改进 Apriori 算法在社交网络好友推荐中的应用①

江三锋, 余建坤

(云南财经大学 信息学院, 昆明 650000)

摘 要: 针对 Apriori 算法在频繁项集自连接中产生大量的候选项集以及多次扫描数据库的不足, 提出了一种改进的算法, 该算法将数据库映射到一个布尔矩阵中, 在矩阵列向量进行“与”运算之后, 删除那些没有意义的项和记录, 改进的算法在时间复杂度和空间复杂度上都有很大的提高. 将改进的算法运用到社交网络好友推荐算法中, 将网络社交平台中用户关注的用户和信息作为记录, 将关注的用户作为交易项, 构建交易数据库, 计算频繁 2 项集, 推荐按支持数排序的前 N 位用户作为好友. 通过实验验证, 改进的算法在社交网络好友推荐中具有较高的准确率和召回率.

关键词:Apriori 算法; 社交网络; 布尔矩阵

Application of Improved Apriori Algorithm in Social Network Friends Recommendation

JIANG San-Feng, YU Jian-Kun

(School of Information, Yunnan University of Finance and Economics, Kunming 650000, China)

Abstract: Considering the limits that the Apriori algorithm produces numerous candidate itemsets during the self-joins of frequent items and scans database time after time, this paper proposed an improved algorithm. This algorithm maps the database to a boolean matrix, and then, deletes those meaningless items and records after the AND operation between matrix columns. This will greatly reduce the time and space complexities. Applying to the friend recommendation algorithm in social networks, this improved algorithm regards the interested users and information as records, takes the concerned users as deal items, builds a transaction database, computes frequent 2-item sets and recommends Top-N users ranked by supporting number as friends . The experiment proves the improved algorithm has higher precision and recall in friend recommendation algorithms of social networks.

Key words: Apriori algorithm; social network; Boolean matrix

随着网络技术的飞速发展, 网络交友已经成为一种趋势, 如国外著名的 Facebook 和国内的腾讯微博、新浪微博等. 在网络社交中, 用户的好友一般都来自亲戚、同学、朋友或者朋友的朋友; 或者通过各种社交平台推荐来添加好友. 目前, 著名的社交网络平台 Facebook 根据用户之间的共同好友数来推荐好友, 当共同好友数超过某个阈值, 则可以推荐为好友; 国内的各种社交平台都是推荐被关注量比较大的用户. 而对于大部分用户来说, 他们更想添加和自己有共同兴趣爱好的的用户, 所以对于大部分用户想要真正找到同自己有共同兴趣爱好的用户变得十分困难.

关联规则挖掘是在数据库中搜索两个项目之间存在的关系. 例如: 数据库中项目 B 在项目 A 中出现, 那么通过关联规则可以表示为 $A \rightarrow B$. 社交网络好友推荐正是这种二元关系. 因此, 可以通过关联规则挖掘算法从社交网络的海量数据中挖掘并向用户推荐具有共同兴趣爱好的好友.

Apriori 算法是最早的关联规则挖掘算法, 他是所有关联规则挖掘算法的核心[1]. 其主要思想是通过一种迭代的方法, 逐层搜索[2-4], 用(k-1)项集去搜索 k 项

① 基金项目:云南省高校商务智能科技创新团队基金

通讯作者:余建坤, E-mail: yjk1102@163.com

收稿时间:2014-11-29;收到修改稿时间:2015-02-11

集. 但 Apriori 算法在执行效率上存在以下缺陷[5,6]: 1)在频繁项集自连接的过程中, 产生大量的候选项集; 2)每次验证候选项集的时都要扫描数据库, 将要耗费大量的时间. 本文针对 Apriori 算法存在的缺陷进行了改进. 该算法将交易数据库映射到一个布尔矩阵中, 之后不再依赖数据库, 在矩阵列向量进行“与”运算之后, 删除那些没有意义的项和记录, 逐渐减少矩阵的行和列. 改进后的算法只扫描数据库一次, 并且矩阵越来越小, 计算效率也大大的提高. 将改进的 Apriori 算法应用到社交网络好友推荐中, 根据用户关注的用户和信息作为交易记录, 关注用户作为交易项, 构建交易数据库, 由于社交网络用户与被推荐的好友之间存在一种二元关系, 只需要计算出的频繁 2 项集, 推荐按支持数排序的前 N 位用户作为好友. 通过实验验证, 改进的 Apriori 算法在社交网络好友推荐算中具有较高的准确率和召回率.

