Scraping SERPs for Archival Seeds: It Matters When You Start

Alexander C. Nwala  
Old Dominion University  
Norfolk, Virginia, USA  
anwala@cs.odu.edu

Michele C. Weigle  
Old Dominion University  
Norfolk, Virginia, USA  
mweigle@cs.odu.edu

Michael L. Nelson  
Old Dominion University  
Norfolk, Virginia, USA  
mln@cs.odu.edu

ABSTRACT
Event-based collections are often started with a web search, but the search results you find on Day 1 may not be the same as those you find on Day 7. In this paper, we consider collections that originate from extracting URIs (Uniform Resource Identifiers) from Search Engine Result Pages (SERPs). Specifically, we seek to provide insight about the retrievability of URIs of news stories found on Google, and to answer two main questions: first, can one “refind” the same URI of a news story (for the same query) from Google after a given time? Second, what is the probability of finding a story on Google over a given period of time? To answer these questions, we issued seven queries to Google every day for over seven months (2017-05-25 to 2018-01-12) and collected links from the first five SERPs to generate seven collections for each query. The queries represent public interest stories: “healthcare bill,” “manchester bombing,” “london terrorism,” “trump russia,” “travel ban,” “hurricane harvey,” and “hurricane irma.” We tracked each URI in all collections over time to estimate the discoverability of URIs from the first five SERPs. Our results showed that the daily average rate at which stories were replaced on the default Google SERP ranged from 0.21 – 0.54, and a weekly rate of 0.39 – 0.79, suggesting the fast replacement of older stories by newer stories. The probability of finding the same URI of a news story after one day from the initial appearance on the SERP ranged from 0.34 – 0.44. After a week, the probability of finding the same news stories diminishes rapidly to 0.01 – 0.11. In addition to the reporting of these probabilities, we also provide two predictive models for estimating the probability of finding the URI of an arbitrary news story on SERPs as a function of time. The web archiving community considers link rot and content drift important reasons for collection building. Similarly, our findings suggest that due to the difficulty in retrieving the URIs of news stories from Google, collection building that originates from search engines should begin as soon as possible in order to capture the first stages of events, and should persist in order to capture the evolution of the events, because it becomes more difficult to find the same news stories with the same queries on Google, as time progresses.

KEYWORDS
Collection building; Web Archiving; Crawling; Discoverability

ACM Reference Format:

1 INTRODUCTION AND BACKGROUND

From elections to natural disasters, web collections provide a critical source of information for researchers studying important historical events. Collections can be built automatically with focused crawlers or manually by an expert user. For example, an archivist at the National Library of Medicine collected seeds on Archive-It for the 2014 Ebola Outbreak event [23]. Collections may also be built by multiple users. For example, the Internet Archive has on multiple occasions requested that users submit seeds via Google Docs to build collection for events such as the 2016 US Presidential Election and the Dakota Access Pipeline (DAPL) event. Depending on when users begin to contribute seeds, URIs (Uniform Resource Identifiers) for early news stories may be difficult to discover via Google after one month, for as we show in this paper they can quickly fall to distant SERPs (Search Engine Result Pages).

Collection building often begins with a simple Google search to discover seeds. This can be done by issuing queries to Google and extracting URIs from the SERP (Fig. 1). For example, the following are two possible URI candidates extracted from the Google SERP (Fig. 1) to include in a collection (or seed list) about the Hurricane Harvey (August, 2017) event:

https://www.cbsnews.com/feature/hurricane-harvey/

A SERP provides an opportunity to generate collections for news stories and events, therefore we focused on assessing the discoverability of news stories on the Google SERP. Queries used to extract news stories are examples of informational queries [9], and we expect their SERP results to change as the news event evolves. We expect the SERP results for transactional (e.g., “samsung galaxy s3”) or navigational (e.g., “youtube”) to be less transient [18], but such queries are not the focus of our collection building effort. The URIs extracted from Google can serve as seeds: the seeds can be crawled to build collections in Archive-It, such as the Archive-It 2017 Hurricane Harvey collection1. It is important to understand the behavior of Google as this will influence the collections or seeds generated from it. This is not easy because the inner workings of Google are proprietary, making it a black box. To build a representative collection about an event, it is important to capture not just a slice of

---

1https://archive-it.org/collections/9323
time, but the various stages of the events [27] - oldest to the newest. For example, on May 25, 2017, we issued the query: “healthcare bill” to Google and extracted links (Table 1, 2017-05-23 – 2017-05-25) from the SERP. Seven months later (January 5, 2017) we repeated the same operation (Table 1, 2017-12-19 – 2017-12-20). The May 2017 healthcare bill collection shows the initial stages of the American Health Care Act of 2017 (AHCA) by highlighting the challenges facing the passage of the bill. On the other hand, the January 2018 collection shows a different (more recent) stage of the bill, highlighting the failure of the bill and the continued existence of Obamacare. These reflect the tendency of Google to highly rank newly created URIs for these kinds of queries. We quantify this behavior by estimating the rate at which new stories occur on the Google SERP. The tendency of Google to return fresh documents can be altered by setting the custom date range parameter on the site. However, the date range information is not always available for the collections we intend to build. We explore how this parameter affects the kinds of documents returned. It is crucial to know the dynamics of finding initial stories on Google, as this would inform the time a collection building task ought to begin; if we know how fast new documents replace old documents on the Google SERP, we can plan collection building to ensure that URIs of older stories are included in the collection and not just the recent ones.

