Advances in Applied Mathematics 应用数学进展, 2021, 10(10), 3412-3420
Published Online October 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1010359

Hans汉斯

基于关联规则和多特征模型的微博好友推荐算法

丁　舟，胡建成

成都信息工程大学应用数学学院，四川　成都

收稿日期：2021年9月21日；录用日期：2021年10月14日；发布日期：2021年10月22日

摘　要

本文通过爬虫程序从新浪微博获得用户的个人信息、关注/粉丝列表信息和微博正文信息等，然后通过分析数据，基于用户社交相似度、用户兴趣相似度、用户间的地理相似度和用户影响力等因素向目标用户进行好友推荐。此外，还利用非目标用户的关注信息，采用关联规则分析技术，为目标用户推荐感兴趣的博主。本文采用的推荐策略经过数据验证，效果良好。

关键词

微博，好友推荐，关联规则，社交网络，数据挖掘

Microblog Friend Recommendation Algorithm Based on Association Rules and Multi Feature Model

Zhou Ding, Jiancheng Hu

College of Applied Mathematics, Chengdu University of Information Technology, Chengdu Sichuan

Received: Sep. 21st, 2021; accepted: Oct. 14th, 2021; published: Oct. 22nd, 2021

Abstract

This paper obtains the user's personal information, attention/fan list information and microblog text information from Sina Weibo through the crawler program, and then recommends friends to the target users based on the user's social similarity, user interest similarity, geographical similarity between users and user influence by analyzing the data. In addition, it also uses the attention information of non target users and association rule analysis technology to recommend in-

文章引用：丁舟，胡建成．基于关联规则和多特征模型的微博好友推荐算法[J]．应用数学进展，2021，10(10): 3412-3420. DOI: 10.12677/aam.2021.1010359

terested bloggers for target users. The recommendation strategy used in this paper is verified by data and has good effect.

Keywords

Microblog, Recommended by Friends, Association Rules, Social Networks, Data Mining

1. 引言

在这个大数据时代，海量数据充斥在微博上，除了寻找符合自己兴趣的内容外，寻找符合自己兴趣的内容发布者和寻找与自己兴趣一致的好友愈发重要。然而现在一些微博用户由于好友数较少，发布的微博曝光率低、互动少，导致信息不能有效传播、用户发布微博意愿降低等。因此，为用户推荐合适的好友能够促进信息传播、增强用户黏性等。如何从广大微博用户中推荐最符合某一用户的好友，已经成为各微博平台关注的问题。本文选取了微博正文、粉丝\关注列表、交互信息、地理位置、性别、注册时间等数据，利用关联规则和神经网络、决策树、回归 3 类机器学习算法综合推荐好友。

2. 相关研究

目前，社交网络好友推荐技术发展迅速，推荐技术的方法也不尽相同。推荐技术所用的因子主要有：用户个人信息(年龄、毕业院校、职业等)、用户兴趣爱好(根据兴趣标签或根据用户的行为推测用户的兴趣)、用户地理位置、用户社交关系。按推荐算法分类主要有：基于内容的推荐、基于关联规则的推荐、基于协同过滤的推荐、基于社交拓扑网络的推荐和混合推荐。

尹云飞等(2021) [1]提出基于认知度与兴趣度的融合好友推荐算法，通过反馈控制方法，动态修正推荐算法；马汉达等(2020) [2]提出了基于 SSD 和时序模型的微博好友推荐算法，构建了基于用户个人信息和图像信息的好友推荐方法，并在图像信息处理上加入时间维度因素；李雪(2018) [3]利用用户身份信息和地理位置信息计算相似度，还选取了用户发表的图片信息来构造兴趣向量；向程冠(2019) [4]将信息碎片相似度达到阈值的两个用户写入交易数据库，再进行关联规则分析，若存在关联规则用户 A→用户 B，则为用户 A 推荐用户 B。张红玉(2018) [5]利用隐语义模型来挖掘用户潜在兴趣特征向量和潜在交友偏好，由此计算相似度，并将此相似度归一化，作为用户在社交拓扑网络中随机游走的转移概率。周浩(2017) [6]引入了局部随机游走算法，结合信任度，计算出目标用户和拓扑网络上其他用户的相似度，构建好友推荐候选集，还提出了基于时间衰减兴趣分类好友推荐算法，并利用此算法计算候选集中的用户兴趣偏好向量，再计算相似度，得出推荐列表；姚彬修(2017) [7]提出了基于 canopy 和粗糙集的 CRS-KNN 文本分类算法，降低了 KNN 算法的数据计算规模，提高了分类效率，分别计算多源信息相似度，再引入时间权重和丰富度权重计算综合相似度。目前国内外对好友推荐有许多研究，但是只专注于单一特征。如何筛选出有效特征，以及如何将多个特征融合起来作推荐，是本文主要研究的问题。

