

Mining Relevant Time for Query Subtopics in Web Archives

Tu Ngoc Nguyen, Nattiya Kanhabua, Wolfgang Nejdl and Claudia Niederée

L3S Research Center / Leibniz Universität Hannover, Germany
{tunguyen, kanhabua, nejdl, niederee}@L3S.de

ABSTRACT

With the reflection of nearly all types of social cultural, societal and everyday processes of our lives in the web, web archives from organizations such as the Internet Archive have the potential of becoming huge gold-mines for temporal content analytics of many kinds (e.g., on politics, social issues, economics or media). First hand evidences for such processes are of great benefit for expert users such as journalists, economists, historians, etc. However, searching in this unique longitudinal collection of huge redundancy (pages of near-identical content are crawled all over again) is completely different from searching over the web. In this work, we present our first study of mining the temporal dynamics of subtopics by leveraging the value of anchor text along the time dimension of the enormous web archives. This task is especially useful for one important ranking problem in the web archive context, the time-aware search result diversification. Due to the time uncertainty (the lagging nature and unpredicted behavior of the crawlers), identifying the trending periods for such temporal subtopics relying solely on the timestamp annotations of the web archive (i.e., crawling times) is extremely difficult. We introduce a brute-force approach to detect a time-reliable sub-collection and propose a method to leverage them for relevant time mining of subtopics. This is empirically found effective in solving the problem.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

Keywords

Anchor Text Mining; Result Diversification; Temporal Ranking; Temporal Subtopic

1. INTRODUCTION

Web search is good in delivering up-to-date or fresh information for topics of all types. Due to its vivid and wide used and participative content creation, the Web is in addition a

good reflection of processes, practices, and topics in all areas of life including politics, society, science. When we regularly take snapshots of the Web at different times as it is done in Web archiving (at least for part of the Web), we can, thus, capture this world reflection at different times as well as its evolution via subsequent versions. Thus, web archives have the potential to provide a rich source of first-hand information from the past - and about how things evolved. It can, for example, be seen how topics such as integration, nuclear power or democracy were discussed in the early 90's - and how this discussion changes over time. In addition, looking at more mundane issues, in some decades from now we can see what people did wear, eat, and talk about in 2015 from archived evidences. Although such content might seem trivial in the first place, it accumulates into an unprecedented form grass-root historical records.

Although such information from the past might still be findable in the current Web, they are typically aggregated, filtered and interpreted from a current perspective. For experts and professionals such as journalists, researchers from political science and sociology, historians, a first-hand and unbiased reflection of the world opens up the investigation of own stories and completely new research questions. It enables them to better understand how and why issues, for example, controversial topics evolved over time. They can also see the context of such discussions and have a first-hand account of change such as the evolution of language.

Anchor texts have been shown as an important factor of the Web that can be used to mimic the behavior of the query logs [3], representing documents [2] or subtopic mining [5]. However, none of the above have studied the temporal dynamics of the anchor texts for the subtopic mining tasks. In a recent work, Kanhabua and Nejdl [6] studied the terminology evolution of entities by means of anchor text in the context of Wikipedia. However, different from Wikipedia, the timestamp annotations in the real web archives are a lot noisier due to the unreliability of the crawling time. In this work, we present our first study of mining the temporal dynamics of subtopics in the web archives for the time-based search result diversification task. The quality of timestamp annotations (i.e., crawling time) in the web archives at the document level is rather unpredictable. Table 1 illustrates an example of how unreliable the crawling time is for certain circumstances. Two documents that mentioned the incident related to the subtopic *late-term abortion* that involved with the Governor *Kathleen Sebelius* are only first crawled 3-4 years after. The actual timing for the trending of the subtopic is however in April 2009. This shows a significant

lagging between the publishing date of a page and when it is actually crawled.