1 改进的Apriori算法

针对 Apriori 算法存在的缺陷, 文献[7]以构建向量矩阵为基础, 提出了一种改进的 Apriori 算法, 该算法将交易数据库映射到一个布尔向量矩阵, 各个向量的长度为交易记录数, 统计各个项目中“1”的数目作为支持数. 将大于或等于最小支持数的任意两个向量进行“与”运算, 直到结果不超过最小支持度, 则结束. 根据社交网络中好友推荐的特性, 在使用关联规则挖掘时, 只需要求解频繁 2 项集. 此算法存在以下缺点: 1) 在算法开始阶段, 构造的布尔向量矩阵会比交易数据库还要大; 2) 在矩阵向量进行“与”运算的时候, 有些没有意义的记录. 这两个缺点必然会导致算法时间复杂度和空间复杂度的增加.

针对以上缺陷, 本文提出一种改进算法. 在产生频繁 k 项集之后, 根据支持数的大小去掉那些非频繁项集, 以免再次组合生成候选项集; 去掉那些没有意义的交易记录, 他们在计算频繁 k+1 项集的时候不用再考虑计数. 此过程类似于在布尔矩阵中逐步的去掉行和列. 这样必然会减少算法的时间复杂度和空间复杂度.

先求频繁 1 项集 L_1. 扫描交易数据库 D, 以交易项集 Itemset 中的每一项作为列, 交易记录作为行, 得到布尔矩阵 $M[m\times n]$, 其中 m 为交易记录数, n 为 Itemset 中的项数. $M[i,j]$=1 表示对 i 个交易记录中含有交易项 j, 否则 $M[i,j]$=0. 统计矩阵 M 中每一列中“1”的总数, 当总数大于或等于 min_sup 时, 为频繁 1 项集; 再统计矩阵 M 中每一行中“1”的总数, 当总数小于或等于 1 时, 删除此记录.

计算频繁 k 项集(k>=2). 对 L_{k-1} 中的任意只有一个项不同的两列自连接, 得到 C_k ,统计 C_k 中每一列中“1”的总数, 当总数大于或等于min_sup时, 为频繁 k 项集; 再统计 C_k 中每一行中“1”的总数, 当总数小于或等于 1 时, 删除此记录. L_k 中的交易记录数一定不会比 L_{k-1} 中的记录数大.

改进算法伪代码如下:

```
/*
 *D:交易数据库
 *M: 布尔矩阵
 * C_k : 候选 k 项集矩阵
 * L_k : 频繁 k 项集矩阵
 *L:所有的频繁项集
 *min_sup: 最小支持数
*/
M=transaction(D);//扫描 D 得到 M
L_1 =find_frequent_1-itemsets(M);//计算频繁 1 项集
for(k=2 ; L_{k-1} !=null; ++k){
    C_k =apriori_gen( L_{k-1} );//计算候选项集
    Csum= C_k 的每一列和;
    if(Csum>=min_sup){
        为频繁项;
        add to  L_k ;
    }
    Rsum= C_k 的每一行和;
    if(Rsum<=1)
        删除行;
    得到 L_k
}
L=频繁项集;
apriori_gen( L_{k-1} ){
    if( L_{k-1} 中两列只有一项不同){
        c=两列进行与运算.
        add c to  C_k ;
    }
    return  C_k ;
}
```

例. 表 1 为一个交易数据库 D, 设置 min_sup=2, 计算频繁项集 L.

表 1　交易数据库 D

TID	Itemset
1	A、C、D
2	B、C、E
3	A、B、C、E
4	B、E
5	A

1) 扫描数据 D, 得到布尔矩阵 M. 如表 2 所示.

表 2　布尔矩阵 M

TID	A	B	C	D	E
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1
5	1	0	0	0	0

2) 计算频繁 1 项集. 对 M 进行列向计数, 得到频繁 1 项集{{A}, {B}, {C}, {E}}, 再对 M 进行行计数, 将总数小于或等于 1 的记录删除, 得到 L_1, 如表 3 所示.

表 3　频繁 1 项集 L_1 矩阵

TID	A	B	C	E
1	1	0	1	0
2	0	1	1	1
3	1	1	1	1
4	0	1	0	1

3) 计算频繁 2 项集. 由 L_1 自连接得到 C_2, 如表 4 所示.

表 4　候选 2 项集 C_2 矩阵

TID	AB	AC	AE	BC	BE	CE
1	0	1	0	0	0	0
2	0	0	0	1	1	1
3	1	1	1	1	1	1
4	0	0	0	0	1	0

对 C_2 进行列计数, 得到频繁 2 项集{{AC}, {BC}, {BE}, {BE}}, 在对 C_2 进行行计数, 将总数小于或等于 1 的记录删除, 得到 L_2, 如表 5 所示.

表 5　频繁 2 项集 L_2 矩阵

TID	AC	BC	BE	CE
1	1	0	0	0
2	0	1	1	1
3	1	1	1	1

4) 根据 L_2 计算频繁 3 项集. 如表 6 所示.

