Why We Need Private Web Archives

Hussam Hallak
WSDL Research Group
Ph.D. student

Types of web pages

- **Publicly Archivable:**
  No paywall (mainly)

- **Not Publicly Archivable:**
  SERPs
  Maps
  Social Media: Facebook, Linkedin, etc
  Commerce: Amazon
  Email sites: Yahoo, Gmail, Hotmail, etc
  Bank statements

www.cs.odu.edu/~zeil/

Dr. Steven J. Zeil

Associate Professor, Old Dominion University

Ph.D., 1981, Ohio State University

vita

Contact Information

- Office: E&CS 3208

Phone: (757) 683-4928
E-mail: zeil@cs.odu.edu

- Office hours and Appointments

Dr. Steven J. Zeil

Associate Professor, Old Dominion University

Ph.D., 1981, Ohio State University

vita

Contact Information

- Office: E&CS 3208

Phone: (757) 683-4928
E-mail: zeil@cs.odu.edu

- Office hours and Appointments


Secure | https://web.archive.org/web/20170210123939/https://www.facebook.com/barackobama
INTERNET ARCHIVE WayBackMachine
https://www.facebook.com/barackobama
Go
1,699 captures
13 Feb 2008 - 15 Apr 2018
JAN 2016
FEB 10 2017
MAR 2018
About this
facebook
註冊
加入或登入 Facebook
電子郵件或電話
密碼
忘記帳號？
登入
是否要加入 Facebook？
註冊
安全驗證
請輸入以下的文字
如此圖片無法顯示，請再瀏覽器上點選重整。
無法分辨上圖文字？
換一組認証碼或撥放驗證碼
請輸入圖中的文字：
這是什麼？
送出
中文(台灣) English (US) Tiếng Việt Bahasa Indonesia ภาษาไทย 日本語 Español Português (Brasil) Français (France) Deutsch Italiano +
註冊 登入 Messenger Facebook Lite 行動版 尋找朋友 用戶 粉絲專頁 地標 遊戲 地點
名人 社團 Moments Instagram 關於 刊登廣告 建立粉絲專頁 開發人員 工作機會 隱私政策 Cookie
Ad Choices 使用條款 使用說明
Facebook © 2017

Problem

- Estimating the amount of archivable web traffic

Few websites get most of the traffic

- Top 100 websites receive %49 of the entire traffic on the web.
- Fact is consistent with a Pareto distribution.

Web Traffic Measures

- **Total visits** is the total number of non-unique visits from last month.
- **Unique visits** is the number of unique visits from last month.
- **Pages/visit** is the average number of visited pages per user's visit.

Approach

- Collect all three web traffic measures for the top 100 websites (Alexa Ranking) from 2 different web traffic analysis services (SimilarWeb, Semrush). (Feb 23rd 2018)

- Classify the top 100 websites:
  Publicly Archivable or **Not Publicly Archivable**

- Calculate the percentage of publicly archivable web traffic using six different formulas that use some or all traffic measures and their combinations in different ways

65% of Web Traffic is NOT publicly archivable

- Using all measures, 65.30% of the traffic of the top 100 sites is not archivable by public web archives
- What about the remaining 1.8 billion live websites?

Best case: If all are publicly archivable,
31% of web traffic is NOT publicly archivable

Worst Case: If all are Not publicly archivable,
83% of web traffic is NOT publicly archivable

Most Likely: The bottom half is similar to the top half
65% of web traffic is NOT publicly archivable

Take Away Message

- Personalized web pages are on the rise.
  Therefore, personal web archiving is crucial.
- There will always be pages that public archives cannot capture no matter how good they get.
- We need personal web archiving tools
- Available tools:
  Web Recorder, Warcreate, WAIL