A Deep Learning Approach for Matching Chinese and American Scholars

Jingyao Tang
517030910309

Le Zhou
517030910361

ABSTRACT

Existing efforts to address one-to-one matching such as User Identity Linkage (UIL) problem is mature, but technique to solve one-to-many is still unexplored. Inspired by the recent successes of deep learning in different tasks, especially in automatic feature extraction and representation, we propose a deep neural network based approach for scholars' matching with auxiliary nodes to help with network composition. We provide a datasets with 6,000 scholars, introduce some auxiliary nodes to help to construct scholars' network, samples the networks and learns to encode network nodes into vector representation to capture local and global network structures which, in turn, can be used to align anchor nodes through deep neural networks. A dual learning based paradigm is exploited to learn how to transfer knowledge and update the linkage using the policy gradient method. Experiments conducted on our dataset can shown our approach is effective, but only by human check.

Keywords

User Identity Linkage, one-to-many matching

1. INTRODUCTION

Scholars with equal status and similar academic experience (due to communication etiquette, knowledge level, social status, etc.) are more likely to build good cooperative relationships and produce better academic results. However, state-of-art scholars matching (in other words, scholars recommendation) usually use keywords of research fields or direct relationships between two scholars(such as cooperations).

Recently, Graph neural networks or graph embeddings has attracted wide attention (Battaglia et al. 2018; Cai, Zheng, and Chang 2018). Graph neural networks have been effective at tasks thought to have rich relational structure and can preserve global structure information of a graph in graph embeddings. Inspired by this, we try to view this problem as a specific form of User Identity Linkage, and use graph embeddings to solve it. A typical solution is Deeplink [?], which use graph embedding to solve one-to-one matching problem in social network. Nevertheless, in real word's datasets, we don't have scholar network, let alone anchor point pairs used in Deeplink. Meanwhile, our problem is a one-to-many matching.

In this work, we propose a graph neural network-based approach for scholars matching. To summarize, our contributions are as follows:

- Provide a datasets with more than 6,000 scholars as well as their schools, rank and research fields. Both raw dataset and processed dataset are available.
- Introduce auxiliary nodes beyond the scholar nodes, in order to facilitate the construction of scholars network as well as the network will have more information for training.
- Modify Deeplink's structure to tackle one-to-many problems more accurately.
- Introduce two-way selection algorithm to get the final matching results.

2. PRELIMINARY BACKGROUND

In this section, we introduce the basic terminology in the setting of our proposed approach, and present a few formal definition in that context.

2.1 Problem Definition

Given the *scholars networks* in China and the United States, find *matching scholars* with *comparable academic ability*, so as to explore potential academic cooperation between China and the United States.

Definition 1. *Scholars Networks.* A graph of all scholar nodes in the scope.

Definition 2. *Matching Scholars.* Binary group (x, y), where x and y belong to the scholars' figure A and B which are not contained.

Definition 3. *Comparable Academic Ability.* The scholars' school has similar ranking in the world, are interested in the same area, has won the same (level) awards and etc.

2.2 Process Model Definition

Definition 1. *Construct Scholars Network.* Given a set of $u_1, u_2, \cdots, u_m$ scholars with their rank, school and interested fields, connect scholars to constuct a network. In this work, the scholars networks not only include scholars nodes, but aslo auxiliary nodes such as rank, school and field nodes. These nodes should be connected into the scholar network as well.

Definition 2. *Network Embedding Model.* Given a set of $u_1, u_2, \cdots, u_m$ scholars in network G, NEM learns to represent each u_i with a vector u_i with a vector $\mathbf{v}(u_i) \in \mathbb{R}^d$, where d is the dimensionality of the latent space.

Definition 3. *Preliminary Scholars Matching.* Given any two scholar networks $\mathcal{G}^s$ and $\mathcal{G}^t$, the goal of PSM is to predict that any user u_s chosen from U_t's top 100 best matches scholars u_t.

