SWINBURNE

CENTRE FOR ADVANCED INTERNET ARCHITECTURES

Dynamics and Cachability of Web Sites: Implications for Inverted Capacity Networks

Sebastian Zander, Grenville Armitage, *Clancy Malcolm*

zander@fokus.fraunhofer.de

{garmitage, cmalcolm}@swin.edu.au

Overview

- Inverted Capacity Networks
- Web Dynamics & Cachability
- Methodology
- Experiment Results
- Conclusions
- Future Work

Inverted Capacity Networks

- Imagine a world where all the bandwidth was around the edges rather than the core
  - E.g. massive fibre to the home deployment
  - Neighbourhoods become local meshes of short-haul high bandwidth connectivity
  - Libraries could host neighbourhood web caches, revitalising their role as information repositories for their communities
- Will this improve the "user experience" enough to justify government or "social good" programs to fund?
  - Need to model likely performance improvements
  - Need to model cachability of content on today's Internet

Web Dynamics & Cachability

- (Still) most important content type: Web
- Move web content closer to the user
  - Limit and smooth traffic into core network
  - Decrease latencies observed by user
- Gain depends on
  - Cachability: how much of the Web is cachable?
    - Cachable according to expiration and validation (HTTP 1.1)
  - Dynamics: how is the Web changing over time?
    - Change: the content as contained in the response has changed between two consecutive visits

Methodology – Active vs. Passive

- Passive
  - Analysis of web server/proxy logs (insufficient information!)
  - Sniffing and analysis of server/proxy traffic
- Active
  - Actively request objects and analyse responses
- ➢ We choose the active approach
  - Unbiased: independent of short term user group behaviour and content popularity
  - Controlled: e.g. regular visit interval
  - No infrastructure concerns: no access to provider network needed
  - No privacy/security concerns
  - Drawback: generated traffic

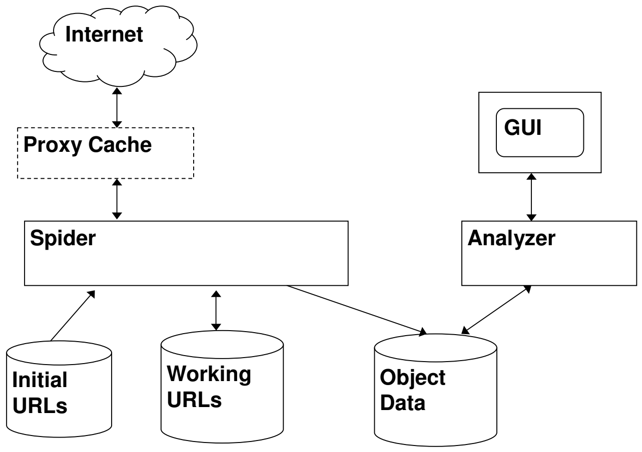
Methodology - Dynamics

- Change Detection
  - Object (response body) may not have a unique identifier
  - Even if it has it one (e.g. ETag) it can't be trusted
  - ➢ Generate a "unique" hash value for an object (CRC32 or MD5)
  - ➢ Object changed if the hash value of the object changed
- In combination with visit timestamps
  - ➢ Time between changes
  - ➢ Visit/change ratio
- ➢ Furthermore
  - ➢ Age of objects
  - ➢ Duplication

Methodology - Cachability

- We consider objects as cachable if expiration or validation (or both) are possible
- Reasons for being not cachable (expiration)
  - Uncachable HTTP method
  - No freshness information
  - Stale (has expired)
  - Cache-Control or Pragma forbids caching
  - Uncachable response
  - Cookies
  - Dynamic URL (? parameter or "cgi-bin" in URL)
- Reasons for being not cachable (validation)
  - Missing Etag and Last-Modified

Methodology - Architecture

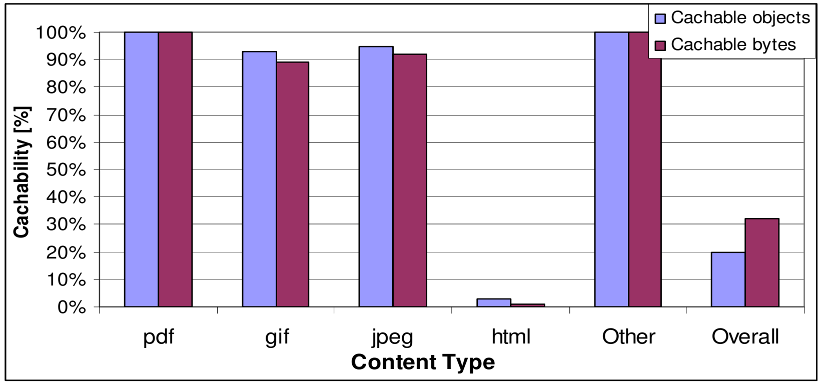


Experiment Results

- Observed 6 web sites for 2 weeks
  - 3 commercial
  - 3 university/government
  - Popular among local users as indicated by a web proxy log
  - >500,000 URLs (URI + parameters)
  - ~15 GB content size
- Actually the sites have been observed for a longer period but only a two week period has been analysed
- Visit interval for all URLS: 1 day

