**STATS 401**

**Final Project Proposal**

**Streamlining Knowledge Acquisition: Enhancing Researchers' Cognitive Abilities in New Domains**

**Ziyu Huang, Lige Zhang.**

# 1. Introduction

For researchers, it can be challenging to grasp a new domain from the start. They are required to read numerous papers, which is time-consuming and difficult to distill key ideas. Additionally, they need to understand industry trends and the insights shared by other experts.

To address this, we propose an integrated solution to the question: "How can researchers' cognitive ability to a research domain be enhanced?" Our workflow is divided into three parts: First, improving researchers' efficiency in reading papers; second, summarizing the research trends within the domain; and finally, organizing the insights from reviewers.

# 2. Datasets and Techniques

For demonstration, we will mainly scrape the data from OpenReview website via this [link](https://openreview.net/group?id=ICLR.cc)[1]. It contains all information related to ICLR submissions, as well as reviewer’s comments from 2013 to 2024. We mainly rely on two potential tools, including python crawler agent, and the scraping software named ParseHub [2].

We aim to scrape the following information: paper name, keyword, abstract, reviewer’s score, confidence, proposed weakness, question. The data cleaning process may involve exploratory data analysis, semantic analysis, and data transformation. Whole data may contain approximately 1,000 items, but we might combine submissions for other conferences as well.

Later visualizations will mainly rely on Python/Python Gallery and its popular visualization packages, including Matplotlib and Seaborn, to formulate visualizations. After deciding what specific visualization idioms to be used, we will use D3 library to recreate our visualizations and publish it online.

# 3.1 How to Improve Researchers' Efficiency in Reading Papers?

Reading is a complex cognitive process that requires lots of brain works. First, we decode symbols into words, then connect those words to their meanings. Next, we apply logical reasoning to link these meanings into coherent thoughts, and finally, we critically evaluate the information to assess its validity and relevance. This entire process can be particularly challenging when dealing with abstract concepts and complex logic, as often encountered in academic writing [3][4]. Our solution focuses on enhancing one aspect of this process by incorporating visualization tools to improve comprehension.

Arguments are a key component of academic papers. According to Copi, Cohen, and McMahon, an argument is the process of arranging premises to support or prove a conclusion [5]. This is essential in any academic paper, as each paper typically presents one or more conclusions that require substantiation. In fact, a paper can be seen as a collection of interconnected arguments, where each argument may function independently as both a premise and a conclusion for other arguments. Such an argument-link graph directly aligns with the logical reasoning process involved in reading.

As Copi, Cohen, and McMahon explain, every argument can be structured in a "premise-conclusion" format, which means it can be represented visually as a link-graph (Fig.1) [5]. In this graph, the nodes represent either premises or conclusions, which has a logical structure known as "propositions”, defined as a statement that can be asserted as either true or false. The directed links in the graph have the premise as the starting point and the conclusion as the endpoint. Through this visualization, we expect to enhance researchers' understanding of the logical structures within academic papers.

The main challenge of this approach lies in distinguishing between premises, conclusions, and other descriptive or non-essential statements. Our current idea is to use indicator sheets containing conjunctive words, such as "because" or "therefore," along with articles like "a" and "the" to help identify logical components. If this method proves insufficient, we may explore natural language processing techniques to perform deeper semantic analysis and improve the identification of argument structures.

# 3.2 Visualizing the Research Trend

For research trend, we intend to create three visualizations. The first visualization is a bar chart, such that counts for the total number of submissions related to specific keyword (Fig.2). The second is a word cloud, serving a similar functional as the first visualization but in a more intuitive manner (Fig.3). From those visualizations, the researchers can identify what are the most popular research topics in the past few years and compare the relative popularity between topics.

# 3.3 Organizing the Insights from Reviewers

For reviewer comments (Attacking Points) visualization, we adopt a box plot that describes the average scores and their standard deviations that reviewers gave in the past few years (Fig.4). Also, we adopt a parallel coordinate plot, where the y-axis is the rating and x-axis contains categorical values (descriptions of reviewer’s proposed weakness) such as lacking experiments or overlooking related works (Fig. 5). From those visualizations, users can analyze and discover what reviewers usually care about and why they give negative scores.

# 4. Sketches

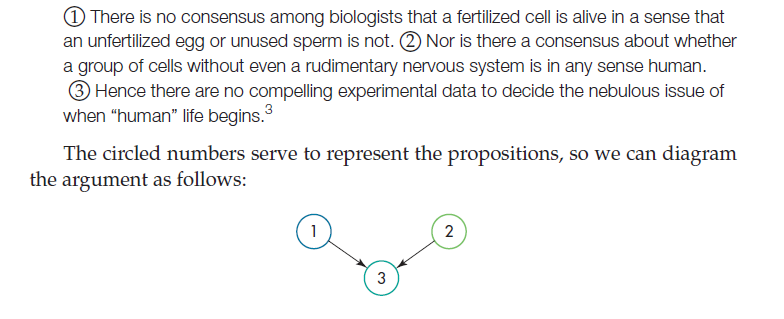


Fig.1 “premise-conclusion” argument link graph [5]: Each note represents either a premise or a conclusion, and the link denote to their logical relation.

