KDD2024
OAG-Challenge

SUNG KYUN KWAN UNIVERSITY

WhoIsWho-IND-KDD-2024

Final Presentation

2024. 06. 11.

Department of Applied Artificial Intelligence

Minsu Kim

Hanyong Kim

SungJun Park

Index

SUNG KYUN KWAN UNIVERSITY

1. Introduction

2. Analysis

(1) Analysis - 1

(2) Analysis - 2

(3) Analysis - 3

3. Result & Conclusion

1. Introduction

KDD 2024 OAG-Challenge

SUNG KYUN KWAN UNIVERSITY

- **Motivation**
  - Academic data mining has potential to unlock enormous scientific, technological, and educational values.
  - However, academic graph mining has been limited by the lack of a suitable public benchmark.
  - Open Academic Graph Challenge(OAG-Challenge) is open to advance the SOTA in academic graph mining.

- **WhoIsWho-IND(Incorrect Assignment Detection) Task**
  - Given the paper assignments of each author and paper metadata, **the goal is to detect paper assignment errors for each author.**

Our Challenge

< Fig 1. Three tasks of KDD 2024 OAG-Challenge >

Dataset

SUNG KYUN KWAN UNIVERSITY

- **train_author.json & ind_valid_author.json**
  - The key is the Author ID and has the **'name'**.
  - **'normal_data'** for owned papers correctly.
  - **'outliers'** for incorrectly assigned papers.
  - 779 authors(train) and 370 authors(valid)

	Iki037dt	ZihzMro7	WXMYBk3c	WrCODHhe	k3uSCGEE
name	atsushi ochiai	mingwu yang	jianzhao huang	xuebiao yao	shunlin tang
normal_data	[YzOCpPTO, AblgcGjH, B5aouLse, u1G7wBEv, W7w6P...	[C58t0yYu, sWIRnfR3, HJW8h2mo, 0Ptx4O5n, fU4vB...	[lJAIOXO4, fYJcce0K, ZaeOFAcI, kg9xDSXm, T37S3...	[3fYoJb1W, wjt8Y8ho, pPx6o7KZ, xgRarLPn, 9w9yz...	[gTeQer76, mVk2vmmN, TLKSll8D, Eg5NcmZ2, kM5Ip...
outliers	[XL3wd3CP, BTKTiJp2, JxSilFxo, 0jy uH	[qK8llKzD, l0eTdaAG, nFqbDDiB	[HwaUxOes, nvELwvhl, 6Z6SRTOb	[OtmIuFFb, wnP8OmXf, lZ1qVc8S	[xPmu4CGB, buwfccml, fBRzgnef

3 rows × 779 columns

	efQ8FQ1i	C97iQ0Fj	VvKR94tE	UF2hch6j	xb6tyRp8
name	chen dong	xiangquan kong	stephen bonner	heping cao	siddhartha chaudhuri
papers	[cGvhkZHC, MmRHIvd2, agExpryu, eBrOqu4i, WEz1t...	[16grt2XV, ddYPEbL2, 4RUXj0dp, bgq0ZUiD, G7me8...	[z9tBeZpB, LBF8V0R7, ve4cZSnG, 4SZH6NMu, jx9x9...	[4Nf5UZBl, zdlvVxoB, L7JbQRWL, ehQ9lY2R, qd1JH...	[ApK8dQmd, oyAtP5qN, 0O83Pooh, 47g7LjvL, 9o8wv...

2 rows × 370 columns

- **pid_to_info_all.json**
  - Paper ID
  - Author info : name, organization
  - Paper info : venue, publication year
  - Text info : paper title, keywords, abstract

Column	Type	Description	Example
ID	string	Paper ID	53e9ab9eb7602d970354a97e
title	string	Paper title	Data mining: concepts and techniques
authors.name	string	Author's name	Jiawei Han
author.org	string	Author's organization	department of computer science University of Illinois at Urbana Champaign
venue	string	Conference or Journal	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	Publication year	2000
keywords	list of strings	Key words	["data mining", "structured data", "world wide web", "social network", "relational data"]
abstract	string	Abstract of a paper	Our ability to generate…

< Fig 2. train_author/ind_valid_author data (left), pid_to_info_all data (right) >

