How to Assess the Exhaustiveness of Longitudinal Web Archives: A Case Study of the German Academic Web

Michael Paris
Humboldt-Universität zu Berlin
Berlin, Germany
michael.paris@hu-berlin.de

Robert Jäschke
Humboldt-Universität zu Berlin
Berlin, Germany
robert.jaeschke@hu-berlin.de

ABSTRACT

Longitudinal web archives can be a foundation for investigating structural and content-based research questions. One prerequisite is that they contain a faithful representation of the relevant subset of the web. Therefore, an assessment of the authority of a given dataset with respect to a research question should precede the actual investigation. Next to proper creation and curation, this requires measures for estimating the potential of a longitudinal web archive to yield information about the central objects the research question aims to investigate. In particular, content-based research questions often lack the ab-initio confidence about the integrity of the data. In this paper we focus on one specifically important aspect, namely the *exhaustiveness* of the dataset with respect to the central objects. Therefore, we investigate the recall coverage of researcher names in a longitudinal academic web crawl over a seven year period and the influence of our crawl method on the dataset integrity. Additionally, we propose a method to estimate the amount of missing information as a means to describe the exhaustiveness of the crawl and motivate a use case for the presented corpus.

CCS CONCEPTS

• **Information systems → Data extraction and integration**; **Digital libraries and archives**; • **Applied computing → Document capture**.

KEYWORDS

dataset; longitudinal; web archive; focused web crawl; exhaustive

ACM Reference Format:
Michael Paris and Robert Jäschke. 2020. How to Assess the Exhaustiveness of Longitudinal Web Archives: A Case Study of the German Academic Web. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media (HT '20), July 13–15, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372923.3404836

1 INTRODUCTION

The web has been extensively used as a resource to motivate new research questions from various disciplines for more than 20 years [6, 7, 9, 13, 14, 17]. The foundation of such research is a suitable corpus of web pages, which is typically harvested by a web crawler traversing specified regions of the web [3, 8]. Different scopes can be addressed with focused crawling [7] and repeated crawling can support the analysis of temporal phenomena [20, 22]. Typically, two types of static datasets are used to analyse the web: On the one hand, datasets can be created by *(focused) crawling* to answer a *specific research question* (RQ). In these cases the crawl policy is implemented by a set of very specific rules and the amount of data that is harvested scales with the number of rules, producing a narrow snapshot of the web. Such an approach is useful when the RQ is precisely known a priori. An important underlying implicit assumption is that the traversed part of the web does not change during the crawling process. This process can be regarded as a conversion of a set of unstructured data on the web into some structured data [1], whereby the location of the data is already known and the data just needs to be harvested. On the other hand, web crawlers are used to *archive web pages* using heuristic rules which are implementing the crawl policy without a specific RQ in mind. Instead, web archives aim to preserve (parts of) the web and implicitly account for a *broad range of potential RQ*. Through repeated crawling with the same rules *longitudinal* web archives can be established. Carefully conducted web archiving can result in a coherent, traceable and complete image of a part of the web. The advantage of web archives over a focused crawl is that a larger variety of related RQs can be investigated with the same dataset. Nevertheless, using a web archive to approach a specific RQ poses crucial challenges: First, the scope of the web archive's crawl must fit to the RQ. Second, the scope of the web archive's crawl should exhaustively cover the data that is available at the time of creation. Therefore, two central questions need to be answered prior to reliably performing analyses using longitudinal web archive data: (1) How can we relate the termination time of the crawl to its exhaustiveness? (2) How can we quantify the exhaustiveness of a web archive relative to a given RQ?

The crawl frontier, that is, discovered but not yet crawled URLs in the queue of the web crawler, can also be useful for gaining insights into the exhaustiveness of a web crawl [12]. However, as it consists of URLs, only statements about entities like domains or hosts can be made, not about entities that require/are contained in the content of the web pages. In addition, for a web archive the crawl frontier is not readily available and can not be used to gain insight about the web archive's exhaustiveness.

