面向微博的个性化内容推荐算法研究*

王洪伟　段友祥
（中国石油大学(华东)计算机科学与技术学院　青岛　266580）

摘　要　随着信息共享时代的发展,海量数据的诞生对推荐系统提出了更高的要求。针对微博的海量数据,提出了一种融合朴素贝叶斯分类和基于用户的协同过滤算法的混合推荐算法。该算法将文本关键字作为特征属性,利用贝叶斯分类法筛选出用户可能感兴趣的数据,缩小推荐结果集;然后采用基于用户的协同过滤算法,通过计算用户相似度,根据最近邻居得到推荐结果列表。实验结果表明,混合推荐算法相比较于单一的推荐算法有着更高的准确率。

关键词　推荐算法;中文分词;贝叶斯分类;协同过滤
中图分类号　TP312　　**DOI**:10. 3969/j. issn. 1672-9722. 2022. 01. 029

Research on Personalized Content Recommendation Algorithm for Micro-blog

WANG Hongwei　DUAN Youxiang
(School of Computer Science and Technology, China University of Petroleum, Qingdao　266580)

Abstract　With the development of the era of information sharing, the birth of massive data puts forward higher requirements for recommendation systems. For the massive data of Micro-blog, a hybrid recommendation algorithm combining naive Bayes classification and user-based collaborative filtering algorithm is proposed. The algorithm uses text keywords as feature attribute, uses Bayesian classification to filter out data that users may be interested in, and narrows down the recommendation result set. Then it uses a user-based collaborative filtering algorithm to calculate user similarity and get the recommended result list based on nearest neighbor. Experimental results show that the hybrid recommendation algorithm has higher accuracy than the single recommendation algorithm.

Key Words　recommendation algorithm, Chinese word segmentation, Bayesian classification, collaborative filtering
Class Number　TP312

1　引言

随着互联网的发展,人们正处于一个信息爆炸的时代。相比于过去的信息匮乏,面对现阶段海量的信息数据,对信息的筛选和过滤成为了衡量一个系统好坏的重要指标。一个具有良好用户体验的系统,会将海量信息进行筛选、过滤,将用户最关注最感兴趣的内容展现在用户面前。这将大大提升系统工作的效率,也会节省用户筛选信息的时间。

微博作为当下信息传播的热门载体,针对微博信息的推荐也成为了当下研究的热点。李敬等[1]基于话题标签来挖掘出有价值的主题信息,有效挖掘出不同微博类型的主题分布;马慧芳等[2]提出了基于多标签关联关系的微博推荐算法,该算法通过挖掘被同一用户标注的多标签的内在关联以及被不同用户标注的多标签外在关联来构建用户的兴趣集;彭泽环等[3]在总结影响用户微博兴趣的基础上,应用潜在因素模型提出了社区热点微博推荐系统。上述研究中,在数据集预处理上均采用了传统的方法,导致用户数据的稀疏性问题没有得到充分解决。本文着眼于对微博数据集的预处理,并基于此提出一种混合推荐算法,基本思想是利用中文分词技术提取文本内容中的关键字,将关键字作为微博内容的特征属性,用于朴素贝叶斯分类。筛选出

* **收稿日期**:2021年6月3日,**修回日期**:2021年7月18日
作者简介:王洪伟,男,硕士研究生,研究方向:人工智能及其应用。段友祥,男,博士,教授,硕士生导师,研究方向:人工智能、图形图像处理、数据科学及应用和理论计算机科学。

与待推荐用户关联度高的微博数据后，通过基于用户的协同过滤算法，计算用户之间的相似度，求解最近邻居，计算得到Top-N的推荐列表。

2 相关研究

2.1 中文分词

针对文本数据的分析，需要将一段连续文字按照一定的规则重新组合成词序列，形成关键字词集合，即中文分词技术[4]，它是文本内容分析和挖掘的基础。

中文分词相较于英文而言，词与词之间缺乏明确的界限和分隔符，使得词组划分上存在不少技术上的困难。目前主流的中文分词算法有基于辞典的方法、基于统计的方法、基于规则的方法等，其中基于辞典的方法是按照一定策略将待分析的汉字串与一个“大机器词典”中的词条进行匹配，从而实现分词，并引入停用词的概念，对于一些功能词汇和常用词汇进行摒除，以获得更具代表性的关键字词集合作为标识句子的特征属性。本文使用基于java语言开发的轻量级的中文分词工具包-IK Analyzer[5]，通过实验表明，该分词工具包具有良好的分词效果。

