Evaluating Unexpected Change in a Distributed Collection

Luis Meneses, 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, furuta, shipman)@cse.tamu.edu

Curating a digital collection is not an easy task. It is not unusual for digital collections to degrade and suffer from problems associated with unexpected change. Many research groups (including ours) have examined the implications of the Web's decentralized administration on the stability of materials. [1-4]. Furthermore, in an analysis of the Association for Computing Machinery (ACM) conference list, we found that categorizing the degree of change affecting a digital collection over time is a difficult task and in part, a characterization of the intent of the change [5]. As a follow up of this study, we carried out a user survey that evaluated how change is perceived within the context of a distributed collection.

The corpus for our user study is the ACM list of conference proceedings (http://dl.acm.org/proceedings.cfm), which we retrieved on 9/27/2014. We were able to extract 1492 documents with a 200 HTTP response code, which we categorized according to their textual content: 917 pages were "clearly correct" and 531 were incorrect (44 didn't provide us enough information to make an accurate assessment). We then divided the 531 incorrect documents into 9 categories: kind of correct, university, directory listings, blank, failed redirects, error, different languages, domain for sale, and deceiving pages.

The user study was administered through the Web and consisted of two sections. The first section of the study prompted the participants for their demographic data. The second section was the main part of the study: participants were asked to go over fifty documents from the ACM corpus. The fifty documents were randomly selected for each session and consisted of 25 correct and 25 incorrect documents. In the end, we collected data from 62 participants – mostly upper-level Computer Science undergraduates – and assessed the validity of 2875 documents.

Interestingly, the participants of the study did not have difficulties identifying documents in the correct category. However, this was not the case in the all the other categories. Table 1 shows a summary of the user responses. Documents in the correct category were correctly identified in 71% of the cases, but users were not able to correctly identify the documents that were incorrect. Surprisingly, documents in the "error page" category were incorrect identified in 98% of the cases. Similar scenarios occurred in instances were it was evident that the documents were incorrect: such as in the "domain for sale" and "hello world" categories. The identification did marginally improve in less evident cases, for example in the "deceiving", "kind of correct" and "not correct" categories. Additionally, the language of the text and explicit error codes did not influence the user responses. On the other hand, layout, presentation and content of the documents were significant factors in the classification.

Taking into account the results of the user evaluation and the inability of the test subjects to identify incorrect pages in the user study raises some alarming issues. For instance, is managing the effects of unexpected changes in digital collection a more severe problem than what was originally anticipated? We believe it is. Taking into account that the study participants for the most part identified incorrect pages as false positives (pages that belonged in the collection because their content was correct) is an indication that there is a need for computer-based methods to identify the validity of documents in digital collections.

1. REFERENCES

[1] M. Klein, J. Ware, and M. L. Nelson, "Rediscovering missing web pages using link neighborhood lexical signatures," in *Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital libraries*, Ottawa, Ontario, Canada, 2011.

[2] H. M. SalahEldeen and M. L. Nelson, "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?," in *Proceedings of Theory and Practice of Digital Libraries 2012*, Paphos, Cyprus, 2012.

[3] L. Francisco-Revilla, F. Shipman, R. Furuta, U. Karadkar, and A. Arora, "Managing change on the web," in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, Virginia, United States, 2001.

[4] L. Meneses, H. Barthwal, S. Singh, R. Furuta, and F. Shipman, "Restoring Semantically Incomplete Document Collections Using Lexical Signatures," in *Research and Advanced Technology for Digital Libraries*. vol. 8092, T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. Farrugia, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 321-332.

[5] L. Meneses, S. Jayarathna, R. Furuta, and F. Shipman, "Grading Degradation in an Institutionally Managed Repository," presented at the Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, Knoxville, Tennessee, USA, 2015.

Table 12: User responses for the document classification in the user study by categories. Shaded: Correct – Not shaded: Incorrect

	Correct	Error	Deceiving	Hello World	Kind of Correct	Not Correct	Domain for Sale	University
Very Much	456	54	99	54	242	56	59	59
Somewhat	567	5	32	7	299	80	5	37
Undecided	74	0	10	7	45	17	3	10
Not Really	178	0	21	0	75	13	5	14
Not At All	163	1	29	4	63	13	6	13
Totals:	1438	60	191	72	724	179	78	133
Correctly Identified	0.71	0.02	0.31	0.15	0.25	0.24	0.18	0.28
Incorrectly Identified	0.29	0.98	0.69	0.85	0.75	0.76	0.82	0.72