An Ensemble of Ranking Strategies for Static Rank Prediction in a Large Heterogeneous Graph

Shu-Kai Chang*, Sui-Tsung Go*, Yueh-Hua Wu†, Yen-Ting Lee*,
Chien-Lin Lai*, Sz-Han Yu*, Chun-Wei Chen*, Huan-Yuan Chen*,
Ming-Feng Tsai‡, Mi-Yen Yeh§, Shou-De Lin*

*Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
†Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan
‡Department of Computer Science, National Chengchi University, Taipei, Taiwan
§Institute of Information Science, Academia Sinica, Taipei, Taiwan

ABSTRACT

Ensemble methods have been wildly used in many data-science competitions for better predictions. This paper proposes a method to aggregate the results of different ranking strategies for the static rank prediction in a large heterogeneous graph, the ranking strategies of which are developed by National Taiwan University, Academia Sinica, and National Chengchi University. The ensemble method achieves a score of 0.761 on the phase 1 public leader board, while achieving 0.641 on the phase 1 private leader board, and obtains the 8th prize of the 2016 WSDM Cup.

Keywords

Learning to Rank; Static Rank; Heterogeneous Graph

1. INTRODUCTION

The WSDM Cup Ranker Challenge is a worldwide data-science competition held by Microsoft in WSDM 2016. The goal of the competition is to assess the importance of scholarly articles by using data from the Microsoft Academic Graph, which is a large heterogeneous graph containing scientific publication records, including citation relationships between publications, authors, institutions, journal and conference venues. To solve the importance ranking problem, traditional methods, such as PageRank [5] and HITS algorithm [4], use only the citation relation as the measure of paper importance, which means that the other abundantly heterogeneous information is ignored in the methods. Therefore, how to effectively leverage the heterogeneous graph among entities has become increasingly important in the context of big data.

This paper proposes an ensemble framework for ranking publication papers in a heterogeneous graph connecting the papers and related entities. In our approach, the studies [6, 7] are referred to construct the heterogeneous graph [2] for storing the complex relationships among the entities. In addition, our ensemble method aggregates the results of the ranking models [1, 3] into a more robust ranking list according to various features from external resources. Furthermore, the proposed method also deals with the scaling problem of ranking scores during the aggregation, and the choice between rank and score for the aggregation.

*{b00902061, d04944007, b02902031, b00902093, r04922007, r04922050, r04922009, sdlin}@ntu.edu.tw
†b02901078@ntu.edu.tw
‡mftsai@cs.nccu.edu.tw
§miyen@iis.sinica.edu.tw

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

DOI: http://dx.doi.org/XXXX.XXXX

2. METHODOLOGY

In order to aggregate different ranking methods, we first need to construct different ranking models. There are three model candidates in our ensemble approach. Below we briefly describe each of the models.

2.1 Weighted PageRank

In [3], we build a PageRank-based model. Since a paper is hardly updated after its publication, a directed citation link should be linked from a new paper to an old one. Consequently, if we simply apply PageRank to the citation graph, the score produced by the algorithm will be time-biased. Therefore, we design a new extension of PageRank to neglect the time bias phenomena. All the papers in the citation graph are given different time-dependent weights before we run PageRank. We also use the weight configurations to integrate some of given paper features such as conference, journal, author and affiliation. The weight of a paper can be represented by the prior knowledge of its importance.

Our weighted PageRank is defined as follows. Let $PR(p)$ be the PageRank score of paper p, which can be iteratively updated using the following formula:

$$PR(p) = (1-d)\frac{W(p)}{\sum_p W(p)} + d \sum_{q \in \mathrm{Pred}(p)} \frac{PR(q)W(p)}{\sum_{r \in \mathrm{Suc}(q)} W(r)} \quad (1)$$

