Not All Mementos Are Created Equal: Measuring The Impact of Missing Resources

Ref.: Ms. No. IJDL-D-14-00041

We would like to thank the reviewers for their feedback. The reviews have highlighted a few shortcomings in our paper, particularly our need to explicitly explain the importance of archival quality assurance for those not familiar. We found all of the recommendations and issues raised to be extremely fair, and addressing these comments has improved the paper.

We have addressed all of the comments from the reviewers, updated our references with newly published material, and summarize our edits in the enumerated list, below.

Reviewer #1:

1. To do justice to the paper's contribution it would make sense to formulate explicitly the take home message for a general IJDL reader who is not working as web archivist, in terms of the importance of quality assessment, or of ways to indicate estimated quality within a DL context, or even the relative need of being "complete" for certain DL use cases and DL applications.

This is an important point, and we have addressed this comment in our introduction section.

2. While the overall goal "understand how missing embedded resources impact user satisfaction" is clear and strong, it is less clear how the precise evidence presented answers this question, even in part. E.g., what is the crucial threshold here? We see two proposed measures, and one of of them is "closer to the perception of users" — but what do we learn from that precisely? To play the devil's advocate, if you propose two different measures it is rather unsurprising that one of them is working 'better'... Clearly, this is not about providing a perfect solution, but about clearly and precisely carving out the strengths but also weaknesses of the current proposal.

While we acknowledge that D_m is not perfect, it solves an important problem better than existing metrics. The Internet Archive has roughly 455 billion web pages (according to https://archive.org/web/), which is far more than can be evaluated using manual or human-in-the-loop approaches. D_m provides archivists an automated method for evaluating their mementos.

We address these concerns further in Section 6.4. While we do not have a specific notion for a threshold of acceptability, we are working on improving our algorithm as much as possible, including a refinement to be completed in the coming months. While our overall goal is to provide a metric in perfect agreement with web users' understanding of memento damage, our evaluation shows that this metric is an improvement (albeit not a perfect solution) rather

than equal or lesser effectiveness over M_m. As such, we accept D_m as an improvement over M_m and also accept that we must work to improve D_m to reach our end state of an algorithm in perfect agreement with turker assessment.

3. Remarkably, one of the proposed measures M_m is never clearly defined. It's introduced as "proportion of missing embedded resources to all requested resources" and later "proportion of embedded missing resources" (which seems not identical). The ground truth seems to not distinguish embedded and other resources — hence why only look at the missing embedded resources? Are we counting unique references or all links to embedded resources? What is the definition of embedded resources in the first place? Why look at proportions and not absolute numbers? How to deal with embedded content on an external site, that may still be available, e.g., think of a standard DTD and CSS to render particular encoded data on the Web, and was likely intentionally left our of this harvest (but included in the harvest of the external site)? Etc. At least more detail is needed about what is exactly (attempted to be) measured by M_m, and why M_m is the obvious measure for this. Otherwise the paper becomes a straw man argument…

We have addressed this recommendation in Section 6. In our original DL2014 paper, we omitted a formal definition of M_m due to a misplaced assumption that we could describe M_m in text rather than an equation. We realize that not only should we have defined M_m (now defined in Equation 2), but also formalized our methods for extracting each dataset (the proportion of missing embedded resources and all requested resources) (now defined in Equation 1). Note that we consider only images/multimedia and CSS because these are elements that can be visually validated. Other elements (e.g., JSON files or other data/metadata files) are consider outside the scope of this research.

The embedded resources were collected using PhantomJS, with the set of unique resources requested constituting the set of all requested resources and the requests that resolve in an HTTP response other than an HTTP 200 or HTTP 300 as a missing embedded resource. We consider proportions because resources that request more resources may be missing more embedded resources, and we do not want to penalize an archive because a particular resource has many embedded resources, but instead evaluate it on its rate of success. For example, a memento that is missing 1 of 2 embedded resources could be considered less successfully archived than a memento that is missing 1 of 100 embedded resources.

4. What is the explanation of the 81% recognition of the damaged pages by Mturkers. Is this due to noisy labels obtained by crowdsourcing (possibly due to malicious workers or due to the colloquial way the question was asked)? Is it due to there being minimal differences between the pages (or unclarity which of the different pages is the "intended" original page? Or is it due to non-essential differences (e.g., difference only visible after scrolling)?

The 81% agreement rate between D_m and the turkers is attributed to more minute differences between the pages. Some resources – when missing – do not have a substantial visual impact on the page, and humans cannot identify these small differences. As such, the missing resources are deemed to be not important. This further shows why our algorithm is needed – some embedded resources deserve to be given a higher weight than others because users find them more important than others, or when a cursory inspection yields no discernible differences

5. While the proposed measure D_m makes certainly sense, it also has an adhoc feel. Please try to separate the general idea of the measure from the precise operationalization to the case at hand. What are general principles and what are good heuristics exploiting the case at hand. Why the restriction to HTML and

CSS/images and multimedia? E.g., stylesheets in XSL/CSS can have a dramatic effect on what's shown and how — think of an XML dataset rendered with complex XSL. Also the visual display makes assumptions on the rendering — many modern web pages render dramatically different under different screen and device settings.

We explain these issues in part in our response to comment #3. Recognizing that CSS has a dramatic impact on the appearance of web content, we attempt to measure this with our D_m_css metric as defined in Equation 6. We only consider “desktop” device settings because this is the same setting used by crawlers and other archival tools during the archival process. To consider a scenario such as a mobile device in our algorithm, we would need to show that archives are capable of performing differentiating between the user-agent string within the archives; we have shown this to be contrary to current possible features in modern archives in our prior work (M. Kelly et al. “A Method for Identifying Personalized Representations in Web Archives”, DLib Magazine, vol. 19 no 11/12).