Definition 4. *Graph Mapping Function.* The function Φ is defined as a mapping from $\mathcal{G}^s$ to $\mathcal{G}^t$, such that for each $u_i \in \mathcal{G}^s$ and its latent space vector $\mathbf{v}(u_i)$, we have $\Phi(\mathbf{v}(u_i)) = \mathbf{v}(u_i'), u_i' \in \mathcal{G}^t$. We also denote the inverse mapping as

$$\Phi^{-1}(\mathbf{v}(u_j)) = \mathbf{v}(u_j'), \text{ where } u_j \in \mathcal{G}^t \text{ and } u_j' \in \mathcal{G}^s$$

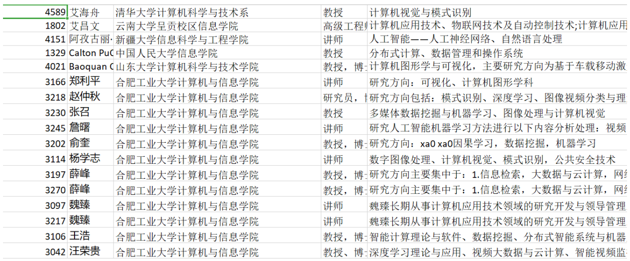
Generally, the mapping function Φ is unknown for a given $\mathcal{G}$ and the objective of our work is to learn a bilateral mapping $(\Phi \text{ and } \Phi^{-1})$ such that the two networks $\mathcal{G}^s$ and $\mathcal{G}^t$ are aligned by maximizing the similarity of all aligned pairs $(\mathbf{v}(u_i), \mathbf{v}(u_j))$.

Definition 5. *Two-way Selection Algorithm.* A and X are valid pairs only if A is in the match of X, and X is in the match of A.

3. THE PROPOSED APPROACH

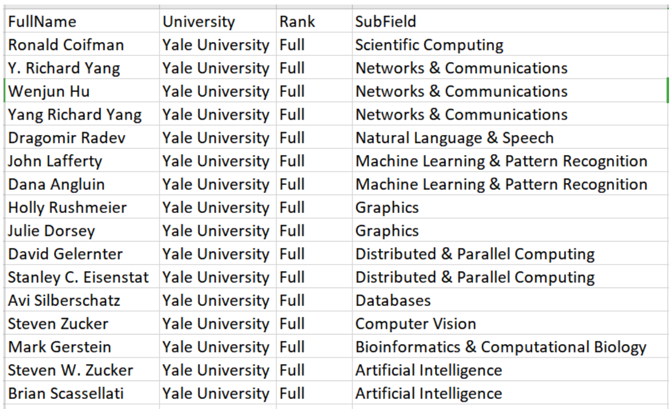
3.1 Data Collection and Processing (Jingyao Tang)

Data Collection. Our Datasets including two sub-datasets: scholars of China and scholars of US. Chinese scholars and their personal details are collected from CCF membership information in Acemap database. Excluding incomplete information, there were 2,734 persons in total. As for US scholars, we use python crawler to collect innformation from https://drafty.cs.brown.edu/ (This website is currently unavailable). Excluding incomplete information, there were 3,449 persons in total.



