Web server traffic in crisis conditions

Andersson, Mikael; Bengtsson, Anders; Höst, Martin; Nyberg, Christian; Holst, Jan

2005

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Web Server Traffic in Crisis Conditions

Mikael Andersson[1], Anders Bengtsson[2], Martin Höst[1] and Christian Nyberg[1]
[1] Department of Communication Systems, [2] Department of Mathematical Statistics
Lund Institute of Technology,
Box 118, SE-221 00 Lund, Sweden,
[1] {mike|martinh|cn}@telecom.lth.se, [2] ab@maths.lth.se

Abstract

During recent years we have seen several large-scale crises. The 9/11 terror attacks, tsunamis, storms, floods and bombings have all been unpredictable and caused a great deal of damage. One common factor in these crises has been the need for information and one important source of information is usually web sites. In this work three new sets of web server access logs are presented and analyzed, one of which represent the traffic to the major news site, Aftonbladet, in Sweden after the bombings in London, 7th of July 2005. The differences in document popularity between the crisis logs and the other logs are investigated.

I. Introduction

Recent years have displayed large-scale crises of several kinds. Since the terror attack in New York September 11, we have seen tsunamis, storms, murders of political leaders and recently the July bombings in the London underground system. These events have in common that it is crucial to spread information about the crisis. Information has to be updated, quick and reliable. Today many people turn to the Internet when they search for information and a crisis situation is no exception. The problem in a crisis situation is that the web servers serving the web sites get overloaded. For example; September 11, 2002, the Swedish foreign minister Anna Lindh was murdered. As a result, one of the major news sites in Sweden crashed because of overload for 10 minutes [1], [2]. The Committee on the Internet Under Crisis Conditions: Learning from September 11 reports similar overload conditions from web sites during the reporting of the September 11 attack in New York [3].

The Swedish Emergency Management Agency (SEMA)[1] has now granted a new research framework program called Risk and Vulnerability Analysis of Technological and Social Systems (FRIVA)[2], that deals with emergency handling on several levels. One of the subprojects focuses on infrastructure problems such as telecommunication systems and the Internet in crisis situations. This work has been funded by FRIVA and SEMA. Crisis-related research in Internet systems is new but nevertheless important. The End-to-End workgroup[3] stated in their 2005 vision for a different world of communications [4] that ”In 10 years, the network itself, and critical applications that run on it, should address the special needs that arise in times of crisis.”

It is beyond the scope of this paper to formally define the term *crisis*. However, informally it could be stated that a crisis is when a combination of events, e.g. accidents and sabotage, result in a situation that negatively affect society in a way that hinders vital society functions. Examples of crises that are included in this definition might be terror attacks, storms, tsunamis, murders etc. Situations that are not considered crises according to this definition are for example the Olympic games, the Monday lunch rush hour, political elections etc. However, there is no clear line that divides crisis situations from normal situations.

In this work we have focused on traffic to a web server during a crisis situation, the bombings in the London underground system 2005. We have analyzed access logs from a web server at the major news site, Aftonbladet[4] in Sweden. We also compare the logs to two other access logs from Aftonbladet. The analyzed metrics are document request arrivals, document popularity and object type distributions.

Section II describes the access logs that have been used in the analysis. In Section III the analysis of the access logs is shown. Finally, Section IV sums up the work.

II. Access logs

Aftonbladet is the largest news site in Sweden with several millions visits per week according to [5]. Their web site consists of several identical web caches in front of ordinary web servers. The logs in this work come from one of the web caches. Three logs were recorded; the London underground bombings (LONDON), the Eurovision Song Contest final (ESC) and a ”normal” Thursday (NORMAL). Table I shows dates, durations (hours), the numbers of gigabytes downloaded and the numbers of millions file objects requested..

