Analyzing the Perceptions of Change in a Distributed Collection of Web Documents

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 documents on the Web to degrade and suffer from problems associated with unexpected change. In an analysis of the Association for Computing Machinery conference list, we found that categorizing the degree of change affecting digital documents 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 paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change in the ACM conference list?

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.

Keywords

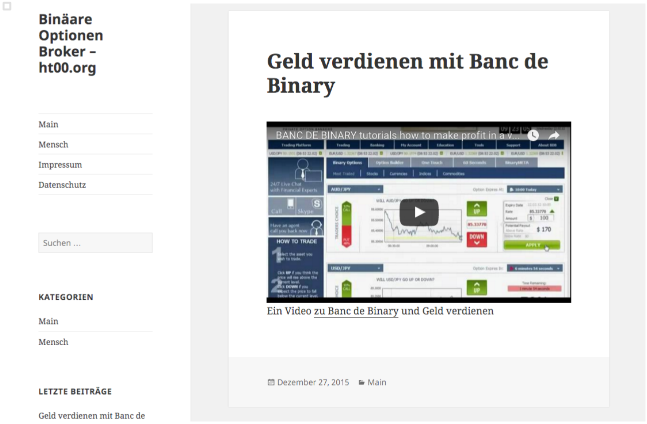
Web resource management, distributed collections, web change classification.

1. INTRODUCTION

Imagine a library filled with books that have missing pages. It might seem as overly exaggerated, but that metaphor can depict the state of digital collections that have been affected by unexpected change. It is not unusual for digital documents to have problems associated with the persistence of links, especially when dealing with references to external resources. External resources on the Web are highly volatile and prone to exhibit unexpected change as cases of “broken links” [1] or “linkrot” [2]. Therefore, our work has been motivated to mitigate the impact of unexpected change in documents stored in decentralized collections [3].

We have found that electronic resources can change, both intentionally and unintentionally, because of different factors and circumstances. More so, Web documents are not static resources and a certain degree of change is expected from them [4]. Our current efforts continue long-standing study of the problems that surface when managing distributed collections and curating missing resources [5-7].

Distributed collections of Web documents that are hosted in institutional repositories operate under the assumption that they are more resilient and able to withstand change. Distributed collections are different from traditional collections in that the curator does not possess the documents. Without possession of the documents, the curator cannot control how they change. By definition, a digital repository must include procedures and tools for curating, organizing, storing, and retrieving the documents and media contained in the collection. Surprisingly, we have found these features – which are often found in digital collections with emphasis in long-term storage – do not fully preserve the referenced documents and make them impervious to change. For example, as of February 2016 the Association for Computing Machinery Digital Library has 19 unique links referencing the different sites of the Hypertext conference series and 10 of them report errors or point to the wrong content. Figure 1 shows a screenshot of http://www.ht00.org – which now displays information about banking and investments.



Binäare Optionen Broker – ht00.org

Geld verdienen mit Banc de Binary

Ein Video zu Banc de Binary und Geld verdienen

Figure 1: Screenshot of http://www.ht00.org. Accessed on 2/19/2016.

In this paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. More so, our study aims to quantify and analyze the perceptions and reactions of users when they come across documents that have

DOI: http://dx.doi.org/10.1145/2914586.2914628

been affected by unexpected change. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change that we found in the ACM conference list? This point becomes increasingly interesting when we take into account that the resources found in a digital collection are often curated and maintained by experts with affiliations to professionally managed institutions.

2. PREVIOUS WORK

Previous work on finding missing resources is based around the premise that documents and information are not lost but simply misplaced [8] as a consequence of the lack of integrity in the Web [9, 10]. Other studies have also focused on finding the longevity of documents in the Web [11] and in distributed collections [12, 13].

Phelps and Wilensky pioneered the use of lexical signatures to locate missing content in the Web [14]. They claimed that if a Web request returned a 404 error, querying a search engine with a five–term lexical signature could retrieve the missing content. Park et al. used Phelps and Wilensky's previous research to perform an evaluation of nine lexical signature generators that incorporate term frequency measures [15]. Additionally, Klein and Nelson have extracted lexical signatures from titles and backlinks to find missing Web resources [16].

