Literature	Source	Size	Avail.
IPOD (this paper)	**Linkedin**	**190K**	**Yes**
Mimno et al., 2008 [14]	Resumes	54K	No
Lou et al., 2010 [13]	Linkedin	67K	No
Paparrizos et al., 2011 [15]	Web	5M	No
Zhang et al., 2014 [26]	Job site	7K	No
Liu et al., 2016 [12]	Social network	30K	No
Li et al., 2017 [10]	Linkedin	-	No
Li et al., 2017 [9]	High tech co.	-	No
Yang et al., 2017 [24]	Resumes	823K	No
Zhu et al., 2018 [28]	Job portals	2M	No
James et al., 2018 [8]	APS	60K	Yes
Yang et al., 2018 [25]	Var. channels	-	No
Xu et al., 2018 [23]	Pro. networks	20M	No
Qin et al., 2018 [17]	High tech co.	1M	No
Lim et al., 2018 [5]	Linkedin	10K	No
Shen et al., 2018 [20]	High tech co.	14K	No

Table 1: A survey of datasets used for related works. No available datasets can be found publicly except a dataset of publications and authors from American Physics Society (APS) [8] that only describes the names and affiliations of physics scientists without titles.

*Also with Forth AI.

†Also with Singapore Uni. of Technology and Design.

IPOD: A Large-scale Industrial and Professional Occupation Dataset

Junhua Liu*
junhua_liu@mymail.sutd.edu.sg
Singapore Uni. of Technology and Design

Yung Chuen Ng
National University of Singapore
e0201912@u.nus.edu

Kristin L. Wood†
kristin.wood@ucdenver.edu
University of Colorado Denver

Kwan Hui Lim
kwanhui_lim@sutd.edu.sg
Singapore Uni. of Technology and Design

ABSTRACT

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. As such, there has been growing interest in utilizing occupational data mining and analysis, and 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 475,073 job titles belonging to 192,295 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.

CSCW '20 Companion, October 17–21, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-8059-1/20/10.
https://doi.org/10.1145/3406865.3418329

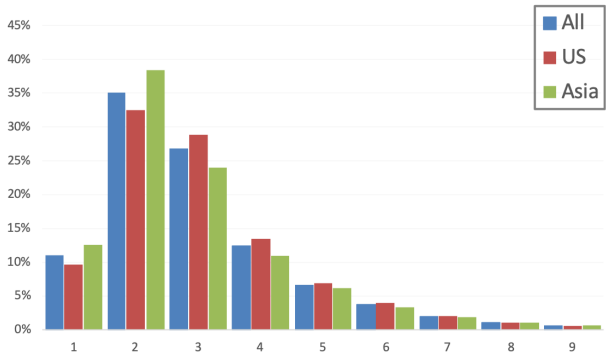

Figure 1: Histogram of occupation entries. The x-axis shows the number of words in the occupation title and the y-axis shows the usage frequency in percentage.

	All	US	Asia
min	1	1	1
max	21	17	21
avg	3.0	3.1	2.9
med	3	3	2

Table 2: Statistics of entries

NE	Count
RES	310570
FUN	255974
LOC	9998
O	66948

Table 3: NE counts

ACM Reference Format:
Junhua Liu, Yung Chuen Ng, Kristin L. Wood, and Kwan Hui Lim. 2020. IPOD: A Large-scale Industrial and Professional Occupation Dataset. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion), October 17–21, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3406865.3418329

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 [25, 27], modelling and predicting career trajectories [12, 14], predicting employee behaviours [3, 4] and various others. A common requirement among these works is the need for an 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 475,073 job titles/positions belonging to 192,295 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.

Related Works

There has been numerous works in recent years that utilize datasets related to occupational data mining and analysis. We performed a literature review of papers since 2008 and identified 15 related works utilizing such datasets. Table 1 shows our survey of 15 related works that utilizes similar types of dataset, of which only one is publicly available [8] (apart from our proposed dataset). The dataset in [8] comprises the names of affiliations of physics scientists without their job titles, whereas our dataset comprises the job titles across the broader industry.