Log	Start time	Duration	Gigabytes	Requests
ESC	Sat 05-05-21 1.00 PM	45	364	80
LONDON	Thu 05-07-07 2.15 PM	26	200	38
NORMAL	Thu 05-10-26 5.15 PM	46	245	59

TABLE I
The access logs

[1] In swedish: Krisberedskapsmyndigheten (KBM)
[2] http://www.friva.lucram.lu.se
[3] http://www.irtf.org
[4] http://www.aftonbladet.se

Figure 1 shows the number of requested files in the LONDON logs. The figure shows that most traffic occurs at midday and afternoon hours, and during the night the traffic drops considerably.

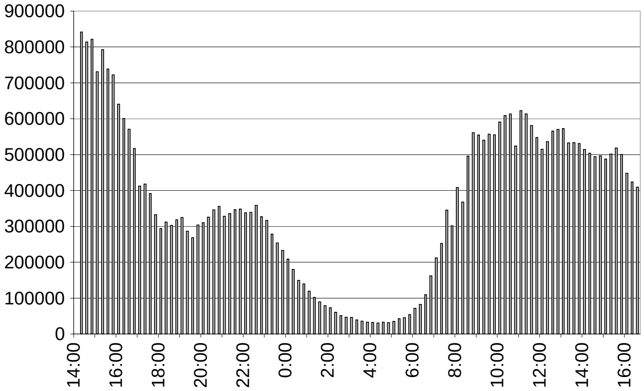

Fig. 1. The number of requests.

A. *Documents*

A user that visits a web site requests several files during his visit. Normally, a request is made for a certain document, which results in many subsequent requests for images, style sheets etc. Denote any file requested by a visitor at the web site by *object*. An object can be a static HTML-file, dynamic files such as Common Gateway Interface (CGI) files, images, style sheet files, flash movies etc. A *document* is for example an article at a news site or similar that contains several objects. The first object that is requested in a document is normally the HTML file that will cause the web browser to request inline images, style sheets etc.

In this work document requests are defined as those requests that start with an HTML request. However, some commercial advertisment parts that are included in article documents also appear in pure HTML format. These commercial advertisement requests were not included in the definition of documents. With this definition documents can be seen as requests for articles on the news site.

III. Analysis

The access logs were analyzed with respect to document popularity which the authors believe to be important for crisis-related traffic. Also, object type distributions and document arrivals were investigated.

A. *Document request arrivals*

The high intensity of document request arrivals makes it hard to measure the time between two consecutive document request arrivals because of the relatively coarse time resolution in the access logs (1 second). Instead the number of arrivals per second can be measured.

Figure 2 shows the distribution of the number of document request arrivals per second compared to a theoretical Poisson distribution with the same intensity. Figure 3 shows the probability plot of the same distributions. The distribution is taken from the first 45 minutes of the LONDON access logs. Since there is a good fit, a Poisson process can be assumed to be a good model for document request arrivals.

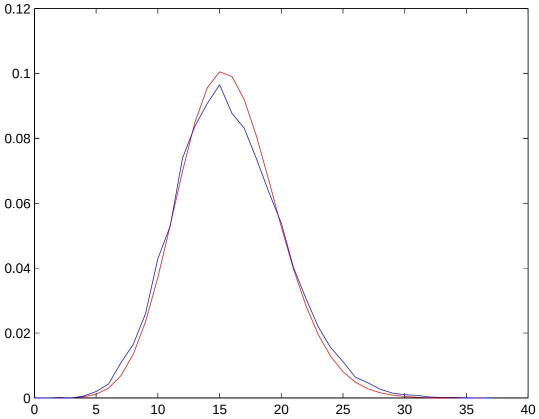

Fig. 2. The distribution of the number of document request arrivals per second compared to a theoretical Poisson distribution.

