Resurrecting My Revolution

Using Social Link Neighborhood in Bringing Context to the Disappearing Web

Hany M. Salaheldeen and Michael L. Nelson

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

Abstract. In previous work we reported that resources linked in tweets disappeared at the rate of 11% in the first year followed by 7.3% each year afterwards. We also found that in the first year 6.7%, and 14.6% in each subsequent year, of the resources were archived in public web archives. In this paper we revisit the same dataset of tweets and find that our prior model still holds and the calculated error for estimating percentages missing was about 4%, but we found the rate of archiving produced a higher error of about 11.5%. We also discovered that resources have disappeared from the archives themselves (7.89%) as well as reappeared on the live web after being declared missing (6.54%). We have also tested the availability of the tweets themselves and found that 10.34% have disappeared from the live web. To mitigate the loss of resources on the live web, we propose the use of a "tweet signature". Using the Topsy API, we extract the top five most frequent terms from the union of all tweets about a resource, and use these five terms as a query to Google. We found that using tweet signatures results in discovering replacement resources with 70+% textual similarity to the missing resource 41% of the time.

Keywords: Web Archiving, Social Media, Digital Preservation, Reconstruction.

1 Introduction

Microblogging services like Twitter have evolved from merely posting a status or quote to an intra-user interaction tool that connect celebrities, politicians, and others to the public. They have also evolved to act as a narration tool and an information exchange describing current publicly recognized events and incidents. In 2011, during the Egyptian revolution, thousands of posts and resources were shared during the 18 days of the uprising. These resources could have crucial value in narrating the personal experience during this historic event, acting as a first draft of history written by the public.

In our previous work, we proved that shared resources on the web are prone to loss and disappearance at nearly constant rate [17]. We found that after

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 333–345, 2013.

only one year we lost nearly 11% of the resources linked in social posts and continued to lose an average of 7.3% yearly. In some cases, this disappearance is not catastrophic as we can rely on the public archives to retrieve a snapshot of the resource to fill into the place of the missing resource. In another study we measured how much of the web is archived and found that 16%–79% of URIs have at least one archived copy [1]. Unfortunately, there is still a large percentage of the web that is not archived and thus a huge amount of resources are not archived and prone to total loss upon disappearance from the live web.

This evolution in the role of social media and the ease of reader interaction and dissemination could be used as a possible solution to metigate or prevent the loss of the unarchived shared resources. Fortunately, when a user tweets or shares a link, it leaves behind a trail of copies, links, likes, comments, other shares. If the shared resource is later gone these traces, in most cases, still persist. Thus, in this paper we investigate if the other tweets that also linked to the resource can be mined to provide enough context to discover similar resources that can be used as a substitute for the missing resource. To do this, in this study we extract up to the 500 most recent tweets about linked URIs and we propose a method of finding the social link neighborhood of the resource we are attempting to reconstruct. This link neighborhood could be mined for identifiers and alternative related resources.

2 Related Work

Social media has been the focus of numerous studies in the last decade. Twitter, for example, was analyzed by Kwak et al. where they aimed to identify the characteristics of the Twittersphere, retweeting, and the diffusion speed of posts by using algorithms like PageRank in ranking users [10]. Bakshy et al. investigated 1.6 million users along with the tweet diffusion events to identify influencers on Twitter and their effect in content spread [2]. To answer questions in regards to the production, flow, and consumption of information on Twitter, Wu et al. analyzed the intra-user interactions and found that nearly 20K elite users are responsible for generation of nearly 50% of URLs shared [18]. Intuitively, this shows that popularity plays an important role in the content disseminated. They also found that type of the content published and the type of users broadcasting this content affect the lifespan of the tweet activity.

