association rules. Anuradhaveleti et al.[9]proposed an incremental algorithm (IPNAR) that mines positive and negative association rules in web usage data. Given algorithm incrementally updates web log association rules by using the metadata of old database transactions as well as old mined rules. So, No need of any multiple scan of dataset there. It performs single scan over the dataset that's why it overcomes the limitations of other mining methods. As described by He Jiang et al.[10] the efficient mining of multiple-level association rule is proposed to resolve the question about the can't mining effective threshold rules due to different frequency of every item. It has high potential to produce rare but informative item rules. Algorithm based on multiple minimum supports is designed simultaneously.

Przemysław kazienkoetal.[8] describes indirect association rules and how to get them using Indirect Association Rule Mining (IDARM) algorithm. Association rules gives relationship between WebPages derived from user behavior which can be used in recommender system. The main intension of recommendation is to suggest to user about web pages that appear to be useful. Two types of indirect associations rules are described in the paper: partial indirect associations and complete indirect association. IDARM algorithm and example given in this paper are used to find indirect association rules in our proposed approach work.

Moreover, such rules may avoid some relationships between pages, which do not occur together in the same user sessions. This concerns especially pages not being connected directly with hyperlinks (Fig. 1). To search important relationships between pages that rarely occur in common sessions but are simultaneously close to other pages (Fig. 1), new patterns indirect association rules are suggested in this paper. Third page can be considered as "indirectly associated" when Two pages, which individually co-occur relatively frequently in sessions with another.

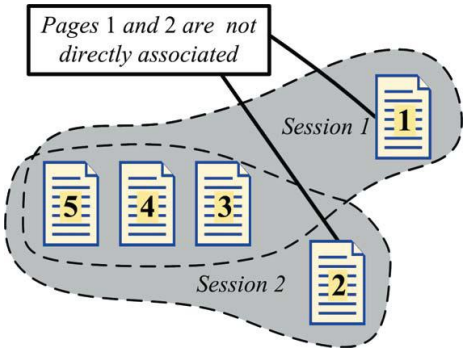

Fig. 1 Sessions with two documents (1 and 2), which are associated only indirectly.[8]

3. PROPOSED WORK:

Analyzing the above existing work, one thing can be proposed that is fusion of Indirect Positive Rules with Negative Association Rules. Block diagram of proposed work is as described below in fig 2.

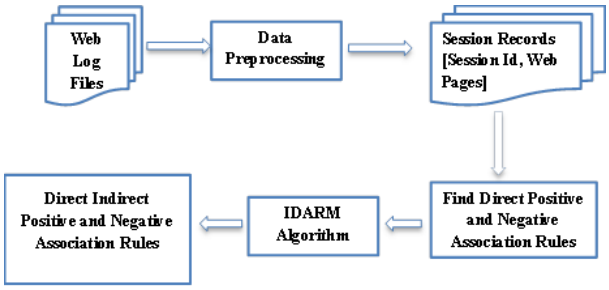

Fig. 2 Proposed Approach – Indirect Positive and Negative Association Rules.

Each Component of above block diagram is described as following.

3.1 Web Log files:

Web log file is a simple plain text file which record information about each user. Presentation of log files data in three different format. W3C Extended log file format, NCSA common log file format and IIS log file format. We are going to use NCSA common log file format [11] as shown in fig.3.

```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
128.159.129.170 - - [01/Aug/1995:00:13:16 -0400] "GET / HTTP/1.0" 200 7280
kgtyk4.kj.yamagata-u.ac.jp - - [01/Aug/1995:00:00:17 -0400] "GET / HTTP/1.0" 200 7280
rpgopher.aist.go.jp - - [01/Aug/1995:00:01:58 -0400] "GET /ksc.html HTTP/1.0" 200 7280
204.248.155.42 - - [01/Aug/1995:00:05:32 -0400] "GET /icons/menu.xbm HTTP/1.0" 200 527
204.248.155.42 - - [01/Aug/1995:00:05:32 -0400] "GET /icons/image.xbm HTTP/1.0" 200 505
143.158.26.50 - - [01/Aug/1995:00:14:07 -0400] "GET / HTTP/1.0" 200 7280
ai.asu.edu - - [01/Aug/1995:00:03:55 -0400] "GET /facts/faq01.html HTTP/1.0" 200 19320
gw1.att.com - - [01/Aug/1995:00:03:56 -0400] "GET /icons/menu.xbm HTTP/1.0" 304 0
gw1.att.com - - [01/Aug/1995:00:03:56 -0400] "GET /icons/text.xbm HTTP/1.0" 304 0
143.158.26.50 - - [01/Aug/1995:00:14:07 -0400] "GET / HTTP/1.0" 200 7280
```

