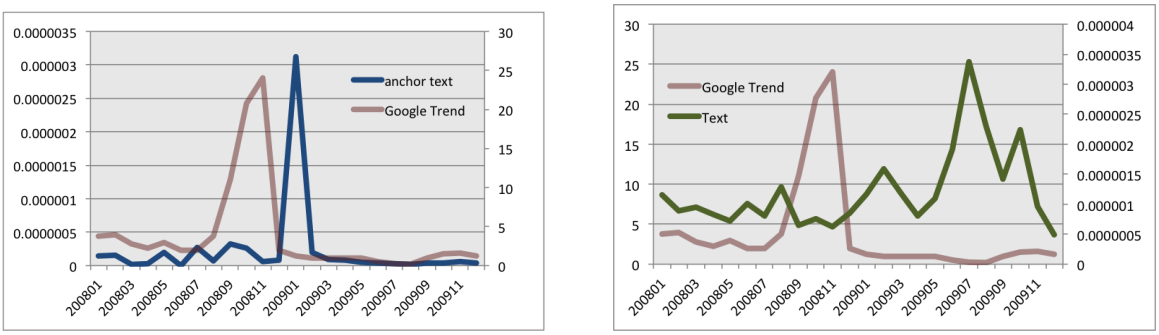

Figure 3: Correlation between time series mined from *anchor text* (left, $ccf = 0.69$, $\tau_{delay} = 2$), *content* (right, $ccf = 0.68$, $\tau_{delay} = 9$) to Google Trend for query electoral college

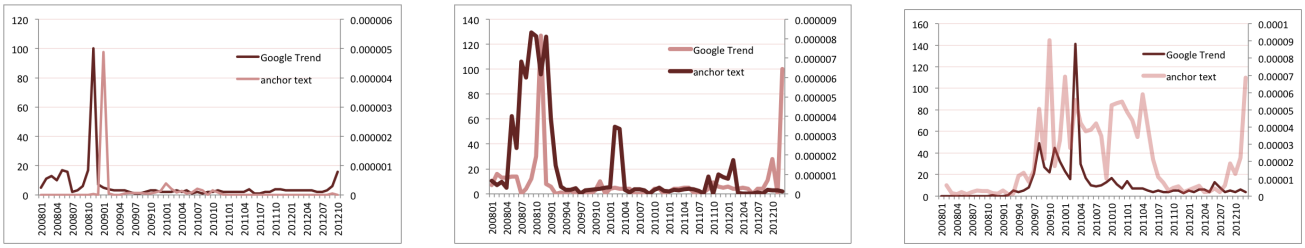

Figure 4: Time series of popular vote ($ccf = 0.94$, $\tau_{delay} = 2$), border fence ($ccf = 0.40$,$\tau_{delay} = 1$) and heath care reform ($ccf = 0.44$, $\tau_{delay} = 2$) from *anchor text* and Google Trend from left to right

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. Nørvåg. 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.

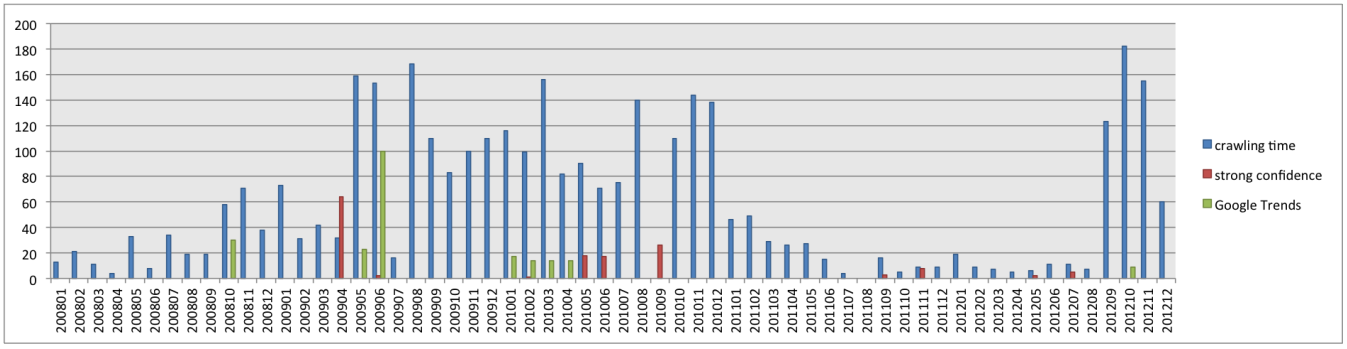

Figure 5: The temporal dynamics - reflexed by the accumulated document frequency of the subtopic late-term abortion from two different time reliability sources (crawling time and strong confidence) and from Google Trend.

