Linea: Tailoring Timelines by Visual Exploration of Temporal Text

Tiago Etiene* Paulo Pagliosa† Luis Gustavo Nonato‡
Modelo Inc. UFMS University of São Paulo

Abstract

Timelines are often used to summarize complex, time-evolving, sequences of events. In this poster, we present the preliminary results of *Linea*, a tool designed to help users build timelines that summarize content from Wikipedia. We combine Natural Language Processing and the PageRank algorithm to automatically build default timelines of any Wikipedia article, giving users a starting point for tailoring them. Then, users can interact with the *event matrix* – a collection of histograms of events – by exploring each matrix cell in search for *peaks*, *i.e.*, proxies of important periods of time. By selecting a peak, the tool shows all events in that period so users are able to add or remove events, thus customizing timelines.

Index Terms: Timeline, Text Summary, Event Matrix, Histogram

1 Introduction

Building meaningful summaries from document collections is a routine task for many professionals. For instance, when preparing a news article about a crime, a political act, or a historical figure, journalists must gather as much information as possible, select events of interest, and organize the flow of events to produce a report. In the era of information overload, semi-automated techniques for text analysis become indispensable to explore the document haystack. Nevertheless, although computational tools for assisting (i) the visualization, and (ii) the summarization of time-stamped documents are available, they are not properly integrated, and the burden of compiling, organizing, and explaining events of interest falls on the shoulders of experts interested in creating timelines.

(i) visualization systems – such as TextFlow and EventRiver – help users understand the evolution of complex events, and illustrate the relationships among them. Nevertheless, because they rely on time-stamped documents, they are tailored to news stream or similar data; a rich source of historical information can hardly be processed, namely, unstructured temporal text, such as text available on Wikipedia. Moreover, these systems *are not* designed to build readable textual summaries from these documents. On the other hand, (ii) automated summarization techniques are capable of generating textual summaries of documents; however, no user-friendly mechanism is available to modify and explore summaries. Once an automated tool builds a summary, it is hard to change it.

We combine both methodologies, that is, a visualization technique that assists the user during the task of text summarization via timelines. Our approach, called *Linea*, combines text analysis techniques – used by the Information Retrieval and Natural Language Processing communities – with the *event matrix*, to build default timelines that can be updated as the user explore the event matrix.

2 Linea

Linea's pipeline can be divided into *event extraction*, *event ranking*, and *event matrix* construction.

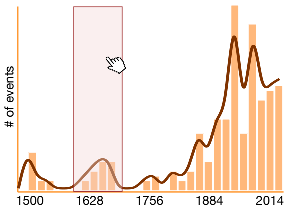
Event extraction When the user loads our tool, a search bar is shown. The user can type the name(s) of Wikipedia article(s) to be summarized (*e.g.*, "United States, History of the United States"). Events are then extracted using a technique similar to the one described in Alonso [1]. The term "event" is used to characterize a happening at a particular time t. We use the dates written in the article body as proxy for important events. The rationale is the following: given that the Wikipedia editor decided that it is worth being specific about when "something" happened, "something" is likely to be relevant. Hence, the algorithm for event detection is straightforward: step (I) detects each sentences in the article; step (II) mark as "event" all sentences that contain a date. Example from an excerpt of the article entitled "Albert Einstein":

> Albert Einstein was born in Ulm, in the Kingdom of Württemberg in the German Empire on 14 March 1879. His father was Hermann Einstein, a salesman and engineer. His mother was Pauline Einstein (née Koch).

Linea detects three sentences, out of which it extracts one event, namely, Albert Einstein's birth date. Currently, we only extract explicit date information that can be normalized to (year), (year-month), or (year-month-day) format. Thus, temporal information, such as "in the 20th century [...] " is ignored.

Ranking Events Having detected all events, we use LexPageRank [2] to rank them. LexPageRank is based on the well-known PageRank algorithm. The PageRank algorithm ranks graph nodes by their "popularity" (see Erkan and Radev [2] for details). The main idea behind LexPageRank is to apply the PageRank algorithm to a graph made of sentences. Each node is a sentence, and two sentences are connected by an edge if their content is similar according to the cosine metric. The result of PageRank is a number that defines the importance of that sentence to the set of sentences. We use that information to show the n most important events of any user-selected topic, by selecting the first n best ranked nodes.