Accordingly, we conducted a longitudinal experiment to understand whether it is possible to retrieve a given URI of a news story over time, to gain insight about the appearance/disappearance of news stories across the pages in Google, and to identify the probability of finding the same URI of a story using the same query over time. This was achieved by issuing seven queries (Table 2) every day for over seven months (2017-05-25 to 2018-01-12), and collecting links within h3 HTML tags from the first five pages (Fig. 1, annotation C) of the Google SERPs (Fig. 1). The queries were issued semi-automatically using a variant of the Local Stories Collection Generator [2].

The longitudinal study resulted in the following contributions that shed some light on the discoverability and behavior of news stories on Google as time progresses. First, the tendency of Google to replace older documents with new ones is well known, but not the rate at which it occurs, our findings quantify this. Given two time intervals, e.g., day 0 and day 1, if we collected a set x stories on day 0 and a set y stories on day 1, the story replacement rate on day 1 is the fraction of stories found on day 0 but not found on day 1 ($|x - y|/|x|$). The daily story replacement rate on the Google SERP ranged from 0.21 – 0.54, the weekly rate ranged from 0.39 – 0.79, and monthly - 0.59 – 0.92. This means if you re-issued a query after one day, between 21% to 54% stories are replaced by new stories. But if you waited for one month, and issued the same query, between 59% and 92% of the original stories are replaced. The News vertical SERP showed a higher story replacement rates: daily - 0.31 – 0.57, weekly - 0.54 – 0.82, and monthly - 0.76 – 0.92. Second, the probability of finding the same URI of a news story with the same query declines with time. For example, the probability of finding the same URI with the same query after one day (from the initial appearance on the General SERP) is between 0.34 – 0.44. This probability drops rapidly after a week, to a value between 0.01 – 0.11. The probability is less for the News vertical (daily - 0.28 – 0.40, weekly - 0.03 – 0.14, and approximately 0 one month later). We provide two predictive models that estimate the probabilities of finding an arbitrary URI of a news story on the General and News vertical SERPs as a function of time. Third, we show that news stories do not gradually progress from page 1 to page 2, 3, etc., and then out of view (beyond the page length considered). The progression we observed is less elegant (e.g., page 5 to 1, 3 to 2). These findings are highly informative to collection building efforts that originate from Google. For example, the results suggest that collections that originate from Google should begin days after an event happens, and should continue as time progresses to capture the various stages in the evolution of the event. Our research dataset comprising of 33,432 links extracted from the Google SERPs for over seven months, as well as the source code for the application utilized to semi-automatically generate the collections, are publicly available [3].

2 RELATED WORK

Since Chakrabarti et al. first introduced the focused crawler [12] as a system of discovering topic-specific web resources, there have been many research efforts pertaining to the generation of collections with some variant of a focused crawler. Bergmark [7] introduced a method for building collections by downloading web pages and subsequently classifying them into various topics in science, mathematics, engineering and technology. Farag et al. [15] introduced the Event Focused Crawler, a focused crawler for events which uses an event model to represent documents and a similarity measure to quantify the degree of relevance between a candidate
Table 1: Rows 1 - 3: Sample collection of URIs extracted from Google on May 25, 2017 with query: "healthcare bill." This shows the initial stages of the AHCA bill, highlighting the struggles to pass the bill. Rows 4 - 6: Sample collection of URIs extracted from Google on January 5, 2018 with query: "healthcare bill." This shows the later (recent) stages of the AHCA bill, highlighting the failure of the bill which happened in September 2017.

<table>
<thead>
<tr>
<th>Publication Date</th>
<th>Title</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-healthcare-saga-shaping-gop-approach-to-tax-bill</td>
</tr>
<tr>
<td>2017-12-19</td>
<td>GOP tax bill also manages to needlessly screw up the healthcare system</td>
<td><a href="http://www.latimes.com/business/lazarus-republican-tax-bill-individual-mandate-20171219-story.html">http://www.latimes.com/business/lazarus-republican-tax-bill-individual-mandate-20171219-story.html</a></td>
</tr>
</tbody>
</table>

URI and a collection. Klein et al. [19] demonstrated that focused crawling on the archived web results in more relevant collections than focused crawling on the live web for events that occurred in the distant past. In order to augment digital library collections with publications not already in the digital library, Zhuang et al. [32] proposed using publication metadata to help guide focused crawlers towards the homepages of authors. Klein et al. [20] also proposed a means of augmenting a digital library collection (the NASA Langley Research Center Atmospheric Science Data Center) with information discovered on the web using search engine APIs. SERPs are useful artifacts in their own right, and can be used for activities such as classifying queries as "scholarly" or "non-scholarly" [24]. This work is similar to these efforts that use focused crawlers as it relates to collection building and using search engines to discover seeds, but we do not use a focused crawler to discover URIs.

