Comparing Topic Coverage in Breadth-First and Depth-First Crawls Using Anchor Texts

Thaer Samar[1(✉)], Myriam C. Traub[1], Jacco van Ossenbruggen[1], and Arjen P. de Vries[2]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
samar@cwi.nl
[2] Radboud University, Nijmegen, The Netherlands

Abstract. Web archives preserve the fast changing Web by repeatedly crawling its content. The crawling strategy has an influence on the data that is archived. We use link anchor text of two Web crawls created with different crawling strategies in order to compare their coverage of past popular topics. One of our crawls was collected by the National Library of the Netherlands (*KB*) using a *depth-first* strategy on manually selected websites from the *.nl* domain, with the goal to crawl websites as completes as possible. The second crawl was collected by the *Common Crawl* foundation using a *breadth-first* strategy on the entire Web, this strategy focuses on discovering as many links as possible. The two crawls differ in their scope of coverage, while the *KB* dataset covers mainly the Dutch domain, the *Common Crawl* dataset covers websites from the entire Web. Therefore, we used three different sources to identify topics that were popular on the Web; both at the global level (entire Web) and at the national level (*.nl* domain): Google Trends, *WikiStats*, and queries collected from users of the Dutch historic newspaper archive. The two crawls are different in terms of their size, number of included websites and domains. To allow fair comparison between the two crawls, we created sub-collections from the *Common Crawl* dataset based on the *.nl* domain and the *KB* seeds. Using simple exact string matching between anchor texts and popular topics from the three different sources, we found that the *breadth-first* crawl covered more topics than the *depth-first* crawl. Surprisingly, this is not limited to popular topics from the entire Web but also applies to topics that were popular in the *.nl* domain.

1 Introduction

The World Wide Web offers rich means for its users to publish, share, create, discuss, collaborate and even earn a living. Web data, however, is surprisingly volatile. Ntoulas et al. found that 80 % of the Web pages disappear within one year [21]. In order to preserve (at least a fraction of) this data, many national libraries and archives have set up Web archiving initiatives. However, it is impossible to archive the entire Web due its increasing size, and the dynamic and ephemeral nature of its content. Therefore, institutes have to make decisions on the websites to be included in the archive, the crawling frequency, and the crawling strategy. One strategy is to crawl a manually selected set of websites (called

N. Fuhr et al. (Eds.): TPDL 2016, LNCS 9819, pp. 133–146, 2016.
DOI: 10.1007/978-3-319-43997-6_11

the crawler's *seeds*) and to harvest these websites in depth (*depth-first* crawl). Another strategy automatically crawls as many websites as possible (usually the national domains), but not in depth (*breadth-first* crawl). Both crawling strategies result in incomplete crawls, as both strategies exclude websites. *Depth-first* ignores websites outside the seeds list, and *breadth-first* archives websites incompletely as it does not follow the links to sub-pages. On top of the content of websites, Web archives also preserve information registered by crawlers such as the date of the crawl, the timestamp of the last modification of the page, the MIME-type, and information that can be derived from the archived pages, for example hyperlinks and anchor texts.

Web archives preserve content which may no longer be available on the Web. We explore how well the collections resulting from different crawling strategies cover content related to topics that were in the focus of Web users in a particular time period. We perform our analysis on two Web archive collections harvested in 2014 using different crawling strategies. The first collection is a crawl from the entire Web harvested by the *Common Crawl* foundation using the *breadth-first* crawling strategy. The second collection is the Dutch Web archive collection preserved by the National Library of The Netherlands[1] (*KB*). Here, the *depth-first* strategy was applied to manually selected websites (KB *seeds*) related to the Dutch history, social, and culture heritage. We propose to use anchor text specified in hyperlinks extracted from the two collections to investigate their coverage of the topics that were of interest to users in the same year (2014). Users of Web search engines express their information needs by issuing queries. User queries collected from major search engines would be the best record of popular topics. However, these queries were not available for us. Therefore, we used different sources as indicators of the trending topics on the Web at the time when the crawls we used were collected (2014). Since our crawls originate from the entire Web (*Common Crawl* crawl) and from the Dutch domain (*KB* crawl), we looked for popular topics both worldwide and on the national level. Our first source is Google Trends. Google provides a list of the top searched terms on the entire Web, and in the given country domain. The second source is the *WikiStats* which aggregates page views of Wikipedia pages. Again we focus on all Wikipedia pages (in all languages), and the pages written in Dutch. Finally, we use queries collected from users searching the Dutch digital newspaper archive via the *KB*'s Delpher[2] interface. These are three heterogeneous sources, the first and the third are real user queries, the second consists of Wikipedia titles associated with their frequency of views over time. We use these sources to represent users interests, which we refer to as topics.