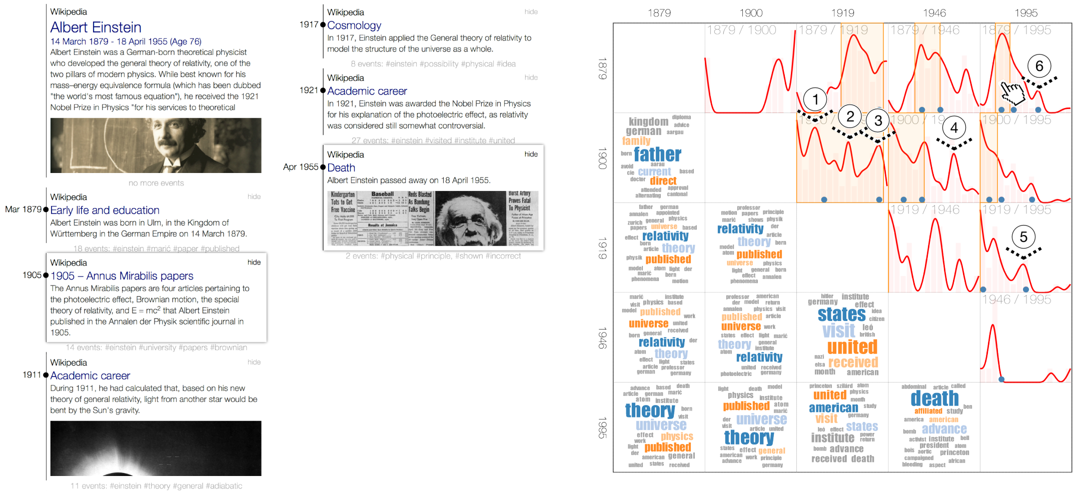
Event matrix The last step is a visualization that allows users to explore all events and modify the timeline. This is necessary because there could be hundreds of events available, making harder to identify any interesting features in the data. A simple solution is to use a histogram. Histogram peaks show time periods that are important and valleys represent the opposite. The inline figure shows a histogram containing 4 locally prominent peaks, which means that a lot has been written about these periods. The user can select one of the peaks and only the events from that period are revealed. Then, the user can update the default timeline.



However, a histogram may hide many interesting peaks. For instance, the tallest bar in the inline figure may refer to collection of two or more important historical facts, not just one, as the single peak suggests. Moreover, since the user will not know how many interesting peaks the data contains, we would like to suggest time intervals to be explored. We solve this problem by building a collection of histograms, each spanning a distinct, but overlapping, time interval. The intervals are computed by clustering the events using

*e-mail: tiago.etiene@gmail.com
†e-mail: pagliosa@facom.ufms.br
‡e-mail: gnonato@icmc.usp.br

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE



Wikipedia

Albert Einstein

14 March 1879 - 18 April 1955 (Age 76)

Albert Einstein was a German-born theoretical physicist who developed the general theory of relativity, one of the two pillars of modern physics. While best known for his mass–energy equivalence formula (which has been dubbed "the world's most famous equation"), he received the 1921 Nobel Prize in Physics "for his services to theoretical

no more events

Wikipedia hide

Mar 1879 Early life and education

Albert Einstein was born in Ulm, in the Kingdom of Württemberg in the German Empire on 14 March 1879.

18 events: #einstein #marić #paper #published

Wikipedia hide

1905 1905 – Annus Mirabilis papers

The Annus Mirabilis papers are four articles pertaining to the photoelectric effect, Brownian motion, the special theory of relativity, and $E = mc^2$ that Albert Einstein published in the Annalen der Physik scientific journal in 1905.

14 events: #einstein #university #papers #brownian

Wikipedia hide

1911 Academic career

During 1911, he had calculated that, based on his new theory of general relativity, light from another star would be bent by the Sun's gravity.

11 events: #einstein #theory #general #adiabatic

Wikipedia hide

1917 Cosmology

In 1917, Einstein applied the General theory of relativity to model the structure of the universe as a whole.

