DIGITALHISTORIAN: Search & Analytics Using Annotations

Dhruv Gupta[1,2], Jannik Strötgen[1], and Klaus Berberich[1,3]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
[2] Saarbrücken Graduate School of Compute Science, Saarbrücken, Germany
[3] htw saar, Saarbrücken, Germany
{dhgupta, jannik.stroetgen, kberberi}@mpi-inf.mpg.de

Abstract

Born-digital document collections contain vast amounts of historical facts and knowledge. However, manual assessment of these large text collections is infeasible. In this paper, we demonstrate a retrieval system, DIGITALHISTORIAN, that analyzes these document collections using semantic annotations in the form of temporal expressions and named entities linked to a knowledge graph. For queries about entities or events DIGITALHISTORIAN utilizes state-of-the-art methods to understand and analyze temporal expressions in the content of documents. It understands uncertainty in temporal expressions and uses them to mine interesting time intervals for keyword queries. These time intervals are further used for re-ranking and diversifying documents, so that the ranked list of documents portray a historic overview of the query. Further, to contextualize the interesting time intervals, we use frequently occurring named entities and display them in informative visualizations. DIGITALHISTORIAN is designed to help scholars in *digital humanities* explore large document collections quickly without having any prior knowledge about interesting time intervals or entities for given keyword query.

1 Introduction

Large born-digital document collections cannot be analyzed by manual human effort. Nowadays, however, they can be automatically annotated with temporal expressions and named entities. Time has been found to be a very important part of generic Web queries; recent studies [9] estimate that around 17.1% of them are *implicitly* time-sensitive in nature. In the context of *digital humanities*, these figures can be expected to be much higher. The *desiderata* that we believe many commercial search engines do not currently meet and which will be useful for scholars in *digital humanities* are:

- ▷ the ability to automatically suggest *interesting* time intervals for *history-oriented* queries;
- ▷ the ability to diversify or re-rank documents using temporal expressions;
- ▷ the ability to establish relationships between the time intervals of interest for query and other evidences in text such as named entities;
- ▷ the ability to visually analyze the different relationships established between the annotations in documents.

In this paper, we demonstrate DIGITALHISTORIAN, a system that leverages the semantic information in documents to retrieve better search results for *history-oriented* queries. We define *history-oriented* queries to consist of keywords that describe an entity (e.g., george w bush) or an event (e.g., economic depression). Our system analyzes temporal expressions in documents to identify interesting time intervals, and subsequently uses them for diversifying search results. The interesting time intervals can also be selected to expand the query to gather search results concerning that particular time interval. Furthermore, DIGITALHISTORIAN can construct visualizations that display frequent named entities in interesting time intervals identified for the history-oriented query.

M. Düring, A. Jatowt, J. Preiser-Kapeller, A. van den Bosch (eds.): Proceedings of the 3rd HistoInformatics Workshop, Krakow, Poland, 11 July 2016, published at http://ceur-ws.org

Organization. The rest of the paper is organized as follows. In Section 2, we give a brief description of the key methods applied for mining interesting time intervals, and how these are used for search result re-ranking and diversification. In Section 3, we describe the technical building blocks of DIGITALHISTORIAN. In Section 4, we describe how DIGTALHISTORIAN can be utilized to explore document collections. We put our system in context to related systems in Section 5 and summarize the contributions in Section 6.

2 History by Algorithms

The underlying methods for temporal search in DIGITALHISTORIAN are derived from our prior research [1, 2, 3, 4]. We next give a brief description of these methods.

Understanding Time. Temporal expressions in documents can be highly uncertain, for example `1990s`. For such temporal expressions it is unclear how the time interval should be constructed for further analysis. In order to represent such ambiguities in time, we use the time model proposed by Berberich et al. [1] which allows for *relaxations* on the begin and end of time intervals. Thus, `1990s` may convey a time interval that can begin anywhere from `[1990,1999]` and end anywhere from `[1990, 1999]`. This new representation thus allows us to perform mathematical manipulations on uncertain and ambiguous temporal expressions.

