Visualization of Literature - Final Report

Yating Ke, Qi Xue, Xiaoyu Gao

1. Introduction

1.1 Team goals

The goal of our project is to help researchers quickly get informed of the current status of research in their subjects and discover research hotspots through author collaboration networks, keyword word clouds, and other visualization methods. The target users are researchers and students who have the need to analyze literature, reduce the cost of reading literature, and improve the efficiency of learning in the subjects. The type of analytical questions we intend to answer include recent research hotspots, recent researchers who have made outstanding contributions and the trend of research in a specific direction in recent years. Our project aims to answer such questions that could potentially help users gain a better understanding of the particular field in a short time period.

1.2 Users' Tasks and Personas

After the previous research, we summarized the following three possible user profiles:

1. Sara, the Overwhelmed PhD Student

- 25 years old, 2nd year PhD candidate in biology,
- Feels overwhelmed trying to keep up with all the latest publications in her field,
- Spends too much time trying to read every new paper, leaving little time for her own research,
- Frustrated because she can never seem to get a big-picture view of current trends,
- Would love a bird's-eye view of current biology research to help guide her own project.

2. Michael, the Tenured Professor

- 47 years old, associate professor of literature at a small liberal arts college,
- Has a wide range of teaching and research interests in Victorian and Romantic poetry,
- Struggles to keep his lectures and curriculum updated with latest publications,
- Wants an efficient way to stay current on the most impactful new researchers and publications,

- Seeks a visualization of trends to see rising stars and guide collaborative opportunities.

3. Jessica, the Ambitious Postdoc

- 32 years old, postdoctoral researcher in the psychology department at a large research university,
- Specializes in social psychology and behavior change interventions,
- Building an independent research program to launch her professorship career,
- Wants to position her work at the leading edge of the latest trends,
- Would benefit greatly from quickly identifying the current research frontiers in her field.

To help users better visualize literature data, they need to do the following as the workflows shown in Fig 1:

1. Select a literature database or upload data.

2. Select entities to be visualized, for example, the author's collaborative network, or keyword word cloud.

3. Adjust the variables to get the required information. For example, select the organization, adjust the time range, etc.

4. Select the desired graphic for download. As shown in the workflows below.

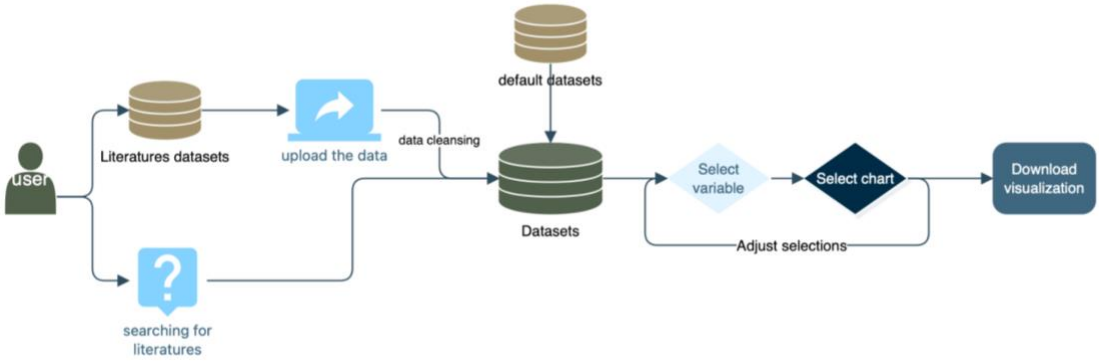

Fig 1: Workflows of visualization for literature

2.Data

2.1 Data source

The data we used were downloaded from Scopus which is a dataset that includes different kinds of articles from various journals and conferences. Scopus provides downloads of literature information, which contains information such as category, year of publication, type of literature, institution, author, grant, country, discipline, etc., which can support us in multitopic visualization. We downloaded data in psychology, anthropology, archeology, sociology and education. The data were downloaded in CSV format.

2.2 Data processing

Due to technical factors, part of the functionality of our project can not be realized directly using the original data, but we need to do some pre-processing and analysis to the original data.

Heat map: The data used to generate the heatmaps do not need to be processed and can be directly used as a CSV file containing the affiliation information downloaded from Scopus.

Line chart: Our data of line chart focused primarily on transforming raw data into a format suitable for linear graph representation. The data, covering publication counts in four disciplines from 1950 to 2023, required careful selection and cleaning to maintain consistency and accuracy. We utilized descriptive statistical methods to reveal trends over time in different disciplines. The goal was to present complex trends in the social sciences in an intuitive and comprehensible manner through a visually engaging chart.

Word cloud: Data of literature references in different subjects was first downloaded from Scopus with information on literature titles and abstracts. We used Python to separate strings into lists of words. We then used NLTK to remove common words and punctuations. The remaining key words were then ranked based on their frequency of occurrence. The transformed data was exported into CSV files. D3.cloud was used to build the word cloud engine in which more frequently used words were displayed with larger fonts. The main transformation of the data was to separate texts into words and calculate word frequencies before further visualization. Several compromises were made due to the nature of our data. First, even with all common words removed, there still remained words that don't particularly convey key information about research studies. Indexed words of literature would better serve this function. Second, for our current design, users have to preprocess the data themselves and carefully check the data to be in the correct format, as we don’t have built-in data transformation function in our webpage.

3.Visualization Designs

The intended visualization is designed to use data from literature databases and user uploads. It shows heat maps of authors who publish together, word clouds of common keywords, relations across different fields, and line charts that show the amount of publication across years. The goal is to help researchers quickly see hot topics and trends in their field without reading every paper. Researchers can filter the maps and clouds in different ways, like by country or year. Having interactive visuals instead of just lists of papers makes it easier to understand the research landscape. This saves new students time so they can focus on their own work instead of reviewing all past papers. The choices of simple visuals and interactions aim to summarize the key information from complicated literature databases. This makes staying current with research faster and more understandable.

Heatmap: We used heatmaps to give a quick visual overview of all the literature data together in one place. Users can filter the heatmap to focus on specific parts they want to understand better. Heat maps will be able to help users explore correlations between elements, including at affiliates, to determine research and collaboration trends. Heatmaps turn complex literature data into color-coded maps that summarize and highlight useful knowledge. The colors help users see insights.

Word cloud: We used word cloud to provide a simplified high-level overview of literature data that distills a large amount of papers into understandable patterns. This supports rapid analysis especially for new student researchers to familiarize themselves with a research area. It allows users to quickly identify key topics and themes. Understanding the prevalent themes facilitates many analytical tasks like identifying research hotspots.

Neural network: We developed neural networks to find hidden patterns and links between topics in literature data. They create visualizations like topic maps that show how topics connect and change over time. This helps researchers explore concepts and trends across many publications and research areas. Neural networks extract useful knowledge from literature and show it visually.

Line chart: We developed a line chart with interactive elements such as checkboxes, enabling users to selectively view data from different disciplines based on their interests. Distinct, vibrant colors were carefully chosen for each discipline to facilitate quick understanding and comparison. This design was selected for its effectiveness in demonstrating long-term trends and its flexibility in offering a customizable view to answer analytical questions.

4. Final Design and Implementation

Due to time and technology constraints, the final product includes a heat map presenting collaborative relationships between affiliates, a line graph showing the number of articles over time, and a word cloud highlighting research hotspots in the discipline.

Home Page: We have introduced the main features included in the project in the home page and provided a navigation bar for the different features. In the introduction, we detail the application scenarios of the different features to help the user select the appropriate chart. Additionally we specify the requirements for uploading files to avoid ambiguity that would make the visualization unavailable. Finally, we provide contact information so that users can ask technical questions and suggest changes.

