Matching Job Market Applicants to Universities Based on Research Interests

Hassan Gudal
hassangudal10@gmail.com

Abstract

In this paper, we present a data-driven approach to match job market applicants to universities based on their research interests and the research profiles of the institutions. We collected data from 100 PhD students or professors who met a set of predefined criteria. Utilizing the top three most frequently occurring keywords in their OpenReview papers, we generated affinity scores between each applicant and 27 distinct research areas using text embedding techniques. Additionally, we created unique profiles for the top 22 US universities listed on the CSRankings website based on their publication counts across each research area. By combining the affinity scores and university profiles, we constructed a similarity matrix to quantify the compatibility between applicants and universities. We explored a wide range of matching algorithms to optimally assign applicants to universities. Our results demonstrate the potential of data-driven approaches for enhancing job market matching in academia and provide valuable insights for both job seekers and hiring institutions.

1 Introduction

The inspiration for this research project emerged from a discussion I had with graduate student Justin Payan and Professor Yair Zick, who served as my mentors for the past year. During a convention, Payan had met and spoken with a colleague regarding the potential optimization of researcher assignment to institutions based on various factors. In a subsequent meeting, Payan proposed this concept as a potential research project that would provide an opportunity for me to develop new skills, given my lack of prior research experience. The idea was compelling, and as a result, our regular meetings focused on the project's progress and receiving guidance from Payan and Zick throughout the course of the project as they both have many publications to their names.

As a senior majoring in computer science at the University of Massachusetts Amherst, this project presented an opportunity to expand my knowledge and skills in several key areas. Firstly, I aimed to gain practical experience in data retrieval and processing by utilizing the OpenReview platform and natural language processing techniques, such as the SBERT model. Secondly, I sought to deepen my understanding of optimization algorithms and their applications in real-world scenarios. Lastly, this project helped to enhance my proficiency in data visualization and presentation using Excel, a valuable skill for effectively communicating research findings. Ultimately, this project served as a stepping stone in my academic journey, equipping me with the necessary tools and knowledge to pursue further research and contribute to the ever-evolving landscape of computer science.

Given the data-driven nature of this project, my primary hypothesis revolved around the potential of leveraging natural language processing techniques and optimization algorithms to effectively match job market applicants to universities based on their research interests and the research profiles of the institutions. By calculating affinity scores between applicants' research keywords and predefined research areas, and subsequently creating university profiles based on publication counts in each area, I hypothesized that a similarity matrix could be constructed to quantify the compatibility between

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

applicants and universities. Furthermore, I posited that exploring different matching algorithms, such as utilitarian and egalitarian approaches, among others, would provide valuable insights into the optimal assignment of applicants to universities, taking into account the overall compatibility scores and the range of scores across all matches. Through this project, I aimed to test these hypotheses and assess the viability of data-driven approaches in streamlining the job market matching process in academia, ultimately contributing to a more efficient and effective hiring process.

1.1 Our Contributions

In this paper, we make the following contributions:

- We propose a novel data-driven approach for matching job market applicants to universities based on their research interests and the research profiles of the institutions. Our methodology leverages natural language processing techniques, similarity measures, and optimization algorithms to create an effective matching system.
- We conduct experiments using real-world data from the NeurIPS 2023 conference and the CSRankings website, utilizing various matching algorithms to find optimal assignments.
- We identify limitations in the current OpenReview tagging system and propose enhancements for future iterations of the project, such as increasing the pool of researchers, focusing on those who have studied or worked in the United States, and exploring alternative keyword systems like DBLP to improve the accuracy and consistency of research interest representation.

1.2 Related Work

While research on matching academics to institutions appears to be limited, several studies have explored the collaboration between academic researchers and industry partners for joint research projects. Manotungvorapun and Gerdsri (2019) proposed a systematic approach for firms to identify ideal academic partners based on the similarities between a researcher's body of work and the industry partner's research interests. This study highlights the potential for extending our research to include corporate and academic collaboration as a next step.

In a related study, Mindruta (2013) investigated the matching process between academic scientists and industry research firms. The findings suggest that academics with specialized focus achieve better results when partnered with research firms that have a more generalized scope, as their skills complement each other. This raises an interesting question for our research: would more specialized computer science researchers benefit from joining universities where people focus on different specialized topics, rather than institutions with similar research interests? This insight could inform hiring decisions, suggesting that institutions should avoid pigeonholing themselves in the hiring process.

These studies provide valuable context for our research on matching job market applicants to universities based on research interests and profiles. While our focus is on academic job market matching, the findings from corporate-academic collaboration research offer potential avenues for future work and highlight the importance of considering the complementarity of skills and research interests in the matching process.

2 Problem Statement

Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of n applicants and $U = \{u_1, u_2, \ldots, u_m\}$ be a set of m universities. Each applicant a_i has a set of top keywords $K_i = \{k_{i1}, k_{i2}, k_{i3}\}$ extracted from their research papers. Let $R = \{r_1, r_2, \ldots, r_p\}$ be a set of p distinct research areas.