Time Intervals of Interest to Queries. Given a history-oriented query such as george w bush, our approach [2] can identify interesting time intervals (e.g., `[2000,2004]`, `[2004,2008]`) by analyzing the temporal expressions in its pseudo-relevant set of documents. This is achieved in two steps. First, by counting the frequency of time intervals in the uncertainty-aware time model described earlier. Second, by weighting each frequent time interval with relevance of the document to the query. This is done recursively to generate interesting time intervals at year, month, and day level granularity.

Re-ranking Documents Using Time. The ranking of the initial set of pseudo-relevant documents can be refined by using one of the interesting time intervals for query expansion. Consider for example the query george w bush reformulated with the time interval `[2000,2004]`. The documents which contain temporal expressions that can generate the time interval in the query more frequently and also have a higher textual relevance to the query will be promoted in the rankings [1]. Hence, all documents that are relevant to the time interval will be higher in the rankings.

Diversifying Documents Using Time. The time intervals of interest can be considered to reflect different temporal aspects underlying the query. The initial pseudo-relevant set of documents can then be diversified so as to contain *at least* one document relevant to the different temporal aspects [4]. The temporally diverse set of documents can thus be viewed as a biography of an entity or a timeline of an event.

Counting Frequent Named Entities. The time intervals of interest can further be used as a basis for aggregating different annotations in text. In particular, we aggregate the occurrence of *unique* named entities. E.g., given the query george w bush and the time interval `[2000,2004]`, we obtain the aggregate counts of co-occurring named entities such as al gore.

3 Architecture

The key building blocks of DIGITALHISTORIAN are: a document collection, semantic annotators, an information retrieval framework, a visualization engine, and the graphical user interface. We describe each of them briefly in the following paragraphs.

Document Collection. We used the *The New York Times Annotated Corpus*,[1] a collection of news articles published in *The New York Times* between 1987 to 2007. It comprises of roughly two million news articles. As metadata, we used only the publication dates.

Semantic Annotators. For annotating temporal expressions we utilized the HEIDELTIME temporal tagger [14]. HEIDELTIME annotates *implicit*, *explicit*, and *relative* temporal expressions. For disambiguating and linking named entities to the YAGO[15] knowledge graph, all documents were processed with AIDA [7].

Information Retrieval Framework. All the documents along with their annotations were indexed with the ELASTICSEARCH[2] framework. As a baseline retrieval model for pseudo-relevant documents we used the *Okapi BM25* method implemented in ELASTICSEARCH. All our methods for temporal search and aggregation were implemented in the JAVA language.

Visualization and GUI. For generating visualizations, we used the BRUNEL VISUALIZATION[3] API. The entire graphical user interface was programmed using JAVA's SWING API.

4 Demonstration

As outlined in the following, there are two key use cases that we would like to demonstrate with DIGITALHISTORIAN. We also describe how the users will be able to interact with DIGITALHISTORIAN using illustrations for the different use cases.

Exploring Search Results. The foremost task that we address with DIGITALHISTORIAN is that of exploring the document collection using the interesting time intervals identified for the query. The main view (Figure 1) of the DIGITALHISTORIAN addresses this by providing the user with a search field to issue keyword queries. Subsequent to the search operation, various interesting time intervals are displayed in a list on the left hand-side of the interface. The list of time intervals are ordered by their *interestingness*, i.e., how frequently they are generated by the temporal expressions in document contents for the given query. A diversified set of documents is shown in the main display which gives a temporal overview of documents for the query. Each document in the list is depicted by its headline, its URL, a snippet from its contents and the normalized temporal expressions in its contents. Furthermore, the users can double-click the various time intervals in the list to expand the query, so as to obtain more documents concerning it. Unlike many commercial search engines, all of this is done automatically, without imposing any sliders or check-boxes to manually specify relevant time intervals.

