

doi: 10.13682/j.issn.2095-6533.2016.06.023

基于友好度的影响力最大化算法

吴 旭,胡一清

(西安邮电大学 计算机学院, 陕西 西安 710121)

摘 要: 提出一种基于友好度的影响力最大化算法。利用社交网络中用户行为信息计算友好度并构建友好关系网络,通过友好度调整节点的扩散概率,同时启发式地选取友好度积累值较高的节点作为启发节点以此提升算法的实现效率,最后采用 CELF++ 算法计算影响力最大的节点集。实验结果表明,该算法与 CELF++ 算法、TIM 算法和 DegreeDiscount 算法相比,在获得较好的扩散效果的同时亦将执行时间有效的控制在一定范围内。

关键词: 影响力最大化;友好关系;社交网络;社会计算

中图分类号: TP393.093

文献标识码: A

文章编号: 2095-6533(2016)06-0118-05

Influence maximization algorithm based on friendliness

WU Xu, HU Yiqing

(School of Computer Science and Technology, Xi'an University of Posts and Telecommunications Xi'an 710121, China)

Abstract: A new influence maximization algorithm based on friendliness is proposed in this paper and named as Friendliness-based Influence Maximization Method (FIMM). FIMM uses behavior and interaction information of nodes in social network to compute friendliness and to build friendly relationship network. FIMM also adjusts the diffusion probability of nodes through the friendliness and improves the efficiency of implementation of the algorithm by selecting nodes. These nodes have higher accumulation of friendliness and have the most potential influence heuristically based on friendly relationship network. Experiment results show that FIMM can get better diffusion effect than CELF++, TIM and DegreeDiscount. FIMM can also limit implement time effectively in a smaller range, which means that using friendliness and selecting nodes heuristically can optimize influence maximization algorithm.

Keywords: influence maximization, friendliness, social network, social computing

社会网络中的影响力最大化算法^[1,2]可以帮助发现现实生活中不易发现的社会现象或者社会问题。其目的是寻找若干个种子节点使影响力扩散最大化,得到影响力最大化的最优解决方案。

利用自然贪心算法计算影响力最大化的效果虽然明显好于基于节点度或中心性的启发式的影响力最大化算法,但其计算效率不高,特别是面对拥有

海量用户节点和边社交网站,低效问题更加明显^[3]。

CELFF(Cost-Effective Lazy Forward selection)算法^[4]在自然贪心算法的基础上,利用影响力最大化的子模块特性提升了贪心算法的效率;CELFF++算法^[5]是对 CELFF 算法的改进,利用堆特性进一步提升了算法效率。NewGreedy 算法^[6]和 CELFF 算法^[4]融合得到的 MixedGreedy 算法,计算效率也有

收稿日期: 2016-6-16

基金项目: 国家自然科学基金资助项目(71501156);中国博士后基金资助项目(2014M560796);陕西省教育厅科研计划资助项目(15JK1679);西安邮电大学创新基金资助项目(114-602080048)

作者简介: 吴旭(1978-),女,副教授,博士,从事可信计算、信任管理和云计算等研究。E-mail: xrdz2005@163.com.

胡一清(1990-),男,硕士研究生,研究方向为计算机网络舆情。E-mail: 172142282@qq.com.

所提高,但是计算得出的种子节点个数与最佳的影响力最大化的节点集仍有差距; DegreeDiscount 算法^[6]利用节点度启发式选取种子节点,并没有考虑到网络中用户间的关系信息,通过牺牲扩散范围来降低时间复杂度;与 DegreeDiscount 算法相似, TIM(Two-phase Influence Maximization)算法^[7]虽然效率较高,但是也没有对个体间的相互关系进行分析和量化,导致扩散效果受到限制。

目前,影响最大化问题中的影响力扩散模型主要有线性阈值模型(Linear Threshold Model, LTM)和独立级联模型(Independent Cascade Model, ICM)^[3]。扩散模型中个体之间的影响力权值默认是已知的常量或者提前赋予的数值,但是在现实环境中这种假设不成立,因此有必要对个体间的影响力因素及其相互关系进行分析和量化。

在信息扩散过程中,用户之间的友好关系是主要影响因素之一^[8-11]。针对用户之间行为和交互信息较少所导致的度量结果与实际情况有偏差的问题,本文拟提出基于友好度的影响力最大化算法(Friendliness-based Influence Maximization Method, FIMM),并基于独立级联模型构建 FIMM 算法框架。利用用户间友好度调整扩散概率,可使友好度较高的用户之间的传播概率更大。同时将启发式的选取友好度积累值较高的节点作为最有潜力的影响力节点以期提升算法的实现效率。

