Grading Degradation in an Institutionally Managed Repository

Luis Meneses, Sampath Jayarathna, Richard Furuta and Frank Shipman

Center for the Study of Digital Libraries and Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112 USA
(ldmm, sampath, furuta, shipman)@cse.tamu.edu

ABSTRACT

It is not unusual for digital collections to degrade and suffer from problems associated with unexpected change. In an analysis of the ACM conference list, we found that categorizing the degree of change affecting a digital collection over time is a difficult task. More specifically, we found that categorizing this degree of change is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. It is, in part, a characterization of the intent of the change. In this work, we examine and categorize the various degrees of change that digital documents endure within the boundaries of an institutionally managed repository.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, systems issues, user issues.*

General Terms

Management, Design, Reliability, Experimentation, Verification

ywords

Web resource management; distributed collections; web change classification

1. INTRODUCTION

Imagine a library filled with books that have missing pages. It might seem as overly exaggerated, but that metaphor can be used to depict the state of the digital repositories that have been affected by unexpected change. We have found that electronic resources can change, both intentionally and unintentionally, because of different factors and circumstances. Change can occur because of deliberate actions on part of the collector, unexpected events or may be due to other uncontrollable factors.

However, Web documents are not static resources and a certain degree of change is expected from them [1]. For example – and taking into account the specific infrastructure of Walden's Paths [2], where decentralized collections are stored and represented as traversable paths containing multiple nodes and documents – we have observed that Web resources suffer from changes in content, layout, presentation and location. However, as members of a larger assembly, these documents are expected to either change little over time or mutate harmoniously and accordingly with the other documents in order to preserve the semantic meaning and systematic order of the collection.

Additionally, distributed collections that are hosted in institutional repositories operate under the assumption that they are more resilient and able to withstand change. Because of their focus on long-term storage, these repositories have different attributes and operate under different principles when compared to the Web as a whole. Surprisingly, we have found these features – which are often found in digital repositories emphasizing in long-term storage – do not fully preserve the referenced documents and make them impervious to change. For example, as of 9/27/2014 the ACM Digital Library had 15 unique links referencing the different sites in the JCDL conference series and 8 of them report errors or point to the wrong content.

More so, we also found that categorizing the degree of change affecting the documents in a digital collection over time is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. Previous work on this topic has and relied on methods based on response codes, monitoring the fluctuation in file sizes and analyzing the documents content. However, we believe there is growing need for a categorization framework that will allow us to tag and sort documents that have been affected by different amounts of change.

2. GRADING DEGRADATION

The corpus for our study is the Association for Computing Machinery list of conference proceedings (http://dl.acm.org/proceedings.cfm), which we retrieved on 9/27/2014 and used as our starting point. Then we followed each hyperlink to a metadata page that displayed basic information for each corresponding conference and workshop, which in turn allowed us to extract the external URLs. As a result of this procedure, we were able to extract 6086 URLs – out of which 2001 were unique.

We then proceeded to inspect and categorize the 1492 pages that were retrieved with a successful HTTP response code. We categorized these pages into three categories by evaluating the relationship between the anchor text and the corresponding retrieved page. As a result of this categorization, we found that 917 pages were "clearly correct" and 531 were incorrect. Additionally, we were unable to evaluate 44 pages because their contents didn't provide us enough information to make an accurate assessment.

Then we analyzed the 531 pages that were incorrect in an effort to understand how conference sites degrade over time. The coding

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.
ACM 978-1-4503-3594-2/15/06.

scheme evolved through examination of the particular collection rather than using a pre-defined classification scheme. In the end, nine categories were used to classify the "incorrect" pages, which we list in (approximate) order of severity. These nine groups provide insight regarding the different stages that conference pages go through until they are ultimately abandoned:

