PostAc®: A Visual Interactive Search, Exploration, and Analysis Platform for PhD Intensive Job Postings

Chenchen Xu[1,2] **Inger Mewburn**[1] **Will J Grant**[1] **Hanna Suominen**[1−4]
1. The Australian National University (ANU) / Canberra, ACT, Australia
2. Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) / Canberra, ACT, Australia
3. University of Canberra / Canberra, ACT, Australia
4. University of Turku / Turku, Finland
Firstname.Lastname@anu.edu.au

Abstract

Over 60% of Australian PhD graduates land their first job after graduation outside academia, but this job market remains largely hidden to these job seekers. Employers' low awareness and interest in attracting PhD graduates means that the term "PhD" is rarely used as a keyword in job advertisements; 80% of companies looking to employ similar researchers do not specifically ask for a PhD qualification. As a result, typing in PhD to a job search engine tends to return mostly academic jobs. We set out to make the market for advanced research skills more visible to job seekers. In this paper, we present PostAc®, an online platform of authentic job postings that helps PhD graduates sharpen their career thinking. The platform is underpinned by research on the key factors that identify what an employer is looking for when they want to hire a highly skilled researcher. Its ranking model leverages the free-form text embedded in the job description to quantify the most sought-after PhD skills and educate information seekers about the Australian job-market appetite for PhD skills. The platform makes visible the geographic location, industry sector, job title, working hours, continuity, and wage of the research intensive jobs. This is the first data-driven exploration in this field. Both empirical results and online platform will be presented in this paper.

1 Introduction

The PhD was originally conceived - and is usually understood - to mark the commencement of an academic career. Yet the degree has never been entirely fit for purpose: as early as 90 years ago, Dale (1930) questioned the role of academic degree. But since then, both academic and industry needs have changed dramatically.

On the academic side, changing workforce structures over the last few decades have meant PhD graduates have faced ever greater difficulties landing academic employment (Bazeley et al., 1996). In Australia, recent research has revealed that up to 60% of students end up working outside of academia, making us ask whether their academic training is really fit for their final purpose (McGagh et al., 2016).

Outside academia, governments are starting to recognize the importance of highly trained graduates to innovation, and are thus putting pressure on universities to re-think PhD curricula so as to target both academic and wider industry needs (Mewburn et al., 2016). Yet limited data-driven research exists to explain how having a PhD actually impacts job seeking in non-academic sectors. Meanwhile, about 80% of the companies looking to employ highly skilled researchers do not specifically ask for PhD qualifications (Mewburn et al., 2018). In this paper, we demonstrate an online platform — PostAc® (short video) — that allows users to explore non-academic career options at scale and is able to accommodate a dynamic industry environment as the model evolves. This educational technology artefact builds on an exploratory study developed through multiple iterations of expert annotation, modelling, and empirical evaluation. The final fine-tuned model is able to correctly categorise jobs requiring PhD level skills at an accuracy of 88%.

We make the following three key contributions: First, we visualize probably the first job posting data set with labels from domain experts showing the intensity of PhD-level research skills. Second, we present a ranking-based model that has been successfully applied to predicting PhD skills intensity from job postings, with empirical performance evaluation. Third, we design and construct a real-world online platform that offers PhD graduates a dedicated job search functionality, as well as helps governments, universities, and employers

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 43–48
Florence, Italy, July 28 - August 2, 2019.