1 基于友好度的影响力最大化算法

1.1 FIMM 算法框架

FIMM 算法的实现框架,如图 1 所示,包括两个阶段:(1)友好关系网络建立阶段。通过计算用户间的友好度得到用户间的友好关系网络。(2)FIMM 算法的具体实现阶段。通过友好度调整节点的激活概率,同时启发式的选取友好度积累值较高的节点作为最有潜力的影响力节点。

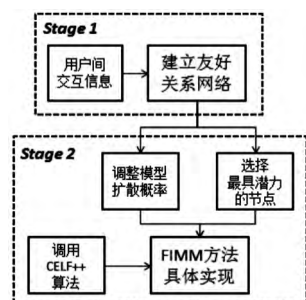


图 1 FIMM 算法的实现框架

1.2 友好关系网络的建立

采集用户间的历史交互行为信息。通过分析这些行为的分布规律和因果关系得到社交网络中用户间的熟悉度与相似度。根据用户间的熟悉度与相似度计算用户间友好度,并利用友好度建立用户间友好关系网络。友好度计算公式为

$$T(a, b) = F(a, b)S(a, b), \quad (1)$$

其中 $T(a, b)$ 为 a 与 b 之间的友好度, $S(a, b)$ 为 a 与 b 的相似度, $F(a, b)$ 为 a 与 b 间的熟悉度。熟悉度计算表达式为

$$F(a, b) = \frac{W_{a, b}}{\sum_{m \in N(a)} W_{a, m}},$$

其中 $W_{a, b}$ 表示 a 与 b 相互回应的消息次数, $N(a)$ 表示 a 的邻居节点集合, $W_{a, m}$ 表示 a 与所有邻居 m 的相互回应的消息次数。用户间的交往时间对熟悉度的计算有影响,将式(1)通过时间因子 t_{Δ} 进行修正,得到

$$F(a, b) = \frac{W_{a, b}}{\sum_{m \in N(a)} W_{a, m}} t_{\Delta}, \quad (3)$$

$$t_{\Delta} = \ln(1 + \frac{\Delta T_{a, b}}{\text{Max}(\Delta T_{\text{universe}})}). \quad (4)$$

其中 $\Delta T_{a, b}$ 表示用户双方首次相互回应的时间点和最新交互回应的时间点的差值。 $\Delta T_{\text{universe}}$ 为友好关系网络中所有节点间首次相互回应的时间点和最新交互回应的时间点的差值。最后将 $\Delta T_{a, b}$ 进行平滑处理得到时间因子 t_{Δ} 。

通过 Pearson 相关系数(Pearson Correlation Coefficient)^[12]调整得到相似度的计算公式为

$$S(a, b) = \left[1 + \frac{\sum_{i \in I_{a, b}} (M_{a, i} - \bar{M}_a)}{\sqrt{\sum_{i \in I_{a, v}} (M_{a, i} - \bar{M}_a)^2}} \times \frac{(M_{b, i} - \bar{M}_b)}{\sqrt{\sum_{i \in I_{a, b}} (M_{b, i} - \bar{M}_b)^2}} \right] \alpha. \quad (5)$$

其中 $M_{a, i}$ 表示 a 对兴趣项目 i 的打分, \bar{M}_a 与 \bar{M}_b 分别为 a 与 b 对兴趣项目集打分的平均分。由于会出现用户双方打分的兴趣项目数量相差过大而导致相似度偏差的情况,因此利用惩罚因子 α 进行修正。将 a 和 b 打过分的项目的数量差值进行平滑处理,取倒数后得到惩罚因子

$$\alpha = \frac{1}{[\ln(e + |I_a - I_b|)]^{1/2}}, \quad (6)$$

其中, I_a 和 I_b 分别表示 a 和 b 打过分的项目的数量。

计算用户对其邻居友好度的积累得到累加的

友好度积累值 $C(a)$, 该值被用来度量用户的潜在影响力, 其表达式为

$$C(a) = \sum_{m \in N(a), m \notin A_r} T(a, m). \quad (7)$$

其中 A_r 表示第 r 轮影响力扩散的种子节点集。友好度积累值 $C(a)$ 较高的节点将作为最有潜力的影响力节点。

1.3 FIMM 算法

利用独立级联模型模拟扩散过程。假设集合 A_r 表示第 r 轮激活的节点集合, 其中 $A_0 = S$, S 表示种子节点集合。对于边 \overline{ab} , $a \in A_r$ 并且 b 是未激活节点。在 $r+1$ 轮时, a 将以激活概率 p 尝试激活节点 b , 并且只有一次尝试的机会。接着激活 a 的邻居节点, 一旦执行激活就会移除这个激活路径(即边 \overline{ab} , 且为无向边)。尝试激活完所有与 a 相连的节点后, 将节点 a 从集合 A_r 中去除。当集合 A_r 为空时, 则这一轮的激活结束。激活概率 p 的计算公式为