2 Related Work

The structure of the Web graph is defined by its links which consist of a source URL, a destination URL and an anchor text describing the link. Several studies

[1] www.kb.nl.
[2] www.delpher.nl.

explored the structure of the Web graph based on crawls from the entire Web [2,19,23]. The link structure was used to study the evolution and the structure of Web crawls of national domains [1,5,25]. Properties of the Web graph, such as the PageRank and out-degree, were used to propose algorithms for seed selection of Web crawlers [24], and to improve the effectiveness of Web search. An empirical study showed that anchor texts exhibit characteristics similar to real user queries [7]. They also showed that anchor texts are similar to titles of webpages. This is based on the observation that titles can be used as an approximation of queries [10]. Anchor texts enrich the representation of a Web page's content to improve Web search effectiveness [4,6,8,11,14–16,18]. Kanhabua and Nejdl studied the evolution of anchor texts extracted from the edit history of Wikipedia [12]. They found that anchor texts with temporal information can be candidates for capturing and tracing the entity evolution.

The link structure and anchor texts constructed from the archived pages play an important role in assessing the completeness of Web archives. It is impossible to archive the entire Web due its increasing size and evolving content. Therefore, the archived parts of the Web are incomplete. Web archiving theorists acknowledge that the archived parts of the Web is both incomplete and over complete [3,17]. It is impossible to crawl the Web in a way that all websites and pages are included, for example the *depth-first* crawling strategy excludes websites not in the seeds list, and the *breadth-first* strategy does not crawl discovered websites in depth. Thus both strategies result in an incomplete crawl. On the other hand, Web archives are over complete, as they do not only contain the raw content but also metadata, such as the MIME-type and the date of the crawling time. More over, information that can be constructed from the archived pages, for example, the link structure and anchor texts. The wealth of information available in the Web archives has been discussed in [22]. Links and anchor texts can be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [13] computed lexical signatures of lost webpages, using the top n words of link anchors, and used these and other methods to retrieve alternative URLs for lost webpages. The use of the link structure and anchor texts to uncover and reconstruct target pages that were not archived was studied in [9], based on a *depth-first* crawl of manually selected websites. They used the link structure extracted from archived Web pages to uncover target URLs that were not archived. Links extracted from the archived pages contain evidence of the existence of unarchived target URLs. Based on the link evidence, Huurdeman et al. found that the number of unarchived Web pages is roughly as high as the number of the archived Web pages. Then, they used link evidence to reconstruct basic representations of target URLs. This evidence includes the aggregated anchor text, crawl date, and source URLs.

3 Setup

In this section we describe the two crawls on which we base our analysis. Then, we introduce the pipeline of extracting hyperlinks and anchor texts from the

crawls. After that, we discus how we zoom in the link structure of *Common Crawl* dataset to generate subsets based on filters synthesized from the *KB* dataset in order to allow a fair comparison. Finally, we introduce the sources that we used to identify popular topics.

3.1 Data

***KB* Dataset:** The *KB* archives a pre-selected set of more than 10,000 websites (*seeds*) with the aim to crawl these websites as complete as possible. The selection is based on categories related to Dutch historical, social and cultural heritage. The websites are categorized by curators of the *KB* using the *UNESCO* classification code. The crawling frequency varies between yearly, biannually, quarterly, and daily, for example news agency websites (such as nu.nl). Our snapshot of the Dutch Web archive between February 2009 and May 2015 consists of 150,557 files in *ARC*[3] format, which contain aggregated web content. Each ARC file contains multiple Web objects, in total, 251,591,618 objects exist in the ARC files. We focus on data crawled in 2014, as we have only access to *Common Crawl* pages crawled in that year.

***Common Crawl* Dataset:** Common Crawl[4] is a non-profit organization aiming to build and maintain an openly accessible repository of archived Web crawls. We use the crawl collected in March 2014, which consists of 2.8 billion Web pages.