Dalal et al. used a different method to find appropriate replacements for missing resources from the Web that belonged to a collection in Walden's Paths [17]. Their approach was based on a two–step process. First, metadata was extracted when the path was created thus preserving the author's intent and vision. Second, the extracted metadata was used to find pages when they cannot be retrieved. In the specific case of collections such as Walden's Paths, each node in a path is destined to make a contribution towards the overall concept and the continuity in the narration. Therefore, finding replacements becomes a critical factor to maintain the integrity of the collections and preserve their semantic meaning.

On the other hand, previous work on link persistence has focused on characterizing the availability of resources over time. Nelson and Allen measured the persistence and availability of documents in a digital library [18]. Koehler found that specialized document collections – such as legal, educational and some scientific citations – tend to stabilize over time [19]. However, citations in some domains have higher rates of failure [20]. McCown et al. also explored other factors that might cause a resource to fail by examining its age, path depth, top-level domain and file extension [21]. Here we extend this body of work by examining a distributed collection of Web documents as a part of an institutional digital library, describe the types of issues found in the collection, and examine the potential to automatically identify these issues when they arise.

3. CHANGE IN A DISTRUBUTED COLLECTIONS

To conduct our experiment we needed a document corpus. For this purpose, we harvested the metadata for the conference proceedings found in the Association for Computing Machinery Digital Library. While the ACM Digital Library stores and maintains the "Full text of every article ever published by ACM" and bibliographic citations from major publishers in computing, it includes the links to the actual conference sites as distributed resources that are hosted externally and therefore more prone to be affected by unexpected change.

3.1 A Distributed Collection within an Institutional Digital Library

The ACM maintains a list of the 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. Additionally, we also stored the metadata associated with each retrieved page. This metadata included the anchor text, URL requested, URL retrieved, HTTP headers and response code. Approximately 75% of the page requests resulted in a response code indicating success (200), which means that no problems were found when trying to fulfill the request. The remaining pages were mostly divided among page not found (404) responses and timeouts.

We then proceeded to inspect and categorize the 1492 pages that were retrieved with a 200 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. These pages could have been placed into the "incorrect" category, but we decided to use an additional category to make our experiment as transparent as possible.

3.2 Categorization of the Types of Change

The 531 pages that were reported by the HTTP server as being correctly retrieved but were clearly not the original contents were then analyzed in an effort to understand how conference sites degrade over time. The coding 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) This case surfaces 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. Here 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)

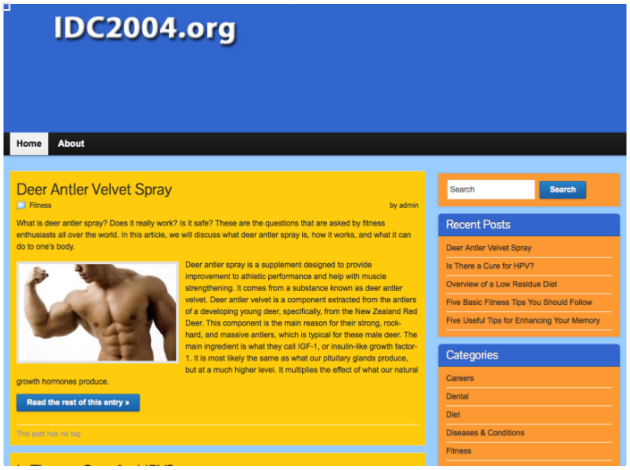
6. **Error pages:** (17 entries) Pages that specifically state that an error has occurred.

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 that these pages were not created to deceive users, but as an attempt to manipulate the PageRank algorithm [22]. Figure 2 shows a screenshot corresponding to the IDC 2004 site that displays an example of this case.

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. 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.



Deer Antler Velvet Spray

What is deer antler spray? Does it really work? Is it safe? These are the questions that are asked by fitness enthusiasts all over the world. In this article, we will discuss what deer antler spray is, how it works, and what it can do to one's body.

Deer antler spray is a supplement designed to provide improvement to athletic performance and help with muscle strengthening. It comes from a substance known as deer antler velvet. Deer antler velvet is a component extracted from the antlers of a developing young deer, specifically, from the New Zealand Red Deer. This component is the main reason for their strong, rock-hard, and massive antlers, which is typical for these male deer. The main ingredient is what they call IGF-1, or insulin-like growth factor-1. It is most likely the same as what our pituitary glands produce, but at a much higher level. It multiplies the effect of what our natural growth hormones produce.

