一种社交网络群组间信息推荐的有效方法

黄震华[1,2],张 波[3],方 强[1],向 阳[1]
(1.同济大学计算机科学与工程系,上海 201804; 2.同济大学嵌入式系统与服务计算教育部重点实验室,上海 201804;
3.上海师范大学信息与机电工程学院,上海 200234)

摘 要: 群组间信息推荐是社交网络中人们传递和分享资讯的重要途径,然而获取精确的最优推荐方案需要指数级时间开销.为此,本文提出一种有效算法EAOORS(Efficient Algorithm for Obtaining Optimal Recommendation Solution),将该指数级时间开销问题等价归约为EST(Extended Steiner Tree,扩展Steiner树)问题,并在多项式时间复杂度内快速获取近似最优推荐方案.理论分析和仿真实验表明,本文所提的算法具有有效性和实用性.

关键词: 社交网络; 信息推荐; 近似算法; 性能评估
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2015)06-1090-04
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.06.008

An Efficient Algorithm of Information Recommendation Between Groups in Social Networks

HUANG Zhen-hua[1,2], ZHANG Bo[3], FANG Qiang[1], XIANG Yang[1]
(1. *Department of Computer and Technology, Tongji University, Shanghai* 201804, *China*;
2. *The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai* 201804, *China*;
3. *College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai* 200234, *China*)

Abstract: Information recommendation between groups is one of the most important ways for information sharing and transmitting in social networks. However, it needs exponential time cost to achieve the exact optimal recommendation solution. Hence this paper proposed an efficient algorithm EAOORS (Efficient Algorithm for Obtaining Optimal Recommendation Solution) which equivalently reduced this exponential time cost problem to extended steiner tree problem, and fast achieved the approximate optimal recommendation solution in the polynomial time complexity. The experimental result shows that the EAOORS algorithm is more efficient than the existing ones.

Key words: social network; information recommendation; approximation algorithm; performance evaluation

1 引言

进入Web 2.0时代,社交网络使人们更容易进行信息的推荐和分享[1].近年来,群组模式被广泛引入社交网络中,群组模式的出发点是基于某种紧密关系将不同用户组合在一个社区中,从而把用户从相对封闭的好友关系疏导至群组,创建一种新的更开放的社交关系,实现信息的传播和分享.

文献[2]通过对社交网络的节点度、聚集系数、特征路径长度以及膨胀率等指标的计算,从多个角度刻画了社交网络群组所具有的特征.文献[3]设计了一个基于终端交互的机器学习系统ReGroup来协助用户创建个性化的社交网络群组.当用户将网络成员添加进某一的社交网络群组时,ReGroup系统基于群组成员特征来学习并获取相关的概率模型,并利用该概率模型来评估网络成员是否适合添加进社交网络群组.作者通过实验评估来表明ReGroup系统能够有效协作用户创建大规模异构社交网络群组.文献[4]利用DBLP和LiveJournal两个网络数据集来测试和分析社交网络群组的规模增长和演化规律.文献[5]将两种不同类型的社交网络World of Warcraft (WoW)和DBLP作为测试平台来建模并预测社交网络群组的动态稳定性.测试结果表明网络成员的多样化程度和社交活动水平是维护社交网络群组稳定性的两个关键因素,同时存在某一特定的网络成员集合,它们在维护社交网络群组的稳定性上扮演重要的角色.

收稿日期:2013-11-25;修回日期:2014-09-01;责任编辑:覃怀银
基金项目:国家自然科学基金(No.61272268,No.61103069,No.71171148);教育部新世纪优秀人才支持计划(No.NCET-12-0413);霍英东教育基金会应用研究课题(No.142002)