where $W(p)$ is the weight of the paper p, d is the damping

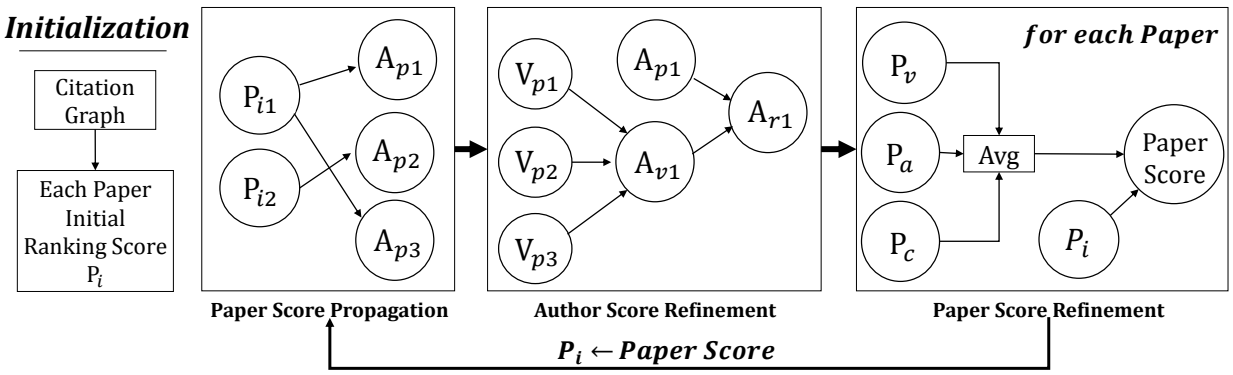

Figure 1: The Score Reinforcement framework: The initial scores of papers are determined by the citation network. In the propagation stage, paper scores are propagated to authors, venues, and the cited papers. In the score refinement stages, the author scores are updated, and then each paper score is refined accordingly. Only the paper scores will be carried on to the next iteration, while the other types of scores are reset, until the paper scores converge.

factor, Pred(p) is the set of predecessors of p, and Suc(q) is the set of successors of q.

For each paper p, we assign each edge of the paper a weight $W_0(p)$, which is the average citation of this paper each year until now and the function can be defined as follows:

$$W_0(p) \equiv \begin{cases} \frac{|\text{Pred}(p)|}{\max_{p \in P} Y(p)+1-Y(p)} & \text{if } |\text{Pred}(p)| > 0 \\ \epsilon & \text{otherwise} \end{cases} \quad (2)$$

where P is the set of all papers, $|\text{Pred}(p)|$ is the in-degree (the number of citations) of p in the citation graph, and $Y(p)$ is the publication year of p. By assigning the weight, we can reduce the time bias since the edges to new papers will be assigned higher weights than others linked older papers.

2.2 Score Reinforcement

We implement the Score Reinforcement framework [1] for ranking publication entities through the heterogeneous information. According to the knowledge, we build several bipartite networks in addition to the citation network G_{PA} which connects a paper to the papers it cites. The bipartite networks are as followed: G_{PA}, a bipartite authorship network that connects a paper to its authors; G_{PCJ}, the bipartite network that connects to paper to the conference/journal it is published on; and G_{ACJ}, a bipartite network that connects authors to the conference/journal where their papers have been published. In particular, we design a scheme using the heterogeneous graph structure to connect those bipartite graphs as a propagation network. Figure 1 shows the basic concept of the framework, which is an iterative procedure through the three components.

2.3 reHITS

In this section, we briefly describe the reHITS method, which is referred to the study [6]. In the method, some modifications are added to reinforce the model to consider some external information. Below We will introduce the modifications and the corresponding modified graph, and we also propose a propagation algorithm for the graph.

2.3.1 The Graph Structure

There are three types of nodes in the reHITS graph: *paper node, author node* and *conference node*. There are also three types of edges in the graph: the *citation edge* is a weighted edge from a paper to another one cited by the paper; the *authorship edge* is an unweighted edge from an author to a paper written by the author; the *conference edge* is an unweighted edge from a conference to a paper that is published on the conference.

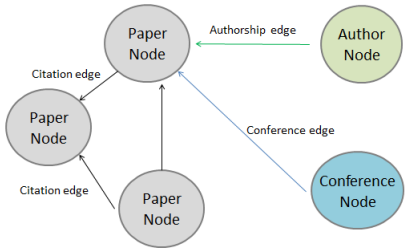

Figure 2: Three node types and three edge types.

Figure 2 is a demonstration of the network structure. However, we can view the graph as a composition of three subgraphs: *citation network, paper-author network* and *paper-conference network*. In the *citation network*, there are only paper nodes and citation edges; in the *paper-author network*, there are paper nodes, author nodes and authorship edges; in the *paper-conference network*, there are paper nodes, conference nodes and conference edges. We can do HITS algorithm on each network, and these three networks will influence each other through paper nodes.

2.3.2 Ranking algorithm

We adopt an reinforcement procedure to update scores of papers, authors and conferences iteratively. The algorithm is conducted by the following steps:

1. Initially, all the authority scores of papers are assigned to a traditional method (such as PageRank or HITS) and then scaled to [0,1] uniformly.
2. Calculate the hub scores of authors by the *paper-author network*

3. Calculate the hub scores of conferences by the *paper-conferences network* and external ranking data.

4. Calculate the hub scores of papers by the *citation network*

5. Update the authority scores of papers, using four types of information, i.e., the authority score it has, the score transferred from authors, the score transferred from conferences, and the score transferred from hub papers. Afterward, the scores are scaled to [0,1] uniformly.

