On the Structure and Characteristics of User Agent Strings

IMC 2017
November 2
London
United Kingdom

Jeff Kline & Aaron Cahn (comScore)
Paul Barford (comScore, University of Wisconsin - Madison)
Joel Sommers (Colgate University)
Introduction and Motivation

About comScore

- We measure and report on audiences for publishers, brands, app developers, etc.

- To measure this, we need the data. To get the data, we partner with brands, publishers, app developers, etc.

- The result is telemetry with worldwide reach. Our telemetry is deployed by major publishers, campaigns and apps.

- Volume on a typical day is ~50B records; each record represents an HTTP(S) request.

- We also maintain a large research panel, we measure TV traffic…

- comScore Labs is the research arm of comScore. It is based in Madison, Wisconsin. We have strong academic roots.
Study Objectives

Describe the User Agent (UA) space from the perspective of a large-scale real-world data corpus

• How large is the space?

• How does it evolve over time?

• How well does the UA fulfill its purpose?

• What about anomalies?
The UA is transmitted as part of the HTTP header

RFC 1945

10.15 User-Agent
The User-Agent request-header field contains information about the user agent originating the request. This is for statistical purposes, the tracing of protocol violations, and automated recognition of user agents for the sake of tailoring responses to avoid particular user agent limitations.

Example:
User-Agent: CERN-LineMode/2.15 libwww/2.17b3

About the study’s data

Archive spanning 2 year time window. Day 0 is January 1, 2015.

Each record in the archive records the Volume of requests that UA issued to our web servers on Day. The schema is:

<table>
<thead>
<tr>
<th>Day</th>
<th>UA</th>
<th>Volume</th>
</tr>
</thead>
</table>
The long tail

The number of distinct UA’s encountered per day is $O(\text{millions})$.

A lot of the mass of the rank-order distribution lives in the tail. The tail is long and there is no clear threshold.

![Graph showing the long tail distribution with daily volume of the millionth-ranked UA is thousands per day.](image-url)
UA Aggregation

Aggregating over UA is a basic task of web log analysis

- Want to get Android or iPhone traffic? This almost works…
  .*Android.*
  .*iPhone.*


- Even for these simple questions, this task is complicated. Long-term maintenance is a challenge.

- Validation?

- comScore’s internal categorization code-base is thousands of lines long
A publicly-visible and independent view that illustrates UA complexity
https://udger.com/resources/ua-list
These are not really “long tail”. Each has millions of records per day.
These are not really “long tail”. Each
has millions of records per day.
Time-dependent features of the UA distribution

Hour-of-day and day-of-week matter in the UA distribution. This matters for results that relate PII to the UA.
Time-dependent features of the UA distribution

The top 1k UA’s churn in a stable manner.

The top 1k week-over-week sets have Jaccard similarity of ~0.7.
The UA space over time

Character Entropy Matrix

This stripe reflects the common prefixes Mozilla, Dalvik. It may be used in conjunction with the legend to help interpret the representation.
Lessons

• UA categorization and parsing is (still) a challenge. This task is basic to web log analysis.

• The UA space is diverse and dynamic.

  • The week-over-week Jaccard similarity of the top 1k is relatively stable at about 0.7.

  • UA distribution depends on time-of-day and day-of-week (among other things)

• Introduce the character entropy matrix. It is simple to construct, interpret and it has been used to expose unexpected features within the UA-space.
If the community expresses interest, we will try to make a portion of our UA set available for academic research.

Thank you.
jkline@comscore.com
Jeffery Kline