3. 基于关联规则的博主推荐

关联规则(Association Rules)是反映一个事物与其他事物之间的相互依存性和关联性，是数据挖掘的

一个重要技术，用于从大量数据中挖掘出有价值的数据项之间的相关关系。一般来说，对于一个给定的交易事务数据集，关联分析就是指通过用户指定最小支持度和最小置信度来寻求强关联规则的过程。关联分析一般分为两大步：发现频繁项集和发现关联规则。

常用的频繁项集的评估标准有支持度，置信度和提升度。

支持度：几个关联的数据在数据集中出现的次数占总数据集的比重

$$\text{Support}(X \to Y) = P(XY)$$

置信度：一个数据出现后，另一个数据出现的概率，或者说数据的条件概率。

$$\text{Confidence}(X \to Y) = P(X|Y) = P(XY)/P(Y)$$

提升度：表示含有 Y 的条件下，同时含有 X 的概率，与 X 总体发生的概率之比。

$$\text{Lift}(X \to Y) = P(X|Y)/P(X)$$

用户不仅希望推荐好友，还希望推荐感兴趣的博主。因此，本文基于爬取的关注数据来进行关联规则分析，以得到“关注了 A 的用户往往会关注 B”的关联规则。若目标用户关注了关联规则左手规则里的博主，则为其推荐右手规则里的博主。

4. 基于多特征的微博好友推荐

4.1. 社交相似度

根据社交网络同质性理论，在社交网络中互相连接的人倾向于相似。这种相似包括身份同质性和价值同质性。这是因为人们通过社会地位的相似性或价值观的相似性进行选择。因此，我们可以通过分析社交拓扑网络情况来分析其相似性，从而利用该相似性为其进行好友推荐。根据微博社交网络的特性，引入出相似度和入相似度的概念。

出相似度(Similarity of Out-degree, OS)。目标用户 U_i，U_j 之间共同关注的用户数占总关注用户数的比值，代表用户间的兴趣相似程度。

$$OS(U_i, U_j) = \frac{|\text{Follow}(U_i) \cap \text{Follow}(U_j)|}{|\text{Follow}(U_i) \cup \text{Follow}(U_j)|}$$

入相似度(Similarity of In-degree, IS)。目标用户 U_i，U_j 之间的共同粉丝用户数占总粉丝用户数的比值，代表用户间微博社交关系的相似程度。

$$IS(U_i, U_j) = \frac{|Fans(U_i) \cap Fans(U_j)|}{|Fans(U_i) \cup Fans(U_j)|}$$

综合出入相似度的定义，给出推荐相似度(Recommendation Similarity, RS)的计算公式。

$$RS(U_i, U_j) = OS(U_i, U_j) + IS(U_i, U_j)$$

4.2. 用户影响力

在社交网络中，影响力高的用户往往是该社交网络的中心，人们更愿意跟随这种有影响力的用户。因此，向微博用户推荐好友时，需要考虑到所推荐用户的影响力。微博用户影响力由活跃度和博文质量构成。

1) 活跃度

在微博上，活跃的用户的影响力往往比不活跃的用户高。这些用户活跃地转发或发表微博，在一定程度上推动了微博信息的传播，具有一定的影响力。微博用户的活跃度通过用户最新发表的 N 条微博的时间跨度的反比来衡量。

$$\text{active}_i = \frac{N_t}{T}$$

其中，active_i 表示用户 i 的活跃度，N_t 表示用户 i 在时间段 T 内所发表的微博数量。

2) 博文质量

博文质量越高，该微博用户的粉丝数，以及所发表微博的互动数也就越高，用户的影响力也相应地越高。但是由于现在微博上存在大量虚假粉丝的情况，而且粉丝数是一个积累量，不能反映最新的用户影响力。所以，本文选择了用户发表微博的互动数来衡量用户的博文质量。

$$\text{quality}_i = w_1 \cdot avg_{\text{retweet}} + w_2 \cdot avg_{\text{comments}} + w_3 \cdot avg_{\text{like}}$$