6. There should be more discussion on the experimental data? What is the exact selection and for what population of web archive data is this representative? How complete are these harvests? (My understanding would be that specific cases of images are not included — e.g. obfuscated by client-side javascript or flash. Do we cover all possible cases of missing embedded content?)

We describe in more detail how selected the URI-Rs for our dataset in Section 5. We also reference our prior work which measure the completeness of the dataset.

7. The paper acknowledges that it is incremental over [33] but the precise extension to that work are never spelled out.

Song et al. used block importance to eliminate unimportant “noise” from web pages to more accurately extract the aspects of pages that users find most important. These important blocks are most frequently large and prominently oriented in the center of the viewport. We utilize this concept in D_m, highlighting the importance of image size and centrality directly in this metric.

Reviewer #2:

1. The paper is excellent. It should be accepted without modification.

We thank Reviewer #2 for the time to review our work, kind words, and acceptance of this paper.

Reviewer #3:

1. The methodology appears sound, though I have some concern about the small number of human raters; with so few, how well could their judgments about many mementos be effectively generalized?

SalahEldeen's work (Reading the Correct History? Modeling Temporal Intention in Resource Sharing) effectively used 5 turkers to establish user opinion. We follow this model in our experiment. We also err on the side of overwhelming agreement between the turkers, considering splits of only 4-1 and 5-0 as agreement. We have cited this paper in Section 4 to justify the use of 5 turkers to assess turker agreement.

2. The findings with respect to the page-at-a-time archiving services are an interesting contrast to the earlier Internet Archive Wayback Machine results, but I have a less

clear idea of why the difference matters - i.e., do the different memento damage profiles suggest that users should prefer one over another for page-at-a-time archiving or access?

This is an excellent point that we overlooked in our current discussion. Archive.today and WebCite were established for different purposes, each offering its own benefits. WebCite was established for the purpose of preventing link rot within academic papers (see Klein et al., Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot). As such, WebCite's target audience is scholars and the target material is academic publications.

Archive.today was established as a more general purpose page-at-a-time archival tool without a specific audience. Archive.today is also a more recently established service and uses more recently developed tools to archive pages. Archive.today does not archive resources such as PDFs or XML, while WebCite makes an attempt to archive such resources. We leave the decision as to whether one is better than the other to the user.

3. The discussion of the impact of JavaScript on memento damage is a great addition to the previous DL manuscript. The quantifiable improvement of PhantomJS plus Heritrix over Heritrix by itself is a finding worth highlighting on its own and perhaps an area worth suggesting for future research.

Thank you for the recommendation. We have included measuring the improvements of pairing PhantomJS with Heritrix as an area of future research in our Conclusion section. We also just submitted a paper to JCDL2015 that measures the performance differences between PhantomJS and Heritrix (Archiving Deferred Representations Using a Two-Tiered Crawling Approach). We have omitted this citation in the paper since the work has not yet been peer reviewed.

4. Reviewing the conclusion and, in the process, referring back to the introduction, I think it could be articulated more clearly that the practical problem that the proposed algorithm helps to solve is the difficulty in measuring, improving, and ensuring the quality of large-scale web archives; the work only means to address the experience of web archive users by solving this other, more tractable problem.

Thank you for this recommendation. Reviewer #1 had similar recommendations, and we have adjusted our presentation of the problem accordingly.

5. In addition to the above comments, here are a few page-specific notes and questions:

   1 - Typo: "...while the missing embedded resources is remaining constant..."
   2 - It's a little unclear what's meant by the parenthetical note: "i.e., the utility of mementos". What's the problem in a general Web context that mementos solve (e.g., substitution of missing embedded resources on "live" websites with mementos)? Or is this just referring to users of web archives?

This is referring to just users of web archives.

6 - The argument that percent embedded resources missing performed worse than random and far worse than optimal performance is compelling but is it reliable with only five turkers providing assessments?

I trust our additions to Comment #1 address this question.

7 - Is image weighting affected by the absence of explicitly-specified height and

width attributes in the HTML?

8 - Incorrect figure citation: "...are shown in their respective thirds in Figure 2."

8 - Incorrect figure citation: "...like the content in Figure 6.2.2."

9 - Should the discussion of multimedia loaded by JavaScript be integrated with the later, more general discussion of the impact of JavaScript on memento damage?

9 - As of 2013, Archive-It started using PhantomJS for all crawls. Given that Archive-It crawl data is added to the primary Wayback Machine index, can you reliably say that all of the Internet Archive mementos did not benefit from JavaScript execution?

Archive-It uses Umbra, which is a service that enlists a headless browsing utility to archive resources. However, Umbra is only used on a hand-selected set of URI-Rs. We do not have the list of URI-Rs selected, but know that the URI-Rs are known to use JavaScript and Ajax and will be difficult to archive using traditional methods (e.g., Facebook and Twitter resources). This is an important point that we have now included in Section 8.1 in response to this very valid recommendation.

11 - I wonder how much of the increasing number of missing important resources in Internet Archive Wayback Machine has to do with the growing predominance of video combined with the file size cap in Internet Archive's crawler configuration, rather than JavaScript?

This is an excellent question. However, it would require a much different experiment. We will keep this question in mind for future experiments.

14 - Should the benefit of a larger crawl frontier be qualified? I am unsure, for instance, whether PhantomJS triggers crawler traps that Heritrix by itself might ignore.

Again, this is an excellent question for a future experiment. We are studying the performance impacts of PhantomJS vs Heritrix, and this is a very interesting research question for our experiment.

15 - Typo: "...we can compare the two page-at-a-time archiver..."

Thank you for the recommendations. We have addressed each of the typos identified on pages 1, 7, 8, and 9 along with the rest of these items.