

Predicting the Behaviour and Interest of the Website Users through Web Log Analysis

Arvind K. Sharma
Ph.D. Scholar
Dept. of Comp. Sc. & Engg
Jaipur National University, Jaipur
Rajasthan-India

P.C. Gupta
Associate Professor
Dept. of Comp. Sc. & Informatics
University of Kota, Kota
Rajasthan-India

ABSTRACT

Web mining is a hot research area of many researchers. Web mining techniques have been widely used to discover interesting and frequent user navigation patterns from the web server logs. The aim of this work is to apply web mining techniques for discovering user's behaviour and interest for an educational institution website usage to reveal previously unknown interesting patterns extracted in order to recommend possible measures for further improvement of the Website. In this paper the web user access and server usage patterns have been analyzed and daily, weekly, monthly web metrics such as number of visits, pages, files, hits and sites have been investigated. An attempt has been made to predict the behaviour and interest of the website users.

General Terms

Web usage mining

Keywords

Web server logs, Web log analysis, AWStats

1. INTRODUCTION

As the size of World Wide Web increases along with number of users, it is very much essential for the website owners and developers to better understand their customers so that they can provide better services, and also enhance the quality of the Website. Web usage mining is the process of data mining technique. It is a process to extract useful and meaningful information from web server logs. It is an automatic discovery of patterns in click streams and associated data collected or generated as a result of user interactions with one or more web sites. Basically web mining[1] refers to overall process of discovering potentially useful and previously unknown information from web documents and services. Web mining could be viewed as an extension of standard data mining to web data. In recent years, web usage mining techniques have been widely used for discovering interesting and frequent user navigation patterns from web server logs. The World Wide Web (known as Web) has influenced a lot to both users that is visitors as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customers 24X7 hours[2].

Paper is organized in different sections: Section 2 explains the web server logs. Literature review is shown in Section 3. Proposed work and its methodology are described in section 4. Section 5 contains observations and results. Conclusion is

shown in section 6 while references are mentioned in the last section.

2. WEB SERVER LOGS

Web server logs are log files which are automatically created and maintained by a server. These are files to which the web server writes information, each time a user requests a Website from that particular server. If user visits several times on the website then it creates entry many times on the server. Web log data offers valuable information insight into web site usage. It represents the activity of many users over a potentially long period of time. A web server when properly configured, can record every click that users make on the Website. For each click in the visit path, the server adds to the log file information about user request. Web logs collect data on the server in the files of specific format [3]. Usually web logs contain data such as: client's IP address, URL of the page requested, time when the request was sent to the server etc. This data is used later as the basis of usage behaviour discovery.

3. LITERATURE REVIEW

Various works have been done by many researchers in the area of web usage mining. In one of the work a novel approach for classifying user navigation patterns was introduced to predict user's future request[4]. In another approach, data from a data warehouse and web data were used to improve marketing activities[5]. Ramya and Kavitha[6] proposed a complete preprocessing methodology for discovering patterns in usage mining to improve the quality of data by reducing the quantity of data. Maheswara Rao et al.[7] had introduced an extensive research frame work capable of preprocessing web log data completely and efficiently. The learning algorithm of proposed research framework could separate human user and search engine access intelligently with less time. The framework reduces the error rate and improves significant learning performance of the algorithm. This framework assists to investigate the user's usage behaviour. Valter Cumbi and Lourino Chemane[8] had given a case study of e-government portal initiative in Mozambique for visitor analysis.

4. PROPOSED METHODOLOGY

In this work web user access and server usage patterns have been analyzed from a web site's main server. The complete work is performed by using AWStats web mining tool[10] that is a popular web analyzer. The typical web log traffic patterns of the Web site are shown, by means of summary of eleven months almost one year from January 2012 to November, 2012 in terms of daily average like number of

visit, pages, files, and hits and monthly average i.e. number of visit, pages, files, hits and sites. This work will enable to understand the behaviour and interest of the users who frequently visit the web site. Today, understanding the interests of users is becoming a fundamental need for web sites owners in order to better serve their users by making adaptive the content and usage, structure of the web site to their preferences. The analysis of web log data permits to identify useful patterns of the browsing behaviour of users, which exploited in the process of navigational behaviour. Web log data captures web browsing behaviour of users from a web site. Academic institutions are good examples that develop website. One such institution of the education sector has been considered in our work. This work presents visitor pattern analysis, page view analysis and time analysis performed through web log data of an educational institution's web site. We have been performed various analyses through collected and preprocessed web server logs to predict the behaviour and interest of the users and determine the effectiveness of the website, including the following:

- Visitor Pattern Analysis
- Page View Analysis
- Time Analysis
- Origin of the Website Visitors
- Portions of the Website that are accessed
- Number of document downloads

In this work, AWStats reports undergo a time analysis and page view analysis. The time analysis looks at the different times of day, days of week, and days of month that the Website receives the most visitors. The page view analysis provides which website pages are most viewed by the visitors. The combination of both statistics will help to predict the characteristics of the web users.

