Exploring Web Archives Through Temporal Anchor Texts*

Helge Holzmann, Wolfgang Nejdl, Avishek Anand
L3S Research Center
Appelstr. 9a
30167 Hannover, Germany
{holzmann,nejdl,anand}@L3S.de

ABSTRACT

Web archives have been instrumental in digital preservation of the Web and provide great opportunity for the study of the societal past and evolution. These Web archives are massive collections, typically in the order of terabytes and petabytes. Due to this, search and exploration of archives has been limited as full-text indexing is both resource and computationally expensive. We identify that for typical access methods to archives, which are navigational and temporal in nature, we do not always require indexing full-text. Instead, meaningful text surrogates like *anchor texts* already go a long way in providing meaningful solutions and can act as reasonable entry points to exploring Web archives.

In this paper, we present a new approach to searching Web archives based on temporal link graphs and corresponding anchor texts. Departing from traditional informational intents, we show how temporal anchor texts can be effective in answering queries beyond purely navigational intents, like finding the most central webpages of an entity in a given time period. We propose indexing methods and a temporal retrieval model based on anchor texts. Further, we discuss several interesting search results as well as one experiment in which we demonstrate how such results can be integrated in a data processing workflow to scale up to thousands of pages. In this analysis we were able to replicate results reported by an offline study, showing that restaurant prices indeed increased in Germany when the Euro was introduced as Europe's currency.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; • **Applied computing → Digital libraries and archives**;

KEYWORDS

Web Archives; Temporal Information Retrieval; Big Data Analysis

1 INTRODUCTION

Web Science is the interdisciplinary endeavor of studying the Web to derive insights into our world [13]. While Web archives constitute extremely valuable datasets for this task by providing access to the Web in a temporal manner, methods and tools as well as best practices to work with these archives have widely been neglected. We address both in this paper: first, we present a live system that provides retrieval models useful for downstream historical analysis, and second, we outline an example use case that exemplifies how historical phenomena on the Web can be investigated using our temporal retrieval approach with Web archives.

Web Archives have been continually archiving the Web in an attempt to document the digital history of human society. In the process, they have been amassing Web content in the order of terabytes and sometimes up to petabytes. What makes such archives attractive is that they are repositories of past contents that encode evolution. Consider the example of German's chancellor *Angela Merkel*, who developed from the leader of a political party to the leader of a country, playing an important role in Europe. This evolution is reflected by the websites representing her on the Web, which used to be a subpage under her party's site and is now her official website and social media profiles (cp. Sec. 5.2). For users of Web archives, such changes make it difficult to lookup past records on the Web.

Supporting keyword search with temporal predicates on these massive datasets typically requires full-text indexing infrastructures. However, due to the large sizes of archives, creating those indexes can be both computationally expensive as well as resource intensive, even with distributed architectures in place. We note that whereas full-text search is beneficial for a wide variety of informational intents, there are specialized intents on archives for which we might not always require indexing full-text. Specifically, most of the intents for information in Web archives is navigational and temporal in nature. Users are often interested in specific URLs or subjects and their evolution over time: versions of people's webpages and social media pages, movements, societies and organizations, or entities in general. In this work, we explore the possibility of identifying surrogate information units that is accurate enough for such information needs.

Towards this, we identify such surrogate information units as anchor texts with their accompanying link structure and propose lean index organization methods to support such temporal, navigational needs. Anchor texts are short, concise, important and non-noisy descriptors of information content typically desired by navigational queries. Another immediate benefit of indexing anchor texts as surrogates of the actual webpages is that they are many orders smaller than the original full text. The accompanying index is leaner, computationally less intensive for construction and storage as well as faster to query.

We introduce a temporal Web model that represents a Web archive as an edge-labeled, temporal, directed graph. Further, we

*This work is partly funded by the European Research Council under ALEXANDRIA (ERC 339233)

WebSci'17, June 25-28, 2017, Troy, NY, USA.

DOI: http://dx.doi.org/10.1145/3091478.3091500

propose a retrieval model based on the semantics of anchor texts towards ranking webpages not only by overall relevance but under consideration of their decisive times and versions. While this retrieval method does not cover pure informational needs, it successfully identifies results beyond navigational needs including representative pages for entities that vary over time. With this retrieval model we present a fully functional indexing and retrieval system based on anchor texts, called Tempas (*Temporal archive search*), that can be accessed live at :

http://tempas.L3S.de

Further, we discuss several interesting search results as well as one experiment in which we demonstrate how such results can be integrated into a data processing workflow to scale up to thousands of pages, showing how restaurant prices increased in Germany when the Euro as a new currency was introduced, as reported by an offline study [11].

2 TEMPORAL WEB ARCHIVE SEARCH

Information needs of users are different on Web archives than on the current Web. The goal of our Tempas system is to satisfy those needs, which are primarily navigational with a temporal aspect, as we discuss in Section 2.1. In Tempas users can formulate their information need by specifying a textual query with the option of selecting a time interval of interest. The screenshot in Figure 1 shows the user interface. Tempas (v2) is based on anchor texts, which feature some specific characteristics that make them well-suited for this our purpose, as listed in Section 2.2. Ultimately, in Section 2.3, we summarize the problem that we address with our system by following the approach presented in Section 4.

2.1 User Intent in Web Archives

User intents formulated as queries and issued to a Web search engine are commonly classified by their information needs into *informational*, *navigational* and *transactional*. Broder [6] analyzed query logs and found that around a half of the queries are informational. The other half is roughly split into 40% navigational and 60% transactional queries.

These proportions are different for Web archives. There is seldom the need to issue an informational query to a Web archive, partly because most informational facts and intents can be served on the current Web as well. Also, transactional queries, which refer to an action that a user wants to perform, e.g., chat or shop online, are typically not applicable in an archive. Therefore, the majority of queries to a Web archive are navigational.

Costa and Silva [9] confirmed this assumption by analyzing query logs of their full-text search engine for the *Portugese Web Archive*. They report more than a half of the queries to be navigational. From the other half a large majority was informational with only 5-10% being transactional. However, what they consider transactional is much more specific than the original definition, such as downloading an old file or recovering a specific website. Similarly, their informational need refers to collecting information about a subject in the past and can often be interpreted as navigational.

Indeed, all information needs of Web archives could be considered navigational in a broad sense. That is, instead of navigating to a specific resource, we want to navigate to a specific information or subject, e.g., an entity. Some of these entities are represented on the Web by their personal or official websites, others by profiles on social networks or sub-pages on related websites, as well as Wikipedia or similar knowledge bases. We refer to these *central* resources as *authority pages* for a subject / an entity. These are dynamic though, and may change over time by some disappearing or moving to a different domain as well as new ones emerging.

2.2 Searching Anchor Texts

Links between pages are typically anchored by a text segment, i.e., the text in the source page that can be clicked on, called *anchor text*. Anchor texts are a special type of text and should not be treated as running text, like articles or similar content. Although those snippets have some disadvantages in being potentially less descriptive than the text surrounding a link or the content of a linked page, they have great advantages for our primary use case of finding *authority pages* as described above:

- Anchor texts describe with a high confidence the linked webpage with **little distracting or unrelated content**. This prevents pages from showing up for unrelated query terms and gives a higher relative relevance to page actually related to a query term.
- Often, anchor texts contain the name of the linked page or a **concise label of the linked content**. Hence, these terms are frequently used to link to authority pages of entities. For instance, the name of a person is outstandingly often used to link to the person's website, which is therefore likely to be ranked high for this query.
- As probably intended by the original reason for hyperlinks on the Web, anchor texts frequently point to pages containing a more detailed description of the contained terms. This way, instead of repeating **descriptions or definitions**, pages link to the most meaningful explanation of the linked text, which often is a social profile, the official website, or an encyclopedic article, e.g., *Wikipedia*.
- Within a website, anchor texts are used as **navigational elements** to refer to certain parts of the site, e.g., the menu on a restaurant site (s. Sec 6). The same term is then commonly used in connection with the site's name to deep link into that part of the website from external pages.

2.3 Problem Statement

The main objective of Tempas is to meet the information need of a user exploring a Web archive and fulfill the user's intent as defined in Section 2.1: Given a textual keyword query together with a time interval we want to identify those webpages that are central for the subject addressed by the query in the specified time period. For instance, before the *European Union* received its own `.eu` top-level domain in 2005, the official website resided under `.eu.int` (cp. Sec. 5.2). Another kind of pages that are of interest when working with Web archives are those in a certain category or with a certain type of contents, such as online shops or restaurant menus. In contrast to queries for *authority pages*, which are rather precision oriented, here recall matters, for instance in data mining tasks (cp. Sec. 6).

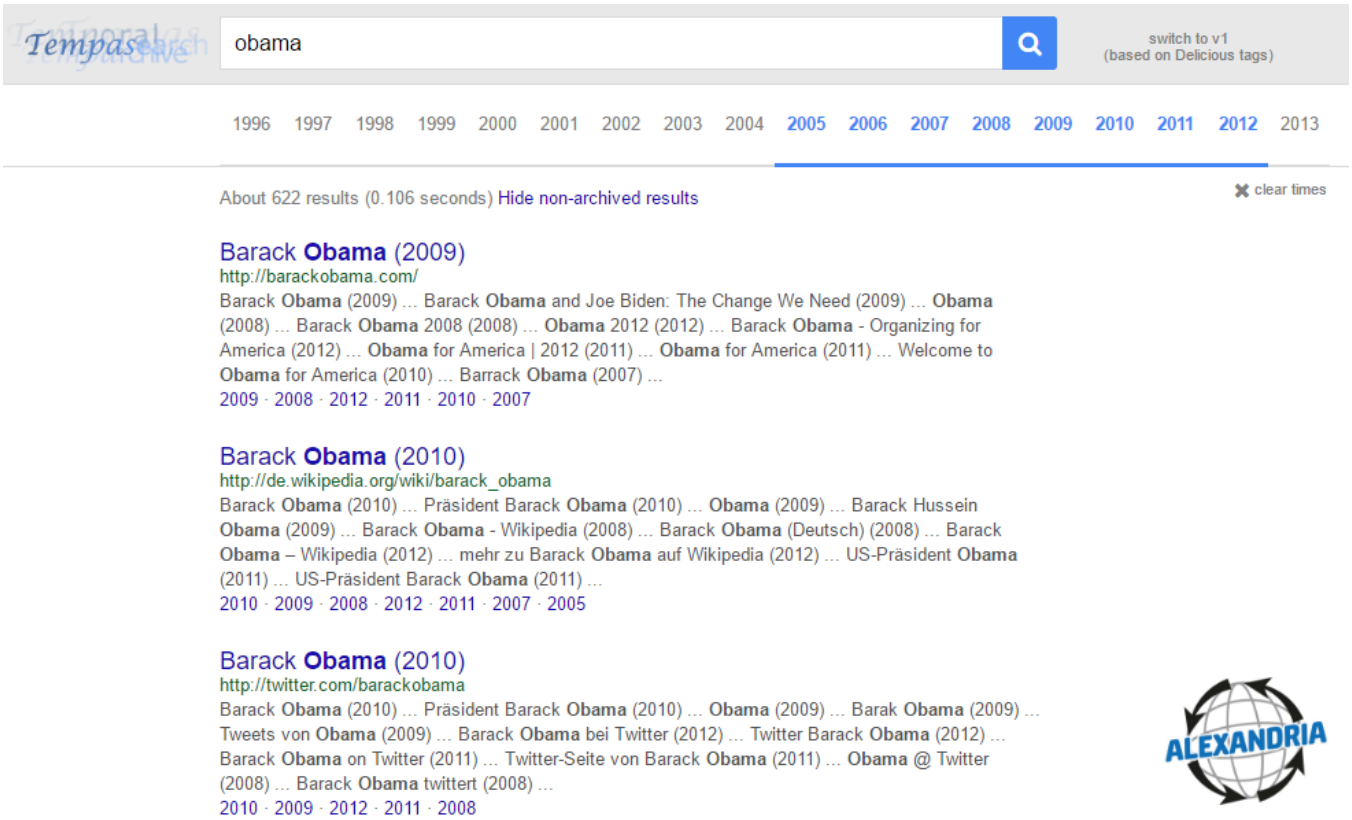

Figure 1: Tempas v2 screenshot for query '*obama*' in period 2005 to 2012.

In summary, both types of navigational queries serve as important entry points into huge Web archives, which is what we are aiming for. Even though users commonly have a subjective understanding of this problem, a quantitative evaluation is not trivial due to the lack of a crisp definition of an authority page or appropriate entry point. Moreover, we found that existing relevance judgments used in Web information retrieval are not suitable for evaluating this task. For instance, in the *TREC 2012 Web Track* adhoc judgements[1], `phoenix.edu` was considered irrelevant for the query *university of phoenix*, which we consider a perfect hit. Therefore, we conduct a qualitative evaluation based on example queries in Section 5 and a data analysis scenario in Section 6.

