History-related Knowledge Extraction from Temporal Text Collections[*]

Yijun Duan

Abstract

There are so many amazing and diverse lessons that we can learn from history. Firstly, studying history makes us understand that people are fundamentally similar to each other, regardless of where, when and how they live. Many differences usually arise because of cultures adapted to different environments. Secondly, contemporary people rarely associate the past with advanced technology, however, looking at technology in the past can still teach us how modern technology connect to and came from the ancient world. Thirdly, looking back on all of the problems that ancient civilizations faced can put modern issues into perspective. Many problems would seem small when we compare them to the ones that humanity has already overcome. Lastly, studying history can break preconceived ideas. Something we thought was a fact could be completely incorrect, and learning history makes us become skillful at admitting these errors and correcting them.

Thanks to accumulated large numbers of digitized documents from the past, it is possible now to employ large scale analyses for uncovering history-related knowledge. In the sea of documents, quite many documents contain chronologically ordered content describing histories of entities or detailed accounts of past events. However, we think average users are facing the following fundamental challenges during learning history.

- **Challenge 1**: Firstly, with the rapid growth of Web, more and more history-related documents are available causing severe information overload. Thus history learning would require tremendous reading and the work is really energy consuming, not to mention manually judging the significance and usefulness of encountered information is difficult.

[*] Doctoral Thesis, Department of Social Informatics, Graduate School of Informatics, Kyoto University, KU-I-DT6960-29-0801, March 2020.

• **Challenge 2**: Secondly, to understand the past information is sometimes hard, especially for the young generation. The general context of articles which originate from different time periods may be fairly different. Intuitively, information in different contexts cannot be understood properly without a solid understanding of the connection (analogies) between their contexts.

• **Challenge 3**: Thirdly, there usually exist underlying patterns embodied in historical documents (e.g., salient and diverse threads in Japanese cities' histories, life stages of American politicians). To manually grasp such high-level and informative information would also require huge cognitive effort.

• **Challenge 4**: Lastly, learning from examples and learning from comparison are both effective strategy extensively adopted in cognition and education. Nevertheless, users may not be able to make good use of such strategies effectively on their own.

Given the research background, to overcome the above challenges, our research interests are listed as follows:

• **Topic 1**: To obtain a good understanding of the history of a category consisting of multiple entities. In other words, to answer questions like *What is the history of Japanese cities? Which events frequently occurred during the life of French scientists in the 19th century?* (see Topic 1 in Chap. 2)

• **Topic 2**: To detect latent entity categories whose members share similar histories effectively, when taking a set of entity-related documents as an input. Namely, to answer questions like *How to group Japanese cities based on the similarities of their historical developments?* (see Topic 2 in Chap. 3)

• **Topic 3**: To provide a condensed and informative document reorganization consisting of major contrasting events chronologically ordered for faster and better understanding of the compared sets of timeline documents. In other words, to answer questions like *how different were the lives of French scientists in the* 19^{th} *century from those of American scientists at the same century?* and *What makes the histories of Japanese cities distinct from the histories of Chinese cities?* (see Topic 3 in Chap. 4)

- **Topic 4**: To automatically generate a summary of corresponding news from two temporally-distant collections of articles based on a user query. For instance, to answer questions like *What are corresponding and important news related to the query "politician" from periods 1980s and 2010s?* (see Topic 4 in Chap. 5)

- **Topic 5**: To automatically discover comparable typical entity pairs from two across-time collections of entities. For example, to answer questions like *What is the relationship between electronic gadgets of 1990s and that of 2010s?* (see Topic 5 in Chap. 6)

- **Topic 6**: To periodize the evolving word semantics embodied in diachronic corpora. That is, to answer questions like *How to split the history of usage of the word "Amazon"? Which other words influence such splitting process?* (see Topic 6 in Chap. 7)

Topic 1 is propose to help tackle the above challenge 1. Topic 2 is proposed to tackle the challenge 3 and challenge 4. To alleviate challenge 1 and challenge 4, topic 3 is proposed. Furthermore, topic 4 is proposed to resolve the challenge 1 and challenge 2. Similarly, topic 5 is proposed to overcome the challenge 2 and challenge 4. Finally, we propose topic 6 to help overcome the challenge 2 and challenge 3.

Keywords: Entity history summarization, History-driven entity categorization, Comparative timeline summarization, Across-time comparative news summarization, Typicality-based across-time entity set mapping, Archive periodization