A Holistic View on Web Archives

Helge Holzmann and Wolfgang Nejdl

Abstract In order to address the requirements of different user groups and use cases of web archives, we have identified three views to access and explore web archives: user-, data- and graph-centric. The user-centric view is the natural way to look at the archived pages in a browser, just like the live web is consumed. By zooming out from there and looking at whole collections in a web archive, data processing methods can enable analysis at scale. In this data-centric view, the web and its dynamics as well as the contents of archived pages can be looked at from two angles: 1. by retrospectively analysing crawl metadata with respect to the size, age and growth of the web, 2. by processing archival collections to build research corpora from web archives. Finally, the third perspective is what we call the graph-centric view, which considers websites, pages or extracted facts as nodes in a graph. Links among pages or the extracted information are represented by edges in the graph. This structural perspective conveys an overview of the holdings and connections among contained resources and information. Only all three views together provide the holistic view that is required to effectively work with web archives.

1 Introduction

By offering unique potential for studying past events and temporal evolution, web archives provide new opportunities for various kinds of historical analysis (Schreibman et al, 2008), cultural analysis and Culturomics (Michel et al, 2010), as well as analytics for computational journalism (Cohen et al, 2011).

Helge Holzmann
Consultant e-mail: consulting@helgeholzmann.de

Wolfgang Nejdl
L3S Research Center, Leibniz Universität Hannover, Germany e-mail: nejdl@L3S.de

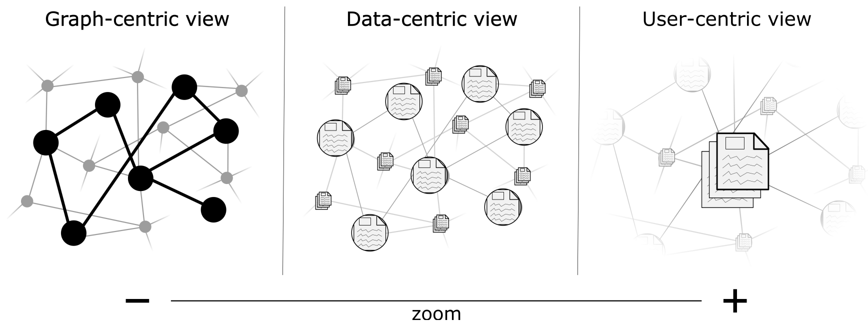

Fig. 1: Three views on web archives, representing different levels of magnification for looking at the archived data.

Consequently, with the growing availability of these collections and the increasing recognition of their importance, an ever larger number of historians, social and political scientists, and researchers from other disciplines see them as rich resources for their work (Hockx-Yu, 2014).

However, as web archives grow in scope and size, they also present unique challenges in terms of usage, access and analysis that require novel, effective and efficient methods and tools for researchers as well as for the average user (Holzmann, 2019). In this chapter, we tackle these challenges from three different perspectives: the *user-centric view*, the *data-centric view* and the *graph-centric view*. One natural way of conceiving these views is as different levels of magnification for looking at the same archival collection, as illustrated in Fig. 1, starting with the user-centric view that targets single documents to be examined by regular users. By zooming out to the data-centric view, one can scale the examination up to the whole archival collection or subsets of it. In contrast, the broadest zoom level, the graph-centric view, does not focus on the individual documents but deals with the underlying structures that span an archive as graphs. These are the foundational models for most downstream applications, such as search and deep data analysis, as well as models to guide users in user-centric views.

Another way of conceiving the relationships among the views is by considering their levels of abstraction. While the data-centric view is rather low level, closest to the data as well as to computational resources, both the graph- and user-centric views may be considered as more abstract. The graph-centric view is a conceptual layer, dealing with underlying conceptual models, facts and information contained in the archive and the relationships among them. The user-centric view, on the other hand, focuses on the users who interact with the archive without any particular technical or data science skills required. Both views attempt to hide the underlying data access and processing complexities from the targeted user group. This understanding leads to another distinguishing factor of the three views, namely, the types of challenges they cause. While we care about usability as well as exploration in the user-centric view, technical and fundamental questions are raised to a much larger extent by both other views. Finally, however, all three views are connected in one form or the other and there exist synergies among them in all the different ways of conceiving the perspectives outlined here, as we shall see later.

According to the above, the following user roles may be assigned to the different views, with which they typically interact:

- User-centric view: web users, close-reading researchers, managers, etc.
- Data-centric view: data engineers, computer scientists, web archive experts, etc.
- Graph-centric view: data scientists, distant-reading/digital researchers, analysts, software (clients, agents, APIs, downstream applications), etc.

