A Project Showcase for Planning Research Work towards Publishable Success

Daheng Wang
University of Notre Dame
Notre Dame, Indiana
dwang8@nd.edu

Meng Jiang
University of Notre Dame
Notre Dame, Indiana
mjiang2@nd.edu

Xueying Wang
University of Notre Dame
Notre Dame, Indiana
xwang41@nd.edu

Tong Zhao
University of Notre Dame
Notre Dame, Indiana
tzhao2@nd.edu

Qingkai Zeng
University of Notre Dame
Notre Dame, Indiana
qzeng@nd.edu

Nitesh V. Chawla
University of Notre Dame
Notre Dame, Indiana
nchawla@nd.edu

ABSTRACT

The goal of the research project is to develop a data-driven approach including a predictive model and a recommender system based on a novel representation learning method to facilitate research planning using public academic datasets.

KEYWORDS

Planning, Behavior modeling, Recommender systems

ACM Reference Format:

Daheng Wang, Meng Jiang, Xueying Wang, Tong Zhao, Qingkai Zeng, and Nitesh V. Chawla. 2018. A Project Showcase for Planning Research Work towards Publishable Success. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

For researchers, especially student researchers, a research work would be considered "a success" if it could be publishable at a good venue, or in other words, be accepted by their targeting international conferences and journals. A research work is actually a product of the interaction between multiple types of contexts such as authors, target conference/journal, datasets, methods, and references. As in [4], we name the components "multi-type contexts" which include operators, goals, resources, spatiotemporal and social dimensions. In order to help researchers achieve publishable success, predictive models and recommendation systems will be very useful and helpful in the early phase of research life-cycle. Specifically, given a public scholarly dataset, we aim at developing two techniques: 1) a predictive model that predicts the probability of a new research on achieving success in terms of paper publication; and, 2) a recommender system that recommends complementary context items (authors, keywords, references, etc.) that will maximize the rate of its publishable success.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

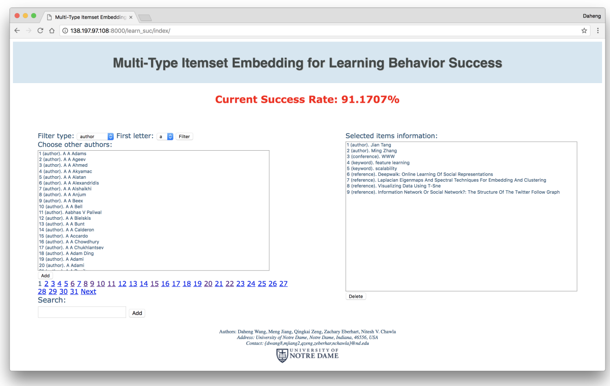

Figure 1: A screenshot of our showcase. Pre-selected items listed on the right-hand side are partially constructed to emulate the composition of the work by Tang et al. [2].

Real world behaviors and events happen in complex environment [3]. Traditional contextual behavior modeling methods suffer from the sparsity of high-dimensional data and the expensive cost of computation. While embedding models proposed recently [1] can be applied on academic networks, they are designed to preserve the characteristics of pair-wise node proximity focusing on the function of clustering and similarity search. This work differs fundamentally from previous work that we focus on a critical property of academic research: "success rate". We formulate a research work as a multi-type itemset structure, i.e., a set of context items of multiple types. We propose a novel *Multi-Type Itemset Embedding* method that efficiently learns representations of the work's context items preserving the itemset structures.

Furthermore, we propose a novel framework, called LearnSuc, that first learns items' low-dimensional representations by optimizing the observational probability of the itemsets, and then feeds the item representations into a logistic regression model to predict the success rate or to recommend complementary items by maximizing the success rate.

As a showcase, we build up a website as an online utility tool based on LearnSuc to show its capability of assisting researchers to establish a better plan or decision-making process towards publishable success. We hope users will enjoy the various functionalities of our demonstration.

We believe our project matches the SIGKDD's primary mission perfectly. Our project will participate as an input of advancement, education, and adoption of knowledge discovery and data mining. It includes merits in basic research and development for KDD researchers, practitioners, and users. A preliminary part of this project has been accepted as a research-track paper with long presentation in KDD 2018 [4].

2 ONLINE SYSTEM

The public website designed for our online demo is located at: http://138.197.97.108:8000/learn_suc/. A screenshot of the current version is shown in Figure 1. Please note that it is still under active construction. More functionalities will be coming soon.