我们发现,目前关于社交网络群组的研究工作主要集中在群组的构建及其演化规律上,关于群组之间信息的传播和分享的研究较少,然而社交网络群组提出的初衷就是为了更有效地实现信息的传播和分享[6].为此,本文基于目前主流的成员信任关系机制,在个体客观信誉度和主观信任度的评价模型[7]下,研究群组间信息推荐的最优方案问题,它是群组之间信息传播和分享的基础.然而,获取精确的最优信息推荐方案是 NP-hard 问题,因此,本文提出一个时间复杂度多项式的有效算法 EAOORS(Efficient Algorithm for Obtaining Optimal Recommendation Solution)来快速获取近似最优推荐方案,其中 n_1和 n_2分别表示社交网络中两个群组 G_1和 G_2的成员个数.EAOORS 算法首先评估群组 G_1和 G_2个体成员的客观信誉度和主观信任度,并将 G_1和 G_2转化为一有向赋权图 G,进而将获取最优信息推荐方案等价归约为 G 上 EST(Extended Steiner Tree,扩展 Steiner 树[8])问题.理论分析和仿真实验表明,本文所提的算法具有有效性和实用性.

2　问题定义

假定社交网络中存在两个群组 G_1和 G_2,其成员集分别为 $G_1=\{v_1^1,\cdots,v_{n_1}^1\}$和 $G_2=\{v_1^2,\cdots,v_{n_2}^2\}$,$G_1$中的每个成员 v_i^1均关联一个客观信誉度,记为 $\mathrm{obj}(v_i^1)$,同时,G_2中的每个成员 v_j^2对于 G_1中的 $c_j\leqslant n_1$个成员 $v_1^1,\cdots,v_{c_j}^1$存在主观信任态度,分别记为 $\mathrm{rep}(v_j^2,v_1^1),\cdots,\mathrm{rep}(v_j^2,v_{c_j}^1)$.那么群组 G_1到 G_2信息推荐的最优方案 IRS 可定义为:获取 $\mathrm{sub}G_1\subseteq G_1$,使得 $\forall v_j^2\in G_2$,均有 G_1中的成员对其进行信息推荐;并确定 $\mathrm{sub}G_1$中成员到 G_2中成员的推荐路径,使得:$\mathrm{IRS}(\mathrm{sub}G_1,G_2)=\sum_{v_i^1\in \mathrm{sub}G_1}\mathrm{obj}(v_i^1)+\sum_{\forall v_j^2\in G_2}\mathrm{rep}(v_j^2,v_i^1)$最大.

从上面问题定义我们可以看出,对于两个信息推荐方案 IRS_1和 IRS_2,如果 $\mathrm{IRS}_1(\mathrm{sub}G_1,G_2)>\mathrm{IRS}_2(\mathrm{sub}G_2,G_2)$,那么 IRS_1比 IRS_2优越,即由 $\mathrm{sub}G_1$对 G_2进行信息推荐的可信度比由 $\mathrm{sub}G_2$对 G_2进行信息推荐的可信度高.

3　EAOORS:快速获取近似最优推荐方案

获取精确最优的信息推荐方案 optIRS 将花费大量的 CPU 时间开销,因此本文将提出一种快速获取近似最优方案 appIRS 的有效方法 EAOORS(Efficient Algorithm for Obtaining Optimal Recommendation Solution),其基本思想可描述如下:首先我们将 G_1到 G_2的信息推荐过程构造成一个加权有向非循环图 $G=(N,E,W)$,其中顶点集合 $N=G_1\cup G_2=\{v_1^1,\cdots,v_{n_1}^1,v_1^2,\cdots,v_{n_2}^2\}$,边集 $E=\{v_i^1\rightarrow v_j^2\mid v_i^1$ 到 v_j^2 存在推荐路径$\}$,在 N 和 E 上定义权函数 $W:N\cup E\rightarrow Z^+$,即对于属于 G_1的顶点 v_i^1,W 取值为 $\mathrm{obj}(v_i^1)$,对于 E 中每条边 $v_i^1\rightarrow v_j^2$,W 取值为 $\mathrm{rep}(v_j^2,v_i^1)$.不难看出,$G$ 是一个二分图,因为顶点集 G_1和 G_2内部不存在相连的边,只有它们之间存在有向加权边.接着,我们在 G 中添加一个控制节点 Θ,并且对于 G_1中的每个顶点 ζ,添加一条由 Θ 指向 ζ 的有向边 $e\langle\Theta,\zeta\rangle$,其权赋值为 $\mathrm{obj}(\zeta)$,并删除 ζ 原先的顶点权值.此时我们得到一个新的加权有向非循环图 $G'=(N',E',W')$,其中顶点集合 $N'=G_1\cup G_2\cup\{\Theta\}=\{v_1^1,\cdots,v_{n_1}^1,v_1^2,\cdots,v_{n_2}^2,\Theta\}$,边集 $E'=E\cup\{\Theta\rightarrow v_i^1\mid 1\leqslant i\leqslant n_1\}$,权值集合 $W'=W\cup\{\forall v_i^1\in G_1,W(\Theta\rightarrow v_i^1)=\mathrm{obj}(v_i^1)\}-\{\forall v_i^1\in G_1,W(v_i^1)=\mathrm{obj}(v_i^1)\}$.最后,我们利用文献[13]的 EST-A 算法产生 G'上的扩展 Steiner 树 $\overline{\mathrm{ESTree}}(\hat{N},\hat{E},\hat{W})$,从而可获得近似最优方案 appIRS:$G_1$中参与信息推荐的成员集合为 $\mathrm{par}G_1=\hat{N}-\{v_1^2,\cdots,v_{n_2}^2,\Theta\}\subseteq G_1$,$\hat{E}$ 中若存在 $\mathrm{par}G_1$成员 $v^{(p)}$到 G_2成员 v_j^2 的一条边 $v^{(p)}\rightarrow v_j^2$,那么这条边即为 $v^{(p)}$到 v_j^2 的推荐路径,且边的权值 $\mathrm{rep}(v^{(p)},v_j^2)$等于 v_j^2 对 $v^{(p)}$的主观信任度.

