Automated Collection and Review of Academic Papers Using AI

Hiroya Taguchi
htaguchi@stud.hs-heilbronn.de
Heilbronn University of Applied Sciences
Heilbronn, Baden-Württemberg, Germany

1 INTRODUCTION

The essential task of gathering relevant literature for new research or a review paper is becoming increasingly daunting in the academic field due to the explosion of published works. This necessity, rooted in the academic principle of building upon existing knowledge, faces significant hurdles in the current landscape. When querying databases like Google Scholar, researchers are swamped with vast numbers of papers, making it challenging to pinpoint those truly pertinent to their specific topic. This challenge is more than volume but also about the relevance and depth of connection between works, which automated search algorithms may not fully capture.

Additionally, the process is complicated by the varied nature of academic databases such as arXiv and IEEE Xplore, each with unique indexing and search functionalities. This inconsistency requires researchers to adapt their search strategies for each platform, leading to increased time investment and the risk of encountering duplicate entries of the same paper across different databases.

To tackle the complexities and inefficiencies associated with the manual collection and review of academic literature, this project has developed a pioneering program that automates this critical process. Users can input specific search criteria, such as keywords and desired journals, and the program swiftly aggregates relevant papers, organizing them into an easily navigable Excel list. This expedites the initial research phase and introduces a level of precision and comprehensiveness previously unattainable through manual methods.

Beyond mere aggregation, the program harnesses advanced clustering algorithms to categorize the literature into thematic groups, thus offering a structured and insightful overview of the research landscape. This feature is precious for discerning prevalent themes, interconnections between studies, and emerging trends within a given field.

To validate the utility of this automated system, a focused case study was conducted within the domain of "Reinforcement Learning with Human Feedback (RLHF)." A meticulously curated list of seminal papers in this area was a benchmark for assessing the program's effectiveness. The primary criterion for evaluation was the program's ability to encompass the entire scope of the benchmark list, demonstrating its capability to capture the relevant literature thoroughly.

Moreover, the clustering functionality provided additional analytical insights by identifying distinct subtopics and patterns among the collected papers. The program also incorporates features for the statistical analysis of the papers' metadata, such as citation counts and publication dates, further enriching the review process by highlighting key works and trends over time.

This project's introduction of an automated literature collection and review system marks a significant leap forward in academic research methodologies. By streamlining this foundational task, the program allows researchers to devote more resources to the substantive literature analysis, thereby enhancing the overall quality and efficiency of academic inquiries.

2 METHODOLOGY

2.1 Data Collection using Findpapers

For the data collection process, Findpapers library[1] streamlines the process for researchers seeking references by performing searches across multiple databases, including ACM, arXiv, bioRxiv, IEEE, medRxiv, PubMed, and Scopus, based on a user-defined query. The tool's process encompasses defining the query, executing the search, refining the results, and obtaining full-text papers.

Upon configuring the search parameters, executing the Findpapers tool automatically generated a JSON file containing detailed information on the research papers. The parameters for the search were as follows: Search Strings: Search queries incorporating relevant keywords and logical operators (AND, OR) to precisely target the desired research papers, Publication Date, Number of Papers, and Database Selection.

2.2 Exporting data from JSON to Excel

Once the JSON data was imported, it parsed it to extract essential information from each paper. This included the title, publication date, abstract, list of authors, databases where the paper was listed, publisher details, journal name, keywords, DOI, and citation count. After the extraction process, it compiled this information into a pandas DataFrame.

Finally, the DataFrame was saved as an Excel file. This file served as the basis for our subsequent data analysis tasks, such as clustering and thematic exploration of the research papers.

2.3 Cleaning Abstract Text

The function is designed to preprocess and clean text data for our analysis. It first checks for and handles any missing values by returning an empty string for NaN or None inputs. It then removes any HTML tags and checks whether the text indicates the absence of an abstract. It also verifies that the text is in English, discarding any non-English text. After these initial checks, the function strips away unnecessary predefined patterns and unwanted characters for our analysis, ensuring that only alphabetic characters remain. Finally, it converts all text to lowercase to maintain consistency in the dataset.

