Mapping the First World War Using Interactive Streamgraphs

Alaa Abi-Haidar\textsuperscript{1,2}\textsuperscript{*}, Bin Yang\textsuperscript{1}, Jean-Gabriel Ganascia\textsuperscript{1,2}

\textsuperscript{1}Laboratory of Computer Sciences, Pierre and Marie Curie University, France
\textsuperscript{2}Labex OBVIL(Observatory Literary Life). Sorbonne University, France

Abstract In this paper, we use unsupervised named entity recognition and streamgraphs in order to visualize massive amounts of unstructured textual stream data, namely, French newspapers (e.g. Le Figaro, La presse, L’humanité) from the first world war period. Such a visualization allows us to identify main characters, events and locations involved in or relevant to the first world war, according to the French press. Furthermore, our visualization technique can help visually identify correlations between major people (e.g. presidents, generals, public figures...), locations (e.g. countries, cities, towns...) and organizations and events (e.g. corporations, battles...) on multiple aligned streamgraphs. Our method can be applied to unstructured data streams of any domain or time period.

Keywords Named Entity Recognition, Information Extraction, Stream Graphs, Information Visualization, Stream Data, Stacked Graphs

1. Introduction

Huge amounts of printed manuscripts from old French news journals (from the 19th and 20th century) have been recently digitized and published by the National French Library, la Bibliotheque Nationale Francaise (BnF). However, the massive amounts of produced textual data are highly unstructured and hard to index, search, or visualize, needless to mention the digitization errors resulting from ill-preserved or damaged manuscripts and imperfect Optical Character Recognition (OCR) techniques.

Named Entity Recognition (NER) is a task of information extraction that aims to identify in-text references to concepts such as people, locations and organizations, mainly in unstructured natural-language text. NER is very useful for text indexing, text summarization, question answering and several other tasks that enhance the experience between humans and literature. Furthermore, advanced NER and disambiguation techniques are capable of dealing with noise resulting from digitization errors.

The unsupervised use of dictionary-lookup is known to enhance NER, however dictionaries have limitations for being finite and ambiguous. On the other hand, supervised NER such as Stanford's NER Classifier that we tested here is known to perform very well but only with the availability of huge amounts of manually annotated training data that is very costly, time consuming and sometimes inaccurate due to inter-annotator inconsistencies.

Therefore, we use an original unsupervised approach for Named Entity Recognition and Disambiguation (UNERD) \cite{8} with a French knowledge-base and a statistical contextual disambiguation technique. We have conducted several studies to assess the performance of our UNERD method in \cite{8-10} on English and French corpora, and compared it to...
state of art unsupervised and supervised NER classifiers, namely DBPedia Spotlight, BaLIE and Stanford NER classifier. We have shown that our method outperforms all unsupervised methods [9] and occasionally Stanford’s supervised method when little training data is available [8]. Details of the UNERD algorithm or performance go beyond the scope of this study.

Finally, we use streamgraphs [1] (or stacked graphs) to visualize the evolving trendline of key figures (people), locations and organizations related to the first world war that are extracted automatically using UNERD from the French newspaper “La Presse” between 1914 and 1919, thus totalling 1820 issues.

In the following sections, we focus on information visualization and the steamgraph method that we use to visualize the trends of main figures, battle locations and organizations involved in the first world war.

2. Information Visualization Overview

The term Information Visualization or InfoVis is referenced in a variety of contexts of meaning; in Computer science, Card, Shneiderman, and Mackinlay define this term in a more narrow sense and referred to it as “the use of computer-supported, interactive, visual representations of abstract non-physically based data to amplify cognition” [2]. Visualization is also an important technique for the analysis of knowledge derived from text mining. Thus many InfoVis techniques have been introduced to visualize documents and text streams in the domain of text mining [13]. However, what method is used depends on the question to answer and the data at hand [11]. According to Risch [5], text visualization is composed of three steps. First, the text is processed in a representation more suitable for sequent operations. Second, in order to render a certain view, a mapping onto a 2D or 3D space is performed. Third, user interaction is enabled. In fig. 1, we show examples of two traditional visualization techniques, namely, Data Mountain (document management visualization) and Topic Island (topic visualization). A Survey on text stream visualization techniques can be also found in [7].

In 2008, Havre et al. introduced Themeriver [3]. In their visualization, a “river” flows from left to right through time with its width varying according to thematic strength of temporally associated documents. Colored “currents” (or streams) flowing within the river narrow or widen to indicate decreased or increased strength of individual topics or a groups of topics in the associated documents. The river is shown within the context of a timeline and a corresponding textual presentation of external events.

Figure 1. Left: Data Mountain with 100 web pages.[6] Right: Topic Island for the Buddhist article showing major themes [4].
In 2008, Byron introduced Streamgraphs [1] to emphasize legibility of individual layers that are arranged distinctly in organic forms. Streamgraphs were applied to last.fm music data as part of an academic project called “Listening Histories”. The Streamgraph design attracted huge interest from both information visualization enthusiasts and music lovers. Streamgraph made it to the New York Times, with a visualization of movies according to box office receipts (see Fig. 2) and an online interactive visualization tool.

Assume the time series as a set of \( n \) real-valued non-negative functions: \( f_1, f_2, \ldots, f_n \) such that the baseline function is given by: 
\[
g_i = g_0 + \sum_{j=1}^{i} f_j
\]

The silhouette is as close as possible to the x-axis and is defined by: 
\[
silhouette(g_0) = g_0^2 + g_n^2
\]
The deviation measure at each value of \( x \), is defined by: 
\[
deviation(g_0) = \sum_{i=0}^{n} g_i^2
\]
The sum of squares of the slopes at each value of \( x \), is defined by: 
\[
wiggle(g_0) = \sum_{i=0}^{n} g_i'^2
\]
Minimizing the deviation and the sum of squares yields the streamgraph’s equation that is detailed in by Lee Byron [1].

In order to easily spot major figures, locations and organizations involved in the first world war we require a visualization tool that is capable of displaying all this time-sensitive information over a period of time, highlighting key entities when necessary. Streamgraph visualization meets our demands, firstly, because it is an
interactive visualization method that could reveal additional information (such as labels and frequencies) when the mouse is rolled over a “stream” which avoids cluttered information. Secondly, Streamgraph can display the importance of an entity based on the relative area of the stream which in our case represents the frequency of each entity in a period of time.

According to Silic’s three important factors of evaluating visualization techniques [12] our NEs Visualization technique using streamgraphs meets Generalizability, Precision and Realism. Our method can be applied to a more general corpus with other types of data sets, bringing the users precise and realistic visual representation of the data set.

Figure 4. Three streamgraphs identifying key figures (people) [blue], locations [pink] and organizations [green] of the first world war extracted automatically as ENAMEX named entities from the French newspaper “La Presse” between 1914 and 1919. The labels for each stream appear when the mouse is rolled over the stream.

3. Visualization Results

The three Streamgraphs in the fig. 4 were automatically generated from extracted ENAMEX named entities labeled as Person (blue), Location (pink) and Organization (green). This streamgraph visualization offers a much more intuitive overview of what happened, when and where during the first world war. Events can be easily correlated and their labels displayed by simply hovering over a stream with the mouse over a certain time period. A publicly accessible interactive demo is accessible at http://alahay.org/labs/ACASA/. For example, at the beginning of 1916, we notice an inflation of the trendline corresponding to “Poincaré”¹ (blue) which happens to inflate along with the inflation of “Verdun” (pink) which is the location of a famous battle and “rente” (green) which is a form of war tax. This observation does not necessary imply any causality, however, it helps identify major co-occurrences of figures, events and/or locations which facilitates the crunching and understanding of huge amounts of textual data.

¹ President of France 1913-1920
4. Conclusions

We have successfully used streamgraphs and unsupervised named entity recognition in order to compress and represent 5 years of daily press related to the first world war in an online interactive visualization tool. Our method can be extended to any domain or field of study over any period of time. In biology, for example, event mining and named entity recognition are very common for identifying protein and gene names, and interactions in between.

Acknowledgements

This work has been done within the LABEX OBVIL project, and received financial state aid managed by the Agence Nationale de la Recherche, as part of the programme "Investissements d'avenir" under the reference ANR-11-IDEX-0004-02. We thank the National French Library, BnF, for the provided data.

REFERENCES


