HyperMap: Analyzing large collections of documents with hypergraphs

Category: Research

Paper Type: application/design study

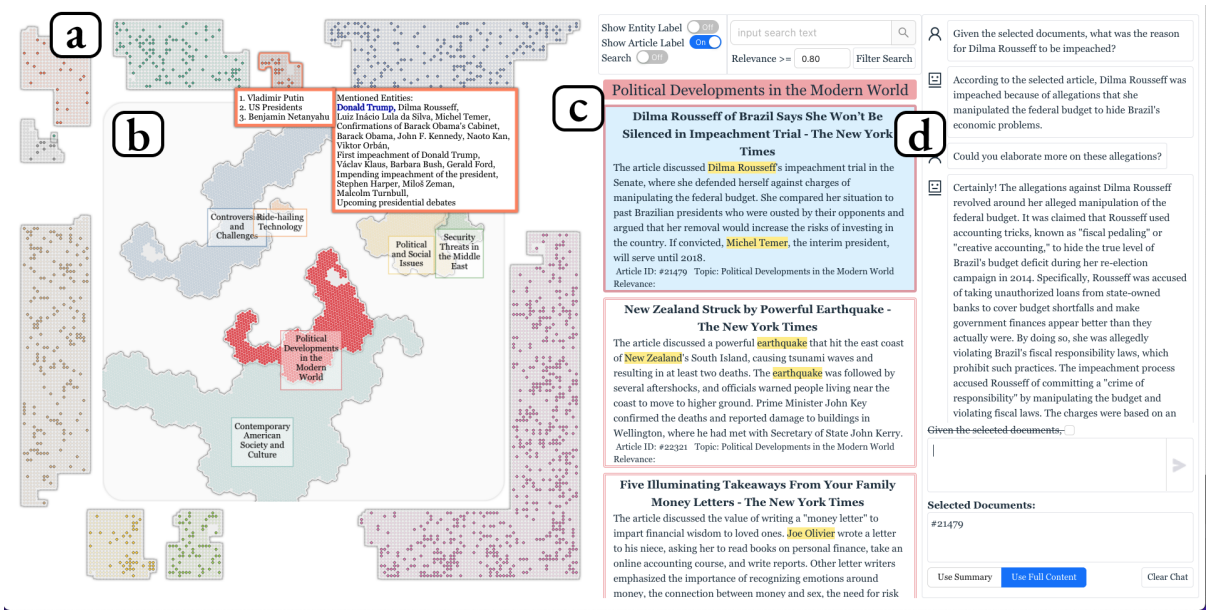

Fig. 1: The HyperMap system. (a) The peripheral area of Cluster View shows the mentioned characters of highlighted documents using Gilbert curves. (b) The center area of Cluster View shows the topic structure of the corpus using Gosper curves. (c) The Document View shows a list of selected documents. (d) The Chatbot View provides a chatbot interface to answer user questions with the option to insert selected documents in the prompt.

Abstract—Sensemaking on large collections of documents (corpus) is a challenging task that analysts often have to perform. Previous works approach this problem either from a topic- or entity-based perspective, but they lack interpretability and trust due to poor model alignment. In this paper, we propose HyperMap, a visual analytics approach that combines topic- and entity-based techniques seamlessly. By leveraging the capability of Large Language Models (LLMs), we model the corpus as a hypergraph that matches the user's mental model when analyzing a corpus. The hypergraph is then hierarchically clustered with an agglomerative clustering algorithm by combining semantic and connectivity similarity. The system is designed to emphasize Model Alignment to foster interpretability and trust. To demonstrate the generalizability and effectiveness of the HyperMap system, we present two case studies on two different datasets: a news article dataset and a visualization publication dataset. We discuss limitations and future work of combining visualization and LLMs to enhance analysts' ability to analyze a corpus.

Index Terms—Text visualization, Sensemaking, Hypergraph, Corpus analysis Large language models

1 Introduction

Text data is ubiquitous. From news articles and social media posts to scientific publications, the tremendous amount of text data that is produced poses not only opportunities but also a great challenge to anyone who needs to analyze them. Visual analytics (VA) mitigates this challenge by combining mathematical models and visualizations to automate the sensemaking process and reduce the cognitive load. Chuang et al. [**?**] proposed that *Model alignment*, the alignment of analysis tasks, visual encodings and model decisions, greatly affects users' interpretation and trust in visual analytic systems. However, in text analysis, the available models often align poorly with analysis tasks. For example, topic models are commonly used to model the topical structure of text documents. Most topic models characterize *topic* as a probabilistic distribution spanning a given vocabulary [**?**]. This transformation from a *topic*, a high-level concept that the user seeks to understand, to a *probabilistic distribution*, a low-level concept that mathematical models can operate on, prevents proper model alignment. The misalignment between analysis tasks and models limits the usage of visual analytics systems for users who are not familiar with the underlying models. Previous works that support users to analyze large collections of documents adopt separate models for analyzing topics and entities, but they are often intertwined in the user's mental model. Model alignment requires the model to operate on documents and entities simultaneously, a capability that existing text analysis models lack.