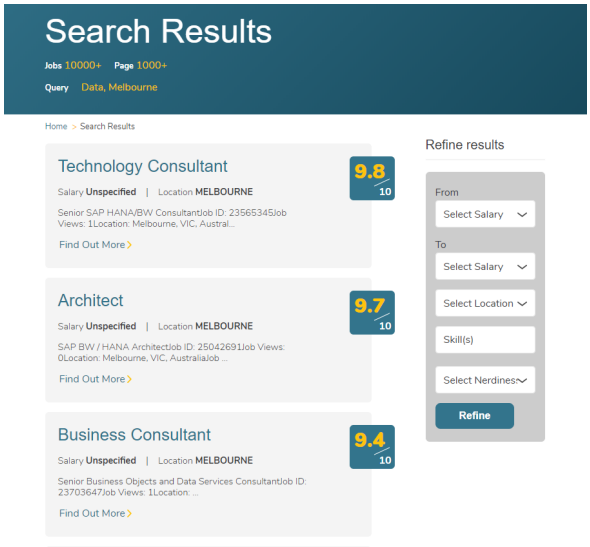

Figure 1: Search Results in the Exploration View

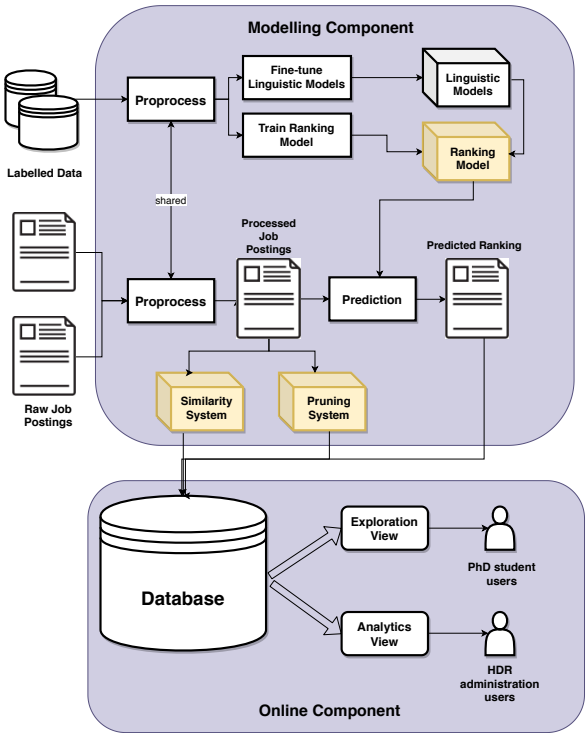

Figure 2: System Architecture

in increasing the understanding of different industries' absorption of PhD graduates.

Since its launch in late 2018, PostAc® has been sharpening the research career thinking of over 1,300 participating PhD students. Its analysis scales out for over 1.2 million job advertisements to quantify the most sought-after PhD skills and educates information seekers about the Australian job-market appetite for PhD skills in terms of geographic location, industry sector, job title, working hours, continuity, and wage. Its 2017 pilot (Mewburn et al., 2018) revealed the hidden job market for research talents to the government.

2 Data Set

To the best of our knowledge, no empirical studies on big data have previously been conducted in this field, so we commenced the work by preparing our own data set. Over 1.2 million jobs postings published during 2015 were collected from Burning Glass International Inc. as the seeding data set. Each posting came with the original job title and job description, as well as 41 unique attributes, including the employer, salary, and discipline codes. As in this study we sought to understand and support PhD graduates finding careers outside academic institutions, academic jobs (university lecturers, fellows, professors, etc.) were removed (approximately 1%).

To facilitate the study of PhD-shaped jobs and the training of our ranking model, human experts manually annotated 1,315 job postings based on an agreed schema (details can be found in (Mewburn et al., 2018)) so that each job is associated with a 'PhDness score (Figure 1) that took values from 1 (least PhD-shaped) to 10 (most PhD-shaped). As expected (not that many jobs require PhD skills), the highly ranked jobs comprised only a small proportion of the entire set. We alleviated this imbalance in the generation of the annotated data set by adding a simple rule-based filter after the random sampling process, resulting in jobs fairly unlikely to require a PhD (e.g., a job paid by the hour) being removed.

3 System Overview

The PostAc® platform is structured into two major components based on the consideration of progressive enhancement of analytic models and platform scalability (Figure 2). More specifically, the model fine-tuning component has been separated out and thus can run in parallel with the continuous integration of data for the online platform component. The general data processing pipeline is shared between the two components to guarantee a consistent process can be applied to data from different sources. After this, the fine-tuning process is invoked on the data set to re-train the ranking model. This process is also responsible for preparing models to handle the extraction of important linguistic attributes, which will later be used in

the construction of the database (e.g., tokenization for full-text searching and semantic embedding for similarity measurement) and fine-tuning of word embeddings. The online component leverages the models from the modelling component to digest the incoming job data set, which is mostly unlabelled. Along with the ranking results, the jobs are enriched with the aforementioned linguistic attributes before finally being saved to the database. Regarded as important principles in the design of any system, the scalability and modularity are examined upon each component to be integrated.

The modelling component is built with Tensorflow, where a scheduling system arranges the fine-tuning work in a distributed manner. Meanwhile, the storage engine is built on top of ElasticSearch, making it possible to handle the digestion of approximately $100,000$ job postings coming monthly, as well as to support future extension.

Since the database is prepared in the backend engine, PostAc® provides two dedicated view flows for the needs of both PhD students and staff members (e.g., careers advisors and curriculum designers). PhD students can use the Exploration view flow to search, compare, and investigate the millions of jobs available on the system. Their behaviors can be analyzed as implicit feedback to further enhance the training data set, and thus contribute to optimization of the modelling component. Staff members from universities and academic institutions will be given access to the Analytics view, allowing them to improve their understanding of the potential job market for PhD graduates, and high degree education policy making.