2 User-centric View: Browsing the Web of the past

The natural way to look at a web archive is through a web browser, just as regular users explore the live Web. This is what we consider the **user-centric view**: access with a focus on users and their needs, without requiring additional infrastructure or knowledge about the underlying data structures.

The most common way to access a web archive from a user's perspective is the *Wayback Machine*[1], the Internet Archive's replay tool to render archived webpages, as well as its open-source counterpart *OpenWayback*[2], which is available for many web archives.

These tools are made for users who want to look up an old webpage as if it is still online, as well as scholarly users who *closely* read individual webpages to understand their content and the context rather than or prior to zooming out and analysing collections in a *data analysis* or *distant reading* fashion (Moretti, 2005). Similar to the live web, where users either directly enter the URL of a webpage in a browser, click a link or utilise search engines to find the desired page, the use of web archives from a user's perspective can be distinguished as **direct access** and *search*:

2.1 Direct access

Direct access to an archived webpage through the Wayback Machine requires the user to enter a target URL first, before selecting the desired version of the corresponding webpage from a so-called calendar view, which gives an overview of all available snapshots of a URL per day, month and year. As

[1] `http://web.archive.org`

[2] `https://github.com/iipc/openwayback`

URLs can be cumbersome, users on the live web often prefer to use search engines rather than remember and type URLs manually. The Internet Archive's Wayback Machine provides search only in a very rudimentary way (Goel, 2016). While the *Site Search* feature is a great improvement over plain URL lookups, this approach is pretty limited as it neither surfaces deep URLs to a specific page under a site nor supports temporal search, i.e., users cannot specify a time interval with their queries.

An alternative to search, if a URL is not known, is to follow hyperlinks from other pages. As web archives are temporal collections, such links need to carry a timestamp in addition to the URL. Within the Wayback Machine, links automatically point to the closest page or capture of the one that is currently viewed. However, it is also possible to link an archived page from the live web. In this case the timestamp needs to be set explicitly.

One way to do this is by manually pointing to a particular capture in a web archive, as is done in Wikipedia in order to keep references working[3]. Another approach to form such temporal hyperlinks is by incorporating time information that can be associated with the link, e.g., when software is referenced by its website in a research paper, the publication time of the paper can be used as a close estimator or upper bound to look up the software's website at the time that best represents the version used in the research (Holzmann et al, 2016d,e). While this example is very domain-specific to software, the same idea can be applied to other scenarios, such as preserving and referencing the evolution of people by archiving their blogs and social network profiles (Kasioumis et al, 2014; Marshall and Shipman, 2014; SalahEldeen and Nelson, 2012).

2.2 Search

Web archives can provide access to historical information that is absent on the current Web, for companies, products, events, entities etc. However, even though they have been in existence for a long time, web archives still lacking the search capabilities that would make them truly accessible and usable as temporal resources. *Web archive search* may be considered a special case of temporal information retrieval (temporal IR) (Kanhabua et al, 2015). This important subfield of IR has the goal of improving search effectiveness by exploiting temporal information in documents and queries (Alonso et al, 2007; Campos et al, 2015). The temporal dimension leads to new challenges in query understanding (Jones and Diaz, 2007) and retrieval models (Berberich et al, 2010; Singh et al, 2016), as well as temporal indexing (Anand et al, 2011, 2012). However, most temporal indexing approaches treat documents as static texts with a certain validity, which does not account for the dynamics in

[3] `https://blog.archive.org/2018/10/01/more-than-9-million-broken-links-on-wikipedia-are-now-rescued`

web archives where webpages change over time, and hence their relevance to a query may also change over time. Furthermore, while information needs in IR are traditionally classified according to the taxonomy introduced by Broder (2002), user intentions are different for web archives, as studied by Costa and Silva (2010). In contrast to the majority of queries on the live web, which are informational, queries in web archives are predominantly navigational, because users often look for specific resources in a web archive by a temporal aspect rather than searching for general information that is commonly still available on the current Web. Costa et al (2013) presented a survey of existing web archive search architectures and Hockx-Yu (2014) identified 15 web archives that already featured full-text search capabilities in 2014. With the incorporation of live web search engines, ArchiveSearch demonstrates how to search a web archive without the expensive indexing phase (Kanhabua et al, 2016).