$$p = p_{ab} = \bar{p} T(a, b), \quad (8)$$

其中 \bar{p} 为整个网络中的平均影响力扩散概率。

FIMM 算法引入了友好度作为 \bar{p} 的修正因子, 不仅能够正确评估行为的发起者对被传播者之间的影响力, 还可以通过友好度捕捉并预测人们在社交网络上的行为。

FIMM 算法将分为 3 步选取 k 个种子节点, 并引入启发因子 $h (0 \leq h \leq 1)$, 启发因子 h 意味着启发阶段选取启发节点个数所占比率。FIMM 算法的具体实现步骤如下。

步骤 1 调整激活概率。

在友好关系网络的基础上, 通过式(8)调整激活概率 p 。

步骤 2 启发式地选取种子节点。

利用式(7)计算节点在友好关系网络上的友好度积累值, 友好度积累值较高的节点被选取作为启发节点。选取 $[hk]$ 个节点作为启发节点。重复执行 $[hk]$ 轮, 每轮选取未激活的友好度累积值较高的节点作为启发节点(种子节点), 然后在保留上一轮扩散的节点的基础上用启发节点进行扩散。

步骤 3 利用贪心算法选取种子节点。

利用 CELF++ 算法计算剩余的 $k - [hk]$ 个种子节点。重复执行 $k - [hk]$ 轮, 直到得到具有 k 个种子节点的集合 S 。输出集合 S 。

当 $h=0$ 时, 算法跳过步骤 2 直接步骤 3。其中启发因子 h 为经验值, 因此需要确定 h 的值才能得到一个确定的影响最大化方法。

2 实验结果及分析

2.1 实验数据集

为验证 FIMM 算法在社交网络中的有效性, 选取两组用户数据集作为仿真验证对象。从社交网络抽取用户节点及节点信息形成社交网络图, 其中用户节点包括用户对电影、音乐、书籍的评分。评分项直接反应了用户的兴趣。通过这些节点数据信息得到节点间的熟悉度与相似度以构建友好关系网络。节点数据为:

(1) 用户信息。包括用户名 ID、用户名、关注用户数和被关注用户数。

(2) 用户关注关系。包括当前用户 ID、当用户关注用户的 ID 和被当前用户关注的用户 ID。

(3) 评分信息。当前用户对电影、书籍及音乐专辑的评分。

最后通过用户关注关系中相互关注的行为形成边。数据集描述如表 1 所示。

表 1 数据集描述

数据集	节点个数	边数	评分项
数据集 1	14 927	56 424	39 872
数据集 2	2 386	10 896	9 113

2.2 实验对比

独立级联模型中平均影响力扩散概率取 $\bar{p} = 0.1^{[3,6]}$, 在贪心算法阶段设置轮数 $R = 20\ 000$ 。

选取种子集合个数 k 作为输入参数, 影响范围和算法执行时间作为种子集合 S 在扩散模型之上的仿真效果的评价指标。其中影响范围表示算法利用种子节点 S 扩散一定的轮数所影响的节点个数。将 CELF++ 算法、TIM 算法和 DegreeDiscount 算法 3 种影响力最大化算法与 FIMM 算法进行比较, 共进行两组实验。

实验 1 选取种子集合个数 k 作为参数, 用以考察在 k 值一定的情况下取不同启发因子 h 所得到的影响范围。数据集 1 的实验结果如图 2 所示。

由图 2 可知, 排除 $h=1$ 的情况, h 取其它值所得到的影响范围均比 $h=1$ 时要大。当 $k=50$, $h=0.5$, FIMM 算法所得的影响范围比 h 的其他取值要高出 3%。当 $h=1$ 时, FIMM 算法变为静态选取潜在影响力最大的节点, 将其作为种子节点, 这种