Zheng et al. [31] demonstrated that seed selection for crawlers is not a trivial problem because different seeds may result in collections that are considered "good" or "bad," and proposed different seed selection algorithms. Similarly, Schneider et al. [28] expressed the difficulty in identifying seed URIs for building thematic collections, and suggested the continuous selection of seeds for collections about unfolding events such as the 9/11 attacks, to accommodate the evolution of the events. Baroni et al. [6] presented their findings about the effectiveness of various seed selection strategies as part of a broader effort to build a large linguistically processed web-crawled corpora. They demonstrated the discovery of a variety of seeds by issuing random queries to search engine APIs. Similar to these efforts, we consider seed selection a vital part of collection building but mainly focus on the temporal considerations when selecting seeds from search engines.

Cho and Garcia-Molina [13] downloaded 720,000 pages (from 270 web servers) daily for four months to quantify the degree of change of web pages over time. They found that about 40% of all web pages changed within a week, and 50% of all the pages changed in about 50 days. Fetterly et al. [16] extended Cho and Garcia-Molina’s work by downloading (once a week for 11 weeks) over 150 million web pages and assessing the degree of change of web pages. They found that the average degree of change varies significantly across top-level domains; larger pages changed more frequently than smaller pages.

In the web archiving community, link rot and content drift [17, 21] are two major reasons for collection building. Comparably, the difficulty in refining news stories on the SERP suggests instant and persistent collection building by efforts that rely on the SERP for seed extraction or collection generation. McCown and Nelson [22] issued queries for five months to search engine web user interfaces (Google, MSN and Yahoo) and their respective APIs, and found significant discrepancies in the results found on both interfaces. They also showed how search results decay over time and modeled the decay rate. Kim and Vitor [18] studied Google, Bing, and Yahoo search engines, and showed that the top 10 results of 90% of their queries were altered within ten days. There are many more studies that examine the evolution of web pages [1, 14, 26] or the web [8, 30]. Our study is specific to news stories found in SERPs and not the evolution of the pages themselves. We sought to find out whether we could retrieve the same URI with the same query over a given period of time, instead of assessing the evolution of the content of individual web pages. This discoverability information is critical to collection building systems that utilize search engines. We tracked individual news stories to find when they were replaced by newer stories and quantified the rate at which older stories were replaced by newer stories. In addition to quantifying the rate of new stories as a function of time, we also quantify the rate of new stories for individual SERP pages. For example, we found that higher numbered pages (e.g., pages 3 - 5) have higher rates of new stories than lower numbered pages (e.g., pages 1 - 2) on Google. Our results enable understanding the dynamics of refining news stories on SERPs and are relevant to efforts that utilize search engines to discover seeds or to build collections.

Teevan et al. [29] illustrated the tendency of users to seek to re-find web pages previously seen at a rate of about 40% of all queries, and demonstrated how changes in search engine results can impede refining links. Aula et al. [4] also studied the prevalence of the "re-find" (re-access) behavior in a group 236 experienced
web users, by investigating the problems with searching and re-accessing information. Their survey showed that the users often used some strategies such as using several browser windows concurrently to facilitate re-access. Capra et al. [11] proposed a search engine use model as part of an effort to provide information to help better understand how users find, refine, and manage information on the web. There are many other studies outlining attempts to refind web resources on search engines, such as the effort of Bainbridge et al. [5] to refind four copies of their published research papers on Google Scholar and ACM Digital Library. Our research is similar to these efforts in the sense that we are interested in quantifying the likelihood of refinding URIs of news stories over time. However, the goal is not to fulfill the informational needs of particular users, but as part of an effort to extract seeds or generate collections from SERPs. Also, it is important to note that we utilize the search engine by issuing queries, it is not a known item search - we do not search for specific URIs to include in the collections we build - we let the SERP give us seeds. Instead we wish to know the rate as which the SERP produces the same URIs for the same query. Knowing the rate of new stories on the SERP for the same query indicates the ability of our collection generation process to refind previous stories and the rate at which the new stories from the SERP are included in the collection.

3 RESEARCH QUESTIONS

Before employing search engines as seed generators we sought to first assess the discoverability of news stories on SERPs (top five pages), and understand the dynamics of refinding news stories on SERPs. Our ability to build representative collections for events from SERPs is tied to the ability of retrieving old and new stories from SERPs. Our primary research question was: what is the rate at which new stories replace old stories on the SERP over time? This rate information may provide the means to approximate our ability to refind older stories about an event and the rate at which the collection we generate from SERPs receives new stories from SERPs. Similarly, if we found a URI of a news story on day \( d_0 \), what is the probability that we would find it on the next day \( d_1 \), or one week later on \( d_7 \)? For example, if we found a URI for a news story on the SERP on Monday, what is the probability that we would refind the same URI with the same query on Tuesday (one day later) or next Monday (one week later)?

The generalization of the first main question led to the second - what is the rate of new stories for the individual SERPs or their pages? For example, does the General SERP (or page 1) possess a higher new story rate than the News vertical SERP (or page 2)? The pages with the lowest rate of new stories may indicate the page with the highest probability of finding the initial reports of an event. Understanding the characteristics of both SERP classes could inform the choice of SERPs when extracting seeds.

Finally, we sought to gain insight about how news stories on SERPs moved across the different pages.

4 METHODOLOGY

Here we describe the generation of our dataset, the measures we extracted from the dataset, and how these measures informed our research questions.