3 RELATED WORK

The predecessor of the system presented in this work, i.e., the first version of Tempas (v1), was built with the same goals in mind: *temporal Web archive search* for *authority pages* given a keyword query together with a time interval [14]. However, instead of using anchor texts, we incorporated tags attached to URLs on the social bookmarking platform *Delicious*. Without evaluating the precision of the ranking, which was based on the frequency of a tag used with a URL, we showed that this approach results in a good recall with respect to URLs clicked on available query logs from AOL and MSN [16]. Switching to anchor texts was the natural consequence, since Delicious is a closed system and available data is limited, with our dataset ranging only from 2003 to 2011. We also showed that it is very biased towards a certain group of users and the shift to anchor texts can deliver better results, for more diverse queries.

Navigational intent in Web archive search: Web archive search can be considered a special case of temporal information retrieval (IR). While information needs in IR are traditionally classified according to the taxonomy introduced by Broder [6], user intents are different for Web archives as studied by Costa and Silva [9]. In contrast to the majority of queries being informational, where users search for information, in Web archives queries are predominantly navigational, because users often look for specific resources in a Web archive under a temporal aspect (cp. Sec. 2.1). The Internet Archive's *Wayback Machine*[2] recently got a *site search* feature based on anchor texts [12], using an approach similar to ours. However, in contrast to our Tempas system, the *Wayback Site Search* has no explicit temporal search support. Users cannot specify a time interval for their queries, and results are limited to homepages, i.e., the hostname of a URL without a path. Thus, it can find *Barack*

[1] http://trec.nist.gov/data/web2012.html

[2] http://archive.org/web

Obama's official website, but not his Wikipedia article or social media profiles.

Temporal indexing and retrieval models: Temporal information retrieval has emerged as an important subfield in information retrieval with the goal to improve search effectiveness by exploiting temporal information in documents and queries [7]. The value of the temporal dimension was clearly identified in [1] and has led to a plethora of work which utilizes temporal features in query understanding [17], retrieval models [5, 22] and temporal indexing [2, 3]. A survey by Campos et al. [7] gives an elaborate overview of the field. Most of the temporal retrieval models either focus on temporal informational intents [5] or are concerned with increasing recall with diversification [4, 22]. Temporal indexing approaches [2, 3] over Web archives assume documents to be versions of full-text content. A survey of existing Web archive search architectures was presented by Costa et al. [8]. We posit that building a suitable temporal full-text index for Web archive data is challenging and expensive though, and has never been shown to be very effective. Contrary to previous approaches that concern themselves with full-text indexes and ad-hoc retrieval tasks we focus on building minimalistic indexes for temporal navigational intents.

Effectiveness of anchor texts: The effectiveness of anchor texts for the task of *site finding* was already shown by Craswell et al. [10], though not in the context of Web archives or a temporal setting. They are reported to be twice as effective as searching the contents of pages. The authors in Kraaij et al. [20] combined anchor texts with content features for the task of *entry page search* and also found that search just based on anchor texts outperforms basic content features. In a similar experiment, Ogilvie and Callan [21] showed that anchor texts are the most effective features among others, such as full text and title, for the task of finding homepages and are only slightly behind full-text search for finding so-called *named pages*. Koolen and Kamps [19] re-evaluated the effectiveness of anchor texts in ad-hoc retrieval and showed that propagated anchor text outperforms full-text retrieval in terms of early precision on the TREC 2009 Web track. The authors in Kanhabua and Nejdl [18] studied anchor texts in a temporal context and analyzed their value in Wikipedia. Similar to our findings presented in Section 5.2, they were able to observe evolutions of entities through anchor texts, such as the transition of *Barack Obama* from senator to president. They also proposed a temporal anchor text model for their study, though specific to Wikipedia.

4 APPROACH

Our approach to temporal Web archive search is based on indexing anchor texts extracted from time-varying Web corpora. We build our work on the observation that anchor texts are crucial text segments, which in many cases succeed well to describe succinctly the target webpage, and hence are a natural choice for navigational queries. In Section 4.1 we first detail the temporal Web model with hyperlinks and anchor texts used as node descriptors. We then describe our retrieval model based on the indexed anchor text statistics in Section 4.2, before we outline the index layout, construction and query processing for efficient retrieval in Section 4.3.

4.1 Temporal Web Model

Web archive model: A Web archive is a collection of pages $p \in \mathcal{P}$, each with one or multiple versions / captures $p = \{c_{p,t_1} \dots c_{p,t_n}\}$ representing the page at different points in time $(t_1 \dots t_n)$, that is when the page was crawled and archived. Each capture c_t represents a response of the web server returned to the crawler at time t for a requested page p. This can be a successful response including the contents of the page or it could be a different status reply hinting at the page being offline or moved. As we are only interested in the hyperlinks on a page, we model a capture as the set of contained links if successful or the empty set otherwise: A capture $c_i \subseteq L$ of page p_i is a set of links where each link $l_{ijm} \in L$ points from p_i to another page p_j, anchored by a text segment $a_m \in A$ (on p_i), i.e., the text that can be clicked on, typically called *anchor text*. In essence each link l_{ijm} connects a source page p_i with the anchor text a_m to a destination page p_j.

Since the **Web is dynamic** and pages change over time, **links can change** as well. While this can happen multiple times between two captures, we only consider those changes than can be observed through one of the captures in the archive. Link changes are constituted by either:

- a link is removed from a page,
- a new link is added to a page,
- the anchor text of a link changes,
- the target page of a link changes.

A change of the source would be considered the link to be removed from its original page and added to a new one. As we identify a link by its properties, i.e., an edge consisting of a source and a target page as well as its anchor text, the result of any of the four types of changes applied to a link would be considered a different link. However, a link l_{ijm} being removed from page p_i and added back later to it with the same target page p_j and the exact same anchor text a_m is treated as the same link again.

Graph model: Let $G = (V, E)$ be a graph with a set of nodes / vertices $v_i \in V$ and a set of directed edges $e_{jk} = (v_j, v_k) \in E \subseteq V \times V$ pointing from node v_j to v_k. For the Web or a Web archive, such a graph represents the links connecting webpages: each node $v_i \in V$ represents a webpage and an edge $e_{jk} \in E$ corresponds to a link from webpage v_j to page v_k. Two or more links on the same page with the same destination and anchor text are considered to be equal and only counted as one. However, it is possible that multiple links $l_{jkm} \in L$ exist for the same edge e_{jk}. Hence, the Web $\mathcal{W}$ can be modeled as a edge-labeled, directed graph, with the labels being the set of anchor texts accompanied by a set of links: $\mathcal{W} = (G, A, L)$.

