HistoryComparator: Interactive Across-Time Comparison in Document Archives

Adam Jatowt
Kyoto University
adam@dl.kuis.kyoto-u.ac.jp

Marc Bron
University of Amsterdam
marc.bron@gmail.com

Abstract

Recent years have witnessed significant increase in the number of large scale digital collections of archival documents such as news articles, books, etc. Typically, users access these collections through searching or browsing. In this paper we investigate another way of accessing temporal collections - *across-time comparison*, i.e., comparing query-relevant information at different periods in the past. We propose an interactive framework called *HistoryComparator* for contrastively analyzing concepts in archival document collections at different time periods.

1 Introduction

The role of history and cultural memory in shaping today's society cannot be overestimated. We often refer to the past for variety of reasons including supporting decision making processes (Gilovich, 1981). Up to recent years analyzing large samples of historical documents was difficult due to the nature of the materials studied, e.g., paper records stored in physically distributed archives. Now, with the availability of large digitized collections, academia and industry are looking into the utility of using computational methods on large samples of digitized records to study human history and culture (Odijk *et al.*, 2012). Understanding and making use of such collections often necessitates across-time comparison to elucidate commonalities and differences between entities existing at different times. Comparison is in fact a common tool used by historians and social scientists for insightful analysis. *Comparative history* (Halperin, 1982), in particular, often relies on across-time comparative analysis presuming that nothing can be correctly understood without proper comparison and grounding (even a timeline is a comparison tool, albeit, a very simple one).

We propose in this paper an interactive framework called *HistoryComparator* for across-time comparison of query result sets within archival document collections. Rather than comparing individual documents, our framework contrasts temporal slices of corpora containing documents related to a query (e.g., entity or event). The proposed system is build on the top of a retrieval engine and offers two basic comparison modes, *contrastive term cloud view* and *contrastive graph view*. The former generates comparative text summaries in the form of term clouds, while the latter aligns networks composed of the top query-related terms. Provided functionalities include, among others, *keyword in temporal context*, *time-based term adjustment* and *sentiment-level correlation* of collection snapshots. In addition, a range of *synchronization facilities* are provided for facilitating effective comparison.

Relatively little research focused on interactively comparing collections of archival text documents. Odijk *et al.* (2012) demonstrated interactive environment to visualize information on volume and correlation of words and documents over time. *Texcavator* (van Eijnatten *et al.* 2014) is a framework integrating analytical tools such as concept clustering, sentiment mining, and named entity recognition to produce world clouds, timelines and other visualizations. The closest work to ours is perhaps an interactive tool for exploratory search in document collections proposed by Bron *et al.* (2012). Like our system, theirs uses the concept of double columns and term clouds for finding interesting information. However, it does not provide the same comparative facilities like our system, neither it offers graph-based contrastive interface.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations,
pages 84–88, Osaka, Japan, December 11-17 2016.

2 HistoryComparator

We employ dual-column visualization (Bron *et al.*, 2012) (see Fig. 1). The side-by-side composition allows immediate comparison alleviating cognitive burden of spotting commonalities and differences. A user enters two queries (they can have the same syntactic form for representing the same concept or they can rather represent two different things) and sets the corresponding time periods in both the left and the right column. The queries are next issued against an underlying document collection subject to the input time constraints. Based on the returned documents the results are displayed in each column.

2.1 Contrastive Term Cloud View

In the first mode the system displays the key terms related to the input query as term clouds of both the collection subsets (see Fig. 1). Term clouds are convenient technique for summarizing large text collections where sizes of terms denote their importance (Bateman *et al.*, 2008). Summarizing a collection of a few thousands documents by selecting whole sentences is not effective since query terms occur in a multitude of possible contexts, and, hence, presenting all such contexts as sentences would lead to prohibitively large summaries. On the other hand, term clouds have been found useful for quick and effortless overview of large portions of textual data (Bateman *et al.*, 2008). In our system, terms in both the columns are color-coded to emphasize differences and similarities. While the font size of each term is bound to its relative frequency in a given time period, the color is associated with the relative difference of frequencies. That is, terms prevailing more in one time period and less in the other are either more red or more blue depending on the column they are shown in (blue for left and red for right). Black color indicates terms with similar rate of occurrence in both the compared periods. Font size and color selection can be set based on either linear or logarithmic scales. Furthermore, the following options are provided:

- Adjustable number of terms to be shown in each column by manipulating sliders
- Choice of term ordering: alphabetically or by frequency
- Grouping separately colored and black terms.