Our contributions in this work are: (1) we address the problem of mining temporal subtopic in a web archive, with the respect to time uncertainty, (2) we introduce a method to extract reliable publication dates from the web archive resources and (3) we exploit anchor text as a good source of mining temporal subtopics in the web archives and propose a method to infer relevant time (or a date) for the temporal subtopic.

2. RELATED WORK

There have been several works on mining aspects from the anchor texts [3, 4, 5], however they only mine on the current snapshot of the Web. The temporal dynamics of subtopics is first studied in [8] and is used to improve the ranking effectiveness of such queries at particular times. Dai et al. [2] also study the trending of the anchor texts by looking back at the historical web snapshot to improve the weighting function for document retrieval task. In the web archive context, there has been no existing work so far studied the temporal dynamics of subtopics.

3. TEMPORAL SUBTOPIC MINING FROM ANCHOR TEXT

Anchor text created by web content editors often reflect high quality summarizations of the destination pages. As anchor texts are often short and descriptive, it is shown to possess similar characteristics with web queries [3]. Anchor text with regards to a topic/query (contain the query terms) often convey diverse aspects of the topic, hence is a good source of subtopic mining. For temporal subtopic mining in the web archives, with the absence of the query logs and the unreliability of state-of-the-art retrieval models (in retrieving top-K relevant documents at a time-period), we observe temporal anchor text reflects a good correlation with the temporal subtopic. The correlation is further elaborated in Section 4.3.1. For a given query q , we first get all anchor texts containing all query terms of q , weight them, and select the most important ones as subtopics. We follow the weighting mechanism proposed in [5], where they observe that the importance of an anchor text is usually proportional to its popularity on the Web, i.e., how many times it is used in web sites or pages. The importance score is also traded off against the length of the anchor text. The anchor text c with respect to query q is weighted as:

$$f(q, c) = freq(c) \times rel(q, c) \\ = [N_c^{site} + \log(N_c^{page} N_c^{site} + 1)] \times \frac{1 + len(q)}{len(c)} \quad (1)$$

whereas $freq(c)$ is the frequency of the anchor text c , N_c^{site} is the number of unique sites contain c and N_c^{page} is the number of pages contain c .

3.1 Subtopic Extraction

In order to construct the set of distinctive and high quality subtopics/aspects of a query, we also need to apply a clustering technique on the anchor texts, as follows [4]:

3.1.1 Similarity measures

Relevant models The similarity measure between anchor text pairs is not effective if based solely on the content

of the anchor text, that often contains few terms. Instead, an anchor text is represented as the accumulated of top-k documents that are relevant to it. Specifically, each anchor text a is represented by the relevance model $P_a(w|R)$ estimated from the top-10 documents returned by the query likelihood retrieval model (we don't need a temporal retrieval model in this step) for a . The similarity of two subtopics c_1 and c_2 is then calculated as the KL-divergence between their relevance models $Pr_1(w|R)$ and $Pr_2(w|R)$. However, building relevance models for every anchor text is relatively computationally expensive.

Co-occurrence At Passage Level Follow Dang et al. [4], we also conduct a more efficient method based on passage analysis. The idea is that two anchor texts are more similar if they co-occur often in the same text passages. Therefore, for every pair of anchor texts c_i and c_j , we compute N_i and N_j - the number of passages in which each of them occurs, and N - the number of passages in which they co-occur. The similarity between c_i and c_j is given by the Jaccard score:

$$sim(c_i, c_j) = \frac{N}{N_i + N_j - N} \quad (2)$$

3.1.2 Clustering Algorithms

Affinity Propagation Algorithm Follow [8, 9], we use Affinity Propagation (AP) for the clustering task. AP has an advantage over other clustering algorithm that it determines the number of clusters automatically.

