IIPC
international internet preservation consortium

Characterizing Change in Web Archiving

Status:	Working draft
Author:	Andrew Boyko, Library of Congress (aboy@loc.gov)
Date of issue:	27 August 2004
Reference:	
Number of pages:	9

Table of Contents

1 Document Control

Issue	***Date of issue***	***Comments***
0.1	27 August 2004	First draft

2 Introduction

We harvest the web because web content changes over time – not merely in the manner of a publication with multiple, well-defined editions at predictable intervals, but in a continuous, fluid way. From one perspective, the Web from its inception to the present is a single entity, continuously changing at every level of detail. Our harvests capture snapshots of the entity, but just as snapshots cannot accurately depict a live event, web crawls only provide an instantaneous slice from the continuum.

By examining and characterizing the change of the Web, we enable harvesters to better focus their interest and harvest more productively, and we enable researchers to better understand and describe the content harvested. While even simple documents available on the web can be expected to change under our watch, the inherently dynamic nature of so much web content requires an even closer watch and a better understanding of the nature of change. In this document, we attempt to define the characteristics and dimensions of change in web content.

3 Goals

3.1 Goals for Analysis of Past Crawls

Given the considerable archives collected by institutions to date, we wish to define approaches, tools, and metrics for describing how content acquired repeatedly may have changed over time. Regardless of how frequently we may have acquired a given resource or site, we have no way of knowing definitively how the actual resource changed.

Nevertheless, we can characterize the change in the content we hold, and attempt to at least gauge what our acquisitions can tell us about the actual rate and type of change.

We set forth the following goals:

1. Given a set of repeated crawls of a given seed, produce measurements for what changes occurred across the crawl output, in order to guide researchers examining the content.

2. Given a large set of varied crawls, develop statistics describing change based on attributes of the content (e.g. content types, type of site, technical implementations), in order to characterize typical patterns across the Web at large.

3.2 Goals for Future Crawls

As we have now begun to amass confidence in our crawl tools and processes with respect to individual crawls, we attend to the on-going task of repeatedly capturing target content that continues to change during our period of interest in it. Until now, the frequency of capture has been largely based on intuitive guesses as to the rate of change of the content, constrained by available time and human resources, and by storage and other computing resources. While we cannot avoid these constraints, our intuitive guesses at frequency can be replaced (or at least augmented) by an understanding of the crawl target, provided by data from previous crawls of that target or other similar content.

Our goals include:

1. Given a set of repeated crawls of a given seed, assess the best frequency of future crawls of the same seed in order to capture as much change as possible without wasting storage, bandwidth, or goodwill.

2. Given a seed not previously crawled, predict an appropriate initial crawl frequency, by relating attributes of the new site to similar sites with known change characteristics.

4 Definitions

Having described our intent, we must now delve into what we mean by *change* above. We consider change at three levels of detail:

- **item change**: change in an individual resource (a single URL referring to a document or media object)

- **site change**: change in the structure of a site (defined as some group of inter-linked documents and embedded content)
- **web change**: change in the structure of a web (defined as a group of inter-linked sites)

4.1 Item Change

The elemental question of item change is: given two copies of a piece of content, retrieved at different times, did it change during that interval? The precise definition of *change* in that question is left open, as the criteria may vary with application; let us consider how they will vary.

The most relevant type of content change, and the kind with the most breadth, is that of text documents. Consider a continuum of possible definitions of *change* of such documents, varying in their tolerance. At one end of the continuum might be a strict definition of change as the alteration of a single bit of the content (the comparison of which might be typically implemented with a hash function like MD5). At the other extreme might be a definition of change as nothing less than a significant semantic alteration of the content (which would require natural language analysis to properly detect). Both extremes, as well as points between them, might be relevant for different applications. We must note that increasing the discrimination of what constitutes change (toward the second extreme given) also increases the complexity of the definition and the difficulty of the implementation.

Assuming HTML as the document format of greatest interest, we consider two instances of an HTML document, the change between which we are trying to assess. In order of increasing complexity, but also increasing abstraction, we might consider:

- has any bit in the stream of octets representing the message changed?
- having interpreted the character encoding of the documents, does the actual stream of characters differ? (abstract of the raw bitstream, thus ignoring changes in encoding)
- is the structure of the actual HTML markup different? (abstract of the character stream, thus ignoring changes in whitespace)
- is the actual document structure different? (abstract of the particular markup and styling used, ignoring cosmetic issues)
- is the content within the page different? (abstract of the page headers, navigation, and other text that appears on the page)

- is the logical *meaning* of content of the page different? (abstract of the particular phrasing of the text)

It is straightforward to conceive of tools that implement these approaches. The Wayback Machine uses one approach; the use of MD5 checksums would be another. It is reasonable to consider and adjust the Wayback approach. In addition to omitting the formatting tags in HTML, we might also attempt to identify unimportant textual content (such as a server-generated representation of today's date, assuming it is possible to detect).

Stepping back from HTML, we consider other document types. PDF or MS Word documents resemble HTML superficially (in that they are documents with markup) but it is reasonable to assume they have significantly different likelihood of change over time based on typical uses; the storage format for these documents is more complex, and difficult to directly compare. Image files (JPEGs or GIFs) might be expected to change even less frequently, and only a simple bit-wise comparison is reasonable without employing sophisticated image processing. Media content such as audio or video provides even more complexity. For all these more complex types, comparison of embedded metadata may provide the only level of abstraction of change above the bitwise comparison of the raw data.

External to the document type, the metadata provided when the document is delivered by HTTP can provide some clues to change, or at least to the hosting server's perception of the document's change. The HTTP Date header typically indicates when the resource last had any change (presumably at the bit level). Relatedly, the ETag header provides an opaque value that is intended to indicate a document's identity. Lars Clausen [2] has computed the correlation between the information provided by these headers and other measures of change. Similarly, syndication formats such as RSS indicate precisely the content that the site's owner perceived to have changed meaningfully, which could be correlated with our own external measurements.

Having described some means for calculating whether a document has changed, we reasonably extend the concept toward describing the magnitude of change. For textual change, we might consider the following measurements:

- The degree of textual change in two instances of a document (using any particular change criterion above), perhaps quantified as one of:
    - the percentage of a document
    - a count of instances of change
    - some other definition of a unitless "change factor".
- An average count of changes in a given document per some unit of time
- Change to the counts of some important elements of the content type of the document (e.g. for HTML documents, outbound links or META tags)
- Change to some selected portion of certain documents of interest

4.2 Site Change

Armed with some insight into item change, we can expand the scope of our interest outward, to a set of items comprising a web site[1], and attempt to characterize change across such a set. Consider the trivial case of a small Web site, consisting of static documents in a simply linked tree, in isolation from the wider Internet. If we consider this site as a group of documents, its change over time includes changes to the documents themselves, changes to the set of documents (addition and removal), and changes to the relationships (linkages) between the documents. If we consider the site as a single entity, we may track changes to the characteristics of the site as a whole, encompassing traits as disparate as legal ownership of the site, document authorship, the site-wide graphic design or technical implementation, or perhaps even a change in the political orientation of the site's editorial perspective.

For the general case of change across a site, we might measure:

- Counts of pages removed, added, and altered within a site
- Change in the counts or characteristics of links, perhaps both within the site of interest and outside of it
- Change to common elements to documents on a site, which can be considered a single change to the site as a whole (e.g. site-wide graphic elements and layout, metadata, or perhaps the technology implementing the site)
- Patterns of change within a site, correlating document change to aspects of the document's position within the site (e.g. its link distance from the site's entry point, or its inward or outward link density)

4.3 Broad Change

Expanding our definitions even further outward, we may consider how a set of sites (an entire crawl, repeated over time, or perhaps even an entire harvested collection) has changed over time, using the same logic. Individual sites may appear or disappear, and the link relationships between sites may change.

As above, metrics might include:

[1] It is challenging to precisely define "web site" in the broad context of archiving; for our immediate purpose, a loose, intuitive definition will suffice.

- Counts of sites added, removed, or altered within the scope of interest
- Change in the counts or characteristics of links between the sites or outside the scope

5 Future Directions

5.1 Expectations of Change

We have certain intuitive expectations of likely patterns of change for typical documents or sites, based on the type of site, the intended use of the particular document, or the particular content type. We enumerate some here, with no strong sense of their validity; by validating these against harvested content using the metrics listed above, we begin to be able to guide future crawl activity.

A small set of starting assumptions that our metrics may validate:

- HTML documents will change more often than PDF, Word, or other more opaque formats.
- Text-oriented documents will change more often than media content (images, audio, video); most media content will remain unchanged or else disappear, rather than change at the same URL.
- The entry page to a site is more likely to change than deeper content.
- Documents with a single date embedded in the page or the URL are likely to remain fairly constant; on the other hand, pages with multiple dates are more likely to be continually updated. In particular, the entry page to sites identifiable as “blogs” is particularly likely to change.

These are conjectures, and we may expect to discover other unexpected patterns and rules as we analyze wider sets of data.

5.2 Toward Implementation

As mentioned in the discussion of item change, the two most visible implementations of document change detection are a simple bit-wise comparison (most commonly by comparing a hash code), and the Wayback Machine’s markup-independent HTML comparator. As Wayback is not open-source, it would be valuable for the community to define and implement a similar comparator algorithm for HTML, which could be integrated into access and analysis tools.

With those two comparators in hand, it should be straightforward to construct tools to compute the various item change metrics enumerated above. Several of the metrics at the site and web scope, however, require more analysis before it is clear how to build appropriate measuring tools.

5.3 Reaching the Goals

The metrics enumerated here do not straightforwardly reach the goals listed in section 3. While they drive us toward being able to characterize change in past crawls, taking the leap to confident predictions about future crawl parameters will require analysis, validation, and refinement of the metrics listed above.

6 References

1. Thorsteinn Hallgrimsson, “Information Required by a Web Harvesting Institution”, July 2004

2. Lars Clausen, "Concerning Etags and Datestamps", July 2004