Along with understanding the nature of the social media, researchers analyzed user behavior on the social networks in general. By analyzing user activity click logs, Beneventu et al. aimed to get a better understanding of social interactions social browsing patterns [5]. Zhao and Rosson aimed to explore the reasons of how and why people use Twitter and this use's impact on informal communication at work [23]. Following the how and the why, Gill et al. attempted to answer the next question of what is the user-generated content is about by investigating personal weblogs to detect the effects of personality, topic type, and the general motivation in published blogs [7]. Yang and Counts investigated the information diffusion speed, scale and range in Twitter and how they could be predicted [22].

This in-depth analysis and study of the social media, its nature, the information dissemination patterns, and the user behavior and interaction paved the way for the researchers to have a better understanding of how the social media played a major role in narrating publicly significant events. These studies prove that user-generated content in social media is of crucial importance and can be considered the first draft of history. Vieweg et al. analyzed two natural hazard events (the Oklahoma grass fires and the red River floods in 2009) and how microblogging contributed in raising the situational awareness of the public [21]. Starbird and Palen analyzed how the crowd interact with politically sensitive context regarding the Egyptian revolution of 2011 [20]. Starbird et al. in another study utilized collaborative filtering techniques for identifying social media users who are most likely to be on the ground during a mass disruption event [19]. Mark et al. investigated weblogs to examine societal interactions to a disaster over time and how they reflect the collective public view towards this disaster [11].

In our previous work we showed that this content is vulnerable to loss. Similar to regular web content and websites, there are several reasons explaining this loss. McCown et al. analyzed some of the reasons behind the disappearance and reappearance of websites [12]. McCown and Nelson also examined several techniques to counter the loss prior to its occurrence in social networking websites like Facebook [13]. As for regular web pages, Klein and Nelson analyzed the means of using lexical signatures to rediscover missing web pages [9]. Given that the web resource itself might not be available for analysis or might be costly to extract, several studies investigated other alternatives to having the resource itself. Other studies investigated the use of the page's URL to aid web page categorization without resorting to the have the webpage itself [4, 8]. Xiaoguang et al. utilized class information from neighboring pages in the link graph to aid the classification [15].

3 Existence and Stability of Shared Resources

We start our analysis by revisiting the experiment conducted in March of 2012, in which we modeled the existence of shared resources on the live web and the public archives. In that experiment, we examined six publicly-recognized events that occurred between June 2009 and March 2012, extracting six sets of corresponding social posts. Each of the selected posts include an linked resource and hashtags related to the events. Consequently, we tested the existence of the embedded resources on the live web and in the public archives. After calculating the percentages lost and archived we estimated the existence as a function of time. In this paper, we start by revisiting this year-old estimation model and checking its validity after a year before proceeding with our analysis of reappearance and extracting the social context of the missing resources. Then we investigated how this context could be utilized in guiding the search in extracting the best possible replacement for the missing resource.

3.1 Revisiting Existence

In the 2012 model, we found a nearly linear relationship between the number of resources missing from the web and time (equation 1), and a less linear relationship between the amount archived and time (equation 2).

$$Content\ Lost\ Percentage = 0.02(Age\ in\ days) + 4.20 \quad (1)$$

$$Content\ Archived\ Percentage = 0.04(Age\ in\ days) + 6.74 \quad (2)$$

As a year has passed, we need to analyze our findings and the estimation calculated to see if it still matches our prediction. For each of the six datasets investigated, we repeat the same experiment of analyzing the existence of each of the resources on the live web. A resource is deemed missing if its HTTP responses terminate in something other than 200, including "soft 404s" [3]. Table 1 shows the results from repeating the experiment, the predicted calculated values based on our model, and the corresponding errors. Figure 1 illustrates the measured and the estimated plots for the missing resources. The standard error is 4.15% which shows that our model matched reality.