Fig. 3Web servers Log file of NCSA format

3.2 Preprocessing:

It consists of data field extraction and data storage, data cleaning and session identification. It eliminates unnecessary records and validates the important records that are saved into the database. Data field extraction is the process of separating out various data fields from single server. After that those data will be stored in database. Data cleaning consists of removing all the data tracked in web logs that are useless and unimportant for mining purposes. *Fig.4* shows the clean data of given web log file. Session identification identified the session of different users as per the usage of website by user.*Fig.5*describes the session identification and session list of web server log file respectively.*Fig.6* and *Fig.7* show the mapping of path to another name and mapping path session wise of data of fig.4 respectively.

Id	Session_Id	site	name
1	1	/	d1
2	1	/shuttle/missions/sts-71/mission-sts-71.html	d2
3	1	/shuttle/resources/orbiters/atlantis.html	d3
4	1	/shuttle/resources/orbiters/challenger.html	d4
5	1	/history/apollo/apollo-17/apollo-17.html	d5
6	1	/shuttle/missions/sts-71/images/images.html	d6
7	1	/shuttle/missions/sts-68/ksc-srl-image.html	d7
8	1	/shuttle/missions/sts-66/mission-sts-66.html	d8
9	1	/shuttle/missions/sts-63/mission-sts-63.html	d9
10	1	/shuttle/resources/orbiters/discovery.html	d10
11	1	/shuttle/resources/orbiters/endeavour.html	d11
12	1	/shuttle/missions/sts-70/mission-sts-70.html	d12
13	1	/facilities/lc39a.html	d13
14	1	/shuttle/missions/sts-26/mission-sts-26.html	d14

Fig.6 Mapping of Path name

	Id	path
1	1	d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,d11,d12,d13,d14,d15,d16
2	2	d17,d18,d19
3	3	d20,d11,d21,d22,d23,d23,d13,d24
4	8	d27,d28,d29,d30
5	9	d31,d24,d32
6	10	d1,d2,d26,d33,d30
7	11	d31,d24,d32
8	12	d13,d34,d2,d15,d35,d26,d36,d37,d38,d39,d40,d41,d33
9	15	d43,d44,d45,d46
10	16	d35,d26,d47,d48,d49,d50,d51,d12,d10,d11,d52,d53,d54
11	17	d22,d26,d26
12	18	d1,d31,d55,d24,d32,d56,d56,d57

Fig.7 Mapping of Session wise Path name from Fig.6

3.3 Direct Positive and Negative Association Rules [2, 9]:

Negative association rules A=>¬B (or ¬A=>B, ¬A=>¬B) are more important for e-commerce and reconstruction of web site. By removing unwanted page links using negative rules appropriate and accurate structure of website can be created. As following Fig.8 describe the method to find the positive and negative association rules.

Id	site	logname	fullname	Date_time	request	path	proto	status	length
2	uplherc.upl.com	-	-	01/Aug/1995:00:00:07	GET	/	HTTP/1.0	304	0
8	slppp6.intermind.net	-	-	01/Aug/1995:00:00:10	GET	/history/skylab/skylab.html	HTTP/1.0	200	1687
15	133.43.96.45	-	-	01/Aug/1995:00:00:16	GET	/shuttle/missions/sts-69/mission-sts-69.html	HTTP/1.0	200	10566
16	kgtyk4.kj.yamagata-u.ac.jp	-	-	01/Aug/1995:00:00:17	GET	/	HTTP/1.0	200	7280
18	d0ucr6.fnal.gov	-	-	01/Aug/1995:00:00:19	GET	/history/apollo/apollo-16/apollo-16.html	HTTP/1.0	200	2743
19	ix-esc-ca2-07.ix.netcom.com	-	-	01/Aug/1995:00:00:19	GET	/shuttle/resources/orbiters/discovery.html	HTTP/1.0	200	6849
28	www-c8.proxy.aol.com	-	-	01/Aug/1995:00:00:24	GET	/shuttle/countdown/	HTTP/1.0	200	4324
32	slppp6.intermind.net	-	-	01/Aug/1995:00:00:32	GET	/history/skylab/skylab-1.html	HTTP/1.0	200	1659
37	uplherc.upl.com	-	-	01/Aug/1995:00:00:43	GET	/shuttle/missions/sts-71/mission-sts-71.html	HTTP/1.0	200	13450
42	133.43.96.45	-	-	01/Aug/1995:00:00:46	GET	/shuttle/resources/orbiters/endeavour.html	HTTP/1.0	200	6168

Fig. 4Cleaned Data of web log

Id	Session_Id	Path
1	1	/, /shuttle/missions/sts-71/mission-sts-71.html, /shuttle/resources/orbiters/atlantis.html, /shuttle/resources/orbiters/challenger.html,
2	2	/history/skylab/skylab.html, /history/skylab/skylab-1.html, /history/skylab/skylab-2.html
3	3	/shuttle/missions/sts-69/mission-sts-69.html, /shuttle/resources/orbiters/endeavour.html, /shuttle/missions/sts-72/mission-sts-72.html
4	4	/
5	5	/history/apollo/apollo-16/apollo-16.html
6	6	/shuttle/resources/orbiters/discovery.html
7	7	/shuttle/countdown/
8	8	/cgi-bin/imagemap/countdown70?285,291, /shuttle/countdown/count.html, /images/
9	9	/history/history.html, /history/apollo/apollo.html, /history/apollo/apollo-13/apollo-13.html
10	10	/, /shuttle/missions/sts-71/mission-sts-71.html, /shuttle/countdown/, /shuttle/missions/sts-71/movies/movies.html, /images/

Fig.5 Session List of Log file data of fig.4

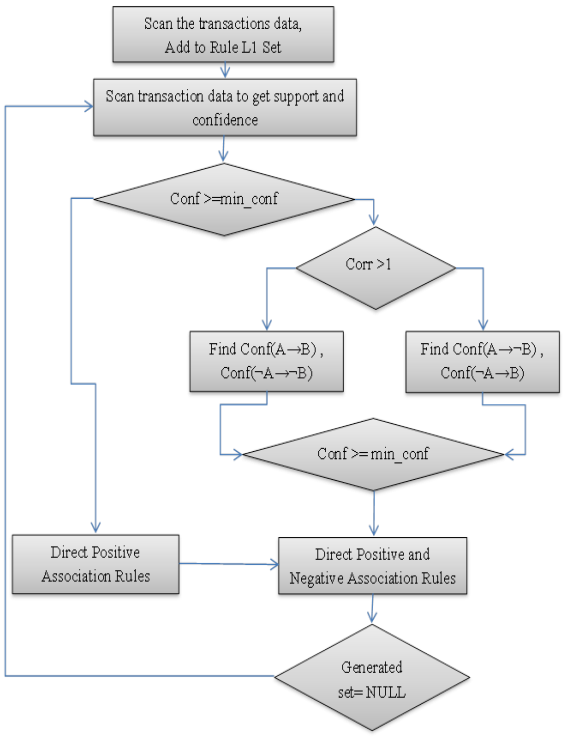

Fig.8 Algorithm for Positive and Negative Association Rules

For the direct association rule two measures are used: support and confidence are described. Here S^S: the set of all user sessions gathered by the system.S_i: i-th user session is the page set containing all pages viewed by the i^{th} user during the i^{th} visit on the web site. [8]

$Sup(X \rightarrow Y)$ =card $(S_i \in S^S : X \cup Y \subset Si\})$/card (S^S)

$Con(X \rightarrow Y)$ =card $(S_i \in S^S : X \cup Y \subset S_i\})$/card$(\{S_i \in S^S : X \subset S_i\}$
Using these measures positive and negative rules (R1 → R2) are found as shown in Fig.9.

	Id	R1	R2	Conf
1	1	d1	~d2	0.66
2	2	~d1	d2	0.75
3	3	~d1	~d3	1
4	4	d1	d4	0.5
5	5	d1	~d6	0.83
6	6	~d1	d6	0.75
7	7	d2	d1	0.4
8	8	d2	d4	1
9	9	d2	d5	0.4
10	10	d2	d6	0.4
11	11	d3	d1	1
12	12	d4	d1	0.42
13	13	d4	d2	0.71
14	14	d4	d5	0.42
15	15	d4	d6	0.42

Fig.9 Values of direct confidence for positive and negative rules of example sessions from Fig.7.

3.4 Indirect Association Rules:

Another approach to associations: indirect Association rules. Association rules gives relationship between Web Pages derived from user behavior which can be used in recommender system. The main intension of recommendation is to suggest to user about web pages that appear to be useful.

