Who and What Links to the Internet Archive

Yasmin Alnoamany, Ahmed Alsum, Michele C. Weigle, and Michael L. Nelson

Old Dominion University, Department of Computer Science
Norfolk VA 23529, USA
{yasmin,aalsum,mweigle,mln}@cs.odu.edu

Abstract. The Internet Archive's (IA) Wayback Machine is the largest and oldest public web archive and has become a significant repository of our recent history and cultural heritage. Despite its importance, there has been little research about how it is discovered and used. Based on web access logs, we analyze what users are looking for, why they come to IA, where they come from, and how pages link to IA. We find that users request English pages the most, followed by the European languages. Most human users come to web archives because they do not find the requested pages on the live web. About 65% of the requested archived pages no longer exist on the live web. We find that more than 82% of human sessions connect to the Wayback Machine via referrals from other web sites, while only 15% of robots have referrers. Most of the links (86%) from websites are to individual archived pages at specific points in time, and of those 83% no longer exist on the live web.

Keywords: Web Archiving, Web Server Logs, Web Usage Mining, Language Detection.

1 Introduction

A variety of research has been conducted for studying web archives in order to answer questions related to user needs and to present web archive data to users [12,5]. However, no previous work has been carried out to answer these questions: What content languages are web archive users looking for? Why do users come to web archives? Where do web archive users come from? Who links to web archives? How do sites link to web archives? Do sites link deeply to specific archived pages or link to the repository? Why do sites link to the past?

The Internet Archive [11] is the first web archiving initiative attempting global scope and currently holds over 240 billion web pages with archives as far back as 1996 [8]. It allows traveling back in time for traversing archived versions of web pages through the Wayback Machine [18]. This paper provides a study of the requests of web archive users, both humans and robots, to gain insight into what users look for, in the context of the language of the requested pages, through an analysis of the server logs of the Internet Archives' Wayback Machine. We also provide an analysis of referring pages of human users to investigate how humans discover the Wayback Machine, why the referrers link to web archives, and how they link to web archives.

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 346–357, 2013.

We found that users of Internet Archive's Wayback Machine request English pages the most, followed by several European languages. We also found that most human users come to the Wayback Machine via links or direct address presumably because they did not find the requested pages on the live web. Of the requested archived pages, 65% do not currently exist on the live web. From analyzing the referrers, we found that more than 82% of human sessions have referrers, while only 15% of robot sessions have referrers. We also found that 86% of the referrers are deep links to archived pages.

2 Related Work

To the best of our knowledge, no prior study has analyzed where web archive users come from nor what they look for in terms of the linguistic context. Furthermore, the usage of web archives in general has not been widely studied. The characterization of search behavior and the information needs of web archive users have been studied by Costa et al. [4,5] based on quantitative analysis of the Portuguese Web Archive (PWA) search logs. In a previous study [1], we provided the first analysis of user access to a large web archive. We discovered four basic access patterns for web archives through analysis of web server logs from the Internet Archive's Wayback Machine. In the study, we applied heuristics for robot detection after data filtering and found that robot sessions outnumber human sessions 10:1. Robots outnumber humans in terms of raw, unfiltered requests 5:4, and 4:1 in terms of megabytes transferred.

Many studies have investigated what is missing from digital libraries and web archives, in addition to the effect of this on the satisfaction of users' needs and expectations [17,3,22,16]. In [17], the Internet Archive's coverage of the web was investigated. The results showed an unintentional international bias through uneven representation of different countries in the archive. Carmel et al. [3] suggest a tool to dynamically analyze the query logs of the digital library system, identify the missing content queries, and then direct the system to obtain the missing data. We investigate what is missing through an analysis of requests with an HTTP 404 status in the Wayback Machine web server logs.

Memento Terminology

In this section, we explain the terminology we adopt in the rest of the paper. Memento [20] is an HTTP protocol extension which enables time travel on the web by linking the current resources with their prior state. Memento defines the following terms:

- URI-R identifies the original resource. It is the resource as it used to appear on the live web. A URI-R may have 0 or more mementos (URI-Ms).
- URI-M identifies an archived snapshot of the URI-R at a specific datetime, which is called Memento-Datetime, e.g., URI-M$_i$= URI-R@t_i.
- URI-T identifies a TimeMap, a resource that provides a list of mementos (URI-Ms) for a URI-R with their Memento-Datetimes, e.g., $URI-T(URI-R) = \{URI-M_1, URI-M_2, ..., URI-M_n\}$.