Detecting User Dissatisfaction and Understanding the Underlying Reasons

Åke Arvidsson
Packet Technologies, Ericsson Research
SE-164 80 Stockholm, Sweden
Ake.Arvidsson@ericsson.com

Ying Zhang
Packet Technologies, Ericsson Research
200 Holger Way, San José, CA 95134, U.S.A.
Ying.Zhang@ericsson.com

Categories and Subject Descriptors

C.2.0 [**Computer Systems Organization**]: PERFORMANCE OF SYSTEMS

General Terms

Measurement

Keywords

Passive monitoring, QoE

1. INTRODUCTION

Today network operators face the challenges of accurately measuring Quality-of-Experience (QoE) for general web applications. QoE is a subjective metric relating to user expectation, satisfaction, and overall experience and despite the rich history in the multimedia community [1, 2, 3, 4] there is no scalable method to quantify QoE for general web applications. Traditional methods relies on users to indicate their satisfaction in surveys in a subjective manner. Such methods can be inaccurate as the results depend on the sampled test subjects and it does not scale.

A recent study proposes to infer user dissatisfaction related to web pages from the behaviour of early interruptions [5]. More specifically, unsatisfactory long times between requests and presentations could make impatient end-users push the STOP or RELOAD buttons in their web browsers thereby initiating end-user cancellations. Such cancellations would result in early closures of the HTTP sessions as well as the underlying TCP connections and may be interpreted as indicators of unsatisfactory QoE. However, naively using all early terminations as signs of bad QoE is not accurate as it may result in many false positives caused by, *e.g.,* user losing interest in the content or closing the browser accidentally. Thus, the first question we explore in this paper is, how to identify the set of terminations that reflects true user dissatisfaction? Once this set is narrowed down, we use it to answer questions like what the acceptable throughput and response time are for most users.

In this paper, we consider transaction-based web applications and develop a method to automatically search for early cancellations from passively collected traces. Our contribution in this paper is twofold. First, we propose a methodology to systematically examine the early cancellations and eliminate false positives. Our key idea is to identify *acceptable performance* by comparing normal transactions to cancelled ones and, since tolerance levels can deviate across users, to do this from user-specific profiles. We show that the filtering method generates much more meaningful results than the naive approach. Second, we conduct comprehensive analysis of data collected in a wireline fibre network and a wireless cellular network. We present results of *acceptable* performance metrics such as throughput and response time in this paper.

SIGMETRICS'13, June 17–21, 2013, Pittsburgh, PA, USA.
ACM 978-1-4503-1900-3/13/06.

2. METHODOLOGY

We will first introduce the data set, followed by our methods of identifying early termination from passively collected traffic traces, and finally describe the filtering methodology to reduce the false positives.

Data set. In this paper, we use three datasets of full-size packets. The first data set (*Wireline0*) is collected from a fibre access network in the Sweden. It is a two-day packet dump with 9.8M HTTP flows, 235K of which have client initiated RST sent. Both wireless data sets (*Wireless0* and *Wireless1*) were collected on a Gn interface between a GGSN and a SGSN in a cellular network. Both sets contain 354 hours of data collected at different locations. The first set contains 3.8M flows, among 783K of which contains reset packets. The second set is at the same scale of 3M flows. To guarantee user privacy, we were not given direct access to the data but our partner kindly ran our code and shared the resulting, anonymised metadata with us.

Identifying user early terminations. Our method to identify users' unsatisfactory experiences is to search for patterns from passively collected packet traces. The intuition is that when a user experiences bad performance, he or she will initiate the closure of the HTTP session and the communication pattern observed in the network will thus be different from the one during normal downloading cases. For example, when a client clicks "STOP" button on the browser before the download is completed, we observed that the client sends a TCP reset packet, *[RST,ACK]*, to close the TCP connection. It then keeps on replying with *[RST]* packet when receiving additional data packets from the web server. We have developed a tool that assembles web pages from packet traces and classify the outcomes by searching for such patterns, the details can be found in [5].

Filtering method. Although user-initiated cancellations are likely indicators of user dissatisfaction, it is clear that cancellations also may be related to other reasons but performance, *e.g.,* web navigation or loss of interest. One way to identify the reasons behind cancellations is to interview the end users but, since it cannot scale, we propose to heuristically classify cancellations as "performance related" or "for other reasons".

In more detail, we identify "acceptable performance" after which cancellations are said to be related to performance if their performance is worse than this threshold, but attributed to other reasons otherwise. Noting that tolerance may vary between different users and different content, we propose to determine "acceptance thresh-