4 System Features

In addition to our major objective of revealing those jobs most likely to require PhD skills, in practice we needed to provide users with similar job postings to assist them in comparing how the recommended ones can fit better. These two targets lead to the two main modules in our system, namely PhDness **ranking model** and job **similarity system**. Acknowledging the nature that jobs of high requirements are hard to satisfy and likely to be reposted, we also elaborate in building a pruning system to cope with it.

4.1 Ranking Model

The ranking model predicts the PhDness for given job postings (Figure 3). This problem can be

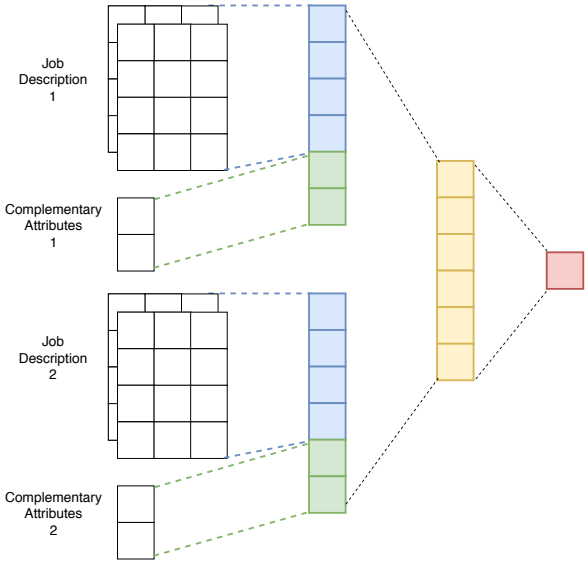

Figure 3: Ranking Model Architecture

treated as a regression task, where each job is evaluated with a numeric PhDness score. Instead, as one of targets in this project is to study what requirements make a job more PhD related, we compose the problem as a ranking task. That is, given any two job postings, the model will learn to judge the one with higher PhDness.

The backbone network for learning the representation of a job posting is modified from the FastText model (Joulin et al., 2017). We incorporate the following input features:

The job description provides key knowledge to PhDness, and we process it the same as the FastText model. Given the input description text of W words, $w = [w_1, w_2, \ldots, w_W]$, a weight matrix $\mathbf{A}$ is a lookup table over the whole vocabulary that maps the individual word tokens to their latent space representation (see Mikolov et al. (2013)). During the seeding phase, we use a pretrained word embedding fine-tuned from Glove (Pennington et al., 2014) to populate $\mathbf{A}$ and then do not change it during the training of the ranking model.[1]

The word representations are then averaged to form the text-level representation:

$$E^w(x_i) = \frac{\mathbf{A}x_i}{W} \tag{1}$$

where x_i refers to the job posting x_i, $i \in \{1, 2, \ldots, I\}$.

[1]The job description can contain many rare or domain-specific words. Even though we have a very large volume of job postings as the training corpus, it may still be insufficient to learn the representation of those words.

We also add the bag of N-grams as additional features to capture some single word-group level knowledge, as this aspect might be blurred in the text-level representation. First, let the bag of N-grams be $N_k = \{w_k, w_{k+1}, \ldots, w_{k+N-1}\}$. Then, similar to word representations (1), the N-gram representations are also averaged:

$$E^{N_k} = \frac{\sum \mathbf{A}\{N_k\}}{N}.$$

The job posting comes with other complementary job attributes, denoted as $a_1, a_2, \ldots, a_A$, also providing useful knowledge to the job's PhDness (e.g., MinimumSalary or Employer), although they are not always presented as the job description. The one-hot encoding f_e is applied and the transformed attributes are normalized and appended to the text representation to form the final input feature, using the concatenation function f_c:

$$E(x) = f_c(E^w, \{E^{N_k}\}, f_e(a_1, a_2, \ldots, a_A)). \quad (2)$$

Now with the input features, the ranking task is defined similarly to the ordinal regression setting (Joachims, 2006). Given any two input job postings x_i and x_j that are comparable, $i \neq j$, i and $j \in \{1, 2, \ldots, I\}$, $y_i \neq y_j$, the target value (i.e., PhDness ranking) is $y_i = \text{sign}(y_i - y_j)$. We apply the hinge loss here as our focus is to learn the comparative rank, and the model then is to minimize, using the final input feature (2):

$$\ell = \sum_{i,j,y_i \neq y_j} \max(0, 1 - y_i f_2(E(x_i) - E(x_j) + b))$$

where b is the bias term and f_2 is a two-layer neural network. The second layer has a linear activation function as the sign is for the hinge loss to learn.