2.4 Generating TF-IDF Feature Matrices

The TfidfVectorizer generates a TF-IDF feature matrix from the cleaned abstract texts. This matrix quantifies the importance of words within the text data, assigning higher values to frequent terms in a particular document but rare across the entire document corpus. This nuanced approach allows us to capture the unique context of each abstract, providing a rich set of features for clustering and text analysis. Transforming the abstracts into this numerical format facilitates sophisticated analytical techniques to uncover patterns and groupings within the data, enabling a deeper understanding of the underlying thematic structures.

2.5 Determining the Number of Clusters and Clustering with MiniBatchKMeans

The elbow method, a widely used technique in cluster analysis, was employed to identify our dataset's optimal number of clusters. This method involves plotting the sum of squared distances of samples to their nearest cluster center against the number of clusters. By observing the point at which the plot starts to flatten, often referred to as the 'elbow,' it can determine a suitable number of clusters that balances compactness and separation.

Then, the MiniBatchKMeans algorithm, an efficient variant of the K-Means clustering algorithm, is applied to the TF-IDF feature matrix generated from the cleaned abstract texts. MiniBatchKMeans is particularly well-suited for large datasets due to its batch-based approach to optimization, which significantly reduces computational time while still delivering robust clustering results.

After fitting the model and predicting cluster labels for each document, the entire dataset containing them was saved to an Excel file.

3 RESULTS AND ANALYSIS

3.1 Conditions for Evaluation and Collection Results

For this evaluation, the query was set as "[Reinforcement learning] AND [Human] AND ([Feedback] OR [Preference])" with the publication date range from 2013 to 2023. The databases selected for the search were ACM, arXiv, IEEE, and Scopus. The initial data collection yielded 1,868 records, which took over 2 hours under the environment of Google Colab's GPU usage. This duration is attributed to the varying data collection speeds across the databases, with arXiv being an exception. The process is inherently slower for the other databases as they require individual data transactions, meaning the more records there are, the slower the process becomes. However, considering that the program can be run in the background, is automated, and excludes duplicates and predatory publishers, it is more resource-efficient and simpler than manually collecting and verifying data from each database individually.

3.2 Clustering

In this study, an attempt was made to determine the optimal number of clusters using the elbow method. However, due to the unclear boundaries in the distribution, a suitable number of clusters could not be identified. Consequently, the decision was made to set the number of clusters to 10. This number was chosen to create clusters that could be narrowed down to around 100 papers, a manageable number for review papers or reference searching.

For the mapping of clustering results, t-SNE was utilized. t-SNE is a technique for dimensionality reduction that is particularly well-suited for the visualization of high-dimensional datasets[2]. It works by converting similarities between data points to joint probabilities. This approach is beneficial as it allows for the visualization of clusters in a two-dimensional space, making it easier to identify patterns and groupings within the data.

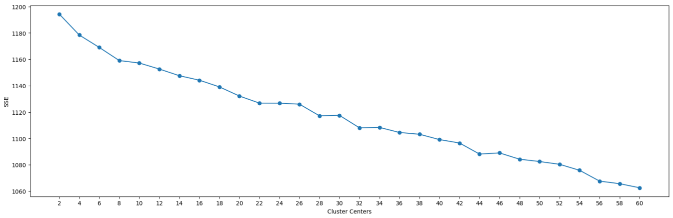

Figure 1: Elbow Method

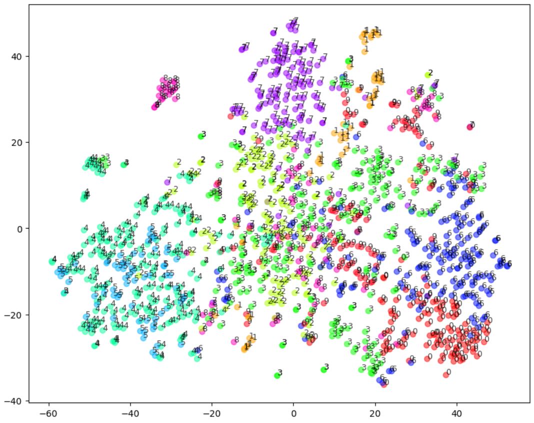

Figure 2: t-SNE Cluster Plot

The parameters for the TfidfVectorizer were set based on standard practices, with the 'min' set to 5, 'max' to 0.95. These parameters were chosen after experimenting with different 'max' and 'min' values, which did not significantly alter the clustering outcomes. Therefore, a general configuration was adopted[3].

