Heterogenous Graph Mining for Measuring the Impact of Research Institutions

Zeyu Qiu
Institute of Automation,
Chinese Academy of Sciences
& University of Chinese
Academy of Sciences
zeyu.qiu@nlpr.ia.ac.cn

Degiang Kong
Beijing Key Laboratory of
Advanced Information Science
and Network Technology,
Beijing Jiaotong University
15120335@bjtu.edu.cn

Zhenfeng Zhu
Beijing Key Laboratory of
Advanced Information Science
and Network Technology,
Beijing Jiaotong University
zhfzhu@bjtu.edu.cn

Hangting Lu
Institute of Automation,
Chinese Academy of Sciences
& University of Chinese
Academy of Sciences
luhq@nlpr.ia.ac.cn

Jian Cheng
Institute of Automation,
Chinese Academy of Sciences
& University of Chinese
Academy of Sciences
jcheng@nlpr.ia.ac.cn

ABSTRACT

Mining influential nodes in a social network for identifying patterns or maximizing information diffusion has been an active research area with many practical applications. In the research community, influential institutions usually attract denser attention than others. Based on the prediction on how many papers will be accepted by some top conferences held in 2016, the KDD Cup 2016 hosts an international competition for evaluating the importance of academic institutions. This paper describes our solution to the competition. Specifically, the proposed scheme involved in the competition mainly comprises of feature engineering and application of decision tree models. Finally, as claimed by the competition organizer, our approach scored 0.6599, 0.8169, 0.7213 with NDCG@20 in phases 1-3, and resulted in 0.7472 in overall score. With the above scores, our team ranked the first place in phase 2 and fourth place in overall rank.

Keywords
Social network, feature engineering, model selection, decision tree

1. INTRODUCTION

With the explosive growth of information worldwide, it becomes much more difficult to collect favorable information for decision making. In many situations, following authorities can yet be regarded as one of good choices, especially for novice or inexperienced persons. The research of evaluating the authority or impact of a person or an institution has received much attention for its wide potential applications [2, 7, 14, 16, 17, 18].

In the research community, it is common that when starting a study in a new research field, people prefer to follow the work of well-known institutions [1, 10]. Therefore, leading academic institutions would enjoy the extra advantage in disseminating new scientific discoveries and technological breakthroughs through aroused people’s concern. There exist various approaches to evaluate the influence of those authoritative institutions. An intuitive way is to directly reference the ranking of research institutions or universities, which was released annually by many popular newspapers, magazines, and academic institutes. However, for this kind of influence judgment, it might be not very convincing since data and methodology employed for the ranking are mostly unknown to the public.

To encourage the development of some effective objective influence evaluating techniques through publicly available datasets and methodologies, the KDD Cup 2016 launches a competition to predict the number of papers will be published by institutions in upcoming top conferences. The prediction can be further applied for measuring the impact of research institutions. To challenge the competition, we propose a solution by mining heterogeneous information from Microsoft Academic Graph (MAG) and blend several tree-based regression models to predict accepted full research paper number of conferences and rank the most influential institutions. Finally, we get the score of 0.6599, 0.8169, 0.7213 with NDCG@20 in three phases respectively and achieve the fourth place on the overall results leaderboard.

2. PROBLEM DESCRIPTION

In order to measure the influence of institutions, the organizer provides an innovative and interesting task: given any upcoming top conferences such as KDD, SIGIR, and ICML in 2016, rank the importance of institutions based on the prediction of how many of their papers will be accepted by full research track. The result will be evaluated by NDCG@20 when conferences release full research papers.
Table 1: The conference data in KDD Cup 2016

<table>
<thead>
<tr>
<th>conference</th>
<th>No. of accepted papers per year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2011</td>
</tr>
<tr>
<td>Phase 1</td>
<td></td>
</tr>
<tr>
<td>SIGIR</td>
<td>108</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>87</td>
</tr>
<tr>
<td>SIGCOMM</td>
<td>32</td>
</tr>
<tr>
<td>Phase 2</td>
<td></td>
</tr>
<tr>
<td>KDD</td>
<td>152</td>
</tr>
<tr>
<td>ICML</td>
<td>147</td>
</tr>
<tr>
<td>Phase 3</td>
<td></td>
</tr>
<tr>
<td>FSE</td>
<td>34</td>
</tr>
<tr>
<td>MobiCom</td>
<td>29</td>
</tr>
<tr>
<td>MM</td>
<td>58</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of accepted paper numbers in different conferences

Although we are encouraged to collect data by ourselves, the organizer does provide the basic dataset for all teams. Firstly, MAG is provided to draw the picture of research community. The MAG is a large-scale heterogeneous graph containing scientific publication records, citation relationships between publications, as well as authors, institutions, journals, conferences and fields of study.