2.2 朴素贝叶斯分类

文本分类技术是组织和管理文本信息的重要和有效手段。利用分词得到的文本关键字、关键词组等属性，采用分类技术筛选数据集中与用户感兴趣的微博(包括用户发表、转发、点赞、评论等)具备强关联关系的微博数据，排除掉大量不相关数据的干扰，以期在推荐算法中得到更好的推荐效果。

主流的分类方法有基于统计的分类方法、基于规则的分类方法和基于连接的分类方法等。朴素贝叶斯分类[6-7]是一种稳定、简单、高效的统计分类方法，在分类问题中被广泛应用。

朴素贝叶斯分类是基于贝叶斯定理与特征条件独立假设的分类法。设$X=\{x_1,x_2,x_3,\cdots,x_m\}$为一个待分类项，$x_i$为$X$的一个特征属性。存在待分类集合$Y=\{y_1,y_2,y_3,\cdots,y_n\}$，其中$P(y_k|X)$表示$X$在分类$y_k$下的概率。

已知贝叶斯公式(1)：

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

要求解X在给定任一分类下的概率$P(y_1|X)$，$P(y_2|X)$，…，$P(y_n|X)$时，由上述贝叶斯公式(1)可知$P(X)$相等，所以只需计算出$P(Y|X)=P(X|Y)P(Y)$，即可根据概率最大值$P(y_k|X)=MAX\{P(y_1|X),P(y_2|X),\cdots,P(y_n|X)\}$得到物品的最终分类结果。

本文中$P(X|Y)$为文本分词后得到的关键字/词组在分类文档中出现的概率，$P(Y)$为某分类下文档数目占数据集总文档数目的比例。

2.3 协同过滤算法

协同过滤是目前应用最广泛的推荐算法，它仅仅通过了解用户与物品之间的关系进行推荐，而不需要考虑到物品本身的属性，该类方法主要有两类：一类是基于用户的协同过滤推荐算法[8-9]，其核心思想是根据用户对物品的偏好计算相似度，找到相邻邻居用户，然后将邻居喜欢的物品推荐给当前用户。另一类是基于物品的协同过滤算法[10-11]，根据用户对物品的偏好找到相似的物品，然后根据用户的历史偏好，推荐相似的物品。两类算法有着鲜明的优缺点和适应场景。其中，基于用户的协同过滤算法适用于用户较少的场合，如果用户很多，计算用户相似度矩阵代价很大，当用户产生新行为时，不一定造成推荐结果的立即变化，具有较强的时效性；基于物品的协同过滤算法适用于物品数明显小于用户数的场合，如果物品特别多，计算物品相似矩阵的代价也会很大，当用户产生新行为时，一定会导致推荐结果的实时变化，适用于用户个性化需求强烈的区域。

根据本文的应用特点，选择采用基于用户的协同过滤推荐算法。用$U=\{u_1,u_2,u_3,\cdots,u_n\}$代表用户集合，用$T=\{t_1,t_2,t_3,\cdots,t_m\}$代表微博文本内容集合，用$S$代表评分项$s_{i,j}$的$n*m$评分矩阵，其中$i\in\{1,2,\cdots,n\}$，$j\in\{1,2,\cdots,m\}$。

通过皮尔逊相关系数计算用户a和用户b之间的相似性，如式(2)。

$$Sim(a,b)=\frac{\sum_{t\in T}(s_{a,t}-\bar{s}_a)(s_{b,t}-\bar{s}_b)}{\sqrt{\sum_{t\in T}(s_{a,t}-\bar{s}_a)^2}\sqrt{\sum_{t\in T}(s_{b,t}-\bar{s}_b)^2}} \tag{2}$$

其中，$s_{a,t}$代表用户a对文本t的评分，$\bar{s}_a$代表用户a对其关联微博文本数据的平均评分，$s_{b,t}$代表用户b对文本t的评分，$\bar{s}_b$代表用户b对其关联微博文本数据的平均评分。用户对其关联微博文本数据的评分基于用户的行为转化而来，众所周知，微博用户可以发表微博、对微博进行点赞、评论、转发等操作，基于不同的动作，本文在对微博内容预处理时为其赋予了不同的评分权重，从而将用户行为转化为量化评分，得到用户、微博内容、评分的三元矩阵。

计算出用户之间的相似度后，挑选相似邻居时

采用了基于相似度门槛的邻居选取法,以当前点为中心,选取距离为K区域中的所有点作为当前点的邻居,该法相较于固定数量邻居选取法,得到的相似邻居数目不固定,但是此法排除了孤立点对于邻居选取的干扰,避免了相似度出现较大的误差。根据邻居的相似度权重以及他们对物品的偏好,预测当前用户可能感兴趣的未涉及物品的评分,根据评分高低得到推荐列表,选取Top-N推荐给当前用户。其中,评分预测公式如式(3)。

$$Pred(a,\ t)=\overline{s_a}+\frac{\sum_{b\in N} Sim(a,\ b)*(s_{b,\ t}-\overline{s_b})}{\sum_{b\in N} Sim(a,\ b)} \tag{3}$$