One specific goal that is often sought by web archive search systems is to provide true *temporal archive search*: given a keyword query together with a time interval we want to find the most authoritative pages, e.g., *"what were the most representative webpages for Barack Obama before he became president in 2005?"*. This would bring up Obama's senatorial website rather than his current website and social media accounts. Such temporal semantics can often not be derived from the webpages under consideration and require external indicators. A proof-of-concept of this approach was implemented by Tempas, which in its first version incorporated tags attached to URLs on the social bookmarking platform Delicious as temporal cues (Holzmann and Anand, 2016). Its ranking was based on the frequency of a tag used with a URL, an approach that we could show results in a good temporal recall with respect to query logs from AOL and MSN (Holzmann et al, 2016b). Unfortunately, since Delicious has now closed, the available data was limited and our dataset only covers the period from 2003 to 2011. We also found that it shows a strong bias towards certain topics, like technology. For these reasons, a second version of Tempas was developed, based on hyperlinks and anchor texts. Using a graph-based query model, Tempas v2 exploits the number of websites and corresponding anchor texts linking to a URL in a given time interval, as shown in Fig. 2. Its temporally sensitive search for authority pages for entities in web archives has been shown to be very effective in multiple scenarios (Holzmann et al, 2017b), like tracing the evolution of people on the web or finding former domains of websites that have moved.

3 Data-centric View: Processing Archival Collections

In contrast to accessing web archives by close reading pages, as users do, archived contents may also be processed at scale, enabling evolution studies and big data analysis reference/mention Part 4 and Part 5 of the book?.

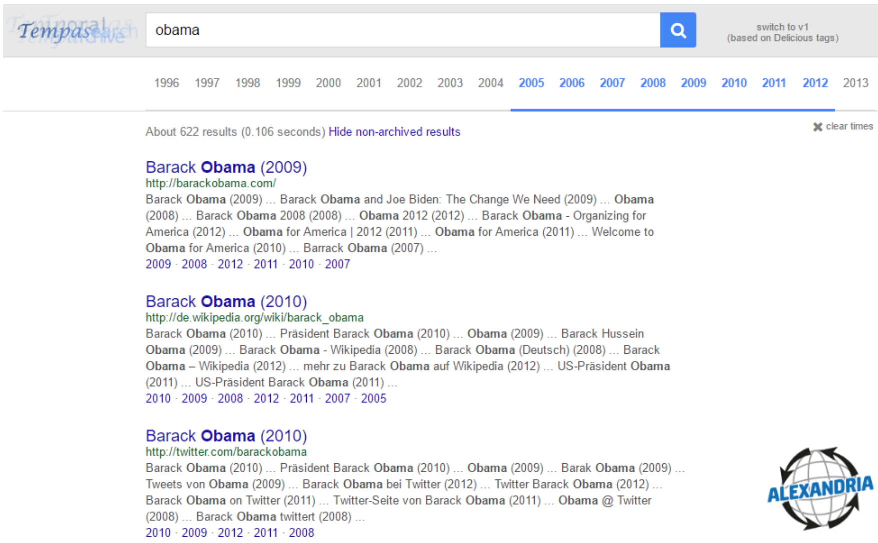

Fig. 2: Tempas v2 screenshot for query '*obama*' in period 2005 to 2012.

However, in the **data-centric view**, webpages are not considered as self-contained units with a layout and embeddings, rather single resources are treated as raw data, such as text or images. Web archives are commonly organised in two data formats: *WARC files* (Web ARChive files) store the actual archived contents, while *CDX files* (Capture Index) are comprised of lightweight metadata records. The data-centric view approaches web archives from the low-level perspective of these files, which is how data engineers would typically look at it. This perspective provides a more global point of view, looking at whole collections rather than individual records. On the downside, we have to deal with higher complexity at this level, instead of pages being nicely rendered in a web browser.

When analysing archived collections, *"What persons co-occur on the archived pages most frequently in a specific period of time?"* is only one example of the kinds of question that can be asked (Shaltev et al, 2016). Studies like this are typical cases for the graph-centric view, which is discussed below. However, answering such questions does not usually require a whole web archive, but only pages from a specific time period, certain data types or other facets that can be employed for pre-filtering the dataset during graph extraction in this data-centric perspective. One way to accomplish this is ArchiveSpark, a tool for building research corpora from web archives that operates using standard formats and facilitates the process of filtering as well as data extraction and derivation at scale in a very efficient manner (Holzmann et al, 2016a). While ArchiveSpark should be considered a tool that operates on the data-centric view, the resulting datasets consist of structured information in the form of graphs that can be used by data scientists and researchers in the graph-centric view.

We distinguish between two sorts of data that can be studied: 1. Derived, extracted or descriptive metadata, representing the web and archived records, which reflects the evolution of the web and its dynamics; 2. Contents of archived webpages, from which can be derived insights into the real world, which is commonly referred to as *Web Science* (Hall et al, 2017). The latter should be considered a graph-centric task, focusing on the required facts, after these have been prepared by data engineers from a data-centric perspective.