Figure 2: Screenshot of the IDC 2004 site showing an example of a "Deceiving Page". Accessed at http://www.idc2004.org on 9/27/2014

4. CLASSIFICATION FEATURES

To develop classifiers for the types of issues identified in Section 3, we explored the potential for a variety of features to help discriminate between the categories observed. In particular, we consider features computed based on the contents and links returned by the initial request (the *base node*) and the contents and links returned by traversing the links in the base node. As we wanted to limit the number of HTTP requests required, this analysis did not examine the potential of the broader network topology to facilitate classification.

We included twelve link-based features. These features are divided into topology features, content-type features, anchor-text features, and child node features. Most of the link-based features were computed for the base node and are based on the number of out-links in that page. In addition we calculated some of the features for child nodes that are the valid out-links in these base-nodes.

The content of resources is generally highly indicative of their purpose. Eight quantitative measures related to the content of the resources were included as features for developing classifiers. These features are divided into image features and text-content features.

The text associated with resources is the most obvious feature for determining the topics. While a deep understanding of the domain of conference web sites could have been used to develop a domain-oriented expectation model (e.g., a discussion of content submission, organizing committees, schedule, keynotes, travel, and hotels), we instead focused on generating quantitative measures based on the text content that could be potentially valuable across domains.

In particular, we generated topic models to examine the similarity among the metadata and the contents of the base node and the metadata and the contents of the child nodes. We used Latent Dirichlet Allocation (LDA) to model the content of the text [23] and the KL-divergence similarity measure to compare them [24].

5. RESULTS

Results from examining the conference web site collection include examining the trends found in the collection, the generation of a test collection for exploring the potential to automatically classify resources based on the types of issues found in Section 3.2, and a comparison of the performance of different classifiers when provided the features in Section 4.

5.1 Classification Algorithms

We performed our binary and category classification with 71 algorithms that are implemented in the Weka toolkit [25]. We report the best classification results based on the F-measures from the following classifiers: K*, Decorate, Random Committee, Rotation Forest, Bagging, Boosting (e.g., LogitBoost) and decisions trees (e.g., Random Forest). The algorithmic details of these classifiers are beyond the scope of this paper and interested readers are referred to [25, 26].

To finalize the topic model, we initially explored how varying the number of topics in the LDA model affected classifier performance. As part of these experiments we varied the number of topics K between 5 and 25. The F-measure of seven classifiers when given the 20 features from Section 4 to classify the incorrect and unsure categories varied between 0.54 to 0.64. After training and testing the category classification data and performing this evaluation, we found that the best performing classifiers exhibit their best performance with 5 topics (K=5). Therefore, we used 5 topics for the remainder of our experiments involving the training and testing datasets and the analysis of the classifier results.

The results of our experiments for the "clearly correct", "incorrect" and "unsure" categories as a binary classification problem, and the performance metrics for the 7 most effective classifiers from our evaluation are presented in Table 1. The majority of our classifiers performed consistently with an F-measure of 0.63; Random Forest was the best performer for the binary classification. Decorate and Random Committee both exhibit a slightly higher F-measure for "correct" category, but Random Forest offers a substantially better F-measure for the "not correct" category.

Two categories of pages in our “incorrect/unsure” dataset, the “unsure” and “different language” groups, are the result of our inability to classify the contents returned for a resource as being either correct or incorrect. Although we originally grouped these two categories into the same "not correct" group, it is possible that some of these pages might contain valid or "correct" pages. The results displayed in Table 2 show the performance of the classifiers when just the “different language” and “unsure” categories were removed from the training/testing corpus.

Table 1: Binary classification (correct, not correct) combining all incorrect categories into single category.

	Precision		Recall		F-Measure	
	C	N	C	N	C	N
Decorate	0.635	0.637	0.641	0.632	0.638	0.635
RandomCommittee	0.606	0.632	0.672	0.562	0.637	0.595
RotationForest	0.626	0.629	0.632	0.623	0.629	0.626
RandomForest	0.643	0.653	0.664	0.632	0.635	0.642
K*	0.633	0.631	0.629	0.635	0.631	0.633
Bagging	0.628	0.633	0.641	0.62	0.634	0.627
LogitBoost	0.61	0.617	0.629	0.597	0.619	0.607

Table 2: Classification of only the “incorrect” categories by removing the “pages in a different language” and “unsure” categories

	Precision	Recall	F-Measure
Decorate	0.824	0.833	0.824
RandomCommittee	0.759	0.781	0.765
RotationForest	0.82	0.826	0.813
RandomForest	0.798	0.807	0.793
K*	0.767	0.785	0.767
Bagging	0.778	0.781	0.758
LogitBoost	0.79	0.8	0.784

6. EVALUATION

One of the main difficulties of the type of our study is finding a way to compare the results obtained using machine learning algorithms against the assessment made by humans. For this purpose, we carried out a user study where human subjects were asked to analyze the contents of Web documents and assess if these documents belonged to a specific collection.

The user study was administered through the Web using a Django [27] backend. In the user study, participants were asked to go over fifty randomly selected documents (25 correct and 25 incorrect) from the ACM corpus that we described in section 3. More specifically, participants were shown a screenshot of the Web document along with its corresponding anchor text – which was extracted form the ACM Digital Library – and a brief questionnaire that was aimed towards identifying the degradation of a document and its possible causes. In the end, we collected data from 62 participants – mostly upper-level Computer Science undergraduates – and assessed the validity of 2875 pages.

Interestingly, the study participants did not have difficulties identifying documents in the correct category. However, this was not the case in the other categories. Table 3 shows a summary of the user responses for the document classification. Documents in the correct category were correctly identified in 71% of the cases, but users were not able 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 with documents were it was evident that the documents are 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. Our analysis indicates that 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. We also analyzed the amount of time users took to classify documents, but we could not find any statistical evidence that would shed insights to support their choices and decisions.

At this point, the question that remains to be answered is whether the classification algorithms perform better than the human subjects. The answer to this question is two-fold: we believe that the algorithms performed more consistently. However, human subjects were able to identify pages in the correct category slightly better than the algorithms, but failed when asked to identify pages in the incorrect categories. On the other hand, the machine learning algorithms outperformed human subjects when identifying pages in the incorrect categories. Table 4 shows a comparison between the classification by humans and the algorithms. The difference in performance is significant, which for us is a clear indication that managing the effects of unexpected changes in collections of Web documents is a more serious problem than we had originally anticipated.

Table 3: 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
Total	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

Table 4: Comparison between the classifications made by human subjects and the classification algorithms.

Classification	Human Subjects		Algorithms	
	Correct	Incorrect	Correct	Incorrect
Precision	0.71	0.24	0.64	0.65
Recall	0.48	0.46	0.66	0.63
F-measure	0.58	0.32	0.65	0.64

7. DISCUSSION

The analysis of the conference website links within the ACM Digital Library shows that institutional archives are not immune to some of the challenges of distributed collection management – which include unexpected changes within the contents of its documents. Knowing when changes to a resource require human attention is not a simple problem. Only once such an assessment has been made can we apply techniques to recover or replace a broken resource.

As we have stated before, some of our findings align with previous work. We coincide that 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.

As the categorization of changes to the collection shows, determining the degree of change affecting a digital collection over time is a difficult task. A web resource may gradually degrade from being correct to one that is still of some use by providing access to related information or information about the institution to contact for more information. Changes in web servers, directory structures, etc. may cause requests to still result in a successful 200 code from the server yet provide no information to the requestor. A number of these categories were purposely left out of the evaluation of the classification algorithms as these cases can be handled by previous work. More specifically, detecting "blank pages", "failed redirects", "directory listings", "domain for sale" and "error pages" are handled with previous work on identifying Soft 404 error pages [6, 28].

Being able to distinguish between the "kind of correct" and "deceiving" pages is important to collection managers. A contribution of our research is detecting when documents change unexpectedly and fall into more problematic categories such as "kind of correct" and "deceiving pages".

This last point lead us to investigate and inquiry: 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 have 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. Examining the variation of other features across categories of change may provide additional insight into the motivations and characterizations of change.

Another potential source of information for identifying the severity of change is web archival services. We explored use of such archives but chose to leave them out of the current approach. In the specific case of the ACM conference list, some conference sites were not crawled at all. Thus, archival services would have provide us with an incomplete index that would not have allowed us to fully answer the research questions for this paper. Additionally, irregular crawling intervals and data embargoes can reduce the value of information from such archives for timely identification of change.

Our research is 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 to cause servers to report errors explicitly. Our analysis focused on the instances that fall between these two extreme cases, which makes their detection more difficult and require the assistance of a classification system and detection framework such as the one that we have described in this paper.

It was not surprising for us that the documents in the "correct" category were more consistently identified during the user study. However, now that data collection phase of the study is over and its results are analyzed, we believe that the premises we established were influential towards its outcome. In this study, subjects were operating under the assumption that they were identifying and categorizing conference websites. Therefore, we hypothesize that the nature of these sites – being institutional and backed up by professional and academic organizations – might have led the subjects to believe that the documents were in the "correct" category despite showing explicit symptoms of change. This effect could potentially explain the incorrect classification of some of the documents the "incorrect categories"; most notably the ones in the "domain for sale", "hello world" and other categories that displayed explicit error codes and content in different languages – which were clear indications of their incorrectness.

There is a wide range of follow-on work related to the classification problem explored here. We limited our approach to features that could be computed quickly with a minimal number of HTTP requests per collection resource. A pragmatic direction of future work is to develop a software package that combines approaches for determining when HTTP error-codes are likely temporary or permanent, recognizing Soft 404 responses, detecting Web spam, and categorizing the remaining changed pages as described here.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Computing Surveys,* vol. 32, pp. 144-173, 2000.

[2] M. K. Taylor and D. Hudson, "" Linkrot" and the Usefulness of Web Site Bibliographies," *Reference & User Services Quarterly,* pp. 273-277, 2000.

[3] P. Logasa Bogen, D. Pogue, F. Poursardar, Y. Li, R. Furuta, and F. Shipman, "WPv4: a re-imagined Walden's paths to support diverse user communities," in *Proceedings of the 11th annual International ACM/IEEE Joint Conference on Digital Libraries*, Ottawa, Ontario, Canada, 2011, pp. 419-420.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] R. Baeza-Yates, I. Pereira, and N. Ziviani, "Genealogical trees on the web: a search engine user perspective," in *Proceedings of the 17th international conference on World Wide Web*, Beijing, China, 2008.

[9] H. Ashman, "Electronic document addressing: dealing with change," *ACM Computing Surveys,* vol. 32, pp. 201-212, 2000.

[10] H. C. Davis, "Hypertext link integrity," *ACM Computing Surveys,* vol. 31, p. 28, 1999.

[11] B. Kahle, "Preserving the Internet," *Scientific American,* vol. 276, pp. 82-83, March 1997 1997.

[12] W. Koehler, "Web page change and persistence---a four-year longitudinal study," *Journal of the American Society for Information Science and Technology,* vol. 53, pp. 162-171, 2002.

[13] D. Spinellis, "The decay and failures of web references," *Communications of the ACM,* vol. 46, pp. 71-77, 2003.

[14] T. A. Phelps and R. Wilensky, "Robust Hyperlinks Cost Just Five Words Each," University of California at Berkeley2000.

[15] S.-T. Park, D. M. Pennock, C. L. Giles, and R. Krovetz, "Analysis of lexical signatures for improving information persistence on the World Wide Web," *Transactions on Information Systems,* vol. 22, pp. 540-572, 2004.

[16] 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.

[17] Z. Dalal, S. Dash, P. Dave, L. Francisco-Revilla, R. Furuta, U. Karadkar*, et al.*, "Managing distributed collections: evaluating web page changes, movement, and replacement," in *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, Tuscon, AZ, USA, 2004, pp. 160-168.

[18] M. Nelson and D. Allen. (2002, January 2002) Object Persistence and Availability in Digital Libraries. *D-Lib Magazine*.

[19] W. Koehler, "A longitudinal study of Web pages continued: a consideration of document persistence," *Information Research,* vol. 9, 2004.

[20] D. H. L. Goh and P. K. Ng, "Link decay in leading information science journals," *Journal of the American Society for Information Science and Technology,* vol. 58, pp. 15-24, 2007.

[21] F. McCown, S. Chan, M. L. Nelson, and J. Bollen, "The availability and persistence of web references in D-Lib Magazine," *arXiv preprint cs/0511077,* 2005.

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

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research,* vol. 3, pp. 993-1022, 2003.

[24] C. M. Bishop, *Pattern recognition and machine learning* vol. 1: springer New York, 2006.

[25] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2005.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.

[27] (2/20/2016). *The Web framework for perfectionists with deadlines | Django*. Available: https://djangoproject.com

[28] 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.