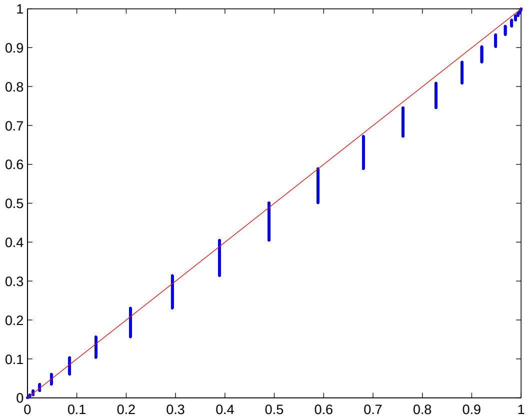

Fig. 3. Probability plot of the number of document request arrivals per second.

B. *Document popularity*

Document popularity can be translated to the relative number of requests a certain document receives at the web site. The access logs were split into 15 minute intervals where document popularity was collected for each document and interval. Documents were defined as described above, except that the most popular document was excluded in each interval, since it was the ”mandatory” entrance page to the web site.

Figure 4 shows the document popularity versus popularity ranking for the first 15 minutes for the three logs. The popularity is expressed as the number of requests to the document divided by the total number of document requests. The LONDON document popularity curve has a more narrow beginning which means that a small number of documents are more popular than the other documents at a higher degree than the other logs.

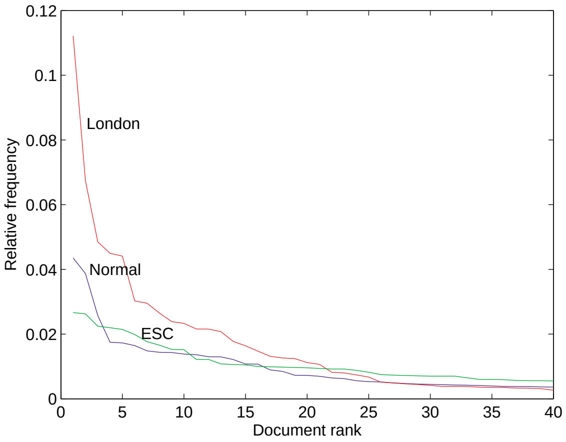

Fig. 4. The document popularity for the first 15 minutes of the three logs.

Figure 5 shows the document popularity versus popularity ranking for three different parts of the logs - 5, 8 and 14 hours after the London bombings. It can be seen that the popularity decreases rather rapidly and the first 15 documents account for between 30 to 60 % of the total number of document requests in the figure. Furthermore, it can also be seen that the steepness of the popularity curve decrease with time.

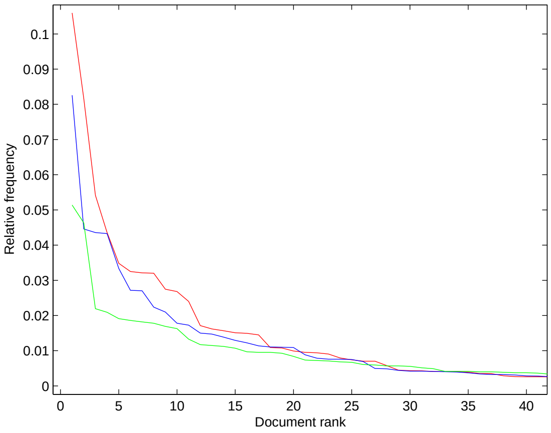

Fig. 5. The document popularity for three different parts of the log. From top to bottom: 5, 8 and 14 hours after the London bombings.

A decreasing steepness of popularity curves seems to be a characteristic in traffic to a web server during a crisis. During a crisis, the interest will be focused on a limited number of documents related to the crisis. This interest should then decline when the news value becomes smaller.

The Zipf law has traditionally been used to model object popularity when human choice is involved [6], [7]. According to this law the popularity is proportional to the inverse of the ranking, that is

$$\text{Popularity} = K \cdot \frac{1}{\text{rank}^{\alpha}}$$

with the exponent α being close to 1. The requirement of $\alpha \approx 1$ may be too strict [8]. If instead α is used as model parameter it is possible to get a better fit and the document popularity can instead be fitted to a regular power-law curve. In Figure 6 the document popularity is plotted on a log-log scale and two power-law curves, one for the 15 most requested documents and a different curve for the rest, are fitted with a standard least squares method. The time interval used in the figure is 2.15 PM to 2.30 PM, July 7. As seen, the fit is very good.