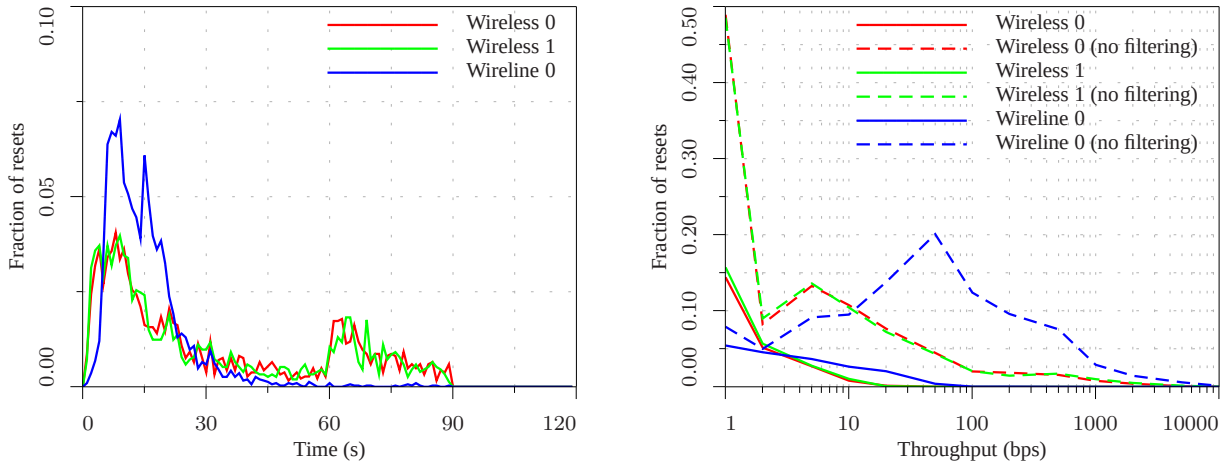

Figure 1: Occurrence of resets at different times (left) and rates (right). The dashed lines in the right graph refer to unfiltered results.

olds" *per user* (to account for different users) expressed *as percentiles* (to account for different content in a way which less sensitive to outliers).

The complete method may be described by the following steps:
1. Remove all requests for which it cannot be verified that they are related to web browsing. (We have used the criterion that the agent name should contain the word "Mozilla". This means that we capture, *e.g.*, Internet Explorer, Chrome and Firefox which correspond to about 46% and 83% of all pages in the wireless data sets and the wireline data set respectively.
2. Build a list of all cancelled requests C and a list of all users (IP addresses) U with at least one cancelled request.
3. Build a list of all normal requests N the users (IP addresses) of which are found in U.
4. Remove all users in U the IP address of which is not seen in at least $R_{\min}$ normal requests. (We have used $R_{\min} = 20$.) Then remove all requests in C and N related to users no longer seen in U.
5. For all users $u \in U$, compute the pth percentile of the response time, $\rho_u(p)$, and the πth percentile of the throughput, $\lambda_u(\pi)$, over all normal requests with the IP address of u. (We have used $p = 75$ and $\pi = 25$.)
6. Remove all requests in C the response time of which is below ρ_u or the throughput of which is above λ_u. Then remove all users in U the IP address of which is no longer seen in C and finally all requests in N related to users no longer seen in U.

3. RESULTS

We now present some results about what users accept in terms of response times and throughputs and how this is linked to the classical measure of QoE known as MOS.

Figure 1 shows at what times and rates resets occur. For **times** (left) we divide time into one second bins and show the number of resets per bin normalised by the number of resets in all bins per data set. It is seen that most resets occur within the first 20 seconds and that there are two peaks. The first, after a few seconds, is likely to be related to "impatient users", and the second one, after about one minute, is probably related to TCP time outs.

For **rates**, shown in the right in Figure 1, we divide rates in approximately logarithmic bins 1, 2 and 5 times 1, 10, 100 *etc.* and show the number of resets per bin normalised by the number of resets in all bins per data set. It is seen that most resets occur for low rates and that the number of resets decreases when the throughput increases (solid lines). To demonstrate the effectiveness of the filtering method, we also plot the same metric using the naive method (dashed lines) and we note that there is no such strong correlation, *i.e.*, we have reset peaks at higher rates which presumably are related to, *e.g.*, navigation or correction rather than performance.

Nothing the difference between filtered and unfiltered data, it is concluded that filtering is necessary to obtain meaningful data. *It is also noted that the two accesses are remarkably similar although we note that resets in the wireline data set are more concentrated with respect to time (left diagram) and less concentrated with respect to rates (right diagram).*

4. CONCLUSION

We have presented a new method to identify a set of characteristics from traffic logs that are related to the user perceived experience. The method identifies such QoE indicators without active involvement of users and without active measurements injected into the network. To reduce the false positives of this approach, we developed a filtering method that identifies and compares to acceptable performance by building a profile for each user. We conducted a set of studies of the correlations between these indicators as well as their correlation with performance metrics.

5. REFERENCES

[1] ITU-T Recommendation, "P. 800. Methods for subjective determination of transmission quality." International Telecommunication Union, 1996.

[2] ITU-T Recommendation, "J. 144. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference." International Telecommunication Union, 2001.

[3] ITU-T Recommendation, "P.862 Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." International Telecommunication Union, 2001.

[4] S. Voran, "The development of objective video quality measures that emulate human perception," in *Proc. IEEE GLOBECOM*, 1991.

[5] Å. Arvidsson, Y. Zhang, and N. Beheshti, "Detecting user dissatisfaction from passive monitoring," in *Proc. of EuroCon*, 2013.