Temporal Anchor Text as Proxy for Real User Queries

Thaer Samar and Arjen P. de Vries
Centrum Wiskunde & Informatica, Amsterdam
firstname.lastname@cwi.nl

Abstract. Web archives preserve the fast changing web. While we can archive the web pages, the popularity of queries in the past has usually not been preserved. Previous studies have observed the importance of anchor text for improving the quality of text search, and have shown that anchor text is similar to real user queries and documents titles. Other studies have shown that documents titles are similar to the real user queries. In this paper, we propose an approach to reconstruct the information that would be provided by query log in the past using temporal anchor text. First, we study the link graph of four years of Web archive in order to show how the target hosts and anchor text evolve over time. Second, we investigate the importance of anchor text over time. Our approach is to rank anchor text based on their popularity in the archive at specific time. Then, we check the importance of the top ranked anchor text in the public Web at the same time. In order to achieve this, we used the *WikiStats* dataset which aggregates page views of Wikipedia pages. Using exact string matching between top ranked anchor text and Wikipedia titles in the *WikiStats* dataset, we find a high percentage of overlap (approximately 57%). Our data strengthens the hypothesis that anchor text may be used as a proxy for actual query volume.

1 Introduction

The World Wide Web (WWW) is the largest and the main source of information nowadays, because of the ease of publishing and sharing data. However, the Web is dynamic and data can be easily lost on the Web. Ntoulas et al. [28] found that 80% of Web pages are not available after one year. Many national libraries and organizations realized the importance of Web archives for future culture heritage. Memory and heritage institutions increasingly recognize that such digital born data are as easily deleted as they are published, thereby introducing unprecedented risks to the world's digital cultural heritage [30]. [2] shows a list of Web archives initiatives undertaken by national libraries, national archives, national and international organizations for preserving the Web.

Despite the important attempts to preserve parts of the web by archiving, a large part of the web's content is unarchived and hence lost forever. In practice it is not feasible to archive the entire web due to its ever increasing size and rapidly changing content. The overall consequence is that our web archives are highly incomplete. On the other hand the Web archive is too complete because it it contains additional information about a Web page, more than its content, such as archived date, outlinks and anchor text.

Queries that represent the past interests of real users, using the archived Web as it was, are usually not available, because they were not preserved. Motivated by studies which showed that anchor text is similar to documents titles and real users queries, we

propose to use the important (popular) anchor text as proxy for queries in the past. In this paper, we study how the link graph evolves over time; specifically, we focus on target hosts and anchor text. We investigate evolution of the anchor text over time in order to understand what was important in Web.

2 Background

Methods using link structure analysis have been widely used, especially in the information retrieval area such as Page Rank [3] and HITS [20]. The links which define the structure of the Web consist of a source URL, a destination URL, and anchor text which is the text used to describe the target page in the link. Anchor text is a well-known resource to enrich the representations of web page content to improve Web retrieval. Craswell et al. [7] first experimented with site finding using aggregated anchor text. Aggregated anchor text for a link target has been used as surrogate documents, instead of the target pages' actual content. They concluded that anchor text can be more useful than content words for navigational queries. Eiron and McCurley [10] have investigated the properties of anchor text in a large intranet corpus in order to understand why using anchor text improves the quality of Web search. First, they showed empirically that anchor text exhibits characteristics similar to real user queries. Second, they hypothesize that anchor text is similar to web page titles, based on the observation by Jin et al. [15] that titles can be used as an approximation of queries. They found that anchor text is indeed similar to documents titles.

Work in this area led to advanced models that combine various representations of page content, anchor text, and link evidence [17]. Kraft and Zien [23] showed that anchor text can produce higher quality query refinement suggestions than content text. Fujii [12] proposed a model for classifying queries into navigational and informational. Their retrieval system used content-based or anchor-based retrieval methods, depending on the query type. Based on their experimental results, they concluded that content of web pages is useful for informational query types, whereas anchor text information and links are useful for navigational query types. Koolen and Kamps [22] concluded that anchor text has added value for ad hoc informational search as well, and can lead to significant improvements in retrieval effectiveness. Dou et al. [9], Kleinberg [21] took the relationship between source and anchor text into account. Their model distinguished between links from the same website and links from related sites to better estimate the importance of anchor text. Similarly, Metzler et al. [26] has overcome the problem of anchor text sparsity by smoothing the influence of anchor text originating from within the same domain by using 'external' anchor text: the aggregated anchor text from all pages that link to a page in the same domain as the page to be enriched.