表 6　频繁 3 项集 L_3 矩阵

TID	BCE
3	1

5) 根据 L_3 计算频繁 4 项集. 而 L_4 为空, 计算结束. 所以 L={{B}, {C}, {E}, {BC}, {BE}, {BCE}}.

改进后的算法将交易数据库映射到布尔矩阵中, 在矩阵进行运算后, 删除那些没有意义的交易项和交易记录, 不但提高算法的时间复杂度, 空间复杂度也大大的提升.

2　社交网络好友推荐算法

目前, 著名的社交网络平台 Facebook 在“可能认识的好友”模块中利用 Friend-of-friend 算法[8]给用户推荐好友, 该算法主要是根据用户之间的共同好友数, 当共同好友数超过设定的阈值时, 则向用户推荐. 显然, 通过此方法并不一定能给用户推荐具有共同兴趣爱好的好友. 众所周知, 在添加好友时, 用户必定会关注另一位用户关注的“人”和“事”, 而 Friend-of-friend 算法只关注了“人”而忽略了“事”[9]. 在文献[9]中提出了一种基于关联规则的社交网络好友推荐算法, 该算法同时关注“人”和“事”, 将被关注的“人”和“事”作为一条交易记录, 而关注这个“人”或者“事”的用户作为一条交易项. 根据候选 2 项集, 按支持数的高低, 推荐 top-N 个用户作为推荐好友. 该算法在给用户推荐好友时, 要求用户必须在交易数据库中, 对于不在交易数据库中的用户来说, 将无法为其推荐好友.

本文对文献[9]算法进行改进, 对于不在交易数据库中的用户(如刚注册的用户), 推荐其关注用户可推荐的好友. 算法思想如下: 用户与被推荐的好友是一种二元关系, 因此在对数据库 D 进行挖掘时, 只需要找出频繁 2 项集 L_2. 将求出的频繁 2 项集 L_2 作为一个数据库 D_l. 如果用户 u_i 在数据库 D_l 中能够找到 $u_i \to u_j$ $(i \neq j)$, 那么可以推荐 u_j 作为 u_i 的一个好友好友, 或者推荐 u_i 作为 u_j 的好友, 如果 u_i 和多个 u_j 存在这种关系, 可以根据支持数的高低推荐前 N 个用户 top-N; 如果用户 u_i 在数据库 D_l 中找不到 $u_i \to u_j$ $(i \neq j)$, 那么可以找 u_i 关注的那个用户可推荐的好友. 算法主要步骤如下:

步骤 1: 根据社交平台中用户关注的“人”和“事”得到交易数据库 D;

步骤 2: 使用改进的 Apriori 算法计算出频繁 2 项集 L_2;

步骤 3: 将 L_2 存入到数据库 D_l 中;

步骤 4: 输入用户号 u_{id} 以及用户所关注的用户号 u'_{id};

步骤 5: 在数据库 D_l 中, 根据支持数的大小查询出 u_{id} 可推荐的好友. 如果没有 u_{id} 可推荐的好友, 则查询 u'_{id} 可推荐的好友;

步骤 6: 向用户 u_{id} 推荐前 N 个好友 top-N.

3 实验验证

3.1 实验数据与环境

实验数据是通过爬虫工具在新浪微博上抓取下来的. 为了让数据更具有说服力, 选取了 5000 条关注用户的数据以及 5000 条关注信息的数据. 通过 Extract、Transform、Load 对这 10000 条数据进行预处理[10], 得到交易数据库 D.

实验运行环境为 core i5 2.5G 双核 CPU, 4G 内存, JDK1.7. 操作系统采用 win7, 数据库使用 MySQL 5.7. 本文所有算法均使用 java 实现.

3.2 实验结果与分析

为了对比 Apriori 算法与改进的 Apriori 算法的效率. 选取最小支持数为 20, 分别选择数据集为 2000、4000、6000、8000 和 10000 进行测试, 测试结果如图 1 所示. 另外, 选取数据集为 5000, 分别取最小支持数为 10、20、30、40、50, 测试运行结果如图 2 所示.

根据图 1, 图 2 对比两个算法的执行时间, 可以看出改进的 Apriori 算法在执行效率上有明显的提高. 由图 1 可知, 当交易数据量比较大时, 改进算法的效率更加显著. 根据图 2, 改进算法的效率几乎是原算法的两倍. 改进后的算法将矩阵中没有意义的交易记录删除, 保证了矩阵的行列数都在减少, 这使得算法在时间复杂度和空间复杂度上都得到了优化.