Table 1. Measured and predicted percentages for missing and archived content in each dataset

		MJ		Iran		H1N1		Obama		Egypt	Syria
Missing	Measured	37.10%	37.50%	28.17%	30.56%	26.29%	31.62%	32.47%	24.64%	7.55%	12.68%
	Predicted	31.72%	31.42%	31.96%	30.98%	30.16%	29.68%	29.60%	28.36%	19.80%	11.54%
	Error	5.38%	6.08%	3.79%	0.42%	3.87%	1.94%	2.87%	3.72%	12.25%	1.14%
									Average Prediction Error		**4.15%**
Archived	Measured	48.61%	40.32%	60.80%	55.04%	47.97%	52.14%	48.38%	40.58%	23.73%	0.56%
	Predicted	61.78%	61.18%	62.26%	60.30%	58.66%	57.70%	57.54%	55.06%	37.94%	21.42%
	Error	13.17%	20.86%	1.46%	5.26%	10.69%	5.56%	9.16%	14.48%	14.21%	20.86%
									Average Prediction Error		**11.57%**

To verify the second part of our model we calculate the percentages of resources that are archived at least once in one of the public archives. Table 1 illustrates the archived results measured, predicted, and the corresponding standard error. Figure 1 also displays the measured and predicted corresponding plots for the archived resources. While the archived content percentages had a higher error percentage of 11.57% and proceeded to become further less linear with time. This fluctuation in the archival percentages convinced us that a further analysis is needed.

3.2 Reappearance and Disappearance

In measuring the percentage of resources missing from the live web, we assumed that when a resource is deemed missing it remains missing. It was also assumed that if a snapshot of the resource is present in one of the public archives the resource is deemed archived and that this snapshot persists indefinitely. Utilizing

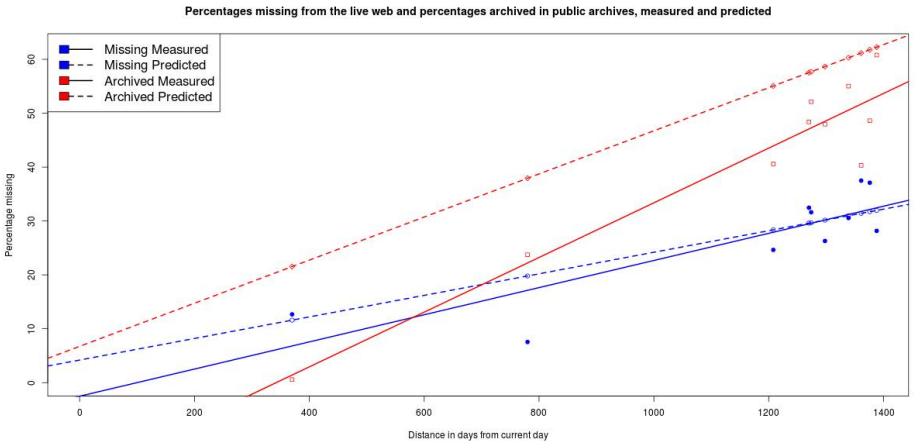

Fig. 1. Measured and predicted percentages of resources missing and archived for each dataset and the corresponding linear regression

the response logs resulting from running the existence experiment in 2012 and in 2013 we compare the corresponding HTTP responses and the number of mementos for each resource. As expected, portions of the datasets disappeared from the live web and were labeled as missing. An interesting phenomena occurred as several of the resources that were previously declared as missing became available again as shown in table 2. A possible explanation of this reappearance could be a domain or a webserver being disrupted and restored again. Another possible explanation is that the previously missing resources could be linked to a suspended user account that was reinstated. To eliminate the effect of transient errors, the experiment was repeated three times in the course of two weeks.

The dotted line in figure 2 shows resources missing in 2012 that reappeared in 2013. Given those percentages we notice a linear relationship with time. By applying linear regression we reach equation 3 describing the reappearance of resources as a function of time.

$$LiveContent\ Reappearing = 0.01(Age\ in\ days) - 1.42 \tag{3}$$

In the same previous study, we modeled the archival existence or the percentage archived as a function of time. The phenomena analyzed in the previous section showed the instability of the resources in the web which influenced us to investigate

Table 2. Percentages of resources reappearing on the live web and disappearing from the public archives per event