In the context of Web archiving, link evidence and anchor text could be used to locate missing webpages, of which the original URL is not accessible anymore. Klein and Nelson [19] 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. Huurdeman et al. [14] first used the link structure extracted from archived Web pages to uncover target URLs that are not archived. Links extracted from the archived pages contain evidence of the existence of unarchived target URLs. Second, they used

link evidence to reconstruct basic representations of target URLs. This evidence includes the aggregated anchor text, crawl date, and source URLs.

So far, we have described works that studied the structure of the Web and how the link structure analysis was exploited for improving retrieval effectiveness. However, all of them focused on using single snapshot of archived websites. Now, we summarize studies that focused on the Web evolution by studying the link development over time. Web link structure is very dynamic and grows following a power law [24]. In the IR community, several works used the temporal information of archived material to improve search effectiveness. Li and Croft [25] proposed a time-based language model based on studying the correlation between time and relevance. Based on the heuristic that the probability of a document being relevant is higher for the most recent documents, they boosted the relevance of recent documents. Jones and Diaz [16] exploited the distribution of document versions over the timeline as an indication of the interval of time relevant to a query. Elsas and Dumais [11] found that documents that are more dynamic over time tend to be more relevant. Finally, Dai and Davison [8] quantified anchor text importance by differentiating pages with different incoming link creation rate over time and different historical incoming link context. They concluded that incorporating the importance of anchor text over time in the ranking model improves the performance, but they also point to the lack of available archived resources (few encountered links were actually available in the Internet Archive).

Costa et al. [6] improved the effectiveness of searching Web archives by incorporating temporal features such as number of versions available for the document in the archive, and life span between first and last version of the document. They studied the relation between Web document persistence and relevance. They presented an approach that learns and combines multiple ranking models specific for each period of time based on their believe that a single generic ranking model cannot predict the variance of Web characteristics over a long period of time. They work on a test collection constructed from the Portuguese Web Archive (PWA) in order to be used as ground truth for Web Archive Information Retrieval (WAIR) research [5]. The dataset is publicly available at [1], including 269,801 assessed Web document versions. The assessed documents were returned by different ranking models in response to 50 navigational queries. Queries were randomly sampled from the PWA's query log. The PWA consists of archived documents from the Portuguese Web in the period from 1996 to 2009. They found that there is no correlation between lifespan and number of versions, but both are correlated with the relevance of documents. They found that 36% of documents have a life span less than one year; notice that this percentage is different from the percentage found by [28] which is 80%.

Kanhabua and Nejdl [18] studied the evolution of anchor text extracted from edit history of Wikipedia. First, they identified a set of entities using the approach introduced by Bunescu and Pasca [4], for each Wikipedia snapshot. The snapshots were generated by partitioning revisions of Wikipedia pages based on one-month granularity. Then, they generate a set of entity-anchor relationships, based on the anchor text derived from links pointing to the entities. They found that anchor texts with temporal information can be candidates for capturing and tracing entities evolution.

In the context of Web archives, the queries that were used are usually not available, especially when the archive was not available for search. Given all the previous work that shows the similarity between anchor text and real users queries, and the similarity between anchor text and titles, we propose to investigate the evolution of anchor text in the past to give an insight about what was important and reconstruct queries over time.