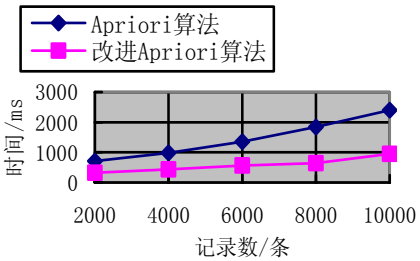


图 1 不同记录数下两种算法执行效率对比

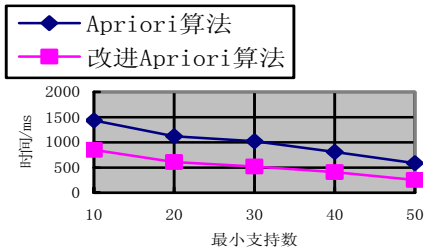


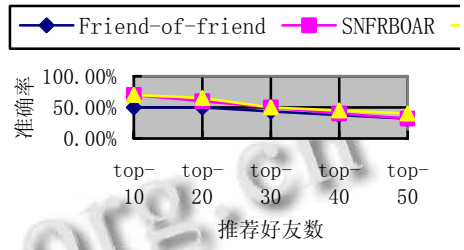
图 2 不同支持数下两种算法效率对比

采用准确率(Precision)和召回率(Recall)两个指标分别对 Friend-of-friend 算法、文献[3]中基于关联规则的社交网络好友推荐算法(SNFRBOAR)和本文提出的社交网络好友推荐算法(SNFRA)进行实验.

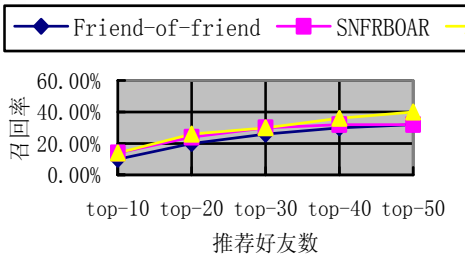
$$\text{准确率 (Pre}cision\text{)} = \frac{\text{算法推荐已成为好友数}}{\text{算法推荐的好友数}}$$

$$\text{召回率 (Re } call\text{)} = \frac{\text{算法推荐已成为好友数}}{\text{好友总数}}$$

设置 Friend-of-friend 算法的阈值和社交网络的好友推荐算法的 min_sup 都为 20. 根据推荐好友个数 top-N 的改变来进行测试对比. 实验结果如图 3 所示.



(a) 准确率



(b) 召回率

图 3 实验结果

通过图 3 结果分析可知, 本文提出的好友推荐算法在准确率和召回率上都比其他两个算法效率要好. 相比于 Friend-of-friend 算法, 本文提出的算法在构建交易数据库时, 不仅关注了人还关注了信息. 而

SNFRBOAR算法要求用户必须在交易数据库中, 对于刚注册的用户不适用. 从实验分析结果来看, 本文提出的好友推荐算法准确率比召回率明显要高, 这也符合现实, 因为推荐的好友“质量”比“数量”更重要.

4 结语

网上交友越来越普遍, 为用户推荐与其具有共同兴趣爱好的好友变得十分有意义. 本文针对 Apriori 算法存在的不足, 提出了一种改进的算法, 该算法将交易数据库映射到一个的布尔矩阵中, 在矩阵列向量进行“与”运算之后, 删除那些没有意义的交易记录. 改进的算法在时间复杂度和空间复杂度上都有很大的提高, 并将改进的算法运用到社交网络好友推荐中. 实验表明, 好友推荐算法在准确率和召回率上均有良好的效率.

参考文献

1 Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufman Publisher, 2001.

2 龙冰莹,陈小惠.改进 Apriori 算法在医院监护中心的研究与应用.计算机技术与发展,2013,23(8):137–140.

3 饶正婵,范年柏.关联规则挖掘 Apriori 算法研究综述.计算机时代,2012,30(9): 11–13.

4 屈展,陈雷.一种改进的 Apriori 算法在电子商务中的应用.西安石油大学学报(自然科学版),2012,27(1):91–98.

5 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large database. Proc. of the 1993 ACM SIGMOD Conference on Management of Data Table of Contents. New York. ACM. 1993. 207–216.

6 Agrawal R, Srikant R, Swami AN. Mining association rules. 20th International Conference on Very Large Data Bases. San Francisco. Margan Kaufmann. 1994. 487–499.

7 元文娟,晏杰.数据挖掘中关联规则 Apriori 算法.计算机系统应用,2013,22(4):121–124.

8 陈克寒,韩盼盼,吴健.基于用户聚类的异构社交网络推荐算法.计算机学报,2013,36(2):349–359.

9 向程冠,熊世桓,王东.基于关联规则的社交网络好友推荐算法.中国科技论文,2014,9(1):87–91,91–98.

10 张素琪,梁志刚,胡丽娟,董永峰. 改进的多维关联规则算法研究及应用.计算机工程与科学,2013,34(9):174–179.