There are 741 academic institutions and companies in total need to be predicted, which cover most of influential organizations in computer science. Then, in order to avoid the impacts such as participants’ research area and randomness, participants are asked to predict multiple conferences in each phase. And there are totally three phases and eight conferences as shown in Table 1. At the end of each phase, the organizer chooses one of conferences to measure participants submitted results. He also provides the full paper list of eight conferences during the past five years for reference. The details about conferences are shown in the Table 1 and Figure 1. Those conferences have a wide range in number of accepted papers. Furthermore, most of institutions just have a few paper accepted by the conferences. It is a difficult problem to accurately predict the number of accepted papers. To tackle this problem, we propose an effective approach is described in the following sections.

3. FEATURE EXTRACTION

The number of accepted paper of each institution is affected by many factors. The top institutions of previous years are prone to retaining their competitiveness. Companies such as Microsoft, IBM, Yahoo are traditional influence institutions in KDD, therefore it is high probability that their paper accepted by KDD 2016. The location of an institution is also an important factor which influences publication number. There is no doubt that America and China are the top two countries in terms of the number of published papers. There are also many other factors that might affect the results, such as the temporal information, relevant conference influence, topic trends of previous years.

In this competition, we focus on mining a heterogeneous graph which is a subgraph of MAG. As shown in Figure 2, the graph contains the relations of papers, authors, institutions and their own characteristics. We generate a large number of features to describe the graph and filter them based on their performance in eight conferences. The selected features are extracted from three levels. In the remainder of this section, we describe them one after another.

3.1 Paper Feature

The task of competition is to predict how many papers will be accepted in upcoming conferences. There have a naive solution to estimate it based on the history information and there are many models that can be employed to implement it, such as ARMA, RNN, LSTM. However we prefer to use features to describe the history information rather than learn them by models. Then following two kinds of paper features are designed for describing them.

- **Track Record** is a feature describing the publication history of an institution. If we record the published papers of an institution in different conferences, we can find out that the records not only indicate the activity of the institutions, but also show their main research field. An affiliation which have published papers in conferences as PAKDD, WSDM and CIKM will have a higher possibility to submit papers to SIGIR rather than SIGCOMM. Meanwhile, an institution will be more likely to submit papers to MM if it has papers...
Figure 3: NDCG@20 of GBRT with different feature groups

Table 2: Summary of feature

<table>
<thead>
<tr>
<th>feature types</th>
<th>dim.</th>
<th>sigir</th>
<th>sigmod</th>
<th>sigcomm</th>
<th>kdd</th>
<th>icml</th>
<th>mm</th>
<th>fse</th>
<th>mobicomb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Track Record</td>
<td>3849</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Institution Ratio</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Conference Ratio</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Institution Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>institution id</td>
<td>741</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Location Indicate</td>
<td>53</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Champion Indicate</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Paper Rank</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Page Rank</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Author Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active Degree</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Continuity Evaluator</td>
<td>15</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

accepted by MM for five consecutive years. The difficulty of extracting such feature is the choice of time interval and feature representation. We finally set the time interval as three years and use an accumulative histogram vector of all conferences which in the MAG dataset to describe each institution.

- **Major Ratio** is a kind of features that indicate the main component of institution and conference. They are calculated as equation (1), (2).

  \[
  \text{ratio}_i = \frac{\text{paper}_{i,c}}{\sum_{c=1}^{C} \text{paper}_{i,c}} \quad (1)
  \]

  \[
  \text{ratio}_c = \frac{\text{paper}_{i,c}}{\sum_{i=1}^{I} \text{paper}_{i,c}} \quad (2)
  \]

  where \(\text{paper}_{i,c}\) denotes the paper number of institution \(i\) in conference \(c\), \(C = 8\) is the number of the chosen conference. \(\text{ratio}_i\) indicates the importance of a conference for institution. Corresponds to \(\text{ratio}_i\), \(\text{ratio}_c\) shows the authority of an institution in the conference while \(I = 741\) is the number of the 741 selected institutions.

3.2 Institution Feature

Institutions are core parts of the task. There exists much information about institution that would help us to predict the number of accepted papers. For example, large academic and research institutions may submit more papers than smaller ones. The institution may follow its publication history with high probability. Institutions in the same area may influence each other by communications and competitions. Based on MAG, we extract the following features to describe institutions.

- **Property Indication** is a suite of features focused on describing the inherent property of an institution. Firstly, name of affiliations is a brand. We use one-hot vector to represent the affiliation id. The vector has 741 dimensions, where each dimension corresponds to one affiliation. Secondly, we combine Google knowledge graph \(^2\) and manually annotations to encode affiliation locations and build another one-hot vector to describe them.

