DOI: http://dx.doi.org/10.14236/ewic/FDIA2015.6

Temporal Information Retrieval Revisited

A Focused Study on the Web

Yue Zhao
Web Information Systems
Delft University of Technology
The Netherlands
y.zhao-1@tudelft.nl

Claudia Hauff
Web Information Systems
Delft University of Technology
The Netherlands
C.Hauff@tudelft.nl

Temporal information retrieval has been an active area of research for a number of years. Most existing works focus on the utility of temporal information in specific types of corpora (such as news archives), specific types of retrieval approaches, and, specific applications that may benefit from temporal information (such as timeline summarization). Underrepresented in existing works are studies that investigate the impact of temporal information analyses on the Web and Web documents. In this paper we (i) describe the research gaps we identified around Web-based temporal information analysis, and, (ii) present an overview of our first results and observations when studying *sub-document timestamping* on the Web.

Keywords: Temporal information analysis, timestamping, sub-doucments, information diffusion

1. INTRODUCTION

One recurring theme in temporal information retrieval and analysis is the use of document creation timestamps for various tasks and applications, such as event detection [Döhling and Leser (2014)], document clustering [Alonso et al. (2009)] and the adaptation of retrieval algorithms to temporal queries [Li and Croft (2003)]. Determining the creation time of a Web document is challenging for a number of reasons: (1) document meta-data is generally unreliable, (2) public sources such as the Internet Archive (https://archive.org/) can only archive a small subset of the Web, and, most important of all, (3) the average Web document may change significantly over the course of its lifetime, in which case a single document creation timestamp is effectively only capturing when the Web document's location (URL) was first established on the Web, instead of when the document's content was created.

Existing works have largely addressed these challenges by relying on the document content itself to estimate a single creation date [de Jong et al. (2005); Kanhabua and Nørvåg (2009); Kumar et al. (2011); Chambers (2012); Ge et al. (2013)]. This approach has been shown to work well for document corpora that are static in nature, such as news corpora — each document is a single news article with few to no changes in content over time. Driven by the lack of research in *sub-document* timestamping (i.e. the labelling of document sentences and paragraphs with individual creation timestamps), we set out to investigate to what extent this simplifying assumption of a single creation timestamp also holds for Web documents.

This initial experiment (described in more detail in Section 2) led us to identify a number of perceived research gaps in Web-based temporal information analysis that we aim to investigate over the course of the next three years:

Web sub-document timestamping: Existing document timestamping algorithms [de Jong et al. (2005); Kanhabua and Nørvåg (2009); Kumar et al. (2011); Chambers (2012)] do not take the special nature of the Web into account, including its link structure and dynamic nature. Are we able to increase the accuracy of sub-document timestamping when utilizing this knowledge?

From News to Web corpora: Existing temporal retrieval approaches have been shown to outperform non-temporally aware approaches on news corpora [Berberich et al. (2010)]. To what extent does temporality aid in the retrieval of general Web documents?

Novel Applications: Assuming that sub-document timestamping is possible at Web scale, what types of novel applications or tasks can we solve? What kind of novel insights can we gain (e.g. about information diffusion [Yang and Leskovec (2010)])?

In the remainder of this paper, we will first summarize our preliminary research on the timestamping of Web sub-documents (Section 2), followed by a broad outline of the research plan (Section 3) and an overview of open questions to be discussed at FDIA (Section 4).

2. SUB-DOCUMENT TIMESTAMPING

Our initial investigation (published in [Zhao and Hauff (2015)]) revolved around the assumption made in existing works utilizing document creation timestamps [Swan and Jensen (2000); Li and Croft (2003); Alonso et al. (2009); Jatowt et al. (2013); Döhling and Leser (2014)]: each document (no matter the corpus) is created at one point in time.

Although it is obvious that for Web documents this assumption generally does not hold, it is not yet known, to what extent the assumption is wrong. We designed an empirical analysis to investigate this issue, answering the following research questions:

RQ1 To what extent do Web documents consist of sub-documents created at different times?

RQ2 What is the timespan between the oldest and most recent sub-document of a document?

RQ3 What fraction of the current document has been created in each version (a version corresponding to a particular timestamp)?

RQ4 To what extent can the timestamp of each sub-document be predicted?

2.1. Data Set

Due to the preliminary nature of our study, we investigated a small sample of Web documents in depth, specifically the 11,075 relevant documents of the ClueWeb12 corpus (http://www.lemurproject.org/clueweb12.php/, TREC Web topics 201-300). This choice ensured that each investigated Web document is at least relevant to some information need.

2.2. Processing Pipeline

We first divided each of these ClueWeb12 documents into sub-documents (each sub-document is simply a paragraph). In order to learn when each sub-document was created, we crawled all historic versions of the document stored before 2012 (the crawl date of ClueWeb12) that are available at the Internet Archive — overall, 64% of our documents were captured by the Internet Archive with on average 17 historical versions.