Recent advances in large language models (LLM) present a promising solution to this problem. LLMs have proven successful in various natural language processing (NLP) tasks, especially in question-answering tasks due to their strong capability to understand user intent. Researchers in visualization have adopted LLMs to assist data transformation [**?**] or directly generate visualization [**?**]. However, they all assumed a clean data format, where the data to be visualized is already in a table format. For unstructured text analysis though, this is rarely the case. Topics [**?**], sentiments [**?**], concepts and entities [**?**, **?**] are

common analysis targets in text analysis that require a data preparation stage to extract them from unstructured text. The capability of LLMs to extract information from documents according to user intent eliminates the need to carefully align the analysis tasks and models in VA systems, because a specific model is no longer needed to prepare the data for the analysis task. In the previous example, instead of relying on abstruse and unfathomable probabilistic models, LLMs can directly process the text data and summarize the topics of the documents. A user can ask a LLM: *'What are the topics of these articles?'*, and the LLM would give a human-like response, such as *'The articles are about Covid-19'*.

In this work, we designed a VA system that models a corpus as a hypergraph, where the nodes are documents and characters (salient entities) extracted from the document using LLMs. We showcase how LLMs are used flexibly to align the data, analysis task and visualization during our design process. The hypergraph is then hierarchically clustered and visualized with enhanced space-filling curve layouts [**?**]. The system supports interactive exploration, reorganization and analysis of the documents. To the best of our knowledge, no published visual analytics system has adopted LLMs to align the data, analysis task and visualization. Using the system, we demonstrate how proper model alignment can be achieved using LLMs.

The contributions of our work are as follows:

- We introduce an LLM-based information extraction pipeline that is capable of extracting topics and salient entities from a given corpus in a way that fosters interpretation.
- We extend space-filling curve layouts to visualize clusters in large hypergraphs.
- We develop a novel VA system that allows users to effectively explore, reorganize and analyze a corpus.

2 Related Works

2.1 Data preparation for large collections of text

Topic-based approaches Topic-based approaches employ certain variations of topic models to organize the documents in a meaningful way. Each topic is often presented as a 'bag of words', which can be in the form of a sequence of words [**?**,**?**,**?**,**?**,**?**,**?**,**?**] or word clouds [**?**,**?**]. The modeling result provides an overview of the dataset for subsequent analysis tasks. Despite their popularity, the use of topic models as an overview, as well as its 'bag-of-word' visualization, is reported by Lee et al. [**?**] to be problematic, especially for non-expert users, in a comprehensive user study. Chuang et al. [**?**] concluded that these problems arise from a misalignment between the analysis task, visual encoding and model. The sensemaking process becomes challenging without a basic understanding of the model because the 'bag-of-words' representation is too far away from the user's mental model of a topic. This misalignment limits the usage of topic models for non-expert users and makes the system prone to produce false positives.

Entity-based approaches A line of work that makes successful model alignments is the entity-based approach. 'Entities' usually include named entities (people, organizations, locations), or meaningful concepts known to an existing knowledge base. The earliest of such approaches is Jigsaw [**?**], where entities are linked if they appear in the same document. FacetAtlas [**?**] generalizes the idea of entity to 'facets' which can be entities or any keywords or user's interest. ConceptVector [**?**] uses 'concept' to represent a similar idea. Generally, entity-based approaches exhibit better model alignments than topic-based approaches [**?**], but the polysemy of natural language makes them prone to produce false positives [**?**]. In our work, we use *Characters* to represent entities or concepts that are discussed in the documents. Characters distinguish themselves from previous works in that they are not only mentioned in a document, but they must also be significantly discussed.

Embedding-based approaches Finally, an important line of work organizes documents by directly modeling their semantic similarity [**?**]. Documents are first projected into a high-dimensional vector space where similarity can be measured, and then a dimensionality reduction technique (e.g.t-SNE) is used to project the dataset onto a two-dimensional space for visualization. Earlier works construct a sparse vector using term-frequency based scores such as *TF-IDF* or BM25 [**?**,**?**]. More recently, the success of pre-trained language models like BERT [**?**] popularizes the idea of embedding documents in a dense vector space [**?**, **?**, **?**]. The embedding can then be used for document retrieval [**?**, **?**] or visualization. Embedding-based approaches also exhibit healthy model alignment, as the vector space directly models the analysis task (finding similar documents). However, the result often lacks explainability and prevents users from trusting the result. Recently, Raval et al. [**?**] proposed to use LLMs to provide explainability to embeddings-mappings visualizations. We adopt a similar approach in our system to provide interpretability to the clustering result.