4589	艾海舟	清华大学计算机科学与技术系	教授	计算机视觉与模式识别
1802	艾昌文	云南大学呈贡校区信息学院	高级工程	计算机应用技术、物联网技术及自动控制技术;计算机应用
4151	阿孜古丽·	新疆大学信息科学与工程学院	讲师	人工智能——人工神经网络、自然语言处理
1329	Calton PuC	中国人民大学信息学院	教授	分布式计算、数据管理和操作系统
4021	Baoquan C	山东大学计算机科学与技术学院	教授，博	计算机图形学与可视化，主要研究方向为基于车载移动激
3166	郑利平	合肥工业大学计算机与信息学院	讲师	研究方向：可视化、计算机图形学科
3218	赵仲秋	合肥工业大学计算机与信息学院	研究员，	研究方向包括：模式识别、深度学习、图像视频分类与理
3230	张召	合肥工业大学计算机与信息学院	教授	多媒体数据挖掘与机器学习、图像处理与计算机视觉
3245	詹曙	合肥工业大学计算机与信息学院	讲师	研究人工智能机器学习方法进行以下内容分析处理：视频
3202	俞奎	合肥工业大学计算机与信息学院	教授，博	研究方向：xa0 xa0因果学习，数据挖掘，机器学习
3114	杨学志	合肥工业大学计算机与信息学院	讲师	数字图像处理、计算机视觉、模式识别，公共安全技术
3197	薛峰	合肥工业大学计算机与信息学院	教授，博	研究方向主要集中于：1.信息检索，大数据与云计算，网
3270	薛峰	合肥工业大学计算机与信息学院	教授，博	研究方向主要集中于：1.信息检索，大数据与云计算，网
3097	魏臻	合肥工业大学计算机与信息学院	讲师	魏臻长期从事计算机应用技术领域的研究开发与领导管理
3217	魏臻	合肥工业大学计算机与信息学院	讲师	魏臻长期从事计算机应用技术领域的研究开发与领导管理
3106	王浩	合肥工业大学计算机与信息学院	教授，博	智能计算理论与软件、数据挖掘、分布式智能系统与机器
3042	汪荣贵	合肥工业大学计算机与信息学院	教授、博	深度学习理论与应用、视频大数据与云计算、智能视频监

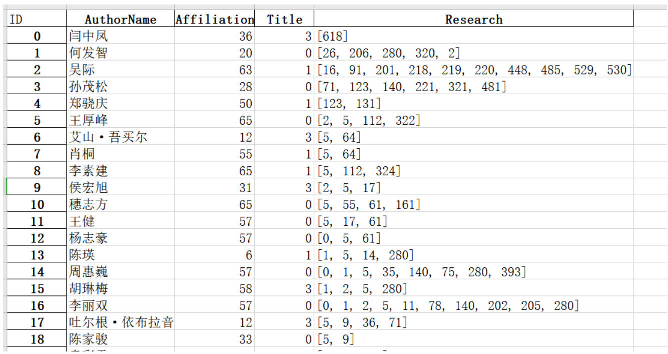
Figure 1: Raw datasets of Chinese scholars



FullName	University	Rank	SubField
Ronald Coifman	Yale University	Full	Scientific Computing
Y. Richard Yang	Yale University	Full	Networks & Communications
Wenjun Hu	Yale University	Full	Networks & Communications
Yang Richard Yang	Yale University	Full	Networks & Communications
Dragomir Radev	Yale University	Full	Natural Language & Speech
John Lafferty	Yale University	Full	Machine Learning & Pattern Recognition
Dana Angluin	Yale University	Full	Machine Learning & Pattern Recognition
Holly Rushmeier	Yale University	Full	Graphics
Julie Dorsey	Yale University	Full	Graphics
David Gelernter	Yale University	Full	Distributed & Parallel Computing
Stanley C. Eisenstat	Yale University	Full	Distributed & Parallel Computing
Avi Silberschatz	Yale University	Full	Databases
Steven Zucker	Yale University	Full	Computer Vision
Mark Gerstein	Yale University	Full	Bioinformatics & Computational Biology
Steven W. Zucker	Yale University	Full	Artificial Intelligence
Brian Scassellati	Yale University	Full	Artificial Intelligence

Figure 2: Raw datasets of American scholars

Data Processing. Raw datasets are full of repeated or invalid information. Moreover, part of the information is expressed in natural language, which is not conducive to the next processing, especially the data of Chinese scholars. For American scholars' field, we split '&' in SubField, each part as a field. For Chinese scholars' field, get rid of the stop words and messy code and split use words with similar meanings as "and", manual error correction is also needed. Since US only has three ranks, we try to align Chinese ranks witn US (Details can be seen in the code). Then, all of this information is converted to ID for later network composition. In the end, there are 67 affilations, 624 fields for Chinese scholars, 90 schools, 38 fields for American scholars.



