Clustering and Mapping Violence News Events

Syed Toufeeq Ahmed, Sukru Tikves, and Hasan Davulcu
Department of Computer Science and Engineering,
Arizona State University,
{Toufeeq,Sukru,Hasan.Davulcu}@asu.edu

***Abstract*—It is difficult to follow the entirity of the global of news sources , and the events happening everyday. If an analyst in her area has to follow and map all these according to the timeline they happen, the task quickly becomes overwhelming. We present an online tool which attempts to ease the analyst's task of finding all news articles about an event, and sorting and mapping them on a timeline. We implemented an incremental clustering algorithm working on real-time incoming news, experimenting with different feature sets, including named entities and sentence overlap methods. We evaluated these approaches using Document Understand Conference (DUC) datasets**

I. Introduction and Related Work

In order to aid the analyst described in the abstract, we developed an online tool that detects new events and clusters articles related to it as an "event thread" and maps it on an easy to use timeline visualizer.

To understand and extract meaningful structure from document continously arriving arriving news streams, [1] presented a "burst of activity" model. Seminal work [2], [3] in event detection and tracking explored clustering algorithms like agglomerative clustering, augmented Group Average Clustering. Well-known idf-weighted cosine coefficient metric method [4] was also used to detect and track topics. A real-time news event extraction system [5] extracts violence and disaster events by extracting grammars on documents in the cluster.

II. Online Tool for Visualizing and Mapping News Clusters and Named Entities

Online system works in three steps: 1) Data Scraping from online sources, pre-processing and Named Entity Recognition (NER); 2) Detecting new *violent events* and then clustering of related items to a thread. 3) Sorting and mapping these related events on a timeline and location visualizer.

For recognizing the violent events (like bombing or kidnapping) in the text, we built an ontology by recursively extracting synonym sets using an initial seed (around 240) from WordNet[1].

Next step is to incremental clustering of news articles using different feature sets (Term vector, Named Entity vector, Mixed (term + named entities) and Sentence Overlap). Our preliminary experiments with sentence overlap method have shown that looking for similar or exact sentences in news articles produces almost perfect precision. However, in order to improve the overall recall, we also utilized a secondary metric based on cosine similarity between document term vectors.

[1]WordNet: wordnet.princeton.edu/

Last step is sorting and mapping the articles on an time-line and location map. The online system utilizes MIT's SIMILE library for timeline visualization. The cluster of related articles, and related entities are shown as filterable components, and a map of mentioned locations are displayed on an interactive map using the Bing maps API.

III. Experiments and Evaluations

The system was evaluated using Document Understanding Conference (DUC)[2] datasets for years 2004, 2005, and 2006. For each year of data, 100 articles spanning 10 different news events have been randomly selected. The comparison has been done using F1-Measure, which summarizes both precision and recall of the algorithm. Table I summarizes the performance results of different approaches. Entity based vectors significantly under performed due to loss of information. We've also experimented combining entities and terms in a single representation as Mixed vector.

TABLE I
F1-Measure performance metrics on datasets DUC 2004, 2005 and 2006.

Method	DUC'04	DUC'05	DUC'06	Avg
Term Vector	0.96	0.71	0.67	0.78
Entity Vector	0.90	0.46	0.34	0.57
Mixed Vector	0.95	0.59	0.52	0.69
Sentence Overlaps	**0.95**	**0.75**	**0.73**	**0.81**

The online website can be accessed at: http://code.azcips.info/DemoSite/. Currently, it hosts 474 event threads and corresponding named entities in the news articles filtered from January 2007 subset of New York Times annotated corpus[3].

References

[1] J. Kleinberg, "Bursty and hierarchical structure in streams," *Data Mining and Knowledge Discovery*, vol. 7, no. 4, pp. 373–397, 2003.
[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in *Proc. of the DARPA broadcast news transcription and understanding workshop*, vol. 1998.
[3] Y. Yang, T. Pierce, and J. Carbonell, "A study of retrospective and on-line event detection," in *Proc. of 21st ACM SIGIR*. ACM, NY, USA, 1998, pp. 28–36.
[4] J. Schultz and M. Liberman, "Topic detection and tracking using idf-weighted cosine coefficient," in *Broadcast News Workshop'99 Proceedings*. Morgan Kaufmann, 1999, p. 189.
[5] H. Tanev, J. Piskorski, and M. Atkinson, "Real-Time News Event Extraction for Global Crisis Monitoring," in *Proc. of the 13th (NLDB 2008), London, UK*, 2008, pp. 24–27.

[2]Document Understanding Conference: http://duc.nist.gov/
[3]NYT Corpus: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T19