其中 quality_i 表示用户 i 的博文质量，avg_{retweet}、avg_{comments}、avg_{like} 分别表示用户最新 N 条博文的平均转发、评论、点赞数，w_1、w_2、w_3 分别表示相应的权重。由于转发最能推动博文的传播，评论次之，故设定的转发数的权重最大。

4.3. 地理相似度

本文使用用户个人信息里的地区数据来计算用户间的地理相似度。首先将各地区数据表示成经纬度数据，若 A、B 两地的经纬度分别表示为 $(LonA, LatA)$，$(LonB, LatB)$，则两地距离为

$$\begin{aligned} &Distance(A,B) \\ &= 6371004 \cdot \arccos\left(\sin\left(RADIANS(LatA)\right)\right. \\ &\quad \cdot \sin\left(RADIANS(LatB)\right) + \cos\left(RADIANS(LatA)\right) \\ &\quad \left. \cdot \cos\left(RADIANS(LatB)\right) \cdot \cos\left(RADIANS(LonB - LonA)\right)\right) \end{aligned}$$

其中 RADIANS 函数将角度转换为弧度。

计算出所有目标用户的地区距离后，再进行归一化，再用 1 减去该数值，得到地理相似度。

4.4. 内容相似度

微博用户发表的博文内容在一定程度上反映了他们的兴趣，通过对微博用户的博文内容分析，可以得到他们的兴趣偏好。而兴趣相似的人更倾向于成为好友。内容相似度的计算方法如下。

1) 文本预处理

通过爬虫获取的微博正文信息中含有许多符号，例如微博表情爬下了之后变成了问号，还有话题标签“#”，提到某个用户“@”，转发标记“@”，以及标点符号。这些符号对分词结果无意义，为了提高分词效率，将其进行删除处理，以获得纯文本。

2) 中文分词

ANSJ 提供了多种分词方式，其中 NlpAnalysis 分词方式的功能最多，特别地，它还具有新词发现功能，对于微博这种网络新词层出不穷的区域，新词发现功能无疑是最满足要求的。所以本文使用的是 NlpAnalysis 分词方式。

3) 停用词过滤

在微博正文中，有许多无实际意义的词，例如语气词、介词、助词等。这些词的存在会影响文本特征词的提取，所以要对其进行剔除。Ansj 分词结果对词性进行了标注，因此可以利用词性来过滤停用词。本文去除了语气词、拟声词、助词、叹词、数词、量词、数量词、代词、副词。另外，本文还生成了停用词表，去除了诸如“转发”、“微博”、“图片”等词。

4) 提取特征词

本文使用 TF-IDF 方法来计算关键词表的权重信息。

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，$n_{i,j}$ 为特征词 i 在微博正文 d_j 中出现的频率，$\sum_k n_{k,j}$ 则是文本 d_j 中特征词的数量。计算的结果即为某个特征词的词频。

$$IDF = \log \frac{|D|}{1+|D_i|}$$

其中，$|D|$ 表示微博正文数据集中微博文本的总数，$|D_i|$ 表示微博正文中特征词 i 出现的频率。考虑到某词语可能在该微博正文中不存在的情况，使用 $1+|D_i|$ 作为分母，以避免为 0。

$$TF-IDF = TF \cdot IDF$$

对每个用户的 TF-IDF 值进行降序排序，保留前 20 个特征词作为该用户的兴趣特征词。

5) 计算相似度

把每一用户的兴趣特征词用向量表示，再利用余弦相似度计算出他们之间的内容相似度。

$$\text{Similarity} = \cos\theta = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

其中，A 和 B 分别是两个不同用户的兴趣特征词向量。

4.5 其他因素

1) 性别因素

性别因素不做相似度的计算，因为交友倾向不一定是性别相同。故作为单独的因素加入模型中。

2) 注册时间

注册时间只取年份数据，同样作为单独的因素加入模型中。

5. 实验及结果分析

首先本文利用爬虫软件抓取了新浪微博 161,740 条非目标用户的关注数据，2816 个目标用户数据。其中包括 42,257 条微博正文信息，41,365 条关注数据，18,751 条粉丝数据。

5.1. 基于关联规则的博主推荐

本文利用爬取的 16 万关注数据，以“原用户－关注用户”为一条事务数据，利用 SAS Enterprise Miner 进行关联规则分析。设置的参数如下：最小置信水平 = 10，最大项数 = 4，支持百分比 = 5。得到的部分规则如表 1。

Table 1. Partial results of association rules
表 1. 关联规则部分结果

规则
联想手机→膳魔师 THERMOS&盗墓笔记重启官微
知否知否应是绿肥红瘦官微→赵丽颖网宣特工队&赵丽颖工作室
丁钰琼→张子凡 Scofield-

5.2. 基于多特征的好友推荐

首先，计算出内容相似度和社交相似度。其次，筛选出社交相似度或内容相似度超过阈值的用户，加入待推荐列表中。

然后，计算出地理相似度、用户影响力，再加上性别信息和注册时间信息后，将数据导入 SAS EM，利用神经网络、决策树、逻辑回归模型分别进行建模。数据设置和流程图如表 2、图 1。目标变量为“是否为好友”。

Table 2. Data settings
表 2. 数据设置

标签	角色	水平
用户 I	ID	列名型
用户 J	ID	列名型
用户 I 的影响力	输入	区间型
是否为好友	目标	二值型
用户 I 的性别	输入	列名型
用户 J 的性别	输入	列名型
用户 I 的注册时间	输入	列名型
用户 J 的注册时间	输入	列名型
用户 J 的影响力	输入	区间型
地区相似度	输入	区间型

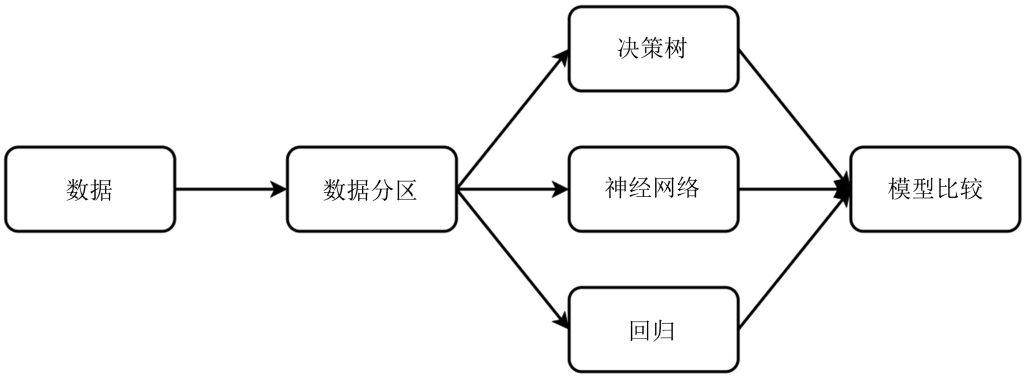

Figure 1. Flow chart
图 1. 流程图

模型结果如图 2，图 3，表 3：

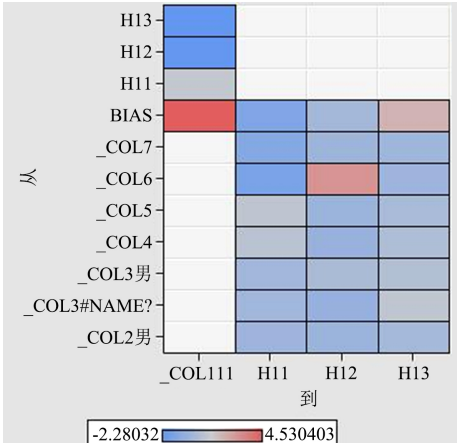

Figure 2. Neural network results
图 2. 神经网络结果

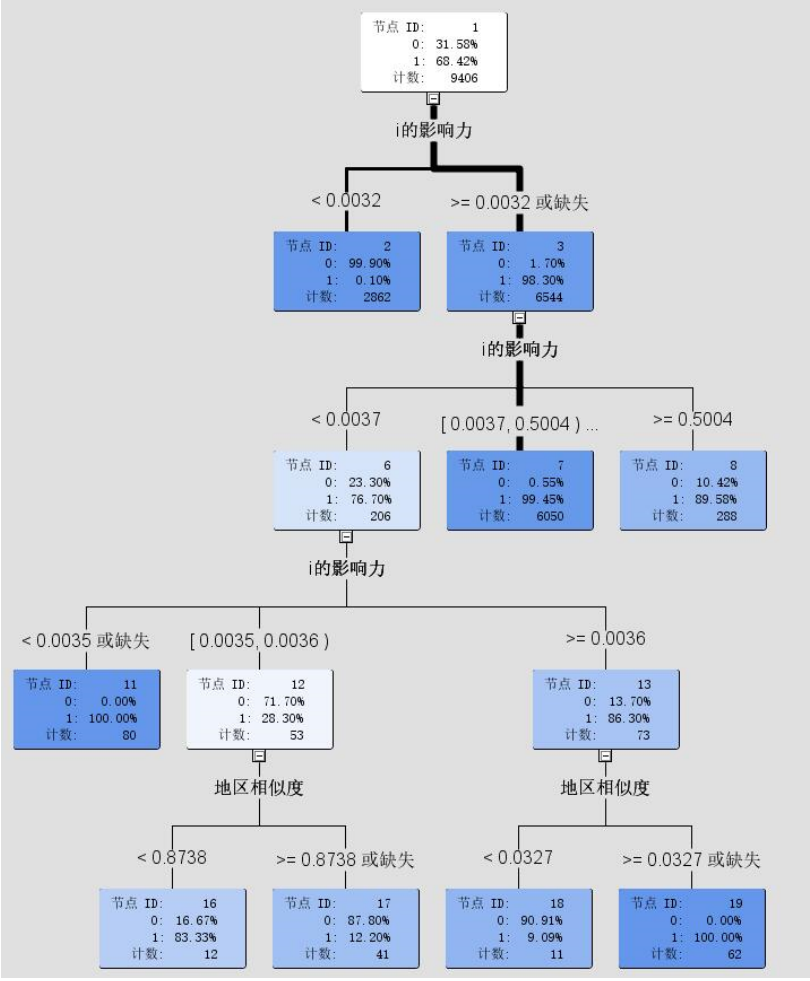

Figure 3. Decision tree results
图 3. 决策树结果

Table 3. Regression results
表 3. 回归结果

Parameter	DF	Estimate	Standard Error
Intercept	1	1.2960	0.0495
用户 j 的影响力	1	1.5919	5.6348
地区相似度	1	0.1702	0.0640

由模型结果可知，在微博好友推荐中，用户影响力比地区相似度更重要。

6. 策略评价

6.1. 评价指标

本文基于多特征的微博好友推荐结果以准确率(Precision)、召回率(Recall)为评价指标。准确率是正确推荐的好友数与推荐列表中好友总数的之比。

$$\text{Precision} = \frac{|R \cap T|}{|R|}$$

召回率是推荐列表中正确推荐的好友数与用户关注总数的之比。

$$\text{Recall} = \frac{|R \cap T|}{|T|}$$

其中，R 为推荐列表，T 为用户现有关注列表。

本文基于关联规则的博主推荐结果以提升度(Lift)和置信度(Confidence)作为评价指标。

$$\text{Confidence}(X \to Y) = P(Y|X)$$

$$\text{Lift}(X \to Y) = \frac{P(Y|X)}{P(Y)}$$

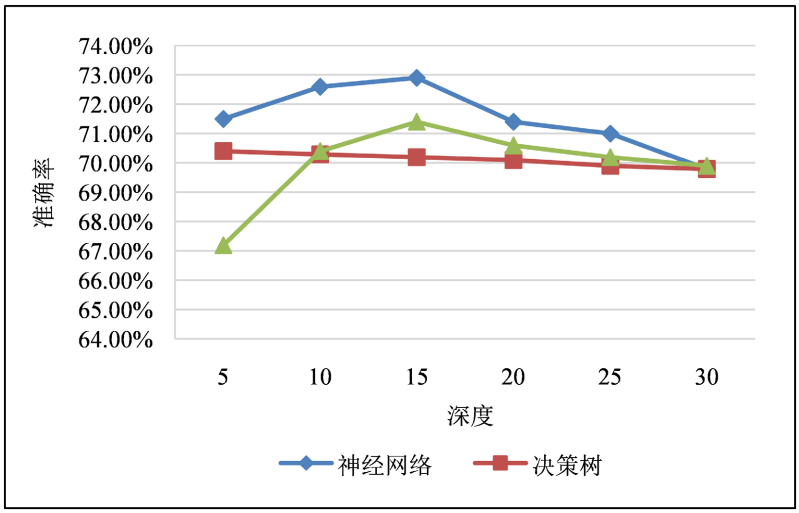

Figure 4. Model accuracy rate
图 4. 模型准确率

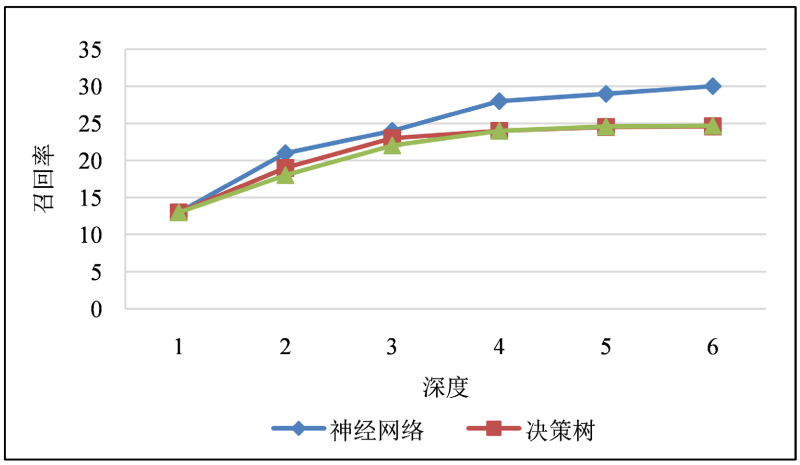

Figure 5. Model recall rate
图 5. 模型召回率

6.2. 结果评价

1) 基于关联规则的博主推荐的结果评价

一共得到 199 条关联规则。提升度最小为 22.94，最大为 50.36，平均为 27.28。置信度最高 87.50%，最低 39.66%，平均为 61.89%。关联规则的效果较好。

2) 基于多特征的微博好友推荐的结果评价

评价结果如图 4、图 5 所示。

可见，模型准确率都是先增加再减少，神经网络模型和回归模型在 15%的深度时达到最高准确率。结合准确率和召回率的评价结果，三种模型中神经网络的效果最好，因此本文选用神经网络作为多因素好友推荐的模型。

7. 总结

以新浪微博为代表的社交平台在中国有数亿活跃用户。大量用户产生大量数据，导致“信息过载”现象。用户花越来越多的时间搜索他们的朋友和他们感兴趣的东西。正是在这样的背景下，本文对好友推荐进行了综合研究。新浪微博平台可以利用该研究结果增强用户黏性，满足用户心理需求以及挖掘社交网络商业价值。

参考文献

[1] 尹云飞, 孙敬钦, 黄发良, 白翔宇. 基于认知度与兴趣度的好友推荐反馈算法[J]. 模式识别与人工智能, 2021, 34(2): 127-136.

[2] 马汉达, 景迪. 基于 SSD 和时序模型的微博好友推荐算法[J]. 计算机工程与科学, 2021, 43(7): 1291-1298.

[3] 李雪. 基于社交图片的精准好友推荐系统的研究与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2018.

[4] 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.

[5] 张红玉. 基于社交网络的好友推荐方法研究[D]: [硕士学位论文]. 天津: 天津理工大学, 2018.

[6] 周浩. 基于大数据分析的微博好友推荐算法研究与应用[D]: [硕士学位论文]. 北京: 北京工业大学, 2017.

[7] 姚彬修. 基于多源信息的个性化微博用户推荐算法研究[D]: [硕士学位论文]. 曲阜: 曲阜师范大学, 2017.