3.2 Time inference for temporal subtopics

Due to the unreliability of the timestamp annotations in the web archives (i.e., the crawling time), detecting the trending period of a subtopic is not straight-forward. We use a brute-force approach (explained in detail in Section 4.2) to extract the most-reliable publication dates out of the web archives to acquire a substantial subset of documents¹ with highly-reliable dates. Our idea is to leverage this high quality temporal sub-collection to infer the relevant date for the temporal subtopics. One can think of mining the subtopics and its relevant times (e.g., via the frequency distribution) directly from this time-reliable sub-collection. However, beside the incompleteness of the sub-collection, it is difficult to infer trending behaviors of the temporal subtopics (as the relevant documents in the sub-collection does not follow a normal distribution).

Temporal language model Given our temporal collection \mathcal{C} with a set of time-partitions $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$, our task is to weight a temporal subtopic c with respect to each time-partition. This time-interval ranking approach is based on the temporal language model presented in [7]. The idea is to assign a probability to a time partition according to word usage or word statistics over time. A normalized log-likelihood ratio is used to compute the similarity between two language models. Here, we expand a subtopic c as the accumulated set of all the anchor text's terms in its cluster (explain in the previous section), removing all the duplicates.

¹since our queries are *informational*, the revisions are mostly of duplicated content, hence we only consider at document-level. In detail, revisions with the same publication date are merged into the oldest revision.

Table 1: Example of relevant documents for the *late-term abortion* subtopic

URL	Crawling time	Actual date	Content
http://www.vitter.senate.gov/...sebelius-appointment	2013-02-22	2009-04-20	Vitter Voices Grave Concerns Over Sebelius Appointment Monday, April 20, 2009 'I was already concerned about the Governor's position on a number of issues, especially those relating to abortion,' said Vitter. 'The fact that Governor Sebelius has accepted thousands of dollars in campaign contributions from George Tiller - a highly-controversial individual who specializes in performing late-term abortions, a practice far beyond those performed in the majority of abortion clinics - provides some insight into her views on abortion that raise many important concerns about her nomination..
http://lamborn.house.gov/...pro-abortion-veto/	2012-11-09	2009-04-24	Lamborn Comments on Governor Sebelius Pro-Abortion Veto Apr 24 2009 Calls yesterday's action preview of extreme agenda Washington, Apr 24 - Congressman Doug Lamborn (CO-05) today released the following statement regarding President Obama's nomination for Secretary of the Department of Health and Human Services, Governor Kathleen Sebelius of Kansas, in response to her latest pro-abortion action. Yesterday, she vetoed a common sense bill that would have required doctors performing late-term abortions to report additional information on those procedures to the Kansas Department of Health and Environment. The bill would have also given women the right to sue the doctors, should they later believe their abortions were illegal..

$$S(c_i, t_j) = \sum_{word \in c_i} P(word|c_i) \log \frac{P(word|t_j)}{P(word|(C))} \quad (3)$$

The $S(c_i, t_j)$ is the probability that a temporal subtopic c_i is relevant to the time period t_j .

Connection with the time-based search result diversification

Search result diversification is meant for diversifying the result list so that the top-k covers all the aspects of an ambiguous query. In the web archive context, the requirement is rather more complicated as its also essential to cover the time-periods where the aspects/subtopics are trending, ranked based on the 'trending weight' of the subtopics. Hence, the objective function of the diversification ranking needs to be re-designed to take this temporal subtopic factor into account. Basically, it needs to diversify over two distinct dimensions (i.e., time and aspects) and present them in a comprehensive way (i.e., *federated/vertical* search). Previous works did not however take the two important factors into account in a unified framework. Berberich et al. [1] only consider diversifying over the time dimension where they consider each time-period is a query aspect. Nguyen et al. [8] take both time and aspect into account but for their recency-favor ranking model. We leave this interesting task in the web archive context for future work.

4. EXPERIMENTS

4.1 Dataset

.gov domain collection We utilized a full corpus of archival web pages in .gov domain collected by the Internet Archive from January 1995 to September 2013. The corpus contains over 900 million of text captures and over 58.8 billion temporal links. Figures 1 and 2 show the document and document/revision ratio distribution of the collection. We extracted the anchor text and its timestamp and built a temporal language model for every monthly bin.