$di \rightarrow^{P\#} dj$, dk is the indirect relationship from di to dj with respect to dj, for which two direct association rulesexist:$di \rightarrow dk$ and $dk \rightarrow dj$ with $sup(di \rightarrow dk) \geq supmin$, $con(di \rightarrow dk) \geq conmin$ and $sup(dk \rightarrow dj) \geq supmin$, $con(dk \rightarrow dj) \geq conmin$, where $di, dj, dk \in D$[where di,dk,dj :independent web pages (documents) and D : web page domain, contains independent web pages];$di \neq dj \neq dk$. The page dk, in the partial indirect association rule $di \rightarrow^{P\#} dj$, dk, is called the *transitive page*s shown in Fig.10. [8]

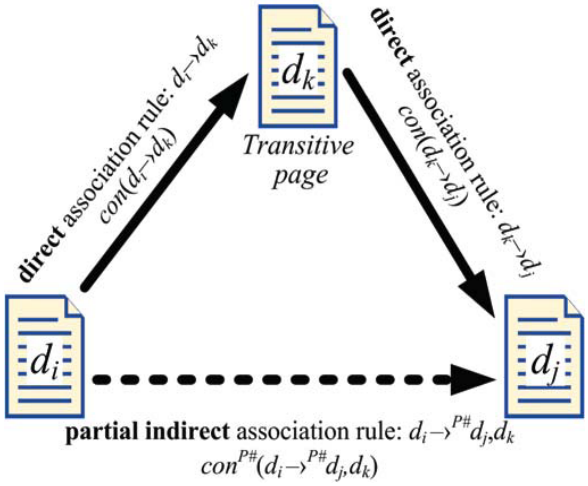

Fig.10 Indirect Association between two web pages [8]

As following, fig. 11 described the method for finding Indirect Positive and Negative Association Rules.

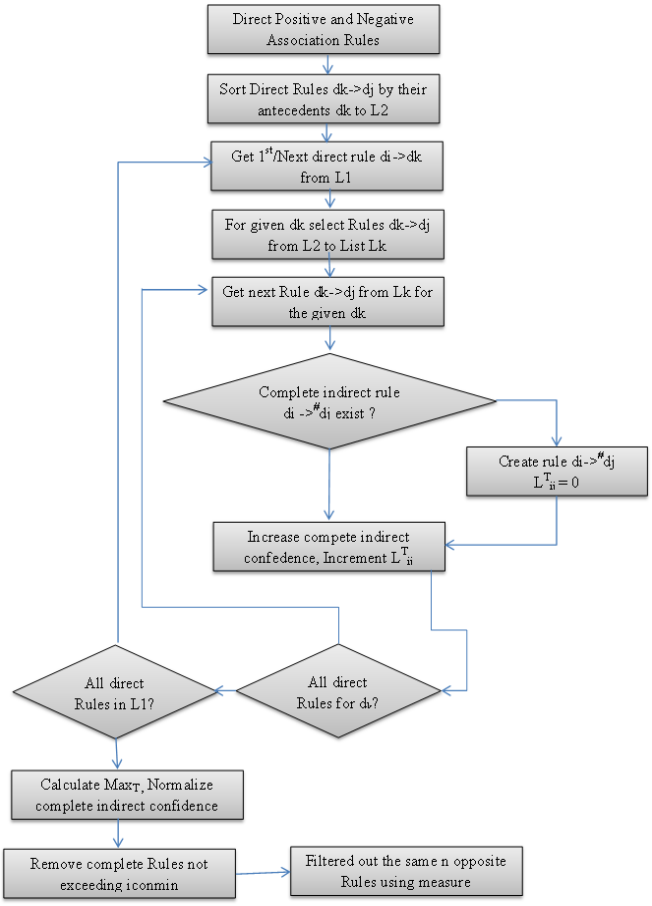

Fig.11Algorithm for indirect Association Rules (IDARM) [8]

3.5 Indirect Positive and Negative Association Rules

Indirect Positive and Negative Association Rules founded from previous Algorithm. Fig.12 Shows (R1 → #R2) the Indirect Positive and Negative Association Rules with their respective confidence.

Id	R1	R2	Count	Conf
1	~d1	d4	2	1.3125
2	~d1	d5	2	0.8625
3	~d1	d6	1	0.3
4	~d1	~d3	4	2.815
5	~d1	~d4	4	2.815
6	~d1	~d5	4	2.815
7	~d1	~d6	2	1.38
8	~d1	~d7	4	2.815
9	~d1	~d8	4	2.815
10	~d1	~d9	4	2.815
11	~d1	~d10	4	2.815
12	~d1	~d11	4	2.815
13	~d1	~d12	4	2.815
14	~d1	~d13	4	2.755
15	~d1	~d14	4	2.815

Fig.12Indirect Positive and Negative Association Rules from fig.9.