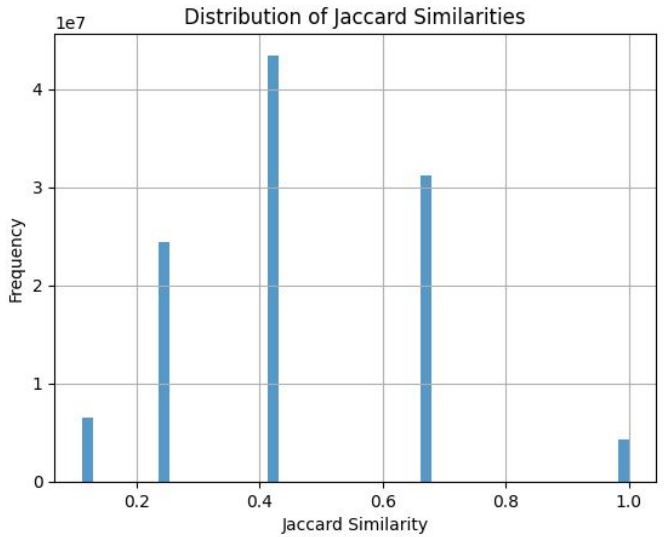
2. Analysis

(1) Analysis – 1 : Graph Learning

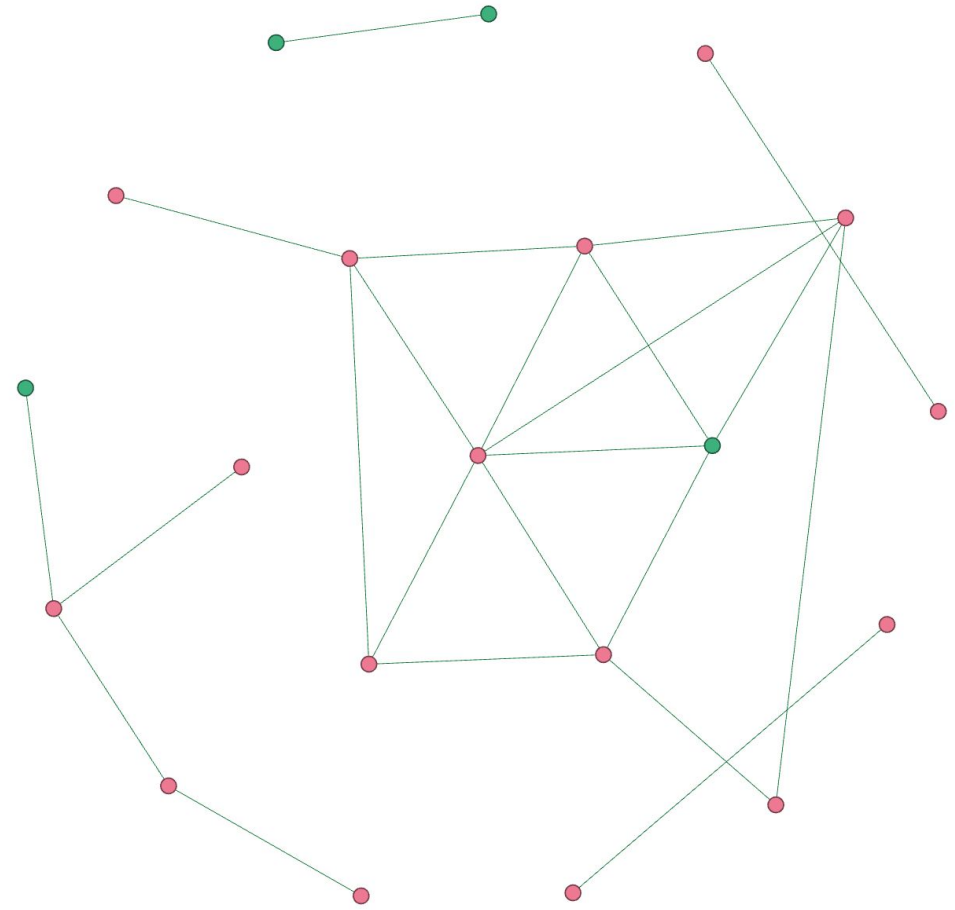
SUNG KYUN KWAN UNIVERSITY

- **Constructed the “Paper-Paper Graph” by author**
  - Stopwords elimination from the title of paper & Embedding
  - Extracted Roberta keywords by using Embedding of title
  - Calculated Jaccard similarities among keywords
  - If Jaccard similarities >= 0.6,

    → Construct the “Paper-Paper Graph” by author
  - Embeddings of title can be used for feature vectors.