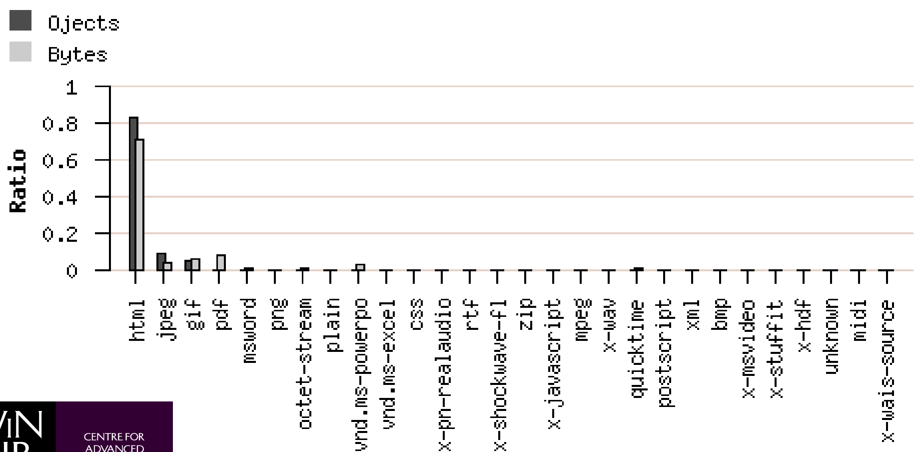
Experiment Results

Cachability of the most common content types



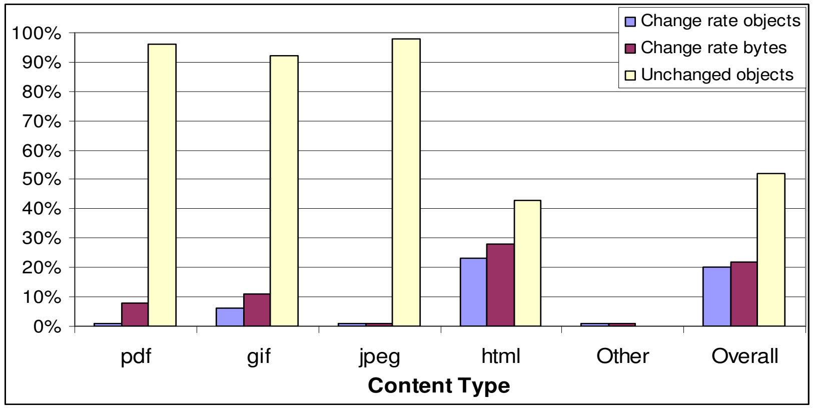
Experiment Results

Content types sorted by object count



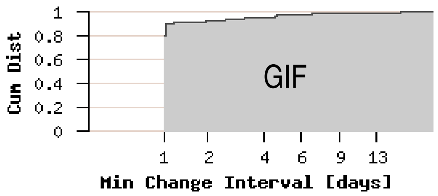
Experiment Results

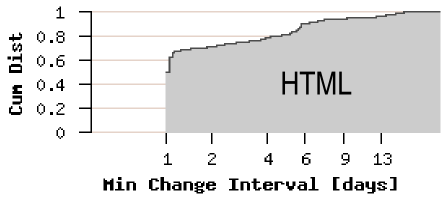
Rate of change and unchanged objects for the most common content types

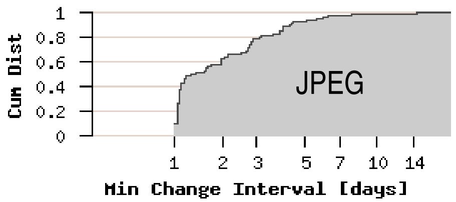


Experiment Results

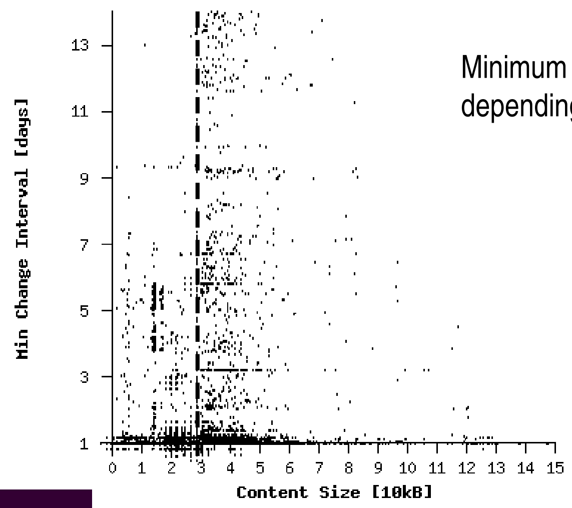
Minimum change intervals for different content types







Experiment Results



Minimum change interval depending on content size

Experiment Results

- Only a small fraction (20%/32%) of the investigated objects/bytes is cachable. The main reason are html objects which we assume are dynamically generated
- On average uncachable objects are smaller (26kB) than cachable objects (40kB)
- 52% of the objects did not change at all in a 2 week period. Of the changed objects 10-40% (depending on content type) changed with a minimum interval of more than 1 day
- Smaller objects seem to smaller minimum time intervals between changes
- 7% of the URLs were duplicated at least twice

Interim Conclusions

- Capacity inverted network infrastructure is advantageous if the content can be cached in the high capacity part close to the user
- Investigated today's Web content distribution
  - For the observed content everything except html has a high cachability
  - Cachability of observed html content is very low; much lower as it could be considering the dynamics
- Limited scope of the experiment but the Web is simply too BIG

Shortfalls

- Active approach uses a large amount of bandwidth and increases load of the sites under investigation
  - Can not handle too many URLs
  - Can not handle small visit intervals
- Spider can not make POST requests
- Spider can not send cookies (although it receives them)
- Spider can not handle HTTP authentication

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

Future Work

- More analysis e.g. growth of sites
- Passive measurement and comparison/combination with active approach
- Hybrid approach
  - Passively obtain URL set
    - Based on user popularity (access logs)
    - Based on the sites themselves (site structure, content, …)
  - Actively scan the URL set (active)
- Adaptive sampling
  - Adjust sampling interval based on observed cachability and dynamics
- Improve cachability of dynamically generated content

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES

Thanks for your attention!

SWINBURNE CENTRE FOR ADVANCED INTERNET ARCHITECTURES