Whose papers are accepted the most: Measuring the impact of research institutions

Group Member: Xueguang Lu & Diyi Wang

I. Problem Description

In the research community finding influential nodes is also highly valued by those who have been longing for mechanisms to effectively disseminate new scientific discoveries and technological breakthroughs so as to advance our collective knowledge and elevate our civilization. For students, parents and funding agencies that are planning their academic pursuits or evaluating grant proposals, having an objective picture of the institutions in question is particularly essential. Against this backdrop we have witnessed that releasing a yearly Research Institution or University Ranking has become a tradition for many popular newspapers, magazines and academic institutes. The published rankings not only attract a lot of attention from governments, universities, students and parents, but also create many debates on the scientific correctness behind those rankings. The most criticized aspect of these rankings is: the data used and the methodology employed for the ranking are mostly unknown to the public.[1]

Input:

Our data is derived from Microsoft Academic Graph (MAG), which include but not limited to:

- Computer Science papers published by each school every year
- Field of Study
- Authors
- Keywords
- References

Output:

Ranking prediction of research institutions in terms of their Computer Science academic influences.

II. Algorithm

We adapt Ranking SVM and proposed a new problem-specific Regression Model.

1. Ranking SVM[6]

Ranking SVM was initially used search engine and information retrieval to optimize ranking based on user clicks as feedback. In our model, the feedback becomes the actual ranking of the research institutions in terms of their academic influences (more specifically, the number of papers that are accepted by the top conferences). The original method updates the ranking weights

for each query, so in our case, we update our weights for each conference-institute-year vector. We use Kendall's τ[2][3] as performance measure in ranking.

$$\tau_S(\mathrm{f}) = \frac{1}{n}\sum_{i=1}^{n}\tau(\mathrm{r}_{\mathrm{f}(\mathrm{q}_i)}, \mathrm{r}_i^*).$$

r*i is the actual ranking according to the accepted paper by conference i.
$\mathrm{f}(q_i)$ is a set of binary relations between institutions decomposed from linear ranking.

$$(\mathrm{d}_i, \mathrm{d}_j) \in \mathrm{f}_{\vec{w}}(\mathrm{q}) \Longleftrightarrow \vec{w}\Phi(\mathrm{q}, \mathrm{d}_i) > \vec{w}\Phi(\mathrm{q}, \mathrm{d}_j).$$

$\overline{w}$ is a weight vector that is adjusted by learning.
Φ(q, d) is a mapping onto features that describe the match between conference q and institution d like in the description-oriented retrieval approach of Fuhr et al.[4][5]

2. Correlation-based Regression Model (CRM)

$$S_r = V_i \otimes C_j$$

We calculate a relevant feature S_r which conveys the relation of institution i and conference j. Where Vi is the vector of specific school and Cj is the vector for a particular conference.

$$S_{i,j,t} = (S_r, S_{0t}, S_{1t}, S_{2t}, S_{3t})$$

S_{ijt} is a vector for each institution (i) regarding specific conference (j) at given year (t). S_{ijt} combines relevant feature S_r with some general year-related features $S_{0t} - S_{3t}$ (such as Top-Conference-Acceptance Rate, Annual budget, Citation-Rank (using pageRank), Computer Science Ranking (US News, etc,.), Where the institute is located, etc.,).

We then calculate a prediction value X_{ijt}' , where w is the weight vector adjusted by learning.

$$X_{i,j,t}{}' = S_{i,j} \cdot w$$

X_{ijt} is the number of actual accepted paper for given ijt.
We suppose to implement squared error as our error loss function. Thus, the target is to minimize amount all the records in training set.

$$Loss = min\sum_{i,j,t}(X_{i,j,t}{}' - X_{i,j,t})^2$$

To minimize the loss, we plan to use stochastic gradient descent (SGD) to update the parameters in W, V and C.

III. Results

Expected our team to be in the top 10% portion of all the teams who were participated in the KDD 2016 competition.

IV. Division of Responsibility

Lu is mainly responsible for data processing and feature extraction, and result demonstration.
Wang is mainly responsible for model derivation, achievement of algorithms.
Both people cooperate in coding and debugging.

V. References

1. KDD Cup 2016, https://kddcup2016.azurewebsites.net
2. M. Kendall. Rank Correlation Methods. Hafner, 1955.
3. H. Lieberman. Letizia: An agent that assists Web browsing. In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI ’95), Montreal, Canada, 1995. Morgan Kaufmann.
4. N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. ACM Transactions on Information Systems, 7(3):183–204, 1989.
5. N. Fuhr, S. Hartmann, G. Lustig, M. Schwantner, K. Tzeras, and G. Knorz. Air/x - a rule-based multistage indexing system for large subject fields. In RIAO, pages 606–623, 1991.
6. Joachims, Thorsten. "Optimizing search engines using clickthrough data." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.