3.2 Anchor Links Extraction

From the two datasets, we extracted hyperlinks from the archived objects with *text/html* as MIME-type. For that we used MapReduce to process all archived web objects contained in the archive's ARC files. During the processing of the archived objects, we used JSoup[5] to extract anchor links (*a*) in order to be able to focus on links between textual content. For each anchor link, we kept the URL of the page that contains the link *source*, the URL of the *target*, and the anchor text specified in the link. Based on the crawl-date, we keep pages crawled in 2014. The anchor texts pointing to the target pages were used in that year. Depending on the source URL and target URL, the link can be an internal link or external link. An internal link has the same domain-name for both source and target (intra-domain), while for an external link the domain-name of the source URL is different from that of the target URL (an inter-domain link). We limit our analysis to the external links as it is of more interest to look into links between different hosts (sites). By discarding internal links we exclude links from menus and other non-content information. The exact URLs may change frequently, while we are really interested in anchor text used by one site to link to another site. Therefore, we replace both the source URL and the target URL by their hosts (site name) before we analyze the data. This pre-processing can be

[3] http://archive.org/web/researcher/ArcFileFormat.php.
[4] http://commoncrawl.org/.
[5] http://jsoup.org/.

viewed as a process to smooth the graph structure to maintain the most salient information. We deduplicate the links based on their values for source, target, and anchor text for *KB* dataset (*Common Crawl* dataset consists of one crawl). This prevents the differences in crawling frequency to influence our analysis. At the end of this pipeline, we keep (*sourceHost, targetHost, anchorText*). We refer to the links extracted from the *KB* dataset as KB_{links}, and links extracted from the *Common Crawl* dataset as CC_{links}.

3.3 Link Subsets from *Common Crawl*

The two crawls differ in terms of size, number of crawled websites and web pages, and the domains of the crawled websites. These differences are reflected in the extracted links structure. The number of links extracted from the *Common Crawl* dataset is $559x$ times larger than the number of *KB* links, (see Table 1). Therefore, in addition to performing one-to-one comparison between the two crawls, we generate subsets from the CC_{links} by mapping it to the Dutch domain in two different ways: First, we focused on pages that originate from the *.nl* domain. This was done by keeping only links from the CC_{links} whose *source hosts* are from the *.nl* domain. We refer to the set as $CC_{links} \cap .nl$. Second, the *KB* crawl is based on a list of manually selected websites (*KB* seeds). We used the *KB* seeds to generate another subset of links from the CC_{links}, based on links with *source hosts* from the *KB* seeds. We refer to this subset as $(CC_{links} \cap KB_{seeds})$. Finally, we investigate the impact of anchor texts associated with targets of links in the *KB* dataset on the topic coverage of the CC_{links}. In order to do that, we dropped links from CC_{links} in which the *target hosts* are targets of links in the KB_{links}. We refer to this set of filtered links as $(CC_{links} \setminus KB_{targets})$. These subsets allow us to investigate whether the *KB* seeds list comprises the part of the Dutch Web that is essential from the perspective of topic coverage, or whether a broader and less deep crawl would still contain sufficient information.

Table 1. Number of unique links in each dataset.

Links dataset	Num. of links
KB_{links}	$3,033,855$
CC_{links}	$1,696,102,933$
$CC_{links} \cap .nl$	$5,128,501$
$CC_{links} \cap KB_{seeds}$	$2,629,765$
$CC_{links} \setminus KB_{targets}$	$1,174,261,413$

3.4 Sources of Topics

Our assumption is that the *Common Crawl* (a *breadth-first* crawl) covers more global topics, and that the *KB* (a *depth-first* crawl) covers more topics from the *.nl* domain. In order to validate our assumption, we use different sources to

identify which topics were popular on the Web, topics that attracted attention in the entire Web (global) and topics that were only picked up in the *.nl* domain.

Google Trends. Google Trends[6] is a public resource, which lists the most searched queries in the global Web or per country in a given year. For our analysis, we use global trends and the trends searched in the Netherlands in 2014 (the year of our crawls).

Wikipedia Page Views Statistics. The *WikiStats* dataset [20] consists of the number of views for Wikipedia pages. The goal is to show how the interest in Wikipedia pages changes over time, and allows comparison between chosen Wikipedia pages. The views are aggregated from the *Page view statistics for Wikimedia projects*[7], which aggregates the request history of articles from Wikimedia projects[8]. For each page, this project provides the page title, the number of requests (on hourly basis), the language in which the page is written, and the name of the project. The *WikiStats* data set consists of the weekly views of Wikipedia pages in the period from January 2008 to January 2015. We select Wikipedia pages viewed in 2014, then aggregate their page view counts, and those pages viewed more than 1,000 times. Finally, we created two datasets: the first contains all Wikipedia pages from all domains (*WikiStats* global), and the second contains only pages written in Dutch language (*WikiStats .nl*).