8 events: #einstein #possibility #physical #idea

Wikipedia hide

1921 Academic career

In 1921, Einstein was awarded the Nobel Prize in Physics for his explanation of the photoelectric effect, as relativity was considered still somewhat controversial.

27 events: #einstein #visited #institute #united

Wikipedia hide

Apr 1955 Death

Albert Einstein passed away on 18 April 1955.

2 events: #physical #principle, #shown #incorrect

Figure 1: Custom timeline of Albert Einstein's life. On the left we show the timeline, and on the right a 5 × 5 event matrix of that article. Blue dots in the event matrix shows the selected timeline events. Each word cloud is associated with the time period in the interval (column j, row i): *e.g.*, the bottom left word cloud refers to the the period (1879, 1995). The following keywords were omitted from the word cloud: *albert* and *einstein*.

k-means. Because we are dealing with 1D data, we used a optimal k-means clustering algorithm. The size of the matrix is $k \times k$, where k is the number of clusters (see Figure 1). After event clustering, an histogram representing each cluster is built. Note that the main goal of the matrix is to show to the user potentially interesting time periods at once, so that the user can be aware of the data features, without having to go through the whole data in the search for peaks. To help the user navigate the event matrix, each matrix cell contains the time period that the histogram refers to. Note a single histogram can be used by setting $k = 1$. Also note that the classic pan+zoom approach can be used for navigation.

3 Result

We show an example with the article entitled "Albert Einstein". The article contains over 330 sentences, of which *Linea* extracted 85 events. Figure 1 shows how *Linea* assists the summarization of documents. When an article is loaded for the first time, the default timeline is built based on the ten highest ranking events. Nine events ranked as the most relevant correspond to Einstein's scientific career: special theory of relativity and Brownian motion (1905); general theory of relativity (1907, 1916); and Einstein's prediction that light would be bent by the Sun's gravity (1911). Figure 1 shows Einstein's timeline after we edited it (in total six events were removed – two of them were redundant – and two were added).

The 5 × 5 event matrix in Figure 1 shows the distribution of events over time. The highest peak in the top right histogram (1879/1995) ranges from 1900 to 1919 (see selection in the top-right matrix cell), which suggests an important period in Einstein's life. By selecting that interval, *Linea* displays only event snippets in that range, which in this case, is mostly related to Einstein's scientific career. Moreover, since the selection is reflected in other matrix cells, the event matrix reveals three smaller peaks, labeled as (1), (2), and (3) in the matrix cell (1900/1919). With further exploration, the user can observe that the three peaks correspond to (1) family matters (marriage and children) and special theory of relativity, (2) academic career, and (3) family matters (divorce and new marriage) and the general theory of relativity.

The selection around the highest peak is not the only range containing interesting features. A small inflection in the curve outside the selection of the (1879/1995) cell, turns out to be another peak inside the (1900/1946) cell, labeled as (4). By selecting that area, the timeline reveals that most of the events correspond to Einstein's imigration to the US. One of the events reads:

> He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences.

The word cloud summarizes it nicely as the words "United States" and "Visit" are prominent. Note that the histogram essentially flattens out after 1960, just after Einstein's death (1946/1995), except by small peaks. Lastly, by exploring the period around his death (labeled as (6)), a curious event is revealed:

> After the death of Israel's first president, Chaim Weizmann, in November 1952, Prime Minister David Ben-Gurion offered Einstein the position of President of Israel, a mostly ceremonial post.

Once an user has explored the collection of events, she/he can build a tailored timeline.

4 Conclusion

Linea is a tool that help users summarize Wikipedia articles in a timeline. In this preliminary work, we decided to use simple techniques for both detect and rank events. More robust techniques designed to extract events and assign importance to them can be used to improve the overall quality of the tool. Furthermore, we also have plans to evaluate the overall quality of the extracted events. Although we have obtained some user feedback, we need to conduct a thorough user study to evaluate the usability of *Linea*. Lastly, the tool can be enriched by extracting actions, such as marriage, birth, death, etc.

Acknowledgements

The authors wish to thank Karina Pupio Raimundo for her valuable insights.

References

[1] O. Alonso and D. University of California. *Temporal Information Retrieval*. University of California, Davis, 2008.

[2] G. Erkan and D. R. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, volume 4, pages 365–371, 2004.