Event	MJ	Iran	Obama	H1N1	Egypt	Syria	Average
% Re-appearing on the web	11.29%	11.48%	6.63%	3.68%	4.21%	1.97 %	6.54%
% Disappearing from archives	9.98%	11.17%	15.65%	5.46%	2.81%	2.25 %	7.89%
% Going from 1 memento to 0	2.72%	2.89%	4.24%	1.96%	0.23%	0.28%	2.05%

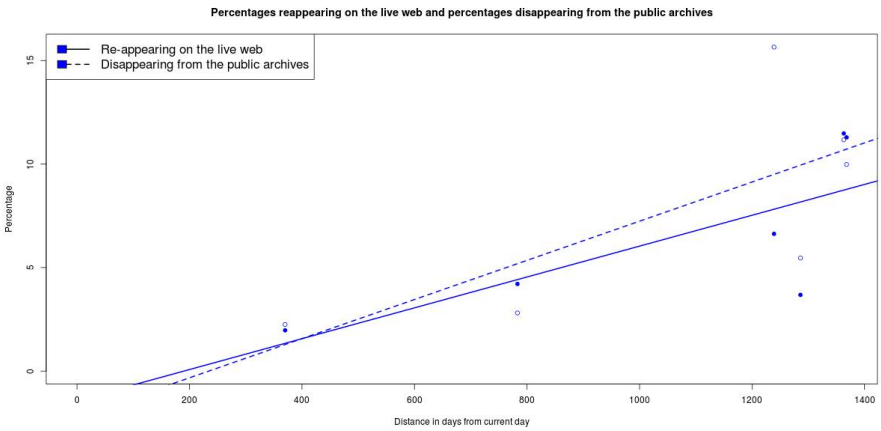

Fig. 2. Percentages of resources reappearing on the live web and the resources disappearing from the public archives

the archived resources as well. We deemed a resource to be archived if there existed at least one publicly available memento of the resource in the archives. For each resource we extracted the memento timemaps and recorded the number of available mementos. The resources are expected to have the same number of mementos or more indicating more snapshots taken into the archives or unarchived resources started to exist in the archives. We notice another interesting phenomena, the number of available mementos of several resources have actually descreased indicating disappearance from the archives as shown in table 2.

Brunelle and Nelson have shown that timemaps shrink 20% of the time [6]. Another possible explanation is that in 2012 the memento aggregator included search engine caches as archives but no longer does so in 2013. We estimate search engine cache only timemaps by measuring the number of resources whose timemaps went from 1 memento to 0 as shown in table 2 as well. Similarly, we plot the percentages of archival disappearance in figure 2. Equation 4 results from applying linear regression in curve fitting. Inspecting figure 2 verifies to a certain degree our explanation of the archival disappearance phenomena as the regression line maintains the same slope of the estimated model as shown in figure 1 while it differs in the Y-intercept. This explains to a certain degree the uniform variation in the estimated function. Unfortunately, we cannot verify this precisely as we do not have the past timemaps of the resources in the datasets.

$$Mementos\ Disappearing = 0.01(Age\ in\ days) - 2.22 \tag{4}$$

3.3 Tweet Existence

After focusing on the embedded resources shared in posts in social media another question arouse, what about the existence of the social post itself? In collecting the dataset that we utilized in our analysis we focused on the embedded resource and the creation dates. Also the Stanford Network Analysis Platform (SNAP)

Table 3. Percentages of missing posts averages

Event	MJ	Iran	Obama	H1N1	Egypt	Syria	Average
Average % of missing posts	14.43%	14.59%	10.03%	7.38%	15.08%	0.53%	10.34%