ID	AuthorName	Affiliation	Title	Research
0	闫中凤	36	3	[618]
1	何发智	20	0	[26, 206, 280, 320, 2]
2	吴际	63	1	[16, 91, 201, 218, 219, 220, 448, 485, 529, 530]
3	孙茂松	28	0	[71, 123, 140, 221, 321, 481]
4	郑骁庆	50	1	[123, 131]
5	王厚峰	65	0	[2, 5, 112, 322]
6	艾山・吾买尔	12	3	[5, 64]
7	肖桐	55	1	[5, 64]
8	李素建	65	1	[5, 112, 324]
9	侯宏旭	31	3	[2, 5, 17]
10	穗志方	65	0	[5, 55, 61, 161]
11	王健	57	0	[5, 17, 61]
12	杨志豪	57	0	[0, 5, 61]
13	陈瑛	6	1	[1, 5, 14, 280]
14	周惠巍	57	0	[0, 1, 5, 35, 140, 75, 280, 393]
15	胡琳梅	58	3	[1, 2, 5, 280]
16	李丽双	57	0	[0, 1, 2, 5, 11, 78, 140, 202, 205, 280]
17	吐尔根・依布拉音	12	3	[5, 9, 36, 71]
18	陈家骏	33	0	[5, 9]

Figure 3: Processing datasets of Chinese scholars

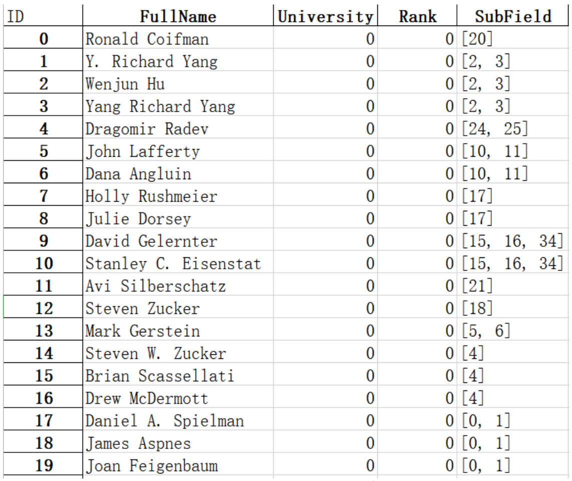
3.2 Build Scholars Network(Jingyao Tang)

3.2.1 Link Scholars in Same Country

Normarlly, in graph construction area, scholars should be linked to each other directly. However, lack of the link information motivates us to introduce auxiliary nodes. Our link rules can be expressed as follows.

- Scholars with top rank directly connect with their schools. Second-rank scholars link to top-rank scholars with same school and same interested fields. If there is no scholars meet the criteria, link them with school. Third-rank scholars are the same, but first try to link with second-ranks, then top-ranks, finally the schools.

- For Chinese schools, split them into C9, 985, 211 and other schools (There is no intersection between these groups). For American schools, use QS ranks to split them into top 50, top 300, top 1000 and others, so



ID	FullName	University	Rank	SubField
0	Ronald Coifman	0	0	[20]
1	Y. Richard Yang	0	0	[2, 3]
2	Wenjun Hu	0	0	[2, 3]
3	Yang Richard Yang	0	0	[2, 3]
4	Dragomir Radev	0	0	[24, 25]
5	John Lafferty	0	0	[10, 11]
6	Dana Angluin	0	0	[10, 11]
7	Holly Rushmeier	0	0	[17]
8	Julie Dorsey	0	0	[17]
9	David Gelernter	0	0	[15, 16, 34]
10	Stanley C. Eisenstat	0	0	[15, 16, 34]
11	Avi Silberschatz	0	0	[21]
12	Steven Zucker	0	0	[18]
13	Mark Gerstein	0	0	[5, 6]
14	Steven W. Zucker	0	0	[4]
15	Brian Scassellati	0	0	[4]
16	Drew McDermott	0	0	[4]
17	Daniel A. Spielman	0	0	[0, 1]
18	James Aspnes	0	0	[0, 1]
19	Joan Feigenbaum	0	0	[0, 1]

Figure 4: Processing datasets of American scholars

as to make those groups has similar size with Chinese groups. Schools in each groups link to each other.