3 Setup

3.1 Dataset

This study uses data from the Dutch Web archive at the National Library of the Netherlands (*KB*). The *KB* currently archives a pre-selected (seed) set of more than 5,000 websites [29]. Websites for preservation are selected by the library per category related to Dutch historical, social and cultural heritage. Our snapshot of the Dutch Web archive consists of 76,828 ARC files, which contain aggregated web content. Each ARC file contains multiple archived records (content plus response header). A total number of 148M documents has been harvested between February 2009 and December 2012, resulting in more than 7 Terabytes of data. Basic harvest metadata is available (crawl dates, page modification dates, etc.). Additional metadata is available in separate documentation, which includes the *KB*'s selection list, date of selection, and manually assigned UNESCO codes (by curators of the *KB*). Table 1 summarizes the number of websites added to the selection list and the total number of Web objects archived over the years.

Table 1: Number of seeds and archived objects over the years

year	# of seeds	# of archived objects
2009	2,491	17,014,067
2010	3,312	38,157,308
2011	3,508	53,604,464
2012	4,085	38,865,673
		147,641,512

3.2 Link Extraction & Aggregation

We extract a link structure from the archived objects that have `text/html` as MIME-type. The main percentage (approximately 70%, per year) of the archived web objects are HTML-based textual content. In order to extract the links from the archive, we use MapReduce to process all archived web objects contained in the archive's ARC files. During processing of the archived objects, JSoup[1] was used to extract anchor links from web objects that have `text/html` as MIME-type. For each found anchor link,

[1] `http://jsoup.org/`

we keep the source URL (which is the URL of the page that has the link), target URL (which is the URL of the page that the link is pointing to), and the anchor text of the link (a short text describing the target page). The archived pages have meta data of about the archived page such as the crawl date. We combine the year and the month of the crawl date with link information (YYYYMM). In addition to that, we keep the hash code (MD5) of the source page. More precisely, we keep the following information:

```
(sourceURL, targetURL, linkType, anchorText, crawlDate,
                        sourceHash)
```

The link type (linkType) indicates whether the link is internal link or external link. An internal link has the same domain-name for both source and target (intra-domain), while an external link the domain-name of the source URL is different from that of the target URL (an inter-domain link).

Different seeds are harvested at different frequencies; while most sites are harvested only once a year, some sites are crawled more frequently. Therefore, we deduplicate the links based on their values for source, target, anchor text, year and a hash of the source's content. We focus on the external links, and partition these links based on one-year, and one-month granularity.

3.3 Wikipedia Page Views Statistics

As evidence of query volume in the past, we used the *WikiStats* project dataset [27], which is an aggregated dataset from the `Page view statistics for Wikimedia projects`[2], which keeps the request history of articles from Wikipedia or from another projects. For each article, it keeps the title and the number of requests. *WikiStats* consists of weekly absolute views for Wikipedia pages in the period from January 2008 and January 2015. This gives the number of page views for the Wikipedia pages, the top-level domain (TLD) of the page (such as NL for the Netherlands), and the page's title. Because our snapshot of the Dutch Web archive covers the period between February 2009 and December 2012, we focused on the same period of the *WikiStats* dataset. We partitioned the dataset in this period based on one-month granularity and one-year granularity, keeping only Wikipedia titles which have more than $1,000$ page views.

4 Hosts Evolution

In Section 3.2, we introduced our approach of extracting the link graph from the archived `text/html` pages combined with metadata such as the crawl date, generating different partitions at different granularities. In this section study the importance of hosts in the archive over time.

First, we experiment with partitions based on the year granularity. For each partition, we generate the host of both the source page and target page in each link, where multiple links from the same source host will be considered one. After that we aggregate the links by target host. Finally, we rank the target hosts based on the number of incoming

[2] http://dumps.wikimedia.org/other/pagecounts-raw/

links; which corresponds to the number of unique source hosts pointing to the target host. Table 3 shows the top ranked hosts per year. We observe that the ranks of the top hosts vary over the years. By considering the top $1,000$ hosts per year, we find no correlation (using Kendall's τ) between the ranked lists of hosts in different years; the strongest negative correlation τ was -0.982 between 2011 and 2012. Table 2 shows the percentage of new hosts in our crawls over the years, considering different thresholds of the top hosts. Here, a host is considered new in a particular year if it does not appear in any previous year.

Next, we experiment with aggregating links by target host, based on the one-month granularity. Table 4 and Table 5 show the top hosts per month in 2009, illustrating that the top hosts vary over the months as well. The number of target hosts varies per month, with an average of $53,215$ hosts per month, where 25% these hosts are new.

Table 2: Percentage of new target hosts over the years considering the top 1,000, 5,000, and 10,000 hosts.

year	Top 1,000	Top 5,000	Top 10,000
2010	37.5	38.3	38.9
2011	26.8	27.3	27.4
2012	19.1	21.2	21.4
Mean	27.80	28.9	29.2

5 Anchor Text Evolution

In this section, we look into the usage of anchor text over time. For each partition $\mathcal{A}_t$ at a given time granularity, we aggregate links by anchor text. The number of links using anchor text a represents the frequency of a in partition $\mathcal{A}_t$. We used this relative frequency to represent the importance of anchor text a in the archive at specific time granularity t (`archive-based popularity`), computing the importance of the anchor text as follows:

$$I(a, \mathcal{A}_t) = \frac{f(a, \mathcal{A}_t)}{max_{\mathcal{A}_t}} \tag{1}$$

where $f(a, \mathcal{A}_t)$ is the frequency of anchor text a in partition $\mathcal{A}_t$, and $max_{\mathcal{A}_t}$ is the maximum frequency of any anchor text in partition $\mathcal{A}_t$.

$$max_{\mathcal{A}_t} = \max_{a} f(a, \mathcal{A}_t) \tag{2}$$

First, we investigate the evolution of anchor text over time. Therefore, for the anchor text in partition $\mathcal{A}_t$, we compute the percentage of new anchor text at the time of t. An anchor text is considered new in $\mathcal{A}_t$ if it does not appear in any previous partition.

$$new(a, t) = \begin{cases} 1, & \text{if } a \notin \bigcup\limits_{i<t} \mathcal{A}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Table 3: Top ranked hosts over the years.

2009	2010	2011	2012
vriezenveners.nl	hetutrechtsarchief.nl	wikipedia.org	**twitter.com**
mi-website.es	wikipedia.org	hetutrechtsarchief.nl	hetutrechtsarchief.nl
startpagina.nl	europa-nu.nl	biblion.nl	wikipedia.org
fd.nl	bibe.library.uu.nl	**twitter.com**	europa.eu
z24.nl	**twitter.com**	europa-nu.nl	bibe.library.uu.nl
wikipedia.org	belastingdienst.nl	europa.eu	wordpress.com
blogspot.com	europa.eu	bibe.library.uu.nl	blogspot.com
deviantart.com	vriezenveners.nl	blogspot.com	europa-nu.nl
co.uk	startpagina.nl	co.jp	youtube.com
volkskrant.nl	minszw.nl	youtube.com	vriezenveners.nl
gencircles.com	uva.nl	co.uk	co.uk
sitestat.com	readspeaker.com	wordpress.com	google.com*
belastingdienst.nl	blogspot.com	leidenuniv.nl	leidenuniv.nl
web-log.nl	co.uk	google.com*	ebay.com
startkabel.nl	google.com*	belastingdienst.nl	rijksoverheid.nl
imageshack.us	sitestat.com	startpagina.nl	marktplaats.nl
readspeaker.com	amazon.com	vriezenveners.nl	overheid.nl
google.com*	wordpress.com	amazon.com	co.jp
hva.nl	youtube.com	readspeaker.com	knaw.nl
digischool.nl	ebay.com	ligfiets.net	volkskrant.nl
nrc.nl	omroep.nl	zijpermuseum.nl	nuzakelijk.nl
trouw.nl	volkskrant.nl	co.cc	zie.nl
wordpress.com	web-log.nl	ebay.com	startpagina.nl
photobucket.com	ligfiets.net	kennisnet.nl	facebook.com
ugo.com	nrc.nl	tue.nl	tue.nl