- **Authority Score** are features representing institutions research strength. We build one feature to indicate the institution whose published paper number is the most. It is an one-dimension feature with the champion set to one and the others zeros. In practice, the champion features are built for three years. In the meantime, rank is also a kind of feature that can suggest the institution ability. For simplicity, we adopt paper number of institutions to rank them, and use the normalized rank as features.

- **Page Rank (PR)** is a high level feature to measure the institution importance. Since using paper number to rank institution is not enough, we also consider the relations between authors, papers and institutions to represent the importance of institutions and use model to extract them from heterogenous graph. One of the major unsupervised methods which can extract graph information is PageRank \(^1\). It offers such paper rank

\(^1\)https://developers.google.com/knowledge-graph/
scores in the MAG dataset, but it is not suitable to representing institutions of this competition because they are calculated from the whole MAG. Inspired by \[15\], we build a heterogeneous graph (Figure 2) limited in one conference. Finally, the score of each affiliation node is chosen as the feature.

As depicted in Figure 2, the structure of the heterogeneous graph includes two subgraphs and they are connected through author nodes. Author and institution nodes are initialized as \[S\] where \(S\) denotes the set of the corresponding nodes. Then, paper nodes are initialized as \[t_n\] where \(t_n\) is the publication year. Since we believe that new publications usually attract more attention, we add time decay to each paper node. The weight of edges between papers and authors are also initialized by another kind of time decay as \(\frac{1}{t_w-t_p}\). The time decay is empirically chosen in experiments. Finally, edges between authors and institutions are directly set to 1.

After initialization, the graph is iterated by updating sub-graph respectively according to equation 3.

\[
P R(v) = (1 - d) \ast P R(v) + \sum_{u \in l i n k(v)} \frac{w(u,v) \ast P R(u)}{\sum_{x \in l i n k(u)} w(u,x)}
\]

where \(\text{link}(v)\) denotes the set of nodes linked to node \(v\), \(w\) is the weight of edges, and \(d\) is a damping parameter in \([0,1]\). \(P R(v)\) is the node score which is calculated via combine its previous score and scores of nodes which around it with damping parameter \(d\).

The left sub-graph is used to obtain authors’ score from citation network. Through the right graph, we use authors’ score to refine affiliation score. Until convergence, the affiliation score is utilized as one of our features.

### 3.3 Author Feature

The research capacity of an institute is always measured by its disciplines, grants and expert prestige, among which experts are the most important. By comparing many experimental results, we design two features about experts for this task.

- **Active Degree** is a feature measuring the institution active degree in one conference according to the number of active authors in the institution. An active author is defined as of author who often publishes papers in the same conference. We judge whether one author is an active author according to one of the following conditions.
  - Published at least 10 papers in the last decade and published at least 2 papers in last 2 years.
  - Published at least 5 papers in recent 3 years.
  - Published at least 4 papers in recent 2 years.
  - Published at least 3 papers in the last year.

- **Continuity Evaluator** is a feature representing how the research field of an institution is insisting on. The consistency of an institution is based on author’s research field alteration. If an author has continuously published papers in the same conference for five consecutive years, there will be high probability that he or she will have papers accepted in this year. We count the number of authors who published in continuous five, four, three, two and one years as features, respectively.

During the competition, we continuously investigated and appended various features to our model. On the other hand, the curse of dimensionality is the critical issue to consider. The performance of various features in the final evaluated three conferences are reported in Figure 3. It is easy to find that different conferences stress different features. SIGIR and KDD tend to depend on institution features while the score of MM is more affected by author features. In the result of KDD, model trained by all feature gets the lowest score. The tendency of mutual suppression may exist between features. Therefore we apply a greedy algorithm to find the best feature combination for each conference. and Table 2 lists all of our features and combination of features in each conference. The Figure 3 shows that greedy feature combination achieves the best performance in all conferences.

### 4. MODEL SELECTION

We regard this competition task as a regression problem of predicting paper number. Regression is one of the classic problems in statistics and there are many approaches that can deal with it. In practice, we use the following approaches to predict the number of accepted papers of each affiliation in one conference one year \([4, 8, 9, 13]\). All models we used are provided by the sklirn-learning package[12].

- **AdaBoost**\([5]\) is one of the ensemble approaches that trains several weak learners and combines the output of each learner by a weighted sum. By assigning each weak learner as a regression tree, we can handle the prediction problem. As shown in Figure 4, Adaboost results in the worst performance for SIGIR.

- **Random Forest (RF)** \([3]\) is a tree based learning method widely used in data mining competitions. The algorithm constructs multiple regression trees using randomly sub-sampled features and samples. RF can get the top performance in the result of the local validation set (Figure 4). In experiment, the training model of an RF changes radically. Its result has large variance. We take this property into consideration when the number of conference paper is less than 70. Since it strongly influences the result.