The recall of crawled entities is typically used to investigate the exhaustiveness of a crawl towards a topic as a function of the crawl policy [10]. In small/focused crawls the required ground truth – the set of all entities falling under the crawl policy – can be the crawl itself, if it has come to a halt after harvesting all available web pages. In contrast, if the crawl has not cleared the crawl frontier, the entities it harvested can only be compared to a previously established ground truth. This can be an exhaustive crawl created

HT '20, July 13–15, 2020, Virtual Event, USA

specifically for the purpose of investigating the exhaustiveness of the crawls with a non-empty crawl frontier. Though this approach is applicable to smaller collections for which the ground truth has been established, it does not adequately translate to a large web archive. Establishing a ground truth for a large web archive would imply extending it with the content that could not by harvested due to time constraints. The existence of such a larger web archive would render the investigated web archive obsolete.

In this paper, we evaluate the scope of a longitudinal web archive. Specifically, we propose and evaluate a heuristic to analyse the exhaustiveness of web archives with respect to entities connected to a class of research questions. For that, we utilize a topical proxy to estimate the relative gain of new relevant entities after a specific crawl time. This approach enables the use of web archives with an increased confidence with respect to a class of research questions. It can also be used as an entry point for similar longitudinal web archives and application scopes.

Our work is organized as follows: In Section 2 we discuss related work and in Section 3 we introduce our datasets and methods. The results are presented in Section 4 and discussed in Section 5.

2 RELATED WORK

Web archiving aims to preserve the information of the live web by means of crawling prioritized resources, thus offering a high quality specialized web archive. Various aspects of quality are subject of extensive research. Cothey [10] investigated the reliability of crawled data to support informetric studies in terms of a changing crawl policy. Cho & Garcia-Molina [8] study the quality of a longitudinal web archive in terms of the freshness of the web pages, thus improving the temporal representation of the harvested web pages. A retrospectively created web archive, such as by Ben David [4] has another aspect to quality. Here the quality of the web data is given by the recall of a set of URLs. The completeness and temporal coherence of the representation of archived the web page is addressed by Ainsworth et al. [2]. Bordino et al.[5] studied the quality of their longitudinal web archive of the .uk-domain in terms of the similarity of results in [18]. Both investigated the temporal change of web pages and links. The aspect of quality expressed here is concerned with the faithfulness of the web archive's representation relative to the .uk-domain. Similar longitudinal web snapshots of university link data have been used to investigate the New Zealand, Australia and United Kingdom university research productivity [20]. Spaniol et al. [21] propose a model for ensuring the temporal coherence of a collection of web data, thereby raising the issue of interpretability of results derived from the underlying data. In a continued effort Weikum et al. [22] constitute the challenges of working with longitudinal web data and mention some use cases for such a web archive. Denev et al. introduce the SHARC-framework [11] for a systematic approach to data quality in web archiving. The framework focuses on the quality measures of blur and coherence of the web snapshot, which are both aspects concerned with the time dependent changes of a longitudinal web archive. These measures are ideal in cases when the snapshot is taken frequently and a high coherence is present throughout a window of time. Frequently taken snapshots become quickly unfeasible when the number of web pages is large and therefore a larger longitudinal web archive tends to be temporally sparse. Such a sparse web archive requires a weaker quality measure to legitimize its temporal integrity. That is where we aim to contribute, by quantifying the exhaustiveness, of potentially undiscovered entities over a time span, of the underlying crawl of the longitudinal web archive with respect to structural and content related entities.

3 METHOD

In this paper, we track mentions of entities that are relevant for a set of research questions to assess the *crawl heuristics* and the *exhaustiveness* of a set of web crawls. Specifically, we use names of researchers as a proxy to investigate these aspects on an exemplary longitudinal web archive of academic web pages. We assess the comprehensiveness of our web archive by comparing it against a set of researchers who were successful in acquiring research funding from the *DFG* (German Research Foundation). First, we describe how the web archive was created and prepared for analyses (Sections 3.1 and 3.2). Then, we introduce the thematically related dataset of researcher names (Section 3.3). Finally, we explain how we use the recall rate to estimate the crawl's exhaustiveness (Section 3.4).