User Queries. Under conditions of strict confidentiality, the *KB* made anonymized user logs available, collected between March 2015 and December 2015 from users visiting the public digital newspaper archive on a webservice called Delpher. The collection consists of newspapers articles published in the Netherlands since 1618. The data set made available consists of 10 million OCRed newspaper pages in DIDL XML format[9].

Sources Summary. We processed all topics from the sources mentioned with the same pre-processing pipeline, which includes lower casing, stopwords (English and Dutch) removal, and the removal of short terms with a length of less than three characters. The resulting dataset statistics are summarized in Table 2.

Table 2. Number of unique topics per source.

Topics source	Count
Google global trends	84
Google *.nl* trends	68
WikiStats global	3,293,749
WikiStats .nl	99,396
Real queries	1,580,386

[6] http://www.google.com/trends/topcharts?hl=en#date=2014&geo=.
[7] http://dumps.wikimedia.org/other/pagecounts-raw/.
[8] These projects are: wikibooks, wiktionary, wikinews, wikivoyage, wikiquote, wikisource, wikiversity, and wikipedia.
[9] http://www.xml.com/pub/a/2001/05/30/didl.html.

4 Analysis

Using anchor texts we investigate the coverage of topics in *Common Crawl* (a *breadth-first* crawl), and *KB* (a *depth-first* crawl). Since anchor texts usually describe target pages, we first provide a deep analysis of them with regard to their hosts and top-level domains (*TLDs*). Then, we present a detailed analysis of anchor texts associated with hyperlinks. Finally, we investigate the anchor texts coverage of topics from the three sources described in Sect. 3.4.

4.1 Target Pages

For all link datasets, the number of unique hosts in the target pages is higher than the number of unique hosts of source pages (see Table 3). In KB_{links}, the number of unique target hosts is 442,296, which is 14 times higher than the number of source hosts (31,829). In CC_{links}, the ratio between the target hosts (30,416,854) and the source hosts (9,715,414) is lower, here, the number of target hosts is only 3 times higher than the number of source hosts. These numbers of source hosts and target host shows the big difference between the two dataset. However, subsets from *Common Crawl* dataset have comparable numbers. The crawling strategy clearly affects the percentage of target hosts that have been crawled. The percentage of the crawled target hosts differ between the link datasets, (see Table 3). For example, only 6.5 % of KB_{links} target hosts were crawled, whereas 23.9 % of target hosts in CC_{links} were crawled. However, both crawling strategies showed that large fractions of target hosts were not crawled, and we cannot find their raw content. This suggests that the use of target hosts, and anchor texts as a means to describe them is a valuable resource. We also looked into the overlap of target hosts between the datasets. A high percentage (71.4 %) of target hosts in KB_{links} were also targets of links in CC_{links}. The percentage of overlap decreases to 38.5 % after subsetting the *Common Crawl* dataset based on source pages from the *.nl* domain ($CC_{links} \cap .nl$), and decreases to 24.2 % after projecting the *KB* seeds on CC_{links} ($CC_{links} \cap KB_{seeds}$). Recall, that there is no overlap between KB_{links} and $CC_{links} \setminus KB_{targets}$, because all links whose target hosts are the same as the target hosts in KB_{links} were dropped from CC_{links}. In terms of the source hosts not only the number of hosts is lower compared to the number of target hosts, but also the overlap between KB_{links} and the other datasets is smaller (see Table 3).

Top-level Domains. Another way of looking at the difference between the link datasets is based on the TLDs of the target pages. The TLDs represent the target domains of the crawled pages. In CC_{links}, a high percentage of links points to the pages from the *.nl* domain, and the majority (60.5 %) of the target pages are from the *.com* TLD, see Table 4. The majority of target pages (45.6 %) in KB_{links} are from the *.nl* domain, which is expected because the *KB* crawl was harvested based on websites mainly from the Dutch Web. The target pages in $CC_{links} \cap .nl$ has the same distribution of top-ranked TLDs of target pages in KB_{links}. In the distribution of TLDs for $CC_{links} \cap KB_{seeds}$, the *.com* is the

Table 3. Analysis of hosts: For both the target and source pages, we present the absolute count of unique hosts (*first row*), the fraction of hosts from KB_{links} that were found in the corresponding dataset in column header (*second row*), and the percentage of target hosts that has been crawled in each link dataset (*third row*).