In Web archives, which have a temporal dimension, there are multiple ways to **extract a representation of the Web** W as defined above **for a time interval** $[t_a, t_b]$. As the graph G as well as the set of anchor texts A in $\mathcal{W}$ can be derived from the set of links L, the temporal representation depends on how we extract the links from the captures during the given interval: $L \subseteq \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t \geq t_a \wedge t \leq t_b\}$. While there exist various ways to define this set, we identified the following three:

1. **Merge**. The most straight-forward way is to merge all links that existed in some capture during the time interval under consideration. Hence, the resulting graph also includes links that existed

in some capture but were deleted later in another capture of the same page during the interval. Two or more links corresponding to the same edge but labeled with different anchor texts at different time points during the interval are all included:

$$L_{\texttt{merge}} = \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t_a \leq t \leq t_b\}$$

This is the most complete representation as it merges all links, regardless of whether they still exist at the end of the interval or not.

2. **Temporal Snapshot**. A temporal snapshot or simply *snapshot* refers to a single point in time as opposed to a period. For a given time interval $[t_a, t_b]$, the snapshot representation at time t_b contains only links from the latest captures c_t for each page from that interval with $t \leq t_b$, i.e., links that actually exist at time t_b:

$$L_{\texttt{snapshot}} = \bigcup_{p \in \mathcal{P}} \underset{c_t \in \{c_t \in p | t_a \leq t \leq t_b\}}{\arg\max} t$$

This model resembles the actual Web at time t_b most closely and is best suited for analyzing the Web or its structure. However, it is not ideal for information retrieval as we miss the intermediate states in the considered time period corresponding to the granularity of our index.

3. **Emergence**. *Emergence* refer to the links posted / created in a given time interval $[t_a, t_b]$. In that respect it is as complete as a *merged* representation if all consecutive intervals are considered. However, it is more space efficient as it does not contain the captures already present before the considered interval:

$$L_{\texttt{emergence}} = \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t_a \leq t \leq t_b\} \setminus \bigcup_{p \in \mathcal{P}} c \in \{c_t \in p | t < t_a\}$$

This representation has two major advantages for temporal search, which is the reason why we chose $L_{\texttt{emergence}}$ to build our indexes: 1) only pages that are actively being linked in the queried time period are considered, while pages that got frequently linked earlier but are not relevant anymore are ignored even if the old links still exist on the Web, 2) the index size is significantly reduced as each link is only included in one interval.

4.2 Temporal Retrieval Model

Based on a given granularity (we use one year), we split the duration of a provided Web archive collection into equally sized time intervals $\{[t_0, t_1], [t_2, t_3], \ldots, [t_{n-1}, t_n]\}$. For each of these time intervals we create a temporal Web representation, derived from the set of links $L_{[t_a, t_b]}$, which is defined as described above (here: $L_{\texttt{emergence}}$):

$$\mathcal{W}_{[t_a, t_b]} = (G_{[t_a, t_b]}, A_{[t_a, t_b]}, L_{[t_a, t_b]})$$

In the following we omit the interval and treat the Web model $\mathcal{W} = (G, A, L)$ for every interval independently. From the graph $G = (V, E)$ of W, each node $v \in V$ represents a page that is either part of the Web archive in the current interval or is linked to from such a page but not necessarily contained itself. Let $freq(v, a)$ be a scoring function used to compute the relevance of the page represented by node v for a given anchor text $a \in A$. We define this function based on the edges $e = (u, v) \in E$ with source $u \in V$ and destination v for which a link $l \in L$ exists with anchor text a (l_{uva}). Instead of counting these edges directly, we count the number of different hosts $host(u)$ of the source nodes u of the edges, i.e., the hostname in the URL of the page corresponding to node u, e.g., `en.wikipedia.org` in `https://en.wikipedia.org/wiki/World_Wide_Web`:

$$freq(v, a) = |\{host(u) | e = (u, v) \in E \wedge (e, a) \in L\}|$$

Host frequencies have turned out to be more resistant against link spam in our experimentation. While many pages linking to a single URL may all belong to the same website and hence, created by the same domain owner, these are counted only once in our system.

To compute the **relevance score** $rel(v, a)$ this frequency score is normalized based on the maximum among all $v \in V$ and all $a \in A$, which results in numbers between 0 and 1. Finally, we introduce a multiplicative factor γ to get positive scores along with a logarithmic function to dampen the differences:

$$rel(v, a) = log(\frac{freq(v, a)}{\max_{v \in V, a \in A} freq(v, a)} \cdot \gamma)$$

This score is used to **boost the textual relevance** of the query, i.e., a double boost means a match is twice as important. The textual relevance is computed among all anchor texts for a page that fall into the same relevance class $\varphi = relc(v, a)$, which we consider the floor of the relevance score, i.e., the greatest integer less than or equal to this score:

$$relc(v, a) = \lfloor rel(v, a) \rfloor$$

For the boosting we multiply the relevance class score with the logarithm of the maximum score to account for time intervals with low overall frequencies. Since it is *easier* for a page to receive a high relevance score if only very few pages are archived in that time period and potentially link to a page, we want to reward those hits for a query that receive a high relevance score in a time period with a larger number of pages in the archive:

$$boost(\varphi) = \varphi \cdot log(\max_{v \in V, a \in A} freq(v, a))$$

The **result ranking** is then computed based on the textual relevance scores retrieved from our index (s. Sec 4.3) for all relevance classes φ and boosted according to the boosting score as defined above. The final score is the linear combination of all boosted scores.

4.3 Index Construction and Retrieval

The indexes for our Tempas (v2) system have been computed according to the models described above on the German Web archive from 1996 to 2013, which was collected and generously provided to us by the *Internet Archive*[3]. It comprises of more than 2 billion distinct archived webpages under the German `.de` top-level domain. These link to a total of 26, 443, 384, 902 URLs, not only under `.de`. After filtering malformed / invalid URLs as well as those that are very infrequently linked, i.e., $rel(v, a)$ is smaller than 1 for any anchor text a that links to the page represented by node v, we are left with 319,574,156 URLs that go into our index. The maximum frequencies of links from distinct hosts to distinct destinations with

[3] http://archive.org

distinct anchor texts, that are used for normalizing the relevance scores as well as for boosting, are listed in Table 1.