3.1 A Dataset of Academic Web Pages – "German Academic Web"

We created the following dataset in order to establish a knowledge base on the "German Academic Web" (GAW).[1] Since 2012, semi-annual focused crawls of the web pages of universities and research institutes in Germany have been performed using Heritrix, the open source web crawler of the Internet Archive [16]. It traverses the web, starting from a list of given seeds, follows newly discovered hyperlinks and stores seen content in the standardised WARC file format [15]. Each crawl began with a seed list of, on average, 150 domains of all German academic institutions with the right to award doctorates.[2] The crawler follows a breadth-first policy on each host, thereby collecting all available pages reachable by links from the homepage. The scope was limited to crawl only pages from the seed domains and certain file types (mainly audio, video, and compressed files) were excluded using regular expressions. Along the crawl, the URL queues were monitored via a web UI. Hosts that appeared to be undesirable, such as e-learning systems or repositories, were 'retired', that is, their URLs no longer crawled. However, previously harvested URLs from retired hosts were not removed. Most crawls were finished (manually) after roughly 100 million pages were collected (according to Heritrix' control console), which took roughly two weeks per crawl, on average. Up to now, 15 crawls have been collected and stored in the WARC file format. Each WARC file contains several WARC records. For each fetched page ('capture'), the HTTP request, the HTTP response and the extracted links are stored as individual WARC records. This collection is a long-term longitudinal focused crawl that is characterised by some aspects that originate from its long-term creation:

- Software updates[3] can influence crawls, for instance, when link extraction from JavaScript is improved.

[1] https://german-academic-web.de/

[2] The seed list is extracted from the current entries on https://de.wikipedia.org/wiki/Liste_der_Hochschulen_in_Deutschland.

[3] First crawls were performed with Heritrix 3.0.0, latest crawls with Heritrix 3.2.0.

- Crawl operators learn over time how to use the crawler and constantly improve its configuration, for instance, by adjusting the crawl delay or deactivating some link extractors to avoid collection of content that is out-of-scope.
- The crawl scope is restricted to the seed domains, although in its default configuration a fractional amount of speculative URLs were collected as well.
- Crawl operators make errors, for instance, forgetting to add seed hosts that were not collected automatically.
- Given an informal scope ("universities in Germany"), the set of seeds can change over time, for instance, due to new universities appearing or existing ones changing their domain name.
- Missing documentation on the used crawl configuration or seed list can make it difficult to pre-process crawls or interpret results.
- Crawling occasionally can cause problems on the server side resulting in special rules added to the configuration to protect certain hosts or domains, for example, by increasing the crawl delay or excluding certain URL patterns.
- An ad-hoc stopping criterion and manual operation of the crawl can result in different crawl lengths and numbers of captures.
- During each crawl non-relevant hosts (like management systems, code repositories, etc.) were identified and blocked, resulting in a growing list of *retired* hosts that are excluded from subsequent crawls.

These aspects and the problems they can cause when using the data for longitudinal analysis can be considered exemplary for such a type of collection.

3.2 Processing of the GAW data

Since the crawl policy changes during and between crawls, a normalisation step needs to be applied to account for these inconsistencies, such as removal of web pages from retired hosts that were collected before the hosts were retired. This is done by reducing the data to heuristically relevant regions. We include the last 13 crawls (since 2013) in our analysis. In all subsequent analyses we only consider WARC records that were retrieved with an *HTTP status code 200*, that are of *MIME type text/html*, and whose compressed *size is below 10 MB*.

The requirements of longitudinal research questions towards the GAW data will generally include a notion of *exhaustiveness* throughout the crawls. Therefore, the following heuristically motivated processing steps aim to derive a suitable subset from the data. This subset should demonstrate that the crawl was performed exhaustively. Thus, for each crawl we create a *processed* subset as follows: (1) We only include WARC records from domains that were seeds in *all* crawls. (2) We remove all WARC records from hosts that were *retired* during the crawl

In the following sections we abbreviate the *unprocessed* dataset with $\mathcal{U}$, and the *processed* subset with $\mathcal{P}$.