	KB_{links}	CC_{links}	$CC_{links} \cap .nl$	$CC_{links} \cap KB_{seeds}$	$CC_{links} \setminus KB_{targets}$
Target hosts	442,296	30,416,854	800,957	529,962	30,100,936
	100.0 %	71.4 %	38.5 %	24.2 %	0 %
Source hosts	31,829	9,715,414	120,498	2,942	8,237,940
	100.0 %	57.8 %	28.5 %	8.5 %	42.9 %
Crawled target hosts	6.5 %	23.9 %	8.5 %	0.4 %	20.5 %

Table 4. TLDs of target pages: The count of unique TLDs, and the top-10 TLDs.

	KB_{links}	CC_{links}	$CC_{links} \cap .nl$	$CC_{links} \cap KB_{seeds}$	$CC_{links} \setminus KB_{targets}$
Count (unique)	293	456	267	268	451
	nl	com	nl	com	com
	com	org	com	org	org
	org	net	org	nl	net
	net	de	net	net	de
	de	info	de	de	info
	be	nl	be	be	nl
	eu	ru	eu	it	it
	info	it	info	ro	ru
	fr	fr	it	fr	fr
	it	pl	fr	info	pl

most prevalent TLD; 49 % of target pages belong to this domain, not all websites in the *KB* seeds were found in *Common Crawl* dataset, only 43.6 % (unique) were found. The *KB* seeds are not all from the *.nl* domain, only 88 % of the seeds belong to the *.nl* domain. The remaining seeds (12 %) belong to different TLDs: 5 % from the *.org* domain, 3.4 % from the *.com* domain, 1.2 % from the *.net* domain, 0.6 from the *.eu* domain, and 0.5 % from the *.info* domain. The distribution of the top TLDs is similar in CC_{links} and $CC_{links} \setminus KB_{targets}$. The only difference is the number of target pages per TLD, which decreases for some TLDs in $CC_{links} \setminus KB_{targets}$ compared to CC_{links}. This is caused by dropping links whose target hosts are the same as the target hosts in KB_{links}. Thus the highest relative decrease was for the *.nl* domain.

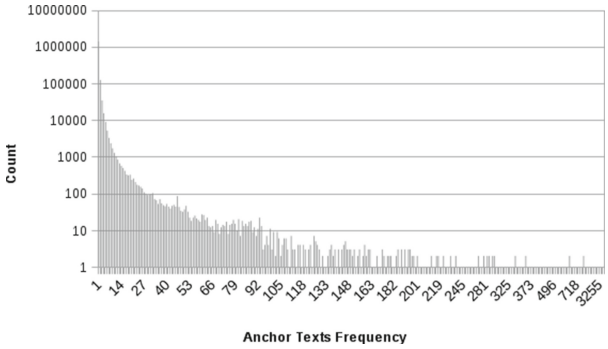

Fig. 1. Anchor texts frequency distribution of KB_{links} in log scale representation.

Table 5. Anchor texts summary: For each link dataset, we present the number of unique anchor texts, and the overlap of anchor texts between KB_{links} and the corresponding dataset. Considering all anchor texts in KB_{links} (%overlap_all), and by considering anchor texts used at least twice in KB_{links} (%overlap_GT1).

Links dataset	Count	%overlap_all	%overlap_GT1
KB_{links}	1,581,013	100.0	13.0
CC_{links}	83,920,299	23.6	49.9
$CC_{links} \cap .nl$	2,613,774	13.7	40.5
$CC_{links} \cap KB_{seeds}$	1,289,803	9.2	26.7
$CC_{links} \setminus KB_{targets}$	61,153,447	15.3	34.4

4.2 Anchor Texts

Some anchor texts are used by multiple links and the frequency of the anchor texts represents their popularity in the archive. We processed the anchor texts with the same pre-processing pipeline we used for the topics (Sect. 3.4) and computed the frequencies of all unique anchor texts for each link dataset. The number of unique anchor texts varies strongly among the datasets (see Table 5). When we compared the percentage of overlap between anchor texts in KB_{links} and all other link datasets based on exact string matching, we found that 23.6 % of the unique anchor texts in KB_{links} exist in the unique anchor texts of CC_{links}. The frequency of anchor texts in KB_{links} shows a long tail distribution (Fig. 1). A high percentage (87 %) of the anchor texts in KB_{links} occurs only once. We investigated the overlap considering only anchor texts with a frequency larger than one. This results in an increase of the percentage of overlap between KB_{links} with all datasets. We can use the frequency as threshold to focus on most popular anchor texts.

Table 6. Topic Coverage: for each link dataset, we present the absolute count and the fraction (%) of found topics in each topic source, where the fraction is the number of matched topics to the total number of topics in the corresponding source. The %*lost* under $CC_{links} \setminus KB_{targets}$ is the relative not found topics, these topics were found in CC_{links} but in $CC_{links} \setminus KB_{targets}$.