- Link scholars with their interested fields.

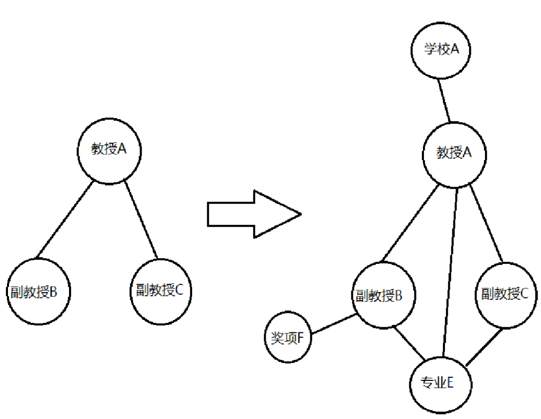


Figure 5: The method of linking edge improvement among scholars in the same country

3.2.2 *Link Scholars in Different Country*

In UIL, datasets usually including anchor points pairs. However, it's hard to decide which two scholars are anchor pairs. Consider this, we only connect auxiliary nodes instead of scholar nodes between two networks. Our link rules can be expressed as follows.

- A school will be linked with the school in other networks that has the most closest QS rank.
- Similar majors will be linked between two graphs.

3.2.3 *Advantages*

- Reduced the number of manually tagged anchor node pairs required.
- The difficulty of anchor node pair construction is reduced effectively.
- Training is faster due to fewer anchor nodes.

3.3 Graph Embedding(Le Zhou)

The central idea of Graph Embedding is to find a mapping function that converts each node in the network into a potential representation of low dimensions. It is convenient for computing and storage, and no need to manually mention features (self-adaptability). In this work, we try to use deep walk, struc2vec and LINE to accomplish graph embedding. Consider training time and effectiveness, we finally choose deep walk as embedding method.

DeepWalk bridges the gap between network embedding and word embedding by treating nodes as words and generating short random walks as sentences. Neural language models such as word2vec can then be applied to these random walks to obtain network embeddings.

The advantage is first that it can generate random walks on demand. As the word2vec model is also optimized for each sample, the combination of the random walk and word2vec makes DeepWalk an online algorithm. Second, DeepWalk is scalable, and the process of generating the random walk and optimizing the word2vec model is highly efficient and trivial parallelism.

3.4 Deep Learning Model(Le Zhou)

After obtaining the embedding vector for each nodes in graph, we turns to learn the mapping functions between any two SNGs based on the anchor nodes by using two CNNs. Given each labeled anchor node pair $(u_i,\ u_j\ )$ and their representation vectors $(v(u_i), v(u_j))$ (i.e. matched auxiliary nodes), we learns mapping $\Phi(v(u_i)$ by minimizing the loss function below:

$$\ell\left(\mathbf{v}\left(u_i\right), \mathbf{v}\left(u_j\right)\right) = \min\left(1 - \cos\left(\Phi\left(\mathbf{v}\left(u_i\right)\right), \mathbf{v}\left(u_j\right)\right)\right)$$

where $\cos(\cdot)$ is the cosine similarity between mapped vector $\Phi\left(\mathbf{v}\left(u_i\right)\right)$ from $\mathcal{G}_s$ and the embedding representation $\mathbf{v}\left(u_j\right)$ in $\mathcal{G}_t$. One of the anchor node in pairs as input and the other as label for each training, and optimized the loss function.