此公式预测用户a对微博内容t的评分,其中Sim(a,b)代表用户a和用户b的相似度,N代表用户a的最近邻居。

3 混合推荐算法

混合推荐算法包括预处理、分类、推荐三部分。

3.1 预处理

对于数据集的预处理[12],目的是将原始数据转化为蕴含一定规则的结构化数据,以便于算法处理。

微博中能反映用户行为的数据分为以下四类:用户发表的微博、用户点赞的微博、用户评论的微博、用户转发的微博。本文使用的微博数据集中包含用户名、微博内容、用户行为等信息,用户行为中,定义publish表示发表,like表示点赞,forward表示转发,comment表示评论,定义$score_{i,j}$为用户i对文本数据j的评分。评分计算公式如式(4)。

$$score_{i,\ j}=a*publish_{i,\ j}+b*like_{i,\ j}+c*forward_{i,\ j}+d*comment_{i,\ j} \tag{4}$$

其中,$publish$,$like$,$forward$,$comment$取值为0或者1,代表该行为存在或者不存在。a,b,c,d代表一组给定的常数,根据经验,用户发表微博的权重应当大于转发微博,转发的权重应当大于评论,评论的权重应当大于点赞,因此本文按照权重大小给定一组常数值,即可计算得到用户i对其关联微博j的评分数据。

现行评分制系统中,大多采用5分制,本文沿用此评分机制,引入式(5)将上述评分结果转化为5分制。

$$score_{i,\ j}=round\left(\frac{score_{i,\ j}}{score_{\max}^{old}}*score_{\max}^{new}\right) \tag{5}$$

其中,$round$函数表示就近取整,$score_{\max}^{old}$表示旧的评分机制中的最大值,$score_{\max}^{new}$表示新的评分机制中的区间上限(本文采用5)。

3.2 分类

预处理完数据集后,得到用户、微博文本数据、评分值的三元矩阵(u,t,s),本文将数据集中获得的评分数据1~5分作为五个分类,将用户相关的微博数据,包括用户发表、点赞、评论、转发的微博作为训练集,计算数据集中其他微博内容所属的分类,其中分类为3、4、5的微博数据与当前用户感兴趣的微博具有强相关性,关键字、词的匹配程度较高,则将该三个分类下的微博数据结合用户训练集作为协同过滤推荐算法的最终数据集。

本文中贝叶斯算法流程如图1所示。

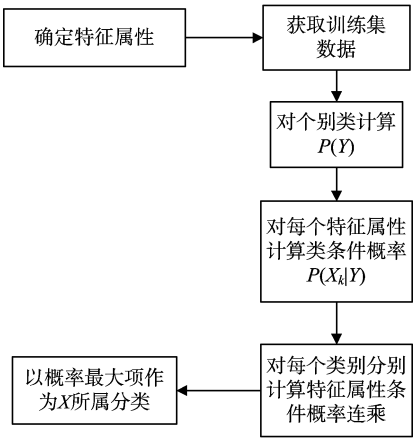

图1 贝叶斯分类流程图

其中,训练集数据中,每个分类下包含若干文档,文档中包含该用户关联的微博内容。计算P(特征属性|分类)时,通过字符串匹配的方式计算出某文档中是否包含该特征属性关键字/词组,给定包含关键字/词组的文本数Nxc,其初始值为0,若文档中包含该关键字/词组,则$Nxc+1$,分类下所有的文档数目Nc为一固定值,因此所得类条件概率P(特征属性|分类)$=Nxc/Nc$,考虑到可能存在的情况是:训练集中,多样本的取值可能并不在其中,但是这并不代表这种情况发生的概率为0,换言之:未被观测到,不代表不会发生。因此引入拉普拉斯修正,即式(6)被修正为式(7):

$$p(c)=\frac{D_c}{D} \tag{6}$$

$$p(c)=\frac{D_c+1}{D+N} \tag{7}$$

其中,N为分类总数,得到本文中拉普拉斯修正后的公式为P(特征属性|分类)$=Nxc+1/Nc+V$,V代表分类总数,本文中$V=5$。

贝叶斯分类通过中文分词法提取文本中的关

键字/词组作为特征属性，计算出每一个特征词在特定分类中的概率，因为特征属性之间相互独立，将计算得到的每一个特征属性的类条件概率相乘后再乘以此分类在数据集文本中所占的比例–先验概率，即可计算得到后验概率，也就是文本属于某个特定分类下的概率。

