Online Processing Algorithms for Influence Maximization

背景

在线社交网络成为信息传播的有效途径，特别是市场营销方面，可以据此制定策略来获取利益。而这关键之处在于选取初始种子集合使得影响的用户数量最多，即影响力最大化问题。这个问题被Kempe等人形式化的概括出来，基于两个基础的传播模型：IC和LT。在IM问题上，有大量的研究工作，但都是离线处理，即如果在算法运行中途停止，则不会得到任何结果，这是缺乏交互性和灵活性的。

解决问题

- 解决了目前IM上的离线算法不具有交互性和灵活性的问题。

创新之处

- 提出了一个在线算法OPIM来处理IM问题，并使其具有交互性和灵活性。
- 设计了一种新的方法提高近似率 α 。

相关工作

- 对于IM问题，它首先被Kempe等人提出，并用贪心算法结合蒙特卡洛模拟返回了近似保证为 $1-1/e-\epsilon$ 的解。之后有很多研究工作，有一部分通过启发式算法减少计算量，还有一部分通过反向影响抽样的方式在保证近似率的情况下减少计算量。
- 反向影响抽样的算法包括Tang等提出的TIM和改进之后的IMM，Nguyen等提出的SSA和D-SSA，之后改进的SSA-Fix和D-SSA-Fix。都是在保证 $1-1/e-\epsilon$ 的近似率下减少计算时间。
- 还有一些IM问题的变体，像Lei等提出的学习影响概率，最大化影响传播；Wang等研究动态传播模型下的IM问题；Tang等研究利益指标，联合优化利益和成本。

解决方法

OPIM定义

在线处理影响最大和问题（OPIM）,给定一个在线社交网络G，一个影响传播模型，一个正整数k，一个失败概率δ，一组时间序列$t_1,t_2,t_3\ldots$，要求在时间t_i找到一个种子集合S_i，在$1-\delta$概率的情况下得到近似率为α_i的解，目标是最大化$\alpha_1,\alpha_2,\alpha_3\ldots$。

OPIM算法

1. OPIM算法采用反向影响抽样(RIS)来评估种子集合的期望影响传播。即采用反向可达(RR)集合：

   （1）从V中随机选择一个节点v

   （2）生存一个样本集合R，每个节点$u\in V$，作为单个种子节点以概率激活v，那么它被加入集合R中。返回R。

   假设选取种子集合为S，可以通过RR集合评估期望影响传播：

   $$\sigma(S)=|V|\cdot Pr[S\cap R\neq\emptyset]$$

2. 将算法生成的RR集合$\mathcal{R}$分成两个不相交的集合$\mathcal{R}_1$和$\mathcal{R}_2$，用$\mathcal{R}_1$来确定种子集合S^*，用$\mathcal{R}_2$来计算S^*的近似保证。该近似保证α通过计算一个比值来得到，通过$\mathcal{R}_1$和$\mathcal{R}_2$计算出S^*期望影响传播的下界$\sigma^l(S^*)$和上界$\sigma^u(S^o)$，返回它们的比值作为α。之后作者优化了一下上界，通过两种方式得到了两个优化后的上界$\hat{\sigma^u}(S^O)$和$\sigma^o(S^o)$。这两个上界更加紧

凑。

3. 用于传统IM问题的OPIM算法过程
    1. 从随机RR集合生存两个不相交且大小相等的集合 $\mathcal{R}_1$ 和 $\mathcal{R}_2$
    2. 应用贪心算法在 $\mathcal{R}_1$ 上生成大小为 k 的种子集合 S^*，计算下界 $\sigma^l(S^*)$ 和上界 $\hat{\sigma^u}(S^O)$，得到近似率 α
    3. 如果 α 达到 $1-1/e-\epsilon$，返回 S^*，否则生成双倍大小的集合 $\mathcal{R}_1$ 和 $\mathcal{R}_2$ 继续求解。

H2 实验

H3 数据集

Dataset	n	m	Type	Avg. degree
Pokec	1.6M	30.6M	directed	37.5
Orkut	3.1M	117.2M	undirected	76.3
LiveJournal	4.8M	69.0M	directed	28.5
Twitter	41.7M	1.5G	directed	70.5

H3 算法

- $OPIM^0$ 和应用两种改进上界的 $OPIM^+$ 和 $OPIM'$
- IMM
- SSA-Fix
- D-SSA-Fix

H3 参数设置

种子集合大小$k=50$，失败概率$\delta=1/n$，RR集合大小$2^i\times 1000$，每个算法运行50次，结果取平均。

H3 结果

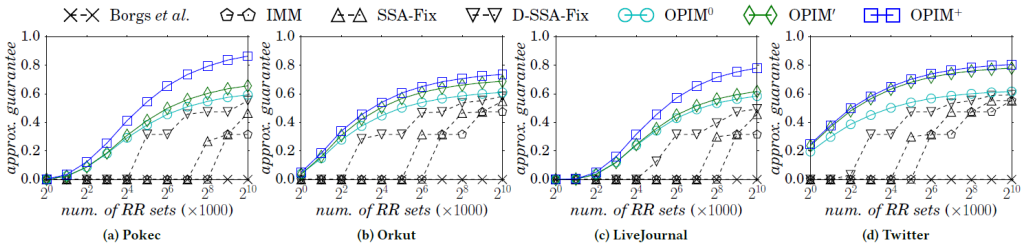

Figure 2: Approximation guarantee on various graphs under the LT model (k = 50).

