

Document Attrition

Data Format

Stephen Wattam

January 11, 2011

1 Introduction

Choosing data for later analysis is a nontrivial task.

One argument is to retain as much information as possible, including full HTTP headers, timeouts and other network parameters. This has the advantage of being prunable later on to pertinent information, however, many factors are likely to reduce the quality of network-level information:

- As the documents begin to ‘die off’ network load will decrease as the number of parallel TCP connections lowers
- Changes in server and the order in which files are downloaded may affect times in a manner which is difficult to predict

Another argument is to store only linguistic content information. This approach would mean storing least, only the content and a marker to show missing pages, but requires removal of page code and other things that are, perhaps, best accomplished during analysis.

A middle ground, then, shall be the final spec: retain enough of the header to judge HTTP changes such as server config or ownership, but retain only basic timing and network information. Retain the whole HTTP return, but do not store images and other non-text content (doing so would punish both the server, local storage, and local network).

1.1 Storage

Metadata regarding a page request will be stored in a database (or flatfile database on-disk). Data must be keyed by URL (or link ID) and request ‘session’ (storing the exact time would lead to incomparable

snapshots through time).

The request session itself should take as little time as possible, and occur as frequently as possible. Practically (though the project is at an early stage) I believe a million URLs may be downloadable in under a day, once a week. This does offer $\pm \frac{1}{7}$ th as jitter, but I shall have to benchmark on the final list before gathering data.

2 Format

- Numeric ID, used as a faster and more compatible form of the URL
- URL, unique
- Sample, the period ID this sample was downloaded at
- Header, the raw HTTP header
- RTT, the download time
- Size, the size of the download, in bytes
- Encoding, the content-encoding
- Content, the raw http content