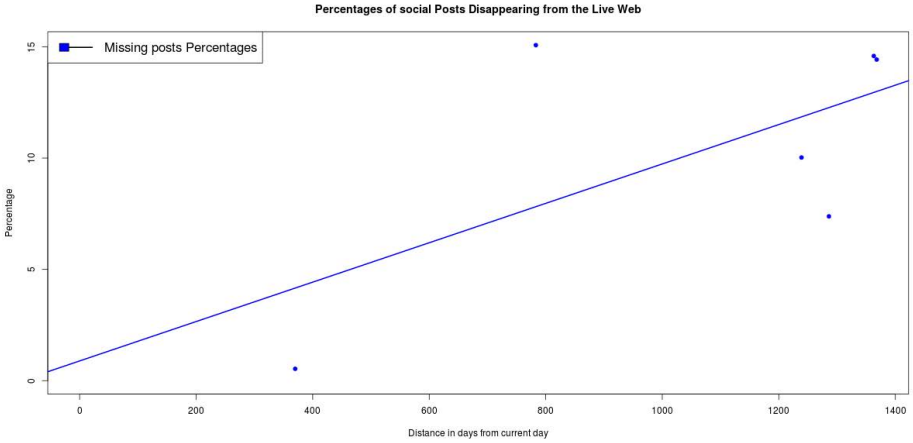

Fig. 3. Percentages of missing posts averages curve fitted using linear regression

dataset we used provides only the tweet text, the author's username, and the creation date with no further information about the tweet or its URI. A social post could face the same fate of the embedded resource by being deleted, service hosting it discontinued, or the author's account getting suspended. Similarly to the resource existence testing, we check the existence of the posts by examining the HTTP response headers. Unfortunately, the datasets we used do not include all the fields and parameters of a tweet, among which is the tweet's URI. To work around the absence of the social post URI we utilized Topsy, a service that mines social media websites like Twitter to provide analytics and insight to topics and resources. Using the API, we can extract all the available tweets that incorporate a given URI with a maximum of 500 tweets. For each resource in the dataset we extract all the tweets and check their existence on the live web accordingly. Given a URI, we can estimate the percentage of social posts that are missing. This number could give an insight to what is the probability that the post itself is missing. Table 3 shows the results for each dataset. Figure 3 illustrates the collective percentages through time. Equation 5 shows the result of curve fitting the percentages of loss as a function of time.

$$SocialPosts\ Missing = 0.01(Age\ in\ days) + 0.88 \tag{5}$$

4 Context Discovery and Shared Resource Replacement

A web resource can fall into one of the categories as shown in table 4. These categories were adopted from the work of McCown and Nelson [14].

Table 4. Web resource categories in regards to archivability and availability

	Archived	**Not Archived**
Available	Replicated	Vulnerable
Missing	Endangered	Unrecoverable

If a resource is available on the live web and also archived in public archives then it is considered replicated and safe. The resource is considered vulnerable if it persists on the web but has no available archived versions. If a resource is not available on the live web but has an archived version then it is considered endangered as it relies on the stability and the persistence of the archive. The worst case scenario occurs when the resource disappears from the live web without being archived at all thusly, be considered unrecoverable. In our study we focus on the latter category and how we can utilize the social media in identifying the context of the shared resource and select a possible replacement candidate to fill in the position of the missing resource and maintain the same context of the social post.

A shared resource leaves traces even after it ceases to exist on the web. We attempt to collect those traces and discover context for the missing resource. Since Twitter for example restricts the length of the posts to be 140 characters only, an author might rely mostly on the shared resource in conveying a thought or an idea by embedding a link in the post and resorting to limiting the associated text. Thusly, obtaining context is crucial when the resource disappears. To accomplish that, we try to find the social link neighborhood of the tweet and the resource we are attempting this context discovery. When a link is shared on Twitter for example, it could be associated with describing text in the form of the status itself, hashtags, usertags, or other links as well. These co-existing links could act as a viable replacement to the missing resource under investigation while the tags and text could provide better context enabling a better understanding of the resource.