We evaluate the ranking model by using the **K-fold** cross validation ($K = 5$ specifically) on the human annotated training data. Two evaluation measures are used to justify the performance from different perspectives: First, the normalized discounted cumulative gain (NDCG) at t (Järvelin and Kekäläinen, 2002) is a widely used measurement for ranking quality in information retrieval measures the usefulness of the top t rated items. We adopt it with t to be 15% of the total number of testing examples. Second, the normalized Kendall's τ distance (Kendall, 1938) is calculated to measure the overall ranking quality by looking at the number of discordant pairs.

Our fine-tuned model is able to achieve $\mathbf{0.89}$ for the NDCG at t score and $\mathbf{0.13}$ for the normalized Kendall's τ distance, showing evidence that the model can both find the most PhD intensive job postings overall and also perform well enough in comparing the PhDness among any randomly chosen pair of job postings.

As for the inference stage, first the model is used to predict a grid of comparative scores for all pairs of candidate job postings. The final prediction of the PhDness score for a given posting is the average of its relative scores against the other postings.

4.2 Similarity System

In addition to the PhDness score prediction from the ranking model, the platform also recognizes similar job postings to help users to perform comparisons. Analogously to the ranking model discussed above, we also incorporate both the text features and complementary attribute features here. However, an unsupervised approach is adopted due to the following considerations: first, the similarity system should not be bounded by the annotation set and thus generalize easily to all job postings; and second, the speed.

Specifically, the term frequency $\times$ inverse document frequency (TF$\times$IDF) features are extracted from the job description text and other textual attributes (e.g., Employer). Numerical attributes (e.g., Salary) are categorized and attached to the feature list with a small normalization factor. The final similarity scoring is calculated using the Euclidean distance.

4.3 Pruning System

The preliminary research reveals a problem that some jobs are re-posted for a few times during the period until being fulfilled. One of the key modules in PostAc® is its pruning system that removes those duplicated postings. The module adopts a heuristic approach to avoid laborious annotation by hand.

For any two job postings from the same employer published within 4 to 16 weeks, a duplication score is calculated for checking. Here we first have the difference in the publish date d as one input. A similarity score s is also evaluated based on the aforementioned similarity system. Similarly to

Xia et al. (2010), the duplication score d' is defined as

$$d' = \frac{s}{\alpha d}.$$

The publish date difference in the denominator (with the normalizer α) acts as a factor that penalizes when two job postings are too far away from each other. Later, the postings whose duplication score is larger than a given threshold are filtered out.

5 Implementation

The PostAc® platform is implemented as a web tool, with the back-end natural language processing systems responsible for ranking, similarity measurement, and pruning built on Python. The front-end website for storing and managing data and users is built using the PHP programming language. This separation enables the interface to be usable from lightweight environment and also support large amount of users. At the back-end side, two major optimizations are applied:

Ranking Model: Although by the nature of a pairwise ranking model the prediction takes $O(I^2)$ time complexity, it is worth noting that the prediction of each individual posting is performed independently of other postings. The fine-tuned ranking model is serialized and replicated for a few copies. The platform now runs a few prediction process in parallel and this can also easily scale up to future extensions.

Similarity System: The output TF×IDF feature matrix can still have a fair number of dimensions even with a pruned vocabulary. Finding the nearest neighbors in this big set can be time-consuming. We saved the extracted feature matrix on a KD-tree data structure (Bentley, 1975) and this is progressively maintained as new data comes into the system. Once the KD-tree is up to date, the nearest neighbor search is performed right after with the results being saved. This incurs a reasonably large cost up front but once made available, it greatly reduces the processing time for the front-end service. The separation of two components makes it possible to perform the process at the back-end side and push the results to the front-end, without interfering it in most of the time.

With regard to the front-end side, as aforementioned, the PostAc® platform contains two major usage ows: Exploration View for PhD graduates to look at individual job advertisements and Analytics View for policy makers and supervisors who are interested in demand for graduates in different industry sectors.

5.1 Exploration View

Individual users seeking PhD-shaped jobs can access the exploration view flow via a search box. The users can enter a few keywords related to the fields of research they are interested in, or words that relate to their existing skill set. The system ranks all jobs based on the combined score from both the keyword matching and the PhDness score predicted by the ranking model.

