Efficient Web Archive Searching

CS4624 - Multimedia/Hypertext
Dr. Edward Fox

Ming Cheng, Xiaolin Zhou, Jinyang Li, Yijing Wu, Lin Zhang
Virginia Tech, Blacksburg VA 24061
April 28, 2020
Outline

1. Project Overview
2. Project Design
3. Implementation
4. Benchmark & Result
5. Future Plans
6. Project Timeline
7. Acknowledgements
Project Overview

a. 25 Petabytes Archived
b. 436 Billion Web Captures
c. 4 Million eBooks
d. 3 Million Hours of TV

HOW TO BUILD AN ARCHIVE

20 years in the making
Project Design

Parquet

APACHE ARROW

Hashing

Apache Zeppelin

BenchMark Result
Implementations

● Our Implementation:
  ○ Simple Hash

● Existing algorithms:
  ○ Spooky Hash
  ○ Murmur Hash
  ○ City Hash
  ○ XXHash
Simple Hash Algorithm Breakdown

1. Protocols + Domain Only

2. Remove Special Characters

3. Simple Hash
   - https -> 0
   - http -> 1
   - first-level domain
   - 2 bytes 0-9, a-z, A-Z
   - 2 * 62 = 3844 > #FLD

4. Result
   - 0Q2canvasvt
Simple Hash

![Graph: Original URL vs. Simple Hash Value](image)
Spooky Hash, Murmum Hash, City Hash, XX Hash
Benchmark

Tool: Zeppelin

1. All hash results + original database → Combined.parquet

2. Run Zeppelin Web Server, and go to: localhost:8080

3. Load parquet file, random pick 1000 samples

4. Clean cache, start query by originalURL and hash results

5. Calculate average single URL query time
Result

Average single URL query time cost with 1000 random selected samples
Future Plans

1. Reduce collisions.
2. Larger benchmark database.
3. Run benchmark multiple times and take the average.
Project Timeline

- **2/7**: Understand project goal
- **2/14**: Collect & review paper
- **2/27**: Decide/define a suitable hashing algorithm
- **3/13**: Algorithm implementation
- **3/20**: Test hashing algorithm implementation
- **4/3**: Interim report
- **4/10**: Benchmark
- **4/26**: Final report
- **5/2**: Finalize VTechWorks
- **5/4**: ICAT Day
Acknowledgements

Xinyue Wang

Dr. Edward A. Fox

Dr. Lenwood S. Heath
Reference