Existing corpora for Named Entity Recognition (NER) tasks [2, 7, 19, 22] typically use general tags such as **LOC**ation, **PER**son, **ORG**anization, **MISC**ellaneous, etc. On the contrary, IPOD provides domain-specific NE tags to denote the properties of occupations, such as ***RES****ponsibility*, ***FUN****ction* and ***LOC****ation*. All named entities are tagged using a gazetteer created by three experts, which reports high inter-rater reliability, achieving 0.853 on Percentage Agreement [21] and 0.778 on Cohen's Kappa [1], with no instances where all three annotators disagree. The labels are further processed by adding prefix using BIOES tagging scheme [18], i.e., **B**egin, **I**nside, **E**nding, **S**ingle, and **O** indicating that a token belongs to no chunk, indicating the positional features of each token in a title.

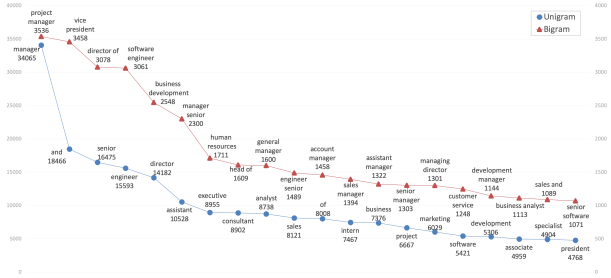

Figure 2: N-grams analysis of the occupation title entries, where the y-axis shows the usage frequency (left axis for unigrams, right axis for bigrams).

Figure 3: Wordcloud of the job titles in IPOD.

IPOD DATASET DESCRIPTION

The IPOD dataset comprises a total of 475,073 job titles/positions that were crawled from the Linkedin profiles of 192,295 users (as of Jun 2020), where 56.7% and 43.3% of these profiles were from the United States and Asia, respectively. Subsequently, the raw data underwent a series of processing, including converting to lowercase, substituting meaningful punctuation to words (i.e. changing & to *and*) and removing special symbols. Figure 3 shows a wordcloud of the job titles in our dataset.

Dataset Exploratory Analysis. The statistics and histogram of the length of job titles can be found in Tables 2/3 and Figure 1 respectively. The corpus comprises of 475,073 English occupation entries from 192,295 unique profiles. These profiles are mainly from United States (56.7%) and Asia (43.3%). Most of the titles fall within five words, contributing to 91.7% of the entries, as shown in both Table 2 and Figure 1). The median statistics and the histogram also suggest that job titles written by Asian professionals tend to be shorter, i.e., within two words, than that by US professionals. Figure 2 shows the distribution of top 20 Unigrams and Bigrams [6] of IPOD, where *manager* is the most popular unigram with 34,065 entries, and *project manager* (3,536 entries) and *vice president* (3,458 entries) are the two most popular bigrams.

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 4, 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.
- 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 is responsible for, which could be for 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 [21] 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.

RESponsibility	Managerial level: - *manager, director, president, etc* Operational role: - *technician, engineer, accountant, etc* Seniority: - *junior, senior, chief, etc*
FUNction	Departments: - *marketing, operations, finance, etc* Scope: - *enterprise, national, international, etc* Content: - *security, education, r&d, etc*
LOCation	Regions: - *Asia, Europe, SEA, etc* Countries/States/Cities: - *USA, Texas, Perth, etc*

Table 4: Examples of tags associated with job titles.

USE CASES

We briefly describe various possible use cases of our IPOD dataset. For a more detailed write-up of the algorithms implemented in these use cases, we refer interested readers to [11].

Embedding for Job Titles. One use case of IPOD is to generate embedding for 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 [16]. *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.

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 . 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.

Occupational Data Mining and Analysis. General tasks in occupational data mining and analysis, such as employee churn prediction [8, 25, 27], professional career trajectory modelling [12, 14] and predicting employee behaviours with various factors [3, 4], typically use one-hot encoding or Bag-of-Words to represent job titles. This form of representation treats each job title as a distinct entity without considering the similarity between them, e.g., "UI developer" and "UI designer" would be modelled as two distinctly different jobs despite their common domain. Our dataset and its associated job title embedding allow us to better model this similarity between jobs via a high-dimensional vector representation, which in turns improves the performance of the earlier mentioned tasks. For occupational analysis, this dataset also allows us to understand basic trends of how jobs are distributed across different responsibility level, function areas and locations.

CONCLUSION

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 475,073 job titles belonging to 192,295 Linkedin users. To the best of our knowledge, IPOD is the largest publicly available dataset that contains occupational information about the general industry.

Acknowledgements. This research is funded in part by the Singapore University of Technology and Design under grants SRG-ISTD-2018-140 and SUTD-UROP-1035.