- **Gradient Boosting Regression Tree (GBRT)** \([6]\) is also a tree-based learning algorithm. Different from RF, GBRT is based on fitting residual and sequentially building the weak regression tree learners. GBRT gets the best performance on all conferences among all the single models in the Figure 4.

With comprehensive investigation of different models, we decide to choose RF and GBRT as the basic model for ensemble. Inspire by Adaboost\([5]\), we blend basic models according to equation 4,

\[
p_{\text{ensemble}} = \sum_{i=1}^{N} \log \left( \frac{\text{val}_{\text{score}_{i}}}{1 - \text{val}_{\text{score}_{i}}} \right) \text{predict}_{i}
\]

\[\text{predict}_{i}\]
where predict\textsubscript{ensemble} denotes final prediction. val\_score is NDCG@20 score in validation set, and predict\_i is the paper number predicted by model i. The results of different ensemble are shown in Figure 4. It is hard to say which is the best. The finally chosen ensemble way is different in every phase, which will be describe in next section.

5. EXPERIMENTS

This competition is different from the tasks in previous years. It has three separated but related phases. Participants can change the strategy according the performance of previous phase. In this section, we discuss our strategies in the three phases.

- **Phase one** is the beginning of the competition. It takes us a lot of time with adaptation of rules and data mining process. We collect some data from The Internet for prediction. In the first two weeks, we focus on crawling information from website like ACM digit library\textsuperscript{3}, DBLP\textsuperscript{4}, the Google knowledge graph. But it is challenging to clean and align data between datasets. Consequently we only use a little of them such as location. The MAG dataset is enough for us to mine information of institution. As shown in Table 1, there are only a small number of samples in each conference.

\textsuperscript{3}http://dl.acm.org/
\textsuperscript{4}http://dblp.uni-trier.de/xml/

We try many ways to increase the number of samples, and finally determine to merge two time interval samples as up sample. After data preprocessing, we train GBRT and RF models for one conference in each time and predict SIGIR, SIGMOD, and SIGCOMM separately. Then, in this phase, we choose ensemble GBRT and RF both to build result by manually tuning the parameters. Finally, we get 0.6599 NDCG@20 score of SIGIR which is chosen by the organizer and rank 109th on the leaderboard. We blame the cause of our low score as that model may overfit in the local data and we realize that the features are not enough.

- **Phase two** is the most important phase in the competition. Both KDD and ICML are the top conferences in machine learning fields and involve many institutions. In this phase, we focus on remedying defects of phase one. Firstly, we concentrate on deep feature engineering and add many useful features, while in the phase one, we only have paper features. Secondly, we find that, with different parameter settings, RF is different from others. We give up ensemble RF. Instead , we train 169 models of GBRT by different parameters (“max\_deep” from 3 to 15, “min\_sample\_leaf” from 3 to 15) and randomly choose one-third of them for ensemble. Model combination is according to equation 4. In this phase, the organizer choose KDD for evaluation, and we get 0.8169 NDCG@20 score and rank
the first place on the leaderboard.

- **Phase three** is the final phase, we continue to remain approaches in phase 2. The different between them is that we picked RF back to ensemble because of its performance in validation set as shown in Figure 4(h). As mentioned above, the prediction between RF models are quite different from each other. However their NDCG@20 score do not show too much variation. As Figure 1 shows, most institutions only publish one or two papers. Specially for the conferences in phase three, almost 90% of the institutions have no more than two papers. It means that different rank of them will have limited influence result. In the evaluation of phase three, we rank 42th and get 0.7213 NDCG@20 score.

According to competition rule, the final score is the weighted ensemble of three phases scores by equation 5.

\[ \text{Score} = 20\% \times \text{Conf}1 + 40\% \times \text{Conf}2 + 40\% \times \text{Conf}3 \]  

where Conf1, Conf2, Conf3 represent the NDCG@20 score in each phase, respectively.

Finally, our approach gets a score of 0.7472 and achieves the fourth rank.

6. CONCLUSIONS

In this paper, we introduce the approaches of team General for KDD Cup 2016 competition. We generate many features to capture factors which may influence the published papers of institutions. We further study institution performance in different conferences and design a unique strategy for each phase. Finally, our solution gets 0.6599, 0.8169, 0.7213 NDCG@20 score in each phase, and achieves 0.7472 final score which ranks the fourth place on the leaderboard. Except regarding this task as a regression problem, we also test some rank methods such as LambdaMART. However their performance is not good enough. We will do more in-depth study and research in the future.

7. ACKNOWLEDGMENT

The authors would like to thank the organizer for hosting this meaningful competition. This work was supported in part by the National Natural Science Foundation of China (No.61332016, No.61532005 and No.61572068), Beijing municipal major science and technology project (16JD0013) and the Program for New Century Excellent Talents in University (No.13-0661).

8. REFERENCES


