Statistical analysis of pageviews on web sites

P.H.A.J.M. van Gelder[1,2], G. Beijer[1], & M. Berger[1]
[1] *NEDSTAT BV, Dalsteindreef 30, 1112 XC DIEMEN, The Netherlands.*
[2] *Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands.*

Abstract

Pageview statistics are useful to describe and predict the behaviour of clients on internet sites. From a theoretical point of view, the number of pageviews during a day should be Poisson distributed. However, violation of stationarity assumptions causes other distribution types to fit pageviews data usually much better. In this paper a procedure is described that explains how to homogenise the data (with detrending techniques) and allows several distribution functions as possible candidates. A goodness-of-fit test will select the optimal distribution for the given dataset. In particular attention will be paid to the occurence probabilities of large numbers of pageviews on different types of slightly correlated websites. The paper furthermore presents models for giving forecasts on the number of pageviews during the rest of the day (given a number of pageviews earlier that day) and for giving uncertainty intervals with that forecast.

1 Introduction

Pageview statistics are useful to describe and predict the behaviour of clients on internet sites. Typical questions that are related to visitor behaviour are the frequency and length of visits during a certain time period, the entrance and exit locations of visitors, the percentage of visitors who reach key pages (such as a sign-up page, cash register, etc), the paths they take, the traffic trend, the prediction of traffic spikes, the accommodation of server space for increased traffic, the adjustment for browser technology, the evaluation of behaviour variations among subsets of customers and the change during sales, etc, etc. However, these questions are difficult to answer because of the existence of several boundary conditions: human behaviour is very stochastic and data can be incomplete or noisy caused by the existence of proxy servers, fire walls, caching,

browser settings, and cookies. Therefore much research has been devoted to the investigation of these issues and several conferences and workshops have special sessions on these issues. Good references may be found in the proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Workshop on Web Usage Analysis (WebKDD), ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD), ACM Workshop on Web Information and Data Management (WIDM), International Conference on Information and Knowledge Management (CIKM), International Conference on E-Commerce and Web Technologies (ECWeb), SIAM International Conference on Data Mining (SDM), IEEE International Conference on Tools with Artificial Intelligence (ICTAI), IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), International Conference on Human Computer Interaction (HCI), Workshop on Information Technologies and Systems (WITS), Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Statistical Models for Data Mining (SMDM, Giudici et al., [3]). An overview of scientific journals with publications on webtraffic analysis and pageview statistics in particular is given by: Journal of Internet Research, Journal of Knowledge and Information Systems, IEEE Bulletin of the Technical Committee on Data Engineering, IEEE Internet Computing, Journal of Applied Artificial Intelligence.

The papers of Kushida [4] and Naldi [6] describe the characteristics of internet traffic, in particular the backbone traffic (number of packets, connection time, number of retransmitted packets, volume of the data transmitted, etc). In this paper a statistical analysis will be presented on the number of pageviews on a website. The first software packages for pageview counters were developed in the mid 1990s. The idea behind counting pageviews was to create a pixel in such a way that every time someone visited a site, the pixel was downloaded from the server and the logfiles were used to track visitors. The pageview counters make the visitor information immediately available by reading the logfiles and displaying a few graphs and tables with the visitor information.

There is nowadays an abundance of software for visualisation of pageview statistics and visitor behaviour on websites. A selection of companies that have developed excellent tools for web traffic analysis:

`http://www.nedstat.nl/`, `http://www.personify.com/`,
`http://www.netgen.com/`, `http://www.webtrends.com/`,
`http://www.nettraffic.de/`, `http://sm6.sitemeter.com/`,
`http://www.stats4all.com/`, `http://www.superstats.com/`.

Each of the packages has its own advantages and disadvantages, and improvement of its performance is an on-going activity.

This paper is organised as follows. In Sec. 2, the pageview data will be described and some first properties of the data will be shown (spectral plot, distribution analysis, correlation matrix). In Sec. 3, the Poisson model will be outlined and its fit to data examined. The tails of the statistical distribution functions will be investigated in Sec. 4, and a comparison will be made with pageview data from other types of websites. A correlation analysis of the traffic between the websites will be part of this comparison. In Sec. 5, two prediction

methods that can be used for estimating the expected number of pageviews for the rest of the day, given a number of visitors that have already been measured until a certain moment in time, will be presented. Finally in Sec. 6 and 7, conclusions, recommendations and references are included.