上图是各个算法在线性阈值模型上，在四个数据集上，所能达到的近似保证随RR集合大小的变化。三种$OPIM$算法表现较稳定，且数量RR集合下，近似保证比其它算法要高，$OPIM^+$最高，大约高出传统算法20到30个百分点。

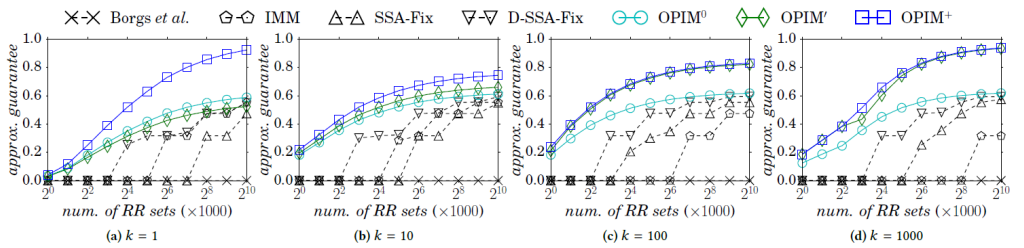

Figure 3: Approximation guarantee for different seed set sizes k on the Twitter dataset under the LT model.

上图是在LT模型上，种子集合大小对近似保证的影响，可以看出$OPIM^+$算法遥遥领先其它算法。

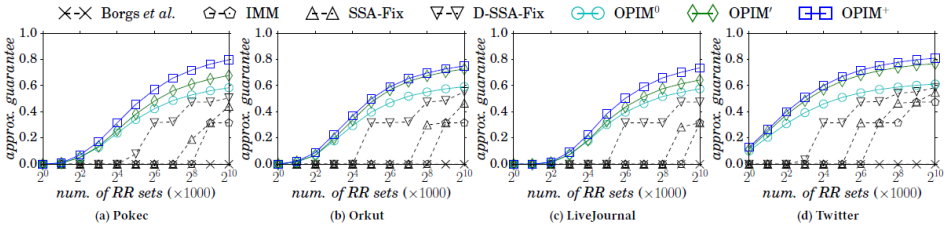

Figure 4: Approximation guarantee on various graphs under the IC model ($k = 50$).

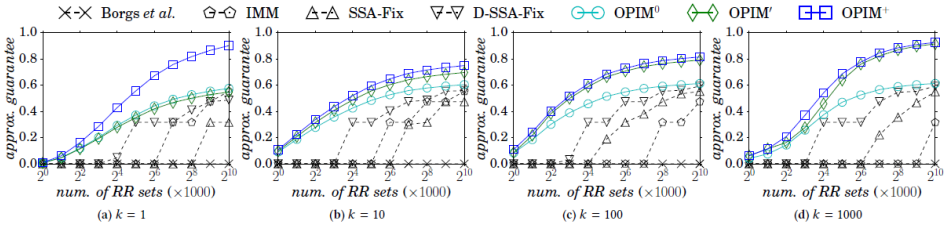

Figure 5: Approximation guarantee for different seed set sizes k on the Twitter dataset under the IC model.

上面的图是在IC模型上又做了一次和LT模型一样的对比实验，结果和LT模型相差不大。

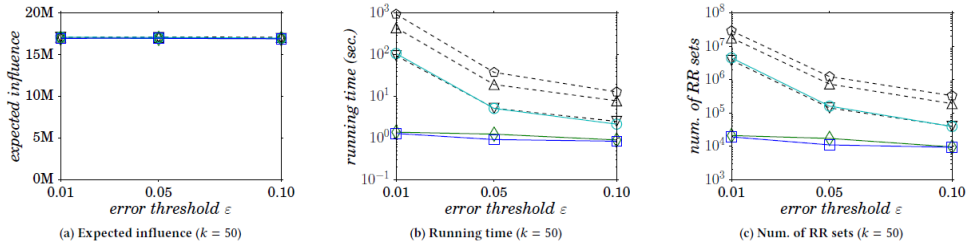

Figure 6: Influence maximization with $(1 - 1/e - \varepsilon)$-approximation on the Twitter dataset under the LT model.

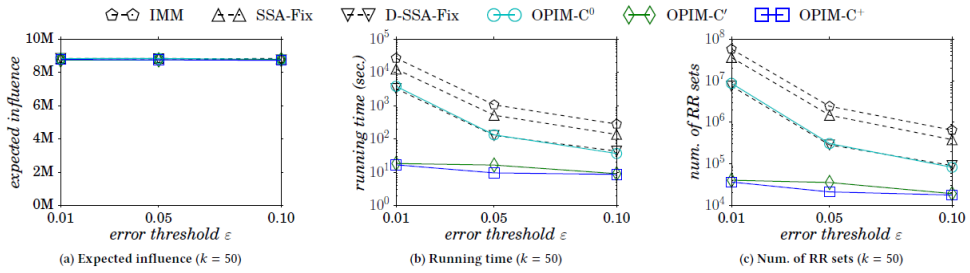

Figure 7: Influence maximization with $(1 - 1/e - \varepsilon)$-approximation on the Twitter dataset under the IC model.

上图是在Twitter数据集上，在两个传播模型下达到$1 - 1/e - \epsilon$近似率运行得到的结果。其中（a）表示误差ϵ的值变化对期望影响传播的影响，结果是对所有算法影响相同；（b）是看对运行时间的影响，$OPIM$算法明显比其它算法效率要高，$OPIM^+$最高；（c）是看对RR集合数量的影响，$OPIM$需要的RR集合数量显然要少很多，$OPIM^+$需要数量最少。

H2 总结

作者开发了一个在线处理影响最大化问题的算法OPIM，避免了算法运行时间长而中途终止算法得不到任何结果的问题。这也算是一种基于交互性的新算法思想。特别是数据量很大，运行时间很长的时候，该算法更能体现出它的优势。

实验表明，作者开发的OPIM算法不仅是一个在线算法，在离线任务上的表现也比先进的算法性能和效率高很多。