Fig. 1 shows the results of example query: `world trade center` (WTC) at the time immediately after the buildings' collapse (left column) and at about 10 years later (right column). We next list and discuss the key components of the contrastive term cloud view.

Keyword in temporal context. Upon clicking on any term displayed in either column, HistoryComparator displays term's contextual information in a popup window as portrayed in Fig. 2. If the same term is also listed in the results of the other column, the second popup window will automatically be shown in that column enabling term's contexts' comparison across the two columns. The context of the term is reflected by three representative text snippets. These are selected from all the snippets that contain both the clicked term and the query words. The selection is done by averaging the minimum distances expressed as the number of words between each of the query words and the selected term with additional penalty in case of missing query terms. The selected term and the query words have backgrounds colored for their easy spotting in the displayed sentences. Additional information about the term includes its frequency and sentiment score. The pop-up windows shown in Fig. 2 have been generated from the results of Fig. 1 after a term `construction` has been clicked. Sentences in the left column refer to the information about construction materials used in the former WTC building or to the National Construction Safety Team investigating the location of the buildings, while the same word in the right column refers to constructing a new building in the area of the former WTC.

Similar terms. Besides information about the context of a selected term, the system also displays terms with similar contexts to it in both the columns (Fig. 2). Term similarities are computed as the overlap of the sets of the top co-occurring words found by applying Jaccard Coefficient. The top 5 similar terms to the target term are displayed in each column.

Time-based term adjustment. Sometimes changes are due to outside-driven effects rather than due to the change of the compared query (e.g., entity) across time. A new term may appear simply because of the time passage and, hence, not due to the change specific (related) to the query. For example, a term computer is considered as novel in the present time in the results of a query unrelated to computers (e.g., tokyo) when compared to some past period for this query, merely, due to the recent significant increase in the use of computers. For capturing the effect of time, we utilize background document collections which are built on the random sample of documents (unrelated to the queries) collected from each time period set by the user. Three options are provided in the system in regards to the time effect:

- *No adjustment*: no adjustment done (default option).
- *Term normalization*: the frequency of each term within the foreground sub-collection is normalized by dividing it by its corresponding frequency in the background sub-collection.
- *Visual indication*: in this option an additional visual signal is added to each term (see Fig. 3) to explicitly inform about term's dependency on the above-discussed time effect. In particular, a rectangle frame surrounding each term is added. The width of the frame is bound to the term's frequency in the corresponding background collection while its color depends on the relative difference of the term's background frequencies across both the compared time periods. These inform users to what extent the term frequency in each column and its column-wise difference are affected by time. Frame sizes and colors can be based on either linear or logarithmic scales.

Note that when selecting the visual indication mode, in total, four signals are visible about each term in either time period (column): term's normalized frequency in the foreground collection (i.e., font size), the difference of term's foreground frequency in the target time period and foreground frequency in the other time period (i.e., font color), term's normalized background frequency (i.e., frame width) and the difference of the term's background frequencies in both the compared time periods (i.e., frame color).

Popularity trend. The popularity of queries in each time period is also shown above the term clouds in the form of two time series corresponding to both the time periods (see Fig. 1). Furthermore, a user can also click on any term in the term clouds to display its popularity across time.

Sentiment analysis. Temporal changes in sentiment associated with query at different times can constitute complementary information for more exhaustive analysis. To study fluctuations in emotional factors across time we utilize SentiWordnet[1]. Sentiment orientations in relation to the query are calculated by summing sentiment scores of terms in the returned results for each time period (each column). The bottom bar displays in each column the rate of positive vs. negative orientations. A user can thus observe the change in sentiment value across time. In the example shown in Fig. 1 we can notice that the recent context in which the world trade center is mentioned is slightly more positive than the context in which it was mentioned during and right after the building's collapse.

Lastly, hovering mouse over the positive (negative) parts of the sentiment bar highlights positive (negative) terms in the corresponding column to explain reasons behind a particular sentiment rate.