Table 1: Maximum frequencies per year

Year	Count	Year	Count	Year	Count
1996	17	2002	3109	2008	28764
1997	166	2003	16222	2009	30132
1998	92	2004	225066	2010	49658
1999	120	2005	42378	2011	64692
2000	128	2006	85691	2012	87444
2001	955	2007	111561	2013	48871

Tempas (v2) is implemented using *Elastic Search*[4] (ES). ES creates a separate **full-text index** for each indexed field in its **schema**. We defined this schema such that a field of a document, i.e., a page / URL, represents a relevance class φ in one time interval. We used a yearly granularity for building our indexes, so one time interval represents a year $y \in \{1996, 1997, \ldots, 2013\}$: $[t_a, t_b]$ = [y/01/01 - 00:00:00, y/12/31 - 12:59:59]. For each of these time intervals we extracted the link list $L_{\texttt{emergence}}$ and corresponding temporal model (s. Sec. 4.1). To compute the relevance scores (s. Sec. 4.2) we set the parameter $\gamma = 10,000$, i.e., we consider four decimal places of the normalized frequencies, and used basis 3 for the logarithm. This results in relevance classes between 0 and 8. Finally, the anchor texts $a \in A$ that describe any of the links to a URL represented by a node $v \in V$ form the document of the corresponding webpage, with a indexed in the field of its relevance class $relc(v, a)$, e.g.:

```
{
  "url": "http://.../wiki/World_Wide_Web",
  "years": {
    "2013": {
      "8": ["World Wide Web", "WWW", ...],
      "7": ["internet", ...]
      ...
    },
    "2012": {
      ...
    },
    ...
  }
}
```

Depending on the selected time period the corresponding fields are queried. Textual relevance is computed by ES among all anchor texts in these field based on a vector space model using a variant of `tf-idf`, i.e., a combination of the *term frequency*, *inverse document frequency* and *field-length norm*[5]. This is boosted and averaged as defined above and the results are ranked accordingly. The title and snippets shown in Figure 1 are generated from matching anchor texts sorted by the boost values of the fields that the anchor text appears in as returned by ES's *highlight* feature. The same order is used for the years listed below each search result, with the first year being selected as the main year linked by the title.

[4]http://www.elastic.co

[5]https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html

5 EMPIRICAL EVALUATION

In the following we give a qualitative evaluation of the Tempas (v2) system. Before we will look at some example queries and analyze the results returned by Tempas for these queries in more detail, we discuss our general observations of the system in Section 5.1.

5.1 General Discussion

In our current Tempas version not all of the returned search results are available in the underlying Web archive, as that is not checked when the indices are built. Instead, we provide a feature to check search results for presence in the Internet Archive's *Wayback Machine* at the displayed years after they have been returned to the user and hide them in case they are not archived. In the following we will ignore this and discuss all results returned by Tempas, regardless of whether they are archived or not.

Anchor text search: Anchor texts have special characteristics and should not be treated as running texts as we discussed in Section 2.2. We found our result rankings often to be highly satisfactory for queries taking those characteristics into account, but less useful when formulated differently. For instance, the name of a website typically yields what is expected, while the topic does not. E.g., the query *google* results in `google.com` and `google.de` at the top ranks, however, the query *search engine* does not even return these URLs on the first page. A reason for this is that famous websites are usually linked by their name instead of a description, while the descriptive terms may be part of another site's name, which is then ranked higher, such as `searchenginewatch.com`.

This meets our objective of finding suitable entry points into a Web archive, such as a specific page even if the **URL changes over time**. However, we do not cover the opposite case where the name of a website has changed, and possibly the URL too. For such renamings one would like to find the former URL under the current name, which would require a deeper evolution analysis of the Web graph, which is out of the scope of this work.

Another assumption is that pages are sufficiently frequently linked to. This is not always the case especially in the earlier periods of our Web archive due to the limited number of available pages (cp. Table 1). As a result, the performance of `Tempas` is better for queries at query intervals starting from around the mid 2000s.

German Web dataset: Although the index of Tempas was build from a German Web archive (s. Sec 4.3), it is not limited to webpages from the German Web but can also find pages under different top-level domains that are linked from a page under `.de`. For multi-lingual websites, such as *Wikipedia*, the German versions are typically preferred, as these are more frequently linked from other German websites and therefore receive higher frequency scores (s. Sec 4.2).

We observed that English query terms or non-German entities work quite well in many cases and return the expected results, but overall result in less hits, e.g., *obama* has only 724 results, while *merkel* has 11,913. This is usually not critical, as similar to what happens with popular search engines like Google or Bing, often only the very first hits in Tempas are relevant for a query. While

Table 2: Selected temporal hits for query '*obama*'

***obama* @ [2005, 2006]**
1. http://obama.senate.gov
2. http://de.wikipedia.org/wiki/barack_obama
***obama* @ [2005, 2007]**
1. http://myspace.com/barackobama
4. http://obama.senate.gov
5. http://youtube.com/profile?user=barackobamadotcom
***obama* @ [2008, 2013]**
1. http://barackobama.com
2. http://de.wikipedia.org/wiki/barack_obama
3. http://twitter.com/barackobama

that is typically the first page, i.e., first ten hits, in Tempas we found that very often only the first one to five results are subjectively very relevant to the query. This can be partially explained by the diversification features of the big, multi-purpose search engines, as well as their goal to meet various kinds of information needs as opposed to the more focused navigational needs we are adressing. Thus, a general query issued to search engines is often multi-faceted and the sought information is scattered among multiple pages, while navigational queries on Google or Bing are usually answered by only a few hits.

Temporal granularity: By temporally searching Tempas, i.e. entering both a textual query and a time interval, we found that the quality of results is much lower when only single years are selected as opposed to selecting a range of consecutive years. Even though our indices are build on a yearly basis (with separate fields for each relevance class), it appears that only combining multiple indices leads to the expected results. By further analyzing this issue, we found that more famous pages are permanently linked over time but often do not show peaks for single years like less popular pages sometimes do. Averaging over multiple years results in a smoothing of these peaks and drops the subjectively less relevant results below these temporally highly frequent hits. For that reason, all example queries shown in Section 5.2 are issues for periods of multiple years instead of single years.

5.2 Example Queries

Let us now discuss a few example queries that we consider to be potentially interesting for the problem of temporally navigational queries in Web archives (s. Sec. 2): *Barack Obama*, *Angela Merkel*, *European Union*, *Creative Common License* and *Wikipedia*. All of them feature temporal characteristics, stressing different aspects.