Although a precise categorization was not achieved, the clustering resulted in a distribution of papers across the clusters, ranging from 52 to 226 per cluster. This distribution is sufficient for writing a review paper, as it provides a broad overview of the literature without being overwhelming.

3.3 Quality of Paper Collection

To assess the quality of the collected dataset, a set of 41 "control papers," considered model cases for the research topic, was prepared. The evaluation focused on how well the collected dataset covered this list and how these papers were distributed across the clusters. Of the 41 control papers, 23, approximately 56%, were included in the generated list. An analysis of the abstracts revealed that the control papers included in the list contained query words in their abstracts, indicating relevance to the search criteria. On the other hand, the control papers not included in the list typically lacked elements relevant to the query. For instance, some papers might have been related to Reinforcement Learning and Human Feedback but did not explicitly include one of these elements in their content or abstract. Others could be related to Large Language Models (LLMs) or GPT, inherently involving RLHF elements but without the keywords explicitly mentioned in their abstracts. Additionally, some papers might have only used abbreviations such as "RL from human feedback," making them miss the query due to the absence of full-term matches.

This assessment highlights the importance of query design in the data collection process and suggests areas for improvement in capturing relevant literature. The presence of relevant terms in the abstract and full terms versus abbreviations plays a significant role in whether a paper is included in the search results, impacting the overall quality and comprehensiveness of the collected dataset.

3.4 Evaluation Regarding the Creation of a Review Paper

In creating a review paper, the PRISMA Framework was referenced to conduct an actual collection of reference literature, through which an evaluation was made[4].

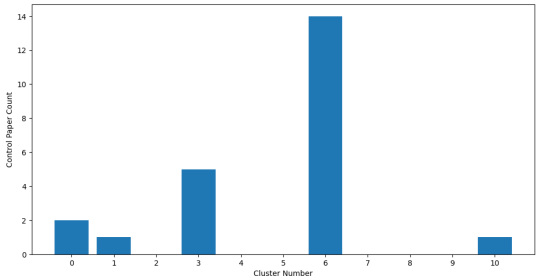

Figure 3: Number of Control Papers in Each Cluster

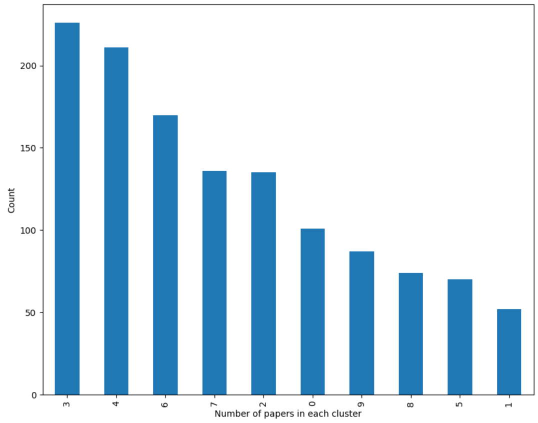

Figure 4: Number of Collected Papers in Each Cluster

Initially, a total of 1,728 papers were collected. Following the clustering into ten groups, the papers were distributed across clusters ranging from 226 to 52 in size. The cluster with the highest distribution of control papers was Cluster No. 6, containing 170 papers.

To understand the theme of this cluster, a review of the top 10 keywords was conducted, revealing terms such as 'rl,' 'feedback,' 'language,' 'preferences,' 'preference,' 'human,' 'models,' 'reward,' 'model,' and 'rlhf.' This suggests that the cluster likely contains papers related to RLHF, similar to the control papers.

Since writing an actual review paper would require around 100 papers, selecting from this cluster should be sufficient. Exploring other similar clusters where control papers are distributed may also be beneficial, thereby enhancing the breadth of literature covered in the review. Consequently, the effectiveness of this program in facilitating the creation of a review paper is affirmed, as it efficiently narrows down a large dataset to a manageable and relevant subset of papers.

4 TREND ANALYSIS

The program used for this study includes code to automatically generate statistical graphs for analysis from the Excel list created. This section summarizes the insights gained from the data related to RLHF.

4.1 Keywords

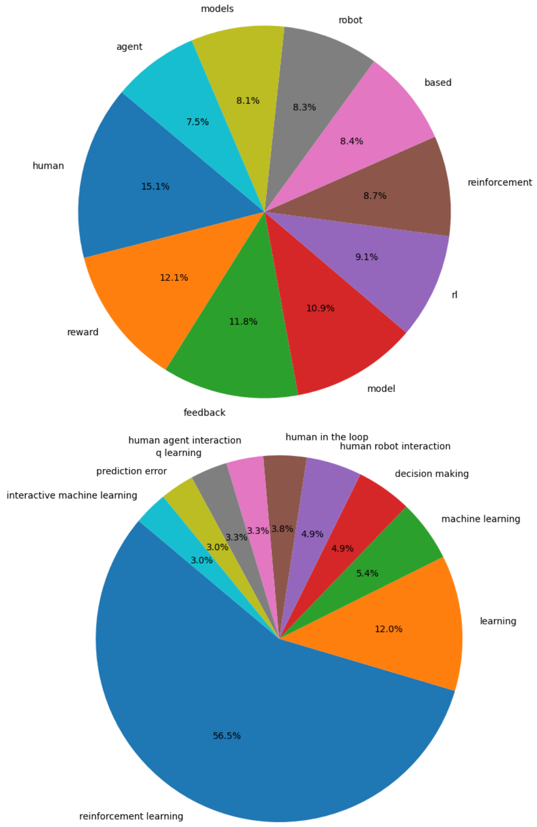

Figure 5: TF-IDF Keywords (Top) and Manually Entered Keywords by the Author (Bottom)

A comparative analysis was conducted between the top 10 words with high TF-IDF scores and the top 10 keywords manually input by the authors of the papers, presented in pie charts. The TF-IDF graph, by its nature, only provides scores for individual words, which makes it challenging to discern whether words like "learning" are used in the context of "reinforcement learning" or other phrases or as standalone terms when compared to the information provided by the authors. This comparison concluded that the TF-IDF-based analysis yielded less helpful information. Standard machine learning terms such as "reward" or "model," frequently appearing with high scores, further obscured the extraction of meaningful insights. Adjusting the TF-IDF parameters did little to mitigate this issue, indicating that a more nuanced approach may be required to extract valuable insights from keyword analysis in this context.

4.2 Publication Number by Year

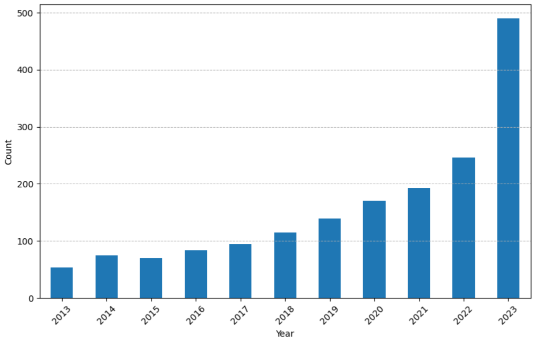

Figure 6: Publication Count by Year

The number of publications has significantly increased from 55 in 2013 to 510 in 2023, nearly a tenfold increase over the decade. Ntably, there was a substantial rise in publications from 266 to 510 between 2022 and 2023 alone, nearly doubling in just one year. This trend indicates a growing interest in RLHF, as the data shows.