选取方式未考虑影响力的传播过程,因此所得的影响范围也最差。而当 $h=0$ 时, FIMM 算法退化为 CELF++ 算法。

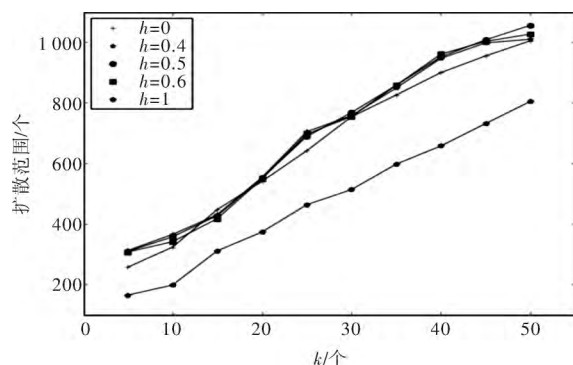


图 2 参数 k 与 h 在数据集 1 中的影响范围曲线

数据集 2 的实验结果如图 3。当 $h \neq 0$ 且取合适的值时,所得的影响范围优于 $h=0$ 情况。这说明在友好关系网络的环境下启发式的选取一部分种子节点优于直接使用 CELF++ 算法。同时,当 $h=0.5$ 时,所取得的影响范围仍高于 h 取其他值的情况。

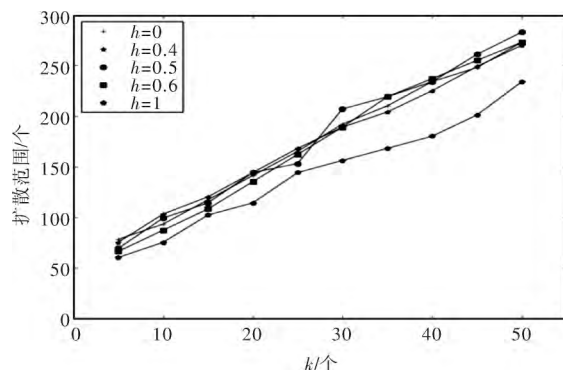


图 3 参数 k 与 h 在数据集 2 中的影响范围曲线

实验 2 将 CELF++ 算法、TIM 算法、DegreeDiscount 算法与 FIMM 算法分别在影响范围和算法执行时间上进行比较。其中 FIMM 算法的启发因子取 $h=0.5$ 。实验结果分别如图 4 和图 5 所示。

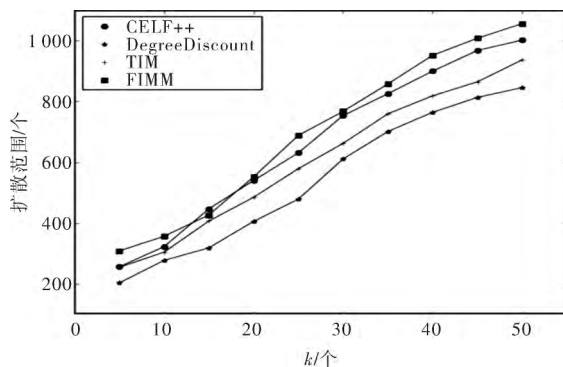


图 4 不同算法在数据集 1 中的影响范围曲线

从图 4 可以看出,在 $k=50$ 时,由 FIMM 算法

得到种子节点的扩散范围(即所影响的种子节点数)比 CELF++ 算法得到的扩散范围高出 7%。这说明 FIMM 算法框架通过友好关系网络有效地模拟了社交网络中影响力传播过程进而使挖掘出的最具影响力的种子节点集合的扩散效果得到提升。由于 FIMM 算法通过计算节点的友好度累积值,启发式的选取种子节点,因此当种子节点个数 k 的取值较小时会影响 FIMM 的扩散效果。从实验结果上看, $k > 25$ 时, FIMM 算法的扩散范围开始优于其他影响力最大化算法。

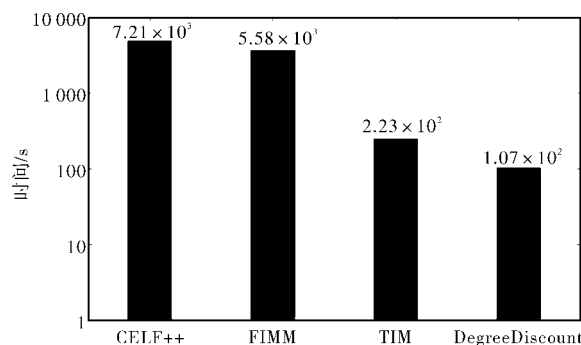


图 5 不同算法在数据集 1 中所消耗的时间

从图 5 中可以看出, CELF++ 算法的执行时间高于其他算法,这是由于该算法在整个社交网络中利用蒙特卡洛模拟估算节点影响力。而 FIMM 算法在构建友好关系网络后利用节点友好度的积累选取启发节点将有效减少算法执行时间,并且最终的扩散范围优于 CELF++ 算法。FIMM 算法与其他算法相比在获得较好的扩散效果的同时亦将执行时间有效的控制在一定范围内。实验结果验证了友好度的引入和启发式的选取种子节点较好地优化了影响力最大化算法。