Barack Obama. One of those entities popular all around the world is the US president. During the later times of our dataset this was *Barack Obama*. Therefore, he will serve as our first example query. Figure 1 shows the results for the time from when he became Senator of Illinois in 2005 until 2012 when he was re-elected as president. Although in this case the query is only his last name *obama*, we receive hits solely for *Barack Obama* as he is more prominently linked on the German Web than for example his wife *Michelle Obama* and search result diversification features are not implemented in our retrieval model.

Table 3: Selected temporal hits for query '*merkel*' and '*angela merkel*'

***merkel* @ [2000, 2004]**
1. http://merkel.de *(university bookstore Merkel)*
2. http://angela-merkel.de
***angela merkel* @ [2000, 2004]**
1. http://angela-merkel.de
2. http://cdu.de/idx-merkel.htm
3. http://cdu.de/ueber-uns/buvo/pv/pv.htm
***merkel* @ [2005, 2010]**
1. http://angela-merkel.de
2. http://de.wikipedia.org/wiki/angela_merkel
***merkel* @ [2010, 2013]**
1. http://angela-merkel.de
2. http://facebook.com/angelamerkel?...
3. http://de.wikipedia.org/wiki/angela_merkel
4. http://twitter.com/search?q=%23merkel

When searching this long time frame of eight years, Tempas finds the overall most prominent authority websites of Barack Obama in these years, as expected: 1. his official website, 2. his Wikipedia article, 3. his Twitter account. More temporally sensitive results are retrieved when meaningful time frames of Barack Obama are queried, as shown in Table 2. For instance, in 2005 / 2006, i.e., Obama's first two years as senator of Illinois, his senate page is the top hit, followed by his Wikipedia article. By extending the time interval to 2007, that page gets pushed to rank 4, caused by the rise of social media with his *Myspace* page taking the lead and his *YouTube* profile on rank 5. In between are German news articles reporting about his run for president (not shown in Table 2). Starting from when Obama was elected president in 2008 we get the same results as discussed above, including his official website and Twitter replacing Myspace as his main social media profile. Before 2005 there are no hits for Barack Obama at all, because he was very famous in Germany and therefore, not sufficiently linked.

Angela Merkel. An equally famous politician, especially in Germany, is the German chancellor *Angela Merkel*. However, it is interesting to search only her last name before 2005. In contrast to *Obama*, which always referred to Barack Obama and was not relevant at all before, *Merkel* was the name of a university bookstore, which received even more links from 2000 to 2004 than the later chancellor, as shown in Table 3. Unfortunately, this website is not present in the archive and the domain is used differently today.

Therefore, to query for the person Angela Merkel, her full name should be used in earlier years. After 2004 this does not make a difference anymore. Angela Merkel exhibits a similar evolution as the US president, which can be observed through Tempas as well. Before she was elected chancellor in 2005 she became the leader of her party *CDU* in 2000. Therefore, next to her official website, pages on the party's site are among the top hits. From 2005, with her election, also her Wikipedia page become more popular. Later, in 2010, her social media profiles on Facebook and Twitter began to gain popularity.

Table 4: Selected temporal hits for query '*european union*'

***european union* @ [1996, 2005]**
1. http://europa.eu.int
***european union* @ [2005, 2013]**
1. http://en.wikipedia.org/wiki/european_union
2. http://europa.eu
3. http://europa.eu.int

European Union. Like many international organizations, the European Union's official website was located under the top-level domain .int: http://europa.eu.int. In 2005 they received their own top-level domain .eu: http://europa.eu. Today the .int URL does not exist anymore and the .eu one replaced it completely. This evolution is reflected by the queries shown in Table 4. In addition, as for most famous entities, their Wikipedia article has become one the most important resources about the EU.

This is a classical example where looking up a website in a Web archive is difficult as we need to be aware of the former URL that was active at the time of interest. Without a system like Tempas the best bet would be the current URL of the EU, which did not exist prior to 2005.

Creative Commons License. *Creative Commons* (CC) is one the most popular copyright and open content licenses. Since the inception of the CC organization in 2001 and the release of the first version in 2002, there have been three updates until its current version 4.0 was released in 2013[6].

The different variants of the CC license, e.g., *BY-NC-SA*, *BY-NC-ND*, ..., are used by many projects on the Web. As they are commonly linked under the name *Creative Commons License*, their version history can be traced through the Tempas search results, shown in Table 5. At any time the query leads to the current version of the license. Even though the URLs change over time, the query together with a timespan can be considered a temporal reference.

Wikipedia. Today Wikipedia is widely known under its domain wikipedia.org or corresponding language versions, e.g., de.wikipedia.org for the German version of Wikipedia. However, when it was launched in 2001, its domain was under .com and moved to .org one year later. Today, the .com domain and subdomains forward to their .org counterparts and no one is aware of the old URLs anymore. Without this information, it is impossible to look up the early website of Wikipedia in a Web archive. Again, this is revealed by the search results in Tempas, shown in Table 6, which make such a lookup very easy.

[6]https://wiki.creativecommons.org/wiki/License_Versions

Table 5: Selected temporal hits for query '*creative commons license*'

***creative commons license* @ [2002, 2003]**
1. http://creativecommons.org/licenses/by-nc-sa/1.0
2. http://creativecommons.org/licenses/by-nd-nc/1.0
***creative commons license* @ [2004, 2006]**
1. http://creativecommons.org/licenses/by-nc-sa/2.0
2. http://creativecommons.org/licenses/by-nc-nd/2.0
***creative commons license* @ [2007, 2013]**
1. http://creativecommons.org/licenses/by/2.5
2. http://creativecommons.org/licenses/by/3.0
3. http://creativecommons.org/licenses/by-nc-sa/3.0

Table 6: Selected temporal hits for query '*wikipedia*'

***wikipedia* @ [2001, 2002]**
1. http://de.wikipedia.com/wiki.cgi?wikipedia_willkommen
2. http://wikipedia.com
3. http://de.wikipedia.com
***wikipedia* @ [2003, 2013]**
1. http://de.wikipedia.org
2. http://de.wikipedia.org/wiki/hauptseite
3. http://wikipedia.de
3. http://wikipedia.org

6 DATA ANALYSIS

Besides manual exploration of Web archives in order to look at webpages from previous times, finding the right entry points into such an archive is crucial for data analysis tasks. Due to the vast sizes of Web archives in the order of hundreds of terabytes or even petabytes, scanning all pages is impossible. Temporal Web search capabilities such as provided by Tempas help to find the right entry points over time, which allows for temporal analyses. From these pages the archive can be scanned further. Although this does not guarantee a full coverage of all interesting contents, it is likely that we find relevant pages and thus, constituting an efficient way to create a representative sample supporting significant analysis results. As an example, we studied the evolution of restaurant prices in Germany during the time when the Euro was introduced as Europe's new currency.

