Supervised Kemeny Rank Aggregation for Influence Prediction in Networks

Karthik Subbian and Prem Melville
IBM TJ Watson Research Center, Yorktown Heights, NY, 10598.
{ksubbian,pmelvil}@us.ibm.com

Identifying influential individuals in a network is commonly addressed through various socio-metrics like PageRank, Hub and Authority scores [1], etc. These metrics are primarily based on the actor's location in the network [2] and often captures only a subset of the critical factors that are usually at play while predicting influence in networks like, relationship of the network (type of edge), degree of relationship (weight of the edge), etc. As each measure captures some aspect of a user's influence in the network, it may be beneficial to combine them in order to more accurately identify influencers. In this paper, we focus on methods for combining such influence measures.

One straightforward way to combine socio-metrics is to aggregate the scores given by each metric to produce an aggregate score for an individual, using methods like Logistic Regression. However, given that the individual influence measures produce an ordering of elements and not just a point-wise score, we can instead leverage approaches of rank aggregation. Methods for rank aggregation have been extensively used in Social Choice Theory, where there is no ground-truth ranking, and as such are unsupervised. In this paper we introduce several supervised approaches to rank aggregation that can be trained on the ground-truth ordering of a subset of elements. We empirically illustrate various advantages of supervised rank aggregation methods through two case studies: (1) social network data from Twitter and (2) citation network data from arXiv.org.

Supervised Rank Aggregation: We consider two key rank aggregation techniques, Borda [3] and Kemeny [4]. Borda aggregation is easy to compute but does not satisfy an important goodness property called Extended Condorcet Criteria [4]. Kemeny satisfies this criterion but is NP-hard to compute [4]. So, we use Local Kemenization (LK) [4], which is a relaxation of Kemeny aggregation that still satisfies the Extended Condorcet Criterion.

Borda and Kemeny aggregations, being motivated from social choice theory, strive for fairness and hence treat all rankers as equally important. However, fairness is not a desirable property in our setting, since we know that some individual rankers (measures) are likely to perform better than others in our target tasks. In fact, given the ordering of a (small) set of candidates, we can estimate the performance of individual rankers and use this to produce a better ranking on a new set of candidates.

In order to accommodate such supervision, we extend Borda and LK aggregation to incorporate weights associated with each input ranking. The weights correspond to the relative utility of each ranker for the task at hand. We refer to such a version of Borda and LK as *Supervised Borda* and *Supervised LK* respectively. Also in the LK approach, instead of using total orderings provided by each ranker, we evaluate partial orderings of only the top K candidates for each ranker. We refer to this variant as LK TopK. Thus, our weighted LK algorithm can be run with varying three options, namely (1) with and without supervision, (2) with total orderings or partial (top K) orderings, and (3) with different initial orderings. We experimented with several combinations of these three options.

Twitter Case Study: Our case study was based on the Twitter discussion around Pepsi. What piqued our interest in Twitter and the role of influencers was the infamous iPhone app called “AMP UP B4 U SCORE”. An avalanche of Twitter users slammed the app ultimately leading to an apology from Pepsi. In this study, we found that the influence of the twitter users heavily depend upon the number of rebroadcasts of his/her messages to millions of other users. In the universe of Twitter, this suggests that a useful task would be to predict which twitterers will be significantly rebroadcasted via retweets.

One obvious indicator of influence could be the number of followers a user has (in-degree of the Follower Graph). However, many users *follow* 100K or more users and therefore this may not be sufficient indication of influence. For this reason, we consider two alternatives, the Retweet Graph and the Mention Graph – where edges correspond to retweets and mentions of users in the past. For our input rankings we use in-degree, out-degree and PageRanks from these graphs.

We extracted the data to generate these graphs over a two week period from 11/11/09 to 11/26/2009. This gives a Follower Graph with 40 million nodes (users) and 1.1 billion edges. We used the sociometrics computed from these graphs to predict which users will have viral outbursts of retweets in the following week. We compare these predictions with the actual amount of retweets in the following week.