4.1 Data Collection

In this work, the user access web log data has been collected from an educational institution web site's server located at <http://www.davkota.org>[9] that stores normally secondary data source in view of the fact that web log keeps every activity of the user regarding to visit of the Website. The web log data contains the information of eleven months almost one year period from January, 2012 to November 2012. During this period 14.59 GB data has been transferred for the complete web log analysis.

4.2 Data Selection

At present, towards Web Usage Mining technique, the main data origin has three kinds: Server-side data, Client-side data, and Proxy-side data (middle data).

In this work, we use the case of the Web server.

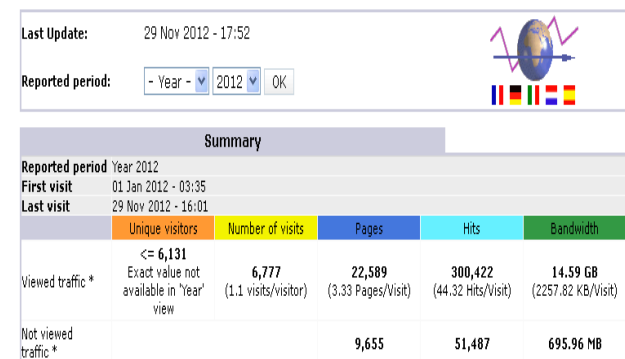
4.3 Tool for Experiment

There are several commercial and freely available tools for web usage mining purposes. In our work, we are going to use AWStats tool that is one of the fast and powerful web log analyzer[10]. It helps to reveal important statistics regarding a web site's usage such as activity of visitors, access statistics, paths through the website, visitors' browsers, etc. It supports W3C extended log format that is the default log format of Microsoft IIS 4.0/.05/6.0/7.0 and also the combined and common log formats of Apache web server. It reads compressed log files (.gz, .bz2 and .zip) and can automatically

detect the log file format. If necessary, log files can also be downloaded via FTP or HTTP. Through AWStats we have been identified visitor statistics like Total Hits, Visitors Hits, Average Hits per Day, Average Hits per Visitor, etc., Page View Analysis like Total Page views, Average Page Views per Day, Average Page Views per Visitor, total Visitors, Total Visitors, Average Visitors per Day, Total Unique IPs, Bandwidth, Total Bandwidth, Visitor Bandwidth, Average Bandwidth per Day, Average Bandwidth per Hit, and Average Bandwidth per Visitor of the Website on monthly and day of the week basis.

5. OBSERVATIONS AND RESULTS

In this work, we took the web server logs of an educational institution website. The website has many web pages with many images, graphics and texts. The pre-processing task has been accomplished firstly then the web log data has been applied through AWStats tool and different experiments have been performed. Our implementation is developed by using AWStats web analyzing tool. Different analyses have done to identify the user web watching behaviour and interests. Website's daily, hourly, weekly, and monthly web data patterns from Jan 2012 to Nov 2012 are presented through different charts, tables and figures respectively. During this period 300,422 Hits were recorded. The AWStats report is shown in figure 1.



Summary					
Reported period Year 2012					
First visit 01 Jan 2012 - 03:35					
Last visit 29 Nov 2012 - 16:01					
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	<= 6,131 Exact value not available in 'Year' view	6,777 (1.1 visits/visitor)	22,589 (3.33 Pages/Visit)	300,422 (44.32 Hits/Visit)	14.59 GB (2257.82 KB/Visit)
Not viewed traffic *			9,655	51,487	695.96 MB

Fig.1: AWStats Report

5.1 Visitor Pattern Analysis

In this work the reported period was taken over almost one year duration, the website's web server receives over 44.32 Hits per Visit and 2257.82 KB per visits bandwidth. The website's total monthly summary from January 2012 to November 2012 is shown in table 1 below.

