Timeline as Information Retrieval and Ranking Unit in News Search

Adam Jatowt

University of Innsbruck, Innrain 15, Innsbruck 6020, Austria

Abstract

News articles are one of the most often read online documents, and the amount of news stories generated daily is quite large. News search engines are then often used to retrieve relevant news articles. The understanding of the returned results can be however impaired when the returned articles are about events being parts of complex or long stories. In this paper we discuss the difficulties resulting from the missing context and information complexity that users searching in temporal news collections may face. To support users in their search and learning we propose the concept of timeline as a retrieval and ranking unit of news search. We believe that automatically generating and ranking timelines could become an effective mechanism to facilitate search and browsing in large news collections.

Keywords

news collections, news search, timeline summarization, news archives

1. Introduction

News are one of the most commonly read types of online documents nowadays. They matter much to all users who want to understand key events in the world or in their localities. In recent years, large amounts of news articles have been published, many of which are about complex and long events or stories. Thus specialized news search engines are available to users helping them to find relevant information.

IR technologies have been adapted to many diverse domains including web search, legal document search, code retrieval, social media search, etc. [1]. We think that novel and effective developments in the field of accessing temporal news collections such as news archives are also needed due to large amounts of news generated these days and their complex characteristics. News articles are a highly temporal type of documents in which time signals (whether in the form of timestamps or embedded temporal expressions) are of key importance, and news tend to be often understood chronologically or in causal order. While text indexing or query suggestion methods have been already studied in the context of temporal search such as search in web archives or in news archives [2, 3, 4], still there is need for research on effective retrieval and ranking approaches. Currently, the usual retrieval approach seems to be basically one based on applying the same access methods as for traditional synchronic text collections such as collections of web pages or Wikipedia articles. In this context, we believe that the unique temporal characteristics of news archives necessitate novel methods for effective information retrieval. Professional users typically know what they wish to find when searching in or interacting with news collections and they also have sufficient skills and knowledge of the collections (or their covered time periods) to effectively retrieve information. On the other hand, general users may lack necessary knowledge and expertise to be able to perform successful searches in large news collections.

Searchers in large news collections face to some extent similar issues as users who search in collections containing documents from unknown and complex domains (e.g., medicine or law). The lack of knowledge causes difficulties in understanding and interpreting search results, updating further queries and continuing search sessions. This is especially a problem when the searched events occurred in more distant past. Imagine a user who wishes to learn about the progress of Iraq War - the event that took place over 10 years ago. Even if the user managed to find all relevant news articles, she or he will have problem to fully understand them, arrange them into meaningful groups as well as continue the searching process by refining subsequent queries. First, this is because the user most likely does not know the context of the times when the war happened and the relations between its individual sub-events. Second, the news articles returned by a typical search engine would be probably ranked just by their relevance despite that the intuitive chronological ordering is useful and intuitive for the user. The lack of chronological or causal order would prevent users from understanding the interrelations between events and the overall progress of the underlying story. Next, the user would also not be able to find important articles as the retrieval method probably does not utilize any notion of

DESIRES 2021 – 2nd International Conference on Design of Experimental Search & Information REtrieval Systems, September 15–18, 2021, Padua, Italy

adam.jatowt@uibk.ac.at (A. Jatowt)
https://ds-informatik.uibk.ac.at/ (A. Jatowt)
0000-0001-7235-0665 (A. Jatowt)

event importance in an explicit way. Hence, less important news articles could be ranked highly with some of them perhaps being returned at the top of the ranked list, effectively preventing the user from understanding the key events of the war. Note that manually arranging articles into meaningful sequences would likely not be realistically feasible. This is because of the presumed lack of knowledge that an average user has about events in arbitrary time periods, and due to potentially large numbers of articles that can be returned, especially for longer events or stories.

Incorporating the notion of event importance into the ranking mechanism could let users receive news articles ordered not just by their relevance but also by the quantified importance levels of the described events. However, still this would not fully help users obtain large picture of all the major events related to their query, as the returned events would not be ordered chronologically. Furthermore, the set of ranked results could contain articles describing events that are parts of different stories or that are related to diverse aspects of the same story. This would likely happen for ambiguous queries such as location names or names of popular entities that took parts in diverse events (e.g., a name of a country's president), or queries about complex and long-evolving events having diverse aspects (e.g., war in Syria).

2. Timeline for Providing Missing Context and Structuring Search Results

We argue that presenting search results in the form of ten blue links ordered by the traditional notion of keyword relevance would not suffice in many cases of news search. Especially, in cases when (a) the sought news are parts of longer ongoing stories, (b) the searched events happened in more distant past, (d) the issued queries are ambiguous, or (c) when the user lacks necessary knowledge of event's context, more effective solutions are required. We think that arranging the returned news results in the form of timelines that summarize the key events and reflect their chronological and/or causal order would be a more user-friendly way towards effective search in longitudinal news article collections. In the above-mentioned example of the search intent of Iraq War the user could receive automatically generated timelines, instead of the usual ranked list of news articles. Such timelines would be then the first step towards combating the lack of context on the user side which prevents successful search experience. After checking the timelines the user should be better informed of the event landscape (e.g. the progress of the war represented by the sequence of its major events) in order to perform further search. This would support conceptualizing or refining queries and let the user better understand what she or he would like to actually retrieve or further explore.

