Jawa:
Web Archival in the Era of JavaScript

Ayush Goel
University of Michigan

Jingyuan Zhu
University of Michigan

Ravi Netravali
Princeton University

Harsha V. Madhyastha
University of Michigan
How to **reduce the storage overhead** of web archives and **improve the quality** of archived pages?
A Couple of Days After the Conference..
A Couple of Days After the Conference..

About me
Ayush Goel is a PhD candidate in the computer science department at the University of Michigan. He is affiliated with the systems lab where he is advised by Harsha V. Madhyastha and co-advised by Ravi Netravali. His research involves applying programming language techniques to optimize modern day web pages on various fronts like performance and storage.

Recent News (all)
- March 2022: Our work Jawa got accepted at OSDI'22.
10 Years From Now (2032)....
High Prevalence of Link Rot on the Web

The Paper of Record Meets an Ephemeral Web: An Examination of Linkrot and Content Drift within The New York Times

Citation

RESEARCH ARTICLE
Scholarly Context Adrift: Three out of Four URI References Lead to Changed Content

Shawn M. Jones\textsuperscript{1,}*, Herbert Van de Sompel\textsuperscript{1,}\textsuperscript{3}, Harihar Shankar\textsuperscript{1,}\textsuperscript{3}, Martin Klein\textsuperscript{1,}\textsuperscript{3}, Richard Tobin\textsuperscript{2,}\textsuperscript{3}, Claire Grover\textsuperscript{2,}\textsuperscript{3}

1 Digital Library Research and Prototyping Team, Research Library, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America, 2 Language Technology Group, The University of Edinburgh, Edinburgh, Scotland, United Kingdom
Web Archives to the Rescue
Web Archives to the Rescue

~700 Billion pages, ~100 Petabytes
How Modern Web Archives Operate?

Web pages on live web

Fetching HTML, JS, CSS, Images

Apply deduplication, compression

Storage

Fetch archived page HTML, CSS, JS, Images

User
Problems with Web Archives: **No Snapshots for Many Pages**

Explore more than 702 billion [web pages](http://icc-cricket.yahoo.net/newsdetails.php?newsId=18836_1327733) saved over time

**Hrm.**

Wayback Machine has not archived that URL.
Click here to search for all archived pages under [http://icc-cricket.yahoo.net/newsdetails.php](http://icc-cricket.yahoo.net/newsdetails.php)
Problems with Web Archives: Poor Page Fidelity
Outline

★ Challenges faced by web archives

★ Understanding the root cause

★ Our insights and design of a new crawler

★ Evaluation
Outline

★ Challenges faced by web archives

★ **Understanding the root cause**

★ Our insights and design of a new crawler

★ Evaluation
Root Cause: **Increasing JavaScript** on Web Pages

**Corpus**: Landing pages of 300 Alexa sites
Root Cause: **Increasing JavaScript** on Web Pages

Per page snapshot **expensive** → **Fewer snapshots**

**Corpus**: Landing pages of 300 Alexa sites
Root Cause: JavaScript Induced Non-Determinism

Resources fetched different from crawled → Poor page fidelity
Outline

★ Challenges faced by web archives
★ Understanding the root cause
★ Our insights and design of a new crawler
★ Evaluation
Key Insight: Archived Page Differs from Live Page

A. No back-end origin server

JavaScript that interacts with server will not work → Removing it will not impact fidelity

Add comments to the article
Key Insight: Archived Page Differs from Live Page

B. Certain sources of non-determinism are absent

JavaScript in certain control flows will never be executed → Removing it will not impact fidelity

Reads client-side cookie: Always returns “no”
Jawa: A New Web Archive Crawler

Improve page fidelity

➔ Eliminate impact of non-determinism

Reduce storage cost

➔ Identify and remove non-functional code

➔ Identify and remove unreachable code
Jawa: A New Web Archive Crawler

- Improve page fidelity
- Reduce storage cost
- Eliminate impact of non-determinism
- Identify and remove non-functional code
- Identify and remove unreachable code
Make JavaScript Execution **Completely** Deterministic

**Original page**

**Deterministic page (broken)**
Understand How Non-determinism Impacts Resources Fetched

Date, Math.random, Performance (DRP)

Client characteristics

Analyze impact on control flow on 3000 pages

- No impact on control flow ➔ retain the non-determinism
- Influences control flow impacting resource fetches ➔ Eliminate non-determinism
Jawa: A New Web Archive Crawler

- Improve page fidelity
- Reduce storage cost
- Eliminate impact of non-determinism
  - Identify and remove non-functional code
  - Identify and remove unreachable code
Insufficient to Save Only JS Executed While Crawling

Michigan Researchers Design a More Efficient Web Archive Crawler

PhD student Ayush Goel, along with his advisor Harsha V. Madhyastha designed a new crawler that significantly lowers the storage overhead while ensuring high page fidelity.