EAOORS 算法依据上述基本思想进行实施,其伪码如算法 1 所示.

算法 1　EAOORS

Input:群组 $G_1=\{v_1^1,\cdots,v_{n_1}^1\}$和 $G_2=\{v_1^2,\cdots,v_{n_2}^2\}$,$G_1$中的每个成员 v_i^1均关联一个客观信誉度 $\mathrm{obj}(v_i^1)$,G_2中每个成员 v_j^2 对于 G_1中的 $c_j\leqslant n_1$个成员 $v_1^1,\cdots,v_{c_j}^1$存在主观信任态度,分别记为 $\mathrm{rep}(v_j^2,v_1^1),\cdots,\mathrm{rep}(v_j^2,v_{c_j}^1)$;

Output:G_1到 G_2信息推荐的近似最优方案 appIRS;

Begin

1.构造加权有向非循环图 $G=(N,E,W)$:

　(a) $N:-G_1\cup G_2=\{v_1^1,\cdots,v_{n_1}^1,v_1^2,\cdots,v_{n_2}^2\}$;

　(b) $E:-\{v_i^1\rightarrow v_j^2\mid v_i^1$ 到 v_j^2 存在推荐路径$\}$;

　(c) $W:-\{\forall v_i^1\in G_1,W(v_i^1)=\mathrm{obj}(v_i^1)\}\cup\{\forall v_i^1\rightarrow v_j^2\in E,W(v_i^1\rightarrow v_j^2)=\mathrm{rep}(v_j^2,v_i^1)\}$;

2.基于 G 产生加权有向非循环图 $G'=(N',E',W')$:

　(a) $N':-G_1\cup G_2\cup\{\Theta\}=\{v_1^1,\cdots,v_{n_1}^1,v_1^2,\cdots,v_{n_2}^2,\Theta\}$;

　(b) $E':-E\cup\{\Theta\rightarrow v_i^1\mid 1\leqslant i\leqslant n_1\}$;

　(c) $W':-W\cup\{\forall v_i^1\in G_1,W(\Theta\rightarrow v_i^1)=\mathrm{obj}(v_i^1)\}-\{\forall v_i^1\in G_1,W(v_i^1)=\mathrm{obj}(v_i^1)\}$;

3.利用 EST-A 算法[13]生成 G'上的扩展 Steiner 树 $\overline{\mathrm{ESTree}}(\hat{N},\hat{E},\hat{W})$;

4. $\mathrm{par}G_1:-\hat{N}-\{v_1^2,\cdots,v_{n_2}^2,\Theta\}$;

5. $\mathrm{rePATH}:-\hat{E}$;

6. $\mathrm{appIRS}(\mathrm{par}G_1,G_2):-\sum_{v_i^1\in \mathrm{par}G_1}\hat{W}(\Theta\rightarrow v_i^1)+\sum_{\forall rp\in \mathrm{rePATH}}\hat{W}(\mathrm{rp})$;

　//计算 appIRS 信息推荐方案的总体可信度

7. $\mathrm{appIRS}:-\langle\mathrm{par}G_1,\mathrm{rePATH},\mathrm{appIRS}(\mathrm{par}G_1,G_2)\rangle$;

8. Return appIRS;

End

EAOORS 算法具有多项式的时间复杂度,如定理 1 所示.

定理 1 EAOORS 算法产生从 $G_1=\{v_1^1,\cdots,v_{n_1}^1\}$ 到 $G_2=\{v_1^2,\cdots,v_{n_2}^2\}$ 近似最优信息推荐方案 appIRS 的时间复杂度为:

$$O(n_1\times n_2+(n_1+n_2)^{e/(e-1)}+(n_1+n_2)\log(n_1+n_2))$$
$$\approx O(n_1\times n_2+(n_1+n_2)^{1.58}+(n_1+n_2)\log(n_1+n_2)).$$

证明 EAOORS 算法构造加权有向非循环图 $G=(N,E,W)$ 的时间开销为 $O(n_1\times n_2)$,而基于 G 产生加权有向非循环图 $G'=(N',E',W')$ 的时间开销为 $O(n_1+n_2)$.其次,依据文献[13]可知,EAOORS 算法使用 EST-A 算法生成 G' 上的扩展 Steiner 树 $\overline{\text{ESTree}}(\hat{N},\hat{E},\hat{W})$ 需要 $O((n_1+n_2)^{e/(e-1)}+(n_1+n_2)\log(n_1+n_2))$ 的时间开销.最后,由 $\overline{\text{ESTree}}(\hat{N},\hat{E},\hat{W})$ 产生近似最优信息推荐方案 appIRS 的时间开销为 $O(n_1+n_2+n_1\times n_2)$.所以可得,EAOORS 算法的总时间为 $O(n_1\times n_2+n_1+n_2+(n_1+n_2)^{e/(e-1)}+(n_1+n_2)\log(n_1+n_2)+n_1\times n_2+n_1+n_2)=O(n_1\times n_2+(n_1+n_2)^{e/(e-1)}+(n_1+n_2)\log(n_1+n_2))$,同时由 $e=2.72$ 代入可得 EAOORS 算法的总时间为 $O(n_1\times n_2+(n_1+n_2)^{1.58}+(n_1+n_2)\log(n_1+n_2))$.所以定理 1 成立.

4 实验评估

本节通过具体实验来评估 EAOORS 算法及其扩展版 X-EAOORS 算法的有效性.在实验中,群组 G_1 和 G_2 的成员个数在 $1\times10^4\sim5\times10^4$ 间变化;G_1 中成员的客观信誉度由 Gauss 分布数据产生器[16]生成,为了方便,我们将客观信誉度的取值标准化在[0,1]之内;G_1 到 G_2 间推荐路径的平均数在 10~50 间变化,即对于 G_2 中的每个成员,G_1 中平均有 10~50 个成员与它存在推荐路径;G_2 中成员对 G_1 中成员的主观信任度同样由 Gauss 分布数据产生器[16]生成,同时,我们将主观信任度的取值标准化在[0,1]之内.仿真实验的机器配置为四核 i5-3450 CPU、4GB 内存和 500GB 硬盘,操作系统为 CentOS Linux 6.4,所有算法的代码编译采用 JDK 1.6.

与之比较的方法有两个:(1)OPTIMAL 算法,通过指数级时间复杂度的穷举来获取 G_1 到 G_2 间精确最优的信息推荐方案;(2)GREEDY 算法,针对 G_2 中的每个成员,获取与之主观信任度最大的 G_1 成员,从而产生贪婪信息推荐方案.本小节实验分为 3 组:(1)固定 G_2 的成员个数为 3×10^4、G_1 到 G_2 间推荐路径的平均数为 30,而 G_1 成员个数在 $1\times10^4\sim5\times10^4$ 间变化;(2)固定 G_1 的成员个数为 3×10^4、G_1 到 G_2 间推荐路径的平均数为 30,而 G_2 成员个数在 $1\times10^4\sim5\times10^4$ 间变化.图 1 给出三个算法产生推荐方案优劣程度和运行时间的评估结果.