4.1 Social Extraction

Given the URI of the resource under investigation, we utilize Topsy's API to extract all the available tweets incorporating this URI. In social media, a resource's URI can be shared in different forms with the aid of URL shortening. To elaborate, a link to Google's web page http://www.google.com could be shared also in several forms like http://goo.gl/xYMol, http://bitly.com/XeRH58, and http://t.co/XFiAkbHnp3. Each of these forms redirects to the same final destination URI. Fortunately, Topsy's API handles this by searching their index for the final target URL rather than the shortened form. A maximum of 500 tweets of

```
Reconstruction:
{
  "URI": "http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html",
  "Related Tweet Count": 290,
  "Related Hashtags": "#history #jan25 #sschat #arabspring #jrn112 #archives #in #revolution
  #iipc12 #mppdigital #egypt #recordkeeping #twitter #egyptrevolution #digitalpreservation
  #preservation #webarchiving #or2012 #1anpa #socialmedia",
  "Users who talked about this": "@textfiles @jigarmehta @blakehounshell] @jonathanglick
  @daensen404: @ryersonjourn @chanders @theotypes) @jwax55 @marklittlenews @ndiipp ...",
  "All associated unique links:": "http://t.co/ZRASTg5o http://t.co/eXhlSTRF
  http://t.co/3GIb6oI3 http://t.co/ArVqCqfP ...",
  "All other links associated:": "http://www.cs.odu.edu/~mln/pubs/tpdl-2012/tpdl-2012.pdf
  http://dashes.com/anil/2011/01/if-you-didnt-blog-it-it-didnt-happen.html",
  "Most frequent link appearing:": "http://t.co/0A1q2fzz",
  "Number of times the Most frequent link appearing:": 19,
  "Most frequent tweet posted and reposted:": "@acarvin You may have seen this already.
  Arab Spring digital content is apparently being lost.",
  "Number of times the Most frequent tweet appearing:": 23,
  "The longest common phrase appearing:": "You may have seen this already Arab Spring
  digital content is apparently being lost",
  "Number of times the Most common phrase appearing:": 28
}
```


Fig. 4. Social Content Extraction using Topsy API

the most recent tweets posted could be extracted from the API regarding a certain URL. The content from all the tweets is collected to form a social context corpus.

From this corpus we extract the best replacement tweet by calculating the longest common N-gram. This represents the tweet with the most information that describes the target resource intended by the author. Within some tweets, multiple links coexist within the same text. These co-occurring resources share the same context and maintain a certain relevancy in most cases. A list of those co-occurring resources are extracted and filtered for redundancies. Finally, the textual components of the tweets are extracted after removing usertags, URIs, social interaction symbols like "RT". We named the document composed of those text-only tweets in the form of phrases the *"Tweet Document"*.

Figure 4 illustrates the JSON object produced from social mining the resource as described above.

4.2 Resource Replacement Recommendation

From the social extraction phase above we gathered information that helps us to infer the aboutness and context of a resource. Given this context, can we utilize it in obtaining a viable replacement resource to fill in the missing one and provide the same context?

To answer this, we utilize the work of Klein and Nelson [9] in defining the lexical signatures of web pages as discussed earlier. First, we extract the tweet document as described above. Next, we remove all the stop words and apply Porter's stemmer to all the remaining words [16]. We calculate the term frequency of each stemmed word and sort them from highest occurring to the lowest. Finally, we extract the top five words to form our tweet signature.

On the one hand, and using this tweet signature as a query, we utilize Google's search engine to extract the top 10 resulting resources. On the other hand, we collect all the other co-occurring pages in the tweets obtained by the API. These

pages combined produce a replacement candidate list of resources. One or more of which can be utilized as a viable replacement of the resource under investigation.

To choose which resource is more relevant and a possibly better replacement we utilize once more the tweet document extracted earlier. For each of the extracted pages in the candidate list, we download the representation and utilize the boilerpipe library in extracting the text within[1]. The library provides algorithms to detect and remove the "clutter" (boilerplate, templates) around the main textual content of a web page. Having a list of possible candidate textual documents and the tweet document, the next step is to calculate similarity. The pages are sorted according to the cosine similarity to the tweets page describing the resource under reconstruction.