Table 4: Top hosts per month. **02-06/2009**

200902	200903	200904	200905	200906
fd.nl	adobe.com	volkskrant.nl	deviantart.com	knnv.nl
z24.nl	ppsi.nl	trouw.nl	imageshack.us	blogspot.com
ugo.com	schoolenveiligheid.nl	nrc.nl	photobucket.com	waarneming.nl
volkskrant.nl	omroep.nl	emancipatie.nl	avs.nl	google.com
volny.cz	minocw.nl	nd.nl	intermediair.nl	europa.eu
digischool.nl	pestweb.nl	szw.nl	intermediairforward.nl	web-log.nl
fateback.com	eu.int	europa-nu.nl	sitestat.com	startpagina.nl
sitestat.com	europa.eu	cgb.nl	independer.nl	volkskrantblog.nl
trouw.nl	aps.nl	mozilla.org	blogspot.com	wordpress.com
aol.com	ez.nl	ad.nl	indymedia.nl	co.uk
nrc.nl	aob.nl	refdag.nl	wikipedia.org	wikipedia.org
wikipedia.org	overheid.nl	mozilla.com	wikimedia.org	decontrabas.com
blogspot.com	kpcgroep.nl	lbr.nl	punt.nl	google.nl
chinesefreewebs.com	kennisnet.nl	wordpress.com	co.uk	vpro.nl
typepad.com	justitie.nl	rotterdamsdagblad.nl	blogger.com	wikimedia.org
szw.nl	telegraaf.nl	parool.nl	wordpress.com	eu.int
pretoriashow.com	volkskrant.nl	wikimedia.org	free.fr	omroep.nl
ad.nl	vfpf.nl	europa.eu	youtube.com	gov.uk
co.uk	havenmuseum.nl	telegraaf.nl	libero.it	americantaskforce.org
freewebtown.com	rutgersnissogroep.nl	cnet.com	aol.com	imageshack.us

Table 5: Top hosts per month. **07-12/2009**

200907	200908	200909	200910	200911	200912
volkskrantblog.nl	seniorweb.nl	readspeaker.com	mi-website.es	vriezenveners.nl	hva.nl
anwb.nl	fietsersbond.nl	belastingdienst.nl	startpagina.nl	gencircles.com	startpagina.nl
wordpress.com	archined.nl	cwi.nl	startkabel.nl	startkabel.nl	blogspot.com
adobe.com	begraafplaats.org	artsennet.nl	fd.nl	startpagina.nl	wikipedia.org
google.com	co.uk	wordpress.com	volkskrant.nl	deviantart.com	google.com
anwbentreebewijs.nl	sitestat.com	w3.org	z24.nl	ugo.com	co.uk
waverunner.nl	drenthe.nl	europa.eu	digischool.nl	readspeaker.com	web-log.nl
wikipedia.org	wikipedia.org	knzb.nl	wikipedia.org	belastingdienst.nl	twitter.com
postbus51.nl	site-id.nl	wikipedia.org	sitestat.com	wikipedia.org	wikimedia.org
pharosreizen.nl	google.com	imageshack.us	trouw.nl	photobucket.com	lexius.nl
live.com	overheid.nl	google.com	web-log.nl	imageshack.us	blogger.com
w3.org	knhb.nl	oreilly.com	nrc.nl	twitter.com	omroep.nl
volkskrant.nl	nai.nl	co.uk	co.uk	youtube.com	wordpress.com
amsterdam.nl	amsterdam.nl	overheid.nl	szw.nl	blogspot.com	creativecommons.org
telekom.at	xs4all.nl	google.nl	members.lycos.co.uk	blogger.com	greenpeace.org
vrom.nl	uitvaartmedia.com	photobucket.com	ifrance.com	avs.nl	hanze.nl
gelderlander.nl	leidenuniv.nl	myspace.com	kennisnet.nl	google.com	technorati.com
google.nl	tudelft.nl	businessweek.com	lycos.nl	co.uk	youtube.com
youtube.com	uitvaartinformatie.nl	uitvaart.nl	ad.nl	wikimedia.org	hszuyd.nl
belastingdienst.nl	volkskrant.nl	blogspot.com	blogspot.com	independer.nl	vpro.nl

where $\mathcal{A}_i$ represents any partition with time granularity less than the time granularity of $\mathcal{A}_t$. Based on the partitions of one-year granularity, with an average of $999,695$ distinct anchor text per year, we find that 59% of anchor text are new (average across the percentage of all years). Based on the partitions of one-month granularity, $17,024$ links with distinct anchor text exist per month. The average percentage of new anchor text per months is 34%.

We have discussed a series of studies that showed that document titles are close to real user queries, and that anchor text is similar to both document titles and real user queries. We therefore hypothesize that we may be able to reconstruct query volume in the past based on anchor text used in the past. Similar to the use of wikipedia in [31], we used the *WikiStats* dataset (described in Section 3.3) in order to find how the important anchor text in the archive were related to popular queries in the past on the public Web. We consider the number of page views of Wikipedia titles that match anchor text to represent the importance of that anchor text in the public Web (`web-based popularity`). We study the similarity between anchor text and Wikipedia titles varying temporal granularity. We used exact string matching to match anchor text with titles of Wikipedia pages in the *WikiStats* dataset, using the same time granularity. Matching was done after transforming both anchor text and Wikipedia titles into lower case. For each partition at time t, we rank the anchor text based on archive-based popularity, after which we check at different thresholds k how many of the top-k anchor text occurrences in the *WikiStats* dataset (in the partition at time t of the *WikiStats* dataset). Table 6 summarizes the percentage of anchor text that have a matched Wikipedia title. As we observe in the Table, a high percentage of the top ranked anchor text has a matching Wikipedia title. For example, 56% of the top-$1k$ anchor text occurrences in the 2009 partition were found also in the 2009 partition of the *WikiStats* dataset. We observe that the percentage of overlap between anchor text and the *WikiStats* dataset partitions decreases as we increase the threshold of the top-k. The percentage reaches

26% (averages across all partitions) when we consider all anchor text in the one-year partition.

Table 6: Absolute count and percentage of anchor text per year that has a Wikipedia title match at different thresholds.

	Top-$1k$		Top-$5k$		Top-$10k$	
year	count	%	count	%	count	%
2009	559	55.9	2488	49.8	4259	42.59
2010	585	58.5	2326	46.5	3350	33.50
2011	572	57.2	2466	49.3	3995	39.95
2012	564	56.4	2340	46.8	4186	41.86

Table 7 shows a comma-separated sample of anchor text taken from the top-$1k$ popular anchor text in 2012 which do not have a match of any Wikipedia titles in 2012 of the *WikiStats* dataset. Some of these are uninformative having a specific purpose, such as 'login' to proceed. Some anchor text have no match because of limitations due to our approach of looking for exact string match between the anchor text and the Wikipedia titles. For example the anchor text 'filmpje' has no match but in the *WikiStats* dataset there is a page with title 'filmpje!'. Likewise, 'nunl' has no exact match, however there is a Wikipedia page with title 'nu.nl'. In the future, our approach should consider these cases by applying additional pre-processing steps like stemming and stopping, and generalizing from exact match to matches with low edit distance. The list of anchor text at the top-$1k$ in 2012 that have a match with Wikipedia title is shown in Table 8. We observe that some of these anchor texts correspond to cities in the Netherlands such as Amsterdam, Rotterdam, Groningen, Utrecht and Den Haag (all are major cities in the Netherlands). Another category of the top anchor text is related to social websites such as twitter, linkedin, flickr, and vimeo. A different category of anchor text consists of the major Dutch daily newspapers such as de Volkskrant, Telegraaf, Trouw, and NRC handelsblad. The 'uitzending gemist' occurrence is related to a web service of the Dutch Public Broadcasting (NPO) that offers a free on demand video for nation broadcasts. The 'belastingdienst' anchor text is about a governmental service related to the Dutch national tax office.

Based on the one-month granularity, on average 26% of all anchor text over all months has an exact match with a Wikipedia title (using all domains). The highest percentage of Wikipedia titles that match the anchor text originate from the 'NL' domain (around 55%). By ranking the anchor text per each one-month granularity based on the archive-based popularity, we find that 42.5% of anchor text in the top-$1k$ has match with Wikipedia titles.

Table 7: List of anchor texts in the top-$1k$ of 2012, that have no matching Wikipedia title.

ga naar website van de fabrikant, word vaste donateur of doneer online via de website van dit goede doel, create your own free blog on wordpresscom, filmpje, vacatures, log in to proceed, wordpresscom, view more information, grotere kaart weergeven, inlichtingen, routebeschrijving, powered by wordpresscom, more information, projectinformatie, volg ons op twitter, nunl, eigen homepage, inschrijven,

Table 8: List of anchor texts the top-$1k$ of 2012 which have matching Wikipedia titles.

twitter, tweet, linkedin, hyves, jaarverslag, onderzoek, persbericht, pdf, weblog, wordpress, flickr, rapport, rss, vimeo, bron, amsterdam, programma, blogger, de volkskrant, brief, trouw, utrecht, details, samenvatting, rotterdam, groningen, joomla, volkskrant, klik, webwinkel, uitzending gemist, belastingdienst, deel, nrc handelsblad, bericht, den haag, de telegraaf, nrc,

6 Conclusions and Future Work

In this study, we looked into the viability of a new approach of using the evolution of anchor text over time to reconstruct information that would be similar to real user queries in the past. Our hypothesis is based on studies that have shown that anchor text behaves similar to both real user queries and documents titles. We used the link structure extracted from the Dutch Web archive to identify the most popular target hosts over time, and to get the most popular anchor text over time. The link structure was extracted from archived `text/html` archived pages in the Dutch Web archive in the period between February 2009 and December 2012. In order to understand the importance of the anchor text, we rely on the *WikiStats* dataset, which provides an aggregation of page views of Wikipedia pages. We investigate the exact matches between anchor text and Wikipedia titles, where both datasets (the link structure and the *WikiStats*) were partitioned based on one-month and one-year granularity. Our analysis of the target hosts shows that target hosts evolve significantly. Based on the one-month granularity, on average 25% among all hosts per month are new. We experiment with finding popular anchor text per time granularity, ranking anchor texts based on their popularity in the archive. We find that a high percentage of anchor text in the top ranks have a match with Wikipedia titles in the *WikiStats* dataset. Based on the one-year granularity, we found that 57% of the top-$1k$ anchor texts have matching Wikipedia titles. We conclude from our data that the most important text provides a view of what are important entities in the Netherlands. We cannot however conclude that evolution of anchor text serves as a proxy for past query logs . There are some limitations that will consider in the future work. First, matching anchor text and Wikipedia titles analysis, suggests a room for improving our approach by applying additional pre-processing steps like stemming and stopping, and generalizing from exact match to matches with low edit distance. Second, we test our approach on a 'deep crawl' which is based on a few thousands of seeds. In

the future, we will test our approach on a 'breadth-first crawl' like the Common Crawl dataset[3].

Bibliography

[1] Dataset for learning to rank for wair research. https://code.google.com/p/pwa-technologies/wiki/L2R4WAIR.

[2] List of web archives initiatives. http://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[4] Razvan C. Bunescu and Marius Pasca. Using encyclopedic knowledge for named entity disambiguation. In Diana McCarthy and Shuly Wintner, editors, *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. The Association for Computer Linguistics, 2006. ISBN 1-932432-59-0. URL http://acl.ldc.upenn.edu/E/E06/E06-1002.pdf.

[5] Miguel Costa and Mário J. Silva. Evaluating web archive search systems. In Xiaoyang Sean Wang, Isabel F. Cruz, Alex Delis, and Guangyan Huang, editors, *WISE*, volume 7651 of *Lecture Notes in Computer Science*, pages 440–454. Springer, 2012. ISBN 978-3-642-35062-7.

[6] Miguel Costa, Francisco M. Couto, and Mário J. Silva. Learning temporal-dependent ranking models. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *SIGIR*, pages 757–766. ACM, 2014. ISBN 978-1-4503-2257-7.

[7] Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *SIGIR*, pages 250–257. ACM, 2001.

[8] Na Dai and Brian D. Davison. Mining anchor text trends for retrieval. In Cathal Gurrin, Yulan He, Gabriella Kazai, Udo Kruschwitz, Suzanne Little, Thomas Roelleke, Stefan M. Rüger, and Keith van Rijsbergen, editors, *ECIR*, volume 5993 of *LNCS*, pages 127–139. Springer, 2010. ISBN 978-3-642-12274-3.