Define an affinity score function $\alpha : A \times R \to [0, 1]$, where $\alpha(a_i, r_j)$ represents the affinity score between applicant a_i and research area r_j, calculated using the SentenceTransformer model ϕ as follows: $\alpha(a_i, r_j) = \frac{1}{|K_i|} \sum_{k \in K_i} \phi(k) \cdot \phi(r_j)$

Let $P = \{p_{u_1}, p_{u_2}, \ldots, p_{u_m}\}$ be the set of university profiles, where each profile p_{u_j} is a vector of publication counts in each research area: $\mathrm{p}_{u_j} = [c_{u_j,r_1}, c_{u_j,r_2}, \ldots, c_{u_j,r_p}]$

Define a similarity score function $\sigma : A \times U \to \mathbb{R}$, where $\sigma(a_i, u_j)$ represents the similarity score between applicant a_i and university u_j, calculated as follows: $\sigma(a_i, u_j) = \frac{\sum_{k=1}^{p} \alpha(a_i, r_k) \cdot c_{u_j, r_k}}{\sum_{k=1}^{p} c_{u_j, r_k}}$

Let S be the similarity matrix of size $n \times m$, where $S_{ij} = \sigma(a_i, u_j)$ represents the similarity score between applicant a_i and university u_j.

Define a matching $\mu : A \to U$, where $\mu(a_i) = u_j$ indicates that applicant a_i is matched with university u_j. Let X be a binary matrix of size $n \times m$, where $X_{ij} = 1$ if $\mu(a_i) = u_j$, and $X_{ij} = 0$ otherwise.

The objective is to find an optimal matching μ^* (or equivalently, an optimal binary matrix X^*) based on different criteria:

1. Most Similar Matching: For each university u_j, find the applicant a_i with the highest similarity score $\sigma(a_i, u_j)$.
2. Least Similar Matching: For each university u_j, find the applicant a_i with the lowest similarity score $\sigma(a_i, u_j)$.
3. Utilitarian Matching: Find a matching μ^* that maximizes the sum of similarity scores: $\mu^* = \arg\max_\mu \sum_{i=1}^{n} \sigma(a_i, \mu(a_i))$
4. Egalitarian Matching: Find a matching μ^* that maximizes the minimum similarity score: $\mu^* = \arg\max_\mu \min_{i=1}^{n} \sigma(a_i, \mu(a_i))$

The utilitarian and egalitarian matching problems can be formulated as linear assignment problems and solved using the Hungarian algorithm.

3 Experiment

The code for this research paper can be found at: https://github.com/hgudal/ResearchPaper

The dataset used in this study consists of 100 researchers from diverse institutions across the globe who submitted papers to the NeurIPS 2023 conference, published at least three research papers, and were actively seeking employment in 2021. These researchers were matched to the top 22 universities according to the CSrankings as of April 2024. It is important to note that the CSrankings data has since been updated, and for more accurate results, it may be beneficial to manually update the university list.

The implementation of our project was done using Python, and the code can be easily executed in an integrated development environment such as VScode or a similar application by simply running the code. The program does not require any manual inputs, but users can easily modify certain values to customize the results according to their preferences.

The execution time of the code is approximately 45 minutes to an hour, resulting in the generation of an Excel file containing the matching results. In the example file provided, we successfully identified three true pairs, demonstrating the effectiveness of our approach. One of these pairs was obtained from the most similar authors matching algorithm, where Fisher Yu was matched to the University of California, Berkeley. The other two pairs were derived from the utilitarian matching algorithm, with Fisher Yu being matched to the University of California, Berkeley, and Marcel Torne Villasevil being matched to the Massachusetts Institute of Technology.

These results highlight the potential of our data-driven approach in accurately matching job market applicants to universities based on their research interests and the research profiles of the institutions. The recovery of true pairs, both from the most similar authors matching and the utilitarian matching algorithm, validates the effectiveness of our methodology in identifying compatible matches between applicants and universities.

4 Conclusion

In this paper, we presented a data-driven approach for matching job market applicants to universities based on their research interests and the research profiles of the institutions. Our experiment, which

utilized a dataset of 100 researchers from various institutions worldwide, resulted in three successful matches between applicants and universities.

Upon analyzing the results, we observed that the scope of the experiment could have been more focused on researchers within the United States. Many of the researchers in the dataset had strong ties to countries such as China, Korea, India and Switzerland, and they may prefer to continue their research within their respective countries. This geographical factor could have influenced the matching process and potentially limited the applicability of the results to the United States job market.

During the analysis of the similarity matrix, we noticed an interesting phenomenon where researchers tended to receive either consistently high or low similarity scores across the board. This observation led us to investigate the tagging system used by OpenReview, which is user-generated. Upon further examination, we found that researchers with low similarity scores typically had longer and more complex keywords, and these keywords were often fewer in number. In contrast, researchers with higher similarity scores had more keywords that were shorter and more generalized. This finding suggests that the OpenReview tagging system may not be ideal for accurately capturing the research interests of applicants. To improve the accuracy of our approach, we propose two enhancements for future iterations of this project. Firstly, increasing the pool of researchers and limiting it to those who have studied or worked in the United States could provide more relevant and targeted results. Secondly, exploring alternative keyword systems, such as those used by DBLP, could potentially improve the quality and consistency of the research interest representation.

Looking ahead, this project has the potential to evolve into a valuable service for both job market applicants and institutions. Applicants could leverage their body of work to find the best-fit institutions for their research interests. Institutions could utilize the software in a manner similar to the National Resident Matching Program which is used to place medical school students into residency training programs, to determine the optimal placement of researchers within their universities.

In conclusion, our data-driven approach for matching job market applicants to universities based on research interests shows promise in improving the efficiency and effectiveness of the academic job market matching process. With further refinements and enhancements, this methodology could revolutionize the way researchers and institutions connect, ultimately fostering better collaborations and driving innovation in the field of computer science.