4.2 Determining time of the crawled web document

Identifying the *publication time* for a crawled web document is a difficult task, as the crawling time is not a totally reliable source. We confide on 5 different sources (ranked by level of reliability). We judge the level of reliability by judging the accuracy level of different source (with random

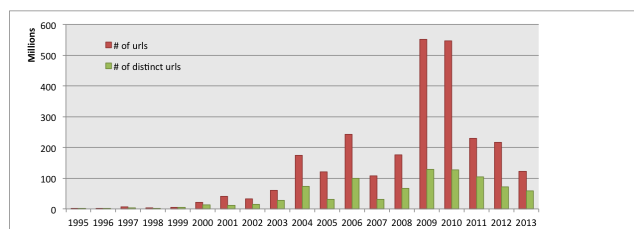


Figure 1: The document distribution

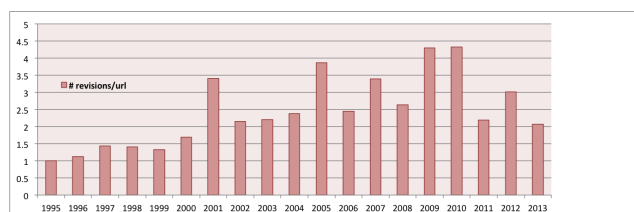


Figure 2: The document/revision ratio distribution

subset of 300 documents). The URL source is of 96%, content source is of 90%, whereas we don't have enough clue to judge the HTTP header sources.

- Date extraction from URLs. Often, a web article's link contains the date of creation (e.g., <http://www.whitehouse.gov/the-press-office/2011/12/24/blah>). We define this source as *1st level* of confidence, called *very strong*. There are also cases that only dates at month granularity are provided in the URL, we mark it as *mildly strong* (that if day granularity- publication date can be extracted from the content then we will use it instead).
- Publication date from document content. The publication date often lies in the first one-two lines of the article content. We use a temporal tagger called Heildtime² to exact the first date (if there is) out of this text snippet. This is *2nd level* of confidence, called *strong*.
- Last-Modified Date from HTTP Header. The Last-Modified Date is however, often not quite different from the creation date in our case as the modification

²<https://code.google.com/p/heildtime/>

Table 2: Examples of the subtopics extraction, weighting and time-relevant for 3 queries *abortion*, *border fence* (graphs not available) and *health care*.







Query	Original anchor text	Subtopic	Weight	Most relevant time
abortion	Arizona governor signs law banning most late-term abortions	late-term abortion	0.15	2009-04 
	Prohibit partial-birth abortion bill	partial-birth abortion	0.03	2011-04 
	Republicans attack Obama ahead of vote on bill to punish sex-selective abortion	sex-selective abortion	0.02	2010-02 
border fence	Bush signs border fence funding into law	border fence funding	0.12	2008-03
	Construction of the Mexico border fence	border fence mexico	0.02	2006-10
	Five years ago, legislation was passed to build a 700-mile double-layer border fence along the southwest border	double-layer border fence	0.07	2011-11
health care	Obama promotes health care reform - at a grocery store	health care reform	0.13	2010-03 
	The health care vote	health care vote	0.11	2010-03 
	Freshmen propose health care amendments	health care amendment	0.03	2010-03 

Table 3: Statistics for date extraction method over revisions

Level	Number of revisions	Percentage
<i>Very strong</i>	1271124	1.23
<i>Strong</i>	12696686	12.22
<i>Mildly strong</i>	1173490	1.13
<i>Acceptable</i>	14179469	13.65
<i>Weakly acceptable</i>	74546697	71.35
<i>Crawling time</i>	435785	0.42

Table 4: Statistics for date extraction method over documents

Level	Number of documents	Percentage
<i>Very strong</i>	139426	0.48
<i>Strong</i>	2968054	10.32
<i>Mildly strong</i>	258591	0.89
<i>Acceptable</i>	5264909	18.32
<i>Weakly acceptable</i>	20102819	69.49
<i>Crawling time</i>	134852	0.46

(if there is, occurs shortly after the article is online), called *acceptable*.