在图 1 评估算法产生推荐方案优劣程度的实验中,我们以 OPTIMAL 算法为基准,因为它产生的推荐方案是精确最优的,即将该精确最优方案的可信度定为 100%.我们从图 1 可以看出,EAOORS 算法产生推荐方案的可信度接近于 OPTMMAL 算法,而 GREEDY 算法产生推荐方的可信度比较差,这主要是因为 EAOORS 算法利用扩展 Steiner 树的近似最优理论来返回推荐方案,因此推荐方案的可信度能够得到很好的保证,而 GREEDY 算法是针对 G_2 中的每个成员,获取与之主观信任度最大的 G_1 成员,来产生贪婪信息推荐方案,因此易于陷入局部最优的问题,从而无法获取全局最优的推荐方案.例如在图 1(*a*)中,当 G_1 的成员数为 1×10^4 时,EAOORS 算法产生推荐方案的可信度是 OPTIMAL 算法的 87.5%,而 GREEDY 算法只有 39.9%;在图 1(*b*)中,当 G_2 的成员数为 3×10^4 时,EAOORS 算法产生推荐方案的可信度是 OPTIMAL 算法的 93.6%,而 GREEDY 算法只有 30.3%.

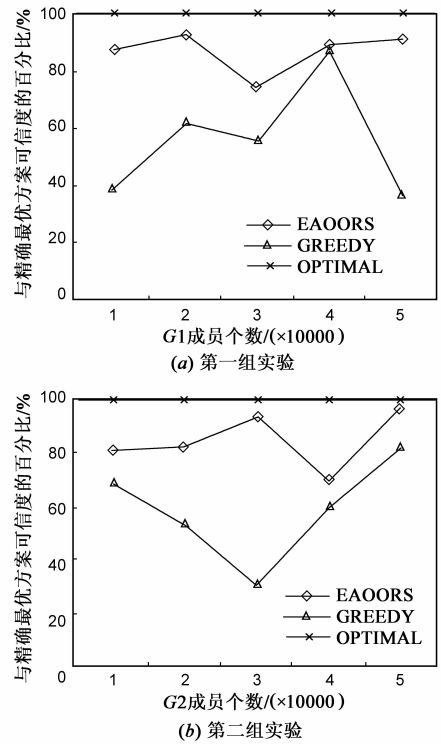

图1 算法产生推荐方案优劣程度实验评估

虽然 OPTMMAL 算法产生精确最优方案的可信度略高于 EAOORS 算法,然而在图 2 评估算法运行时间的实验中,我们可以发现 OPTMMAL 算法在每种实验环境下的运行时间多非常巨大,这主要是因为 OPTMMAL 算法为了获取精确最优方案,需要遍历推荐方案空间的所有实例,因此需要指数级的时间开销.而 EAOORS 算法只

需要多项式时间开销即可返回近似最优推荐方案,而无需遍历推荐方案空间的所有实例,它的运行时间比较接近于 GREEDY 算法.例如在图 2(a)中,当 G_1的成员数为 5×10^4时,EAOORS 算法的运行时间是 OPTIMAL 算法的 1.7%,而 GREEDY 算法为 0.56%;在图 2(b)中,当 G_2的成员数为 5×10^4时,EAOORS 算法的运行时间是 OPTIMAL 算法的 0.49%,而 GREEDY 算法为 0.21%.

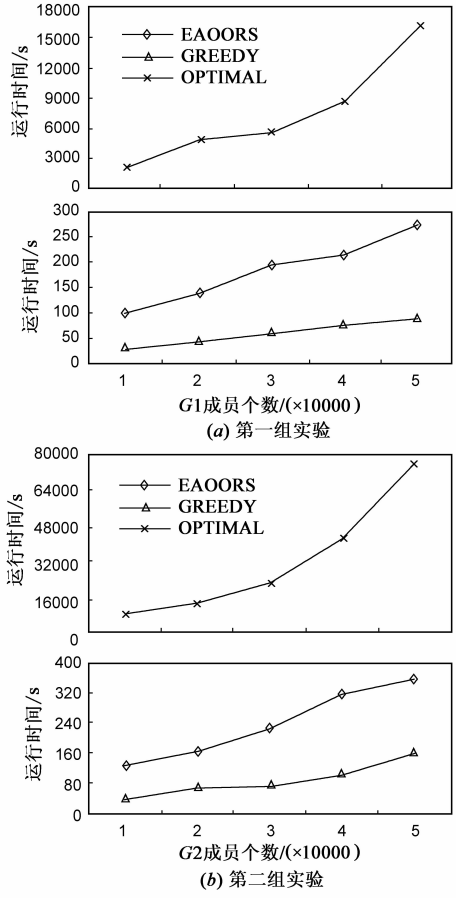

图2　算法运行时间实验评估

因此,综合图 1 和图 2 的实验评估,我们可以得出本文的 EAOORS 算法能够很好平衡推荐方案可信度与运行时间,而且具有很好的可扩展性.

5　结论

群组间信息推荐技术是近年来社交网络领域的一个研究热点和重点,这主要是因为群组模式的出发点是基于某种紧密关系将不同用户组合在一个社区中,从而把用户从相对封闭的好友关系疏导至群组,创建一种新的更开放的社交关系,实现信息的传播和分享.

然而我们发现获取精确的最优推荐方案是指数级时间代价问题.本文在个体客观信誉度和主观信任度的评价模型下提出了一种多项式时间复杂度的算法 EAOORS 来高效获取近似最优推荐方案.理论分析和仿真实验表明,本文所提的算法具有有效性和实用性.

参考文献

[1] S Cohen, L Ebel, B Kimelfeld. Asocial network database that learns how to answer queries[A]. Proceedings of the Biennial Conference on Innovative Data Systems Research[C]. Asilomar: ACM Press, 2013. 1 – 4.

[2] SP Borgatti, A Mehra, DJ Brass, G Labianca. Network analysis in the social sciences[J]. Science, 2009, 323(5916): 892 – 895.

[3] SAmershi, J Fogarty, D Weld. Regroup: interactive machine learning for on-demand group creation in social networks[A]. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems[C]. Austin: ACM Press, 2012. 21 – 30.

[4] L Backstrom, D Huttenlocher, J Kleinberg, X Lan. Group formation in large social networks: membership, growth, and evolution[A]. Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining[C]. Philadelphia: ACM Press, 2006. 44 – 54.

[5] A Patil, J Liu, J Gao. Predicting group stability in online social networks[A]. Proceedings of the 22nd International Conference on World Wide Web[C]. Rio: ACM Press, 2013. 1021 – 1030.

[6] 冷作福.基于贪婪优化技术的网络社区发现算法研究[J].电子学报,2014,42(4):723 – 729.
ZF Leng. Community detection in complex networks based on greedy optimization[J]. Acta Electronica Sinica, 2014, 42(4): 723 – 729. (in Chinese)

[7] A Mislove, M Marcon, P Gummadi, B Bhattacharjee. Measurement and analysis of online social networks[A]. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement [C]. Kyoto: ACM Press, 2007. 29 – 42.

[8] 梁东敏,马绍汉.一类扩展的 Steiner 树优化问题及其应用[J].计算机学报,1996,19(12):895 – 902.
DM Liang, SH Ma. An extended steiner tree optimization problem and its applications[J]. Chinese Journal of Computers, 1996, 19(12): 895 – 902. (in Chinese)

作者简介

黄震华　男,1980 年生,博士,副教授,软件行业协会系统工程分会理事,CCF 会员.主要研究方向为云计算、信息服务、数据挖掘与知识发现等.
E-mail: huangzhenhua@tongji.edu.cn

张　波　男,1978 年生,博士,副教授.主要研究方向为社交网络数据分析、数据挖掘与语义计算等.

方　强　男,1991 年生,硕士研究生.主要研究方向:数据挖掘、模式识别和服务发现等.

向　阳　男,1962 年生,教授,博士生导师.主要研究方向为智能计算、数据仓库、数据挖掘、决策支持系统与语义网等.