3. Ranked Timelines as Output

TimeLine Summarization (TLS) research helps alleviate the problem of redundancy and complexity inherent in news article collections, thereby supporting users to better understand the news landscape. Usually a timeline is in the form of short descriptions of major events which are presented in chronological order, and it may also reflect causal relations. In traditional setting, the TLS approaches [5] work on a homogeneous type of datasets (i.e., collections of documents about the same event or about events of the same story) and produce just a single timeline as output. For heterogeneous document collections such as the set of relevant search results obtained in news search, one would need to apply the generalization of TLS.

Recently, Multi-TimeLine Summarization (MTLS) [6] has extended the typical TLS settings by allowing heterogeneous document collections as an input and by generating the output in the form of multiple timelines. Outputting multiple timelines as a retrieval result in news search would be suitable to effectively organize and present search results, especially when the query is ambiguous, relates to complex or diverse stories, or the related news have multiple different aspects. The timelines would summarize different relevant stories or could reflect different aspects of the same story (e.g., timeline of the military actions, timeline of the economical aspects of the Iraq war, timeline of the societal responses towards the war, etc.). These timelines could be ranked based on their relevance degrees to the user query (e.g., computed as the sum of relevance scores of their constituent events) or on their importance (e.g., computed as the sum of importance scores of the constituent events, or simply bound to the number of events, i.e., the timeline length). In short, rather than ranking individual news articles, the search engine would construct on-the-fly and rank timelines to be presented as ranked results to users. These timelines would provide missing context to individual events and help structure the entire news landscape that is relevant to user query. Finally, the timelines used as retrieval units could be then further expanded based on user clicks to let the users view the detailed articles. This style of result presentation bears resemblance to search result clustering by their temporal aspects and ranking in Web search as proposed by Alonso *et al.* [7].

Regarding the actual algorithm that could be used for MTLS task, Yi *et al.* [6] have proposed a two stage Affinity Propagation approach to automatically generate the set of timelines from a heterogeneous news collection.

An advantage of that approach is that the number of the generated timelines does not need to be known or set beforehand, as it is dynamically obtained based on the underlying document collection. This flexibility is important in search scenarios as users can input arbitrary queries that could necessitate varying numbers of constructed timelines.

4. Conclusions and Open Research Directions

In this position paper we propose considering timeline as an atomic retrieval[1], ranking and presentation unit for news search. We think that incorporating Multiple Timeline Summarization methods as well as adding the timeline ranking component would provide useful alternative to users besides the usual retrieval method that centers on fine-grained retrieval units such as individual documents. This novel approach should be especially useful in cases of searching in longitudinal news collections. Users could, for example, benefit from the interwoven combination of providing timelines as search results and of outputting the ranked news articles in a traditional way. For example, upon understanding the news landscape based on the constructed timelines, the users could switch to the usual document relevance-based search in order to locate particular articles and satisfy detailed search needs. Timelines would then not only provide the "global picture" useful for conducting search, but could also let users narrow down the results (e.g., by selecting a sub-part of some timeline for subsequent search).

There are many open problems that need to be approached for using timelines as retrieval and ranking units including: (a) effective and efficient generation of timelines for arbitrary user queries, (b) timeline labelling[2] and presentation[3] for easy and quick result understanding, as well as (c) the actual ranking mechanism for arranging the generated timelines.

Finally, worth investigation are user interaction models and methods with search results composed of timelines. This encompasses timeline zooming in and out, expanding timeline events to read their detailed articles, or switching between timeline search results and usual search results. One can also imagine the possibility of users updating/constructing their own timelines (e.g., by removing irrelevant events/articles or adding relevant ones, perhaps the ones discovered through subsequent search episodes).

[1] Actually, to be correct, in the context of our proposal, the term "retrieval" should be substituted by "construction" or "generation"

[2] For fast overview of the ranked timelines, each timeline could have its generated label or short description

[3] Some of the constructed timelines may also contain shared events, the explicit indication of which could be perhaps added for more effective result presentation

Acknowledgments

We thank anonymous reviewers for their useful comments and suggestions.

References

[1] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, volume 520, Addison-Wesley Reading, 2010.

[2] K. Berberich, S. Bedathur, T. Neumann, G. Weikum, A time machine for text search, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 519–526.

[3] N. K. Tran, A. Ceroni, N. Kanhabua, C. Niederée, Back to the past: Supporting interpretations of forgotten stories by time-aware re-contextualization, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015, pp. 339–348.

[4] Y. Zhang, A. Jatowt, S. Bhowmick, K. Tanaka, Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 645–655.

[5] D. G. Ghalandari, G. Ifrim, Examining the state-of-the-art in news timeline summarization, arXiv preprint arXiv:2005.10107 (2020).

[6] Y. Yu, A. Jatowt, A. Doucet, K. Sugiyama, M. Yoshikawa, Multi-timeline summarization (mtls): Improving timeline summarization by generating multiple summaries, in: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021.

[7] O. Alonso, M. Gertz, R. Baeza-Yates, Clustering and exploring search results using timeline constructions, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 97–106.