- Creation date from HTTP Header. This is ranked as *3-level*, called *weakly acceptable*.
- Crawling date, the weakest reliable source.

Table 4 describes the statistics of our date extraction method. Even the percent of number of *strongly*-confident date extractions are not high (approx. 15%), we believe that this is still accountable as the *important* documents appear in top-K results tends to have a good template and are easier to determine the publication time.

4.3 Preliminary results

4.3.1 Correlation with the query logs

Figure 3 illustrates the time-series (represented as normalized frequency in monthly bins) between of the query *electoral college*, mining from three different sources (i.e., anchor text, content and query logs (from Google Trend)). We use *cross correlation* (ccf) $f \star g(\tau)$ to measure time series of the two time series. The lagging time τ_{delay} is calculated as $argmax_t(f \star g(t))$. This preliminarily shows

the rather ineffectiveness of the accumulated document frequency in capturing the temporal dynamics of the controversial queries ($ccf = 0.68$ with $\tau_{delay} = 9$). Instead, even with some lagging ($\tau_{delay} = 2$), the correlation between anchor text $f(t)$ and the query logs $g(t)$ is rather high ($ccf = 0.69$). This correlation is further illustrates in Figure 4. This rather shows the value of the anchor text in capturing the temporal dynamics in the web archives. The correlation is not clear for every queries, as it is affected by many factors (e.g., the event-relatedness and its impact). We leave a quantitative evaluation and deeper analysis for future work.

4.3.2 Analysis on the temporal subtopic mining

Figure 5 depicts the temporal dynamics - reflexed by the accumulated document frequency of the subtopic *late-term abortion* from two different time reliability sources. The first one is from the crawling time, the second one is extracted only from our date extraction process with strong level of confidence. Our assumption is high quality pages often follow standard templates and hence their publication dates are often easily extracted by our process. When an event happens (which leads to the trending of some subtopics), the amount of high quality pages issued also gets higher. Hence, we can partially use this second source of time reliability to represent the trending of a subtopic. This brings us a subtopic with the high reliability of timestamp. Figure 5 shows that in the real query log, the subtopic *late-term abortion* get bursted starting from April 2009, and peaked in June 2009. While looking at the crawling time, it starts getting trend from May 2009 to September 2009, peaked in August 2009. However, investigating on the documents crawled in August 2009, we empirically found out that mostly they were published before and only being re-crawled or late-crawled till then.

Figure 6 shows the temporal dynamics of the subtopics underlined the query *abortion*. We see a clear alignment in peaks of the subtopic *late-term abortion* with the main query *abortion* in April 2012 (according to crawling time). It shows that there is temporal correlation in trending of both at a time-point but it is unsure when it is due to the time uncertainty (lagging) in the web archive.

We present another studied query, *health care* in this analysis. Figure 7 depicts the temporal dynamics of the subtopics underlined *health care* over the 2008-2012 period. Even though