2.2 Contrastive Graph View

Term clouds cannot capture changes in relations between terms over time. To compare the inter-word relations, we provide the second view, *contrastive graph view*, as portrayed in Fig. 4. In this view the top frequent words are positioned as nodes in two force-directed graphs in the columns. To inform about the term importance, the node size is dependent on the term frequency. Terms that frequently co-occur with each other are connected by the edge whose width is determined by the value of Jaccard coefficient computed on their co-occurrence and occurrence rates. The graphs are then composed of the top important nodes and the top high-scored edges linking them. Same as in the contrastive term cloud view, the color of a node in a given column depends on the relative frequency difference of a term underlying the node across both the columns. In addition to the node coloring, the color of edges conveys information on whether the connected nodes have similar or different affinity across the compared periods. To enable effective comparison, the positions of nodes in both the graphs are aligned. In other words, the terms which are same in both the graphs are placed in the same relative positions.

[1] http://sentiwordnet.isti.cnr.it/

Selecting any node triggers automatic selection of an identical node/edge in the other column. The nodes can also have their positions rearranged. As the graphs are synchronized, any displacement in one column results in an equal displacement in the other column.

Dynamic adjustment of node and link counts. Too many nodes or links may clutter the view. The proposed system provides then an easy way for adjusting the number of nodes or edges by manipulating sliders in each column. Less important nodes or edges can be then increasingly added to either graph by incrementing sliders. When the sliders are synchronized (synchronization option), the change in one column triggers the same change in the other column. This allows for synchronized comparison of node and edge importance. Such progressive edge increment permits also observing gradual additions of edges starting from the most important to less important ones.

Similar nodes detection. Like in the contrastive term cloud view, a user can select a node by double clicking on it in order to see its most similar nodes in the other column.

3 Architecture

The system is implemented using Perl 5.10 and works in the client-server mode. The Web interface depends on Mojolicious Web Application Framework[2]. We use jQuery as foundation for JavaScript design together with D3[3] visualizations. The time plots are generated using jQuery plugin called jqPlot[4]. The user-specified time periods are used for constructing time-constrained queries. These time periods can be set to be divided into L non-overlapping equal-size time units (by default, $L=1$). L queries would be then sent to the underlying search engine together with associated time constrains. The latter are defined by the starting and ending points of each of L units. By issuing L ($L \gg 1$) queries over smaller, consecutive time units, instead of a single query over the entire chosen time period, the system effectively "forces" the search engine to retrieve documents more or less uniformly over time rather than from only one of few time points. As underlying data sources, currently, our system uses the New York Times Article Archive on Solr, Google News Archive and Google Blog search engines. The content of each collected document is subject to stop word removal, tokenization and normalization.

4 Conclusions

To support effective search and discovery in text archives, we have introduced in this paper a novel framework for the comparative analysis of historical document collections both on the term (contrastive term cloud view) and term association (contrastive graph view) levels. In future, we plan to suggest relevant and interesting time periods for contrasting entities by comparing term distributions over time.

Acknowledgments

This work has been partially supported by Grant-in-Aid for Scientific Research (No. 15K12158) from MEXT of Japan and by the JST research promotion program Presto/Sakigake.

References

Bateman, S., Gutwin, C., Nacenta, M.: Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. Proceedings of the HT 2008, pp. 193–202, 2008.

Bron, M, van Gorp, J., Nack, F., de Rijke, M., Vishneuski, A., and de Leeuw, S. A Subjunctive Exploratory Search Interface to Support Media Studies Rearchers. Proceedings of the SIGIR 2012, pp. 425-434, 2012.

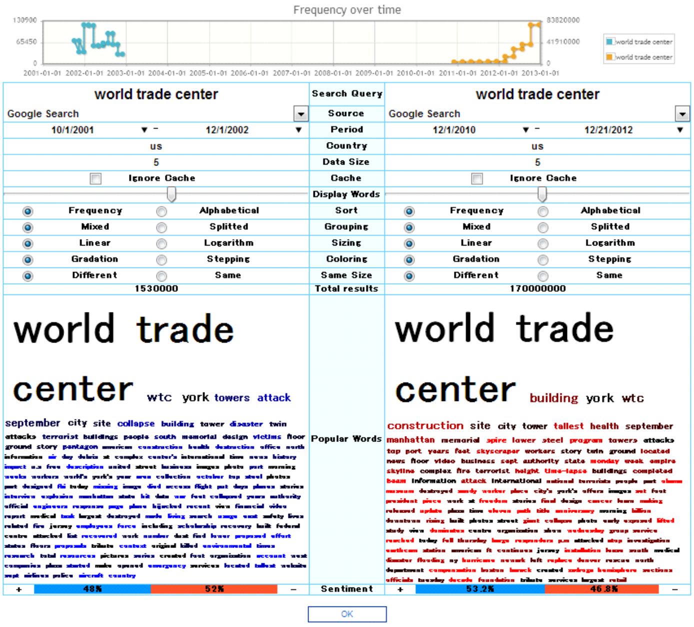
van Eijnatten, J., Verheul, J., Pieters, T. TS Tools: Using Texcavator to Map Public Discourse. TS: Tijdschrift voor Tijdschriftstudies, issue 35, pp. 59-65, 2014.

Gilovich, T. 1981. Seeing the past in the present: The Effect of Associations to Familiar Events on Judgments and Decisions. Journal of Personality and Social Psychology, 40(5):797.

[2] http://mojolicio.us/
[3] https://d3js.org/
[4] http://www.jqplot.com/

Halperin C.J. et al. 1982. Comparative History in Theory and Practice: A discussion. The American Hist. Rev., 87(1):123–143.

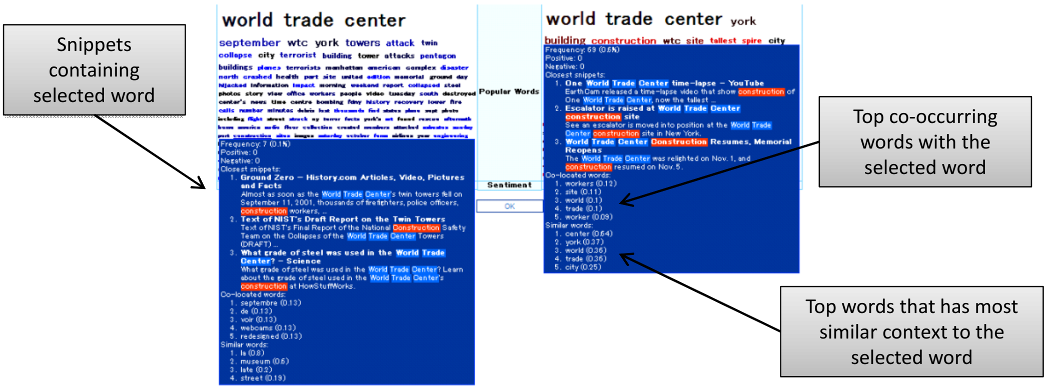
Odijk, D., Santucci, G., de Rijke, M., Angelini, M., and Granato, G. Exploring Word Meaning through Time. Proceedings of the TAIA 2012 Workshop in conjunction with SIGIR 2012. Portland, USA.



world trade center	Search Query	world trade center
Google Search	Source	Google Search
10/1/2001 – 12/1/2002	Period	12/1/2010 – 12/21/2012
us	Country	us
5	Data Size	5
□ Ignore Cache	Cache	□ Ignore Cache
	Display Words	
● Frequency ○ Alphabetical	Sort	● Frequency ○ Alphabetical
● Mixed ○ Splitted	Grouping	● Mixed ○ Splitted
● Linear ○ Logarithm	Sizing	● Linear ○ Logarithm
● Gradation ○ Stepping	Coloring	● Gradation ○ Stepping
● Different ○ Same	Same Size	● Different ○ Same
1530000	Total results	170000000
world trade center wtc york towers attack september city site collapse building tower disaster twin …	Popular Words	world trade center building york wtc construction site city tower tallest health september …
+ 48% 52% –	Sentiment	+ 53.2% 46.8% –

OK

Figure 1: System interface and output for query `world trade center.`


Snippets containing selected word
Top co-occurring words with the selected word
Top words that has most similar context to the selected word
world trade center
september wtc york towers attack twin
collapse city terrorist building tower attacks pentagon
Popular Words
Sentiment
OK
world trade center york
building construction wtc site tallest spire city

Frequency: 7 (0.1%)
Positive: 0
Negative: 0
Closest snippets:
1. Ground Zero – History.com Articles, Video, Pictures and Facts
Almost as soon as the World Trade Center's twin towers fell on September 11, 2001, thousands of firefighters, police officers, construction workers, ...
2. Text of NIST's Draft Report on the Twin Towers
Text of NIST's Final Report of the National Construction Safety Team on the Collapses of the World Trade Center Towers (DRAFT) ...
3. What grade of steel was used in the World Trade Center? – Science
What grade of steel was used in the World Trade Center? Learn about the grade of steel used in the World Trade Center's construction at HowStuffWorks.
Co-located words:
1. septembre (0.13)
2. de (0.13)
3. voir (0.13)
4. webcams (0.13)
5. redesigned (0.13)
Similar words:
1. la (0.8)
2. museum (0.6)
3. late (0.2)
4. street (0.19)

Frequency: 69 (0.6%)
Positive: 0
Negative: 0
Closest snippets:
1. One World Trade Center time-lapse – YouTube
EarthCam released a time-lapse video that show construction of One World Trade Center, now the tallest ...
2. Escalator is raised at World Trade Center construction site
See an escalator is moved into position at the World Trade Center construction site in New York.
3. World Trade Center Construction Resumes, Memorial Reopens
The World Trade Center was relighted on Nov. 1, and construction resumed on Nov. 5.
Co-located words:
1. workers (0.12)
2. site (0.11)
3. world (0.1)
4. trade (0.1)
5. worker (0.09)
Similar words:
1. center (0.64)
2. york (0.37)
3. world (0.36)
4. trade (0.36)
5. city (0.25)

Figure 2: Popup windows due to highlighting term `construction` in the left column of Fig. 1.

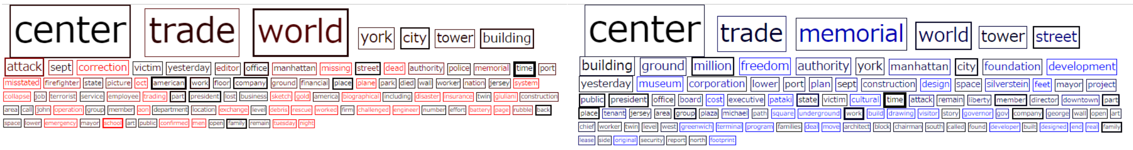

center trade world york city tower building
attack sept correction victim yesterday editor office manhattan missing street dead authority police memorial time port
center trade memorial world tower street
building ground million freedom authority york manhattan city foundation development
yesterday museum corporation construction design space silverstein feet mayor project

Figure 3: Snapshot of term clouds with visual adjustment of time passage effect.

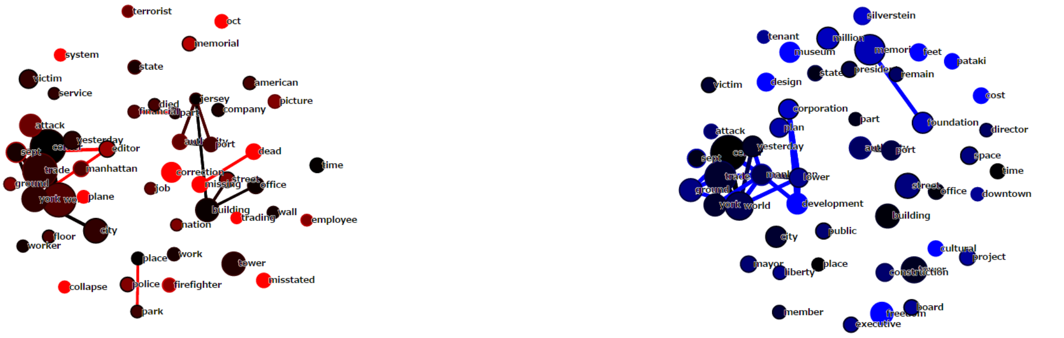


Figure 4: Snapshot of the contrastive graphs in the contrastive graph view.