Table 1. Monthly Summary of the Website

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2012	699	758	3,544	50,838	2.09 GB
Feb 2012	687	739	2,638	36,898	1.77 GB
Mar 2012	825	899	3,789	48,513	2.18 GB
Apr 2012	709	770	2,495	31,598	1.61 GB
May 2012	862	979	2,951	38,332	2.01 GB
Jun 2012	522	556	1,606	19,896	1.07 GB

Jul 2012	394	466	1,296	16,959	922.56 MB
Aug 2012	131	152	405	5,550	280.97 MB
Sep 2012	383	424	1,241	16,536	914.08 MB
Oct 2012	451	513	1,254	16,857	912.60 MB
Nov 2012	468	521	1,370	18,445	933.49 MB
Total	6,131	6,777	22,589	300,422	14.59 GB

The maximum average size is in the month of March which is 2380026 Kbytes.

5.1.1 Monthly Usage of the Website

The daily web usage data for the month of November 2012, tells about the number of Hits occurs as well as about Files, Pages, Visits, and Kbytes. For the month of November 2012, the maximum Hits were 18,445, the maximum Files opened were 1,370, the maximum Pages accessed were 1,370, the maximum Visits were 521 and the maximum downloading in Kbytes were 933.49. The total summary of monthly use of the Website is shown in figure 2.

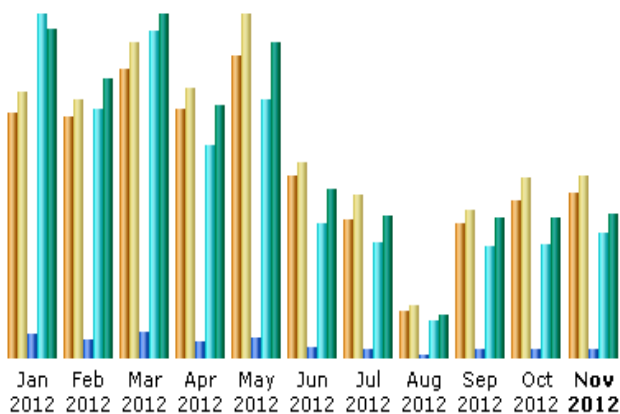


Fig. 2: Monthly Usage of the Website

5.1.2 Daily Usage of the Website

In each month, one can also see the daily usage chart-an illustration of the visits received per day. Figure 3, displayed below, shows the website which is accessed in each day of the month, it receives far more visits from Monday to Saturday. The increased visits received by the Website on Monday to Saturday, reinforces the earlier finding that this Website is mainly used by working executives and/or students at the office or school. Recreational users of the Website, with Internet access at home, would likely visit it more on the weekends. Daily summary by the month is shown in table 2 (see Appendix-I).

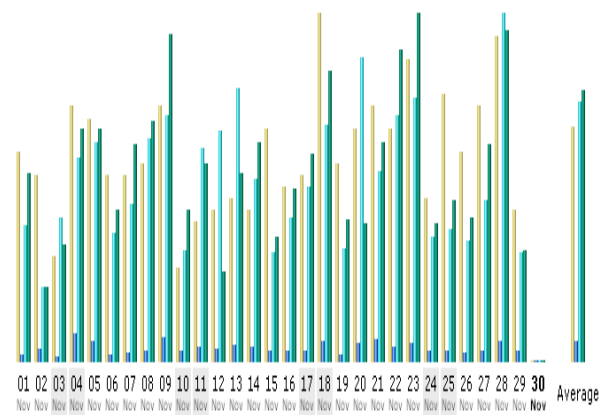


Figure 3: Daily Usage of the Website

5.1.3 Day of Week Usage of the Website

The day of week usage of the website has been shown in figure 4 that illustrates the number of hits and pages, on average, at every day of week for November 2012.

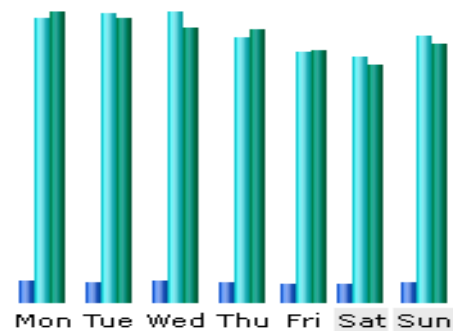


Fig. 4: Day of Week Usage of the Website

And the total summary of day of week usage of the website for November 2012 is summarized in table 3.

Table 3. Day of Week Usage of the Website

Day	Pages	Hits	Bandwidth
Mon	69	941	48.43 MB
Tue	68	961	47.40 MB
Wed	72	962	45.90 MB
Thu	69	880	45.52 MB
Fri	63	830	42.21 MB
Sat	59	816	39.60 MB
Sun	68	882	43.07 MB

5.1.4 Hourly Usage of the Website

The hourly usage of the web site provides the number of hits and pages, on average, at every hour of the day. Figure 5 displays a drastic increase in the hits starting from 8:00 am. The increase then continues steeply until 11:00 am, though not as steeply from 7:00 am to 8:00 am. Between noon and 1:00 pm, there is a dip in the usage, but by 2:00 pm there is a peak usage of the web site. From this time until late at night there is a gradual decline in the usage which is illustrated in figure 5 below.

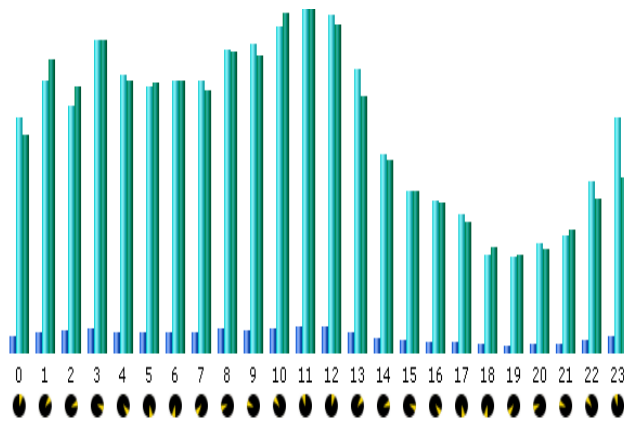


Fig. 5: Hourly Usage of the Website

And the total summary of hourly usage of the website is described in table 4.

Table 4. Hourly Usage of the Website

Hours	Pages	Hits	Bandwidth
0	900	12,702	599.66 MB
1	1,102	14,743	808.88 MB
2	1,205	13,406	730.80 MB
3	1,313	16,974	857.92 MB
4	1,085	15,092	750.66 MB
5	1,102	14,441	744.76 MB
6	1,055	14,770	747.53 MB
7	1,138	14,763	720.70 MB
8	1,276	16,407	826.49 MB
9	1,222	16,748	815.28 MB
10	1,284	17,698	934.76 MB
11	1,392	18,884	958.40 MB
12	1,452	18,268	900.73 MB
13	1,111	15,351	707.27 MB
14	754	10,732	532.28 MB
15	659	8,763	444.81 MB
16	626	8,248	411.81 MB
17	564	7,546	359.57 MB
18	473	5,345	291.92 MB
19	352	5,208	269.29 MB
20	468	5,941	286.41 MB
21	510	6,380	338.46 MB
22	652	9,306	422.05 MB
23	894	12,706	482.18 MB

The high usage of the website between 8:00 am and 5:00 pm is indicative to the working executive visitors and is reaffirmed by the dip at noon and 1:00 pm, lunchtime. This hourly usage provides information on what type of person visits the website.

5.2 Page View Analysis

The analysis of specific pages viewed adds another dimension into ascertaining the identity of the typical website visitor. Here we have attempted to discern the visitor by the type of services which they seek on the website, rather than purely by what time they access the website. This analysis is useful in the formulation of new services to offer on the website. These pages are those pages that are the first requested in a visit (called as Entry page), and the last requested (called as Exit page). When a visit is first triggered, the requested page is counted as an Entry page, and whatever the last requested URL was, is counted as an Exit page.

5.2.1 Top Entry Pages

An entry page is the first page requested in a visit. This page, triggered by the visit statistic, is where people start navigating on the website. The top entry pages table contains a list of many pages and documents on the website. The top entry page is “/” which, unsurprisingly, signifies the initial page. The other pages are most likely bookmarked i.e. their links are stored by the browser for easy access. This data on entry pages can also be applied to the amelioration of the user experience for the Website. Knowing where users first access the Website means services can be placed on these pages that encourage the user to navigate to other pages. It means that specific care should be taken to ensure that these pages load quickly and have little chance of containing errors. The most viewed pages with their average size are shown in table 5 (see Appendix-II).

5.2.2 Top Exit Pages

An exit page is the last page requested in a visit by the user. This page, triggered by the visit statistic, is where people end navigating on the Website. The top exit pages are described in table 6 that contains a list of many pages and documents on the Website. The top entry page is “/” which, unsurprisingly, signifies the web page. The most exit pages with their average size are shown in table 6 (see Appendix-III).

5.2.3 Top URLs

The top Uniform Resource Locators (URLs) most visited pages are shown in table 7 (see Appendix-IV), the main thing is that the dominance of “.css”, “.js”, “.jpg” and some other files as the most requested URLs. The reason for this is that these files are used by the visitors’ browsers to present the Website web page. While the visitor never sees these pages, every computer that accesses the Website downloads them. So, for the purpose of this document, these files are irrelevant. Two most requested and relevant URLs communicate a lot about the Website users. The most important information is that the *contactus* URL is one of the most requested.

5.2.4 Top Search Keyphrases

The search keyphrases (known as search strings) obtained from examining the referrer string and looking for known patterns from various search engines. The collected web log data of the website has been analyzed through AWStats and it has been found that the top 10 total search keyphrases are used by the users to access the website through the search engine. The most of the search keyphrase *dav kota* has been searched by the many users when they have accessed the

website. The total search keyphrases with their percentage of Hits are shown in table 8.

Table 8. Top Search Keyphrases of the Website

Search Keyphrases (Top 10)		
779 different keyphrases	Search	Percent
dav kota	1205	27.3 %
dav public school kota	534	12.1 %
davkota	345	7.8 %
dav school kota	311	7 %
www.davkota.org	152	3.4 %
mahatma hansraj	116	2.6 %
http://www.davkota.org/	64	1.4 %
dav public school kota website	57	1.2 %
davkota.org	55	1.2 %
dav public school kota rajasthan	46	1 %
Other phrases	1527	34.6 %

6. CONCLUSION

In this paper, we have collected the web access log data from the web server of an educational institution web site to predict several web metrics and statistics such as user's daily access, top hosts, most popular requested pages, most popular images, browsers, hits per each day etc. The complete work is done by analyzing web log data for a long period of time. Several analyses like Visitor Pattern Analysis, Page View Analysis and Time Analysis have performed by using preprocessed web log data. The results after experiments are satisfactory and contained vital information about the users who have visited the website. The obtained results can be used to predict the behaviour and interest of the website users by which the loyalty of the website will be measured.

7. REFERENCES

- [1] Arvind K. Sharma and P.C. Gupta, "Exploration of efficient methodologies for the Improvement in web mining techniques: A survey", *International Journal of Research in IT & Management (ISSN 2231-4334)* Vol.1, Issue 3, July 2011.
- [2] Aditi Shrivastava and Nitin Shukla, "Extracting Knowledge from User Access Logs", *International Journal of Scientific and Research Publications*, Volume 2, Issue 4, April 2012.
- [3] <http://www.w3.org/TR/WD-logfile.html>
- [4] Liu, H., and Keselj, V., "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting user's future requests", *Data and Knowledge Engineering*, 2007, Vol 61, Issue 2, pp.304-330.
- [5] Arya, S., and Silva M. "A methodology for web usage mining and its applications to target group identification", *Fuzzy sets and systems*, 2004, pp.139-152.
- [6] G.C. A.K, "An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network", *Fifth International Conference on Information Processing* 2011. Springer-Verlag.
- [7] Maheswara Rao et al, "An Enhanced Pre-Processing Research Framework for Web Log Data Using a Learning Algorithm," *Computer Science and Information Technology*, DOI, pp. 1-15, 2011. 10.5121/csit.2011.1101.
- [8] Valter Cumbi and Lourino Chemane, "Mozambican Government Portal Case Study: Visitor Analysis", *IST-Africa 2007 Conference Proceedings* Paul Cunningham and Miriam Cunningham (Eds) IIMC International Information Management Corporation, 2007 ISBN: 1-905824-04-1.
- [9] <http://www.davkota.org>
- [10] <http://www.awstats.sourceforge.net>

Appendix–I

Table 2. Daily Usage of the Website

Day	Number of visits	Pages	Hits	Bandwidth
01 Nov 2012	18	26	469	30.99 MB
02 Nov 2012	16	41	257	12.26 MB
03 Nov 2012	9	20	499	19.31 MB
04 Nov 2012	22	98	708	38.24 MB
05 Nov 2012	21	68	760	38.23 MB
06 Nov 2012	16	24	447	24.77 MB
07 Nov 2012	16	30	546	35.83 MB
08 Nov 2012	17	40	775	39.51 MB
09 Nov 2012	22	86	856	53.73 MB
10 Nov 2012	8	40	383	24.99 MB
11 Nov 2012	12	47	739	32.60 MB
12 Nov 2012	13	45	801	14.63 MB
13 Nov 2012	14	55	946	31.04 MB
14 Nov 2012	13	51	632	35.85 MB
15 Nov 2012	20	34	377	20.30 MB
16 Nov 2012	15	39	496	28.50 MB
17 Nov 2012	16	37	605	34.21 MB
18 Nov 2012	30	68	822	47.73 MB
19 Nov 2012	17	25	394	23.39 MB
20 Nov 2012	20	65	1,056	22.68 MB
21 Nov 2012	22	75	657	35.83 MB
22 Nov 2012	20	49	851	51.15 MB
23 Nov 2012	26	61	912	57.08 MB
24 Nov 2012	14	39	433	22.69 MB
25 Nov 2012	23	35	461	26.38 MB
26 Nov 2012	18	30	419	23.61 MB
27 Nov 2012	22	40	560	35.58 MB
28 Nov 2012	28	67	1,204	54.26 MB
29 Nov 2012	13	35	380	18.12 MB
30 Nov 2012	0	0	0	0
Average	20	67	896	44.60 MB
Total	521	1,370	18,445	933.49 MB

Appendix–II

Table5. Top Entry Pages of the Website

Entry					
different pages-url	Viewed	Average size	Entry	Exit	
/	5,586	12.68 KB	4,437	1,734	
/founder.htm	755	16.30 KB	466	459	
/faculty.htm	937	28.95 KB	246	359	
/admissionrules.htm	905	14.79 KB	236	373	
/aboutschoo.htm	658	14.45 KB	153	214	
/sports.htm	675	12.02 KB	142	189	
/contactus.htm	1,471	13.91 KB	139	766	
/achievements.htm	1,738	14.99 KB	123	503	
/aboutkotacity.htm	383	11.54 KB	98	129	
/index.htm	990	13.17 KB	72	323	
/presidentsmessage.htm	287	11.36 KB	66	69	

Appendix–III

Table 6. Top Exit Pages of the Website

Exit					
different pages-url	Viewed	Average size	Entry	Exit	
/	5,586	12.68 KB	4,437	1,734	
/contactus.htm	1,471	13.91 KB	139	766	
/achievements.htm	1,738	14.99 KB	123	503	
/founder.htm	755	16.30 KB	466	459	
/admissionrules.htm	905	14.79 KB	236	373	
/faculty.htm	937	28.95 KB	246	359	
/index.htm	990	13.17 KB	72	323	
/results.htm	799	10.84 KB	28	321	
/aboutschoool.htm	658	14.45 KB	153	214	
/album/slides.xml	403	36.28 KB	2	207	
/sports.htm	675	12.02 KB	142	189	

Appendix–IV

Table 7. Most requested URLs of the Website

Pages-URL (Top 25) - Full list - Entry - Exit					
45 different pages-url	Viewed	Average size	Entry	Exit	
/	5,586	12.68 KB	4,437	1,734	
/achievements.htm	1,738	14.99 KB	123	503	
/contactus.htm	1,471	13.91 KB	139	766	
/index.htm	990	13.17 KB	72	323	
/faculty.htm	937	28.95 KB	246	359	
/admissionrules.htm	905	14.79 KB	236	373	
/results.htm	799	10.84 KB	28	321	
/founder.htm	755	16.30 KB	466	459	
/sports.htm	675	12.02 KB	142	189	
/aboutschoool.htm	658	14.45 KB	153	214	
/album/	474	1.03 KB	16	22	
/cirriculum.htm	454	11.58 KB	45	112	
/album/res/play.swf	417	49.16 KB		9	
/album/slides.xml	403	36.28 KB	2	207	
/celebrations.htm	400	13.62 KB	38	81	
/divisionofhouse.htm	396	12.78 KB	40	73	
/aboutkotacity.htm	383	11.54 KB	98	129	
/culturalactivities.htm	373	11.37 KB	14	47	
/educationtours.htm	330	10.55 KB	3	41	
/examination.htm	322	12.83 KB	19	102	
/computerlab.htm	321	12.49 KB	13	32	
/library.htm	318	11.48 KB	20	35	
/musicroom.htm	309	11.11 KB	16	51	
/physicslab.htm	296	11.10 KB	13	30	
/discipline.htm	290	11.02 KB	35	57	
Others	2,589	10.89 KB	363	507	