To improve the effect, we replicate the dual learning in Deeplink, which extends the training set to all nodes, extracting non-anchored nodes every few steps to train both networks simultaneously, making the mapping function Φ and Φ^{-1} cyclic consistent. The process of dual learning can be expressed as: one node vector $v(u_i)$ use $\Phi(v(u_i))$ to map into another network, and then use $\Phi^{-1}(\Phi(v(u_i)))$ to map back. The aim is to minimize loss function below:

$$\ell(\mathbf{v}(u_i), \Phi^{-1}(\Phi(\mathbf{v}(u_i)))) = \min(1-\cos(\mathbf{v}(u_i), \Phi^{-1}(\Phi(\mathbf{v}(u_i))))$$

3.5 Two-way Selection Model(Le Zhou)

From deep learning model, for any scholar node, we choose top 100 scholars with the most similar vector using cosine similarity. In two-way selection model, only when both scholars are in each other's matching list, can they be seen as successefully match.

4. EXPERIMENTS

We now describe a real-world datasets that we provided above used in our experiments. Since this area is unexplored so that lack of true labels, we will try to use human check to express the effectiveness of our approach.

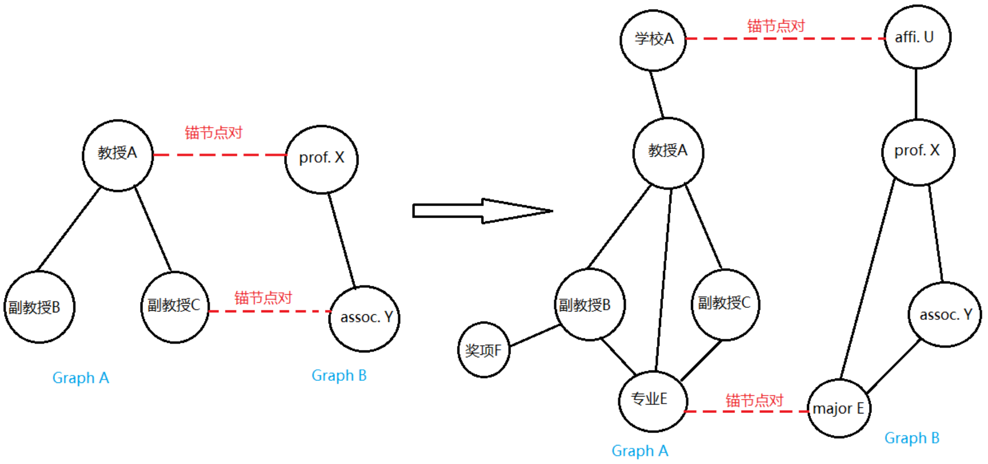
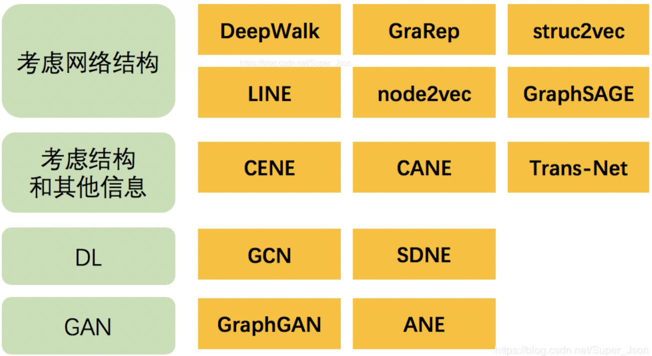


Figure 6: The method of linking edge improvement among scholars in the different country



考虑网络结构	DeepWalk LINE	GraRep node2vec	struc2vec GraphSAGE
考虑结构和其他信息	CENE	CANE	Trans-Net
DL	GCN	SDNE	
GAN	GraphGAN	ANE	