通过贝叶斯分类法将计算属于3、4、5分类下的微博数据保留并结合待推荐用户的训练集数据形成新的数据集，后续的协同过滤算法在此数据集基础上进行微博个性化内容推荐。

3.3　推荐

在微博推荐内容领域[13]，每天都会产生海量的实时数据，相比较而言，用户的数目是相对固定的，因此本文选用基于用户的协同过滤算法进行微博内容的个性化推荐。针对筛选后的数据集，数据内容由三元矩阵（微博用户，微博内容，用户评分）组成，算法流程如图2所示。

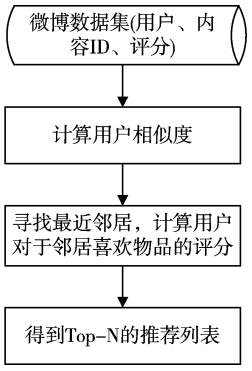

图2　基于用户的协同过滤算法流程图

本文在实验中使用Apache Software Foundation（ASF）旗下的一个开源项目–Mahout来实现基于用户的协同过滤推荐算法，算法中对于用户和物品的处理是基于long类型数据处理的，因此需要对三元矩阵中的微博用户和微博内容进行再处理，策略如下：

1）构造微博用户ID发号规则，设定发号规则为5位数字，从“00001”开始顺番发号，将该映射关系记录到数据库表中。

2）同理，可构造微博内容ID的发号规则，设定发号规则为6位数字，从“000001”开始顺番发号，同样将该映射关系记录到数据表中。

将微博用户名和微博内容分别与long类型数据映射完毕后，得到如下格式的三元矩阵数据集：

00001，000001，3

00001，000002，5

……

00010，000001，3

00010，000010，5

……

Mahout提供的基于用户的协同过滤算法中，通过传入的用户ID和待推荐的物品数量计算得到推荐列表–微博内容ID，通过关联数据表匹配到最终的待推荐微博内容列表。

4　实验及分析

本文使用的微博数据集来源于爬虫抓取的实时微博数据，包含32004个用户的160076条微博内容，其中微博内容中包含相关用户行为。

本文根据实验结果设定当评分值大于2时，代表用户对该微博内容感兴趣；当评分值小于等于2时，代表用户对该微博内容不感兴趣。实验进行N（本文N = 10）组，每组随机取100名不同用户分别计算得到前100条推荐结果，计算推荐准确率，即推荐列表中评分值 > 2的微博内容占推荐列表的比例。针对单一的基于用户的协同过滤推荐算法和本文的混合推荐算法所得的准确率结果（百分比）统计如表1所示。

表1　单一和混合算法所得的准确率结果统计

算法	N=1	N=2	N=3	N=4	N=5
单一	86.4	88.2	87.3	80.7	92.1
混合	92.5	96.3	94.2	91.7	97.5

N=6	N=7	N=8	N=9	N=10	平均
85.9	90.4	85.6	84.7	83.7	86.5
93.0	94.5	91.9	92.4	90.1	93.4

综合实验结果的准确率统计表明，混合算法的平均推荐结果的平均准确率为93.4%，而单一的基于用户的协同过滤算法平均准确率只有86.5%，混合算法[14~15]的平均推荐准确率提高了7%左右，实验表明混合算法在提高推荐精度上有着良好的效果。

5　结语

单一的推荐算法一般情况下存在的缺点比较明显，混合算法的优势在于取长补短，通过结合其他算法来弥补单一算法中的不足，从而实现推荐准确率的提高。不同的场景下，存在不同的数据结构特点，需要根据不同的场景设计不同的推荐算法，以求达到最理想的推荐效果。本文基于微博数据的个性化结构，结合贝叶斯分类与协同过滤推荐提出了混合推荐算法，实验数据表明，该算法在微博内容推荐上实现了比单一算法更高的准确率。

参考文献

[1] 李敬, 印鉴, 刘少鹏,等. 基于话题标签的微博主题挖掘[J]. 计算机工程, 2015, 41(4):30-35.
LI Jing, YIN Jian, LIU Shaopeng, et al. Topic mining of microblog based on topic tagging [J]. Computer Engineering, 2015, 41 (4): 30-35.

[2] 马慧芳, 贾美惠子, 李晓红, 等. 一种基于标签关联关系的微博推荐方法[J]. 计算机工程, 2016, 42(4): 197-201.
MA Huifang, JIA Meihuizi, LI Xiaohong, et al. A microblog recommendation method based on tag association relationship [J]. Computer Engineering, 2016, 42 (4): 197-201.

