EXTRACTING EVENT-CENTRIC DOCUMENT COLLECTIONS FROM LARGE-SCALE WEB ARCHIVES

Gerhard Gossen, Elena Demidova; L3S Research Center, Leibniz Universitat Hannover; Thomas Risse; University Library J.C. Senckenberg, University Frankfurt

Introduction

Web archives created by the Internet Archive (IA) (https://archive.org), national libraries and other archiving services contain large amounts of information collected for a time period of over twenty years. These archives constitute a valuable source for research in many disciplines, including the digital humanities and the historical sciences by offering a unique possibility to look into past events and their representation on the Web.

Most Web archive services aim to capture the entire Web (IA) or national top-level domains and are therefore broad in their scope, diverse regarding the topics they contain and the time intervals they cover. Due to the large size and the broad scope it is difficult for interested researchers to locate relevant information in the archives as search facilities are very limited. Many users are more interested in studying smaller and topically coherent event-centric collections of documents contained in a Web archive [1,2]. Such collections can reflect specific events such as elections, or natural disasters, e.g. the Fukushima nuclear disaster (2011) or the German federal elections.

Event-Centric Collection Extraction

Events are typically characterized through a certain date or a time interval. Nevertheless, event-related documents also appear outside of this time interval. For planned and regularly recurring events such as sports competitions or elections, relevant documents are often published in advance of the actual begin of the event during the event lead time, and are still published after the event completion during the cool-down time period. For non-recurring events such as natural disasters, event-related documents are published from the start of the event and during the cool-down time. Given an event of user interest and a large-scale broad-scope Web archive, our goal is to generate an interlinked collection of documents relevant to this event. A naive approach to create an event-centric collection is to iterate through all documents in a Web archive and check their relevance using an automatic method. However, this is computationally expensive and does not scale. While a full-text index could reduce the iteration cost, it requires high up-front computation [3]. Furthermore, such an index can only be used to retrieve individual documents instead of interlinked documents. In [4] we proposed an alternative approach that uses the hypertext characteristics of the archived documents by adapting focused Web crawling. A focused Web crawler collects documents by recursively following the prioritized links from a Web document to other documents within the archive.

Relevance Estimation

For Web Archive re-crawling we need to prioritize the URLs during the focused crawl to effectively extract event-centric collections based on a relevance function. The relevance function we proposed is a linear combination of the temporal and topical relevance of a Web document.

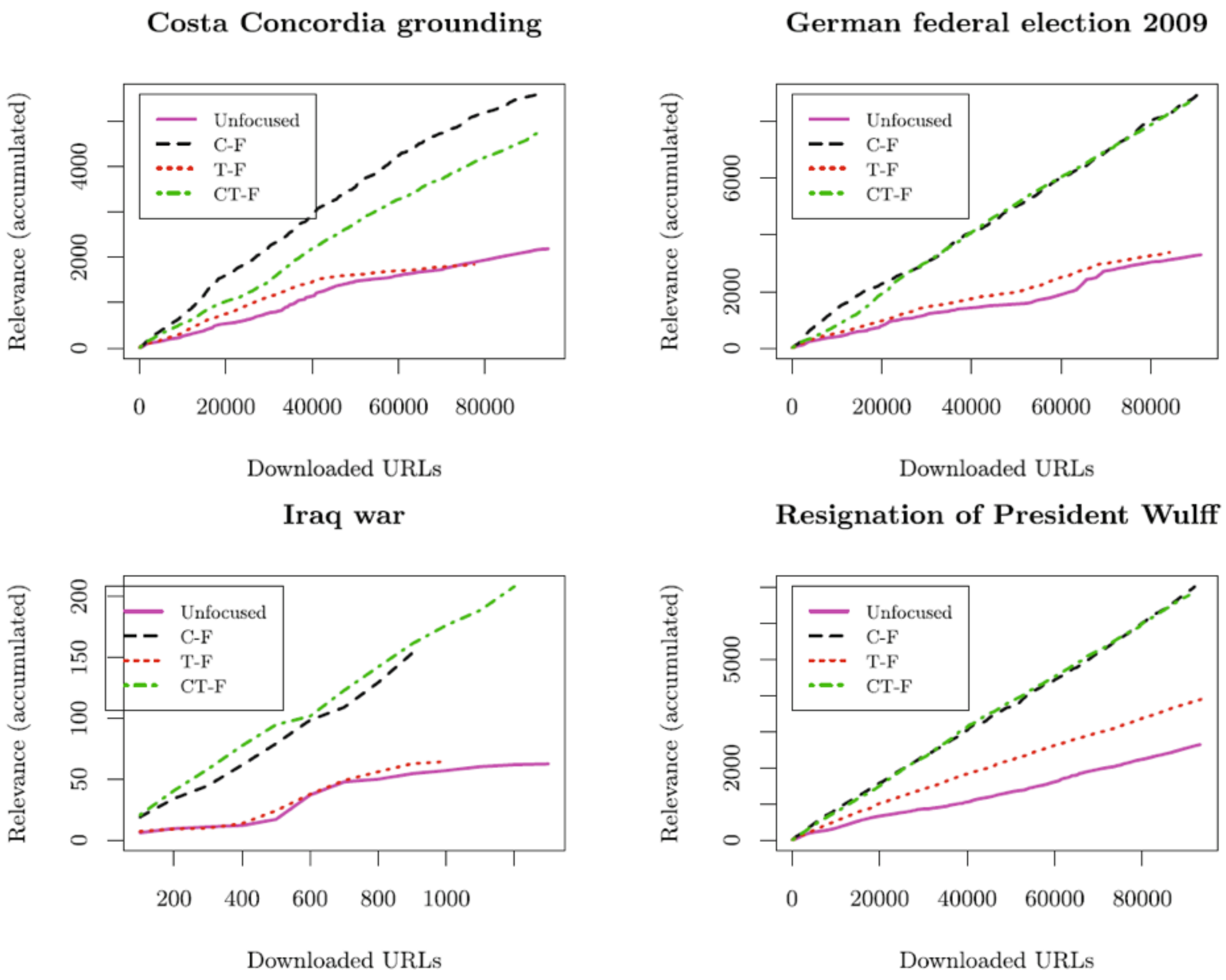
Temporal relevance is estimated based on a time point associated with the Web document (e.g. the creation or capture date). We assume that in general the relevance of documents decreases rapidly as the distance to the event time increases and therefore define a temporal relevance function based on the exponential decay function.