Figure 7: Methods for graph embedding

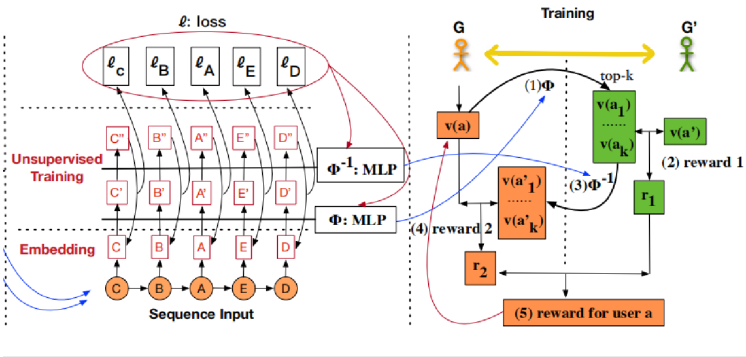
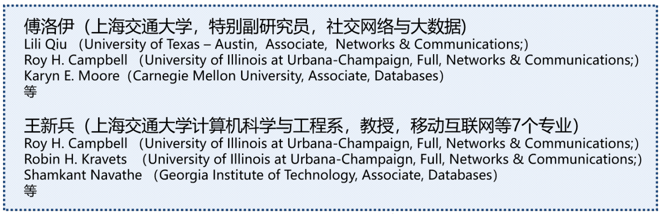


Figure 8: Deep Learning Model



傅洛伊（上海交通大学，特别副研究员，社交网络与大数据)
Lili Qiu (University of Texas – Austin, Associate, Networks & Communications;)
Roy H. Campbell (University of Illinois at Urbana-Champaign, Full, Networks & Communications;)
Karyn E. Moore (Carnegie Mellon University, Associate, Databases)
等

王新兵（上海交通大学计算机科学与工程系，教授，移动互联网等7个专业)
Roy H. Campbell (University of Illinois at Urbana-Champaign, Full, Networks & Communications;)
Robin H. Kravets (University of Illinois at Urbana-Champaign, Full, Networks & Communications;)
Shamkant Navathe (Georgia Institute of Technology, Associate, Databases)
等

Figure 9: Results

4.1 Datasets

In this work, we use a datasets collecting and processing based on above methods. Including auxiliary nodes, there is 3,369 nodes and 80,290 edges in Chinese scholar networks, and 3,556 nodes with 37,630 edges in American scholar networks. 360 anchor points are created for training, which only accounts for about 5% of the total number of nodes.

4.2 Results

To illustatre, we enumerate two examples here (Show as Fig.9). And the full match results will be submitted as individual file.

As Fig.9 shown, for instance, for Lili Qiu, who is a match for Teacher Fu, her school is ranked 65 by QS, similar to Jiao Tong University, which is ranked 60. Her title Associate is the same level as Teacher Fu. Moreover, they are work in similar fields.

5. CONCLUSION & FUTURE WORK

In this work, we provide a scholars network datasets, introduce auxiliary nodes for network construction, proposed a novel deep reinforcement learning based approach for one-to-many matching problem, specifically speaking, for Chinese and American scholars matching problem.

Our approach in networks construction can be used to tackle similar graph embedding problem that lack of manually tagged datasets. Moreover, the whole process can be used to accomplish one-to-many matching or one-to-many recommendation problem.

There are several directions to be investigated in the future.

- For graph embedding, we have tried DeepWalk and struc2vec. However they have similar performance while DeepWalk is trained faster. Some other methods for capturing the network structure can be tried, such as LINE which based on BFS, may learn more details of our network structures. Therefore, some other sample network methods may have better performance.
- For dual learning, mature Cycle GAN structure may

be used to get performance improvement.

- For selection algorithm, other mathcing algorithms such as stable match algorithm can be used.

6. GROUP DIVISION

Limited by structrure of paper, we divide the writing task as very small parts, and only label the parts using name. The student number of us is: Jingyao Tang(517030910309), Le Zhou(517030910361). Parts without labels mean that we write together.

The group division of us are:

- Jingyao Tang:
  - Data Collection and Processing
  - Build Scholars Network
- Le Zhou:
  - Graph Embedding
  - Deep learning model
  - Two-way selection model