We construct our prediction task from our data by dividing users in our test period into two classes – people who have been retweeted 100 or more times within a week, and those who have not. We treat this as a binary classification problem, where the ranking produced by each measure is used to predict the potential for viral retweeting in the test time period. Since we are primarily concerned with how well these measures perform at ranking users, we compare the area under the ROC curve (AUC) based on using each measure by itself.

We compared all individual and aggregate measures of influence in 20 trials of random stratified train-test splits. We find that 9 of the 13 individual measures by themselves are quite effective at ranking the top potentially viral twitterers with an AUC > 0.8. Not surprisingly, the total number of times that someone has been retweeted in the recent past (Weighted Retweet Graph Indegree) and the number of followers (Follower Graph Indegree) produces a very good ranking. However the Spearman rank correlation between recent past retweets and followers is not high (0.43), suggesting that there are multiple forces at work here.

Next, we compared the different supervised and unsupervised rank aggregation techniques. As expected, the supervised versions performed better than the unsupervised versions. We also observe that all aggregation techniques improve over the individual ranking measures. The exception here is LK on total orderings, which can often perform worse than an individual measure. However, the real benefit to using LK can be seen when it's applied only to the partial ordering of the top K candidates. When applied to partial orderings, LK TopK performs better than Borda. These results are improved by using Supervised LK TopK; which are further improved by using Supervised Borda as the initial ranking. Thus supervised locally optimal order-based ranking proves to be advantageous over Borda and unsupervised methods.

Citation Network Case Study: In addition to Twitter data, we also performed a case study on publication citation networks. For this we used a collection of papers with their citations that was used in the KDD Cup contest held in 2003. This data consists of 1,716 papers in the field of High Energy Physics Theory (*hep-th*), published on arXiv during a 6 month period. The data set also contains the number of times each paper was downloaded during the 60 day period after it was published. As such, we define the task of predicting highly influential papers, as measured by downloads, based on the citation data of the papers. If a paper received 600 or more downloads, we consider it as a high-influence paper (77 papers); else we consider it to have little or no influence.

First, we constructed a citation graph based on all publications in *hep-th*, which was also provided as part of KDD Cup 2003. In this citation graph, each node represents a paper and each edge represents a citation. As of May 1, 2003, there were 29,014 papers and 342,427 citations in total in the *hep-th* data. Next, for each of the 1,716 papers with download information, we used this citation graph to compute 5 influence measures – Indegree, Outdegree, Pagerank, Hub and Authority score.

We ran experiments as before, using 20% of the data (343 papers) for training the supervised methods. As expected, the number of papers citing a given paper (in-degree) is a good indicator of how often the paper will be downloaded. Furthermore, having more citations from highly-cited papers, as captured by PageRank is a better indicator of influence in this data. These results also substantiate the fact that supervised rank aggregation algorithms perform better than their unsupervised counterparts. Notably, Supervised LK TopK (Supervised Borda) (AUC = 0.8170) outperforms all individual measures (0.6107-0.8109) and other rank aggregation techniques (0.7747-0.8142).

Conclusion: In this paper, we have addressed the problem of identifying influence in networks by casting it in the form of predictive tasks; which allows us to assess the effectiveness of different measures of influence in light of standard classification and ranking metrics. We have evaluated our approach on two case studies, social and citation networks. We demonstrated that combining aspects of different measures produces a composite ranking mechanism that is most effective for a desired task. In particular, we demonstrated the merits of supervised locally-optimal order-based rank aggregation.

REFERENCES

[1] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, 1999, pp. 604-632.

[2] D. Knoke and R.S. Burt, "Applied Network Analysis," *Applied Network Analysis*, Newbury Park, CA: Sage, 1983.

[3] J.C. Borda, "Memoire sur les elections au scrutin," *Histoire de l'Academie Royale des Sciences*, 1781.

[4] C. Dwork, R. Kumar, R. Naor, and D. Sivakumar, "Rank Aggregation Methods for the Web," *WWW*, 2001.