The search results (Figure 1) are displayed as a list of job titles that can be further refined by adding more filters. A user can click on a job posting to navigate to the page with more detailed information. In this page, complementary job attributes are provided along with top-ranked similar jobs from the similarity system. Users' navigation and click-through behaviors are recorded as feedback to complement the seeding training data and fine-tune the ranking system in the future.

5.2 Analytics View

One of the main aims of the PostAc® platform is to improve PhD graduate awareness of the demand for their research skills. Aggregated data can also help universities and policy makers to target policy interventions and education efforts appropriately. The Analytics View is designed to help users by showing a range of visualisations of the data set characteristics as well as the key factors our ranking system has been able to discover from the data.

The **Time Series Graph** visualizes the seasonal changes in the demand for jobs in various industries, reflecting the industry and market level changes, month by month, over a year.

The **Distribution Graph** (Figure 4) visualizes how the PhD-shaped jobs are distributed among different areas. Users from government agencies can use these graphs to support regional policy making.

The **Skill Set Graph** visualizes the commonly requested skill sets and abilities for jobs requiring PhD level skills.

6 Conclusion

In this paper, we have presented PostAc®, an evolving platform that makes high level research jobs in the Australian economy more visible. The platform is based on evaluating a human experts'

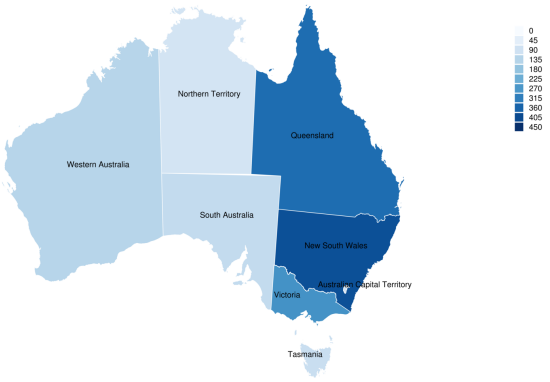

Figure 4: Distribution of PhD Jobs in the Healthcare Sector by a Region: urban Australian areas tend to call for more PhD graduates for healthcare jobs.

hand-annotated data set and its results give empirical evidence of the underlying ranking model being scalable and effective. Currently, the online platform enables a visual-interactive search, exploration, and visualisation of the findings from our machine learning model. This platform will help boost the awareness of the value of PhD level skills and better match PhD graduates to great jobs outside academia.

Acknowledgments

This research partnership of the ANU and Data61/CSIRO was supported by the ANU Discovery Translation Fund 2.0, Australian Government Department of Industry, Innovation and Science (DIIS), On Prime program, and data provided by Burning Glass International Inc. We gratefully acknowledge the funding from the Data61/CSIRO for the first author's PhD studies.

References

Pat Bazeley, Lynn Kemp, Kate Stevens, Christine Asmar, Carol Grbich, Herb Marsh, and Ragbir Bhathal. 1996. *Waiting in the Wings: A Study of Early Career Academic Researchers in Australia*. Canberra, Australia: Australian Government Publishing Service.

Jon Louis Bentley. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.

Edgar Dale. 1930. The training of Ph.D.'s. *The Journal of Higher Education*, 1(4):198–202.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 217–226. New York, NY, USA: Association for Computing Machinery.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Valencia, Spain: Association for Computational Linguistics.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

John McGagh, Helene Marsh, Mark Western, Peter Thomas, Andrew Hastings, Milla Mihailova, and Matt Wenham. 2016. *Securing Australias Future: Review of Australia's Research Training System*. Melbourne, Australia: Australian Council of Learned Academies.

Inger Mewburn, Will J. Grant, Hanna Suominen, and Stephanie Kizimchuk. 2018. A machine learning analysis of the non-academic employment opportunities for Ph.D. graduates in Australia. *Higher Education Policy*, pages 1–15.

Inger Mewburn, William Grant, Hanna Suominen, and Stephanie Kizimchuk. 2016. What do non academic employers want? A critical examination of 'PhD shaped' job advertisements for doctoral employability. *Society for Research into Higher Education (SRHE) Annual International Conference 2016*, pages 1–3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems — Volume 2*, NIPS'13, pages 3111–3119. Lake Tahoe, NV, USA: Curran Associates Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Chaolun Xia, Xiaohong Jiang, Sen Liu, Zhaobo Luo, and Zhang Yu. 2010. Dynamic item-based recommendation algorithm with time decay. In *2010 Sixth International Conference on Natural Computation*, volume 1, pages 242–247. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers.