Predicting Paper Acceptance Rank in citation graph

Yuntao Du
East China Normal University
10153903105@stu.ecnu.edu.cn

Abstract

Measuring research impact and having an objective picture of the institutions is essential for students, parents and funding agencies. We tend to apply several machine learning methods to rank research institutions based on predicting the number of accepted papers at upcoming top conferences. In our proposal, We aim at a three-phase experiment, starting with a simple average method and then extend our training dataset by finding the similarity of conferences, engineer trend features and use linear regression, rank SVM and ensemble models to improve our predictions.

1 Introduction

Ranking academic institutions and researchers is not a new topic. Many approaches have been developed to measure the information diffusion, such as author impact factor (AIF), H-Index and so on. Many issues in academic network have been investigated and several systems have been developed, such as DBLP, Google Scholar and Aminer. However, compared to scholar's personal impact, it is much more difficult to measure the institution's academic achievement and research impact. Specifically, given a research area, such as Data Mining, how to rank the future relevance of research institutions?

Microsoft provides a public large and heterogeneous academic graph dataset, called Microsoft Academic Graph(MAG)[Sinha *et al.*, 2015], which contains extensive scientific publication records, citation relationships, and fields of study. We utilize *KDD 2016* version, includes 19,843 institutions, 114,698,044 authors, and 126,909,021 publications from 2000 to 2015.

It is unnecessary and time-consuming to rank every affiliation in every field of study. So we narrow the problem and select 8 top computer science conferences as our target, which are SIGIR, SIGMOD, SIGCOMM, KDD, ICML, MM, MobiCom, and FSE.

The evaluation is performed on the number of paper acceptance of selected conferences in next year. Grand truth ranking is determined by all the **full research papers** accepted in each conference. The simple policy is described as following:

1. Each accepted paper has an equal vote (i.e., they are equally important).
2. Each author has an equal contribution to a paper.
3. If an author has multiple affiliations, each affiliation also contributes equally.

According to the Matthew effect, we think it is reasonable to rank only top 20 institutions in each conferences, so NDCG@20[Järvelin and Kekäläinen, 2002] is used as the evaluation metric:

$$\begin{aligned} DCG_n &= \sum_{i=1}^{n} \frac{rel_i}{\log_2^{i+1}} \\ NDCG_n &= \frac{DCG_n}{IDCG_n} \end{aligned} \tag{1}$$

Where i is the rank of an institution, and rel_i is this' institution's relevance score. For a perfect ranking algorithm, IDCG is equal to DCG producing an NDCG of 1.0. For the calculation of true relevance scores, the number of accepted full research papers from every institution is used. The detail of NDCG can be seen in [1].

2 Background

Previous work by [Zimmermann, 2012] summarized traditional methods on academic ranking problems. Various ranking criteria can be used in these rankings such as citation counts, h-index, and discounted impact factors. Besides, a new learning-based method has been proposed recently by [Liu, 2009] in the construction of ranking models for information retrieval systems, known as learning to rank.

During the WSDM Cup 2016, teams mined the MAG and calculated scores for each paper using the number of citations and reference links inside the graph. Our feature selection approach is inspired by previous work by [Wade *et al.*, 2016].

3 Data exploration

3.1 Datasets

Because the massive academic graph provided, it is essential to look into the dataset and get intuitive feature to set a good baseline. We list several datasets extracted from MAG that we utilize in this paper in Table 1. First, We traverse the whole graph to get all accepted papers of top 8 conferences between 2000 and 2015. Then, we use *Papers, Authors and Affiliations* datasets to join affiliations information for each paper

[1] https://www.kdd.org/kdd-cup/view/kdd-cup-2016/Rules

of above conferences. And we also include all the keywords for papers selected.

Dataset	Attributes
Affiliations	Affiliation ID Affiliation name
Authors	Author ID Author name
ConferenceSeries	Conference series ID Full name
ConferenceInstances	Conference series ID Conference instance ID Full name
Journals	Journal ID Journal name
Papers	Paper ID Paper publish year
PaperAuthorAffiliations	Paper ID Author ID Affiliation ID
PaperKeywords	Paper ID Keyword name ID mapped to keyword
PaperReferences	Paper ID Paper reference ID

Table 1: MAG datasets

3.2 Visualize Trend

The first thing we can think of is to check for any obvious trend for the accepted papers of 8 conferences. So we select *KDD* as a sample to visualize the trend in the last five years. Because the top affiliations hardly changed in several year, and it is unlikely that affiliations which have little impact could be present in the top 20 places in a short time, so we only consider the trend of top 20 affiliations.

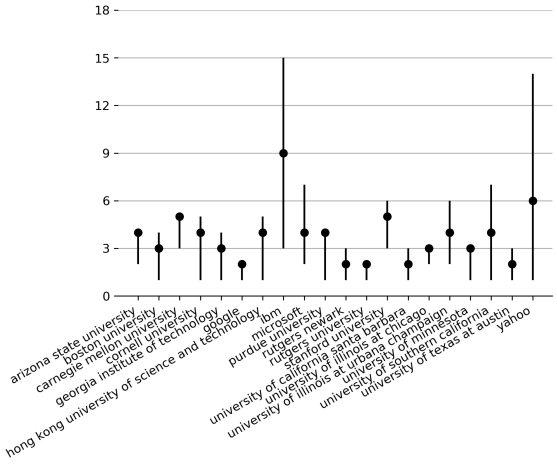

Figure 1: The range and mean value of the number of accepted full papers for top 20 affiliation at KDD between 2011 and 2015

We visualize the accepted full search paper trend at *KDD* between 2011 and 2015 in Figure 1. The vertical line of each affiliation indicates the maximum and minimum number of accepted papers during the last five years. And the mean value of each affiliation is marked as black dot on each line. For most affiliations, the variance is relatively small. So the mean number of accepted full papers is a good predictor to show how an affiliation will perform in near future.

Five-year full accepted papers is a small dataset compared with the large heterogeneous graph. Apart from the full research papers, one conference also has many other types of papers, such as journal papers, workshop papers or posters. So we add more data into practice in phase II. Specifically, we investigate the possible correlation of the number of full research papers with the total number of papers at one conference for the same year. Figure 2 shows the numbers between all papers with full research papers share the similar trend. So it is possible to enlarge dataset with all papers to get better prediction.

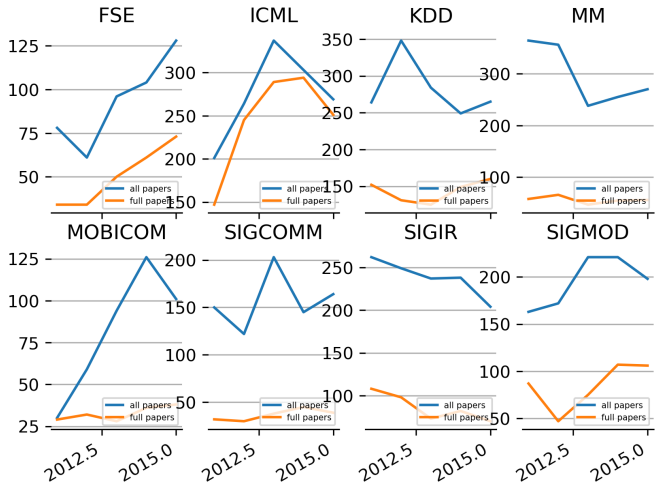

Figure 2: All papers vs full research papers for 8 conferences selected

3.3 Three phases

For better understanding the data, we process three phases to complete this task. Each phase is composed of feature engineering and prediction model selection. As we go further into the progress, more features has been extracted, and more complex models are used.

4 Phase I

In the first stage, our goal is to set up a good baseline for next complex models. Figure 1 shows that it is natural to use the average number of papers accepted in the past five years to predict next year result of the same affiliation. For phase I, we do not intend to use any model, so the affiliations are ranked according to the mean value and calculate NDCG@20 as the result.

5 Phase II

5.1 Prediction Target

In phase I, we use the number of accepted papers per year per affiliation as the predictor, and rank the value to calculate NDGG@20. However, this method is not feasible for complex model output because of it is discrete. And it also misses the fractional contributions of authors to the final affiliation rank. On the other hand, the relevance score, as calculated in the competition *KDD 2016 Cup*[2], distributes the score to each author, thus each involved affiliation gets a fraction of overall score.

Because relevance score is more informative than the number of accepted papers, so we decide to use the relevance scores as the prediction target in all models. Since the evaluation metric NDCG@20 is also calculated using the relevance scores, so it should be good to predict the relevance scores directly.