REFERENCES

[1] Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[2] Łukasz Borchmann, Andrzej Gretkowski, and Filip Gralinski. 2018. Approaching nested named entity recognition with parallel LSTM-CRFs. *Proceedings of the PolEval2018Workshop* (2018), 63.

[3] Suleyman Cetintas, Monica Rogati, Luo Si, and Yi Fang. 2011. Identifying similar people in professional social networks with discriminative probabilistic models. In *Proc. of SIGIR*. 1209–1210.

[4] Zhenyu Chen. 2012. Mining individual behavior pattern based on significant locations and spatial trajectories. In *Proc. of PerCom Workshops*. 540–541.

[5] Meng-Fen CHIANG, Ee-peng LIM, Wang-Chien LEE, Yuan TIAN, and Chih-Chieh HUNG. 2018. Are you on the right track? Learning career tracks for job movement analysis. *Proc. of DSHCM* (2018), 1–16.

[6] Marc Damashek. 1995. Gauging similarity with n-grams: Language-independent categorization of text. *Science* 267, 5199 (1995), 843–848.

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proc. of ACL*. 363–370.

[8] Charlotte James, Luca Pappalardo, Alina Sîrbu, and Filippo Simini. 2018. Prediction of next career moves from scientific profiles. *arXiv:1802.04830* (2018).

[9] Huayu Li, Yong Ge, Hengshu Zhu, Hui Xiong, and Hongke Zhao. 2017. Prospecting the career development of talents: A survival analysis perspective. In *Proc. of KDD*. 917–925.

[10] Liangyue Li, How Jing, Hanghang Tong, Jaewon Yang, Qi He, and Bee-Chung Chen. 2017. Nemo: Next career move prediction with contextual embedding. In *Proc. of WWW Companion*. 505–513.

[11] Junhua Liu, Chu Guo, Yung Chuen Ng, Kristin L Wood, and Kwan Hui Lim. 2019. IPOD: Corpus of 190,000 industrial occupations. *arXiv preprint arXiv:1910.10495* (2019).

[12] Ye Liu, Luming Zhang, Liqiang Nie, Yan Yan, and David S Rosenblum. 2016. Fortune teller: predicting your career path. In *Proc. of AAAI*.

[13] Yu Lou, Ran Ren, and Yiyang Zhao. 2010. *A machine learning approach for future career planning*. Technical Report.

[14] David Mimno and Andrew McCallum. 2008. Modeling career path trajectories.

[15] Ioannis Paparrizos, B Barla Cambazoglu, and Aristides Gionis. 2011. Machine learned job recommendation. In *Proc. of RecSys*. 325–328.

[16] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv:1802.05365* (2018).

[17] Chuan Qin and et al. 2018. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *Proc. of SIGIR*.

[18] Lev Ratinov and Dan Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proc. of CoNLL*. 147–155.

[19] Erik F Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050* (2003).

[20] Dazhong Shen, Hengshu Zhu, Chen Zhu, Tong Xu, Chao Ma, and Hui Xiong. 2018. A joint learning approach to intelligent job interview assessment.. In *IJCAI*. 3542–3548.

[21] Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med* 37, 5 (2005), 360–363.

[22] Ralph Weischedel and et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA* 23 (2013).

[23] Huang Xu, Zhiwen Yu, Bin Guo, Mingfei Teng, and Hui Xiong. 2018. Extracting Job Title Hierarchy from Career Trajectories: A Bayesian Perspective.. In *IJCAI*. 3599–3605.
[24] Shuo Yang, Mohammed Korayem, Khalifeh AlJadda, Trey Grainger, and Sriraam Natarajan. 2017. Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. *Knowledge-Based Systems* 136 (2017), 37–45.
[25] Yang Yang, De-Chuan Zhan, and Yuan Jiang. 2018. Which One Will be Next? An Analysis of Talent Demission. (2018).
[26] Yingya Zhang, Cheng Yang, and Zhixiang Niu. 2014. A research of job recommendation system based on collaborative filtering. In *Proc. of ISCID*, Vol. 1. 533–538.
[27] Yue Zhao, Maciej K Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu. 2018. Employee turnover prediction with machine learning: A reliable approach. In *Proc. of IntelliSys*. 737–758.
[28] Chen Zhu, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. 2018. Person-Job Fit: Adapting the Right Talent for the Right Job with Joint Representation Learning. *ACM TMIS* 9, 3 (2018), 12.