Academic Field and Future Influence Prediction for Scholar Profile Construction

Jiaying Tang, Hongying Jin, Daling Wang(✉), Shi Feng, and Yifei Zhang
School of Computer Science and Engineering, Northeastern University, Shenyang, China
tangjiaying1996@foxmail.com, 757232699@qq.com, {wangdaling, fengshi, zhangyifeig}@cse.neu.edu.cn

1 Introduction

Collecting scholar information from massive academic resources to construct scholar profiles can provide a reference for various academic activities. For a scholar profile, except basic attributes such as age, gender, job title, some potential attributes such as academic field and future influence need to be predicted (Fig.1). Towards that, we propose two models to predict them.

Fig. 1. Basic structure of scholar profile

2 Methods

After extracting the basic information of scholars, we predict scholars' academic fields and academic influence in the next decade. The future influence of a scholar is defined as: his all papers citation number within a certain period in academic circles (in the next 10 years).

2.1 Academic field prediction

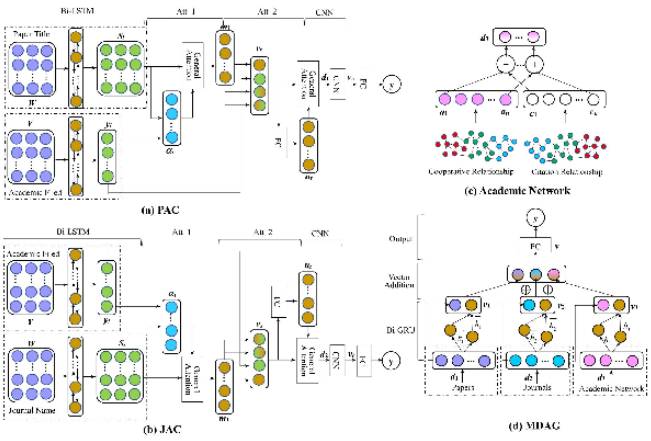

Fig. 2. The model of academic field prediction

We propose a model framework shown in Fig. 2, which connects three different data sources - scholar paper information (PAC), journal information (JAC), and scholar network relationships (Academic Network) as input, and obtains the high-level feature representation through the sub-model (MDAG).

2.2 Future academic influence prediction

We use the PageRank algorithm with time perception to calculate the impact of the paper or the journal, and newly published papers will be given a higher score.

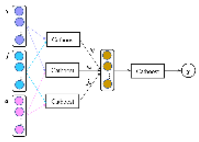

Fig. 3. The framework of CSCF model

Two-level catboost fusion regression model . The first level takes different combinations of three features as input (paper timing features s, journal timing features j, and scholar feature space a) (Fig. 3).

3 Experiment

The dataset for our experiment comes from Aminer.

Table 1. Academic field dataset

Task	Data set	Scholar	Paper	Citations	Journal	Label
Academic field	Raw data	1,715,170	3,074,071	7,038,195	23,640	1,467
	Task data	11,367	432,383	808,195	11,364	790
Academic influence	Raw data	1,715,170	3,074,071	7,038,195	23,640	-
	Task data	101,458	372,348	808,195	18,564	-

The results are as follows.

Table 2. The results of academic field

Method		Recall	Precision	F1
Single data source	LDANE	0.498	0.412	0.468
	PAC	0.503	0.472	0.487
	JAC	0.483	0.458	0.470
Multiple data sources	MDC	0.598	0.596	0.596
	MA	0.601	**0.607**	0.604
	MDAH	0.613	0.605	0.608
	MDAG	**0.622**	0.606	**0.614**

Table 4. The results of academic influence

Method		MAE	RMSE	R^2
One-level model	LR	0.807	0.092	0.881
	XGBoost	0.622	**0.072**	0.848
	Light_GBM	0.992	0.291	0.913
	CatBoost	0.602	0.091	0.874
Two-level model	CSXGBF	0.620	0.081	0.861
	CSLGBF	0.854	0.147	0.753
	SCF	0.623	0.082	0.846
	CSCF	**0.542**	0.077	**0.901**

For academic field, we study whether the sequence order of high-level features of the three data sources before sharing the Bi-GRU layer will affect the experimental results. Table 3 shows the results of six different combinations.

Table 3. The results of different sequence orders

Order	Recall	Precision	F1
$d_1 - d_2 - d_3$	0.616	0.596	0.605
$d_2 - d_1 - d_3$	**0.622**	**0.606**	**0.614**
$d_3 - d_1 - d_2$	0.618	**0.606**	0.612
$d_1 - d_2 - d_3$	0.612	0.604	0.607
$d_2 - d_3 - d_1$	0.610	0.598	0.604
$d_1 - d_3 - d_2$	0.608	0.601	0.606

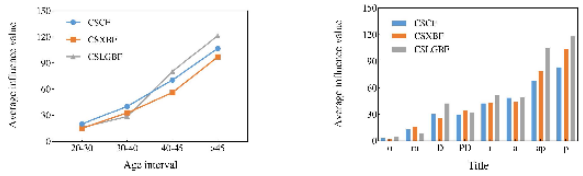

Fig. 4. The effect of age interval or title on the future academic influence

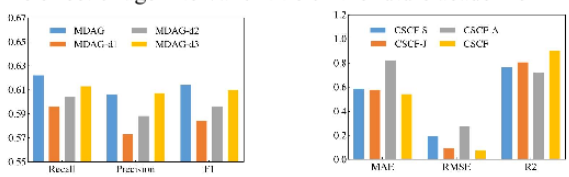

Fig. 5. Evaluation index results of ablation experiments

4 Conclusion

We propose a double-layer attention model of paper information and journal information representation model for predicting scholars academic field based on multiple data sources. We also propose a two-level fusion model based on feature combination for predicting scholars' future academic influence. The experimental results prove that our models are superior to the state-of-the-art methods.