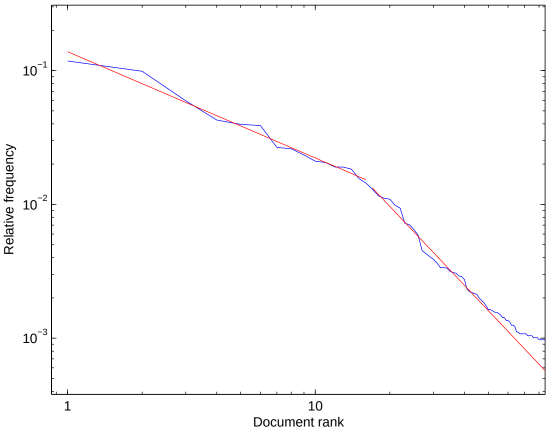

Fig. 6. Log-log plot of the document popularity. Power-law fits shown as straight lines.

Another way of determining the impact of the interest of crisis-related documents is to look at the popularity of the ten most popular documents at the web site. Figure 7 shows the popularity of the 10 most popular documents over time for the three logs. The LONDON logs show much higher popularity than the other two logs in the beginning. Up to 50 % of the document requests were made to one of the 10 most popular document in the LONDON logs. After a few hours the popularity decreases to normal levels, which gives support for the theory above that the distribution of document popularity during a crisis is more narrow and that this effect declines over time after the crisis.

C. *Object type distributions*

Object type distributions have also been investigated. Figure 8 shows the distribution of the most popular object types. For example, GIF images contribute to more than 45 percent % of the total number of requests. The relative number of downloaded bytes are also shown as comparison. While GIF images are very popular, they only represent around 8 % of the total number of downloaded bytes. If both GIF and JPEG images are considered, they stand for about 75 % of the number of requests and 47 % of the bytes. 3 % of the requests

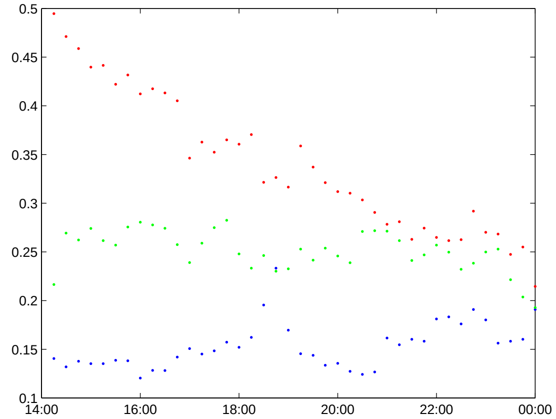

Fig. 7. The popularity of the 10 most popular documents over time. From top to bottom: LONDON, NORMAL and ESC.

were not recognized by their object type and are shown as ”unknown” in the figure. About 1 % of the requests account for the remaining object types. The object type ”directory” refers to when a user has requested a directory, for example ”www.site.com” or ”www.site.com/section1”. This normally results in that an HTML object is returned.

Interestingly, while not representing very many requests, HTML objects account for 26 % of the number of downloaded bytes. The other types shown in the figure are CSS (Cascading Style Sheets that contains template information for HTML pages), FLG and SWF that are Flash-related objects and finally XML that are used for sending data to web pages.

In a crisis it is not unusual for a web server to become overloaded. Different mechanisms [9]–[11] has been investigated and designed for controlling the overload, however, not many deal with so called *content adaption*, which means that the complexity and size of the documents are reduced during overload instead of completely rejecting requests. In a crisis situation it is probably more interesting to distribute as much information as possible to as many visitors as possible instead of distributing full pages to a limited amount of visitors.

The object type distributions shown here clearly show that much of the requests are made for image files. This suggests that a content adaption system could benefit much from reducing both the number and the size of returned images in crisis situations, since images do not add very much information to a page compared to text-based information.

