InsightLink: Empowering News Analysis with Graph RAG and LLM Integration

Bing-Chen Chiu, Jeng-Yue Liu, Cheng-Shuan Lee, Peng-Jen Chen, Yu-Ting Chen, Ding-Hong Chen
Advised by Dr. Chia-Yen Lee & Dr. Tzu-Fen Chang

Background & Motivation

In the realm of social sciences, content analysis is a widely employed research method for qualitative data, particularly news articles. However, this method often demands substantial time from researchers. They must meticulously read **extensive volumes of texts** and **manually code** the data to derive and summarize research findings. Consequently, we embarked on the development of InsightLink, a system designed to **streamline the extraction of insights** from **large qualitative datasets**.

In InsightLink, we employ **Retrieval-Augmented Generation (RAG)** (Lewis et al., 2021) for its ability to integrate parametric memory with non-parametric memory, enabling more specific, factual, and diverse language generation while reducing hallucinations. However, **traditional RAG struggles with global questions**, such as the inquiry into the *main themes* of the given large text corpus. To overcome this, we adopt **Graph RAG** (Edge et al., 2024), which constructs **a graph-based text index** using a large language model (LLM). This approach enhances the model's ability to generate comprehensive and contextually diverse answers to global queries.

InsightLink is a specialized tool designed to analyze a **vast volume of news articles.** It enables efficient comparisons across various dimensions, such as publication date and media type. By analyzing the results, we can derive valuable insights. These insights not only assist researchers in comprehending how narratives evolve over time and across different outlets but also empower decision-makers to make more informed choices and policies.

Case Study

In our case study, we examine both mainstream and non-mainstream media reports on discrimination against Asian Americans in the United States from March 2020 to December 2021. We specifically focus on the critical incident of the **Atlanta Spa Shootings** that occurred on March 16, 2021. The media outlets include in our study are *The Washington Post, USA Today, New York Times, Star Tribune,* and over nine other news organizations, which collectively published more than **600 news articles**.

Our research seeks to explore the reporting of anti-Asian discrimination in the United States differs based on the timing of coverage, the type of media outlet, and the background of the authors. For instance, we can utilize InsightLink to address the following research question:

RQ: *Examine whether the articles published before March 16, 2021 (the occurrence of Atlanta Spa Shooting) and those published after the date (including March 16, 2021) areare different or similar in terms of types of racism reported.*

A peek of the model pipeline generated by InsightLink when addressing RQ on news articles before March 16, 2021 (dataset 1, *N* = 300) and after March 16, 2021 (dataset 2, *N* = 334) . We utilized GPT-4o-mini as our main model for all stages in InsightLink. In the case study below, we also concatenate a predefined *concept tree* with racism types into our prompt to enable the LLM to identify the types from the provided tree.

"Identify the **types of racism** being reported in the news articles."

Graph RAG Response for Dataset 1

The news articles reflect various forms of racism, categorized into macro-level and individual-level racism. Below are the key points summarizing the types of racism identified:

1. Macro-Level Racism

- **Systematic Racism:** The articles highlight systemic issues that contribute to racial injustice and inequity faced by Asian communities, particularly during the COVID-19 pandemic [Data: Reports (419, 540, 371)].
- **Racial Injustice/Inequity and Oppression:** There are discussions on the broader societal implications of racial injustice, including the historical context of discrimination against Asian Americans .....

Graph RAG Response for Dataset 2

The analysis of recent news articles reveals various forms of racism, categorized into macro-level and individual-level racism. Below are the key points summarizing the types of racism identified:

1. Macro-Level Racism

- **Systematic Racism:** Reports highlight the presence of systematic racism, particularly in the context of anti-Asian violence during the COVID-19 pandemic, reflecting broader societal issues of racial injustice and inequity [Data: Reports (607, 338, 415, 384, 386)].
- **Racial Injustice/Inequity and Oppression:** The articles discuss the impact of racial injustice and the oppression faced by Asian communities, ...

Comparison between Dataset 1 & Dataset 2

Commonalities

- **Macro-Level Racism:** Both datasets emphasize systematic racism, discussing its implications, particularly in relation to anti-Asian violence during the pandemic ...
- **Cultural Complicity:** Both datasets address cultural complicity ...
- **Individual-Level Racism:** The response from both datasets highlights the use of derogatory terms like "China virus" and "Kung flu", ...
- **Racism Towards Asian Women:** Both datasets acknowledge the intersection of racism and misogyny directed at Asian women ...
- **Physical Attacks/Violence:** Both datasets report a rise in anti-Asian hate crimes characterized by physical attacks ...

Differences

- **Link to Advocacy and Legislative Changes:** Dataset 2 emphasizes the call for advocacy and legislative changes in addressing racial injustice, while Dataset 1 does not explicitly mention advocacy and legislative responses.
- **Focus on Xenophobia:** Dataset 2 specifically mentions xenophobia as a prominent concern, ... This notion is less emphasized in Dataset 1, ...
- **Mention of Specific Incidents:** Dataset 2 specifically references incidents such as the Atlanta spa shootings as examples of the intersection of misogyny and racism, ...

Practical Policy Insights for Federal Officials

Develop a comprehensive anti-racism framework that addresses both macro-level and individual-level racism targeted at Asian communities.

Enhance awareness and responsiveness to the intersectionality of racism and misogyny directed at Asian women.

Prioritize the exploration of xenophobia as a unique aspect of anti-Asian racism in policy responses and public discourse.

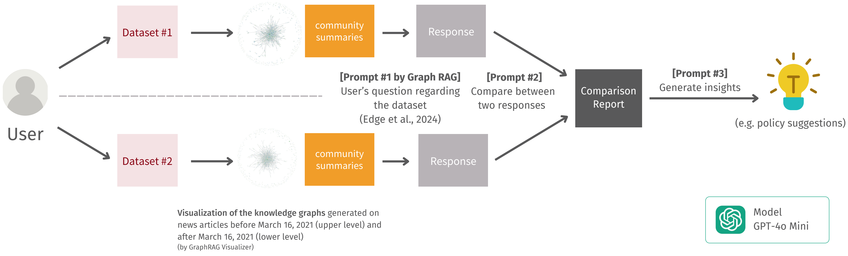
Promote advocacy and supportive legislative changes as essential components of combating anti-Asian racism.

Establish a systematic reporting mechanism for anti-Asian hate incidents to enhance data collection and inform policy.

Conclusion

InsightLink streamlines the manual process of analyzing large amount of qualitative data, providing detailed reports tailored to users' needs for comparing two datasets. **This significantly saves time and human effort.** However, a more extensive evaluation of the model's response is needed, but this is hindered by the substantial amount of time required for human labeling. Despite this limitation, we believe InsightLink has the potential to support a wider range of tasks and domains, enabling qualitative data researchers to obtain insights from vast amounts of data in a timely manner.

System Framework



Visualization of the knowledge graphs generated on news articles before March 16, 2021 (upper level) and after March 16, 2021 (lower level)
(by GraphRAG Visualizer)

InsightLink integrates Graph RAG with LLMs to analyze and compare datasets effectively. Users upload two datasets, which undergo processing to generate community summaries that unveil key insights and structural patterns within the data. Leveraging these summaries and user-provided prompts, the system generates personalized responses for each dataset. Subsequently, it conducts comparison of the responses to identify differences and commonalities. Users can further request the model to extract insights from the previous results, such as policy recommendations for decision-makers.

Reference

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

[2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130 [cs].

國立臺灣大學資訊管理學系
Dept. of Information Management
National Taiwan University