4. Experimental Result:

Data sets (logfiles) from the [13] repository are used in the experiments. Table 1 describes the analysis of result from different logfile. Fig.13 shows the graphical view of analysis result and log1, log2, log3, log4 are logfiles as shown in table 1 respectively.

Table 1 Analysis of Result

Algorithm		UofS_access _log	nasa_access _log_aug95	clarknet_acce ss_log_Sep	Clarknet_ access_aug
Records		54584	53300	52083	36185
IDARM		1744	1402	1668	1067
Proposed Approach	+ve Rules	1081	1140	1355	867
	-ve Rules	1046	968	1150	736
	Total	2127	2108	2505	1603

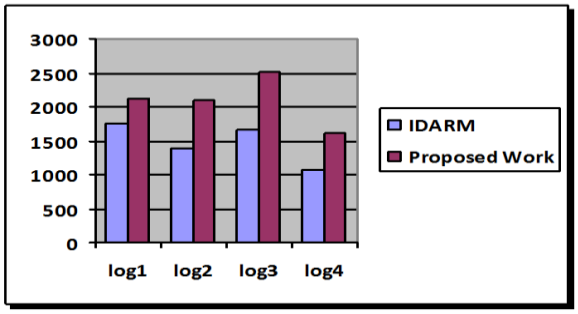

Fig. 13Analysis of Result

5. CONCLUSION:

We have proposed approach for finding Indirect Positive and Negative Association Rules in web usage mining .This Given Approach combines the Rules of IDARM algorithm and Positive and Negative Association Rules. We have verified the approach practically using various Rule Measures. It gets the accurate and more association rules then the other algorithms.

6. REFERENCES

[1] Brijendra Singh, Hemant Kumar Singh2010.WEB DATA MINING RESEARCH: A SURVEY. 978-1-4244-5967-4/10/$26.00 ©2010 IEEE

[2] XiufengPiao, Zhanlong Wang, Gang Liu2010. Research on Mining Positive and Negative Association Rules Based on Dual Confidence.Fifth International Conference on Internet Computing for Science and Engineering.978-0-7695-4339-0/10 $26.00 © 2010 IEEE.DOI 10.1109/ICICSE.2010.28

[3] PriyankaPatil,UjwalaPatil 2012. Preprocessing of web server log file for web mining.World Journal of Science and Technology 2012, 2(3):14-18ISSN: 2231 – 2587.

[4] ChintandeepKaur, RinkleRaniAggarwal 2012. "WEB MINING TASKS AND TYPES: A SURVEY ". IJRIM Volume 2, Issue 2 (February 2012) (ISSN 2231-4334).

[5] R. Kosala, H. Blockeel. 2000. "Web mining research: A survey," ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000

[6] MohdHelmyAbdWahab, MohdNorzali Haji Mohd, HafizulFahriHanafi, MohamadFarhanMohamadMohsin 2008. Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm.World Academy of Science, Engineering and Technology 24 2008.

[7] C.p. sumathi, R. Padmajavalli 2011.An Overview of Preprocessing OfWeb Log Files for Web Usage Mining.Journal of Theoretical and Applied Information Technology31st December 2011. Vol. 34 No.2.© 2005 - 2011 JATIT & LLS.

[8] PRZEMYSŁAW KAZIENKO 2009. Mining Indirect Association Rules for Web Recommendation. Int. J. Appl. Math. Comput. Sci., 2009, Vol. 19, No. 1, 165–186 DOI: 10.2478/v10006-009-0015-5

[9] Anuradhaveleti, T.Nagalakshmi 2011.Web Usage Mining: An Incremental Positive and Negative Association Rule Mining Approach. ISSN: 0975-9646 (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6), 2011, 2862-2866

[10] He Jiang, Yuanyuan Zhao, Chunhua Yang, Xiangjun Dong 2008. Mining both Positive and Negative Weighted Association Rules with Multiple Minimum Supports. International Conference on Computer Science and Software Engineering 2008.

[11] Apache NCSA Log File http://httpd.apache.org/docs/1.3/logs.html

[12] NASA Web server log File http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html

[13] Web server Logfiles

http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.htm