Robots Still Outnumber Humans in Web Archives in 2019, But Less Than in 2012

Himarsha R. Jayanetti, Kritika Garg, Sawood Alam, Michael L. Nelson, Michele C. Weigle
Department of Computer Science, Old Dominion University

Introduction

- Comparison of robots vs. humans in web archive access logs [1].
- Identified user sessions as human or robot based on browsing behavior.
- Examined user access patterns and temporal preferences.
- Extension of AlNoamany et al. [2]

Bot Identification

- Known bots: A compiled list of User-Agents that are known to be used by bots.
- Number of UA per IP: The IPs that update their User-Agent (UA) more than 20 times.
- robots.txt: A session that requested for the robots.txt file.
- Image-to-HTML Ratio (IH): A session that requested less than one image file for every 10 HTML files.
- Browsing Speed (BS): A session with a browsing speed faster than one HTML request every two seconds (BS >= 0.5 requests/sec).

Conclusion

- Percentage of robots requests in IA decreased over time (2012 – 91%, 2015 – 88%, 2019 – 70%)
- Robots account for 98% of requests in PT2019.
- Robots are almost entirely limited to Dip and Skim access patterns in IA 2012 and 2015, but exhibit all the patterns and their combinations in IA 2019.
- Both humans and robots show a preference for web pages archived in the near past.

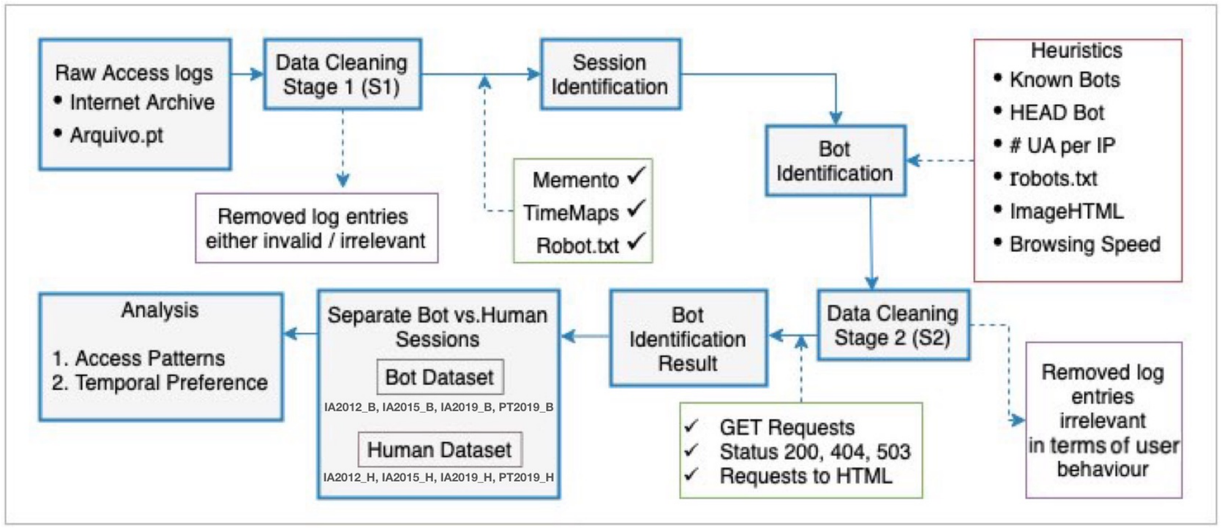
References

[1] Himarsha R. Jayanetti, Kritika Garg, Sawood Alam, Michael L. Nelson, and Michele C. Weigle. Robots still outnumber humans in web archives, but less than before. In *Proceedings of the Theory and Practice of Digital Libraries Conference (TPDL)*, September 2022.

[2] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. Access patterns for robots and humans in web archives. In *JCDL '13: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 339–348, 2013.

Methodology

Datasets: Full-day access logs from Internet Archive's Wayback Machine (IA) and Portuguese Web Archive (Arquivo.pt)
$*IA2012\ Logs,\ 02/02/2012$ $*\ IA2015\ Logs,\ 02/05/2015$ $*\ IA2019\ and\ PT2019\ Logs,\ 02/07/2019$



Results

The bot identification results based on the total number of sessions and the total number of requests for each dataset (the header for each column displays the total number of sessions and requests).

Heuristics	IA2012 Sessions 1,527,340	IA2012 Requests 22,302,090	IA2015 Sessions 1,355,286	IA2015 Requests 27,424,389	IA2019 Sessions 2,658,637	IA2019 Requests 42,868,048	PT2019 Sessions 3,680	PT2019 Requests 613,672
Known Bots	21,423 *(1.40%)*	398,053 *(1.78%)*	19,441 *(1.43%)*	639,335 *(2.33%)*	322,379 *(12.13%)*	4,969,187 *(11.59%)*	884 *(24.02%)*	67,453 *(10.99%)*
#UA per IP	5,050 *(0.33%)*	756,801 *(3.39%)*	1,824 *(0.13%)*	683,138 *(2.49%)*	5,475 *(0.21%)*	1,442,574 *(3.37%)*	3 *(0.08%)*	2,636 *(0.43%)*
robots.txt	1,958 *(0.13%)*	11,074 *(0.05%)*	2,992 *(0.22%)*	11,061 *(0.04%)*	9,296 *(0.35%)*	31,452 *(0.07%)*	404 *(10.98%)*	4,236 *(0.69%)*
IH Ratio	1,327,896 *(86.94%)*	19,893,394 *(89.20%).*	1,034,404 *(76.32%)*	22,308,925 *(81.35%)*	1,746,989 *(65.71%)*	24,056,112 *(56.12%)*	2,916 *(79.24%)*	589,363 *(96.04%)*
Browsing Speed	237,271 *(15.53%)*	4,563,851 *(20.46%)*	239,120 *(17.64%)*	8,108,851 *(29.57%)*	514,878 *(19.37%)*	21,176,163 *(49.40%)*	1,694 *(46.03%)*	162,068 *(26.41%)*
Total Robots	**1,340,318** *(87.76%)*	**20,281,301** *(90.94%)*	**1,083,830** *(79.97%)*	**24,132,614** *(87.99%)*	**1,854,282** *(69.75%)*	**29,968,059** *(69.91%)*	**3,584** *(97.39%)*	**603,654** *(98.37%)*

The heuristics are not mutually exclusive.

arXiv Technical Report: https://arxiv.org/abs/2208.12914

Contact: {hjaya002, kgarg001}@odu.edu, sawood@archive.org, {mln,mweigle}@cs.odu.edu

Access Patterns

* The basic access patterns for users of web archives proposed in 2012.

* The access patterns of robots and humans in our datasets. The color of the stacked bar distinguishes between requests for mementos and TimeMaps.