6.1 ArchiveSpark

ArchiveSpark was presented in Holzmann et al. [15] and is available open source[7]. This distributed data computing framework for efficient Web archive processing and analysis exploits the fact that archive data is commonly organized with corresponding metadata in order to randomly access archived resources. The *Wayback Machine* uses these metadata records to locate a archived webpages for a given URL and timestamp as well as its embedded resources, such as images and scripts, in the Web archive. In ArchiveSpark these metadata records are incorporated as surrogate of an actual Web archive collection. Because of their small size and easily parsable structure this enables much more efficient loading and processing compared to the full records. After performing the necessary operations on the available metadata, like filtering, sorting and grouping based on URLs, status codes and mime types, the actual content is seamlessly integrated in a so-called *enrichment* step. During this phase, third-party libraries can be applied to extract or derive additional information.

[7]https://github.com/helgeho/ArchiveSpark

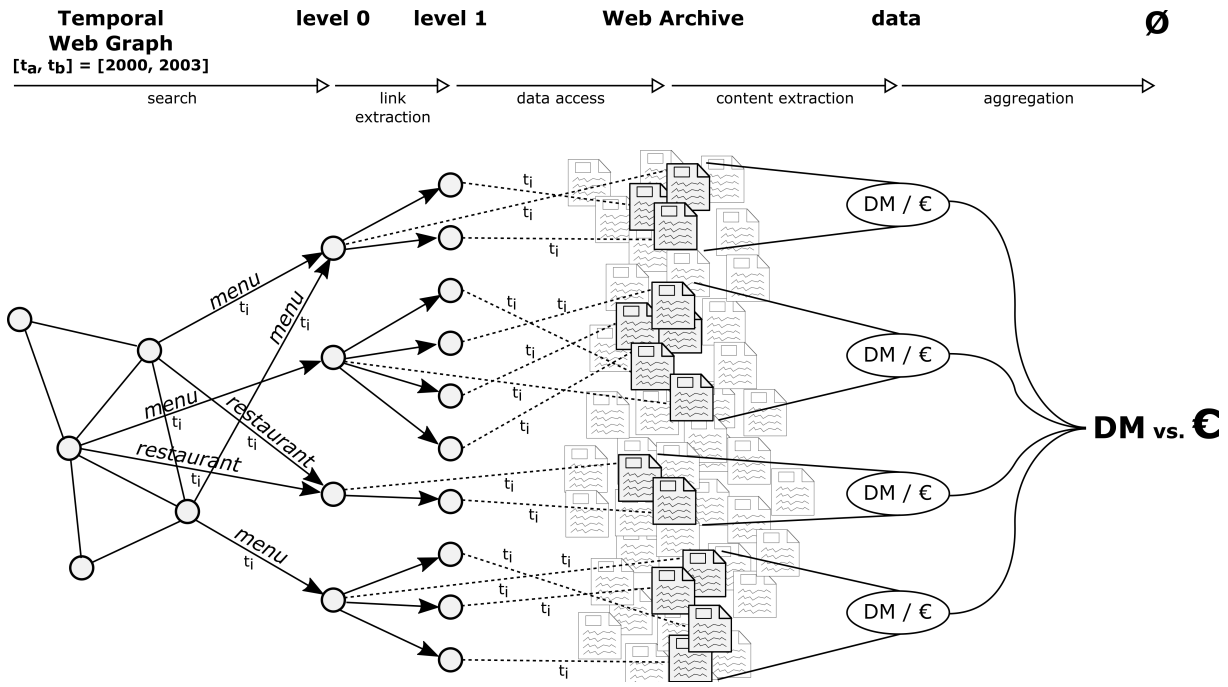

Figure 2: Data analysis pipeline from Web archive search on temporal anchor texts over selective Web archive data access and extraction to the result as shown in Section 6.2: Restaurant prices increased when the Euro replaced the DM as the currency in Germany.

We later generalized this approach to work with any kind of metadata as well as various data sources. For instance, metadata can be loaded from a database system or search engine now, and the data records can be integrated from remote services through HTTP instead of a local file store, for example directly from the *Wayback Machine*. In addition to that, we also added support for arbitrary new data types other than Web archives. This allows for very flexible data analysis and *distant reading* workflows. In contrast to *close reading*, which refers to looking at every record individually in a manual fashion, *distant reading* refers to the analysis of larger collections of documents through aggregations and statistical methods.

For the work described in this paper, we integrate Tempas as an alternative metadata provider, while the archived webpages corresponding to the search results are loaded from the *Wayback Machine*. Hence, a pre-filtering of the Internet Archive's Web archive is performed by keyword as well as time through the Tempas index, before content is loaded for returned hits. The logic for this data loading and integration step is defined in separate modules for ArchiveSpark, called data specifications (dataspecs). The dataspec used in this experiment has been published under *Tempas2ArchiveSpark*[8] together with the code that describes the pipeline for this particular analysis as illustrated in Figure 2. As metadata and data records are loaded remotely, the study can be repeated or similar studies can be conducted by anyone even without having a suitable dataset available.

[8]https://github.com/helgeho/Tempas2ArchiveSpark

6.2 Example Study

In this small example study we utilize ArchiveSpark to analyze menus of German restaurants at the time when the Euro was introduced as the new currency in Europe and replaced the former German currency *Deutsche Mark (DM)* in 2001/2002. According to many voices in Germany this resulted in increased prices particularly in restaurants. In 2011 the German federal office of statistics published a report in which they studied the effect of the Euro in various areas and categories [11]. According to that study, restaurant prices increased around the time of currency reform by about 4% on average based on more than 700 examined restaurants. However, while some restaurants even reduced their prices, others increased them by up to 20% to 40%.

To get entry points into the archive for our analysis, we queried Tempas for the keywords '*speisekarte*', which is the German word for menu, as well as '*restaurant*' to find menus after one additional hop, in case it is not linked with this anchor text. Both queries were issued for the time period from 2000 to 2003 and we considered the first 10 pages with 100 results per page, hence 1,000 hits each. As some of the returned URLs were found in multiple years, we started off with 3,567 hits of URL / timestamp pairs for which we integrated contents from the *Wayback Machine*. By manually investigating a few of the result pages on Tempas, we found that they often do not show the prices directly, but link to subpages, such as for starters or main courses. Therefore, we extracted all links under the same hostname and fetched the contents for those URLs as well, which resulted in additional 6,028 pages. From each of these pages we extracted the amounts of money mentioned on the pages as floating

point numbers prefixed or suffixed with either *DM*, *Euro*, *EUR* or the Euro symbol €. To verify that we were actually dealing with restaurant menus, we required the pages to mention the term *speise*, which is the German word for meal / dish, and to contain at least three prices. This filtering left us with menus for 49 entry point URLs.

