Event HyperGraph for analyzing large collections of text

Sam Yu-Te Lee*
University of California, Davis

Kwan-Liu Ma†
University of California, Davis

ABSTRACT

Sensemaking on large collections of text is a challenging task that analysts have to perform. Previous works approach this problem either from a topic- or entity-based perspective, but they lack interpretability and trust. In this paper, we propose a visual analytics system that allow analysts to explore and reorganize a corpus to suit their needs and quickly make sense of the 4Ws from the organized corpus. The system first organizes a large corpus into a hypergraph by combining topic- and entity-based extraction techniques. Then the hypergraph is hierarchically clustered and visualized for analysts to explore and reorganize interactively. Finally, an event graph visualization method based on storyline enables analysts to quickly make sense of the 4Ws. The whole pipeline is designed to foster interpretability and trust by providing semantic context of the visualization and by supporting curating interactions. Case studies and a task-based evaluation are used to demonstrate the effectiveness and trustworthiness of the system.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Treemaps; Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

2 RELATED WORKS

3 DESIGN RATIONALE

What DRs are needed for analysts to explore and reorganize a corpus for their analytical tasks and why?

3.1 Design Considerations

- **DC1: An overview of the topic structure**
- **DC2: Support for curation through user interaction**
- **DC3: Detailed analysis of the curated result**

4 METHODOLOGY

4.1 Preprocessing

The Methodology can work for any unstructured dataset

4.1.1 Summarization

Chatgpt for summarization

4.1.2 Document Embedding

OpenAI's embedding API

4.1.3 Major Participant Extraction

Chatgpt for major participant extraction and another model for entity linking

4.2 Modeling

4.2.1 Hierarchical HyperGraph Clustering

We organize the data into a hypergraph: nodes, hyperedges.

The hypergraph is clustered based on semantic and connectivity similarity: dual, Ravasz algorithm

Chatgpt to assign topics to each cluster

5 VISUALIZATION

5.1 Space Filling Curves

Introduce Gosper curve and generalized Hilbert curve, and how they are used for large graph layout

5.2 SFC for HyperGraph

Using the Gosper curve to layout the article graph

Concatenating four generalized Hilbert curve to layout the entity graph on the peripheral

5.3 Spacing Strategy

5.4 Border Approximation

5.5 Edge Bundling

ACKNOWLEDGMENTS

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ.

REFERENCES

*e-mail: ytlee@ucdavis.edu
†e-mail: klma@ucdavis.edu