At this stage we have extracted contextual information about the resource and a possible replacement. The next step is to measure how well the reconstruction process was undergone and how close is this replacement page is to the missing resource.

5 Evaluation

Since we cannot measure the quality of the discovered context or the resulting replacement page to the missing resource, we have to set some assumptions. We extract a dataset of resources that are currently available on the live web and assume they no longer exist. Each of these resources are textual based and neither media files nor executables. Each of these resources has to have at least 30 retrievable tweets using Topsy's API to be enough to build context.

We collect a dataset of 731 unique resources following these rules. We perform the context extraction and the replacement recommendation phases. We download the resource under investigation ($R_{missing}$) and the list of candidate replacements from the search engines (R_{search}) and the list of co-occurring resources ($R_{co-occurring}$). For each we use the boilerpipe library to extract text and use cosine similarity to perform the comparisons. For each resource, we measure the similarity between the ($R_{missing}$) and the extracted tweet page. For each element in (R_{search}) we calculate the cosine similarity with the tweet page and sort the results accordingly from most similar to the least. We repeat the same with the list of co-occurring resources ($R_{co-occurring}$). Then we calculate the similarity between ($R_{missing}$) and ($R_{search}(first)$) indicating the top result obtained from the search engine index. Then, we compare ($R_{missing}$) with each of the elements in (R_{search}) and ($R_{co-occurring}$) to demonstrate the best possible similarity. Figure 5 illustrates the different similarities sorted for each measure and shows that 41% of the time we can extract a significantly similar replacement page ($R_{replacement}$) to the original resource ($R_{missing}$) by at least 70% similarity. Finally, we needed to validate the effectiveness of using the tweet signature as a query string to the search engine. Using the tweet signature extracted from tweets associated with an existing resource against the search engine API and locating the rank in which the resource appear in the results list, we calculate the mean reciprocal rank to be 0.43.

[1] `http://code.google.com/p/boilerpipe/`

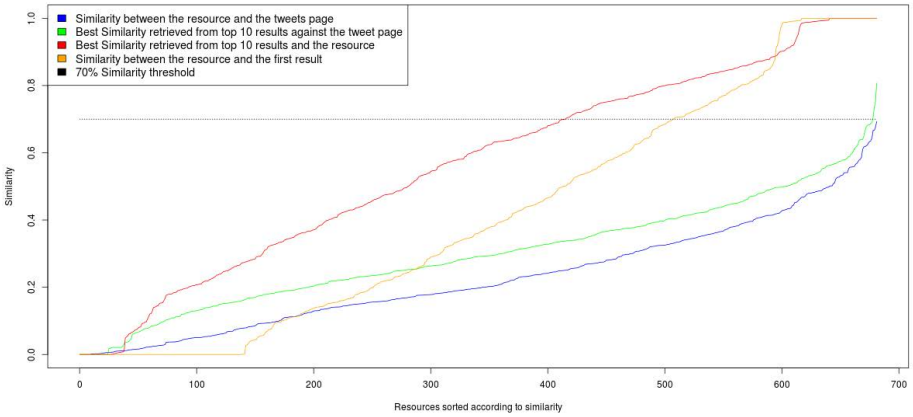

Fig. 5. Similarities with the original resource $R_{missing}$

6 Conclusions and Future Work

In this study we verify our previous analysis and estimation of the percentage missing of the resources shared on social media. The function in time still holds in modeling the percentage disappearing from the web. As for the model estimated for the amount archived it showed an alteration. The slope of the regression line in the model stayed the same while the y-intercept varied. We deduce that a possible explanation to this phenomena is due to timemap shrinkage. Previously, timemaps incorporated search engine caches as mementos which was removed in the most recent Memento revision. Next, we classified web resources into four different categories in regards to existence on the live web and in public web archives. Then we considered the unrecoverable category where the resource is deemed missing from the live web whilst not having any archived versions. Since we cannot perform a full reconstruction or retrieval, we utilize the social nature of the shared resources by using Topsy's API in discovering the resource's context. Using this context and the co-occurring resources we apply a range of heuristics and comparisons to extract the most viable replacement to the missing resource from its social neighborhood. Finally, we performed an evaluation to measure the quality of this replacement and found that for 41% of the resources we can obtain a significantly similar replacement resource with at least 70% similarity. For our future work, we would like to expand our investigation to incorporate other resources of different types like images and videos. A further investigation is crucial to better rank the results and account for the different types of resources.