5.2 Feature Engineering

For a large dataset like this, feature selection is not an easy thing. We create a *Conference-Affiliation-Year* vector for each affiliation per year as the input feature.We classify our feature into three class: time-based features, statistics-based feature, and author-based feature.

Time-based features. The rank problem can be seen as a time-series prediction problem. So it is essential to capture the trend across past years. We utilize weighted moving-average(MA) relevance score to indicate the influence of past. More specifically, different weight is used according to closeness: higher weights for recent years and lower weights for further years. The weights are normalized to 1 and decrease linearly with past years.

Statistics-based features. Basic statistics of past relevance scores cross five year: standard deviation, sum, minimum, maximum, median and mean.

Author-based features. We count the number of first and second authors because the first author is the main contributor and second authors is usually the mentor or correspond author. Those feature can have a big influence of the ranking of an affiliation.

The overall selected feature can be found in Table 2.

Feature	Description
relevance score	weighted moving-average method
statistics feature	standard deviation, sum, minimum maximum, median and mean
#paper	Number of papers published
#author	Number of authors who published at least one paper this year
#1st author	number of first author
#2nd author	number of second authors

Table 2: Features in Conference-Affiliation-Year vectors

[2]https://www.kdd.org/kdd-cup/view/kdd-cup-2016

5.3 Enlarging Dataset

Figure 2 shows that the number of full research papers share the similar trend with the number of all papers, regardless of their types. So we extend our dataset more, using records from all papers that accepted in each conference from 2011 to 2015.

5.4 Models

In the second phase, we experiment with two classes of models: regression model and gradient boosted decision trees. The former is more interpretable and faster to calculate and the latter is more complex and powerful.

Linear regression model. The most popular machine learning method in supervised learning is regression, because of the efficiency and simplicity. Given M dimensional feature vector $(x_1, x_2, ..., x_M)$ as continuous input variables, the goal of regression is to predict continuous target **Y** on it. The simplest regression is linear regression which only contain a linear combination of input variables:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + ... + w_M x_M \tag{2}$$

where **x** = $(x_1, x_2, ..., x_M)$. The common loss function is the sum-of-squares error, which is equivalent to the maximum-likelihood estimation:

$$L(\mathbf{w}; \mathbf{x}) = \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \mathbf{w^T} \phi \left(\mathbf{x_n} \right) \right)^2 \tag{3}$$

We use stochastic gradient descent(SGD) to optimize parameter $\mathbf{w}$.

Gradient boosting decision trees. The gradient boosting decision trees (GBDT) is a powerful model in dealing with a large number of features and non-linear correlations between feature and target. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function[3]. Meanwhile, we use cross-validation to search best parameters.

6 Phase III

Based on the previous work, we develop our method in two ways: extending the dataset by adding records from similar conferences and extending the range of years. Meanwhile, we reveal more important features by using time-series analysis method, such as exponential smoothing.

6.1 Similar conferences Features

In order to expand our collection of features and utilize plenty of other information, we find it is helpful to add the records from most similar and representative conferences to each given conference according to the semantic similarity. Because of the uncertainty of paper submission, it is not enough for using only the targeted conference data to predict. Features extracted from similar conferences can make a more stable measure.

[3]https://en.wikipedia.org/wiki/Gradient_boosting

The main problem is how to measure the similarity. Most researchers publish their work at different conferences, but most researchers specialize in a specific area, so the conferences they prefer must have some similarities. We simply use authors and keywords from the papers to distinguish similar conferences. More specifically, we compute the Jaccard similarity for both authors and keywords for pairs of conferences. From this, we can rank the values to determine which conference is most similar to another in terms of common authors and common keywords.

The similar conferences matrix shows in Table3.

Conference	By authors	By keywords
SIGIR	KDD	ICML
KDD	ICML	SIGIR
SIGMOD	ICML	KDD
SIGCOMM	MobiCom	SIGIR
MM	SIGIR	KDD
MobiCom	SIGCOMM	SIGCOMM
FSE	SIGMOD	SIGMOD

Table 3: Most related conference based on authors and keywords

6.2 Feature Engineering

We keep all the features used in Phase II, and add more sophisticate features to get better prediction in Table 4.

Time-based features. We have used weighted moving-average method to measure the average trend of relevance score in one particular conference. However, drift trend of historical relevance scores captures the increase or decrease of the correlation over time, which is equally important. Meanwhile, the trend ploted in Figure 2 is not stationary, so we try autoregressive integrated moving average(ARIMA) model[Adhikari and Agrawal, 2013] with second order differencing to get the one-step-ahead forecasts.

