A Large-scale Industrial and Professional Occupation Dataset

Junhua Liu[1], Yung Chuen Ng[2], and Kwan Hui Lim[1]

[1] Singapore University of Technology and Design
junhua_liu@mymail.sutd.edu.sg, kwanhui_lim@sutd.edu.sg
[2] National University of Singapore
e0201912@u.nus.edu

Abstract. There has been growing interest in utilizing occupational data mining and analysis. In today's job market, occupational data mining and analysis is growing in importance as it enables companies to predict employee turnover, model career trajectories, screen through resumes and perform other human resource tasks. A key requirement to facilitate these tasks is the need for an occupation-related dataset. However, most research use proprietary datasets or do not make their dataset publicly available, thus impeding development in this area. To solve this issue, we present the Industrial and Professional Occupation Dataset (IPOD), which comprises 192k job titles belonging to 56k LinkedIn users. In addition to making IPOD publicly available, we also: (i) manually annotate each job title with its associated level of seniority, domain of work and location; and (ii) provide embedding for job titles and discuss various use cases. This dataset is publicly available at https://github.com/junhua/ipod.

Keywords: Occupational Data Mining · Social Networks · Professional Networking Sites · LinkedIn

1 Introduction

Occupational data mining and analysis is a popular research topic in recent years. There are many lines of research within occupational data mining and analysis, including predicting employee turnover [10,11], modelling and predicting career trajectories [6,7], predicting employee behaviors [3,2] and various others. A common requirement among these works is the need for a occupation-related dataset, which could be derived from professional networking sites (e.g., LinkedIn), scraped from online resumes or other sources. However, most of these datasets are not publicly available, thus impeding future research in this area. To address this problem, we curate and make publicly available the Industrial and Professional Occupation Dataset (IPOD), which comprises 192,295 job titles/positions belonging to 56,648 users on LinkedIn. To the best of our knowledge, IPOD is the largest publicly available occupation-related dataset. This dataset will be useful for researchers and industry practitioners who are interested in occupational data mining and analysis.

Fig. 1: Wordcloud of the job titles in IPOD.

1.1   Related Works

There has been numerous works in recent years that utilize related datasets. We performed a comprehensive literature review of papers since 2008 and identified 15 related works utilizing such datasets.[3] However, out of these 15 works, only one [4] made their dataset publicly available. This dataset comprises the names of affiliations of physics scientists without their job titles, whereas our dataset comprises the job titles across the broader industry.

2   Dataset Description

The IPOD dataset comprises a total of 192,295 job titles/positions that were crawled from the LinkedIn profiles of 56,648 users. 56.7% and 43.3% of these profiles were from the United States and Asia, respectively. Figure 1 shows a wordcloud of the job titles in our dataset.

   Structure of Dataset. The job titles represent the various positions held by a person, particularly the level of seniority, domain of work and location. As shown in Table 1, we label each job title with the following tags:

 – Responsibility (RES). This tag indicates the level of responsibility associated with a job. This RES tag can be further divided into managerial level, operational role and seniority. For example, "Senior Director" corresponds to seniority and managerial level, respectively, while "Junior Technician" corresponds to seniority and operational role, respectively.

---

[3] Due to space constraints, we are unable to list all 15 papers but details are provided in [5].

Responsibility (RES)	Managerial level: *manager, director, president, etc*
	Operational role: *technician, engineer, accountant, etc*
	Seniority: *junior, senior, chief, etc*
Function (FUN)	Departments: *marketing, operations, finance, etc*
	Scope: *enterprise, national, international, etc*
	Content: *security, education, r&d, etc*
Location (LOC)	Regions: *Asia, Europe, SEA, etc*
	Countries/States/Cities: *Singapore, California, Melbourne, etc*

Table 1: Examples of tags associated with job titles.

- Function (FUN). This tag indicates the typical business functions in organizations. Similarly, the FUN tag can be further divided into the department, scope of work and content.
- Location (LOC). This tag indicates the geographic locality that the job title is responsible for, which could be at the scope of a region (e.g., Europe) or a smaller area like a city, state or country (e.g. Singapore).
- Others (O). This tag is for any other tokens which do not fall into the earlier three categories.