3 结语

FIMM 算法利用独立级联模型模拟扩散过程,采集用户间的历史交互行为信息计算用户间的友好度,以此建立用户间的友好关系网络;启发式的选取友好度累积值较高的节点作为最有潜力的影响力节点,以减少寻找种子节点所消耗的时间。与 CELF++ 算法、TIM 算法和 DegreeDiscount 算法对比实验结果表明, FIMM 算法在获得较好的扩散效果的同时亦将执行时间有效的控制在一定范围内,这表明友好度的引入和启发式的选取种子节点较好地优化了影响力最大化算法。

参 考 文 献

- [1] DOMINGOS P, RICHARDSON M. Mining the Network Value of Customers[C/OL]//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2001: 57-66 [2016-5-20]. <http://dx.doi.org/10.1145/502512.502525>.
- [2] RICHARDSON M, DOMINGOS P. Mining Knowledge-sharing Sites for Viral Marketing[C/OL]//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2002: 61-70 [2016-5-20]. <http://dx.doi.org/10.1145/775056.775057>.
- [3] KEMPEL D, KLEINBERG J, TARDOS. Maximizing the Spread of Influence Through a Social Network[C/OL]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA: ACM, 2003: 137-146 [2016-5-20]. <http://dx.doi.org/10.1145/956750.956769>.
- [4] LESKOVEC J, KRAUSE A, GUESTRIN C. Cost-effective Outbreak Detection in Networks[C/OL]//Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, 2007: 420-429 [2016-5-20]. <http://dx.doi.org/10.1145/1281192.1281239>.
- [5] GOYAL A, LU WEI, LAKSHMANAN L V S. CELF++: Optimizing the Greedy Algorithm for Influence Maximization in Social Networks[C/OL]//Proceedings of the international conference companion on World Wide Web. New York, NY, USA: ACM, 2011: 47-48 [2016-5-20]. <http://dx.doi.org/10.1145/1963192.1963217>.
- [6] CHEN W, WANG Y, YANG S. Efficient Influence Maximization in Social Networks[C/OL]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA: ACM, 2009: 199-208 [2016-5-20]. <http://dx.doi.org/10.1145/1557019.1557047>.
- [7] TANG Y Z, XIAO X K, SHI Y C. Influence Maximization: Near-Optimal Time Complexity Meets Practical Efficiency[C/OL]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, New York, NY, USA: ACM, 2014: 75-86 [2016-5-20]. <http://dx.doi.org/10.1145/2588555.2593670>.
- [8] VACCARO A, ADARMS S K, KISLER T S, et al. The Use of Social Media for Navigating the Transitions Into and Through the First Year of College[J/OL]. Journal of the First-Year Experience & Students in Transition, 2015, 27(2): 29-48 [2016-5-20]. <http://eric.ed.gov/?q=social+AND+media+AND+school&pg=4&id=EJ1102760>.
- [9] LUO H, MA L. Empirical research on consumers' post-transaction general trust in B2C E-business[C/OL]// 2013 10th International Conference on Service Systems and Service Management, Hong Kong: ICSSSM, 2013: 208-213 [2016-5-20]. <http://dx.doi.org/10.1109/ICSSSM.2013.6602639>.
- [10] 吴旭. 基于增强稳定组模型的移动 P2P 网络信任评估方[J/OL]. 计算机学报, 2014, 37(10): 2118-2127 [2016-5-20]. <http://dx.chinadoi.cn/10.3724/SP.J.1016.2014.02118>.
- [11] ZHANG H, WANG Y, YANG J. Space Reduction for Contextual Transaction Trust Computation in E-Commerce and E-Service Environments[C/OL]// 2015 IEEE International Conference on Services Computing (SCC), 2015: 680-687 [2016-5-20]. <http://dx.doi.org/10.1109/SCC.2015.97>.
- [12] AHLGREN P, JAMEVING B, ROUSSEAU R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient[J/OL]. Journal of the American Society for Information Science and Technology, 2003: 54(6): 550-560 [2016-5-20]. <http://dx.doi.org/10.1002/asi.10242>.

[责任编辑:祝剑]