2 Pageview data at Nedstat

In this section an overview is given of the available datasets. Nedstat tracks currently (February 2002) about 700,000 websites on its pageviews.

Table 1. The number of websites and its daily activity in the number of pageviews #PVs, (measured on January 30, 2002).

#PVs	<=0	<=1	<=2	<=5	<=10	<=15	<=20	<=25	<=30
#sites	306501	69548	42127	70535	56284	29853	19263	13491	9927
<=40	<=50	<=100	<=200	<=500	<=1000	<=2000	<=5000	<=10^4	<=inf
14256	9526	22569	14447	11110	4662	2710	1935	741	629

Notice that for about 44% of the websites there is no traffic at all, and that for only 0.1% of the sites there is more than 10,000 pageviews per day. These percentages remain fairly stable in time. Furthermore, the data of table 1 follows remarkably a very nice curved line on double logarithmic paper (Fig. 1). Such a surprising relation has not been found in the earlier literature.

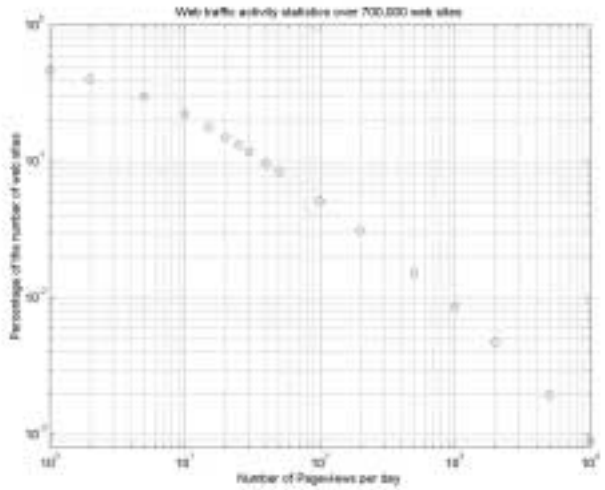
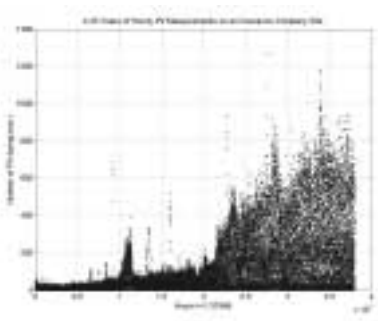


Figure 1: Relationship between the number of websites and its daily activity (measured by the number of pageviews).

The pageview data of a particular website consists of hourly counts and is updated every hour. The longest measurements of hourly pageviews are currently up to 5 years. Figure 2 shows the measurements for a large insurance company in the Netherlands.



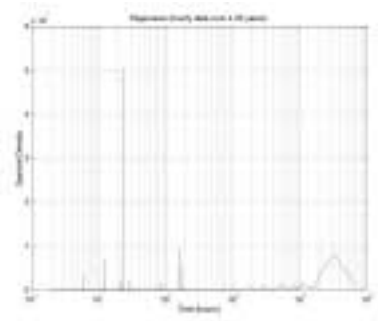


Figure 2 (left)**:** Growth of the number of pageviews over time (October 1997 - January 2002). Figure 3 (right)**:** A spectral density plot of web traffic measured in hourly intervals.

Applying a spectral FFT analysis to the above time series[1] leads to the identification of a few peaks: 6 hours, 12 hours, 24 hours (first significant), and 168 hours (second significant), as is illustrated in Fig. 3. Indeed, the first significant peak is caused by the huge difference in traffic during the day and during the night. Internet traffic between 0.00am and 7.00am takes in avarage only 2% of the total traffic during the day. Heaviest traffic takes place between 10am and 5pm. There is a dip around 6pm and a second relative high traffic intensity around 9pm (up to 5% in one hour of the total traffic during the day). This traffic distribution differs from the type of website (commercial, governmental, entertainment, etc), and also depends strongly on the origin of its visitors. Evening visitors from the USA are measured as night visitors for European sites. The correlation between the number of day and evening pageviews is quite high (>0,6), whereas the correlation in the night is small (< 0.4).

3 Data preparation and model fits

A process in which points occur randomly in time is called a stochastic process. It turns out that under some basic assumptions that deal with independence and uniformity in time, a single, one-parameter probability model governs all such random processes. This is an amazing result and because of it the Poisson process (named after Simeon Poisson) is one of the most important in probability

[1] After detrending. A partial autocorrelation test ensured adequate stationarity of the resulting time series.

theory. It is often used as a model for the number of events (such as the number of telephone calls at a business or the number of accidents at an intersection) in a specific time period. It is also useful in ecological studies, e.g., to model the number of prairie dogs found in a square mile of prairie.
The assumption that is made can be described intuitively (but imprecisely) as follows: If a time t is fixed, then the process after time t is independent of the process before time t and behaves probabilistically just like the original process. Thus, the random process has a regeneration property. This assumption enables us to derive the distribution of each of the following in turn:

The interarrival times
The arrival times
The number of arrivals in an interval

The interarrival times X1, X2, ... are continuous, independent random variables, each having the exponential probability density function:

$$f(t) = re^{-rt}, t \$ 0 \quad (1)$$

The k'th arrival time is simply the sum of the first k interarrival times:

$$Tk = X1 + X2 + \cdots + Xk \quad (2)$$

The density function of the k'th arrival time is

$$fk(t) = (rt)^{k-1} re^{-rt} / (k - 1)!, \quad t > 0 \quad (3)$$

This distribution is the gamma distribution with shape parameter k and rate parameter r. The number of arrivals Nt in an interval (0, t] are Poisson distributed. At least k arrivals come in the interval (0, t] if and only if the k'th arrival occurs by time t:

$$Nt \$ k \text{ if and only if } Tk \# t \quad (4)$$

The density function of the number of arrivals in the interval (0, t] is

$$P(Nt = k) = e^{-rt} (rt)^k / k! \text{ for } k = 0, 1, ... \quad (5)$$

The number of pageviews data of Sec. 2 has to be prepared in order to fulfill the above assumptions. First of all the hourly variation will be avoided by summing up the pageviews to daily observations. The growth over time will be substracted from the data by performing a (piecewise or segmented, if necessary) regression analysis (Fig. 4).

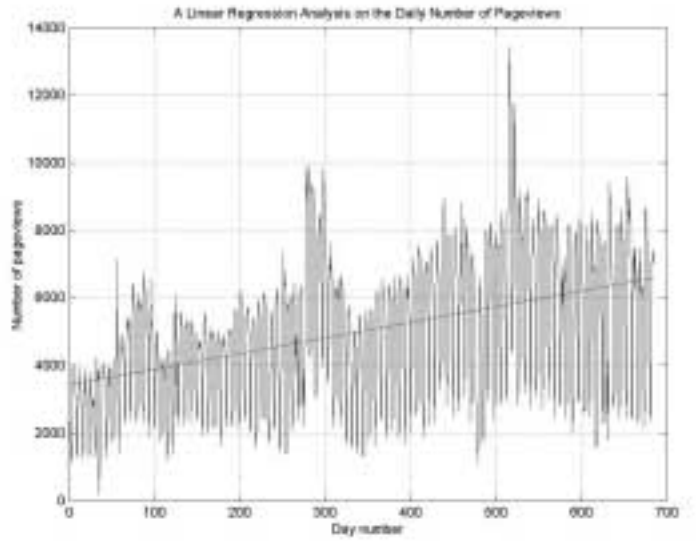


Figure 4: Linear regression of the daily PV data of the insurance company data from Fig 2.

Finally the clear difference between web traffic during weekdays and weekends is separated, leading to two datasets of which the weekdays dataset is presented in Fig. 4. The weekend dataset should be analysed separately.

According to the theory of Poisson processes the prepared dataset of Fig. 4 should follow a Poisson process (as well as the dataset of pageviews during the weekends). However, the results from Table 2 show that, according to a Chi-Square criterion on the differences between fit and observations, the Logistic and Normal distributions model the data much better than the Poisson process (which has rank 10 in the list of Chi-Square results).

Table 2: Results (489 observations).

Function	Chi-Square	Rank	K-S Test	Rank	A-D Test	Rank
Logistic(0.59,8.34e-2)	0.097938	1	0.067942	1	3.254982	1
Normal(0.59,0.15)	0.324778	2	0.079696	2	5.656272	2
Triang(2.92e-2,0.57,1.23)	0.573779	3	0.212218	8	36.729396	8
Lognormal2(-0.51,0.57)	1.109844	4	0.265259	9	56.512201	9
Weibull(4.60,0.70)	1.285934	5	0.20609	7	32.881735	7
Poisson(0.59)	2.674916	10	NA		NA	

Roadknight et al. [7] came through the analysis of usage logs at a range of caches also to the observation that WWW traffic is not a Poisson arrival process.