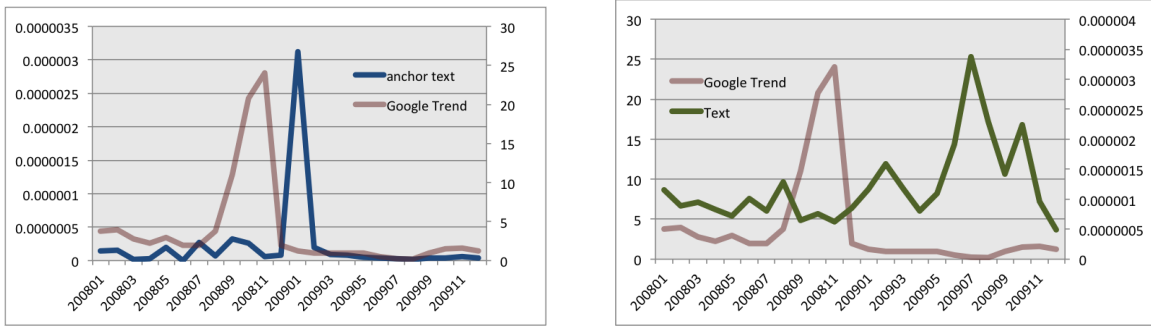


Figure 3: Correlation between time series mined from *anchor text* (left, $ccf = 0.69$, $\tau_{delay} = 2$), *content* (right, $ccf = 0.68$, $\tau_{delay} = 9$) to Google Trend for query electoral college

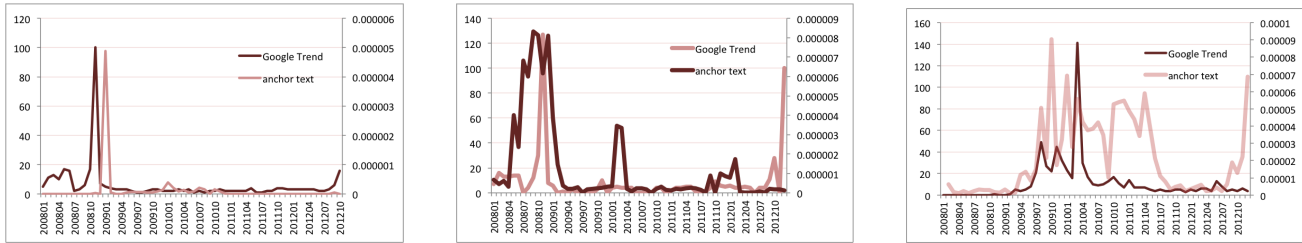


Figure 4: Time series of popular vote ($ccf = 0.94$, $\tau_{delay} = 2$), border fence ($ccf = 0.40$, $\tau_{delay} = 1$) and health care reform ($ccf = 0.44$, $\tau_{delay} = 2$) from *anchor text* and Google Trend from left to right

health care is a broad topic and being discussed all over again, its subtopics however are time-sensitive and trended at certain time-points. Although the crawling time does not provide any time-certainty but it can capture the dynamics of such subtopics, as shown in Section 4.3.1. Figure 8 then illustrates the development of the subtopic *health care reform* with two different time sources, crawling time and the strong confidence. Interestingly, both *crawling time* and the query log become bursty in January 2010. However, a deeper look into the development of the subtopic provided by the reliable time source show that the subtopic is already on trend 2 months earlier. Hence, both the real query log and the crawling time fail to detect the right relevant time for the subtopic. The lagging in the query log can be intuitively understood that the topic has been emerged and discussed in the .gov domain before it receives public attention.

4.3.3 Inferring date for the temporal subtopics

This section provides some insights on determining the relevant time points for the temporal subtopics (mined by the temporal anchor texts), using our methods described in Section 3.2. Table 2 shows the temporal subtopic mining for 3 queries: *abortion*, *border fence* and *health care*. For each subtopic, we also show its corresponding temporal dynamics in Google Trend. For the subtopics of *border fence* the graphs are omitted due to the insufficiency of search volume. We can see that all the subtopics represented show a strong degree of burstiness and hence indicate their time sensitivity. However, identifying these time-points based solely on the timestamp annotations provided by the crawlers is difficult due to the natural lagging of the web archives. Our method that infers the relevant time periods by leveraging the part with strong level of time confidence is shown to be an effective indicator to solve the problem.

5. CONCLUSIONS

In this paper, we have studied the problem of mining temporal subtopics in the web archive. In future work, we will extend it to the time-aware search result diversification task in the web archive context. A further interesting problem is detecting the underlying ‘topic drift’ in this huge longitudinal of multi-modal data collection.