[9] Zhicheng Dou, Ruihua Song, Jian-Yun Nie, and Ji-Rong Wen. Using anchor texts with their hyperlink structure for web search. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *SIGIR*, pages 227–234. ACM, 2009. ISBN 978-1-60558-483-6.

[10] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR*, pages 459–460, 2003.

[11] Jonathan L. Elsas and Susan T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *WSDM*, pages 1–10. ACM, 2010. ISBN 978-1-60558-889-6.

[12] Atsushi Fujii. Modeling anchor text and classifying queries to enhance web document retrieval. In Huai et al. [13], pages 337–346. ISBN 978-1-60558-085-2.

[13] Jinpeng Huai, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins, and Xiaodong Zhang, editors. *Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008*, 2008. ACM. ISBN 978-1-60558-085-2.

[14] HugoC. Huurdeman, Jaap Kamps, Thaer Samar, ArjenP. de Vries, Anat Ben-David, and RichardA. Rogers. Lost but not forgotten: finding pages on the unarchived

[3] https://commoncrawl.org/

web. *International Journal on Digital Libraries*, pages 1–19, 2015. ISSN 1432-5012. doi: 10.1007/s00799-015-0153-3. URL http://dx.doi.org/10.1007/s00799-015-0153-3.

[15] Rong Jin, Alexander G. Hauptmann, and ChengXiang Zhai. Title language model for information retrieval. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 42–48. ACM, 2002. doi: 10.1145/564376.564386. URL http://doi.acm.org/10.1145/564376.564386.

[16] Rosie Jones and Fernando Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3), 2007.

[17] Jaap Kamps. Web-centric language models. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *CIKM*, pages 307–308. ACM, 2005. ISBN 1-59593-140-6.

[18] Nattiya Kanhabua and Wolfgang Nejdl. On the value of temporal anchor texts in wikipedia. In *SIGIR 2014 Workshop on Temporal, Social and Spatially-aware Information Access (TAIA'2014)*, 2014.

[19] Martin Klein and Michael L. Nelson. Moved but not gone: an evaluation of real-time methods for discovering replacement web pages. *Int. J. on Digital Libraries*, 14(1-2):17–38, 2014.

[20] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, 1999.

[21] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, 1999. ISSN 0004-5411. doi: 10.1145/324133.324140.

[22] Marijn Koolen and Jaap Kamps. The importance of anchor text for ad hoc search revisited. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *SIGIR*, pages 122–129. ACM, 2010. ISBN 978-1-4503-0153-4.

[23] Reiner Kraft and Jason Zien. Mining anchor text for query refinement. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 666–674, New York, NY, USA, 2004. ACM. ISBN 1-58113-844-X. doi: 10.1145/988672.988763.

[24] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *TKDD*, 1(1), 2007. doi: 10.1145/1217299.1217301. URL http://doi.acm.org/10.1145/1217299.1217301.

[25] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *CIKM*, pages 469–475. ACM, 2003. ISBN 1-58113-723-0.

[26] Donald Metzler, Jasmine Novak, Hang Cui, and Srihari Reddy. Building enriched document representations using aggregated anchor text. In *SIGIR*, pages 219–226, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571981.

[27] Hannes Mühleisen. Wikistats – Wikipedia page views, 2013. URL http://wikistats.ins.cwi.nl.

[28] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web?: the evolution of the web from a search engine perspective. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *WWW*, pages 1–12. ACM, 2004. ISBN 1-58113-844-X.

[29] M. Ras. Eerste fase webarchivering. Technical report, Koninklijke Bibliotheek, 2007.

[30] UNESCO. Charter on the preservation of digital heritage (article 3.4), 2003.

[31] Stewart Whiting, Joemon M. Jose, and Omar Alonso. Wikipedia as a time machine. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 857–862, 2014. doi: 10.1145/2567948.2579048. URL http://doi.acm.org/10.1145/2567948.2579048.