By Kenneth Chang
July 20, 2012

Sign up for Science Times Get stories that capture the wonders of nature, the cosmos and the human body. Get it sent to your inbox.
Challenge: Code Executed Depends on How User Interacts
Challenge: Code Executed Depends on Order
Challenge: Code Executed Depends on **Order**

➔ **Read-write dependencies** between different events on page

◆ **JavaScript/DOM state**

➔ **Can not predict** order of events
One result

1000s of results

Exploring all possible orderings/inputs impractical
Analysis Framework: Understand State Dependencies

Inject instrumentation code

function foo(a,b) {
  var p = Proxy(win, handler);
  p.c = _(a) + _(b);
  return _ret(true);
}

foo(1,2);

Trigger each interaction

Monitor state accesses

Corpus: 3000 pages from 300 Alexa sites
Findings From the Analysis

➔ Order mostly influences coverage for analytics events.

➔ Input mostly influences coverage for text-based events.

Both categories of events irrelevant for archived pages

Implication:
Estimate code coverage with **single execution of events with default inputs**
Outline

★ Challenges faced by web archives

★ Understanding the root cause

★ Our insights and design of a new crawler

★ Evaluation
Evaluation

Storage
- Storage for web resources
- Storage for crawling/serving indices

Fidelity
- Failed Resource fetches
- Visual comparison
- Functional interactions

Throughput
- Crawling throughput
- Overhead of each technique
Evaluation

Storage
- Storage for web resources
- Storage for crawling/serving indices

Fidelity
- Failed Resource fetches
- Visual comparison
- Functional interactions

Throughput
- Crawling throughput
- Overhead of each technique
Storage: Jawa Reduces Storage Overhead by **41%**

**Corpus:**
1 million snapshots of pages on 300 sites archived by Internet Archive
Fidelity: Jawa Eliminates Almost All Failed Resource Fetches

Corpus: 3000 web pages from 300 sites
Fidelity: Jawa Eliminates Almost All Failed Resource Fetches

Corpus: 3000 web pages from 300 sites
Conclusion

1. JavaScript on web pages negatively impacts web archival

2. Fundamental differences b/w live and archived pages can be exploited to overcome such negative impacts

https://github.com/goelayu/Jawa

goelayu@umich.edu
Evaluation.
Backup
Jawa: A New Web Archival Crawler

Web pages on live web

Crawler

Computed lightweight crawling

Reduced Storage Overhead

Storage

High Fidelity Replay

User
Improve Page Fidelity: Strawman #1

Screen capture

No client interactions
Jawa: A New Web Archival Crawler

High Fidelity

Archived page

Original page

Visual

Functionality

Low Cost

Reduced storage

Computationally lightweight
Web Archives to the Rescue

~700 Billion pages, ~100 Petabytes
Web Archives to the Rescue

To tackle this problem of ephemeral web, various web archives have been established.
Web Archives to the Rescue
Web Archives to the Rescue
Web Archives to the Rescue
Approach to Fix Fidelity

➔ Do not make DRP APIs deterministic

➔ **Leverage server-side matching** techniques to account for such URL differences

➔ Enforce same client characteristics for loads as the ones used for crawling
How Modern Web Archives Operate

Crawling
- Initialize w/ list of URLs
- Visit & download all web resources

Storing
- For each resource store headers, payload
- Apply deduplication

Replay
- Export timestamps of snapshots
- Replay each page by iteratively fetching resources
Non-determinism Resulting in Poor Fidelity

Taken from Ronny et al., NSDI'21
Strawman: Preserve Code Executed During Page Load
Make JavaScript Execution Completely Deterministic
Key Insight #1: Archived Page ≠ Live Page

B. Certain sources of non-determinism are absent

- **Print login button**: Absent on Live Page
- **IsUserSubscribed()**: Present on Archived Page, absent on Live Page
- **Print account info**: Present on Archived Page, absent on Live Page
- **Queries server-side state**: Present on Archived Page, absent on Live Page
  - **Always returns the same response across loads**: Present on Archived Page
- **enableTelemetry()**: Absent on Archived Page
- **AddTelemetryEventHandlers()**: Absent on Archived Page