3.3 A Dataset of Researcher Names – GEPRIS

Since we are interested in quantifying the extent to which the GAW data contains information about *researchers*, we create a dataset of names of researchers who were successful in acquiring research funding. Therefore, we extracted from the GEPRIS[4] database of *projects funded by the DFG* information for projects of type "Sachbeihilfe". This type mainly comprises basic research projects of individual researchers (or small groups). We only considered projects that started in or after 2012 and where at least one of the associated researchers is affiliated with an institution whose web presence is contained in all crawls. Finally, we extract the names of all researchers associated with these projects. The resulting list of researcher *names* does not include academic titles. Furthermore, common names referring to different persons are excluded to reduce the amount of potential ambiguity, which yields a total of 10 433 entities. As a benchmark, we additionally created a list of *random names* from the list of *names* by shuffling the last names, since we can assume that among the names we selected the first and last names are independent. For the rest of our contribution, we will refer to these two lists as *names* and *random names*.

3.4 Evaluation Approach

Discovery of entities and construction of time series. For each person name we want to find the earliest web page of each crawl that contains the name. 'Earliest' in that case refers to crawl time t_c. Person names are discovered in the pages that were extracted from the WARC records by searching for their string representation using exact string matching. All host and person names are associated with a unique timestamp. For each name, the crawl time of the web page that contains it and was crawled earliest in the corresponding crawl is used.

Quantifying potentially undiscovered entities. While the number of hosts that were *not* crawled is unknown to us for both $\mathcal{U}$ and $\mathcal{P}$, we are able to measure which fraction of *names* from GEPRIS were not discovered. As we are interested in the recall of entities regarding an unknown ground truth, we assume that exhaustiveness is characterized by a decline in the number of discovered entities over time and can be modeled as a random process for the final stage of a crawl. Therefore, we investigate how well sampling from the observed occurrence counts estimates the behavior for a crawl time t_c larger than the end of the chosen observation interval t_e. The end of the observation interval is not necessarily the end of the crawl. We perform this comparison of discovery of entities by modeling the potentially gained number of entities by a random process X_t. As such, the outcome of measuring X_t reflects the number of observed entities over the course of an hour. The random variable X at time t is sampled k-times. The underlying distribution of X_t is given by the observed entity counts between t and t_e, that is, the end of the observation interval. The sum of X_t over k measurements corresponds to the total gain of entities after k hours. We choose $k = 100$ and $t_e = 200$ hours and compare the sampled entity counts with the crawled entity counts for $t_e \leq t_c = t + k \leq t_e + k$. The expectation is that the relative gain of crawled entities decreases with time and the counts sampled from the last part of the observation interval can be related to the crawled entities after t_e. We then plot the gain of entities relative to the number of entities in $\mathcal{U}$ at t_e to display how much would be additionally gained at t_c had we kept

[4] https://gepris.dfg.de/

Table 1: Overview on the two datasets, each based on 13 crawls between December 2013 and December 2019.

	$\mathcal{U}$				$\mathcal{P}$			
# Seeds	149.6	±	1.7		128			
# Hosts	35.8	k ±	5.1	k	21.6	k ±	0.5	k
# Pages	67.2	M ±	4.3	M	52.6	M ±	5.9	M
# Links	3.9	G ±	0.2	G	3.1	G ±	0.4	G
Duration	340	h ±	222	h	340	h ±	222	h

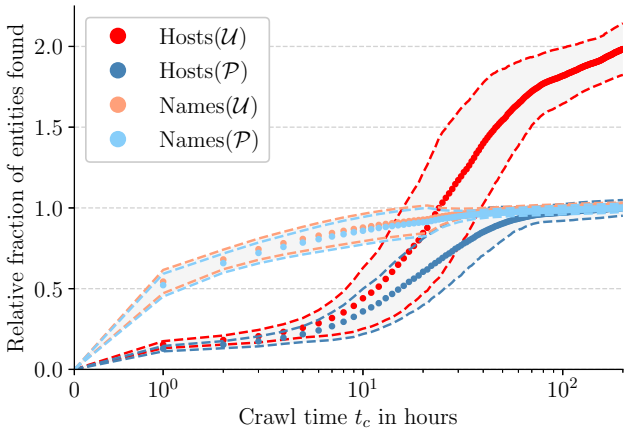

Figure 1: Cumulative occurrence of *hosts* and *names* normalized by the average entity count at t_e over all crawls in $\mathcal{P}$.

crawling. In summary, we leverage freely available structured data (*researcher names*) about our topic of interest (*successful researchers*) to infer the scope of exhaustiveness of the created web archive by quantifying the recall rate in the late stages of the crawl process.

4 RESULTS

We first have a look at the basic properties of our datasets in terms of the means and standard deviations presented in Table 1.

While the row # Pages indicates the number of web pages that have been *crawled*, the row # Links indicates the number of links that have been *extracted* from those crawled pages. We can see that of the, on average, 150 seed domains, only 128 remain in $\mathcal{P}$. Analogously, the mean number of hosts, pages, and links is reduced.

Figure 1 presents the mean crawl progression of *hosts* and *names*, where the gray shaded areas represent the standard deviation over the 13 crawls. The plot allows us to observe the difference in the fraction of entities (*hosts* and *names*) collected over time between $\mathcal{P}$ and $\mathcal{U}$. We see that the processing in $\mathcal{P}$, which is a simple transformation of the data, changes the amount of discovered *hosts* significantly. However, the amount of discovered *researcher names* is practically the same in $\mathcal{P}$ and $\mathcal{U}$ over the whole crawl duration. The ratio between the number of hosts in $\mathcal{U}$ and $\mathcal{P}$ is approximately 2. For the *hosts* progressions we can also see that the standard deviation decreases towards the end of the observation interval t_e and that the slope of *Hosts*($\mathcal{P}$) tends to zero in contrast to *Hosts*($\mathcal{U}$). We note that 50% of all *names* are found within the first hour of crawling in both $\mathcal{U}$ and $\mathcal{P}$, whereas it required more than 10 hours to collect 50% of the *Hosts*.

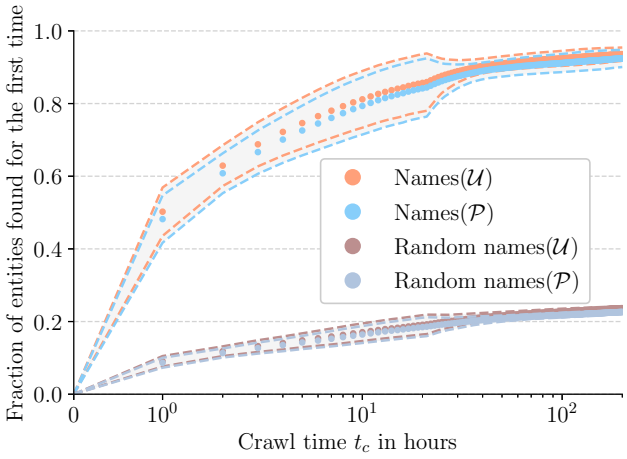

Figure 2: Progression of name collection in all crawls.

The progression of collecting *names* is juxtaposed to the progression of *random names* and normalized against the total number of 10 433 known names in Figure 2. Among all *names* from GEPRIS, 10 174 were found, whereas only 2941 *random names* were found in the longitudinal web archive. At the end of the crawls, on average 95% of all *names* were discovered whereas less than 25% of the *random names* were found. A similar behavior can be observed at the beginning of the observation interval: After the first crawl hour, on average 10% of the *random names* were discovered, while 50% of the *names* were found (as already observed in Figure 1). We also note that the difference between the number of entities found in $\mathcal{U}$ and $\mathcal{P}$ is at no point larger than 2% of the total number of entities.

The results of our sampling experiment are shown in Figure 3. For all three types of entities we can observe that the sampled gain is larger than the measured gain and that both display a downwards trend. We further see that there is no significant difference between $\mathcal{U}$ and $\mathcal{P}$ for *names* and *random names*. At the same time, *names* tends to gain 1% over k hours, whereas *random names* gains on average an additional 5% of entities of the total. The *hosts* show a different behavior: The ratio between gains in $\mathcal{U}$ and $\mathcal{P}$ for the projected and measured series differs on average by a factor of 7.07 and 7.42, respectively. The projections suggest that the average gain of *hosts* increases the average number of hosts in $\mathcal{U}$ and $\mathcal{P}$ by 7% and 1%, respectively.

5 DISCUSSION

In this paper we asked how exhaustiveness of a large web archive can be quantified. We presented a case study for the "German Academic Web" and expressed this in terms of the gain of a specifically relevant set of entities (researcher names) relative to the total gain by the end of the observation interval t_e. As we see in Figure 3, the estimated gain from sampling the last observed entity counts is greater than the respective measured gain. This implies that the expectation value of the random process $\mathbb{E}[X_t]$ shifts to smaller values with increasing time and that the sampled progressions seem to provide a good estimate of missing entities. We combine this with the behavior of the progression of *names* shown in Figure 2, for which we can state an average recall of 95%. This expresses

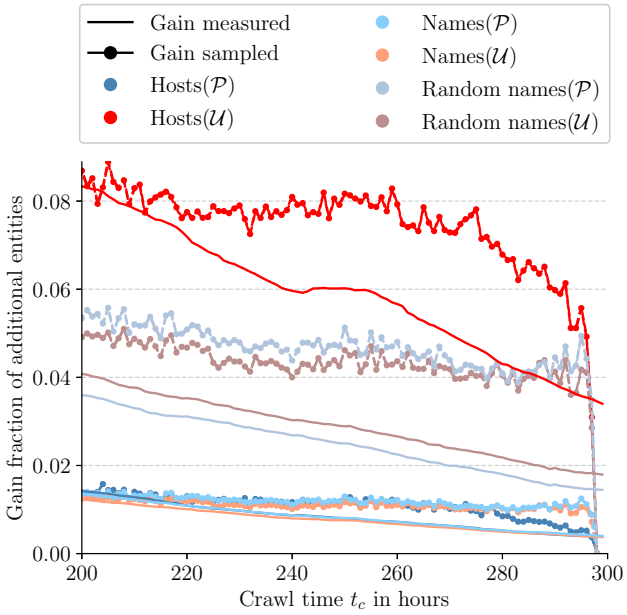

Figure 3: Additionally gained entities over 100 hours. Gain relative to total entity count at t_e.

how exhaustive the web crawler has been up to t_e. Such a reference point allows us to state an expected upper bound of missing entities over a time span of 100 h. The missing entities amount to 1% of the total number of entities at time t_e.

The heuristic processing of $\mathcal{U}$ lead to a similar relative gain of *hosts* and *names* in $\mathcal{P}$ for both measured and sampled progressions. Therefore, the underlying random process for these gains is the same up to a scaling factor. In Figure 1, we see that using our heuristic rule, we are able to decrease the gain of hosts in $\mathcal{U}$ from 7% to 1% in $\mathcal{P}$. On the one hand, we boost our confidence about the exhaustiveness of each of the snapshots of our longitudinal web archive. On the other hand, we reduced the standard deviation of *hosts* by a factor of approximately 10, which suggests a better comparability among the snapshots. What we mean by comparability is that the change observed by the crawler is the change of the web and not an artifact of the crawl policy. This implies that posing RQs related to the entities at hand yields more faithful observations on $\mathcal{P}$ than on $\mathcal{U}$. The *random names* are given as a benchmark to approximate the behavior of random entities being found in the web archive. A gain below *random names* in Figure 3 can be interpreted as relatively exhaustive, in the previously mentioned sense, whereas a gain above *random names* can be taken as an indicator that a faithful representation of the web has not yet been achieved. In the future, scientometric research about the behavior of researchers in the academic ecosystem can be conducted with the presented web archive. This research includes migration patterns of researchers, entity linking and the evolution of academic collaborations, all of which necessitate an understanding of the underlying completeness of the data for the interpretability of the generated results. In general, our approach can be applied to longitudinal web archives by leveraging ground truth data to determine the utility of a given web archive for a research question.

Parts of the dataset metadata (namely URLs and timestamps) are available online [19].

ACKNOWLEDGMENTS

Parts of this research were funded by the German Federal Ministry of Education and Research (BMBF) in the REGIO project (grant no. 01PU17012D).

REFERENCES

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. 2009. The web changes everything. In *Proc. WSDM*. ACM. https://doi.org/10.1145/1498759.1498837
[2] S. G. Ainsworth, M. L. Nelson, and H. Van de Sompel. 2015. Only One Out of Five Archived Web Pages Existed as Presented. In *Proc. Hypertext*. ACM. https://doi.org/10.1145/2700171.2791044
[3] R. A. Baeza-Yates and C. Castillo. 2004. Crawling the Infinite Web: Five Levels Are Enough. In *Proc. Int. Workshop on Algorithms and Models for the Web-Graph*. 156–167. https://doi.org/10.1007/978-3-540-30216-2_13
[4] A. Ben-David. 2019. 2014 not found: a cross-platform approach to retrospective web archiving. *Internet Histories* 3, 3-4 (Aug. 2019), 316–342. https://doi.org/10.1080/24701475.2019.1654290
[5] I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. 2008. Temporal evolution of the UK web. *Proc. Int. Conf. Data Mining / ICDM Workshops* (2008), 909–918. https://doi.org/10.1109/ICDMW.2008.88
[6] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. 1999. Mining the Web's link structure. *Computer* 32, 8 (Aug. 1999), 60–67. https://doi.org/10.1109/2.781636
[7] S. Chakrabarti, M. van den Berg, and B. Dom. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks* 31, 11 (1999), 1623 – 1640. https://doi.org/10.1016/S1389-1286(99)00052-3
[8] J. Cho and H. Garcia-Molina. 2000. The Evolution of the Web and Implications for an Incremental Crawler. In *Proc. VLDB*. 200–209.
[9] S.-C. Chu, L. C. Leung, Y. V. Hui, and W. Cheung. 2007. Evolution of e-commerce Web sites: A conceptual framework and a longitudinal study. *Information & Management* 44, 2 (2007), 154 – 164. https://doi.org/10.1016/j.im.2006.11.003
[10] V. Cothey. 2004. Web-crawling reliability. *J. Assoc. Inf. Sci. Technol.* 55 (2004), 1228–1238.
[11] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. 2011. The SHARC Framework for Data Quality in Web Archiving. *The VLDB Journal* 20, 2 (April 2011), 183–207. https://doi.org/10.1007/s00778-011-0219-9
[12] M. Ester, M. Groß, and H.-P. Kriegel. 2001. Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies. In *Proceedings of 27th International Conference on Very Large Data Bases*. 321–329.
[13] J. Feise. 2001. An Approach to Persistence of Web Resources. In *Proc. Hypertext*. ACM, 215–216. https://doi.org/10.1145/504216.504267
[14] G. Grefenstette and L. Muchemi. 2016. Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler. *CoRR* abs/1605.09564 (2016). arXiv:1605.09564
[15] International Internet Preservation Consortium (IIPC). [n.d.]. The WARC Format 1.1. https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/#warc-file-name-size-and-compression. Online; Last accessed 1 Apr 2020.
[16] Internet Archive. [n.d.]. Heritrix – The Internet Archive's open-source, extensible, web-scale, archival-quality web crawler project. https://github.com/internetarchive/heritrix3. [Online; Last accessed 1 Apr 2020].
[17] D. Liben-Nowell and J. Kleinberg. 2003. The link prediction problem for social networks. In *Pro. Int. Conf. Information and Knowledge Management – CIKM*. ACM Press. https://doi.org/10.1145/956863.956972
[18] A. Ntoulas, J. Cho, and C. Olston. 2004. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In *Proc. Int. Conf. World Wide Web*. ACM, 1–12. https://doi.org/10.1145/988672.988674
[19] Michael Paris and Robert Jäschke. 2020. *Summary GAW*. https://doi.org/10.5281/zenodo.3843507
[20] N. Payne and M. Thelwall. 2008. Longitudinal trends in academic web links. *Journal of Information Science* 34, 1 (2008), 3–14.
[21] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. 2009. Data Quality in Web Archiving. In *Proc. 3rd Workshop on Information Credibility on the Web (WICOW '09)*. ACM, 19–26. https://doi.org/10.1145/1526993.1526999
[22] G. Weikum, N. Ntarmos, M. Spaniol, P. Triantafillou, A. Benczúr, S. Kirkpatrick, P. Rigaux, and M. Williamson. 2011. Longitudinal Analytics on Web Archive Data: It's About Time!. In *Pro. Conf. Innovative Data Systems Research*. 199–202.