Credit Allocation

May 2, 2018

1 Credit Allocation Problem

$$\mathrm{PR}_j = \frac{1-d}{N} + d \sum_{k \in I_j} \frac{\mathrm{PR}_k}{|O_k|}$$

where N is the total number of nodes, I_j is the incoming nodes to j, O_k is the outgoing nodes from j, $d \in (0,1)$ is the damping factor, it's the probability of following the out-links to the next node.

2 Intrinsic Credits

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ be the entire set of articles, $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ be all the authors appeared in $\mathcal{P}$. Each author may have multiple works, and each paper may have multiple coauthors. Let A_i be the coauthors of p_i. We assume that the credit of a paper is allocated among all the authors in $\mathcal{A}$. The credit allocation can be considered as the mechanism of intrinsic value allocation, reflecting the contribution, directly or indirectly from each researchers. Let $C \in \mathbb{R}^{n \times m}$ be the intrinsic credit allocation matrix of m papers over n authors.

The Shen-Barabasi credit allocation approach allocates credits based on the recognition from a community of the co-cited papers. The co-cited papers play as a committee, each of them has an independent credit allocation over the coauthors of the target paper. However, the credit allocation values the relative contribution from the coauthors in subject but completely ignoring other authors' contribution. The credit allocation is definitely inconsistent with the true allocation.

To construct an intrinsic credit allocation schema, we proposed an iterative approach based on the infrastructure of SB model. Let $S = (s_1, s_2, \ldots, s_m)$ be the strength matrix, and each column vector is defined as the strength vector in SB model, i.e. s_i is the strength vector of paper p_i. The strength indicates the relevance of a pair of papers, and it's measured with the co-citation count.

Based on SB model, we have $C = CS$, which is equivalent to solve from $C(S-I) = 0$ the matrix C. Also, we expect the credits allocated to non-authors being minimized. Let B be the indicator matrix presenting that $b_{ij} = 0$ if a_i is a coauthor of p_j, otherwise $b_{ij} = 1$. The minimization terms would be

$$\sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij}b_{ij} = \mathrm{tr}(B^T C).$$

Integrating both components, we define the objective function

$$L(C) = \|C(S-I)\|_F^2 + \mathrm{tr}(B^T C) = \mathrm{tr}[(S-I)^T C^T C(S-I)] + \mathrm{tr}(B^T C).$$

Obviously, we can compute the gradient

$$\begin{aligned}\nabla_C L(C) &= \partial[\|C(S-I)\|_F^2 + \mathrm{tr}(B^T C)]/\partial C \\ &= \partial \mathrm{tr}[(S-I)^T C^T C(S-I)]/\partial C + B \\ &= 2C(S-I)(S-I)^T + B.\end{aligned}$$

The matrix $(S-I)(S-I)^T$ is positive semi-definite. Let the gradient be zero, we directly solve the optimal credit allocation

$$C = -B[(S-I)(S-I)^T]^{\dagger}/2,$$

where $A^{\dagger}$ stands for the pseudo-inverse.

However, when the number of papers or the authors becomes very large, even the storage of these matrices becomes prohibitively expensive, it's infeasible to conduct the pseudo-inverse computation.

Considering the sparsity of these matrices, the iterative updating mode may be a better choice. Let the initial guess C_0 be a column stochastic matrix, with each coauthor of the paper the same amount of credits. With the GD method, it's updated following the rule:

$$C_t \leftarrow C_{t-1} + \alpha[2C_{t-1}(S-I)(S-I)^T + B],$$

where $\alpha \in (0,1)$ is the learning rate.