Acknowledgments. This work was supported in part by the Library of Congress and NSF IIS-1009392.

References

1. Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., Nelson, M.L.: How Much of the Web Is Archived? In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL 2011, pp. 133–136 (2011)
2. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Identifying 'Influencers' on Twitter. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011 (2011)
3. Bar-Yossef, Z., Broder, A.Z., Kumar, R., Tomkins, A.: Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay. In: Proceedings of the 13th International Conference on World Wide Web, WWW 2004, pp. 328–337 (2004)
4. Baykan, E., Henzinger, M., Marian, L., Weber, I.: Purely URL-based topic classification. In: Proceedings of the 18th International Conference on World wide web, WWW 2009, pp. 1109–1110 (2009)
5. Benevenut, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing User Behavior in Online Social Networks. In: Proceedings of ACM SIGCOMM Internet Measure- ment Conference, SIGCOMM 2009, pp. 49–62 (2009)
6. Brunelle, J.F., Nelson, M.L.: An Evaluation of Caching Policies for Memento TimeMaps. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2013 (2013)
7. Gill, A.J., Nowson, S., Oberlander, J.: What are they blogging about? Personality, topic and motivation in blogs. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, ICWSM 2009 (2009)
8. Kan, M.-Y.: Web page classification without the web page. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, WWW Alt. 2004, pp. 262–263 (2004)
9. Klein, M., Nelson, M.L.: Revisiting lexical signatures to re-discover web pages. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 371–382. Springer, Heidelberg (2008)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 591–600 (2010)
11. Mark, G., Bagdouri, M., Palen, L., Martin, J., Al-Ani, B., Anderson, K.: Blogs as a collective war diary. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW 2012, pp. 37–46 (2012)
12. McCown, F., Marshall, C.C., Nelson, M.L.: Why web sites are lost (and how they're sometimes found). Communications of the ACM, 141–145 (November 2009)
13. McCown, F., Nelson, M.L.: What happens when facebook is gone. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2009, pp. 251–254 (2009)
14. McCown, F., Nelson, M.L.: A framework for describing web repositories. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2009, pp. 341–344 (2009)
15. Qi, X., Davison, B.D.: Knowing a web page by the company it keeps. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, pp. 228–237 (2006)
16. Porter, M.F.: An algorithm for suffix stripping. Program: electronic library and information systems 14, 313–316 (1980)

17. SalahEldeen, H.M., Nelson, M.L.: Losing my revolution: how many resources shared on social media have been lost? In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 125–137. Springer, Heidelberg (2012)
18. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who Says What to Whom on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, WWW 2011, pp. 705–714 (2011)
19. Starbird, K., Muzny, G., Palen, L.: Learning from the Crowd: Collaborative Filtering Techniques for Identifying On-the-Ground Twitterers during Mass Disruptions. In: Proceedings of the 9th International ISCRAM Conference, ISCRAM 2012 (2012)
20. Starbird, K., Palen, L. (How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW 2012, pp. 7–16 (2012)
21. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 1079–1088 (2010)
22. Yang, J., Counts, S.: Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. In: 4th International AAAI Conference on Weblogs and Social Media, ICWSM 2010 (2010)
23. Zhao, D., Rosson, M.B.: How and Why People Twitter: The Role that Micro- blogging Plays in Informal Communication at Work. In: Proceedings of the ACM 2009 International Conference on Supporting Group Work, GROUP 2009, pp. 243–252 (2009)