1. **Kind of correct:** (197 entries) Pages that contain related content, but they do not fully match the semantic concept encapsulated in the anchor text. When taking into account conference proceedings, these pages often link to a different year in the conference series. For example: Anchor text "Conference X 2006" references the Conference X 2009 site.
2. **University/institution pages:** (36 entries) These are cases that surface when a site has been taken down, but the server configuration redirects the user to its parent institution. In cases dealing with conference sites, servers would usually redirect the user to the website of the University that hosted the conference or to a related professional organization.
3. **Directory listings pages:** (18 entries) Pages displaying a listing of files or a "Hello World" page. Probably caused by an error in the server configuration. In these cases, the original content looks to still be available but the new web server does not recognize homepage.html as a default page to view for this URL.
4. **Blank pages:** (141 entries) Pages that returned no content.
5. **Failed redirects:** (30 entries)
6. **Error pages:** (17 entries) Pages that specifically state that an error has occurred.
7. **Pages in a different language:** (32 entries) Pages that didn't match the language found in the anchor text. Most of these pages were in a language different than English.
8. **Domain for sale pages:** (17 entries) Pages that indicated that the domain name registration has lapsed and it is being sold by a registrar, or taken over by a third party in order to profit from the sale.
9. **Deceiving pages:** (43 entries) Pages that have been taken over by a third party. The content displayed in these pages is totally unrelated to the original purpose of the site. We believe many of these pages were not created to deceive users, but as an attempt to manipulate the PageRank algorithm [3].

Many of these links still lead to information related to the original purpose but clearly not to the originally intended materials. There are a number of categories that result when no content is available depending on how the servers are configured – blank pages, failed redirects, some directory listings, error pages, and university/institutional pages. In some of these cases, these pages can be detected with previous work on identifying Soft 404 error pages [4, 5]. The remaining pages are perhaps the most problematic, when the web address has been taken over and is either for sale or being used for other purposes.

3. DISCUSSION

The analysis of the conference website links within the ACM Digital Library shows that institutional archives are not immune to the challenges of distributed collection management. Upon our initial assessment, 404 HTTP errors were more prevalent in our corpus. However, upon further inspection we found that 36% of the pages that were supposed to be correct as reported by their HTTP response code were actually incorrect. For us this is a clear indication that the correctness of a web page is relative and that there is a growing need for methods to categorize and locate likely problematic resources that might require the attention of collection managers.

Being able to distinguish between the "kind of correct" and "deceiving" pages is important to collection managers. So, why are the documents in the "deceiving pages" category created? Although the pages in this category are very diverse in content and presentation, they do share two characteristics. First, the number of links that point to other pages within the site is much greater than the number of out-links. On average, pages in the "deceiving" category had 66 links, which is more than twice the average in the "correct" and "kind of correct" categories (20 and 27 links respectively). And second, the domain names that host these pages once belonged to a reputable institution for number of years (i.e., a conference series) before being abandoned. Consequently, these abandoned domain names have value – not necessarily due to current network traffic but in the perception of their authority/validity. We could hypothesize that these pages are created to manipulate pageRank scores by utilizing a large number of links from a page that once had a high PageRank but has been taken over by a third party. This problem becomes increasingly interesting when we consider that the cost of creating a web page is small and that some search engines (most notably Google) do not share the overall rankings for their indexed sites, which can lead some parties to abuse these malicious techniques.

Our research focuses on investigating methods to detect unexpected changes in Web documents within a collection. However, the degree of change that we are focusing on falls within a specific range: not as subtle as a few terms substitutions in the body of a Web page and not as dramatic causing servers to report errors explicitly. Detecting the instances that fall between these two extreme cases is difficult and therefore requires the assistance of a classification system.

4. ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grants DUE–0840715 and DUE–1044212.

5. REFERENCES

[1] P. L. Bogen, R. Furuta, and F. Shipman, "A quantitative evaluation of techniques for detection of abnormal change events in blogs," presented at the Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC, USA, 2012.

[2] P. Dave, U. P. Karadkar, R. Furuta, L. Francisco-Revilla, F. Shipman, S. Dash, *et al.*, "Browsing intricately interconnected paths," in *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia - HYPERTEXT '03*, Nottingham, UK, 2003, pp. 95 - 103.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford University1999.

[4] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, "Sic transit gloria telae: towards an understanding of the web's decay," in *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, 2004.

[5] L. Meneses, R. Furuta, and F. M. Shipman, "Identifying "Soft 404" Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections," in *Proceedings of Theory and Practice of Digital Libraries 2012*, Paphos, Cyprus, 2012.