4.3 Author Analysis

An attempt was made to aggregate data on the first authors of RLHF papers. However, challenges were encountered due to variations in author name representations, such as abbreviations or omissions of parts of names in different journals. This issue suggests a need for more sophisticated methods, such as GPT, to standardize author names and accurately aggregate data, minimizing discrepancies in author name representations.

4.4 Journal Distribution

Analyzing the journals that publish RLHF papers reveals the diversity of fields engaging with this topic. Many publications in neuroscience-related journals indicate a strong interest in RLHF within this discipline. Other notable fields include robotics and psychophysiology, showcasing the interdisciplinary application of RLHF. This distribution highlights the wide-ranging impact and relevance of RLHF across various scientific and research domains.

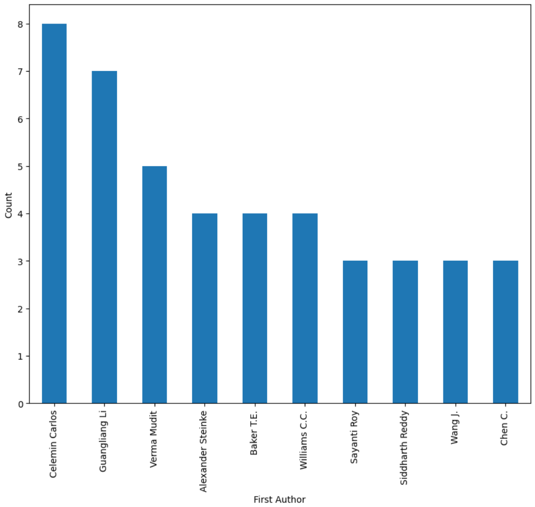

Figure 7: Number of papers per author

5 DISCUSSION

In this study, the tool "Findpaper" was utilized to collect research papers. However, there were instances where other crucial information was missing despite the titles being available. Direct scraping from individual websites might recover such missing data.

Additionally, the function intended to eliminate duplicates automatically might have needed to operate more effectively, as there were instances of papers being included that were duplicated across different databases. This necessitates a verification process to ensure data integrity.

When generating statistical data based on names, inconsistencies were observed, such as variations in the order of first and last names or using initials. Accounting for the diversity and cultural differences in naming conventions is challenging, and logical settings have limitations. Exploring natural language processing might offer a solution to these challenges.

This study performed clustering based on abstracts, but it was observed that not all relevant papers contain the search query keywords in their abstracts. Expanding clustering from abstracts to full texts could potentially enhance the process by incorporating a broader range of information for analysis.

While TF-IDF is a fundamental technique that has been around for quite some time, there may be more effective approaches for this case than using word frequency for clustering. This is primarily because, in this context, phrases consisting of two or more words hold significant importance. For instance, rather than treating "human" and "feedback" as separate entities, "human feedback" forms a crucial keyword in its entirety.

Furthermore, the ultimate goal is not just to cluster papers based on the proximity of individual words but to identify semantically similar papers in their entire textual content. This approach aims to understand papers' overall meaning and context, which may share common themes or research focuses even if they do not necessarily use the exact keywords.

In practice, using TF-IDF followed by K-means clustering did not yield an optimal cluster distribution for this study. This outcome suggests the potential benefit of exploring alternative methods, such as DBSCAN, LLMs, or BERT (Bidirectional Encoder Representations from Transformers), which could provide more meaningful results by capturing the semantic similarities between papers more effectively.

In conclusion, adapting to more advanced and contextually aware methodologies could significantly enhance the process of clustering and analyzing research papers, particularly in fields where the contextual meaning and comprehensive understanding of textual content are crucial.

REFERENCES

[1] https://github.com/jonatasgrosman/findpapers

[2] https://www.kaggle.com/code/jbencina/clustering-documents-with-tfidf-and-kmeans

[3] https://www.kaggle.com/code/meuge672/tf-idf-and-knn-in-people-wikipedia-dataset

[4] http://www.prisma-statement.org/PRISMAStatement/FlowDiagram