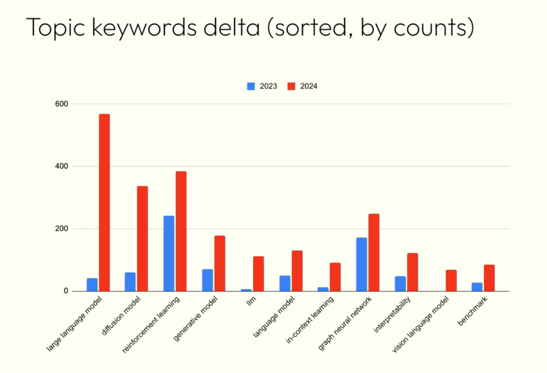


Fig.2 topic trend example: x-axis represents research topics and y-axis represents total number of submissions.

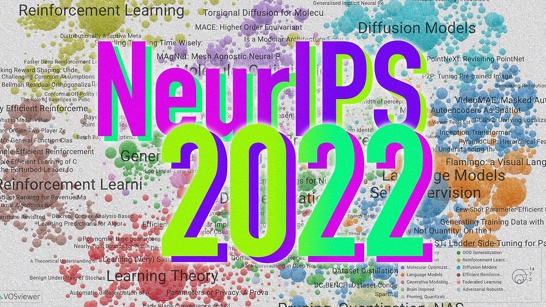


Fig.3 Word cloud example: A word cloud showing hot research topics.

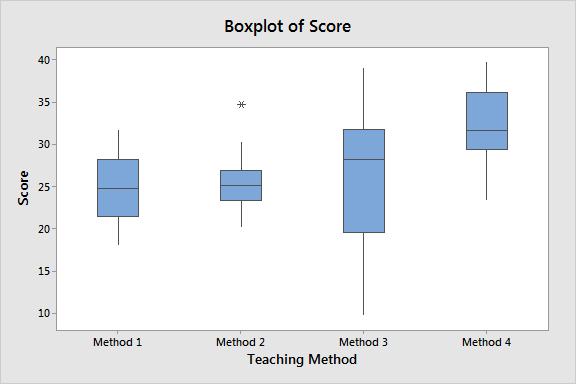


Fig.4 review score distribution example: x-axis represents years and y-axis represents scores (average and standard deviation).

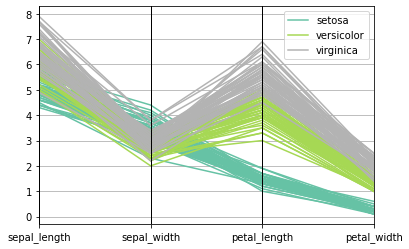


Fig.5 Reviewer comments distribution example: x-axis represents weakness pointed out by the reviewer; y-axis represents score of the associated paper.

# 5. Roles and Responsibilities of each Group Member.

|  |  |
| --- | --- |
|  | Tasks |
| Ziyu | 1. Scrape dataset 2. Conduct part 1 of the research question |
| lige | 1. Scrape dataset 2. Conduct part 2&3 of the research question |

# 6. Deliverables for the Interim Presentation.

1. Scraping data and some simple visualizations

2. Simple complementation for question 1 (semantic analysis & feasibility analysis)

3. More visualizations for question 2&3, and some simple discussion

# 7. Timeline and Milestones.

|  |  |  |
| --- | --- | --- |
|  | ZIYU | LIGE |
| Week5 | Finish scraping data and data preprocessing | Finish scraping data and data preprocessing |
| Week6 | Simple complementation for question 1 | Finish question 2&3 visualizations |
| Week7 | Develop an integrated solution for question 1 | Publish visualization online and wrap up |

# References

1. OpenReview.net. ICLR - International Conference on Learning Representations. <https://openreview.net/group?id=ICLR.cc>
2. "ParseHub." ParseHub, <https://www.parsehub.com/>.
3. Hoover, Wesley A., and Philip B. Gough. "The simple view of reading." *Reading and writing* 2 (1990): 127-160.
4. Rayner, Keith, et al. *Psychology of reading*. Psychology Press, 2012.
5. Copi, Irving M., Carl Cohen, and Kenneth McMahon. *Introduction to logic*. Routledge, 2016.