We plan to have 4 small sessions to show interactively with the audience, each for around 3 minutes. In the first session, we will walk through the basic usage of our system, namely, how to add/delete/search items to compose an arbitrary research. We will briefly introduce the system's underlying implementation architecture; and we will talk about the public academic dataset we used in backstage. Second, we plan to show a few motivating examples to demonstrate the efficiency and effectiveness of our system. Then, in the third session, we plan to ask for a few voluntary audiences and show the results about research compositions they are interested in for testing. Finally, we would discuss on future work and listen to suggestions for improvement.

The open GitHub repository of this research project is hosted at: https://github.com/dmsquare/learnsuc. With different practice purposes, there are 3 versions of our Multi-Type Itemset Embedding method inside the repository: basic Python, Tensorflow, and C programming language. The usage and datasets are provided for each of the three. For more information, please refer to the README file in each folder.

3 METHODOLOGY

Planning research work towards publishable success can be seen as planning behavior success in general case [4]. Our LearnSuc framework has two modules for learning behavior success. The first module is a multi-type itemset embedding model that learns item representations from behavior data based on a novel metric of measuring the success rate of a behavior. The second module is to feed the itemset representations into a logistic regression model to predict the probability of a future behavior's success. It can also recommend complementary items to maximize the probability.

In the multi-type itemset embedding model, the representation of each behavior $\vec{b}$ can be computed as the weighted sum of its context item representations $\sum_{c \in b} w_{t(c)} \cdot \vec{c}$. We define b's estimated success rate as $r(b) = \tanh \frac{\|\vec{b}\|_2}{2}$ where $\|\vec{b}\|_2 \in [0, \infty)$ is the Euclidean norm of $\vec{b}$ in the d-dimensional space. Then, to preserve the success properties, we choose to minimize the KL-divergence of empirical and estimated success-rate probability distributions as follows:

$$O = -\sum_{b \in \mathcal{B}} \hat{r}(b) \log r(b). \tag{1}$$

We adopt the asynchronous stochastic gradient algorithm (ASGD) for optimizing Eqn. (1). In each step, the ASGD algorithm samples one behavior b, and the gradient w.r.t. the embedding vector $\vec{c}$ of a context item c in b will be calculated as:

$$\frac{\partial O}{\partial \vec{c}} = \frac{\hat{r}(b)}{\sinh \|\vec{b}\|_2} \cdot \frac{\partial \|\vec{b}\|_2}{\partial \vec{c}} = \frac{w_{t(c)} \hat{r}(b)}{\|\vec{b}\|_2 \sinh \|\vec{b}\|_2} \cdot \vec{b}. \tag{2}$$

Once the representations of context items have been learnt by the multi-type itemset embedding model preserving the behavior success property, we can use those low-dimensional feature vectors for prediction and recommendation tasks. Specifically, we we train a logistic regression model using the itemset's representation vector with its empirical success label and apply the model to predict the success probability of testing instances. For recommending complementary items to a potential behavior/itemset, the goal is to maximize the predicted probability of being successful/observed. We hide one item from each testing itemset and enumerate all itemset candidates and compute their probability scores.

We use a public dataset from the Microsoft Academic project including 10,880 papers in the field of computer science, whose context items contain one conference in the field of data science, at least one author, at least one keyword and at least one reference.

4 CONCLUSION

This work aims at developing an effective and efficient data-driven approach to facilitate research planning. It includes a predictive model and a recommender system based on a novel representation learning method to help researchers determining how plausible a research work would be accepted by its targeting conference. Future work include: 1) taking the cost of context items into consideration seems a promising direction; and, 2) modeling the complementarity between different context items is interesting to explore.

ACKNOWLEDGMENTS

This Research was supported in part by the Army Research Laboratory under the Cooperative Agreement Number W911NF-09-2-0053, and NSF Grants IIS-1447795 and CNS-1622914. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 135–144.

[2] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*. International World Wide Web Conferences Steering Committee, 1067–1077.

[3] Daheng Wang, Meng Jiang, Xueying Wang, Nitesh Chawla, and Paul Brunts. 2018. Multifaceted Event Analysis on Cross-Media Network Data. In *Proceedings of the 1st International Workshop on Heterogeneous Networks Analysis and Mining (HeteroNAM)*.

[4] Daheng Wang, Meng Jiang, Qingkai Zeng, Zachary Eberhart, and Nitesh V Chawla. 2018. Multi-Type Itemset Embedding for Learning Behavior Success. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM.