

[文章编号] 1003—4684(2020)01-0033-05

基于 FP-Growth 的社交好友推荐方法研究

熊才权, 陈 曦

(湖北工业大学计算机学院, 湖北 武汉 430068)

[摘 要] 针对基于关系的好友推荐中偏离共同兴趣以及推荐好友数量不足的问题,将数据挖掘中 FP-Growth 关联规则算法应用于社交网络好友推荐中,对用户间的相互关注关系进行深度挖掘,将不同用户同时被关注的事件作为一个项集,挖掘其频繁模式,再根据设定支持度,推荐用户感兴趣 Top-N 组合好友。63641 条实验结果表明,算法具有良好的性能,可实现较高的召回率与准确率。

[关键词] 社交网络; 关注关系; 频繁模式; FP-Growth

[中图分类号] TP18 [文献标识码] A

随着 Web2.0 时代的发展,用户成为网络内容的主导者。人们通过网络社交活动,例如评论关注转发收藏等形式参与网络社交获取相应的信息。然而随着网络社区的发展,用户数量规模的急剧扩增,人们对社交的需求更加丰富,人们不仅仅想要通过社交平台跟线下的好友进行交互,还想通过网上社交活动拓展自己的朋友圈,以获取一些自己需要的资源和时兴的动态消息。在如今主流的社交网络上的好友推荐大多采取 FOF^[1]理论拓展朋友圈,如果两个用户拥有很多共同好友,那么他们是朋友的概率将会很大。Armentano^[2,3]等人将社交网络结构和用户特征结合,将邻域内满足条件的用户推荐给目标用户。Akbari^[4]等人提出了一种基于图拓扑和人工蜂群的新方法,改进了 FOF 为三度分离的好友推荐,有效的扩大了朋友推荐的范围。基于已有的好友进行推荐,虽然具有较高的可信度,但是更多的是推荐线下可能认识的好友,无法推荐社交网络上基于兴趣相同的好友,同时也存在冷启动问题。YANG^[5]等人使用社会化标签将社交网络拓扑结构和用户兴趣网络结合起来进行个性化推荐。Wu W^[6]等人使用词频反文档 TF-IDF 算法提取用户文本关键词,为每个用户自动添加兴趣和标签,为目标用户推荐兴趣相似的用户。肖晓丽^[7]等人提出基于用户兴趣和社交信任的聚类推荐算法,对于用户数据矩阵稀疏和扩展性差问题提出了解决方案,在一定程度上解决了冷启动问题。基于兴趣的好友推荐

可以很好的解决社交网络中朋友圈拓展问题,但是容易忽略潜在好友。王兵辉^[8]提出社交网络中潜在好友推荐算法。向程冠^[9]等人改进 AprioriTid 算法应用于好友推荐。基于关联规则算法的好友推荐可以解决复杂的好友关系网络问题,同时也可以挖掘潜在好友。因此,本研究通过分析用户相互关注关系挖掘用户的好友关系,通过实验比较分析关联规则算法 FP-Growth 和 Apriori 算法计算好友关系数据,挖掘社交网络中的好友频繁模式,为好友推荐提供有效的建议。

1 好友关系复杂网络问题

微博作为国内成熟的社交平台,由于其主要的社交功能属性,及其作为信息交流的集散地,大量的用户活跃在其中,从而提供了可获取的社交关系网络,通过研究该平台的社交网络关系,不仅可以实现推荐系统的应用,同时通过对复杂的社会关系的解析可以提供高价值的数据分析。

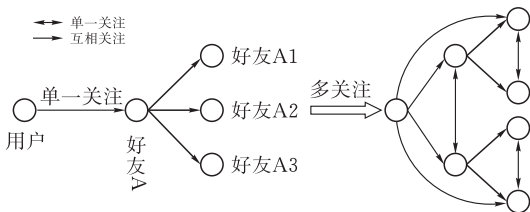


图 1 单一/多好友关注关系网络结构图

用户可以通过关注这种形式形成强关联模式如

[收稿日期] 2019—09—30

[基金项目] 国家重点研发计划项目(2017YFC1405403); 国家自然科学基金(61075059); 湖北工业大学绿色工业科技引领计划项目(CPYF2017008)

[第一作者] 熊才权(1966—), 男, 湖北鄂州人, 湖北工业大学教授, 研究方向为模型识别与智能系统

[通信作者] 陈 曦(1994—), 男, 福建宁德人, 湖北工业大学硕士研究生, 研究方向为模型识别与智能系统

图 1,一个用户可以根据兴趣关注某个好友,同时可以看到该好友还关注了哪些好友,而好友的好友还有其好友,如此,可以扩增好友圈,但寻找的过程中,通过浏览好友短文本信息,耗时久且效率低,并且一旦用户根据兴趣标签关注多个好友,那么好友的好友的数量将激增,用户将很难找到“志同道合”的好友,而数据挖掘知识能够很好解决这类复杂问题,通过设定阈值,利用关联规则算法挖掘出其中哪几个好友常常是同时被关注的,也就是说某个好友被关注的同时还会连带关注某些好友,以此强关联模式帮助用户找到相似兴趣的好友具有较高的推荐可信度。

2 基于 FP-Growth 的好友推荐方法

1.1 FP-Growth 算法

FP-growth 算法是深度优先算法中最新最高效的本质上不同于 Apriori 算法的经典算法,将数据库的信息压缩成一个描述频繁项相关信息的频繁模式树。Apriori 算法有两步骤:一是发现所有的频繁项集;二是生成强关联规则。发现频繁项集是生成强关联规则的前提也是算法中关键的步骤。在 Apriori 算法中利用“频繁项集的子集是频繁项集,非频繁项集的超集是非频繁项集”这一性质有效对频繁项集进行修剪^[10]。Apriori 算法有两个致命的性能瓶颈:1)产生的候选集过大(尤其是 2-项集),算法必须耗费大量的时间处理候选项集 2)多次扫描数据库,需要很大的 I/O 负载,在时间、空间上都需要付出很大的代价。FP-Growth^[11] 算法不同于 Apriori 算法,主要采取数据结构中树存储的概念对数据集进行挖掘,其中有两个关键步骤:一是生成频繁模式树 FP-tree;二是在频繁模式树 FP-tree 上挖掘频繁项集,FPTree 算法在不生成候选项的情况下完成 Apriori 算法的功能。

FP-Growth 算法的效率优于一般的类 Apriori 算法,因为 FP-Tree 算法的整个过程只需要遍历两次事务数据库,并且把大量的数据压缩存储在树中,时间与空间的开销都优于 Apriori 算法。

1.2 基于 FP-Growth 的好友频繁模式挖掘

在复杂的社交网络中,通过 FP-Growth 算法可以挖掘用户-好友的关注关系。如表 1 好友关系数据集中,suid 表示用户编号,tuid 表示用户关注的好友。

- 好友频繁模式挖掘步骤:
- 1)设定最小支持度为 2。
 - 2)tuid 列中得到频繁项的集合和每个频繁项的支持度并降序排序{ $T_2:7, T_1:6, T_3:6, T_4:2, T_5:2$ }

记为 L,依据 L 对 tuid 列重排序,如{ T_1, T_2, T_5 }重排序后为{ T_2, T_1, T_5 } ,以此类推。

3)构建好友关系 FP-Tree,以 null 为根节点,将 tuid 重排序后的项集依次插入树中,如果项集中元素在 FP-Tree 中没有节点,则重新建立节点。如果节点已经存在,则在原有节点上计数加 1,直到项集中所有元素插入到树中,则好友关系 FP-tree 树构建完成图 2。

表 1 好友关系数据集

suid	tuid	suid	tuid
S001	T_1, T_2, T_5	S006	T_2, T_3
S002	T_2, T_4	S007	T_1, T_3
S003	T_2, T_3	S008	T_1, T_2, T_3, T_5
S004	T_1, T_2, T_4	S009	T_1, T_2, T_3
S005	T_1, T_3		

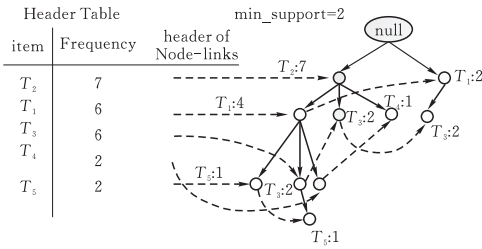


图 2 构建好友关系 FP-Tree

4)调用 FP-growth(Tree,null) 开始进行频繁模式挖掘。伪代码如下

```
输入:好友关系 Tree,模式后缀 a
过程:
if Tree 含单个路径 R then
for 单路径 R 每个节点(b) do
    频繁模式 b U a 组合,
    该组合支持度 support=b 支持度(舍去小于设定支持度为 2 的组合);
end for
else
for 多路径项头表 D
b=D ∪ a 组合为模式后缀,生成其条件模式基 c
产生一个频繁模式 c U b 组合,其支持度 support = c 支持度 > 2;
end for
end if
if Treeb≠∅ then
调用 FP_growth (Treeb, b);
end if
输出: 满足支持度的好友频繁模式
```

由以上好友关系 FP-Tree 树,可以根据模式后缀挖掘频繁模式。如果条件 FP-Tree 是单路径的,则可以通过简单排列组合得到该后缀树的频繁模式。以模式后缀为 T_5 的条件模式树为例:它的条件模式基为($T_2 T_1:1$),($T_2 T_1 T_3:1$), 通过组合变

成 $\{T_2:2,T_1:2,T_3:1\}$,由于 $\{T_3:1\}$ 支持度小于 1 舍去,则通过排列组合后可以得到模式后缀为 T_5 并且支持度大于 2 的频繁模式: $\{T_2 T_5:2,T_1 T_5:2,T_2 T_1 T_5:2\}$ 。

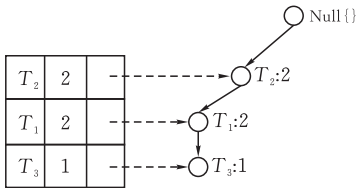


图 3 模式后缀为 T_5 的条件模式树

单路径条件模式树可以直接采用排列组合挖掘频繁模式,但是对于多路径情况的条件,模式树需要另外的考虑。以模式后缀为 T_3 的条件模式树为例,它的条件模式基为 $(T_2 T_1:2),(T_2:2),(T_1:2)$,这是一个多路径树,首先通过模式后缀 T_3 和项

表 2 好友频繁模式生成表

item	条件模式基	产生的频繁模式
T_5	$\{(T_2 T_1:1),(T_2 T_1 T_3:1)\}$	$T_2 T_5:2,T_1 T_5:2,T_2 T_1 T_5:2$
T_4	$\{(T_2 T_1:1),(T_2:1)\}$	$T_2 T_4:2$
T_3	$\{(T_2 T_1:2),(T_2:2),(T_1:2)\}$	$T_2 T_3:4,T_1 T_3:4,T_2 T_1 T_3:2$
T_1	$\{(T_2:4)\}$	$T_2 T_1:4$

从好友模式生成表中,挖掘出好友关系数据集中所有的频繁模式,其中 $\{T_2 T_5:2\}$ 表示 T_2,T_5 用户同时被关注,可以看作是一类组合,它的支持度为 2。 $\{T_2 T_1 T_3:2\}$ 表示 T_2,T_1,T_3 用户同时被关注,该组合的支持度也为 2。 $\{T_2 T_1:4\}$ 表示 T_2,T_1 用户同时被关注,它的支持度为 4。对于满足最小支持度的组合也就是频繁模式都可以挖掘出来,如果在社交平台上需要某一类组合的最小支持度为 20,那么通过 FP-Growth 递归挖掘,再通过降序排序就可以推荐 Top-N 好友组合,这就使得社交网络复杂关注关系的好友推荐具有推荐意义。

3 实验评估

3.1 实验数据集

为了验证算法的有效性,将新浪微博平台作为实验的对象,实验数据集从新浪微博平台抓取,通过清理、集成、变换 3 个步骤对数据集进行处理,将处理后的 63641 条新浪微博信息,1391718 条用户好友关系存储数据库。实验中算法实现采用 JAVA 语言编写,推荐规则的存储采用 Mysql5.5 数据库,实验运行时的硬件配置为 Intel(R) Core(TM) i5-6200U CPU @ 2.30 GHz (4 CPUs),8G 内存,256G 固态硬盘,操作系统为 Win10。

3.2 评价指标

评价好友社交平台采用召回率(Recall)、准确

头表中的每一项做组合得到一组频繁模式 $\{T_2 T_3:4,T_1 T_3:4\}$,然后递归调用 FP-Growth,模式后缀为 $\{T_1,T_3\}$,它的条件模式基为 $\{T_2:2\}$,它是单路径条件模式树,通过组合可得 $\{T_1 T_2 T_3:2\}$ 。最后还需要挖掘模式后缀为 $\{T_2,T_3\}$ 的频繁模式,由于该模式后缀为空,则递归调用结束。最终得出模式后缀为 T_3 的频繁模式为 $\{T_2 T_3:4,T_1 T_3:4,T_1 T_2 T_3:2\}$ 。

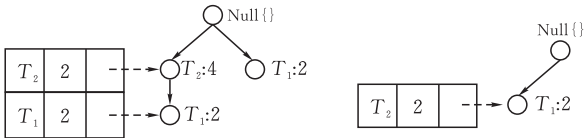


图 4 模式后缀为 T_3 的条件模式树

基于 FP-growth 好友推荐算法,最终得到支持度不小于 2 的好友频繁模式见表 2。

率(Precision)和 F_1 度量(F_1 -measure)三个指标进行评估,召回率主要用于观察推荐算法推荐的好友集合 U_{rc} 与实际可推荐好友集合 U_{real} 交集中好友数量占实际可推荐的好友比率

$$Recall = \frac{|U_{real} \cap U_{rc}|}{|U_{real}|} \times 100\%$$
 (1)

准确率主要用来衡量算法的正确性、可行性,反应推荐出的准确好友数所占推荐好友数的百分比,越高准确性越好,计算如下:

$$Precision = \frac{|U_{real} \cap U_{rc}|}{|U_{rc}|} \times 100\%$$
 (2)

F_1 度量(F_1 -measure)将准确率和召回率的调和平均数作为一个评价标准。 F_1 度量综合考虑了推荐系统的性能,其表达式如下:

$$F_1 = \frac{2 \times Recall \times Precision}{Recall + Precision}$$
 (3)

3.3 实验结果与分析

在本次实验中,将 FP-Growth 算法和 Apriori 算法在好友推荐的实例中进行对比实验,主要通过以下三个方面:

1)设定支持度为 20,从微博用户关注数据集中生成 1 k、2 k、3 k、4 k、5 k、6 k、7 k、8 k 数据集,比较算法运行效率如图 5 所示。

从图 1 FP-Growth 算法和 Apriori 算法数据增大时执行时间比较所示,在相同支持度的情况下,随着数据量的增大,FP-Growth 算法在执行时间上总

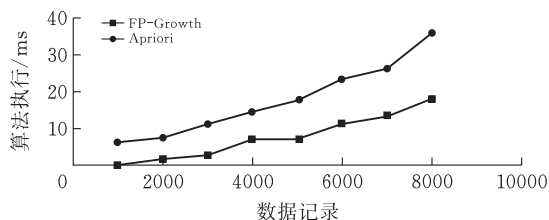


图5 FP-Growth算法和 Apriori 算法数据增大时执行时间比较

是优于 Apriori 算法, 由于 FP-Growth 利用 FP-Tree 树的数据结构进行存储和计算数据, 不需要频繁扫描数据库, 利用计算机内存计算, 大大减少了计算数据时间。

2) 设定恒定的用户关注数据集, 通过调整最小支持度为 10、20、30、40、50、60、70、80 比较算法运行效率如图 6 所示。

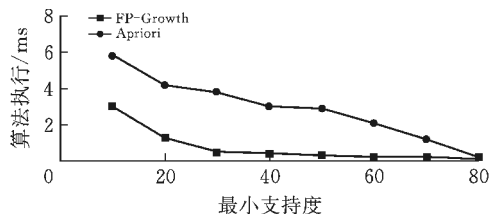


图6 FP-Growth算法和 Apriori 算法支持数据增大时执行时间比较

本次实验从数据集中筛选 2k 数据量, 通过调整不同的支持度比较 FP-Growth 算法和 Apriori 算法执行时间, 从图表中可以看出 FP-Growth 算法执行效率较高。随着最小支持度增大, 候选项集将会相应减少, 两种算法运行效率都会相应提高。

3) 为了验证本文 FP-Growth 好友推荐模型的有效性, 本实验将其与现有的社交网络好友推荐模型: 基于关系的两阶段好友推荐模型^[12] (简称为 M1 模型)、以及基于协同过滤的好友推荐模型^[13] (简称为 M2 模型) 进行对比实验, 推荐目标用户 Top10, Top20, Top30, Top40 好友, 并且比较、观察三个模型在不同推荐好友数量下的 F_1 度量值随推荐好友数量变化的结果见图 7。

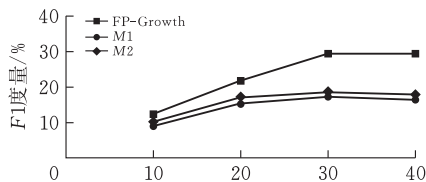


图7 对比实验结果图

从实验结果看出 FP-Growth 好友推荐算法在推荐好友数量为 30 时, F_1 值达到最高, 在准确率和召回率上都相对优于其他两种模型。

4 结论

FP-Growth 算法利用内存运算比较 Apriori 算法在执行时间上有明显的优势, 在准确率和召回率上这两个算法结果相同。本实验通过 FP-Growth 好友推荐模型与基于关系的两阶段好友推荐模型、基于内容的好友推荐模型作比较, 其 FP-Growth 好友推荐算法相较于其他两种算法具有较高的 F_1 度量, 同时由于 FP-Growth 关联规则算法, 是根据用户间的关注关系进行数据挖掘, 在实际应用场景推荐中, 更具有可信度, 在社交网络中可以很好为用户提供好友推荐序列, 提高用户交友体验, 但是基于 FP-Growth 算法本身是基于内存的计算, 对于大数据运行还是需要采取并行化计算, 同时由于本文只是对所有的用户关注关系进行数据挖掘, 并没有涉及用户文本, 推荐结果比较多样, 所以在今后的研究工作中还需要抽取用户文本主题, 通过对主题聚类更加精确地给目标用户推荐好友。

[参考文献]

- [1] Silva N B, Tsang R, Cavalcanti G D C, et al. A graph-based friend recommendation system using genetic algorithm[C]//IEEE congress on evolutionary computation. IEEE, 2010: 1-7.
- [2] Armentano M G, Godoy D, Amandi A. Towards a followee recommender system for information seeking users in twitter[C]//Proceedings of the Workshop on Semantic Adaptive Social Web (SASWeb 2011). CEUR Workshop Proceedings, 2011, 730: 27-38.
- [3] Armentano M G, Godoy D, Amandi A. Topology-based recommendation of users in micro-blogging communities[J]. Journal of Computer Science and Technology, 2012, 27(3): 624-634.
- [4] Akbari F, Tajfar A H, Nejad A F. Graph-based friend recommendation in social networks using artificial bee colony[C]//2013 IEEE 11th International Conference on Dependable, Autonomic and Secure Computing. IEEE, 2013: 464-468.
- [5] Tan Y, CUI Y, JIN Y. BPR-UserRec: a personalized user recommendation method in social tagging systems [J]. The Journal of China Universities of Posts and Telecommunications, 2013, 20(1): 122-128.
- [6] Wu W, Zhang B, Ostendorf M. Automatic generation of personalized annotation tags for twitter users[C]//Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010: 689-692.
- [7] 肖晓丽, 钱娅丽, 李旦江, 等. 基于用户兴趣和社交信

任的聚类推荐算法[J]. 计算机应用, 2016, 36(5): 1273-1278.

[8] 王兵辉. 社交网络中潜在好友推荐算法研究[D]. 昆明:云南大学, 2013.

[9] 向程冠,熊世桓,王东,熊伟程.基于关联规则与相似度的社交好友推荐算法[J].计算机工程, 2019, 45(4): 175-180.

[10] Agrawal R, Srikant R. Fast algorithms for mining as-sociation rules[C]//Proc. 20th int. conf. very large da-ta bases, VLDB. 1994, 1215: 487-499.

[11] Han J, Pei J, Yin Y, et al. Mining frequent patterns without candidate generation: A frequent-pattern tree approach[J]. Data mining and knowledge discovery, 2004, 8(1): 53-87.

[12] 张中峰, 李秋丹. 社交网站中潜在好友推荐模型研究[J]. 情报学报, 2011, 30(12): 1319-1325.

[13] Agarwal V, Bharadwaj K K. A collaborative filtering framework for friends recommendation in social net-works based on interaction intensity and adaptive user similarity[J]. Social Network Analysis and Mining, 2013, 3(3): 359-379.

Research on Friend Recommendation Method at Social Network Based on FP-Growth

XIONG Caiquan, CHENG Xi

(School of Computer Science, Hubei Univ. of Tech., 430068, China)

Abstract: In view of the problem of deviation from common interests and insufficient number of recom-mended friends in relationship-based friend recommendation, FP-Growth association rule algorithm in data mining is applied to social network friend recommendation to conduct deep mining of mutual concern a-mong users. Taking events that different users are concerned about at the same time as an item set, their frequent patterns are mined. Top-N group friends whom users are interested in are then recommended ac-cording to the set support degree. Experimental results show that the algorithm has good performance and can achieve high recall and accuracy.

Keywords: social network; concern relationship; data item; FP-Growth

[责任编辑: 张岩芳]

(上接第 32 页)

Simulation and Optimal Allocation of Agricultural Water Resources System in Xuanhua District

ZHANG Sijun, HE Li, WU Shuang, DU Yu, ZHANG Zhaolong

(School of Electrical and Electronic Engineering, Hubei Univ. of Tech., Wuhan 430068, China)

Abstract: Aiming at the demand-supply conflict and pollution problems caused by improper allocation of agricultural water resources, agricultural water structure at Xuanhua District was analyzed by establishing a system dynamics model to simulate and analyze sensitive factors. A multi-objective model with aims of the minimum supply-demand difference and the maximum agricultural benefit was established, and the constraint of water quality was considered. Finally, the data of Year 2016 were optimized to verify the model, based on which the water resources for the year of 2020 were optimized. The results show that the supply-demand difference and COD were reduced by 6.9718 million cubic meters and 2.75% respectively, and that the growth rate of agricultural benefit was maintained at 10.77%. The optimization plan alleviate the contradiction between supply and demand, which could provide reference of utilization and develop-ment for agricultural water resources.

Keywords: Xuanhua District; agriculture; water resources; simulation; optimization

[责任编辑: 张岩芳]