A First Study on Temporal Dynamics of Topics on the Web

Aécio Santos
aecio.santos@nyu.edu

Bruno Pasini
bmpasini@nyu.edu

Juliana Freire
juliana.freire@nyu.edu

Tandon School of Engineering
New York University

ABSTRACT

While much work has been devoted to understanding Web dynamics and using this knowledge to efficiently maintain the freshness of the indexes of generic search engines, the same is not true for domain-specific indexes constructed by focused crawlers. For the latter, the problem is compounded by the fact that it is important not only to maintain already-crawled pages fresh, but also to identify new relevant content and expand the collection. In this paper, we discuss the challenges involved in this problem and describe our preliminary efforts in building a testbed to better understand the dynamics of specific topics and characterize how they evolve over time. We propose a data collection methodology and a set of experiments to answer important questions about temporal dynamics and evolution of topics. We also present the results of the experimental analysis we carried out using data collected over a period of four weeks using two distinct topics. These results suggest that topic-specific refreshing strategies can be beneficial for focused crawlers.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Experimentation

Keywords

Web crawling, Focused crawling, Web temporal dynamics, Web analytics

1. INTRODUCTION

Studies on temporal dynamics and evolution of pages on the Web have uncovered important characteristics about the behavior of the Web [1, 2, 11, 12, 13]. Understanding the temporal dynamics of Web pages is crucial to the design of predictive models and efficient refreshing strategies. For example, by predicting the change rate of a page, a crawler can adopt an efficient refreshing policy that maximizes freshness and minimizes the crawling overhead [5]. Furthermore, as new pages are created at a very fast pace, web crawlers can discover new links that lead to these pages efficiently by revisiting a well-chosen subset of previously-crawled pages [11]. These studies, however, considered the whole Web and examined strategies applicable to generic crawlers. For focused crawlers, one key question that remains open is how specific *topics* evolve over time.

Focused crawling provides a scalable and effective alternative to search for pages on a specific topic that represents a small segment of the Web [6]. Instead of attempting to cover all web pages, a focused crawler tunes its strategy to search for a target topic—it attempts to maximize the number of on-topic pages it retrieves and minimize the number of irrelevant pages visited. Focused crawlers thus bring many benefits: they lead to substantial savings in hardware and network resources compared to searching the *whole* Web; they make it cheaper to maintain the crawl up-to-date; crawls can go deeper and obtain a better coverage for a given topic; and the derived index, being focused, is more likely to return a higher fraction of relevant pages, reducing the information overload.

In this paper, to the best of our knowledge, we present the first detailed study that tries to shed light on the temporal dynamics and evolution of topics on the Web from the perspective of a focused crawler. Similar to general crawlers, focused crawlers can benefit from revisiting strategies that maximize freshness of a page collection while minimizing resource consumption. These crawlers, however, have an additional requirement: to maintain *topic freshness*, they need to discover new relevant content. Thus, in our study, we attempt to answer the following questions:

1. Do pages in different topics have different change patterns? What are the implications of such differences for revisiting strategies that aim to maximize page freshness?

2. Are web pages stable with regard to relevance to a topic? In other words, do on-topic pages become off-topic, and if so, how often does this happen?

3. How do topics evolve? How does the creation and deletion rate of on-topic pages vary for different topics? Is revisiting of previously crawled pages a good strategy to discover new links to on-topic pages?

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889291.

We built a dataset that contains web pages in two distinct topics. Over a four-week period, we monitored these pages for changes in both their textual content and in the content of neighboring on-topic pages, and analyzed the change patterns. Our main findings include: the studied topics display different change patterns; although most pages tend to be stable regarding its relevance to a topic, some pages change its relevance at different rates; and revisiting already downloaded pages, a focused crawler can discover new pages, but many of these are off-topic. These findings suggest that topic-specific refreshing policies may benefit focused crawlers both in maintaining page freshness and in increasing the revisitation harvest rate.