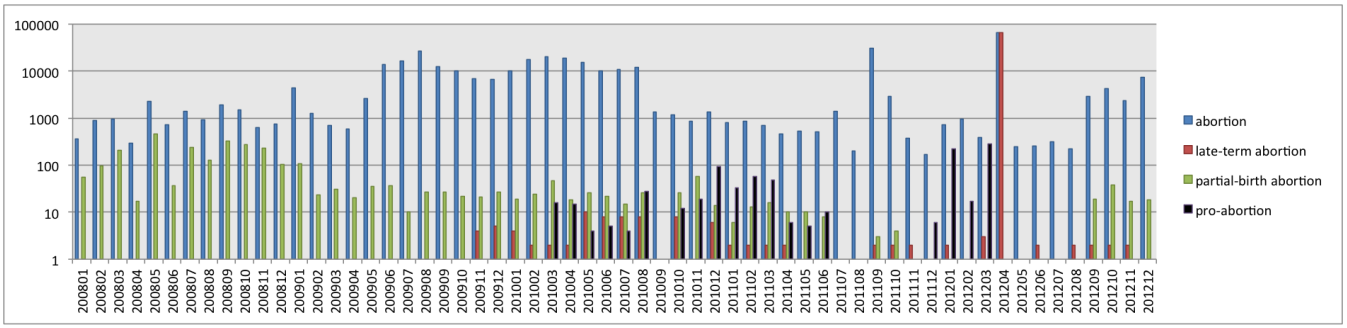

Figure 6: The temporal dynamics of the query abortion and its subtopics over time - reflexed by the accumulated frequency of anchor texts.

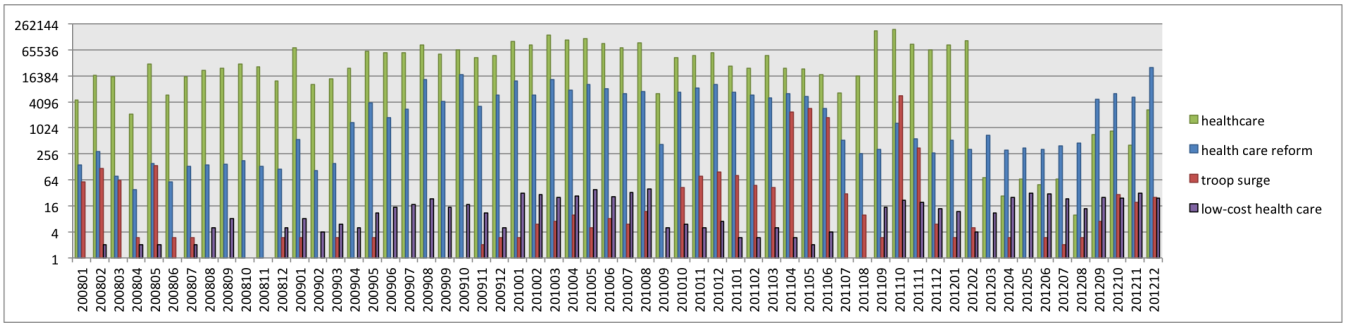

Figure 7: The temporal dynamics of the query health care reform and its subtopics over time - reflexed by the accumulated frequency of anchor texts.

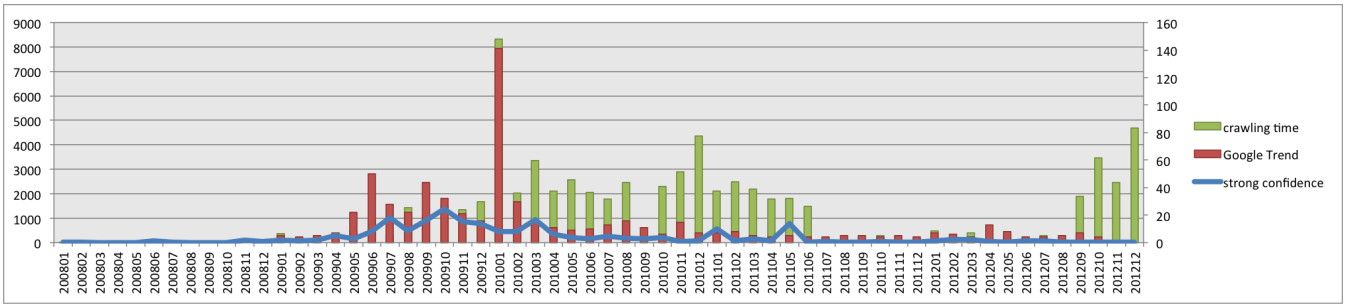

Figure 8: The temporal dynamics - reflexed by the accumulated document frequency of the subtopic health care reform from two different time reliability sources (crawling time and strong confidence) and from Google Trend.