IV. Conclusions

A new set of web server access logs were presented and analyzed. The logs come from Aftonbladet, the largest news site in Sweden during a.o. their reporting on the London underground system bombings. Power-law functions were fitted to the document popularity curves and it could be seen that the popularity of the 10 most popular documents where much higher during the crisis compared to the other two logs. The Poisson distribution is shown to be a good fit for number of arriving documents requests, which makes the Poisson process a good candidate for the arrival process of document requests.

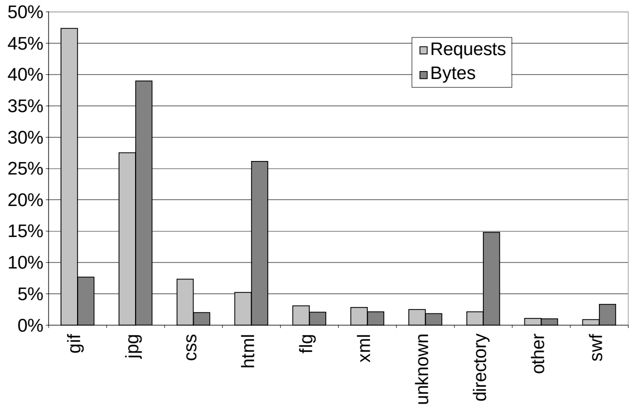

Fig. 8. Object type distribution.

Also, the distribution of object types were investigated. Image files account for about 75 % of the requests and 47 % of the bytes during the crisis. This indicates that a content adaptation system would benefit much from targeting images during a crisis overload situation.

Acknowledgments

The authors wish to express their gratitude to Aftonbladet for contributing the access logs and especially Jesper Wahlberg. This work has been funded by the Swedish Emergency Management Agency (SEMA).

References

[1] “Webben sviker när det gäller,” *Computer Sweden*, http://computersweden.idg.se/ArticlePages/200309/15/20030915132253_cs038/20030915132253_cs038.dbp.asp(InSwedish.

[2] L. Larsson, “Ministermordet,” kBM:s temaserie 2004:4, Swedish Emergency Management Agency.

[3] “Internet under crisis conditions: Learning from september 11,” *National Research Council*, 2003.

[4] D. D. Clark, C. Partridge, R. T. Braden, B. Davie, S. Floyd, V. Jacobson, D. Katabi, G. Minshall, K. K. Ramakrishnan, T. Roscoe, I. Stoica, J. Wroclawski, and L. Zhang, “Making the world (of communications) a different place,” End-to-End Research Group, Tech. Rep., 2005.

[5] “Kiaindex - insight xe, 2005 v36, unika webbläsare,” http://www.annons.se/?get=content&action=view&id=127-98.

[6] C. Williamson, R. Simmonds, and M. Arlitt, “A case study of web server benchmarking using parallel wan emulation,” *Performance Evaluation*, vol. 111-127, no. 49, 2002.

[7] M. Crovella and A. Bestavros, “Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes,” in *Proceedings of SIGMETRICS’96: The ACM International Conference on Measurement and Modeling of Computer Systems.*, May 1996.

[8] “File popularity characterisation,” http://www.ee.ucl.ac.uk/~imarshal/sigreview.pdf.

[9] M. Andersson, “Performance modeling and control of web servers,” Department of Communication Systems, Lund Institute of Technology, Tech. Rep. 160, 2004, lic. Thesis.

[10] L. Cherkasova and P. Phaal, “Session-based admission control: A mechanism for peak load management of commercial web sites,” *IEEE Transactions on computers*, vol. 51, no. 6, pp. 669–685, June 2002.

[11] C. Lu, T. Abdelzaher, J. Stankovic, and S. So, “A feedback control approach for guaranteeing relative delays in web servers,” in *Proceedings of the 7th IEEE Real-Time Technology and Applications Symposium*, 2001, pp. 51–62.