[3] 彭泽环, 孙乐, 韩先培, 等. 社区热点微博推荐研究[J]. 计算机研究与发展, 2015, 52(5): 1014-1021.
PENG Zehuan, SUN Le, HAN Xianpei, et al. Community hot microblog recommendation research[J]. Computer Research and Development, 2015, 52 (5): 1014-1021.

[4] 许高建, 胡学钢, 王庆人. 文本挖掘中的中文分词算法研究及实现[J]. 计算机技术与发展, 2007(12): 128-130,178.
XU Gaojian, HU Xuegang, WANG Qingren. Research and implementation of Chinese word segmentation algorithm in text mining [J]. Computer Technology and Development, 2007 (12): 128-130,178.

[5] 柴洁. 基于IKAnalyzer和Lucene的地理编码中文搜索引擎的研究与实现[J]. 城市勘测, 2014(06):50-55.
CHAI Jie. Research and implementation of Chinese geocoding search engine based on ikanalyzer and Lucene [J]. Urban Survey, 2014 (06): 50-55.

[6] 史琬莹. 朴素贝叶斯方法在文本分类中的运用[J]. 电子技术与软件工程, 2018,133(11):208.
SHI Wanying. Application of naive Bayesian method in text classification[J]. Electronic Technology and Software Engineering, 2018,133(11): 208.

[7] 张亚萍, 陈得宝, 侯俊钦, 等. 朴素贝叶斯分类算法的改进及应用[J]. 计算机工程与应用, 2011, 47(15): 134-137.
ZHANG Yaping, CHEN Debao, HOU Junqin, et al. Improvement and application of naive Bayesian classification algorithm [J]. Computer Engineering and Application, 2011, 47(15): 134-137.

[8] 王成, 朱志刚, 张玉侠, 等. 基于用户的协同过滤算法的推荐效率和个性化改进[J]. 小型微型计算机系统, 2016, 37(03):30-34.
WANG Cheng, ZHU Zhigang, ZHANG Yuxia, et al. Recommendation efficiency and personalized improv- ement of user based collaborative filtering algorithm [J]. Minicomputer System, 2016,37 (03): 30-34.

[9] 邱爽, 葛万成, 汪亮友, 等. 个性化推荐中基于用户协同过滤算法的优化[J]. 信息技术, 2016(03):70-71, 80.
QIU Shuang, GE Wancheng, WANG Liangyou, et al. Optimization of user collaborative filtering algorithm in personalized recommendation [J]. Information Technology, 2016 (03): 70-71,80.

[10] 汪静, 印鉴, 郑利荣, 等. 基于共同评分和相似性权重的协同过滤推荐算法[J]. 计算机科学, 2010(02): 105-110.
WANG Jing, YIN Jian, ZHENG Lirong, et al. Collaborative filtering recommendation algorithm based on common score and similarity weight [J]. Computer Science, 2010 (02): 105-110.

[11] 石京京, 肖迎元, 郑文广. 改进的基于物品的协同过滤推荐算法[J]. 天津理工大学学报, 35(01):35-39.
SHI Jingjing, XIAO Yingyuan, ZHENG Wenguang. Improved collaborative filtering recommendation algorithm based on items [J]. Journal of Tianjin University of Technology, 35 (01): 35-39.

[12] 孔钦, 叶长青, 孙赟. 大数据下数据预处理方法研究[J]. 计算机技术与发展,2018,28(5):7-10.
KONG Qin, YE Changqing, SUN Yun. Research on data preprocessing method under big data [J]. Computer Technology and Development,2018,28(5):7-10.

[13] Zhang M, Wang S, Wang Z. 个性化微博推荐算法[J]. 计算机科学与探索, 2012(10):895-902.
ZHANG M, WANG S, WANG Z. Personalized microblog recommendation algorithm [J]. Computer Science and Exploration, 2012 (10): 895-902.

[14] 郭艳红, 邓贵仕. 协同过滤系统项目冷启动的混合推荐算法[J]. 计算机工程, 2008(23):17-19.
GUO Yanhong, DENG Guishi. Hybrid recommendation algorithm for cold start of collaborative filtering system project [J]. Computer Engineering, 2008 (23): 17-19.

[15] 何丽, 李熙伟. 基于朴素贝叶斯与协同过滤的分布式推荐模型研究[J]. 北方工业大学学报, 2017(05): 96-102.
HE Li, LI Xiwei. Research on distributed recommendation model based on Naive Bayes and collaborative filtering [J]. Journal of North University of Technology, 2017 (05): 96-102.