Topics source	KB_{links}		CC_{links}		$CC_{links} \cap .nl$		$CC_{links} \cap KB_{seeds}$		$CC_{links} \setminus KB_{targets}$		
	count	%	count	%	count	%	count	%	count	%	%lost
Google global trends	24	28.6	51	60.7	25	29.8	23	27.4	51	60.7	0.0
Google *.nl* trends	22	32.4	27	39.7	25	36.8	18	26.5	24	35.3	−11.1
WikiStats global	80,043	2.4	1,376,222	41.8	122,659	3.7	116,259	3.5	1,122,767	34.1	−18.4
WikiStats .nl	24,726	24.9	48,825	49.1	31,742	31.9	19,098	19.2	43,304	43.6	−11.3
Real queries	26,099	1.7	77,152	4.9	38,033	2.4	15,839	1.0	66,874	4.2	−13.3

4.3 Topic Coverage

An anchor text describes the target page with a brief text which is known to resemble user queries. Therefore analyzing the anchor texts' overlap with queries is a good proxy for assessing whether the crawls are likely to contain answers to user queries and popular topics. Not all target pages that are linked to from the crawled pages are harvested by the crawler. As mentioned earlier, Web archives are incomplete, and the advantage of anchor texts is their availability for both crawled and not crawled target pages. In order to investigate the topic coverage, we used exact string matching between pre-processed anchor texts from the five link datasets with topics from the sources (described in Sect. 3.4). Topic coverage varies among the datasets for the different sources of topics, (see Table 6). For some cases we found high coverage, for example anchor texts from CC_{links} matched 60.7 % of Google global trends and 49.1 % of the Dutch Wikipedia pages in the *WikiStats*. After sorting the anchor texts in descending order based on their frequencies, we investigated the relation between the percentage of topics covered and the frequency (popularity) of anchor texts. We report on exact string matches between the top anchor texts and both, Wikipedia titles and real user queries, considering different rank cutoffs c; $c = 1k$, $10k$, and $100k$. The percentage of matched anchor texts with *WikiStats* (global and *.nl* domain) decreases as c increases for all link datasets (see Table 8). The lowest overlap corresponds to the case when all anchor texts are used to match Wikipedia titles.

In general, the percentage of overlap between anchor texts from the different datasets and the user queries is low. For example, we found that only 1.7 % of the user queries had a match in KB_{links} when we applied exact string matching with all anchor texts. We found the highest percentage of overlap with user queries (4.9 %) for anchor texts in CC_{links} (see Table 6). When we compared the top-c anchor texts instead of the complete set of anchor texts, we found a relation between the top ranked anchor texts and the percentage of the topic coverage. A high percentage of the most frequently used anchor texts matched user queries, and the percentage of overlap decreases while the cutoff c increases, see Table 9.

Table 7. Unique Topic Coverage in KB_{links}: in comparison with topics found in other datasets. Under every link dataset x, we present the percentage of topics found in the KB but not found in x, and the percentage of topics found in x but not found in KB_{links}.

Topics source	CC_{links}		$CC_{links} \cap .nl$		$CC_{links} \cap KB_{seeds}$		$CC_{links} \setminus KB_{targets}$	
Google global trends	0.0 %	54.9 %	29.2 %	32.0 %	33.3 %	30.4 %	0.0 %	54.9 %
Google *.nl* trends	0.0 %	18.5 %	9.1 %	20.0 %	31.8 %	16.7 %	13.6 %	20.8 %
WikiStats global	6.1 %	94.5 %	46.1 %	64.8 %	56.3 %	69.9 %	12.2 %	93.7 %
WikiStats .nl	16.5 %	57.7 %	28.2 %	44.0 %	53.6 %	39.9 %	26.4 %	58.0 %
Real queries	22.3 %	73.7 %	31.4 %	53.0 %	60.4 %	34.7 %	31.8 %	73.4 %

Table 8. The fraction of top-c ranked anchor texts matching document titles in *WikiStats*, fraction in percentage (*num. matches/c*). In addition to the percentage of matching when all anchor texts are used (*num. matches/num. all anchor texts*).