Acknowledgments The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233 and the FP7 project ForgetIT under grant No. 600826.

References

- [1] K. Berberich and S. Bedathur. Temporal diversification of search results. In *TAIA'2013*.
- [2] N. Dai and B. D. Davison. Mining anchor text trends for retrieval. In *Proceedings of ECIR'2010*.
- [3] V. Dang and B. W. Croft. Query reformulation using anchor text. In *Proceedings of WSDM'2010*.
- [4] V. Dang, X. Xue, and W. B. Croft. Inferring query aspects from reformulations using clustering. In *Proceedings of CIKM'2011*.
- [5] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of WSDM'2011*.
- [6] N. Kanhabua and W. Nejdl. On the value of temporal anchor texts in wikipedia. In *TAIA'2014*.
- [7] N. Kanhabua and K. Nørvg. Determining time of queries for re-ranking search results. In *Proceedings of ECDL'2010*.
- [8] T. N. Nguyen and N. Kanhabua. Leveraging dynamic query subtopics for time-aware search result diversification. In *Proceedings of ECIR'2014*.
- [9] W. Song, Y. Zhang, H. Gao, T. Liu, and S. Li. HITSCIR system in NTCIR-9 subtopic mining task. 2014.

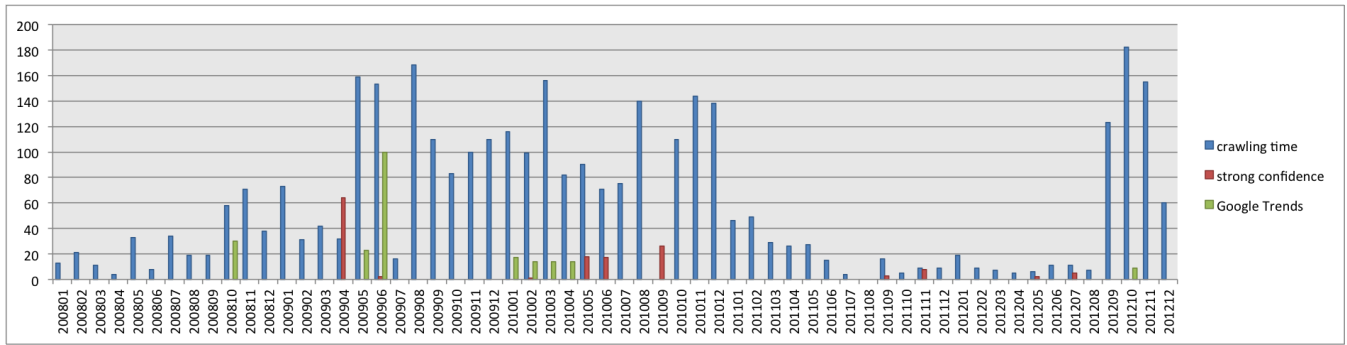


Figure 5: The temporal dynamics - reflexed by the accumulated document frequency of the subtopic *late-term abortion* from two different time reliability sources (crawling time and strong confidence) and from Google Trend.