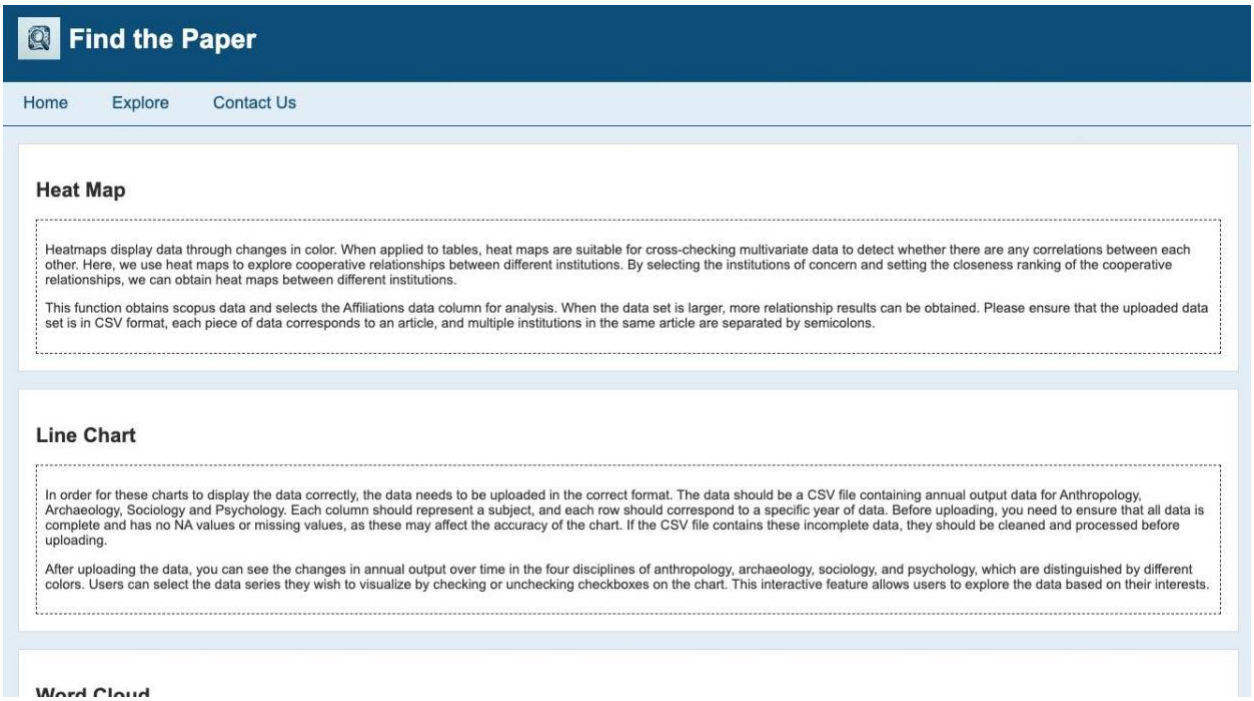

Fig 2: Screenshot of the home page

Heatmap: The final rendering of the heat map is shown below. After uploading the file, the drop-down box will automatically grab all the affiliated organizations, where users can select the ones they want to focus on to quickly get the information of their needs. At the same time, the heatmap provides an input box for selecting the number of presentations, so as to avoid the situation where there are a lot of collaborators and the information is too complicated.

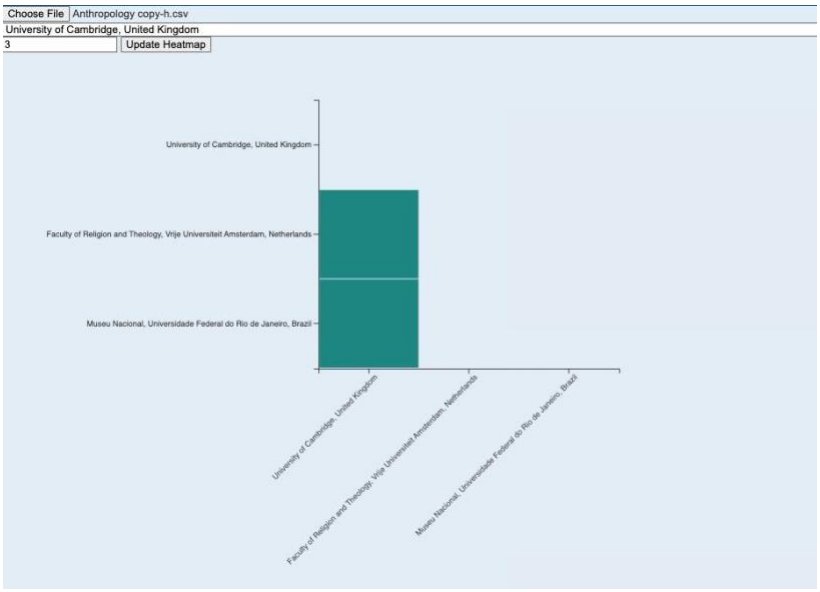

Fig 3: Screenshot of the heatmap display

Line chart: Our final outcome is an interactive line chart where users can interact with checkboxes to selectively view trends in various disciplines. The chart updates in real-time in response to user interactions, reflecting trends in different fields. Our design focus was to ensure that the chart not only presented complex data clearly but was also engaging and easy to understand.

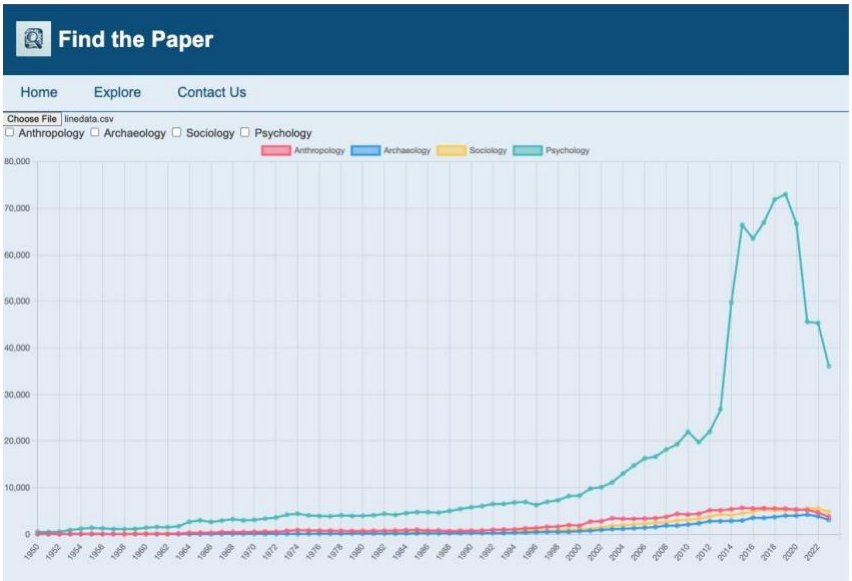

Fig 4: Screenshot of the line chart display

Word cloud: Below is the final design of the word cloud. Users can upload their CSV files and a corresponding word cloud would be generated from the file.

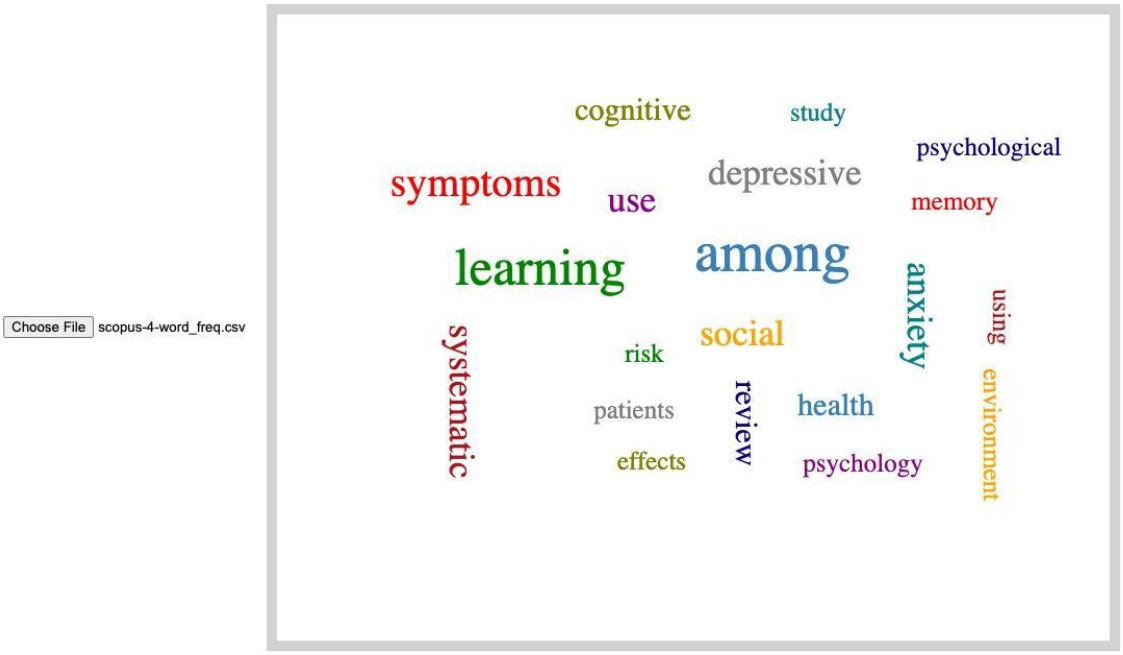

Fig 5: Screenshot of the word-cloud display

5. User Scenarios

Our projects are designed based on user requirements, different functions are proposed for a variety of user needs, with additions and subtractions made on this basis.

Heatmap: Heatmap focuses on users' need to explore partnerships. In order to help users organize a quick search for partnerships, we selected heatmaps to present the number of collaborations between different affiliates, which helps users to explore the collaboration areas centered on the target affiliates. For example, when a user wants to apply for a teaching position or a research position in a specific university, he or she would like to explore which institutions collaborate with them in this field, so that he or she can broaden the scope of job search, and can use this as a basis to explore the academic achievements of the relevant institutions in recent years, in order to achieve the purpose of rapid retrieval of academic outputs of the leading institutions in a specific field.

Line chart: For instance, in psychology, users can observe the discipline's development over the years compared to others like anthropology, archaeology, and sociology by selecting or deselecting relevant checkboxes. This feature allows users to quickly identify research hotspots and trends in psychology, such as the rise from the early 2000s to the peak in 2019 and the recent decline.

Word cloud: This word cloud was built to provide quick visual summaries of trends and themes across large literature databases. For example, a researcher interested in "What are the current research hotspots in my field?" could upload a dataset of abstracts or titles

to the visualization. They would instantly see key topic words sized proportionally to their frequency across the literature. This would highlight rising trends with larger word sizes, allowing the researcher to identify emerging popular research areas.

6. Project Management

Looking back at our project, we had to change some of our original plans, especially the complex network graphs. This was a tough choice, but it showed us that being flexible is important when managing a project. We ran into big problems figuring out the algorithms and getting the data ready. So we had to focus on other parts of the project that were more practical based on our time and skills. As the semester ended, we learned that it's important to pick goals that match what we can actually do. We also saw that we need to be able to adapt when unexpected things happen. This project wasn't easy, but working through those challenges taught us a lot about analyzing and visualizing data. Even though we didn't get to everything we first wanted to, we still grew a lot from having to rethink our plans.

7. Ethical and Societal Considerations

Our literature visualization project aims to broadly increase access to academic research. However, there are ethical considerations regarding how the information is presented, shared, and used.

In this project, we placed a strong emphasis on ethical and societal considerations. Ensuring the integrity and privacy of the data was paramount. We were careful not to misrepresent or oversimplify the complex trends within these academic disciplines. Recognizing the potential impact of our tool on shaping perceptions of research trends, we strived for transparency in our data sources and methodologies. We also considered the diverse backgrounds of our users, ensuring our design was accessible and inclusive. This ethical approach underpinned our project, ensuring it served as a responsible and valuable tool for the academic community.

On the positive side, by summarizing vast amounts of literature, our tool may allow more people to access academic insights. This could promote scientific literacy. But there are risks of oversimplification—reducing nuanced studies to data points may perpetuate misconceptions. In the future, we can mitigate this by allowing users to access fully published articles.

Overall, our goal is increased open access and scientific communication. But in lowering barriers, we must thoughtfully consider who may be excluded. With careful design and stakeholder input, we're optimistic our project can disseminate knowledge ethically. But we'll continually evaluate the societal impacts to address issues responsibly.