Links dataset	*WikiStats* global				*WikiStats .nl* domain			
	top-1k	top-10k	top-100k	all	top-1k	top-10k	top-100k	all
KB_{links}	44.6	39.1	22.1	5.1	32.4	24.3	10.2	1.6
CC_{links}	51.3	56.9	38.2	1.6	11.5	14.4	7.4	0.1
$CC_{links} \cap .nl$	45.2	37.3	24.7	4.7	32.0	23.9	11.8	1.2
$CC_{links} \cap KB_{seeds}$	70.6	65.6	33.9	9.0	32.0	23.3	8.3	1.5
$CC_{links} \setminus KB_{targets}$	71.2	63.5	28.1	1.8	19.5	14.7	4.2	0.1

Table 9. The fraction of top-c ranked anchor texts matching user queries, same notation as in Table 8.

Links dataset	Real queries			
	top-1k	top-10k	top-100k	All
KB_{links}	26.4	23.2	9.7	1.7
CC_{links}	17.2	18.2	7.0	0.9
$CC_{links} \cap .nl$	33.5	26.0	12.8	1.5
$CC_{links} \cap KB_{seeds}$	30.0	20.5	6.9	1.2
$CC_{links} \setminus KB_{targets}$	28.5	23.0	6.9	0.1

We found the highest overlap of topics and anchor texts in CC_{links}, suggesting that the *breadth-first* crawl covers more topics than the *depth-first* crawl. This result holds for both, the global and the national (*.nl*) topics. Focusing on the Dutch part of the *Common Crawl* dataset ($CC_{links} \cap .nl$) showed that this part covers more topics than topics covered in KB_{links}. However, the comparison is based on the absolute count of found topics in each links dataset. That does not necessarily mean that all topics covered by KB_{links}, are identical with those found, for instance, in CC_{links}. For all topic sources, we analyzed the topics that

were found in KB_{links} but not in the other datasets (see Table 7). For example, we found that all Google trends (both the global and the *.nl* domain) that were found in KB_{links}, were also found in CC_{links}. On the other hand, 54.9 % of Google's global trends and 18.4 % of Google's *.nl* trends found in CC_{links} were not found in KB_{links}. Regarding the *WikiStats* dataset, not all topics found in KB_{links} were found in CC_{links}. The percentage of topics that were found in KB_{links} is higher for the Wikipedia pages from the *.nl* domain (16.5 %), while 6.1 % of Wikipedia pages (global) found in KB_{links} were also found in CC_{links}.

These results suggest that anchor texts can be used as a resource for finding topics that were popular with users from the past. The coverage of topics was higher for the most frequently used anchor texts in the crawls. Anchor texts from the *breadth-first* crawl cover more topics than the anchor texts from the *depth-first* crawl. However, some topics were only covered by the *depth-first* crawl.

5 Conclusions

We studied the influence of the crawling strategy on the coverage of topics that were of interest to users on the Web. We performed our analysis on two Web crawls created by following different crawling strategies; the *Common Crawl* dataset, (a *breadth-first* crawl) collected from the entire Web, and the *KB* dataset (a *depth-first* crawl) harvested by the *KB* based on manually selected websites). We made use of anchor texts to investigate the topic coverage in the two crawls. We extracted anchor texts from the raw content of documents in crawls, and compared them with other sources that identify popular topics on Web at the time of the crawls (2014). The two crawls differ in terms of scope. While *Common Crawl* covers domains from the entire Web, *KB* covers mainly the Dutch domain. Therefore, we used different sources as a proxy of topics that were popular in 2014, both worldwide (entire Web) and national (*.nl* domain).

Using exact string matching between anchor texts and topics from different sources, we found that the percentages of matches vary between the topic sources and the two crawls. For example, CC_{links} covers 61 % of Google global trends, and 5 % of real queries (submitted by users to the search system of the Dutch digital newspaper archive). KB_{links} covers 32 % of Google *.nl* trends, and 2 % of the real queries. This suggests that anchor texts are a useful resource for investigating popular topics from the past. We found a correlation between the frequency of anchor texts in the archive and the percentage of topic matches.

When we compared the topic coverage between the *Common Crawl* and the *KB* datasets, we found that the percentage of overlapping topics is higher in the *Common Crawl* dataset, for both global and *.nl* topics. This result holds for the $CC_{links} \cap .nl$ (only focusing on links in *Common Crawl* originating from the *.nl* domain). More over, using the $CC_{links} \cap KB_{seeds}$ (was created using *KB* seeds to subset CC_{links}) has comparable result to KB_{links}. However, not all topics found by the *depth-first* crawl were found by the *breadth-first* crawl. We conclude that the coverage in the *breadth-first* crawl is higher even for topics of national interest, but there are topics that are covered only by *depth-first* crawl.

In future work, we can investigate the topic coverage in the crawls taking the importance of topics into account, in this analysis all topics were weighted equally.