The topical relevance of Web documents is estimated by computing the similarity of the textual content of Web documents to the topical scope of the collection. The topical scope is specified primarily through a set of reference documents that describe the event (e.g. Wikipedia pages). In case of an ambiguous event description or if the scope of the collection should be narrowed down further, keywords can be provided to clarify the topical intent.

Evaluation

The Web Archive we used for the evaluation consists of 4.05 billion captures of Web pages from the .de top-level domain as collected by the Internet Archive between 1994 until 2013. We manually defined 28 events to be extracted from the Web archive likely to be represented in the archive like the Fukushima nuclear accident or German federal elections. The topical scope of the events is described by relevant Wikipedia pages. We also defined a start and end date, as well as an estimated event lead and cool-down time. The outgoing links of the Wikipedia pages were used as seed URLs.

The goal of the evaluation is to assess the precision of the proposed collection extraction method in light of different event types and to better understand the influence of this method on the quality of the resulting event-centric collections. We compare our combined relevance function (CT-F) with two baselines that use state-of-the-art functions for topical [5] (C-F) or temporal [6] (T-F) relevance estimation. We also use an unfocused crawl that does not use any relevance estimates as an additional baseline. For each of the 28 events we started a crawl using each of the configurations described above. Each crawl ran until it had retrieved 100,000 documents or until the crawler queue was empty.


Costa Concordia grounding

German federal election 2009

Iraq war

Resignation of President Wulff

Figure 1: Accumulated relevance of different event collections: combined relevance (CT-F), topical relevance [5] (C-F), temporal relevance [6] (T-F).

Figure 1 shows the accumulated relevance of document collections for selected events in relation to the number of documents crawled. This function should ideally start with a strong incline, meaning that the crawler fetches many relevant documents early on, flattening into a plateau when no relevant documents are available anymore. We see that for all topics the combined function outperforms the temporal function and the unfocused baseline both in terms of average relevance of documents retrieved at any given point and total relevance. The content function often performs slightly better than the combined function. The relevance focused strategies manage to uncover more potentially relevant URLs even if they are not contained in the locally available Web archive.

We crawled each event using an exponential decay function with a fixed decay and compared it to the crawl using the user-specified lead and cool-down times. We see that the event-specific parameters cause a significant improvement for most of the events (see [4]).

Conclusions

We presented a summary of our work in [4] to create interlinked event-centric document collections from Web archives by adapting focused Web crawling. We showed that our re-crawling method can effectively retrieve event-centric collections. Further research is needed to better understand the influence of extraction methods, relevance functions and parameters in regard to different events, time periods and Web archives.

References

[1] Gossen, G., Demidova, E., Risse, T.: Analyzing web archives through topic and event focused sub-collections. In: WebSci 2016. pp. 291–295, May 2016

[2] Risse, T., Demidova, E., Gossen, G.:What do you want to collect from the web? In: Proceedings of the Building Web Observatories Workshop (BWOW) 2014

[3] Jackson, A., Lin, J., Milligan, I., Ruest, N.: Desiderata for exploratory search interfaces to web archives in support of scholarly activities. In: JCDL2016 (2016)

[4] Gossen, G., Demidova, E., Risse, T.: Extracting Event-Centric Document Collections from Large-Scale Web Archives. Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings (p. 116--127). Springer, Cham (2017)

[5] Pant, G., Srinivasan, P., Menczer, F.: Crawling the web. In: Web Dynamics (2004)

[6] Pereira, P., Macedo, J., Craveiro, O., Madeira, H.: Time-aware focused web crawling. In: ECIR 2014. LNCS, vol. 8416, pp. 534–539. Springer, Cham (2014).