Statistics-based features. Basic statistics of past relevance scores cross five year: standard deviation, sum, minimum, maximum, median and mean.

Author-based features. Apart from target conference, we use the most similar conference collection to count the authors' information, such as the number of first author in each affiliation. We also try some interesting metrics such as author's impact factor (AIF)[Pan and Fortunato, 2014] in the scientific community to measure the influence of one affiliation in a whole.

Feature	Description
time-series relevance score	ARIMA model
more statistics feature	add from similar conference
AIF	overall author's impact factor

Table 4: New Features in Conference-Affiliation-Year vectors

6.3 Enlarge Dataset More

In the final phase, we enlarge our dataset by prolonging the year range and adding auxiliary collections from similar conference.

Prolonging the year range. In previous work, we only use last five years data, because the full research papers are not explicitly marked in the MAG before 2011. But since we use all papers instead of full research papers as input, this problem is trivial. So we extend the dataset by using papers from 2000 to 2015.

Auxiliary collections. Another efficient way to enlarge dataset is using collections from most similar conference. The intuition is described in selection 6.1.

6.4 Ranking SVM Model

Learning to rank refers to the machine learning approaches of training models in a ranking task. A ranking SVM is a variant of the support vector machine algorithm, which is used to solve certain ranking problems[4].

We use a pairwise approach Ranking SVM[Joachims, 2002] to transforms the learning-to-rank problem into a classification problem – given a pair of items, learning a binary classifier to tell which one should be ranked higher. Then the goal is to minimize average number of inversions in ranking. In Ranking SVM, we train SVM as the classifier.

Suppose the training data is given as $(x_i^{(1)}, x_i^{(2)}, y_i), i = 1, \ldots, m$ where each instance contains two feature vectors $x_i^{(1)}$ and $x_i^{(2)}$, and a label $y_i \in \{+1, -1\}$ indicates which feature factor should be ranked ahead. m is the size of training data.The learning task is to solve a Quadratic Problem:

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \\ &\text{s.t. } y_i(w, x_i^{(1)} - x_i^{(2)}) \geq 1 - \xi_i \\ &\quad \xi_i \geq 0 \\ &\quad i = 1, \ldots, m \end{aligned} \tag{4}$$

It is equivalent to a non-constrained optimization problem:

$$\min_{w} \sum_{i=1}^{m} \max\left(0, 1 - y_i \left\langle w, x_i^{(1)} - x_i^{(2)} \right\rangle\right) + \lambda\|w\|^2 \tag{5}$$

where $\lambda = \frac{1}{2C}$.

7 Experiments

7.1 Data Set

Table 5 shows the dataset size we utilize in every phase. First we only use full research papers for five years, so the data size is rather small. Then we extend our dataset by prolonging the year range and finding similar conference. Meanwhile, we use all types of papers instead of only full research papers as auxiliary collections. In the last phase, our dataset is 80 times larger than original dataset.

[4]https://en.wikipedia.org/wiki/Ranking_SVM

Item	Data size
Phase I: full research papers(five years)	1296
Phase II: full research papers(all years)	8501
Phase II: all papers(five years)	10903
Phase III: all papers(five years) + similar confs	20136
Phase III: all papers(all years) + similar confs	80341

Table 5: Data set used in every phase

7.2 Results of Phase I

It turns out the mean of accepted papers over years is really representative and get good results. We also find using more years data can generate more accurate results. Table 6 shows the NDCG@20 metrics results for three conferences for 2013 to 2015. We find that more data used in next year, more accurate results would be.

Conference	2013	2014	2015
SIGCOMM	0.83	0.85	0.9
SIGMOD	0.87	0.84	0.82
SIGIR	0.9	0.92	0.89

Table 6: Phase I NDCG@20 results for 2013, 2014, 2015

7.3 Results of Phase II

In this phase, we train two models on data from all accepted papers for 2011 to 2015 vs full research papers all range of years. At the same time, three different types of features mentioned in section 5.2 is used. The results shows in figure 3. The detail of data size is described in Section 7.1, considering the range of years as well as all papers. We use the best score of two models as the final results.

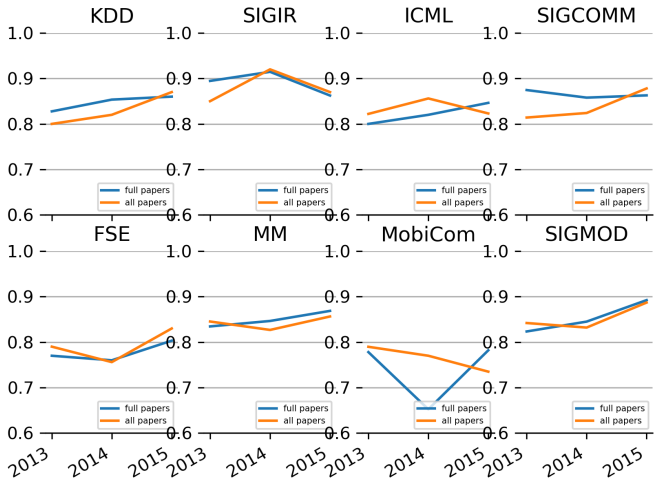

Figure 3: Phase II NDCG@20 results for full papers vs all papers

It is hard to tell which plays an essential role because the results are nearly the same. So we think combine the two types of data can get better results, which would be tested in Phase III.

7.4 Results of Phase III

We summarize the features we utilize in Table 4. Most of the used features in this phase have been experimented in previous phase. As described in Section 6.3, we enlarge dataset by containing all the data in MAG from 2000 to 2015. The final feature size is around 40 after using ARIMA model to get the one-step-ahead forecasts.

Apart from linear regression model and GBDT, we also experiment a ranking SVM model which is commonly used in *learning to rank*[5] problem. Then we use NDCG@20 as the metrics to compare the fitness of models. Figure 4 shows all the models used in the paper, and GBDT outperforms others across all years.

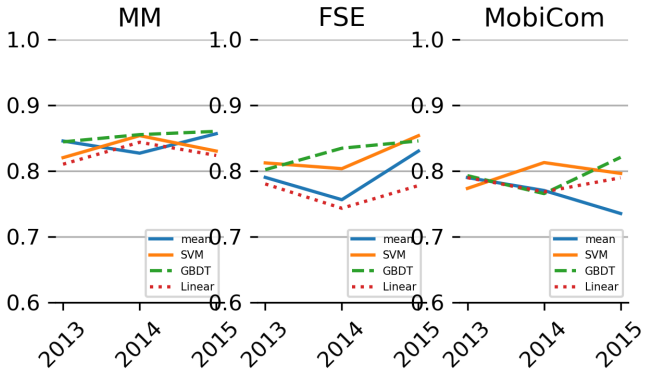

Figure 4: Results for different models

8 Discussion

In this paper, we have proposed and investigated several methods to rank the influence of affiliation at future conferences by predicting their number of accepted full research papers. The data mining process is very interesting, we have found the strong impact of the affiliations' past relevance. Not only the short term influence trends matters, the similar conference trend and the longer term also contribute the prediction of the current relevance of the affiliation.

We believe the feature engineering is the most important part across the process. The idea of three different types of features really helps us a lot. Based on this, we can look through each type of feature and extend our dataset easily. On the other hand, using relevance score as target rather than ranking is also inspired. We think those are the key points that our method could work properly.

Last but not least, there are still many ways that we could improve in the future. From models aspect, we could try some deep Learning networks such as RNN and LSTM, which are more powerful in solving time-series problems. Another approach would be to use all conference of computer science to

[5]https://en.wikipedia.org/wiki/Learning_to_rank

find the most similar conference related to the target conference, rather than just 8 conferences, which should improve the performers but also time-consuming as well.

References

[Adhikari and Agrawal, 2013] Ratnadip Adhikari and R. K. Agrawal. An introductory study on time series modeling and forecasting. *CoRR*, abs/1302.6613, 2013.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.

[Joachims, 2002] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[Liu, 2009] Tie-Yan Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, March 2009.

[Pan and Fortunato, 2014] Raj Kumar Pan and Santo Fortunato. Author impact factor: tracking the dynamics of individual scientific impact. *Scientific Reports*, 4:4880 EP –, 05 2014.

[Sinha *et al.*, 2015] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. May 2015.

[Wade *et al.*, 2016] Alex D. Wade, Kuansan Wang, Yizhou Sun, and Antonio Gulli. Wsdm cup 2016: Entity ranking challenge. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 593–594, New York, NY, USA, 2016. ACM.

[Zimmermann, 2012] Christian Zimmermann. Academic rankings with RePEc. Technical report, 2012.