Acknowledgments. We would like to thank the National Library of the Netherlands for their support. This research was funded by the Netherlands Organization for Scientific Research (NWO CATCH program, WebART project), and Dutch COMMIT/ program (SEALINCMedia project). Part of the analysis work was carried out on the Dutch e-infrastructure with the support of the SURF Foundation.

References

1. Baeza-Yates, R.A., Poblete, B.: Evolution of the Chilean web structure composition. In: LA-WEB, pp. 11–13 (2003)
2. Broder, A.Z., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Janet, L.: Wiener.: graph structure in the web. Comput. Netw. **33** (1–6), 309–320 (2000)
3. Brügger, N.: Historical network analysis of the web. Soc. Sci. Comput. Rev. **31**(3), 306–321 (2013)
4. Craswell, N., Hawking, D., Robertson, S.: Effective site finding using link anchor information. In: SIGIR, pp. 250–257. ACM (2001)
5. Donato, D., Leonardi, S., Millozzi, S., Tsaparas, P.: Mining the inner structure of the web graph. In: WebDB, pp. 145–150 (2005)
6. Dou, Z., Song, R., Nie, J.-Y., Wen, J.-R.: Using anchor texts with their hyperlink structure for web search. In: SIGIR, pp. 227–234 (2009)
7. Eiron, N., McCurley, K.S.: Analysis of anchor text for web search. In: SIGIR, pp. 459–460 (2003)
8. Fujii, A.: Modeling anchor text and classifying queries to enhance web document retrieval. In: WWW, pp. 337–346 (2008)
9. Huurdeman, H.C., Kamps, J., Samar, T., de Vries, A.P., Ben-David, A., Rogers, R.A.: Lost but not forgotten: finding pages on the unarchived web. Int. J. Digit. Libr. **16**, 247–265 (2015)
10. Jin, R., Hauptmann, A.G., Zhai, C.: Title language model for information retrieval. In: SIGIR, 11–15 August 2002, Tampere, Finland, pp. 42–48 (2002)
11. Kamps, J.: Web-centric language models. In: CIKM, pp. 307–308 (2005)
12. Kanhabua, N., Nejdl, W.: On the value of temporal anchor texts in wikipedia. In: SIGIR Workshop on Temporal, Social and Spatially-Aware Information Access (2014)
13. Klein, M., Nelson, M.L.: Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. Int. J. Digit. Libr. **14**, 17–38 (2014)
14. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**, 604–632 (1999)
15. Koolen, M., Kamps, J.: The importance of anchor text for ad hoc search revisited. In: SIGIR, pp. 122–129 (2010)
16. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proceedings of the 13th International Conference on World Wide Web, WWW 2004, pp. 666–674. ACM, New York (2004). doi:10.1145/988672.988763. ISBN: 1-58113-844-X
17. Masanès, J.: Web Archiving. Springer, Berlin (2006)

18. Metzler, D., Novak, J., Cui, H., Reddy, S.: Building enriched document representations using aggregated anchor text. In: SIGIR (2009)
19. Meusel, R., Vigna, S., Lehmberg, O., Bizer, C.: Graph structure in the web - revisited: a trick of the heavy tail. In: WWW, pp. 427–432 (2014)
20. Mühleisen, H.: Wikistats – wikipedia pageviews (2013). http://wikistats.ins.cwi.nl
21. Ntoulas, A., Cho, J., Olston, C.: What's new on the web? The evolution of the web from a search engine perspective. In: WWW, pp. 1–12 (2004)
22. Rauber, A., Bruckner, R.M., Aschenbrenner, A., Witvoet, O., Kaiser, M.: Uncovering information hidden in web archives: a glimpse at web analysis building on data warehouses. D-Lib Mag. **8**(12), 1082–9873 (2002). doi:10.1045/december2002-rauber
23. Ángeles Serrano, M., Maguitman, A.G., Boguñá, M., Fortunato, S., Vespignani, A.: Decoding the structure of the www: a comparative analysis of web crawls. ACM Trans. Web (TWEB) **1**(2), 10 (2007). doi:10.1145/1255438.1255442
24. Zheng, S., Pavel Dmitriev, C., Giles, L.: Graph-based seed selection for web-scale crawlers. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM, Hong Kong, China, 2–6 November 2009, pp. 1967–1970 (2009)
25. Zhu, J.J.H., Meng, T., Xie, Z., Li, G., Li, X.: A teapot graph and its hierarchical structure of the Chinese web. In: WWW, pp. 1133–1134 (2008)