In the next section, a more detailed analysis will be performed upon the tails of the probability distributions; in particular the right tail.

4 Tails of the pageview distribution

In this section an investigation will be presented on the occurence probabilities of extreme high webtraffic loads. Such information is important for the design aspects and performance evaluation of networks.

The normal distribution that was fitted to the pageview data from Sec. 2, shows a relative unsatisfactory behaviour to the extreme observations (Fig. 5). The model estimates occurence probabilities that are up to a factor 10 smaller than what is observed (12000 pageviews estimated with a 10^{-4} probability, instead of 10^{-3}). On semilogarithmic paper, the observations follow quite reasonably a straight line, indicating that an exponential distribution should be adopted for the tail.

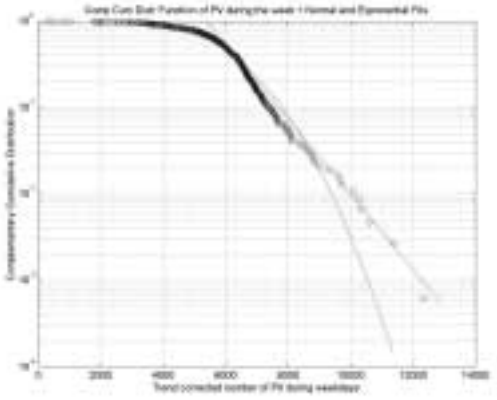


Figure 5: Complementary cumulative distribution function of pageviews.

In order to check if this property also holds for other types of websites, apart from the insurance company data, pageview data from a newspaper site, Ministry site, a webtraffic counter site, and a publisher site were included in this study. Indeed, it is noted that the tails follow very well a straight line and one may assume that the hypothesis of expontial tail behaviour will be accepted. For completeness, a correlation matrix of the hourly pageview data among the different website types is included in table 3.

Table 3. Correlation matrix.

	Publisher	Newspaper	Counter	Insurance	Ministry
Publisher	1	0.88	0.80	0.73	0.92
Newspaper		1	0.84	0.85	0.91
Counter			1	0.63	0.84
Insurance				1	0.74
Ministry					1

Apparently there is no need for full correlation between sites to obtain comparable extreme tail behaviour.

5 Prediction models

In this section a few prediction models are presented that can be used for providing information to the web site visitor about the number of pageviews which can be expected at the end of the day. Assume that T is the time (in hours) at which a prediction of the expected number of pageviews at the end of the day (denoted by M*) is desired. Assume that already N pageviews took place that day until time T. From historical records it can be derived how the occurence time of a pageview is distributed over the day from time 0 to 24 hours. This distribution will be given by F(t) in which t is the time parameter in hours. So F(t) = P(occurence time of a pageview is before time t). Note that F(0) = 0 and F(24) = 1. For convenience we write F(T) = p in which p is the ratio of the number of pageviews that occur within the period 0 to T and the total number of page views that occur in a day. An intuitive prediction of M* would be:

$$M^* = \frac{N}{p} \tag{6}$$

However, this prediction has some disadvantages in case of small samples or no samples at all. Division by a small number will lead to unrealistic high predictions M*. An intuitive prediction of the uncertainty of M* would be:

$$\sigma(M^*) = \frac{1}{\sqrt{M^*}} \tag{7}$$

Alternative expressions are given by the following formulae:

$$M^* = N + D\int_T^{24} f(t)\,dt \tag{8}$$

in which D is the expected number of pageviews on a whole day and f(t) = d/dt F(t). D can be estimated on basis of the observed daily number of pageviews over the last few months (under the assumption of stationarity). An unbiased estimator for D on basis of n historical values of D is:

$$D^* = \Sigma Di / n \tag{9}$$

The uncertainty in M* (under the assumption of normality[2] a 68.3% confidence interval; 2.sigma corresponds to a 95.45% interval; 3.sigma to 99.73%).

$$\sigma(M^*) = s(D)\int_T^{24} f(t)\,dt \text{ ; in which s(D) an estimator for } \Sigma(D) \text{ and given by:} \tag{10}$$

$$s(D) = \sqrt{\frac{\sum_{i=1}^{n}(D_i - D^*)^2}{n-1}} \tag{11}$$

[2] Central Limit Theorem can be applied here. Note that a location and scale parameter do not affect the distribution type of a normal distribution.

Disadvantage to the above approach is that the calculations are quite time consuming. Furthemore, a trend correction on the historical dataset has to be performed before the predictions can be made.

If both prediction methods are applied to hourly measured data and predictions are provided at 5 am, given the observed pageviews between 0 and 5 am of that day, the following results are obtained (Fig. 6).

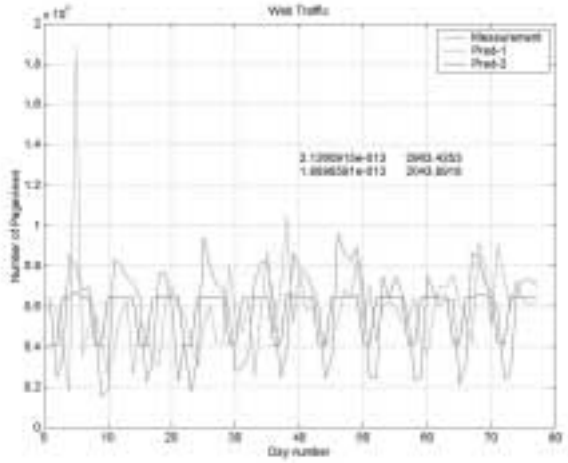


Figure 6: Comparison of two prediction methods with actual measurements.

The bias of both prediction methods is negligable (10^{-13}), and the RMSE (root mean square error) of the first prediction method is about 50% higher than the second method (2963 versus 2043). However, if calculation time is considered in a judgement analysis, certainly the first method would be prefered.

Finally, ARIMA models (Bragg, [1]) will be able to decrease the RMSE in predictions even more, but calculation time for these types of models can be very large (especially in case of large datasets).

8 Conclusions and Recommendations

In this paper, an extensive analysis is presented of statistical properties for pageview data of websites. Apart from distribution function analysis, also spectral and correlation analyses have been outlined in this paper.

From a theoretical point of view, the number of pageviews during a day should be Poisson distributed. However, violation of stationarity assumptions causes other distribution types to fit pageviews data usually much better. It is recommended to homogenise the data (with detrending techniques, and seperation of data from weekends and weekdays) and to allow several distribution functions as possible candidates. A goodness-of-fit test will select the optimal distribution for the given dataset.

The behaviour of pageview distributions in tail can be very well described by exponential functions, irrespective to the type of websites. Such information is very useful when occurence probabilities have to be provided for extreme webtraffic loads.

To give forecasts on the number of pageviews during the rest of the day (given a number of pageviews earlier that day) and to give confidence intervals with that forecast, several methods are described in this paper. Accurate methods with low bias and RMSE values are available, but require long calculation time. A fast calculation method is also described, for which the mistakes are still acceptable.

Acknowledgements

The authors thank Peter Bel, Michiel Schaap, and Karen Aerts from Nedstat BV for their assistence in providing the datasets.

References

[1] Bragg, A.W.: The locality and transitional behavior of ARIMA network traffic models, Technical Report North Carolina State University, 1997.

[2] Feelders, A.J.: Statistical Concepts. Chapter 2 in Intelligent Data Analysis. Editors: Berthold M., and Hand D.J., Springer, 1999

[3] Giudici, P., Castelo R.: Association Models for Web Mining. Data Mining and Knowledge Discovery 5 (2001) 183-196

[4] Giudici, P., Heckerman D., Whittaker J.: Statistical models for data mining. Data Mining and Knowledge Discovery 5 (2001) 163-165

[5] Kushida, T.: An Empirical study of the characteristics of internet traffic. Computer Communications 22 (1999) 1607 – 1618

[6] Liu Z., Niclausse N., Jalpa-Villanueva C.: Traffic model and performance evaluation of Web servers, PERFORMANCE EVALUATION 46 (2-3): 77-100 OCT 2001

[7] Naldi, M.: Measurement-based modelling of Internet dial-up access connections. Computer Networks 31 (1999) 2381-2390

[8] Roadknight C., Marshall I., Bilchev G.: Network performance implications of variability in data traffic. BT TECHNOLOGY JOURNAL 18 (2): 151-158 APR 2000

[9] Siliani S.: Data Mining for e-intelligence; Understanding Customer Behavior on the Web. Technical Report SAS - Italy, February 2001

[10] Spiliopoulou M., Pohle C.: Data mining for measuring and improving the succes of web sites. Data Mining and Knowledge Discovery 5 (2001) 85-114

[11] Van der Mei R.: Modelling network traffic (In Dutch: Het modelleren van netwerkverkeer). NAW 5/1 nr. 4 December 2000, pp. 390 - 396