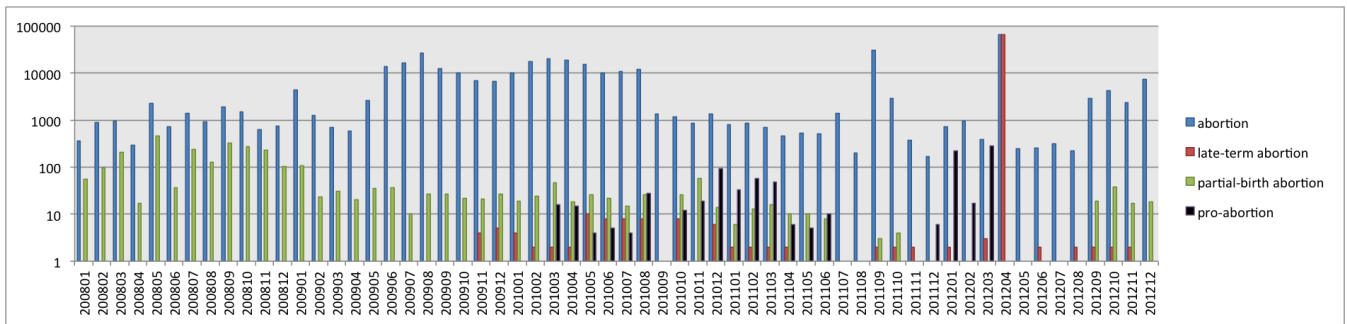


Figure 6: The temporal dynamics of the query *abortion* and its subtopics over time - reflexed by the accumulated frequency of anchor texts.

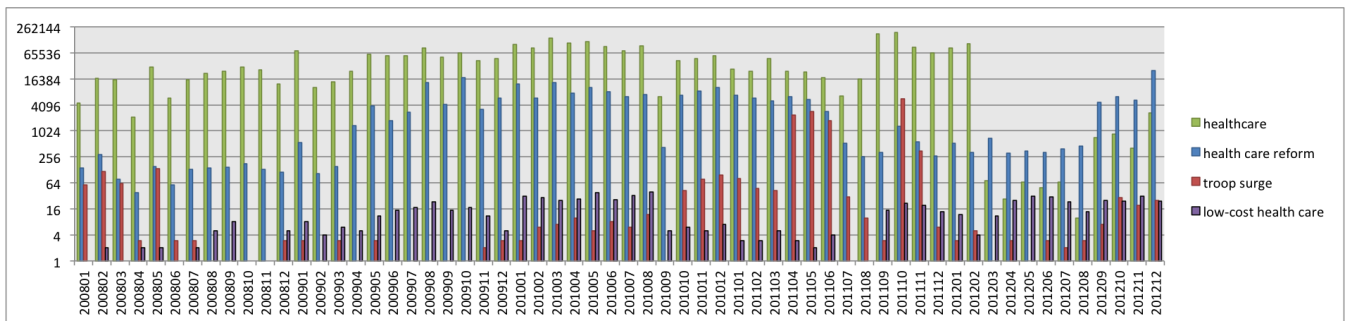


Figure 7: The temporal dynamics of the query *health care reform* and its subtopics over time - reflexed by the accumulated frequency of anchor texts.

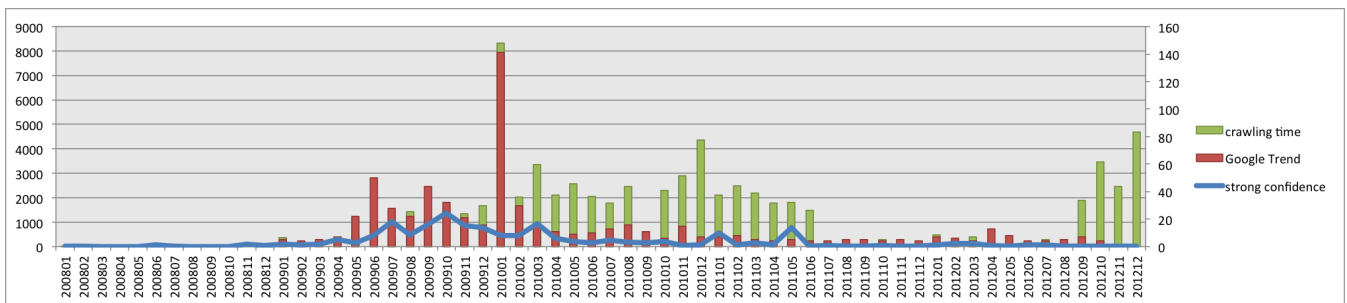


Figure 8: The temporal dynamics - reflexed by the accumulated document frequency of the subtopic *health care reform* from two different time reliability sources (crawling time and strong confidence) and from Google Trend.