Our main contributions can be summarized as follows:

- We explore a new problem: understanding temporal dynamics of topics on the Web.
- We describe a methodology to collect and monitor change patterns and evolution of pages that belong to a topic.
- We present a study of the changes for two different topics over a four-week period. This study suggests that topic-specific revisitation strategies are likely to be more effective for focused crawlers than policies adopted by generic crawlers.

The remainder of this paper is organized as follows. We discuss related work in Section 2. The data used in our study and the methodology used to collect them are presented in Section 3. In Section 4, we describe our experiments and discuss our main findings. We conclude in Section 5, where we outline directions for future research.

2. RELATED WORK

Several studies have considered the temporal aspects of pages on the Web [5, 8, 9, 10, 11, 12, 15, 16, 17]. Coffman et al. [10] postulated that web page change events follow a Poisson process, which means that changes occur randomly and independently. Cho and Garcia-Molina [8] studied web page changes considering the Poisson model, and proposed efficient change frequency estimators for various scenarios, including the web crawling scenario in which information about all change events is not available. Other works found that features extracted from the content of web pages, web link structure, and web search logs can be effectively used for predicting change patterns [5, 17, 15]. Barbosa et al. [5] were the first to exploit the use of static features extracted from the content of a page to predict its change behavior. Based on this idea, Tan and Mitra [17] proposed the use of new dynamic features, and other features extracted from the web link structure and web search logs to group pages with similar change behavior. Radinsky and Bennett [15] went a step further and proposed a change prediction framework that uses not only features from the content, but also the degree and relationship among the observed changes to a page, the relatedness to other pages, and the similarity in the kinds of changes they experienced.

The temporal dynamics and evolution of the Web content have been studied in [11, 13, 2, 1]. Ntoulas et al. [13] observed that web pages are created and retired at a very fast pace. Olston and Pandey [14] studied the longevity of information found in web pages and proposed recrawl scheduling policies that allow crawlers to target persistent content, instead of ephemeral content that will be quickly overwritten by subsequent changes. Adar et al. [1] studied how Web content changes both with respect to time intervals (hourly and sub-hourly crawls) and the actual changes (page-level changes, DOM tree changes and term-level changes). Bar-Yossef [2] studied the decay of the Web and showed that not only do some web pages exhibit a rapid death but also large subgraphs of the Web decay significantly. Dasgupta et al. [11] studied the extent to which new pages can be efficiently discovered by a crawler. They proposed algorithms that use historical statistics to estimate which pages are most likely to yield links to new content. They showed that, with perfect foreknowledge of where to explore for links to new content, it is possible to discover 90% of all new content by monitoring a small set of well-chosen pages, whereas substantially more effort is required to find the remaining 10%.

In contrast to these works, which have studied the dynamics of the whole Web graph, we consider specific topics that represent (small) sub-graphs of the Web. In this paper, we study topics that are derived by page classifiers commonly used by focused crawlers.

3. DATA COLLECTION

In order to carry our experimental analysis, we built a dataset that contains web pages in two distinct topics: Ebola and movies. Note that a challenge in creating a testbed to study the dynamics of topics on the Web is to define the topics and configure a process to collect the relevant data. We selected these two topics for two main reasons. First, they are very different in nature – this is confirmed in the different behaviors observed in our experimental analysis (Section 4). Second, there is human-edited data available in DMOZ[1] as a starting point to obtain a list of seed URLs and labeled training data for training page classifiers for these topics.

We used the open-source ACHE focused crawler[2] to collect the data [3, 4]. ACHE supports the customization of crawling policies – it can perform both generic and focused crawls. We followed a multi-step process to select and collect the data.

Step 1. First, for each topic, we provided as input to ACHE a list of seed URLs created using data from DMOZ and URLs manually collected by a user from a commercial search engine. Starting from these URLs, we ran ACHE using breadth-first policy (i.e., with no restrictions in what links to follow) to obtain the content of the seed pages as well as URLs for neighboring pages.

Step 2. In the second step, we manually inspected random samples of the pages collected to label the positive and negative examples required to train page classifiers for the two topics. Table 1 shows the training set size and evaluation metrics of the models. Precision, recall and F-measure values were computed using stratified 10-fold cross-validation. We used Support Vector Machine (SVM) as the learning classifier, and as features, we used the term frequencies of the text extracted from the page's HTML `body`, `title` and `meta` tags. To select only the best terms as features for our models, we used document frequency as a simple feature

[1] http://www.dmoz.org/

[2] https://www.github.com/ViDA-NYU/ache

Table 1: Training data and classifiers metrics

	Ebola	Movies
# positive examples	1,594	1,202
# negative examples	2,912	3,324
% positive examples	35.38%	26.56%
% negative examples	64.62%	73.44%
% precision	91.48%	85.87%
% recall	99.06%	97.59%
% F-measure	95.12%	91.36%

selection method to choose the top 5,000 most important terms.

Step 3. After training the models for Ebola and movies, we created a final set of seed URLs to be monitored daily. The set consists of URLs randomly sampled from the respective DMOZ categories and from the data collected in the previous crawls (Step 1). Note that this seed set is not the same set used in the Steps 1 and 2. Because we also downloaded all URLs extracted from these seed web pages (as detailed in Step 4), we limited the monitored seed set to contain a small number (at most 6) of URLs per domain. We did this for two main reasons. First, we wanted to avoid downloading too many links from web sites that have a large number of links per page, as this would lead to a collection with a disproportionate number of pages from these large web sites, and thus statistically bias the results. Second, we wanted to ensure politeness and avoid overloading the web servers where the monitored pages reside. Because the time in one day is limited and we can not send requests to the same web server too often, the number of web pages downloaded per web server also needs to be limited.

Step 4. We monitored the final seed daily for approximately one month, between December 2015 and January 2016. The monitoring process was carried out as follows. Every day, the crawler downloaded all seed URLs, and then retrieved all links in the neighborhood of the seeds, using a breadth-first search with depth 1 starting from the seeds. This way, we ensured that all outlinks from our seed URLs are also downloaded and classified as on-topic or off-topic. A summary of the collected data is given in Table 2. For our experimental analysis in Sections 4.1 and 4.2, we considered only pages that were successfully downloaded every day. Table 2 show the numbers of pages before (raw) and after removing these pages (filtered). The percentage of on-topic/off-topic web pages, was measured before cleaning the data, in order to preserve the ratio as it is on the Web. Table 2 also shows the number of on-topic web sites, which is the number of distinct domains included in the filtered data used to compute our metrics.

4. EXPERIMENTAL ANALYSIS

In this section, we discuss a series of analyses we performed over the collected data (as described in Section 3). Our goal is to answer the questions outlined in Section 1: (1) how does the change rate of pages vary across different topics? (Section 4.1); (2) does the relevance of pages with respect to a topic change over time? (Section 4.2); and (3) what page creation patterns can be observed in different topics? (Section 4.3).

Table 2: Overview of the dataset.

Topic	Ebola	Movies
Monitoring period	29 days	28 days
Start date	12/15/2015	12/19/2015
End date	01/12/2016	01/15/2016
% on-topic web pages	24.21%	70.73%
% off-topic web pages	75.79%	29.27%
# on-topic web pages (raw)	80,437	67,687
# on-topic web pages (filtered)	22,200	27,353
# on-topic web sites	2,177	1,422
# monitored seed URLs	4,239	2,219
Average webpages/site	39.34	19.24

Table 3: Average Change Rate for on-topic and off-topic pages in each topic

Topic	Ebola	Movies
On-topic	0.469773	0.373650
Off-topic	0.458534	0.245847

4.1 Change Rate of Page Contents

To address our first question, we analyzed the rate of change of the web page content over time. To quantify how often a page p changes, we adopted the Change Rate (CR) metric, which is defined by:

$$CR = \frac{n}{X}$$

where X is the number of times the page was revisited and n is the number of times the content of page p changed in the X revisits. We consider that a page changed if the text extracted from the HTML has changed. Although we did not apply any sophisticated method to remove irrelevant content such as ads, sidebars, and menus, we considered only tokenized terms in the comparison (all non-textual content such as spaces, line breaks and other characters were removed).

We computed the Change Rate for each web page that was successfully downloaded every day. Table 3 shows the average Change Rate by topic for all on-topic and off-topic pages. Note that the different topics have very different Change Rates. In our data, pages classified as belonging to the topic Ebola change much more frequently than pages that belong to the Movies topic. The table also shows that even within the same topic, pages classified as off-topic may have a change rate that is different from that of pages classified as on-topic. This was observed for Movies.

Another aspect we investigated was the distribution of change rate for all pages. Figure 1 shows a histogram of change rates for all on-topic and off-topic pages in each topic. First of all, we can see that most pages are concentrated in the ranges $[0.0, 0.1]$ and $(0.9, 1.0]$, which means that most pages either have very high or very low change rates. Indeed, as can be seen in Table 4, if we consider only the topic Movies, 32.98% of pages never changed in all revisits ($CR = 0.0$) and 24.91% of pages changed in every revisit ($CR = 1.0$). Another difference we observed is that Movies has a larger proportion of pages that never change (32.98%) than the topic Ebola (24.91%).

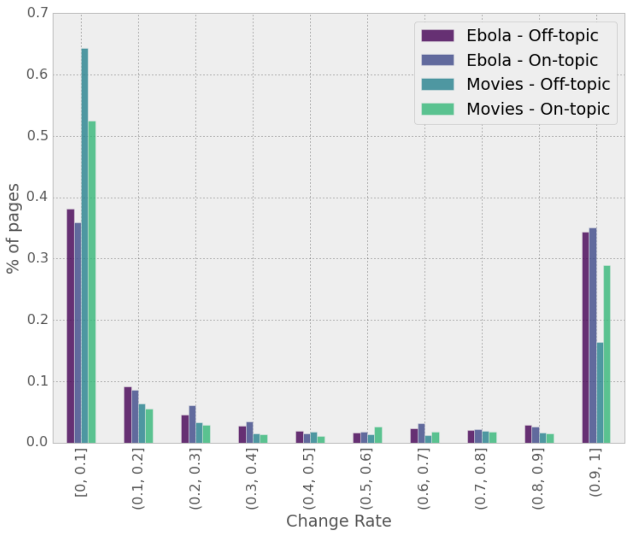

Figure 1: Distribution of Change Rate for on-topic and off-topic pages in each topic

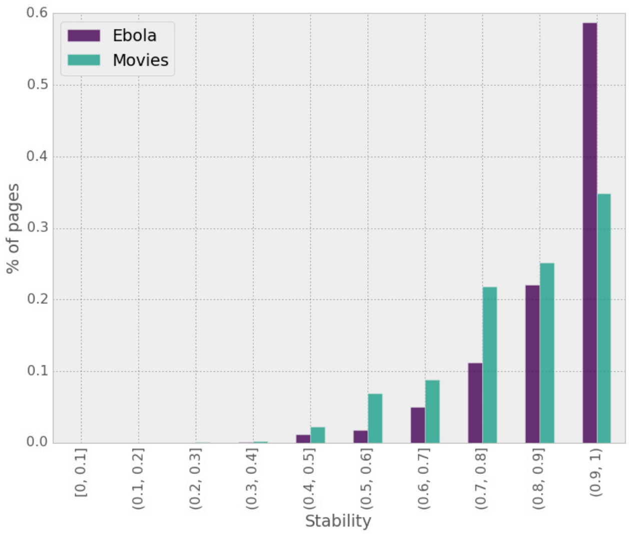

Figure 2: Distribution of Stability for each topic disregarding stable pages ($stability = 1.0$)

Table 4: Pages with Change Rate (CR) equal to 0.0 or 1.0

Topic	Ebola	Movies
Total	22,200	27,353
CR=0.0	3,900	9,022
CR=1.0	6,365	6,813
% CR=0.0	17.97	32.98
% CR=1.0	28.67	24.91

Table 5: Stability of pages by topic

Topic	Ebola	Movies
Total	91,129	40,737
# stable pages	88,941	37,458
% stable pages	97.60	91.95

4.2 Stability of Page Relevance

To study the stability of the relevance of pages to a topic, we analyzed how frequently the relevance changes over time. To quantify how often a page changes its relevance, we defined the following metric:

$$stability = 1 - \frac{\sum_{i=1}^{X} relevance\_changed(i)}{X}$$

where X is the number of times the page was revisited, and $relevance\_changed(i)$ is a function that returns 1 if relevance of the page p changed in revisit i, and 0 otherwise. We consider that the relevance of a page changes if the class predicted by the page classifier in revisit i is different from the class predicted in revisit $i-1$.

We computed the stability for every web page present in our dataset. We found that most pages are stable regarding page relevance (i.e., $stability = 1.0$), although the actual number of stable pages may be very different for different topics. For example, as shown in Table 5, while the percentage of pages that have stability equal to 1.0 for the topic Ebola is 97.6%, the percentage of stable pages for topic Movies is much lower: 91.95%. This is an interesting result, since this contrasts with what we have found Section 4.1: although web pages about Ebola change more frequently (have a higher change rate) than pages in Movies, the relevance of an Ebola page rarely changes.

Since most of the pages are stable, to better visualize the behavior of unstable pages, in Figure 2, we plotted a histogram of the stability of web pages for each topic disregarding stable pages. Although the stability curves for the two topics show that most pages are more stable than unstable, the stability decreases at different rates, with minimum values close to the range $[0.3, 0.4]$.

4.3 Topic Evolution

To better understand topic evolution, we examined the rate at which new links to relevant pages appear or disappear in a web page. We defined the following metrics to quantify evolution-related changes:

- new_links_rate: the average number of new links found in page p after X revisits.
- $new\_relevant\_links\_rate$: the average number of new links found in page p after X revisits that point to pages classified as on-topic by the page classifier.
- $links\_gone\_rate$: the average number of links that disappeared from page p after X revisits.
- $relevant\_links\_gone\_rate$: the average number of links that disappeared from page p after X revisits that point to pages classified as on-topic by the page classifier.

We computed these metrics for all monitored seed web pages, as these are the only ones guaranteed to have all outlinks downloaded. Table 6 shows the mean and median values for these metrics by topic in our dataset, where X is 29 for Ebola and 28 for Movies. Once again, we can observe differences in behavior for the different topics. For example, while pages about Ebola have an mean new_links_rate of

Table 6: Mean and median values of topic evolution metrics by topic.

	Topic	new_links_rate	new_relevant_links_rate	links_gone_rate	relevant_links_gone_rate
Mean	**Ebola**	7.06	1.36	7.05	1.36
	Movies	2.48	2.10	2.47	2.08
Median	**Ebola**	1.07	0.14	1.04	0.14
	Movies	0.00	0.15	0.00	0.15

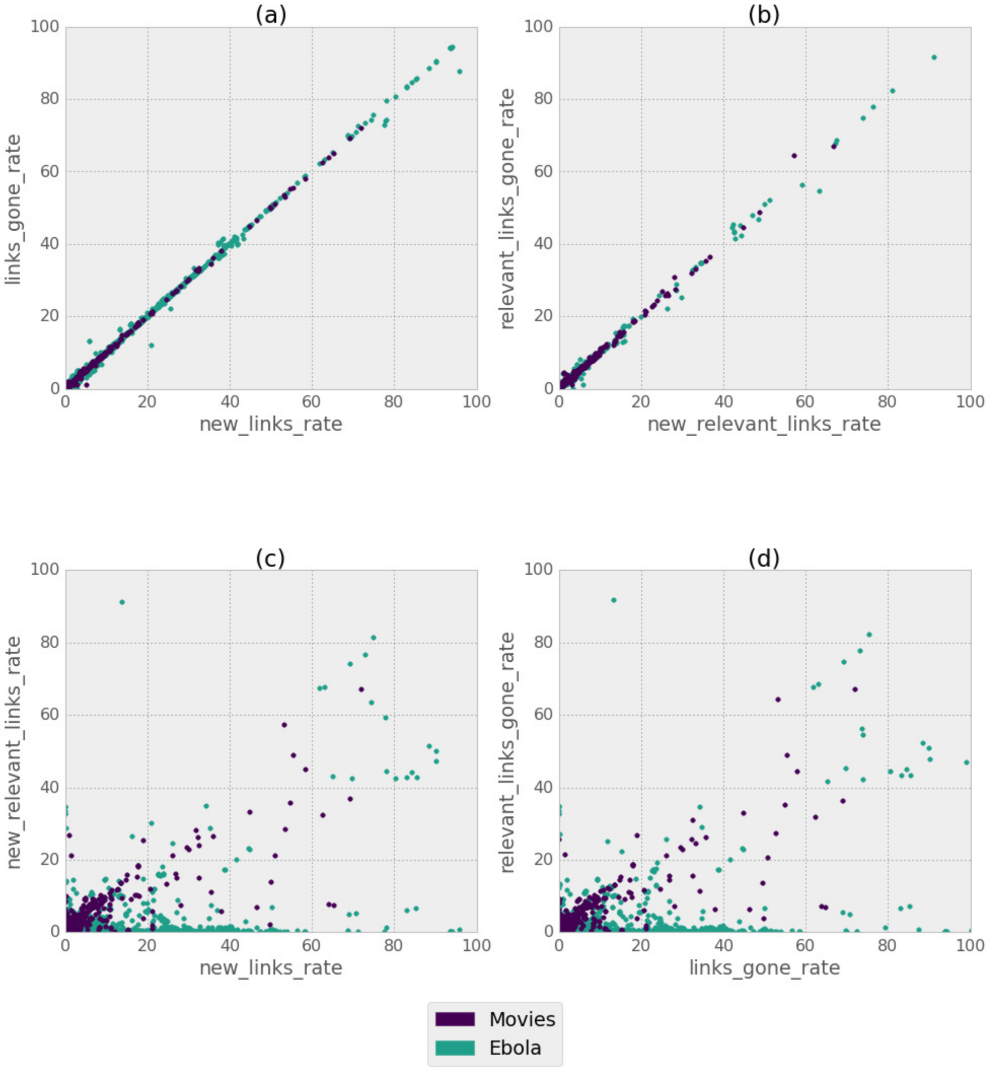

Figure 3: Scatter-plot of topic evolution metrics by topic.

7.06, pages about movies have a mean of 2.48. This means that every time we revisit page in the topic Ebola, we will find, on average, 7.06 different new links compared to the previous visit. Table 6 also shows the *new_relevant_links_rate*. We can see that the rate of new relevant links is smaller than the new links rate: not all links found in a revisited page will lead to relevant pages. This reinforces the importance of using a revisit strategy that takes link selection into account. A similar idea has been used in crawling strategies, which in addition to identifying whether a page is relevant, rank links to crawl in the order of their predicted distance to a relevant page [7, 4].

Table 6 also shows the average values for the rate of links gone and relevant links gone. One interesting fact to note is that, for both topics, the values of the *links_gone_rate* is similar to the *new_links_rate*. This can be explained by the fact that some pages are generated automatically by content management systems (CMS) that present a list of the top n most recently created pages. When new pages are created on the web site, some items of these lists are replaced by links to the new pages. Figures 3(a) and 3(b) provide additional evidence of this behavior. These figures show scatter-plots of the new links rate by links gone rate. Note that most of the pages have similar values for both x and y axis.

Another pattern noticeable in Figures 3(c) and 3(d) is that the pages about Ebola have a higher *new_links_rate* around the $[0, 60]$ interval, while pages about movies are more concentrated around $[0, 15]$. We can also see that despite the fact that pages about Ebola tend to have a higher new links rate, the rate of relevant links is smaller than for Movies, staying around the range $[0, 5]$.

5. CONCLUSIONS

We explored a new and important problem for focused crawling: understanding temporal dynamics and evolution of topics on the Web. To this end, we developed a methodology to monitor web pages that belong to a topic, collected pages over four weeks, and analyzed the collected data. The results of our analysis uncovered interesting properties of topics on the Web:

- Distinct topics have different change patterns;
- Although most pages are stable with respect to relevance for a topic, the relevance of some pages very over time and at different rates;
- By revisiting previously downloaded pages, a focused crawler can discover links that lead to new pages, but many of these can be off-topic.

These findings suggest that a naïve refreshing policy that does not consider the change patterns intrinsic to a topic, may be insufficient to ensure an efficient recrawling strategy that maximizes freshness, minimizes resource consumption, and yields a high harvest rate. For the latter, we observed that while some web pages change and produce new links, not all of these links lead to on-topic pages. Thus, if the crawler is able to predict which pages yield a larger number new links as well as the likelihood of these links to point to relevant pages, efficiency can be substantially improved.

Our preliminary results are promising and provide insights into how to construct effective focused revisiting policies that take into account the change rate and new link discovery. There are several avenues we plan to explore in future work. Besides exploring additional topics and larger page collections, we would like to track their evolution over a longer period of time. In addition, we will explore features that can lead to effective change predictors and evaluate different revisit strategies.

Acknowledgments. This work was funded by the Defense Advanced Research Projects Agency (DARPA) MEMEX program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

6. REFERENCES

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second International Conference on Web Search and Web Data Mining*, pages 282–291, 2009.

[2] 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*, pages 328–337, 2004.

[3] L. Barbosa and J. Freire. Searching for hidden-web databases. In *Proceedings of the Eight International Workshop on the Web & Databases (WebDB 2005)*, pages 1–6, 2005.

[4] L. Barbosa and J. Freire. An adaptive crawler for locating hidden-web entry points. In *Proceedings of the 16th International Conference on World Wide Web*, pages 441–450, 2007.

[5] L. Barbosa, A. C. Salgado, F. de Carvalho, J. Robin, and J. Freire. Looking at both the present and the past to efficiently update replicas of web content. In *7th ACM International Workshop on Web Information and Data Management*, pages 75–80, 2005.

[6] S. Chakrabarti. Focused web crawling. In *Encyclopedia of Database Systems*, pages 1147–1155. Springer, 2009.

[7] S. Chakrabarti, K. Punera, and M. Subramanyam. Accelerated focused crawling through online relevance feedback. In *Proceedings of the 11th International Conference on World Wide Web*, pages 148–159, 2002.

[8] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Transactions on Internet Technology*, 3:256–290, 2003.

[9] J. Cho and A. Ntoulas. Effective change detection using sampling. In *Proceedings of the 28th International Conference on Very Large Data Bases*, pages 514–525, 2002.

[10] E. G. Coffman, Z. Liu, and R. R. Weber. Optimal robot scheduling for web search engines. *Journal of Scheduling*, 1(1), 1998.

[11] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins. The discoverability of the web. In *Proceedings of the 16th International Conference on World Wide Web*, pages 421–430, 2007.

[12] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*, pages 669–678, 2003.

[13] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: The evolution of the web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, pages 1–12, 2004.

[14] C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In *Proceedings of the 17th International Conference on World Wide Web*, pages 437–446, 2008.

[15] K. Radinsky and P. Bennett. Predicting content change on the web. In *6th ACM International Conference on Web Search and Data Mining*, 2013.

[16] A. Santos, C. de Carvalho, J. Almeida, E. de Moura, A. da Silva, and N. Ziviani. A genetic programming framework to schedule webpage updates. *Information Retrieval Journal*, 18(1):73–94, 2015.

[17] Q. Tan and P. Mitra. Clustering-based incremental web crawling. *ACM Transactions on Information Systems*, 28:17:1–17:27, 2010.