6. Perform Step 5 iteratively until convergence is encountered.

2.3.3 Calculating hub scores

In the above Step 2-4, the hub scores are calculated. We use the following function to compute the hub scores of *Authors, Conferences* and *Papers* in three subgraphs independently:

$$H(N_i) = \theta * \frac{\sum_{N_j \in Neighbor_+(N_i)} S(N_j) * W(E_{N_i,N_j})}{|Neighbor_+(N_i)|} + (1-\theta) * R(N_i), \quad (3)$$

where $H(N_i)$ is the hub score of node N_i, $S(N_j)$ is the authority score of node N_j, $Neighbor_+(N_i)$ is the collection of nodes pointed by the node N_i, and $|Neighbor_+(N_i)|$ is the size of $Neighbor_+(N_i)$. Notes that in *paper-conferences network* and *paper-author network*, there is no edge starting from paper nodes, so the hub scores of papers in these two graphs are 0.

In addition, $W(E_{N_i,N_j})$ is the weight of the edge E_{N_i,N_j}. In citation graph, we set the weights to be what we used in the weighted PageRank [3]. In other two graphs, the weights are set to 1 for each edge of the graph. Furthermore, $R(N_i)$ is the external ranking score (from Google Scholar, Microsoft Academic Graph, etc.) of node N_i. Similarly. We mix the external ranking and the output of our score function by a hyper parameter θ, where $0 < \theta < 1$.The score of external ranking data is map to [0,1] first.

Unlike the tradition HITS algorithm, which uses sum of authority scores, we use the average of authority scores. This adjustment enables us to appropriately estimate the paper importance since there is no node been overwhelmingly influenced.

2.3.4 Calculating authority scores

After we update the hub scores of the nodes. The new authority score of a paper can be calculated as follows:

$$\begin{aligned} S(P_i) = {} & \alpha * S(P_i) \\ & + \beta * Author(P_i) \\ & + \gamma * Conference(P_i) \\ & + \delta * Citation(P_i) \\ & + (1-\alpha-\beta-\gamma-\delta)/N_p \end{aligned} \quad (4)$$

where α, β, γ and δ are hyper parameters which range in [0,1] and $\alpha + \beta + \gamma + \delta \leq 1$.

- $Author(P_i)$ is the authority score of paper P_i propagated from the corresponding authors.

$$Author(P_i) = \sum_{A_j \in Neighbor_-(P_i)} H(A_j) \quad (5)$$

The computation is done at *paper-author network*. When it is done, all the scores will divide their maximum to guarantee each of them ≤ 1. If $Neighbor_-(P_i) = \emptyset$, we will assign mean value of others to it to make the scores smooth.

- $Conference(P_i)$ is the authority score of paper P_i propagated from the corresponding conferences.

$$Conference(P_i) = \sum_{A_j \in Neighbor_-(P_i)} H(C_j) \quad (6)$$

The computation is done at *paper-conference network*. When it is done, all the scores will be divided by their maximum to guarantee each of them ≤ 1. If $Neighbor_-(P_i) = \emptyset$, we will assign mean value of others to it to make the scores smooth.

- $Citation(P_i)$ is the authority score of paper P_i propagated from the corresponding Conference.

$$Citation(P_i) = \sum_{A_j \in Neighbor_-(P_i)} H(P_j) \quad (7)$$

The computation is done at *citation network*. When it is done, all the scores will be divided by their maximum to guarantee each of them ≤ 1. If $Neighbor_-(P_i) = \emptyset$, we will assign it the mean value of the others to make the scores smooth.

- $(1-\alpha-\beta-\gamma-\delta)/N_p$ denotes the probability of random jump, where N_p is the number of papers in the network.

2.4 Ensemble Methodology

Below is the flowchart of our ensemble approach:

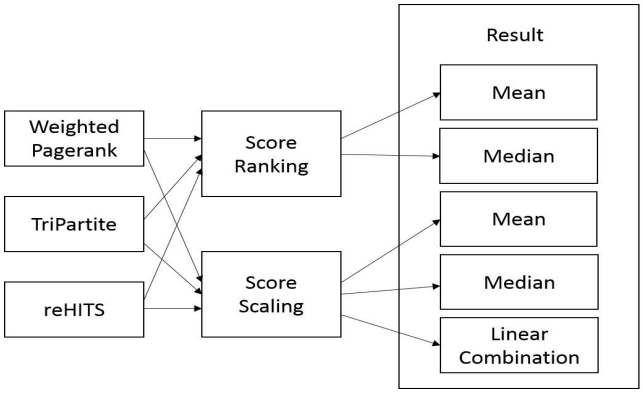

Figure 3: Flowchart of the ensemble method.

Before using the scores of the each mentioned models directly, we have to scale their prediction values into a specific range so as to avoid some model dominating over others. Hence, here we simply scale the prediction scores into [1e-4, 1e-5], much smaller than 1, to avoid that the final score of a paper exceeds the score upper bound when we try some strategies like linear combination. After scaling, we use several statistical methods to aggregate the scores of these three models. We try mean, median and linear combination for aggregation. The following formula is the aggregation we used to obtain the final ranking list from the three methods:

$$
\begin{aligned}
E(M_i) = \alpha * M(Hits) \\
+ \beta * M(Weighted\ Pagerank) \\
+ \gamma * M(ScoreReinforcement) \\
subject\ to\ \alpha + \beta + \gamma = 1.
\end{aligned}
\tag{8}
$$

Moreover, we also use the ranks of a paper instead of its score for aggregation. We calculate the rank by sorting the origin paper score and get its rank in the three mentioned model. After aggregating the rank, we inverse the aggregated ranks and scale them into [0, 1] for submission.

3. EXPERIMENT

3.1 Dataset

In our experiments, the dataset is provided by the WSDM challenge. Since the citation information gives us a simple and useful estimation of how important a paper is, we only focus on the smallest subgraph which contains all the citation relationship. Table 1 lists the statistics of the dataset.

Table 1: Data Statistics

Item	Amounts
Papers	50,011,348
Authors	123,017,488
Conference series	1,275
Citations	757,462,733

3.2 Baselines

In this section, we show the experimental results of the three ensemble candidates. The following scores are from the phase 1 public leader board. The score is in [0, 1], and a higher score indicates a better result. The results of three models are shown in Table 2.

Table 2: Ranking Baseline

Method	Score
Weighted PageRank	0.747
Score Reinforcement	0.741
reHITS	0.763

3.3 Ensemble experiment result

Table 3 lists the results of different ensemble methods according to various aggregation criteria. After getting the results, the submissions score of using rank are a little lower than score with any aggregation methods. That is probably because the aggregation by ranks will lose too much information about distance between each other. On top of that, public scores are mostly similar for mean, median and linear combination of scores. Eventually, we decide to submit the linear aggregation one due to the assumption that it is robust for aggregating the three models.

For the linear aggregation, the scores of the three models are almost the same, and just lower than the reHITS method. However, in order to avoid the potential overfitting problem and give more importance to model that has higher submission score, we set the aggregation parameters (α, β, γ) as (0.4, 0.3, 0.3) as our final submission.

Table 3: submission scores

Method	Score
Ranking by mean	0.737
Ranking by median	0.755
Mean	0.761
Median	0.759
LB	0.761

4. DISCUSSION

We achieve a relatively acceptable score on public data. However, the score on the WSDM Cup hidden data is not as good as what we expected. Since the weighted PageRank and Score Reinforcement methods are doing well on hidden data, we think the reason might be the usage of external resources and meticulous tuning in the reHITs model, which therefore causes our ensemble a serious overfitting problem.

5. CONCLUSION

This paper proposes an ensemble ranking method for static rank prediction in a scientific heterogeneous graph. In addition to introducing the weighted PageRank, Score Reinforcement, and reHITS methods, we also present a simple rank aggregation method to combine the results of the three models. Furthermore, we have learned a lot from the challenge, including how to deal with the big data, how to sort needed information out from the noisy data, how to find out reasons behind phenomenons, and how to find out useful features for solving a rank prediction problem.

6. REFERENCES

[1] M.-H. Feng, K.-H. Chan, H.-Y. Chen, M.-F. Tsai, M.-Y. Yeh, and S.-D. Lin. An efficient solution to reinforce paper ranking using author/venue/citation information - the winner's solution for wsdm cup 2016. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, WSDM '16, 2016.

[2] J. Han. Mining heterogeneous information networks: The next frontier. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, 2012.

[3] C.-C. Hsu, K.-H. Chan, M.-H. Feng, Y.-H. Wu, H.-Y. Chen, S.-H. Yu, C.-W. Chen, M.-F. Tsai, M.-Y. Yeh, and S.-D. Lin. Time-aware weighted pagerank for paper ranking in academic graphs. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, WSDM '16, 2016.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, Sept. 1999.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[6] Y. Wang, Y. Tong, and M. Zeng. Ranking scientific articles by exploiting citations, authors, journals, and time information. In *Proceedings of the 2013 AAAI Conference on Artificial Intelligence*, 2013.

[7] E. Yan, Y. Ding, and C. R. Sugimoto. P-rank: An indicator measuring prestige in heterogeneous scholarly networks. *J. Am. Soc. Inf. Sci. Technol.*, 62(3):467–477, Mar. 2011.