After converting DM to Euro amounts with an exchange rate of 1 EUR = 1.95583 DM, we averaged the amount per currency and URL as well as among URLs. Finally, we ended up with an average of 9.58 Euros and 15.24 DM = 7.79 Euros. That is an increase of 23%. Although this is higher than the average 4% as reported by the study mentioned above, it is in the range between 20% and 40% that was found in the statistics report. We can therefore consider it a reasonable result, although not significant due to the limited number of analyzed menus. The experiment does show, however, how Web archives can be used for studying social issues with temporal anchor texts serving as entry points to access and analyze these extremely big collection, which are otherwise almost impossible to process. With 49 menus out of the initially 3,567 hits and 6,083 analyzed URLs, studying Web archives often is like finding a needle in a haystack. This shows the need of effective ways to identify entry points for both manual search / *close reading* as well as *distant reading* tasks such as the study presented here. Anchor texts have been proven suitable as starting points for exploring Web archives. Thus, given a better domain knowledge of the studied subject and by investing some additional effort, we should be able to produce interesting and significant results with this approach.

7 CONCLUSION AND OUTLOOK

In this paper we propose an alternative to full-text indexing for navigational query intents on Web archives. Full-text indexing over Web archives, which are massive collections (typically in the order of TB and PB), is computationally intensive. To lower the entry barrier for search and recognizing that most interesting query intents are navigational in nature, we can to a large extent satisfy them using surrogates such as anchor texts. The results of such navigational queries can then be used as hooks into more detailed temporal exploration of the Web archive.

Our Tempas system supports these queries and provides a fully functional indexing and retrieval infrastructure for Web Archives. It is based on anchor texts and the underlying Web graph structure. Tempas makes it easy to search for entities in a Web archive and through its results provides useful hooks for further exploration of these entities and aspects related to them.

We also note that *TREC* (Text REtrieval Conference) style empirical evaluations of such temporal navigational tasks are difficult to accomplish, even in a controlled user setup. One promising direction for future work is to improve the effectiveness and soundness of Tempas by building on query logs. We will evaluate our retrieval models on our logs by characterizing specific navigational intents in Web archive search. Further, we plan to outline other informational intents and adapt our retrieval models accordingly.

REFERENCES

[1] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. 2007. On the Value of Temporal Information in Information Retrieval. *SIGIR Forum* 41, 2 (Dec. 2007), 35–41. DOI: https://doi.org/10.1145/1328964.1328968

[2] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. 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*.

[3] Avishek Anand, Srikanta Bedathur, Klaus Berberich, and Ralf Schenkel. 2012. Index maintenance for time-travel text search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in Information Retrieval*.

[4] Klaus Berberich and Srikanta Bedathur. 2013. Temporal diversification of search results. In *SIGIR 2013 Workshop on Time-aware Information Access (TAIA 2013)*.

[5] Klaus Berberich, Srikanta Bedathur, Omar Alonso, and Gerhard Weikum. 2010. A Language Modeling Approach for Temporal Information Needs. In *Proceedings of the 32Nd European Conference on Advances in Information Retrieval (ECIR'2010)*. Springer-Verlag, Berlin, Heidelberg, 13–25. DOI: https://doi.org/10.1007/978-3-642-12275-0_5

[6] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[7] Ricardo Campos, GaÃćËĘåąÃĆÂťl Dias, AlÃćËĘåąÃćâĂřÅäpio M. Jorge, and Adam Jatowt. 2014. Survey of Temporal Information Retrieval and Related Applications. *ACM Comput. Surv.* 47, 2 (Aug. 2014), 15:1–15:41. DOI: https://doi.org/10.1145/2619088

[8] Miguel Costa, Daniel Gomes, Francisco Couto, and Mário Silva. 2013. A Survey of Web Archive Search Architectures. In *Proceedings of the 22nd International Conference on World Wide Web (Companion)*.

[9] Miguel Costa and Mário J Silva. 2010. Understanding the Information Needs of Web Archive Users . In *Proceedings of the 10th International Web Archiving Workshop*.

[10] Nick Craswell, David Hawking, and Stephen Robertson. 2001. Effective Site Finding using Link Anchor Information. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[11] German Federal Statistical Office (Statistisches Bundesamt, Destatis). 2011. Fast zehn Jahre Euro - Preisentwicklung vor und nach der BargeldeinfÃČÂijhrung. Article number: 5611105119004 (Decenber 2011). https://www.destatis.de/DE/Publikationen/Thematisch/Preise/Verbraucherpreise/Fast10JahreEuro5611105119004.html [Accessed: 16/03/2017].

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

[13] Wendy Hall, Jim Hendler, and Steffen Staab. 2017. A Manifesto for Web Science @10. *arXiv:1702.08291* (2017).

[14] Helge Holzmann and Avishek Anand. 2016. Tempas: Temporal Archive Search Based on Tags. In *Proceedings of the 25th International Conference Companion on World Wide Web*.

[15] Helge Holzmann, Vinay Goel, and Avishek Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*.

[16] Helge Holzmann, Wolfgang Nejdl, and Avishek Anand. 2016. 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*.

[17] Rosie Jones and Fernando Diaz. 2007. Temporal Profiles of Queries. *ACM Trans. Inf. Syst.* 25, 3 (July 2007). DOI: https://doi.org/10.1145/1247715.1247720

[18] Nattiya Kanhabua and Wolfgang Nejdl. 2014. On the Value of Temporal Anchor Texts in Wikipedia. In *SIGIR 2014 Workshop on Temporal, Social and Spatiallyaware Information Access (TAIA)*.

[19] Marijn Koolen and Jaap Kamps. 2010. The importance of anchor text for ad hoc search revisited. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 122–129.

[20] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. 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*. ACM.

[21] Paul Ogilvie and Jamie Callan. 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*. ACM.

[22] Jaspreet Singh, Wolfgang Nejdl, and Avishek Anand. 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, New York, NY, USA, 183–192. DOI: https://doi.org/10.1145/2854946.2854959