4.1 Dataset generation, representation, and processing

Seven queries representing public interest stories were selected: “healthcare bill,” “manchester bombing,” “london terrorism,” “trump russia,” “travel ban,” “hurricane harvey,” and “hurricane irma.” The queries represent various events that happened (or are happening) in different timelines. Consequently, the dataset extraction duration varied for the queries as outlined by Table 2. The dataset extraction process lasted from 2017-05-25 to 2018-01-12. For each query, we extracted approximately 50 links within h3 HTML tags from the first five pages of the Google SERP from the default (All) and News vertical SERPs (Fig. 1, annotation A & B). To avoid confusion, in this research we renamed the All SERP to General SERP. The first five pages were considered in order to gain better insight about the rate of new stories across pages, as considering a few pages (e.g., 1 or 2) may present an incomplete view. In total, 73,968 (13,708 unique) URIs were collected for the General SERP and 77,634 (19,724 unique) for the News vertical SERP (Table 2).

Table 2: Dataset generated by extracting URIs from SERPs (General and News vertical) for seven queries between 2017-05-25 and 2018-01-12.

<table>
<thead>
<tr>
<th>Collection (Query/Topic)</th>
<th>Start date (duration in days)</th>
<th>News story count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>General SERP count (unique count)</td>
</tr>
<tr>
<td>healthcare bill</td>
<td>2017-05-25 (232)</td>
<td>12,809 (2,559)</td>
</tr>
<tr>
<td>manchester bombing</td>
<td>2017-05-25 (232)</td>
<td>12,451 (1,018)</td>
</tr>
<tr>
<td>london terrorism</td>
<td>2017-06-04 (222)</td>
<td>10,698 (1,098)</td>
</tr>
<tr>
<td>trump russia</td>
<td>2017-06-06 (220)</td>
<td>12,311 (4,638)</td>
</tr>
<tr>
<td>travel ban</td>
<td>2017-06-07 (219)</td>
<td>12,830 (2,849)</td>
</tr>
<tr>
<td>hurricane harvey</td>
<td>2017-08-30 (135)</td>
<td>6,666 (685)</td>
</tr>
<tr>
<td>hurricane irma</td>
<td>2017-09-07 (127)</td>
<td>6,203 (861)</td>
</tr>
<tr>
<td>Subtotal</td>
<td></td>
<td>73,968 (13,708)</td>
</tr>
<tr>
<td>Collections Total</td>
<td></td>
<td>151,602 (33,432)</td>
</tr>
</tbody>
</table>

In previous work with the Local Memory Project (LMP) [25], we introduced a local news collection generator [2]. The local news collection generator utilizes Google in order to build collections of stories from local newspapers and TV stations for US and non-US news sources. Unlike LMP, in this work we did not restrict the sources sampled to local news organization, but still utilized Google in order to discover seeds. The local news collection generator was used to scrape links from the Google SERP, and it was adapted to include the ability to extract all kinds of news stories from Google (not just from local news organizations). The Google search interface is meant for humans and not for robots and it presents a CAPTCHA when it is used too frequently in order to
discourage automated searches. Consequently, the dataset collections were all generated semi-automatically with the use of the local news collection generator. The input provided to the extension was the query and the maximum number of pages to explore (five), and the output was a collection of URIs extracted from the SERPs.

The URIs collected daily from the SERPs were represented as JSON files. For a single query, two JSON files per day were generated, each file represented the URIs extracted from the General SERP and News vertical SERP. This means for a given day, a total of 14 (two per query) JSON files were generated. Each URI in a JSON file included metadata extracted from the SERP such as the page number and the rank which is the position across all SERP pages. Additionally, each file included the date the data was generated.

At the center of the analysis was the ability to track the URI of a news story over time. A URI is a unique identifier for a resource, however, URIs often have aliases (multiple URIs identifying the same resource). For example, the following pair of URIs identify the same resource:

(a) https://www.redcross.org/donate/disaster-donations?campname=irm&campmedium=aspot
(b) https://www.redcross.org/donate/disaster-donations

As a result, we transformed all URLs before matching by trimming the scheme and all parameters from the URIs, using a method suggested by Brunelle et al. [10]. The parameters in URIs often express a reference source such as origin and callback, or session parameters such as session. The transformed version of the URI was used to track the individual news stories. Subsequently, for each news story we recorded all the dates and pages it was observed on the SERP.

### 4.2 Primitive measures extraction

The following measures were extracted from the dataset and provided information to help answer our research questions.

#### 4.2.1 Story replacement rate, new story rate, and page level new story rate.

Given that at time point $t_0$ we observed a set of URIs for news stories $u_0$, and at time point $t_1$ we observed a set of URIs for news stories $u_1$, the story replacement rate at $t_1$ is \( \frac{|u_0 - u_1|}{|u_0|} \). For example, if we observed URIs \{a, b, c\} at $t_0$ and URIs \{a, b, x, y\} at $t_1$, then the story replacement rate at $t_1$ is \( \frac{|\{a, b, c\} - \{a, b, x, y\}|}{|\{a, b, c\}|} = \frac{3}{3} = 0.3 \).

This means that at $t_1$ c was replaced. Similarly, the rate of new stories going from $t_0$ to $t_1$ is \( \frac{|u_0 - u_1|}{|u_0|} \). For example, if we observed URIs \{a, b, c\} at $t_0$ and URIs \{a, b, c, d, e\} at $t_1$, then the new story rate from $t_0$ to $t_1$ is \( \frac{|\{a, b, c, d, e\} - \{a, b, c\}|}{|\{a, b, c, d, e\}|} = \frac{3}{5} = 0.4 \).

This means that at $t_1$ we observed new stories $d$ and e. We calculated the story representation and new story rates using different temporal intervals (daily, weekly, and monthly) for the individual first pages of the General and News vertical SERPs. The daily story representation rate indicates the proportion of stories replaced on a daily basis. This is similar to the daily new story rate because the SERP returns a similar number of results (mean = median = mode = 10 links and $\sigma = 0.43$). The daily new story rate approximately indicates the rate of new stories that replaced previously seen stories on the SERP on a daily basis. The higher the story replacement and new story rates, the lower the likelihood of re-finding previously seen stories.

#### 4.2.2 Probability of finding a story.

Given a collection of URIs for news stories for a topic (e.g., "hurricane harvey"), consider the URI for a story $s_0$ that was observed for the first time on page 4 of the SERP on day $d_0$. We represent this as $s_0^{d_0} = 4$. If we find $s_0$ on page 2 on the next day $d_1$ and then it disappears for the next two days, we represent the timeline observation of $s_0$ as $\{4, 2, 0, 0\}$. Therefore, given a collection (e.g., "hurricane harvey") of $N$ URIs for news stories, the probability $P(s_k^d)$ that the URI of a story $s$ is seen after $k$ days ($d_k$), is calculated using Eqn. 1.

$$P(s_k^d) = \frac{\sum_{m=1}^{N} T(s_k^d)}{N} \cdot T(s_k^d) = \begin{cases} 0 \quad \text{if } s_k^d = 0 \\ 1 \quad \text{if } s_k^d > 0 \end{cases}$$

The probability $P(s_k^d = m)$ that the URI of a story $s$ is seen after $k$ days ($d_k$) on page $m$, is calculated using Eqn. 2.

$$P(s_k^d = m) = \frac{\sum_{m=1}^{N} T(s_k^d)}{N} \cdot T(s_k^d) = \begin{cases} 0 \quad \text{if } s_k^d \neq m \\ 1 \quad \text{if } s_k^d = m \end{cases}$$

#### 4.2.3 Distribution of stories over time across pages.

For each story URI, we recorded the dates it was observed on the SERP. For each date, we recorded the page where the story was found. The collection of stories and the date/page observations were expressed using the notation introduced in Section 4.2.2. For example, the following list of three URIs for news stories $s_0, s_1$, and $s_2$ were observed for the first time (first day - $d_0$) on pages, 4, 1, and 1, respectively. On the last day ($d_3$), the first story ($s_0$) was not seen on any of the pages ($s_0^{d_3} = 0$), however both the second ($s_1$) and third ($s_2$) stories were found on the first page ($s_1^{d_3} = 1$ and $s_2^{d_3} = 1$): $s_0 = \{4, 2, 0, 0\}$, $s_1 = \{1, 2, 0, 1\}$, and $s_2 = \{1, 1, 1, 1\}$.

#### 4.2.4 Overlap rate and recall.

Given two sets of collections of URIs, $A$ and $B$, the overlap rate $O(A, B)$ quantifies the amount of URIs common within both sets without considering the size disparities of the sets. This was calculated using the Overlap coefficient: $O(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$. The standard information retrieval recall metric $r(A, B)$ for two sets of collections $A$ and $B$ with respect to $A$, quantifies the amount of stories present in $A$ and $B$ (as a fraction of $A$) was calculated as $r(A, B) = \frac{|A \cap B|}{|A|}$.

Our dataset was generated without setting any parameters on the Google SERP. However, the Google SERP provides a date range parameter that attempts to restrict the documents returned on the SERP to documents published within the date range. For example, setting the date range to 2017-06-01 and 2017-06-30, attempts to restrict the documents in the SERP to those published between June
Table 3: Average story replacement rate for General and News vertical SERP collections. Column markers: \textit{minimum} and \textit{maximum}.\
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Collection & \multicolumn{2}{c|}{General SERP} & \multicolumn{2}{c|}{News vertical SERP} \\
& Daily & Weekly & Monthly & Daily & Weekly & Monthly \\
\hline
healthcare bill & 0.42 & 0.60 & 0.76 & 0.44 & 0.71 & 0.85 \\
manchester bombing & 0.37 & 0.39 & 0.59 & 0.31 & 0.34 & 0.46 \\
london terrorism & 0.34 & 0.41 & 0.60 & 0.43 & 0.66 & 0.84 \\
trump russia & 0.54 & 0.79 & 0.92 & 0.42 & 0.71 & 0.90 \\
teavel ban & 0.43 & 0.63 & 0.84 & 0.43 & 0.62 & 0.83 \\
hurricane irma & 0.21 & 0.41 & 0.61 & 0.49 & 0.77 & 0.91 \\
\hline
\end{tabular}

Table 4: Average new story rate for General and News vertical SERP collections. Column markers: \textit{minimum} and \textit{maximum}.\
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Collection & \multicolumn{2}{c|}{General SERP} & \multicolumn{2}{c|}{News vertical SERP} \\
& Daily & Weekly & Monthly & Daily & Weekly & Monthly \\
\hline
healthcare bill & 0.42 & 0.58 & 0.62 & 0.44 & 0.70 & 0.82 \\
manchester bombing & 0.27 & 0.37 & 0.46 & 0.31 & 0.52 & 0.66 \\
london terrorism & 0.38 & 0.40 & 0.51 & 0.43 & 0.65 & 0.84 \\
trump russia & 0.74 & 0.78 & 0.83 & 0.42 & 0.70 & 0.85 \\
teavel ban & 0.43 & 0.62 & 0.71 & 0.45 & 0.61 & 0.75 \\
hurricane irma & 0.21 & 0.38 & 0.51 & 0.49 & 0.76 & 0.82 \\
\hline
\end{tabular}

1, 2017 and June 30, 2017. To assess the effect of setting the date range parameter on discovering older stories that fall within a specific timeframe, we took the following steps. First, from our original dataset, we selected five collections of stories for queries about topics that occurred before June 2017: "healthcare bill," "trump russia," "travel ban," "manchester bombing," and "london terrorism." This set of five collections was called \textit{June-2017}. Second, we removed all stories from \textit{June-2017} that were not published in June 2017. Third, we issued the selected five queries to the Google SERP without setting the date range to generate five additional collections (from the first five pages). This set of five collection was called \textit{Jan-2018} (control test collection). Fourth, we issued the same five queries to the Google SERP, but this time, we set the date range to \textit{Jan-2018-Restricted-to-June}, as well as \textit{June-2017} and \textit{Jan-2018-Restricted-to-June} collections for the pairs of collections with the same query.

5 RESULTS AND DISCUSSION

Here we present the results for each of the respective measures introduced in Subsection 4.2.

5.1 Story replacement rate, new story rate, and page level new story rate

Table 3 and 4 show the average story replacement rate and new story rate, respectively over time (daily, weekly, and monthly) for both the \textit{General} and News vertical SERPs. For both \textit{General} and \textit{News} vertical SERPs, we can see that the average story replacement rate was similar to the new story rate, and both increased with time. They also show that the story replacement and new story rates are strongly dependent on the topic. For example, the Hurricane Harvey natural disaster showed a lower daily average story replacement rate (0.21) and new story rate (0.21) compared to the Trump-Russia event. This event maintained the highest daily (0.54), weekly (0.79), and monthly (0.92) average story replacement and new story rates.

Figure 2: Page-level temporal distribution of "manchester bombing" \textit{General} SERP collection showing multiple page movement patterns.
These results show us that the timing of collection building efforts undergo multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Similar to the political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Similar to the political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - from the firing of the FBI Director James Comey in May 2017 to the indictment of former Trump Campaign Chair Paul Manafort in October 2017. Unlike natural disasters which have a well-defined timeframe, this ongoing political event does not have a well-defined timeframe and has undergone multiple event cycles - Figs. 4a & c show that the average story replacement rate and average new story rate differed across various pages for the General SERP. There was a direct relationship between page number and story replacement rate (or new story rate) - the higher the page number, the higher the story replacement rate (or new story rate), and vice versa. The direct relationship may be due to fact that higher order pages (e.g., pages 4 and 5) are more likely to receive documents from lower order pages (e.g., page 1 – 3) than the opposite. For example, the probability of going from page 1 to page 5 was 0.0239 while the probability of going from page 5 to page 1 was 0.0048. The lower order pages have the highest quality on the SERP, thus, there is high competition within documents to retain their position on a lower order page (high rank). The competition in the higher order pages is less, therefore, when documents from the lower order pages lose some rank, they may fall into the higher order pages thereby increasing the new story rate of higher order pages. The News vertical SERP showed an inverse relationship between the page number and the story replacement rate (or new story rate) (Fig. 4b & d) even though the probability of going from a page 1 to page 5 (0.0801) was more likely than the opposite (0.0009). This may be due to some unseen mechanism in the News vertical SERP.

5.2 Probability of finding a story

Table 5 shows the probability of finding the same story after one day, one week, and one month (from first observation) for General and News vertical SERP collections. The probability of finding the same URI of a news story with the same query decreased with time from a page 1 to page 5 (0.0801) was more likely than the opposite. For example, the probability of going from page 1 to page 5 was 0.0239 while the probability of going from page 5 to page 1 was 0.0048. The lower order pages have the highest quality on the SERP, thus, there is high competition within documents to retain their position on a lower order page (high rank). The competition in the higher order pages is less, therefore, when documents from the lower order pages lose some rank, they may fall into the higher order pages thereby increasing the new story rate of higher order pages. The News vertical SERP showed an inverse relationship between the page number and the story replacement rate (or new story rate) (Fig. 4b & d) even though the probability of going from a page 1 to page 5 (0.0801) was more likely than the opposite (0.0009). This may be due to some unseen mechanism in the News vertical SERP.

Table 5: Probability of finding the same story after one day, one week, and one month (from first observation) for General and News vertical SERP collections. Column markers: minimum and maximum.

<table>
<thead>
<tr>
<th>Collection</th>
<th>General SERP</th>
<th>News vertical SERP</th>
</tr>
</thead>
<tbody>
<tr>
<td>healthcare ball</td>
<td>a day 0.35</td>
<td>a week 0.04</td>
</tr>
<tr>
<td>manchester bombing</td>
<td>0.54</td>
<td>a month 0.15</td>
</tr>
<tr>
<td>london terrorism</td>
<td>0.37</td>
<td>a day 0.04</td>
</tr>
<tr>
<td>trump russia</td>
<td>a day 0.39</td>
<td>a week 0.16</td>
</tr>
<tr>
<td>travel ban</td>
<td>a day 0.43</td>
<td>a week 0.06</td>
</tr>
<tr>
<td>hurricane irma</td>
<td>a day 0.34</td>
<td>a week 0.07</td>
</tr>
</tbody>
</table>

Table 6: Comparison of two collections against the June-2017 collection (documents published in June 2017). The collection Jan-2018, which was created (2018-01-11) without modifying the SERP date range parameter has a lower overlap than the collection (June-2018-Restricted-to-June) created the same day (2018-01-11) by setting the SERP date range parameter to June 2017. Even though setting the date range parameter increases finding stories with common publication dates as the date range, the recall is poor due to the fixed SERP result. Column markers: maximum.
when the same query is issued one day after it was first observed ranged from 0.34 – 0.44. When the query was issued one week after, the probability dropped to from 0.01 – 0.11, one month after 0.01 – 0.08. The probability of finding the same story with time is related to the rate of new stories: for a given time interval, the higher the rate of new stories, the lower the chance of observing the same story, because it is more likely to be replaced by another story. For example, compared to the manchester bombing collection, the hurricane irma collection produced a lower (0.34) probability (vs. manchester bombing 0.44) of finding the same story after one day due to its higher (0.79) new story rate after one day (vs. manchester bombing 0.52). The probability of observing the same news story on the News vertical SERP declined with time, but at a much faster rate compared to the General SERP. In fact, Table 5 shows that for all seven topics in the dataset, the probability of finding the same story on the News vertical when the query was re-issued one month after was marginal (approximately 0). This is partly because the News vertical SERP collections produced higher story replacement and new story rates than the General SERP collections.

In order to generalize the probability of finding an arbitrary URI as a function of time (days), we fitted a curve (Fig. 4g) over the union of occurrence of the URIs in our dataset with an exponential model. The probability \( P_{s,sp}(k) \) of finding an arbitrary URI of a news story \( s \) on a SERP \( sp \in \{\text{General, NewsVertical}\} \), after \( k \) days is predicted as follows:

\[
P_{s,\text{General}}(k) = 0.0362 + 0.9560e^{-0.9159k}
\]

\[
P_{s,\text{NewsVertical}}(k) = 0.0469 + 0.9370e^{-0.9806k}
\]

Also, similar to the story replacement and new story rates, for the General SERP, the results showed a direct relationship with the page number and probability of finding news stories over time (Fig. 4e). For the General SERP, higher order page numbers (e.g., 4 and 5) produced higher probabilities of finding the same stories compared to lower order (e.g., 1 and 2) pages. This might be because during the lifetime of a story, the probability of the story going...
from a lower order (high rank) page to a higher (low rank) order page is higher than the opposite - going from higher order page to lower order page (climbing in rank). For example, the probability of going from page 1 to page 5 was higher (0.0239) than the probability of going from page 5 to page 1 (0.0048). However, collections from News verticals showed that the lower the page number, the higher the probability of finding news stories (inverse relationship) even though the probability of falling in rank (lower order page to higher order page) is higher than the probability of climbing in rank (higher order page to lower order page).

5.3 Distribution of stories over time across pages

Fig. 3 shows how the temporal distributions differs typically between General and News vertical SERP collections. There are two dimensions in the figure: days (x-axis) and URIs of stories (y-axis). A single dot in the figure indicates that a specific story occurred at that point. The temporal distribution is a reflection of the new story rate, but at a granular (individual) story level. General SERP collections had lower new story rates, thus produced stories with a longer lifespan than News vertical SERP collections. In Fig. 3, this is represented by a long trail of dots. Since News vertical collections had higher story replacement and new story rates, they produced documents with shorter lifespans. For example, Fig. 3a contrasts the denser (longer lifespan) temporal distribution of the “hurricane harvey” General SERP collection to the sparser “trump russia” General SERP collection (Fig. 3c). The “trump russia” collection produced new documents on average at a rate of 0.54 (daily) to 0.83 (monthly), compared to the “hurricane harvey” collection (daily - 0.21, and monthly - 0.51). Similarly, since documents from the “trump russia” collections were rapidly replaced (story replacement rate: 0.54 – 0.92) with newer documents, they mostly did not persist on the SERP.

Fig. 2 shows how URIs moved across pages over time. The rows represent the URIs and the columns represent the pages in which the URIs were observed on a specific day. A single cell represents the page in which a URI occurred on a specific day. For example, the first cell (row 0, column 0) of Fig. 2 is 1. This means the URI at row 0 was first observed on page 1. Some of the same URIs persist over time within the same page. For example Fig. 2a, row 0, shows that the highly ranked Wikipedia page\(^2\) of the Manchester bombing event was seen for 24 consecutive days on the first page of the SERP, was not seen (within page 1-5) on the 25th day, and then seen for 13 consecutive days (still on page 1). Fig. 2a also shows the increase/decrease in ranks for stories. For example, in Fig. 2a, row 4, the URI\(^3\) was first observed on page 5, the next day it increased in rank to page 1, skipping (2-4). The page-level temporal distribution also shows that some stories go directly from page 5 to 1. In contrast with General SERP collections, the temporal distribution of News vertical collections is shorter (Fig. 2b) and reflect the higher story replacement and new story rates of News vertical collections.