Based on these historical versions, we determined the earliest version in which each sub-document occurred and assigned the Internet Archive crawl

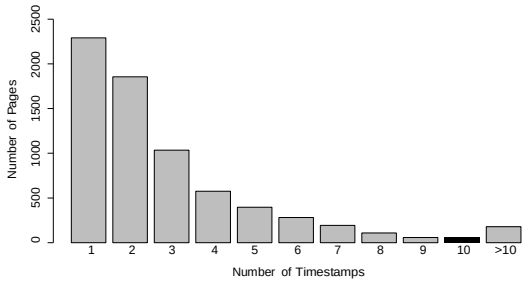

Figure 1: *Overview of the number of documents containing content created at different points in time.*

date of that version as sub-document creation timestamp. First, we extract all textual contents and divide them into sub-documents by HTML tags (e.g. <p> and). Subsequently, these sub-documents are compared to sub-documents in the historical versions. If two sub-documents have a high similarity, they are treated as being the same and sub-documents in ClueWeb12 documents are timestamped by their earliest occurrence in the Internet Archive.

Lastly, we also extracted more than 20 features (such as the length and position of sub-documents and the number and value of temporal expressions in sub-documents) from each sub-document in order to train & test our timestamp classifier. Based on the ground truth sub-document timestamps, we created five timestamp classes by dividing them in balance, which means that each class has ∼55K instances. For example, the time intervals of the first two classes are $A = [0, 20.5], B = (20.5, 311.5]$, which means that the sub-documents in class A have been created no later than 20.5 days before the ClueWeb12 document's crawl date, while sub-documents in class B have been created between 20.5 and 311 days before the document's crawl date respectively. We aim to predict each sub-document's timestamp class correctly.

2.3. Results

The results shown in Figure 1 indicate that the majority of Web documents have indeed more than one creation timestamp (answering **RQ1**): 67% of Web documents consist of sub-documents with at least two different creation timestamps, with only a small percentage (less than 4%) having more than 8 creation timestamps.

When considering the difference in days between a document's oldest and most recent sub-document creation timestamp (**RQ2**) we find a surprisingly large gap in Figure 2: the median difference is 400

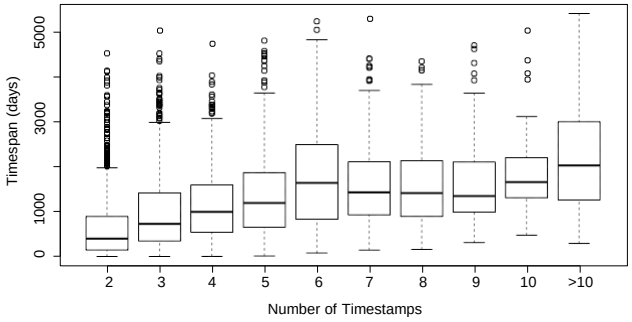

Figure 2: *The document set is partitioned according to the number of creation timestamps (documents with a single creation timestamp are ignored). Shown is the difference (in days) between the oldest and most recent creation timestamp.*

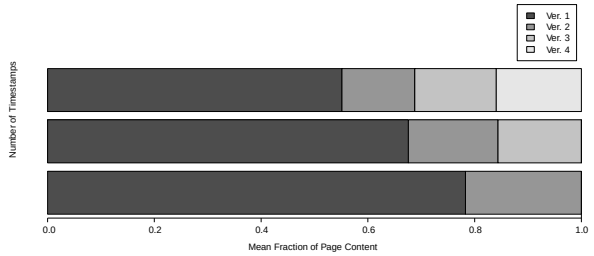

Figure 3: *Overview of content created at different points in time for documents with 2, 3 or 4 creation timestamps. Each bar shows the mean fraction of content available at each creation timestamp.* Ver. 1 *indicates the content created at the oldest timestamp,* Ver. 2 *the content created at the second oldest timestamp and so on.*

days, i.e. 50% of our investigated Web documents contain content created more than one year apart.

Lastly, we turned our attention to the amount of content created at each point in time (**RQ3**), restricting our analysis to those documents with 2, 3 or 4 creation timestamps as shown in Figure 3. Evidently, most of a Web document's content is created initially; the more creation timestamps a document has, the lower the percentage of initially created content. Moreover, we find that the average contents updated in each time are similar, which has also been found in earlier studies [Fetterly et al. (2004); Ntoulas et al. (2004)].

Having analysed the extent of a Web document's content being created at different points in time, we also experimented with the prediction of the correct timestamp class (**RQ4**). In a 5-class setup and our 20+ features we were able to classify 64% of all sub-documents correctly, significantly better than our baseline classifier which considered only the temporal expression within each sub-document for classification purposes (39% accuracy).

These results indicate that utilizing the creation times of sub-documents (instead of a single creation time per document) are likely to have a significant effect on Web retrieval tasks that utilize this type of temporal information.

3. RESEARCH PLAN

Based on these encouraging results, we developed a research plan for the coming three years, revolving around temporal information analysis on the Web.

We aim to investigate the following research themes:

Web sub-document timestamping: Our first goal is to drastically scale up the analysis of sub-document timestamping along the lines of the preliminary experiments. Ideally, instead of investigating 11,000

Million ClueWeb12 documents. This change in magnitude will allow us to also effectively investigate novel features for the prediction of sub-document timestamps such as the link structure *between* sub-documents and time-series based features (due to the dynamic nature of the Web). Additionally, we aim to move beyond standard classification towards a more fine-grained approach using sequential labeling methods [Lafferty et al. (2001); Dietterich (2002)] which can exploit the relationships among a document's sub-documents.

From News to Web Corpora: While past works have mostly investigated news corpora (partially due to their unambiguity in creation timestamps), we aim to extend these works by employing proposed (as well as newly developed) retrieval approaches to Web corpora, investigating the impact of sub-document timestamps on retrieval effectiveness.

Novel Applications: The envisioned large-scale nature and at the same time fine-grained analysis of sub-document timestamping will allow us to consider novel applications, such as investigating the effect of the rise (or decline) of particular portals on information diffusion on the Web. Information diffusion on the Web might be influenced by specific websites. For example, programmers prefer to talk about programing problems on StackOverflow nowadays rather than writing their problems on their blogs as before.

4. OPEN QUESTIONS

There are a number of open questions, that would be particularly interesting to discuss during FDIA.

(1) The main limitation of our work is the restricted accuracy of the sub-document timestamps we generate with our Internet Archive-based

methodology. While for popular Web documents the crawling frequency is high, unpopular documents are crawled infrequently and thus, the timespan between the crawling time and the real creation time of a sub-document may be large. Are there ways to improve the methodology to generate more accurate timestamps for unpopular Web documents?

(2) We expect that in contrast to other research areas where humans achieve a very high accuracy (e.g. face recognition or image content labelling), sub-document timestamping to be very challenging for human labellers. This leads to the questions of (i) how to measure when a prediction is accurate enough, and, (ii) how to determine whether or not there is a hidden ceiling for the prediction accuracy of sub-document timestamping?

(3) In previous works on temporal information retrieval, the temporal information is leveraged as filters or additional conditions. Is it possible to combine temporal and other features (e.g. in a learning to rank setting) as innate features rather than additional restrictions?

5. ACKNOWLEDGEMENT

We thank the ESF ELIAS for the awarded scholarship for participating in ESSIR 2015 and the FDIA Symposium.

REFERENCES

Alonso, O., M. Gertz, and R. Baeza-Yates (2009). Clustering and exploring search results using timeline constructions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 97–106. ACM.

Berberich, K., S. Bedathur, O. Alonso, and G. Weikum (2010). *A language modeling approach for temporal information needs*. Springer.

Chambers, N. (2012). Labeling documents with timestamps: Learning from their time expressions. In *ACL '12*, pp. 98–106.

de Jong, F., H. Rode, and D. Hiemstra (2005). Temporal language models for the disclosure of historical text. Royal Netherlands Academy of Arts and Sciences.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. In *Structural, syntactic, and statistical pattern recognition*, pp. 15–30. Springer.

Döhling, L. and U. Leser (2014). Extracting and aggregating temporal events from text. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pp. 839–844. International World Wide Web Conferences Steering Committee.

Fetterly, D., M. Manasse, M. Najork, and J. L. Wiener (2004). A large-scale study of the evolution of web pages. *Software – Practice & Experience 34*(2), 213–237.

Ge, T., B. Chang, S. Li, and Z. Sui (2013). Event-based time label propagation for automatic dating of news articles. In *EMNLP '13*, pp. 1–11.

Jatowt, A., C.-M. Au Yeung, and K. Tanaka (2013). Estimating document focus time. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 2273–2278. ACM.

Kanhabua, N. and K. Nørvåg (2009). Using temporal language models for document dating. In *Machine Learning and Knowledge Discovery in Databases*, pp. 738–741.

Kumar, A., M. Lease, and J. Baldridge (2011). Supervised language modeling for temporal resolution of texts. In *CIKM '11*, pp. 2069–2072.

Lafferty, J., A. McCallum, and F. C. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Li, X. and W. B. Croft (2003). Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, pp. 469–475. ACM.

Ntoulas, A., J. Cho, and C. Olston (2004). What's new on the Web?: the evolution of the Web from a search engine perspective. In *Proceedings of the 13th International Conference on World Wide Web*, New York, NY, USA, pp. 1–12. ACM.

Swan, R. and D. Jensen (2000). Timemines: Constructing timelines with statistical models of word usage. In *KDD Workshop on Text Mining*, pp. 73–80.

Yang, J. and J. Leskovec (2010). Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 599–608. IEEE.

Zhao, Y. and C. Hauff (2015). Sub-document timestamping of web documents. In *Proceedings of the 38th international ACM SIGIR conference on Research & development in information retrieval*. ACM.