< Fig 3. Distribution of Jaccard Similarities among keywords >



< Fig 4. Sample graph of Paper-Paper Graph on KDD Dataset >

(1) Analysis – 1 : Graph Learning

SUNG KYUN KWAN UNIVERSITY

- **Graph Modeling : GCN (Graph Convolution Network)**
  - Inductive Learning (Dataset split → Train Set : Validation Set = 7 : 3)
  - Feature Extraction with two of GCN Conv. Layers from the graph
  - Adam Optimizer & FC Layer for the final output & binary output with Sigmoid activation

- **Hyper-parameters**
  - Hidden : 768
  - Epochs : 50
  - Learning rate : 0.0005
  - Evaluation metric : AUC

Valid AUC	Public Board
0.592	**0.583**

*Ref. Lecture Slide-05_nc, p.5, Interactive Graph Mining, SKKU, 2024 Spring

(2) Analysis – 2 : Machine Learning

SUNG KYUN KWAN UNIVERSITY

- **Data Preprocessing**
  - Stopwords elimination from the title and abstract
  - Text embedding for the title and abstract with RoBERTa
  - Combining embeddings, features(title, abstract, keywords, authors, venue), and year
  - Get ready with Training dataset(148,409) and Validation dataset(62,229)

- **Method**
  - LightGBM learning with stratified K-Fold cross validation
  - Train:Test = 80:20
  - Optimizing hyper-parameters by using grid search
  - Evaluation metrics: ROC-AUC, Accuracy, Precision, Recall, F1-score

Valid AUC	Public Board
0.764	0.638

(3) Analysis – 3 : GCCAD Modeling

SUNG KYUN KWAN UNIVERSITY

- **GCCAD modeling with WhoIsWho-IND Baseline code**
  - **Build Graph**
    - Eliminating stopwords from the title
    - Building Paper-Paper Graph by author → Edge weights with co-author, co-work years and venues
    - Embeddings of title are used for feature vectors.
  - **GCCAD Modeling (Graph Contrastive Learning for Anomaly Detection)**
    - GraphCAD is a complex graph neural network model aimed at Outlier Detection in graph structures.
    - Designed to exploit the characteristics of graph data to detect anomalies at node, edge, and system levels
  - **Experiment**
    - epochs : 40 and default value of baseline code

Valid AUC	Public Board
0.693	0.682

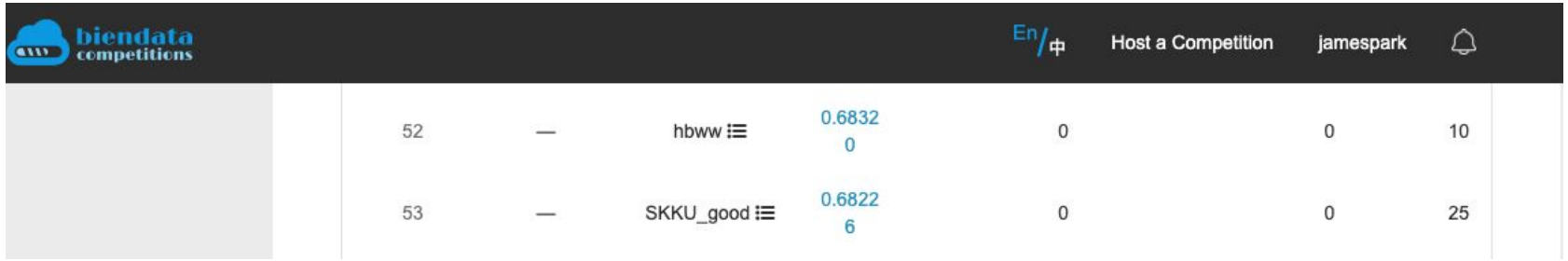
*Ref. Lecture Slide-05_nc, p.5, Interactive Graph Mining, SKKU, 2024 Spring

3. Result & Conclusion

Result & Conclusion

SUNG KYUN KWAN UNIVERSITY

- **Result**
  - Ranked 53rd on the Leaderboard with an accuracy 0.68226 (scored by GCCAD)



< Fig 5. Screenshot of the competition leaderboard (June 9, 2024) >

- **Conclusion & Limitations**
  - We tried both Machine Learning and Graph Learning.
    - ML performs well in general.
    - GNN is huge and heavy. → GNN needs enormous computing power and resources.
  - There was a difficulty on modeling and running codes due to the lack of computing resources. (Oh, C'mon Colab!)
  - Proper topic selection and utilization for GNN are very important.

References

[1] Ravipati, R. D., & Abualkibash, M. (2019). Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets-a review paper. International Journal of Computer Science & Information Technology (IJCSIT) Vol, 11.

[2] Li, Y., Fang, B., Guo, L., & Chen, Y. (2007, March). Network anomaly detection based on TCM-KNN algorithm. In Proceedings of the 2nd ACM symposium on Information, computer and communications security (pp. 13-19).

[3] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152-160.

[4] Chiang, W. L., Liu, X., Si, S., Li, Y., Bengio, S., & Hsieh, C. J. (2019, July). Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 257-266).

[5] Jiang, J., Chen, J., Gu, T., Choo, K. K. R., Liu, C., Yu, M., ... & Mohapatra, P. (2019, November). Anomaly detection with graph convolutional networks for insider threat and fraud detection. In MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM) (pp. 109-114). IEEE.

Thank you!

Q&A