5.4 Overlap and recall

Table 6 shows that setting the Google date range parameter improves finding stories with respect to the set date range for both General and News vertical collection. For example, for the “healthcare bill” General SERP collection, the Jan-2018 collection which was created (2018-01-11) by making a default search (without) setting the date range had an overlap rate of 0.06 with respect to the collection of documents created in June 2017 (June-2017). In contrast, the collection created the same day (2018-01-11) by setting the date range parameter to June 2017 (2017-06-01 to 2017-06-30) had a much higher overlap rate of 0.60. This is the case across all collection topics, especially for topics with lower new story rates (0.27 - 0.46) such as “manchester bombing” (0.82 overlap rate). The News vertical collections had lower overlap rates compared to the General SERP collections since News vertical collection have higher story replacement and new story rates.

Irrespective of the increase in finding (overlap) new stories that occurs when the date range parameter is set, the recall is poor. Since the SERP only produces a fixed number of documents per page, we only get a small fraction of the documents relevant to the specified date range. The “healthcare bill” June-2017 General SERP collection contains 460 documents published in June 2017, collected by extracting URIs from the first five pages of the SERP. A query (“healthcare bill”) issued to the SERP in January 2018, with the date range parameter set to June 2017 increased overlap (refinding stories), but did not increase the number of results - we could only extract at most approximately 50 URIs (first five page). Consequently, across all topics in Table 6, both Jan-2018 and Jan-2018 Restricted-to-June collections had recall of under 0.10 except for the “london terrorism” topic (maximum recall 0.20). This reaffirms the that idea that collection building or seed selection processes that rely on the SERP must start early and persist in order to maximize recall. To further aid selection of seeds, a simple set of heuristics could identify most of the likely stable URIs (e.g., wikipedia.org, nasa.gov, whitehouse.gov) as well as URIs likely to quickly disappear from the top-k SERPs (e.g., cnn.com or nytimes.com, followed by a long path in the URI). The archivist could give priority to the latter URIs, knowing that the former URIs will continue to be discoverable via Google.

6 FUTURE WORK AND CONCLUSIONS

Our findings motivate the need for instant and persistent collection building. The next stage of this research is the implementation of a collection building system that extracts URIs from SERPs. The system may be triggered semi-automatically by a curator for an important event or automatically by an event detection system. An event detection system could listen to traffic in news and social media for clusters of interest, identify the events of interest, and initiate a collection building process from SERPs. In addition to the implementation of such a collection building system, it is important to investigate the kinds of collections, topics or events most suitable for SERPs. For example, this research focused on news collections, but further research is required to assess the effectiveness of using SERPs for other kinds of collections.

Collection building offers a way of preserving the historic record of important events. This involves collecting URIs of web pages

\(^2\)https://en.wikipedia.org/wiki/Manchester_Arena_bombing

\(^3\)http://www.dailymail.co.uk/news/article-4578566/Evidence-Nissan-linked-Manchester-bombing.html
that are relevant to a predefined set of topics. It is crucial for collections to capture all the stages (oldest to newest) of events and not only the recent stories. Search engines provide an opportunity to build collections or extract seeds, but tend to provide the most recent documents. As a first step toward a larger effort of generating collections from SERPs, we sought to gain insight on the dynamics of refining stories on SERPs. Our findings illustrate the difficulty in refining news stories as time progresses: on average, the daily rate at which stories were replaced on the Google General SERP ranged from 0.21 – 0.54, weekly – 0.39 – 0.79, and monthly - 0.59 – 0.92. The Google News vertical SERP showed even higher story replacement rates, with a daily range of 0.31 – 0.57, weekly - 0.54 – 0.82, and monthly - 0.76 – 0.92. We also showed that the probability of finding the same news story diminishes with time and is query dependent. The probability of finding the same news story with the same query again, one day after the first time the story was first seen ranged from 0.34 – 0.44. If one waited a week, or a month and issued the same query again, the probability of finding the same news story drops to 0.01 – 0.11. The probability declines even further if we used the News vertical SERP due to its higher story replacement and new story rates. The probability for finding the URI of a news story was further generalized through our provision of two predictive models that estimate this probability as a function of time (days). Discoverability may be improved by instructing the search engine to return documents published within a temporal range, but this information is not readily available for many events, and we discover only a small fraction of relevant documents since the count of search results are restricted. These findings collectively express the difficulty in refining news stories with time, thus motivates the need for collection building processes that utilize the SERP to begin early and persist in order to capture the start and evolution of an event. Our research dataset comprising of 151,602 (33,432 unique) links extracted from the Google SERPs for over seven months, as well as the source code for the application utilized to semi-automatically generate the collections are publicly available [3].

ACKNOWLEDGEMENTS
This work was made possible by IMLS-LG-71-15-0077-15, and help from Christie Moffat at the National Library of Medicine.

REFERENCES