2.2 LLMs for Information Extraction

Information Extraction aims to identify structured information of interest from unstructured text data. Some of its subtasks include Named Entity Recognition (NER), Relation Extraction (RE) and Event Extraction (EE) [**?**, **?**]. Although LLMs have proven successful in many NLP tasks, their application to IE is non-trivial. First, the *faithfulness* of LLMs needs to be carefully evaluated. Faithfulness refers to the ability of a model to adhere to the provided information and not use parametric knowledge learned during training to answer user questions [**?**]. When conducting information extraction, it is necessary to ensure that the extracted information is actually from the provided text and not from the model's parametric knowledge. Second, LLMs are known to produce *hallucination*, where LLMs provide answers factually contradicting to input text (intrinsic) or even factually false (extrinsic). In the context of IE, we mainly focus on the intrinsic hallucination problem. A recent evaluation conducted by Bang et al. [**?**] found that ChatGPT rarely exhibits intrinsic hallucinations, including the abstractive summarization task from which neural models usually suffer.

More specifically, Li et al. [**?**] comprehensively evaluated the capabilities of ChatGPT for common IE tasks. They found that ChatGPT excels under the Open-IE setting, where the model relies solely on user input to extract information from documents, but performs poorly under the Standard-IE setting, where ChatGPT is instructed to choose a correct label. Their findings agree with Zhang et al. [**?**] where ChatGPT is reported to perform poorly on extractive summarization. A common reason for the poor performance of ChatGPT in these tasks is that they are essentially supervised learning tasks, and ChatGPT is not trained to perform them. To make the best use of ChatGPT (or more generally, LLMs) for IE tasks, we need to carefully design the extraction tasks as question-answering tasks instead of supervised learning tasks.

3 Limitations and Future Work

The case studies demonstrate that HyperMap provides highly interpretable visualization of the topic structure and character connections and flexible interaction to reorganize and analyze a corpus. During the usage, the clusters and borders help users clearly distinguish different clusters. The topic labels assigned by LLMs coupled with character connections provide rich semantic information for users to make sense of the topics. Most importantly, the user is able to focus on their domain-specific analysis tasks and does not need to understand the underlying model. Analysts from all disciplines are not expected to understand LLMs, hypergraphs or clustering algorithms, but they can still explore, reorganize and analyze the corpus. We attribute this outcome to a good model alignment. Hypergraphs allow us to directly map user interactions on both documents and characters, two seemingly unrelated units of analysis, to operations on one unified model. LLMs assist in generating clean and interpretable preprocessed data for the modeling process. Combined, the visualization can be designed in a way that all preprocessing and model details are irrelevant to the user during analysis.

However, there are still several limitations in the current approach. To use LLMs for information extraction, the prompts need to be redesigned for every dataset. Although the general template structure is the same, optimizing prompts is a time-consuming trial-and-error process. To make it worse, the evaluation of the accuracy of a prompt relies on human judgment. This means that the preprocessing result is

not guaranteed to be error-free. Addressing the evaluation challenge is a promising direction for future work. We envision a combination of LLMs, computational evaluation metrics and visualizations to assist prompt evaluation. Another limitation is the amount of user interactions needed to converge on a satisfying corpus organization. it would be better if the system could automatically generate a satisfying organization based on user feedback. Our current automatic cluster expansion strategy only deals with expansion, and it does not consider user intent. We believe LLM's ability to understand user intent can be utilized to address this limitation. Overall, with high interpretability and generalizability, it is possible to posit the users as the center of the analysis, opening up many possibilities for future work.

4 Conclusion

In this paper, we propose HyperMap, a visual analytics system that assists users in exploring, reorganizing and analyzing a corpus. Previous works have shown that topic- and entity-based analysis are essential to sensemaking on a corpus, but existing text analysis tools do not provide a unified interface for both types of analysis. We fill in this gap by combining state-of-the-art large language models and hypergraph analysis and visualization techniques. We introduce an LLM-based pipeline to extract topics and characters from unstructured text documents. We model the corpus as hypergraphs and apply an agglomerative clustering algorithm. The clusters are visualized by building upon existing space-filling curve layouts, which exhibit a high level of visual scalability and aesthetics. Moreover, we emphasize the importance of Model Alignment in the design of visual analytics systems. The generalizability and effectiveness of HyperMap are demonstrated in two case studies, analyzing datasets from two different domains. Our future research aims to utilize the capabilities of LLMs to understand user intent to support more advanced analysis tasks.