Annotation Process. Our labelling is performed by three annotators who are highly experienced with such job titles and the tagging task, namely a Human Resource personnel, senior recruiter and business owner. From our corpus of job titles, we extracted 1,500 tokens of the most frequently occurring uni-grams which are labelled by the three annotators. The Inter-Rater Reliability scores based on two inter-annotator agreements show a score of 0.853 for Percentage Agreement [9] and 0.778 for Cohen's Kappa [1], which represents a *Strong* level of agreement among annotators. We also observe that 77.9% of labelled tags are agreed by all three annotators, 22.1% are between two annotators, while there are no cases where all three annotators disagree on a label.

3 Use Cases

In this section, we briefly describe various possible use cases of our IPOD dataset. For a detailed write-up of the algorithms in these use cases, we refer readers to [5] that is currently under review at the Applied Data Science track of ECML-PKDD 2020.

3.1 Embedding for Job Titles

One use case of our IPOD dataset is to generate embedding for the job titles, which will enable us to perform various occupational data mining tasks. For this purpose, we also develop and release an embedding for job titles, *Title2vec*, which we generate using a deep bidirectional language model (biLM) that is fine-tuned from pre-trained ELMo embeddings on a large text corpus [8]. *Title2vec* is useful for numerous tasks, such as understanding similar job titles across different companies, or as the input to career trajectory prediction problems, job turnover prediction problem and other similar tasks.

3.2   Occupational Named Entity Recognition

Another use case of IPOD is for the Occupational Named Entity Recognition (NER) task. Traditional NER tasks uses general tags such as **PER**son, **ORG**anization, etc, whereas in our occupational NER task, we have more specialized and domain-specific tags such as **RES**ponsibility, **FUN**ction and **LOC**ation as previously described in Section 2. With the ever-changing industry landscape and cultural difference between international workplaces, this occupational NER tasks allow us to better understand the profile of emerging job titles and identify similar job positions across different countries.

4   Conclusion

In this paper, we present the IPOD dataset for occupational data mining and analysis tasks, comprising the job titles, manually annotated tags and a *Title2vec* embedding for the job titles. This dataset comprises 192k job titles belonging to 56k LinkedIn users. To the best of our knowledge, IPOD is the largest publicly available dataset that contains occupational information about the general industry.

5   Acknowledgement

References

1. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics **34**(4), 555–596 (2008)
2. Cetintas, S., Rogati, M., Si, L., Fang, Y.: Identifying similar people in professional social networks with discriminative probabilistic models. In: Proc. of SIGIR. pp. 1209–1210 (2011)
3. Chen, Z.: Mining individual behavior pattern based on significant locations and spatial trajectories. In: Proc. of PerCom Workshops. pp. 540–541 (2012)
4. James, C., Pappalardo, L., Sîrbu, A., Simini, F.: Prediction of next career moves from scientific profiles. arXiv:1802.04830 (2018)
5. Liu, J., Guo, C., Ng, Y.C., Wood, K.L., Lim, K.H.: Ipod: Corpus of 190,000 industrial occupations. arXiv preprint arXiv:1910.10495 (2019)
6. Liu, Y., Zhang, L., Nie, L., Yan, Y., Rosenblum, D.S.: Fortune teller: predicting your career path. In: Proc. of AAAI (2016)
7. Mimno, D., McCallum, A.: Modeling career path trajectories (2008)
8. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv:1802.05365 (2018)
9. Viera, A.J., Garrett, J.M., et al.: Understanding interobserver agreement: the kappa statistic. Fam med **37**(5), 360–363 (2005)
10. Yang, Y., Zhan, D.C., Jiang, Y.: Which one will be next? an analysis of talent demission (2018)
11. Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X.: Employee turnover prediction with machine learning: A reliable approach. In: Proc. of IntelliSys. pp. 737–758 (2018)