Analytical Visualizations. We further construct informative visualizations by contextualizing the interesting time intervals with co-occurring named entities. Currently, there are two analytical visualizations available. Both of them show the frequency of various named entities that occur in different time intervals. The first visualization is a *chord diagram* (Figure 2a), where an arc is drawn between a time interval and a corresponding named entity that occurs in that time interval. The thickness of the arc is in proportion to the frequency of the named entity in that time interval. The user can also hover over to each individual chord in the graph to see the time interval and the entity it connects and the corresponding aggregate count. The second visualization is a *heatmap diagram* (Figure 2b), where on the *x-axis* the different time points and on the *y-axis* different named entities are plotted. The intensity of the cell in the heatmap shows the frequency of that named entity in that time point. The user can further drill up from years to decades and drill down from years to days by scrolling on the time axis to inspect the different time intervals with their respective frequency of named entities.

[1] https://catalog.ldc.upenn.edu/LDC2008T19
[2] https://www.elastic.co/
[3] https://github.com/Brunel-Visualization/Brunel

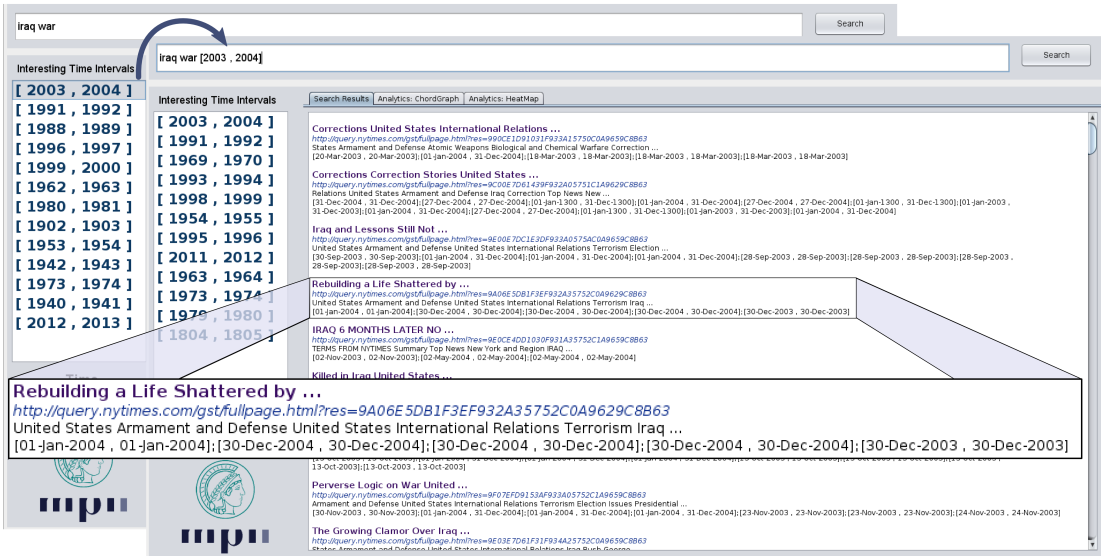
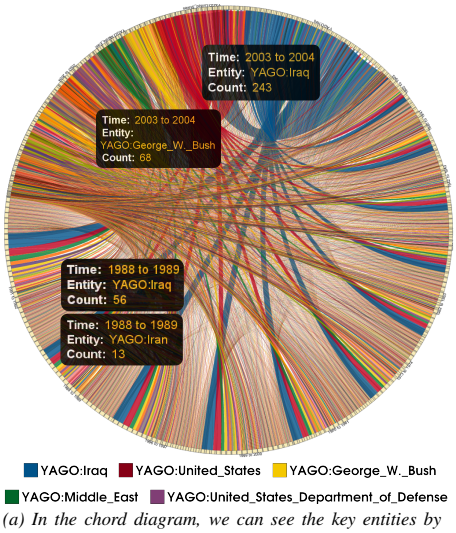
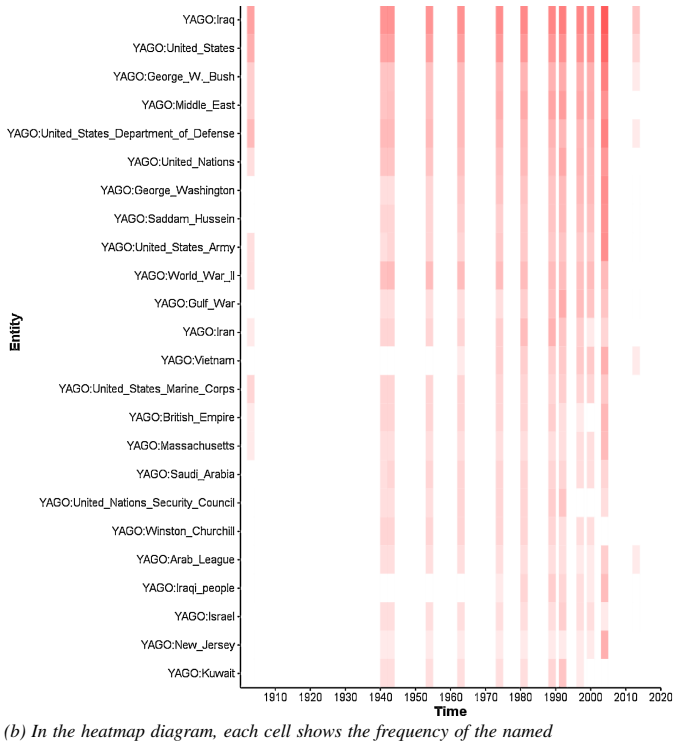


Figure 1: The GUI of DIGITALHISTORIAN. *Users can type in keyword queries in the search text field and* DIGITALHISTORIAN *will automatically determine interesting time intervals for it. The users can also double-click one of the many intervals in the list to expand the query and retrieve the search results with that time interval. In the illustration it's shown how the user selects the time interval* `[2003,2004]` *to expand the query* `iraq war` *and obtains search results for that particular time interval.*



(a) In the chord diagram, we can see the key entities by large chords corresponding to `iraq`, `george w bush`, *and* `iran`. *Iraq was involved in multiple conflicts which can be seen in the highlighted boxes, i.e., the American invasion of Iraq in 2003 and the Iran-Iraq war of 1980-1988. The most frequent entities are highlighted in the legend*



(b) In the heatmap diagram, each cell shows the frequency of the named entity in the corresponding time point.

Figure 2: Various analytical visualizations of frequent entities in different time intervals for the query `iraq war`.

5 Related Work

There exists few demonstrations that utilize named entities and temporal expressions for search and analysis of documents. None of them put any emphasis on the understanding and analysis of temporal expressions in document contents as we have done in this paper. Some of the challenges that we addressed, have been described in detail by Tahmesebi et al. from the *digital humanities* perspective [16]. Hoffart et al. [6] demonstrated a search system that utilized named entities in the YAGO knowledge graph for retrieval of documents. To this end they use named entities and their categorical types for *auto-complete* suggestion to queries. They, however, rank documents based on publication dates. In their subsequent work, Hoffart et al. [5] perform analytics by using a combination of document publication dates and the entities contained therein. Yeung and Jatowt [10] use LDA topics over time in text to assist historians in answering various queries. In contrast to these systems, we have looked at temporal expressions in document contents in order to generate a deeper analysis for search results. Similarly, Strötgen and Gertz [13] extract content temporal expressions and geographic locations to anchor news articles on a map, and Odjik et al. [12] present an interface to explore different document collections using temporal expressions and text. However unlike both these systems, we have utilized disambiguated named entities in a knowledge graph to contextualize interesting time intervals. Other systems such as WAHSP [8] use *sentiment* in text, and the system HISTOGRAPH uses *social relations* in photographic collections [11].

6 Summary

In this paper, we demonstrated DIGITALHISTORIAN, a system that is able to analyze temporal expressions in document content to generate interesting time intervals which are subsequently used to re-rank and diversify documents to give a historic overview for the issued query. It also offers capabilities to analyze frequent named entities in the YAGO knowledge graph for informative visualizations. DIGITALHISTORIAN thus provides scholars in *digital humanities* an informative and innovative way of exploring semantically annotated document collections.

References

[1] Berberich, K., Bedathur, S., Alonso, O. and Weikum, G. A language modeling approach for temporal information needs. In *32nd European Conference on IR Research (ECIR)*, (Milton Keynes, UK, 2010). Springer, 13–25.

[2] Gupta, D. and Berberich, K. Identifying time intervals of interest to queries. In *23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, (Shanghai, China, 2014), ACM, 1835–1838.

[3] Gupta, D. and Berberich, K. Temporal query classification at different granularities. In *22nd edition of the International Symposium on String Processing and Information Retrieval (SPIRE)*, (London, UK, 2015), Springer, 156-164.

[4] Gupta, D. and Berberich, K. Diversifying search results using time. In *38th European Conference on Information Retrieval (ECIR)*, (Padova, Italy, 2016), Springer, 789-795.

[5] Hoffart, J., Milchevski, D. and Weikum, G. AESTHETICS: Analytics with strings, things, and cats. In *23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, (Shanghai, China, 2014), ACM, 2018-2020.

[6] Hoffart, J., Milchevski, D. and Weikum, G. STICS: Searching with strings, things, and cats. In *37th ACM International Conference on Research and Development in Information Retrieval (SIGIR)* , (Gold Coast, Australia, 2014), ACM, 1247–1248.

[7] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S. and Weikum, G. Robust disambiguation of named entities in text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Edinburgh, UK, 2011), ACL, 782–792.

[8] Huijnen, P., Laan, F., de Rijke, M. and Pieters, T. A digital humanities approach to the history of science - eugenics revisited in hidden debates by means of semantic text mining. In *Social Informatics - SocInfo 2013 International Workshops, QMC and HISTOINFORMATICS*, (Kyoto, Japan, 2013), 71–85.

[9] Kanhabua, N., Blanco, R. and Nørvåg, K. Temporal information retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208, 2015.

[10] Yeung, C.M.A. and Jatowt, A. Studying how the past is remembered: towards computational history through large scale text mining. In *20th ACM International Conference on Information and Knowledge Management (CIKM)*, (Glasgow, UK, 2011), ACM, 1231–1240.

[11] Novak, J., Micheel, I., Melenhorst, M.S., Wieneke, L., Düring, M., Moron, J.G., Pasini, C., Tagliasacchi, M. and Fraternali, P. Histograph - A visualization tool for collaborative analysis of networks from historical social multimedia collections. In *18th International Conference on Information Visualisation, (IV)*, (Paris, France, 2014), IEEE Computer Society, 241–250.

[12] Odijk, D., Gârbacea, C., Schoegje, T., Hollink, L., de Boer, V., Ribbens, K., and van Ossenbruggen, J. Supporting exploration of historical perspectives across collections. In *19th International Conference on Theory and Practice of Digital Libraries, (TPDL)* , (Poznań, Poland, 2015), Springer, 238–251.

[13] Strötgen, J. and Gertz, M. Event-centric search and exploration in document collections. In *12th ACM/IEEE-CS Joint Conference on Digital Libraries, (JCDL)*, (Washington, DC, USA, 2012), ACM, 223–232.

[14] Strötgen, J. and Gertz, M. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[15] Suchanek, F.M., Kasneci, G. and Weikum, G. Yago: A large ontology from wikipedia and wordnet. *Web Semantics*, 6(3):203–217, 2008.

[16] Tahmasebi, N., Borin, L., Capannini, G., Dubhashi, D., Exner, P., Forsberg, M., Gossen, G., Johansson, F.D., Johansson, R., Kågebäck, M., Mogren, O., Nugues, P. and Risse, T. Visions and open challenges for a knowledge-based culturomics. *International Journal on Digital Libraries*, 15(2):169–187, 2015.