3.1 Metadata analysis

Web archives that span multiple years constitute a valuable resource for studying the evolution of the Web as well as its dynamics. In previous works on web dynamics, suitable datasets had to be crawled first, which is tedious and can only be done for shorter periods (Cho and Garcia-Molina, 2000; Fetterly et al, 2003; Koehler, 2002; Ntoulas et al, 2004; Adar et al, 2009). With access to existing archives, more recent studies of the Web were conducted retrospectively on available data (Hale et al, 2014; Agata et al, 2014; Alkwai et al, 2015), commonly with a focus on a particular subset, such as national domains or topical subsets. These kinds of works are typical data-centric tasks as they require access to archived raw data or metadata records.

An example of such a study is an analysis we conducted in 2016 of the dawn of today's most popular German domains over 18 years, i.e., the top-level domain `.de` from 1996 to 2013, with the required data provided by the Internet Archive (Holzmann et al, 2016c). This investigation was carried out purely by studying metadata describing the archived records, without analysing actual payloads. Based on that, we introduced properties to explore the evolution of the web in terms of age, volume and size, which can be used to replicate similar studies using other web archive collection. These properties deliver insights into the development and current state of the Web. One of our findings was that the majority of the most popular educational domains, like universities, have already existed for more than a decade, while domains relating to shopping and games have emerged steadily. Furthermore, it could be observed that the Web is getting older, not in its entirety, but with many domains having a constant fraction of webpages that are more than five years old and ageing further. Finally, we could see that popular websites have been growing exponentially since their inception, doubling in volume every two years; and that newborn pages have become bigger over time.

3.2 Web archive data processing

In order to analyse deeper structures and characteristics of the web, or to study its contents, the actual archived payloads have to be accessed. Because of the sheer size of web archives, in the order of multiple terabytes or even petabytes, this requires distributed computing facilities to process archived web data efficiently. Common operations, like selection, filtering, transformation and aggregation, may be performed using the generic *MapReduce* programming model (Dean and Ghemawat, 2010), as supported by *Apache Hadoop*[4] or *Apache Spark*[5] (Zaharia et al, 2010). AlSum (2014) presents *ArcContent*, a tool developed specifically for web archives using the distributed database *Cassandra* (Lakshman and Malik, 2010). In this approach, the records of interest are selected by means of the CDX metadata records and inserted into the database to be queried through a web service. The *Archives Unleashed Toolkit (AUT)*, formerly known as Warcbase, by Lin et al (2014), used to follow a similar approach based on *HBase*, a Hadoop-based distributed database system, which is an open-source implementation of Google's *Bigtable* (Chang et al, 2008). While being very efficient for lookups, major drawbacks of these database solutions are their limited flexibility as well as the additional effort to insert the records, which is expensive both in terms of time and resources. In its more recent version, AUT loads and processes (WARC) files directly using Apache Spark in order to avoid the HBase overhead, for which it provides convenient functions to work with web archives. With the *Archives Unleashed Cloud (AUK)*, there even exists a hosted service for a limited number of analyses based on WARC files.

In contrast to that, ArchiveSpark introduces a novel data processing approach for web archives and other archival collections that exploits metadata records for gains in efficiency while not having to rely on an external index (Holzmann et al, 2016a). ArchiveSpark is a tool for general web archive access based on Spark. It supports arbitrary filtering and data derivation operations on archived data in an easy and efficient way. Starting from the small and lightweight CDX metadata records it can run basic operations, such as filtering, grouping and sorting very efficiently, without touching the actual data payloads. In a step-wise approach, the records are enriched with additional information by applying external modules that can be customised and shared among researchers and tasks, even beyond web archives (Holzmann et al, 2017a). In order to extract or derive information from archived resources, third-party tools can be integrated. It is only at this point that ArchiveSpark seamlessly integrates the actual data for the records of interest stored in WARC files. Internally, ArchiveSpark documents the lineage of all derived and extracted information, which can serve as a source for additional filtering and processing steps or be stored in a convenient output format

[4] https://hadoop.apache.org

[5] https://spark.apache.org

to be used as a research corpus in further studies. Benchmarks show that ArchiveSpark is faster than competitors, like AUT/Warcbase and pure Spark in typical use case scenarios when working with web archive data (Holzmann et al, 2016a).

4 Graph-centric view: exploring web archive content

The final perspective, besides the *user-centric* and *data-centric views*, is referred to as the **graph-centric view**. This view enables the exploration of web archives from a more structural perspective, which constitutes the foundational model for most downstream applications and studies run by researchers, data scientists and others who are not as close to the data as engineers. In contrast to the views discussed above, the focus here is not on content or individual archived records, but on the facts and information contained within them and the relationships among them. In the context of the Web, the most obvious relationships are the hyperlinks that connect webpages by pointing from one to another. However, there is a lot more valuable data on the Web that is less obvious. Looking at hyperlinks from a more coarse-grained perspective, multiple links can be combined to connections among hosts, domains or even top-level domains, revealing connections among services, organisations or the different national regions of the Web. Furthermore, by zooming out to the graph perspective after processing the archived data from a data-centric view, even relationships among persons or objects mentioned on the analysed pages can be derived (Shaltev et al, 2016; Fafalios et al, 2017, 2018).

The holistic view of archival collections provided by graphs is very helpful in many tasks and naturally generates synergies with the other views. The broad zoom level is crucial to get an overview of available records in an archive and to find the right resources as well as to run analyses and power downstream applications. Hyperlinks among the archived pages can point users or algorithms in search or data analysis tasks to the desired entry points within the big and often chaotic web archive collections. As shown before, we make use of this with our web archive search engine Tempas (see Sec. 2). The effectiveness of hyperlinks and attached anchor texts for this task has already been shown by previous works (Craswell et al, 2001; Kraaij et al, 2002; Ogilvie and Callan, 2003; Koolen and Kamps, 2010).

4.1 Data analysis

The approaches for exploring web archives through graphs that are described here allow for queries on a structural level (cf. Fig. 1). Once a set of documents

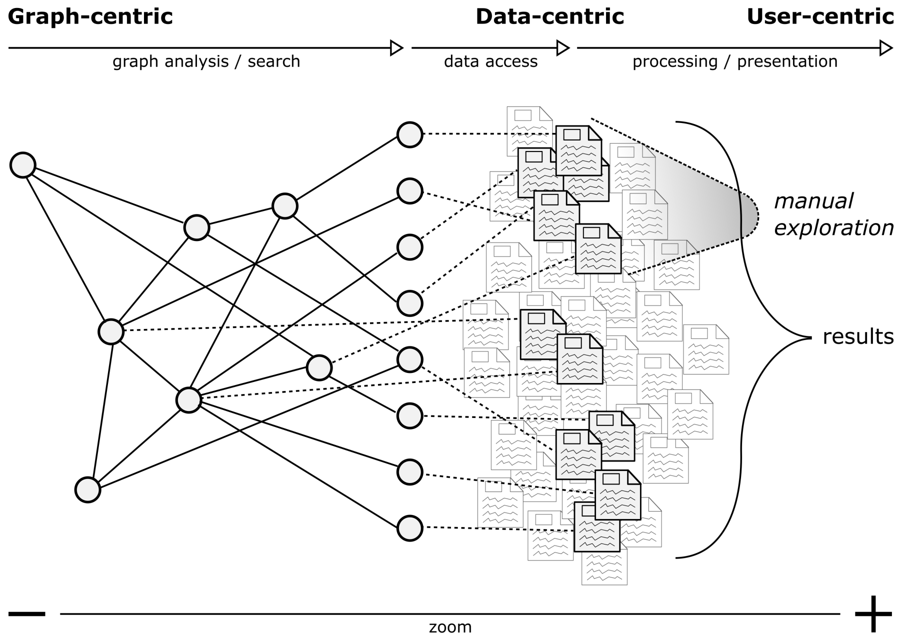

Fig. 3: Combining different views on web archives for systematic data analysis.

that match a query has been identified, a data engineer might be involved to zoom in to the contents in order to extract new structured knowledge graphs from a data-centric perspective, to be processed further by data scientists or the like. Quite commonly, such workflows also involve manual inspection of the records under consideration from a user-centric view. This is helpful to get an understanding of the data that is being studied. Ultimately, derived results need to be aggregated and presented to the user in an appropriate form.

Fig. 3 shows this generic analysis schema which outlines a systematic way to study web archives. This schema can be adopted and implemented for a range of different scenarios. In such a setting, the graph-centric view is utilised to get an overview and find suitable entry points into the archive. This may initially be done manually by the user, to get a feeling for the available data using a graph-based search engine like Tempas, but can also be integrated as the first step in a data-processing pipeline to (semi-)automatically select the corpus for further steps. Next, the selected records can be accessed from a data-centric view at scale, using a tool like ArchiveSpark (see Sec. 3), to extract the desired information, compute metrics or aggregate statistics. Finally, the results are presented to the user. A concrete implementation of this pipeline is outlined in Holzmann et al (2017b), where we describe the example of analysing restaurant menus and compare prices before and after the introduction of the Euro as Europe's new currency in Germany in 2001-2.

4.2 Open challenges

The reason for addressing the graph-centric view at the end of this chapter is because it requires a certain understanding of the eventual task or study in order to evaluate its utility. While there are many synergies between graphs and the challenges discussed above, in which this structural perspective is very helpful, they also raise new issues and questions. Graphs enable completely different kinds of analysis, such as centrality computations with algorithms like *PageRank* (Page et al, 1999). However, scientific contributions in this area specific to web archives are very limited and results are less mature. Although scientists have looked into graph properties of the web in general, both in static (Albert et al, 1999; Broder et al, 2000; Adamic et al, 2000; Suel and Yuan, 2001; Boldi and Vigna, 2004) and evolving graphs (Huberman and Adamic, 1999; Leskovec et al, 2005, 2007), we found that certain traits of web archives lead to new kinds of questions. For instance, as we show in Holzmann et al (2018, 2019), the inherent incompleteness of archives can affect rankings produced by graph algorithms on web archive graphs.

5 Summary

Web archives have been instrumental in the digital preservation of the Web and provide great opportunities for the study of the societal past and its evolution. These archival collections are massive datasets, typically in the order of terabytes or petabytes, spanning time periods of up to more than two decades and growing. As a result of this, their use has been difficult, as effective and efficient exploration, and methods of access, are limited. We have identified three views on web archives, for which we have proposed novel concepts and tools to tackle existing challenges: user-, data- and graph-centric. Depending on who you are, these provide you with the right perspective from which to approach archived web data for your needs, with suitable abstractions and simplifications. Switching between roles and combining different views provides a holistic view on web archives.

Acknowledgements This work was partially funded by the EU Horizon 2020 under ERC grant "ALEXANDRIA" (339233).

References

Adamic LA, Huberman BA, Barabási AL, Albert R, Jeong H, Bianconi G (2000) Power-Law Distribution of the World Wide Web. Science 287(5461):2115, DOI 10.1126/science.287.5461.2115a

Adar E, Teevan J, Dumais ST, Elsas JL (2009) The Web Changes Everything: Understanding the Dynamics of Web Content. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining - WSDM '09, ACM Press, pp 282–291, DOI 10.1145/1498759.1498837

Agata T, Miyata Y, Ishita E, Ikeuchi A, Ueda S (2014) Life Span of Web Pages: A Survey of 10 Million Pages Collected in 2001. Digital Libraries pp 463–464, DOI 10.1109/JCDL.2014.6970226

Albert R, Jeong H, Barabási AL (1999) Internet: Diameter of the World-Wide Web. Nature 401(6749):130–131, DOI 10.1038/43601

Alkwai LM, Nelson ML, Weigle MC (2015) How Well Are Arabic Websites Archived? In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '15, ACM, pp 223–232, DOI 10.1145/2756406.2756912

Alonso O, Gertz M, Baeza-Yates R (2007) On the Value of Temporal Information in Information Retrieval. ACM SIGIR Forum 41(2):35–41, DOI 10.1145/1328964.1328968

AlSum A (2014) Web Archive Services Framework for Tighter Integration between the Past and Present Web. PhD thesis, Old Dominion University

Anand A, Bedathur S, Berberich K, Schenkel R (2011) Temporal Index Sharding for Space-time Efficiency in Archive Search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '11, ACM, New York, NY, USA, pp 545–554, DOI 10.1145/2009916.2009991

Anand A, Bedathur S, Berberich K, Schenkel R (2012) Index Maintenance for Time-travel Text Search. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '12, Portland, Oregon, USA, pp 235–244, DOI 10.1145/2348283.2348318

Berberich K, Bedathur S, Alonso O, Weikum G (2010) A Language Modeling Approach for Temporal Information Needs. In: Proceedings of the 32Nd European Conference on Advances in Information Retrieval (ECIR), Springer-Verlag, Berlin, Heidelberg, ECIR'2010, pp 13–25, DOI 10.1007/978-3-642-12275-0_5

Boldi P, Vigna S (2004) The WebGraph Framework I: Compression Techniques. In: Proceedings of the 13th Conference on World Wide Web - WWW '04, ACM, ACM Press, Manhattan, USA, pp 595–602, DOI 10.1145/988672.988752, URL http://law.di.unimi.it/datasets.php

Broder A (2002) A Taxonomy of Web Search. In: ACM Sigir forum, ACM, Association for Computing Machinery (ACM), vol 36, pp 3–10, DOI 10.1145/792550.792552

Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph Structure in the Web. Computer Networks 33(1):309–320, DOI 10.1016/s1389-1286(00)00083-9

Campos R, Dias G, Jorge AM, Jatowt A (2015) Survey of Temporal Information Retrieval and Related Applications. ACM Computing Surveys (CSUR) 47(2):15

Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE (2008) Bigtable: A Distributed Storage System for Structured Data. ACM Transactions on Computer Systems (TOCS) 26(2):4, DOI 10.1145/1365815.1365816

Cho J, Garcia-Molina H (2000) The Evolution of the Web and Implications for an Incremental Crawler. In: Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00

Cohen S, Li C, Yang J, Yu C (2011) Computational Journalism: A Call to Arms to Database Researchers. In: Proceedings of the 5th Biennial Conference on Innovative Data Systems Research, pp 148–151

Costa M, Silva MJ (2010) Understanding the Information Needs of Web Archive Users. In: Proceedings of the 10th International Web Archiving Workshop

Costa M, Gomes D, Couto F, Silva M (2013) A Survey of Web Archive Search Architectures. In: Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion, ACM Press, New York, NY, USA, pp 1045–1050, DOI 10.1145/2487788.2488116

Craswell N, Hawking D, Robertson S (2001) Effective Site Finding Using Link Anchor Information. In: Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '01, ACM, ACM Press, DOI 10.1145/383952.383999

Dean J, Ghemawat S (2010) Mapreduce: A Flexible Data Processing Tool. Communications of the ACM 53(1):72–77, DOI 10.1145/1629175.1629198

Fafalios P, Holzmann H, Kasturia V, Nejdl W (2017) Building and Querying Semantic Layers for Web Archives. In: Proceedings of the 17th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '17, IEEE, DOI 10.1109/jcdl.2017.7991555

Fafalios P, Holzmann H, Kasturia V, Nejdl W (2018) Building and querying semantic layers for web archives (extended version). International Journal on Digital Libraries DOI 10.1007/s00799-018-0251-0, URL `https://doi.org/10.1007/s00799-018-0251-0`

Fetterly D, Manasse M, Najork M, Wiener J (2003) A Large-scale Study of the Evolution of Web Pages. In: Proceedings of the 12th International Conference on World Wide Web - WWW '03, pp 669–678, DOI 10.1002/spe.577

Goel V (2016) Beta Wayback Machine - Now with Site Search! URL `https://blog.archive.org/2016/10/24/beta-wayback-machine-now-with-site-search`, [Accessed: 16/03/2017]

Hale SA, Yasseri T, Cowls J, Meyer ET, Schroeder R, Margetts H (2014) Mapping the UK Webspace: Fifteen Years of British Universities on the Web. In: Proceedings of the 2014 ACM Conference on Web Science - WebSci '14, ACM Press, WebSci '14, DOI 10.1145/2615569.2615691

Hall W, Hendler J, Staab S (2017) A Manifesto for Web Science @10. arXiv:170208291

Hockx-Yu H (2014) Access and Scholarly Use of Web Archives. Alexandria: The Journal of National and International Library and Information Issues 25(1-2):113–127, DOI 10.7227/alx.0023

Holzmann H (2019) Concepts and Tools for the Effective and Efficient Use of Web Archives. PhD thesis, Leibniz Universität Hannover, DOI 10.15488/4436

Holzmann H, Anand A (2016) Tempas: Temporal Archive Search Based on Tags. In: Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion, ACM Press, DOI 10.1145/2872518.2890555

Holzmann H, Goel V, Anand A (2016a) Archivespark: Efficient Web Archive Access, Extraction and Derivation. In: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '16, ACM, New York, NY, USA, pp 83–92, DOI 10.1145/2910896.2910902

Holzmann H, Nejdl W, Anand A (2016b) On the Applicability of Delicious for Temporal Search on Web Archives. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16, ACM Press, Pisa, Italy, DOI 10.1145/2911451.2914724

Holzmann H, Nejdl W, Anand A (2016c) The Dawn of Today's Popular Domains - A Study of the Archived German Web Over 18 Years. In: Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '16, IEEE, ACM Press, Newark, New Jersey, USA, pp 73–82, DOI 10.1145/2910896.2910901

Holzmann H, Runnwerth M, Sperber W (2016d) Linking Mathematical Software in Web Archives. In: Mathematical Software – ICMS 2016, Springer International Publishing, pp 419–422, DOI 10.1007/978-3-319-42432-3_52

Holzmann H, Sperber W, Runnwerth M (2016e) Archiving Software Surrogates on the Web for Future Reference. In: Research and Advanced Technology for Digital Libraries, 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, Hannover, Germany, DOI 10.1007/978-3-319-43997-6_17

Holzmann H, Goel V, Gustainis EN (2017a) Universal Distant Reading through Metadata Proxies with Archivespark. In: 2017 IEEE International Conference on Big Data (Big Data), IEEE, Boston, MA, USA, DOI 10.1109/bigdata.2017.8257958

Holzmann H, Nejdl W, Anand A (2017b) Exploring Web Archives through Temporal Anchor Texts. In: Proceedings of the 2017 ACM on Web Science Conference - WebSci '17, ACM Press, Troy, New York, USA, DOI 10.1145/3091478.3091500

Holzmann H, Anand A, Khosla M (2018) What the HAK? Estimating Ranking Deviations in Incomplete Graphs. In: 14th International Workshop on Mining and Learning with Graphs (MLG) - Co-located with 24th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), London, UK

Holzmann H, Anand A, Khosla M (2019) Delusive pagerank in incomplete graphs. In: Complex Networks and Their Applications VII, Springer International Publishing

Huberman BA, Adamic LA (1999) Internet: Growth Dynamics of the World-wide Web. Nature 401(6749):131–131

Jones R, Diaz F (2007) Temporal Profiles of Queries. ACM Transactions on Information Systems 25(3):14–es, DOI 10.1145/1247715.1247720

Kanhabua N, Blanco R, Nørvåg K, et al (2015) Temporal Information Retrieval. Foundations and Trends® in Information Retrieval 9(2):91–208, DOI 10.1145/2911451.2914805

Kanhabua N, Kemkes P, Nejdl W, Nguyen TN, Reis F, Tran NK (2016) How to Search the Internet Archive Without Indexing It. In: Research and Advanced Technology for Digital Libraries, Springer International Publishing, Hannover, Germany, pp 147–160, DOI 10.1007/978-3-319-43997-6_12

Kasioumis N, Banos V, Kalb H (2014) Towards Building a Blog Preservation Platform. World Wide Web Journal 17(4):799–825, DOI 10.1007/s11280-013-0234-4

Koehler W (2002) Web Page Change and Persistence-a Four-year Longitudinal Study. Journal of the American Society for Information Science and Technology 53(2):162–171, DOI 10.1002/asi.10018

Koolen M, Kamps J (2010) The Importance of Anchor Text for Ad Hoc Search Revisited. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10, ACM, ACM Press, pp 122–129, DOI 10.1145/1835449.1835472

Kraaij W, Westerveld T, Hiemstra D (2002) The Importance of Prior Probabilities for Entry Page Search. In: Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '02, ACM, DOI 10.1145/564376.564383

Lakshman A, Malik P (2010) Cassandra: A Decentralized Structured Storage System. ACM SIGOPS Operating Systems Review 44(2):35–40, DOI 10.1145/1773912.1773922

Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In: Proceeding of the 1th ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05, ACM, ACM Press, pp 177–187, DOI 10.1145/1081870.1081893

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1):2, DOI 10.1145/1217299.1217301

Lin J, Gholami M, Rao J (2014) Infrastructure for Supporting Exploration and Discovery in Web Archives. In: Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, ACM Press, DOI 10.1145/2567948.2579045

Marshall CC, Shipman FM (2014) An Argument for Archiving Facebook As a Heterogeneous Personal Store. In: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '14, IEEE Press, pp 11–20, DOI 10.1109/jcdl.2014.6970144

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J, Pinker S, Nowak MA, and ELA (2010) Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331(6014):176–182, DOI 10.1126/science.1199644

Moretti F (2005) Graphs, Maps, Trees: Abstract Models for a Literary History. Verso

Ntoulas A, Cho J, Olston C (2004) What's New on the Web?: The Evolution of the Web from a Search Engine Perspective. In: Proceedings of the 13th Conference on World Wide Web - WWW '04, ACM Press, pp 1–12, DOI 10.1145/988672.988674

Ogilvie P, Callan J (2003) Combining Document Representations for Known-item Search. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval - SIGIR '03, ACM, ACM Press, DOI 10.1145/860435.860463

Page L, Brin S, Motwani R, Winograd T (1999) The PageRank Citation Ranking: Bringing Order to the Web. InfoLab

SalahEldeen HM, Nelson ML (2012) Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? In: Theory and Practice of Digital Libraries, Springer, Paphos, Cyprus, TPDL'12, pp 125–137, DOI 10.1007/978-3-642-33290-6_14

Schreibman S, Siemens R, Unsworth J (2008) A Companion to Digital Humanities. Blackwell Publishing

Shaltev M, Zab JH, Kemkes P, Siersdorfer S, Zerr S (2016) Cobwebs from the Past and Present: Extracting Large Social Networks Using Internet Archive Data. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '16, Pisa, Italy, DOI 10.1145/2911451.2911467

Singh J, Nejdl W, Anand A (2016) History by Diversity: Helping Historians Search News Archives. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval - CHIIR '16, ACM Press, pp 183–192, DOI 10.1145/2854946.2854959

Suel T, Yuan J (2001) Compressing the Graph Structure of the Web. In: Data Compression Conference

Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: Cluster Computing with Working Sets. In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing, vol 10, p 10