IRIE: A Scalable Influence Maximization Algorithm in Social Networks

Finding influential users in a social network is essential for viral marketing and social media marketing. Influence maximization problem is defined as finding a node set S of given size K in a social network to maximize their *influence spread* — the expected total number of activated nodes under a certain diffusion process initiated from the set S. In this work, we propose a novel scalable and memory-efficient Influence Rank Influence Estimation (IRIE) algorithm for the influence maximization problem under the popular independent cascade (IC) model [1] and its extension. In the IC model, each activated node has a single chance to activate each of its outgoing neighbor with a probability assigned to the edge. The IC model can be identified with the Susceptible/Infective/Recovered (SIR) model for the epidemic spreading [2]. Kempe et al. [1] showed that finding optimum solution for the influence maximization under the IC model is NP-hard, and proposed a Greedy algorithm that obtains $(1-1/e)$-approximation for the problem. A number of follow-up works tackle the problem by designing more efficient and scalable optimizations and heuristics [3, 4, 5]. Among them PMIA [3] has stood out as the most efficient heuristic so far.

In the greedy algorithm as well as in PMIA, each round a new seed with the largest marginal influence spread is selected. To select this seed, the greedy algorithm uses Monte-Carlo simulations while PMIA uses more efficient local tree based heuristics to estimate marginal influence spread of every possible candidate. These are especially slow for the first round where the influence spread of every node needs to be estimated. Instead of estimating influence spread for each node at each round, we devise a global influence ranking method, Influence Rank(IR), derived from a belief propagation approach. By integrating IR with tree-based influence estimation IE, we propose our scalable and memory-efficient IRIE algorithm.

Let $\sigma(S)$ be the expected total number of activated nodes given a seed set S. When $S=\emptyset$, our algorithm computes estimate $r(u)$ of influence $\sigma(\{u\})$ of a node u by the following equation.

$$r(u) = 1 + \alpha \cdot \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right), \tag{1}$$

where N^{out} is a set of our-neighbor of u, P_{uv} is the probability that u activates its out-neighbor v, and $\alpha \in (0,1]$ is a damping factor. We first prove that $r(u)$ with $\alpha = 1$ is very close to $\sigma(\{u\})$ in any tree graph, and show that $r(u)$ is a good estimate of $\sigma(\{u\})$ in any graph. After selecting some seed node, we consider the influence from the selected seed node. Let $AP_S(u)$ be the probability that a node u is activated when the diffusion process begins from the seed set S. To estimate $AP_S(u)$, we adopt a tree-based approximation to influence of each seed node [3]. Then, we compute estimate $r(u)$ of marginal influence $\sigma(S \cup \{u\}) - \sigma(S)$ of a node u by the following equation.

$$r(u) = (1 - AP_S(u)) \cdot \left(1 + \alpha \left(\sum_{v \in N^{out}(u)} P_{uv} \cdot r(v) \right) \right). \tag{2}$$

The factor $(1 - AP_S(u))$ indicates the probability that a node u is not activated by the seed set S. Note that (1) and (2) are exactly same when $S = \emptyset$. We compute iterative computations of (2) up to t times and obtain $r^{(t)}(u)$, which computes the estimate of marginal influence of u within distance t from u.

We conduct extensive experiments using synthetic networks as well as six real-world networks such as Amazon and DBLP whose size ranging from $29K$ to $69M$ edges, and different IC model parameter settings. In the experiments comparing IRIE with the state-of-the-art algorithms such as Greedy with Cost-effective Lazy forward(CELF) [5], PMIA [3], and SA [4], our results show that IRIE has matching or sometimes even better influence spread as the CELF and PMIA, and generates much better influence spread than SA. For the scalability and memory usage, IRIE achieves up to two orders of magnitude speedup comparing with PMIA (much more with CELF) while using significant less memory than PMIA, especially for large and relatively dense networks. To show the wide applicability of our IRIE approach, we also suggest a variant of IRIE to the IC-N model that incorporates negative opinion propagations [6]. Our simulation results show that IRIE has better influence coverage with less running time than the known state-of-the-art MIA-N heuristic [6].

References

1. D. Kempe, J. Kleinberg, E. Tardos, in *KDD* (2003), pp. 137–146
2. Newman, M.E.J, SIAM Review **45**, 167256 (2003)
3. W. Chen, C. Wang, Y. Wang, in *KDD* (2010)
4. Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, K. Xie, in *AAAI* (2011)
5. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N.S. Glance, in *KDD* (2007), pp. 199–208
6. W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, W. Yuan, in *SDM* (2011)