文章编号：1003-0077(2019)07-0128-08

基于相似主题和 HITS 的微博用户推荐算法研究

王嵘冰，徐红艳，冯 勇，安维凯

（辽宁大学 信息学院，辽宁 沈阳 110036）

摘 要：为了准确地为微博用户推荐相近兴趣领域的重要用户，有效提高用户对微博平台的依赖度。该文对传统的 HITS 算法进行了改进：通过分析微博用户社交网络结构，运用改进算法将微博用户划分为 3 类，在微博主题相似度计算中引入用户的权威度和中心度，最后根据用户类别进行微博用户推荐。实验中，使用爬取的微博数据对传统的推荐算法和该文的改进算法进行对比实验，由于所提算法在分析过程中考虑了用户结构信息、用户的权威度与中心度等多种因素，因而在准确率、召回率、F1 值上均有明显提高。

关键词：微博用户推荐；HITS；权威度；中心度；主题相似度

中图分类号：TP311 **文献标识码**：A

Microblog User Recommendation Algorithm Based on Similar Topics and HITS

WANG Rongbing, XU Hongyan, FENG Yong, AN Weikai

(College of Information, Liaoning University, Shenyang, Liaoning 110036, China)

Abstract: To recommend important users in similar interest areas for micro-blog users, the improved HITS method is used to classify user categories based on the analysis of the micro-blog users' network structure. Since the user's authority and centrality is already introduced into micro-blog topic similarity calculation, the micro-blog users are recommended according to the category of users. Using the crawled micro-blog data, the proposed algorithm has significant improvement compared with the traditional recommendation algorithms.

Keywords: micro-blog user recommendation; HITS; authority; centrality; topic similarity

0 引言

Web 2.0 时代的到来，使各类应用于社交的新媒体不断涌现。目前应用较为广泛的社交新媒体如：新浪微博、Facebook、Twitter 等。使用者通过这些新媒体可以得到自己感兴趣的资源与信息。国内社交新媒体的领军者新浪微博因用户群逐渐壮大而导致微博内容也随之激增，“信息迷航”问题[1]日益严峻。个性化推荐技术被认为是解决该问题的有效手段[2]。

本文在改进原始 HITS(hyperlink-induced topic search)算法时将微博用户划分为不同的类别，在计算微博主题相似度时融入中心度及权威度，以达到提升推荐结果准确率的目的；对微博用户间的社交网络关系以及用户发布的微博主题内容进行分析，进而将兴趣、爱好相近的用户推荐给使用者。用户可以有效地利用这些社交平台提供的推荐功能来发现新朋友，进而使用户对社交平台产生依赖性，提高忠诚度。庞大的用户数量可以增强该社交平台的社会影响力，继而为平台带来可观的经济效益。

微博拥有数量庞大的用户群，只利用微博主题相似性去发现符合需求的用户推荐是难以实现的。本文以提高用户推荐准确度为目的，提出了一种基于相似主题和 HITS 的微博用户推荐算法。算法首先根据微博用户的粉丝数量、原创微博数量、转发微博数量来划分用户类别；然后，将改进后的 HITS 算法应用到微博用户权威度以及中心度的计算中。在本文中将重要用户划分为专家用

收稿日期：2018-07-18 **定稿日期**：2018-10-15

基金项目：国家自然科学基金(71771110)；中国博士后科学基金(2018M631814)；教育部重点实验室资助项目(93K172018K01)；辽宁省社科规划基金(L18AGL007)

户(Authority user)—Authority(权威度)值高、中枢用户(Hub user)—Hub(中心度)值高；最后按类别计算用户间的微博主题相似性，并进行兴趣相似微博用户的推荐。

本文的内容组织如下：第 1 节是相关工作的介绍；第 2 节介绍基于相似主题的微博用户推荐算法框架；第 3 节对改进的推荐算法核心环节进行详细介绍；第 4 节介绍实验环节的设计及实验结果分析；第 5 节为本文的总结。

1　相关工作

1.1　微博重要用户发现方法研究

随着学者们对微博重要用户发现方法的深入展开，有研究者在推荐算法中运用微博主题相似性的计算来提升推荐质量[3]。代表性研究成果有：仲兆满等[4]选择分别来自文化、企业管理、军事、时尚、教育 5 个领域的微博数据，利用这些数据来挖掘用户兴趣、计算用户相似度，但算法没有考虑到用户兴趣领域中的微博用户可分为不同的类型：有的以原创为主，有的以转发为主，还有的以浏览为主。本文将上述三种类型用户命名为：专家用户、中枢用户及普通用户。缺乏对用户所属类型的分析是导致推荐结果准确性不高的原因之一。姚彬修等[5]先对微博用户实施类型划分，接着对微博用户多源信息相似度进行计算，在计算时引入时间权重值和丰富度权重值。但该算法缺乏对微博用户之间社交网络结构的分析，因而该算法的推荐准确率依旧有待提高。彭泽环等[6]在设计推荐算法时考虑如下四类信息：个人信息、社交网络结构信息、交互信息以及微博主题内容信息，但此推荐算法缺少对微博主题权重的进一步区分，因而推荐结果质量提升有限。

1.2　HITS 算法

分析网页重要度的 HITS 算法是由美国康奈尔大学的 Jon Kleinberg 教授提出。该算法通过分析网页的链接关系来计算每个页面的 Hub 属性值和 Authority 属性值，并根据值的大小将网页分为 Hub 页面和 Authority 页面。

目前，HITS 算法被广泛地应用在搜索引擎、自然语言处理、社交分析等领域。吴树芳等[7]针对传统的 HITS 算法，提出在计算微博用户可信度时融入博文内容以及用户交互行为的改进 HITS 算法。喻依等[8]通过 HITS 算法计算期刊的权威度和中心度来反映期刊的权威性和中心性，产生更权威性的期刊排序。苗家等[9]首先基于特征计算出评论的权重，然后结合图模型使用 HITS 算法得到正文句子权重，进而得到文摘句。上述算法都是使用 HITS 算法来计算链接图中节点的权威度和中心度。由于原始 HITS 算法仅仅考虑用户之间的链接关系，因而存在主题偏移、容易作弊等缺陷。

而本文所研究的微博用户之间的关注模型与网页的链入、链出模型很相似，因此建立链接图时，图中的节点为微博用户，图中的有向边为微博用户间的关注关系，通过改进 HITS 算法来准确计算微博用户的权威度和中心度，有效地解决了原始 HITS 算法的不足，从而提高微博用户推荐的准确率。

2　基于相似主题的微博用户推荐算法框架

随着微博用户数量的激增，将导致了多种用户类型的出现。例如：微博认证用户以及普通微博用户。在社交平台中，普通微博用户通过浏览微博认证用户的微博来获取自己感兴趣的资源。而微博认证用户是浏览与自身兴趣相似或相近的其他微博认证用户的微博来获取资源。为使各类微博用户均能有效获取与自身兴趣相似的用户或信息，本文提出了一种基于相似主题的微博用户推荐算法。改进的算法包括如下四个环节[10]：①用户社群划分—专家/中枢类、普通类；②用户类别划分—专家及中枢；③设定用户推荐关系；④微博主题相似度计算及 Top-N 推荐。其中环节②和环节④是算法的核心。改进算法框架描述如下：

(1)用户社群划分。根据用户拥有的粉丝量、转发微博量以及原创微博量划分出两大类用户社群：专家/中枢类、普通类。计算如式(1)所示。

$$\begin{cases} F \geqslant \alpha \\ \dfrac{Y-Z}{Y+Z} \geqslant 50\% \\ Y+Z \geqslant \beta \end{cases} \tag{1}$$

其中，F 为用户粉丝数量，Y 为原创数量，Z 为转发数量，α 和 β 的取值根据实验爬取的数据集确定。满足式(1)的用户不仅粉丝数量众多，而且微博信息量庞大，本文将该类用户归类为专家/中枢类。其他用户归类为普通类。

(2) 用户类别划分。对于步骤(1)中的专家/中枢类用户社群,本文将采用改进的 HITS 算法将其进一步划分为专家用户和中枢用户,此环节内容的实现过程详见 3.1 节。

(3) 用户推荐关系设定。步骤(1)和步骤(2)划分出了不同类型的用户,算法按照不同用户的需求进行推荐。各类型用户间的推荐关系设定如图 1 所示。在图中所示的推荐关系中,仅将专家用户推荐给中枢用户和其他专家用户,为普通用户仅推荐中枢用户。在推荐关系设定中,为普通用户推荐中枢用户的好处如下:①普通用户无须关注过多的中枢用户即可获取中枢用户转发的多个专家用户的微博信息,从中发现与自身兴趣相同的信息。②专家用户兴趣单一,而中枢用户兴趣具有多样性,可将中枢用户所关注的专家用户的微博信息推荐给普通用户,这些信息分属不同兴趣领域,进而扩展普通用户的兴趣领域。

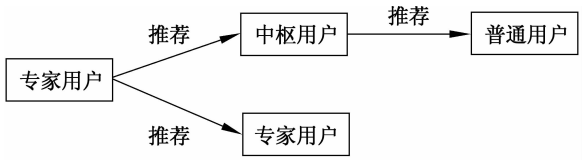

图 1 用户推荐关系设定

(4) 微博主题相似度计算及 Top-N 推荐。该环节完成三类微博主题向量的计算:原创微博主题向量(专家类)、转发微博主题向量(中枢类)以及所有微博主题向量(普通类)。对图 1 所示的三类推荐关系(专家推荐专家、专家推荐中枢、中枢推荐普通用户),本文使用余弦相似度来实现用户微博相似度的计算,根据计算结果完成 Top-N 推荐,此环节内容的具体实现详见 3.2 节。

3 推荐算法核心环节

3.1 用户类别划分

式(1)仅能将微博用户划分为两大类用户社群:专家/中枢类、普通类。本文还需进一步区分专家用户和中枢用户,这个区分过程的实现可以借鉴 HITS 算法。传统 HITS 算法衡量检索到网页的重要度是通过如下两个指标来实现:网页中心度(Hub)、权威度(Authority)[11]。据此,本文规定专家用户为 Authority 值高的微博用户,而中枢用户则为 Hub 值高的用户。受 HITS 算法中隐含假设的启发,我们同理认为社交平台中的专家用户与中枢用户之间的关注是相互的。本文将改进传统 HITS 算法,并使用改进后的算法来计算微博用户的 Authority 值和 Hub 值,以实现对专家用户和中枢用户的区分。

$A(u_i)$ 为用户 u_i 的 Authority 值,$H(u_i)$ 为用户 u_i 的 Hub 值。传统的 HITS 算法在计算用户 Authority 值时,仅累加该用户粉丝用户的 Hub 值,如式(2)所示;而计算用户 Hub 值时,也仅累加该微博用户所关注用户的 Authority 值,如式(3)所示。综上分析,传统的 HITS 方法因无法体现微博用户间的差异性而使计算所得的指标值准确率较低。

$$A(u_i) = H(d_1) + H(d_2) + \cdots + H(d_j) \quad (2)$$

$$H(u_i) = A(t_1) + A(t_2) + \cdots + A(t_k) \quad (3)$$

其中,$d_1, d_2 \cdots, d_j$ 为用户 u_i 的粉丝用户集合,$t_1, t_2, \cdots, t_k$ 为用户 u_i 关注的用户集合。

每个微博用户类型具有自身的特点:专家用户转发微博数量要远远小于原创微博数量,而中枢用户则恰恰相反。因而,在对 HITS 算法进行改进时将引入如下参数:用户原创微博比例、用户转发微博比例。改进算法中对 Authority 值和 Hub 值的计算如式(4)、式(5)所示。

$$A(u_i) = H(d_1)\frac{Z_{d_1}}{N_{d_1}} + H(d_2)\frac{Z_{d_2}}{N_{d_2}} + \cdots + H(d_j)\frac{Z_{d_j}}{N_{d_j}} \quad (4)$$

其中,Z_{d_j} 为粉丝用户 d_j 转发微博用户 u_i 的微博数量,N_{d_j} 为用户 d_j 的原创及转发微博数量和。

$$H(u_i) = A(t_1)\frac{Y_{t_1}}{N_{t_1}} + A(t_2)\frac{Y_{t_2}}{N_{t_2}} + \cdots + A(t_k)\frac{Y_{t_k}}{N_{t_k}} \quad (5)$$

其中,Y_{t_k} 为用户 u_i 转发微博用户 t_k 的原创微博数量,N_{t_k} 表示用户 t_k 的原创及转发微博数量总和。

在 HITS 算法中使用式(4)和式(5),通过多次迭代计算所得的用户 Hub 值和 Authority 值更加精确,与传统的 HITS 算法相比能将用户之间的差异性体现得更为准确。

3.2 微博主题相似度计算及推荐

本文在计算用户微博主题相似度时,不同类别的用户将采用不同的计算方法。

(1) 针对普通用户:整合该用户的所有转发和原创微博信息到一个文档中,继而运用 LDA[12] 对该普通用户的微博主题向量进行计算[13]。

(2) 针对专家用户：计算其原创微博主题向量，如式(6)所示。

$$A(u_i)=H(d_1)\overrightarrow{Z_{d_1}}+H(d_2)\overrightarrow{Z_{d_2}}+\cdots+H(d_j)\overrightarrow{Z_{d_j}} \tag{6}$$

其中，$d_1,d_2,\cdots,d_j$ 为微博用户 u_i 的粉丝用户集合，$H(d_j)$为微博用户 d_j 的 Hub 值，$\overrightarrow{Z_{d_j}}$为微博用户 d_j 转发微博用户 u_i 的微博主题向量。

(3) 针对中枢用户：计算其转发微博主题向量，如式(7)所示。

$$H(u_i)=A(t_1)\overrightarrow{Y_{t_1}}+A(t_2)\overrightarrow{Y_{t_2}}+\cdots+A(t_k)\overrightarrow{Y_{t_k}} \tag{7}$$

其中，$t_1,t_2,\cdots,t_k$ 为微博用户 u_i 关注的用户集合，$A(t_k)$为用户 t_k 的 Authority 值，$\overrightarrow{Y_{t_k}}$为微博用户 u_i 转发微博用户 t_k 的微博主题向量。式(6)和式(7)表明用户微博主题向量的计算会受到该微博用户的 Authority 值和 Hub 值的影响。

为了发现并推荐与使用者兴趣相似的其他微博用户，本文算法首先依据式(1)完成目标用户的社群划分[14]，若目标用户划分到专家/中枢类用户社群，则继续使用式(4)、式(5)对 Authority 值和 Hub 值进行计算，以确定该用户是专家用户还是中枢用户；为了发现与目标用户兴趣相近的其他微博用户，微博主题相似度的计算使用余弦相似度[15]来实现。根据图 1 设定的微博用户推荐关系，为目标用户选择合适的用户类型进行微博主题相似度的计算。

上述微博主题相似度计算如式(8)所示。

$$\mathrm{sim}(u_i,u_j)=\frac{\overrightarrow{U_i}\cdot\overrightarrow{U_j}}{\sqrt{|\overrightarrow{U_i}|\cdot|\overrightarrow{U_j}|}} \tag{8}$$

其中，$\overrightarrow{U_i}$表示微博用户 u_i 的微博主题向量，$\overrightarrow{U_j}$表示微博用户 u_j 的微博主题向量。将计算所得的微博主题相似度值排序后进行 Top-N 推荐[16]。

3.3　核心算法设计

算法 1　相似主题的微博用户推荐算法

Input：所有用户集 U，粉丝数量阈值 α，微博数量阈值 β，用户粉丝数量 F，原创微博数量 Y，转发微博数量 Z
Output：微博用户间的微博主题相似性 *sim*

```
//步骤 1：用户社群划分—专家/中枢及普通用户群划分
1：for i=1 to n do   //n 为所有用户数量
2：   if u_i.F>α&&u_i.Y+u_i.Z>β&&(u_i.Y-u_i.Z)/(u_i.Y+u_i.Z)>50%
3：      L_1.AddUser(u_i)；//添加用户到 L_1
4：   else
5：      L_2.AddUser(u_i)；//添加用户到 L_2
6：   end if
7：end for
//步骤 2：改进的 HITS 算法
8：for k=1 to n do
9：   u_k.auth=1；//Authority 初始值设置为 1
10：u_k.hub=1；//Hub 初始值设置为 1
11：end for
12：for j=1 to n do
13.   u_j.auth=Get_Auth (u_j.hub,Y)；//用户 u_j 的 Authority
14.   u_j.hub=Get_Hub(u_j.auth,Z)；//用户 u_j 的 Hub
15：end for
//步骤 3：用户类别划分—专家用户与中枢用户
16：for k=1 to n do
17：   if u_k.auth>>u_k.hub&&u_k∈L_1
18：      U_zj.AddUser(u_k)；//u_k 添加为专家用户 U_zj
19：   else if u_k.auth<<u_k.hub&&u_k∈L_1
20：      U_zs.AddUser(u_k)；//u_k 添加为中枢用户 U_zs
21：   else if u_k∈L_2
22：      U_pt.AddUser(u_k)；//u_k 添加为普通用户 U_pt
23：   end if
24：end for
//步骤 4：使用 LDA 抽取主题向量并计算微博主题相似度
25：for i=1 to n do
26：   Z_F=Get_ZF(u_i)；//u_i 的转发微博量
27：   Y_C=Get_YC(u_i)；//u_i 的原创微博量
28：   Z_U_i=Get_ZT(Z_F)；//u_i 的转发主题向量 Z_U_i
29：   Y_U_i=Get_ZT(Y_C)；//u_i 的原创主题向量 Y_U_i
30：end for
31：for k=1 to n do
32：   A_uk=Get_ZJ(U_zj.Z_U_k,u_k.hub_k)；//计算专家用户原创微博主题向量 A_Uk
33：   H_uk= Get_Z (U_zs.Y_U_k,u_k.auth_k)；//计算中枢用户转发微博主题向量 H_Uk
34：   P_T=Get_WB(U_pt)；//普通用户所有微博量
35：   P_uk=Get_ZT(P_T)；//计算普通用户微博主题向量
36：end for
37：for m=1 to n do
38：   sim=Get_ZT_Sim(U_m, U_m+1)；//计算主题相似度
39：end for
```

在改进的 HITS 算法中：引入用户微博转发率和用户微博原创率作为链接系数，用户之间的链接系数越大表明用户之间有更多的相似兴趣，因而有效解决主题偏移问题；此外，算法中还引入用户之间的微博转发率和原创率，即使作弊用户关注大量专家用户，但是作弊用户与专家用户之间的链接系数却很低，有效降低作弊用户的 Hub 值，使容易作弊问题得到有效的解决。

3.4 改进算法收敛性分析

由于本文计算用户 Authority 值只是将原始 HITS 算法计算权威度的每一个 Hub 值乘以用户的微博转发率，计算用户 Hub 值也仅在原 HITS 算法计算中心度的每一个 Authority 值乘以用户的原创率(微博转发率、原创率均为介于 0 到 1 之间数值)。即若原始 HITS 算法计算所得的权威度记为 $A(i)$、中心度记为 $H(i)$，本文算法计算所得的用户权威度和中心度分别为 $A'(i)$和 $H'(i)$，则有 $A'(i)\leqslant A(i)$，$H'(i)\leqslant H(i)$，改进算法的迭代规则与原始算法相同，因此，改进后的 HITS 算法依旧保持收敛。在算法设计中最大迭代次数 max_iterations=150，最小误差阈值 $\varepsilon=0.001$，在算法执行过程中，在进行 105 次迭代计算后收敛，得到稳定的数值。

4 实验实施与结果分析

4.1 实验环境配置

本文实验所需的软硬件环境配置：操作系统为 Windows7、CPU 为 i5-4460、主频 3.20GHz、硬盘空间 50GB 及以上、内存 2GB 及以上，编程语言为 Java。在 Eclipse 环境下搭建实验平台并完成实验方案设计。Mahout 为 Apache 的一个开源工具，提供大多数经典推荐算法的代码。在实验过程中使用和改编的对比算法均为封装在 Apache Mahout 框架中的协同过滤推荐算法。

4.2 实验数据

本文通过新浪微博提供的 API 接口和爬虫工具[17]，从新浪微博中选取 8 个常见的主题领域：体育领域、美食领域、科技领域、健康领域、汽车领域、情感生活领域、房产领域和娱乐领域，实验中从这些领域中各随机选取 10 名种子用户进行辐射以爬取实验数据，最终采集微博用户信息 15 万条、微博主题信息 5 000 万条。按照式(1)将 15 万名用户划分到不同的用户社群：专家/中枢类包含 10 516 名用户，剩下的用户均为普通类；进而继续使用用户类别划分算法将 10 516 名用户划分成 4 275 个专家用户、6 241 个中枢用户。

实验中根据爬取的数据集确定参数 α 和 β 的取值。α=sum(数据集中所有用户的粉丝数量)/count(数据集中的用户)；β=sum(数据集中所有用户的微博数量)/count(数据集中的用户)。针对本文数据集的计算结果 $\alpha=1000$、$\beta=300$。

为了验证本文参数 α 和 β 取值的合理性，在实验过程中另外选取 α 为 600、800、1200，β 为 100、200、400，再加上本文选取的 $\alpha=1000$，$\beta=300$。根据不同的 α 和 β 取值组合进行实验，对每一组实验的 $F1$ 值进行计算，结果如表 1 所示。

表 1 不同 α 和 β 值对应的 *F1* 度量值

β \ α	600	800	1 000	1 200
100	0.39	0.44	0.51	0.47
200	0.42	0.49	0.57	0.52
300	0.57	0.61	0.68	0.63
400	0.5	0.52	0.60	0.55

从表 1 的实验结果可以看出，当参数 $\alpha=1000$、$\beta=300$ 时，$F1$ 度量值最大，实验效果最好。

实验选取的数据分布如表 2 所示。

表 2 各领域实验数据分布

序号	主题领域	微博数量/万	用户数量
1	体育	640	21 680
2	美食	580	15 738
3	体育	640	21 680
4	健康	632	19 238
5	汽车	606	16 295
6	情感生活	634	19 057
7	房产	638	19 639
8	娱乐	645	20 428

4.3 实验结果与分析

实验中分别从不同类别的微博用户中随机选取 1 名用户以验证用户分类的准确性：

① 中枢用户 1772598673，微博主题领域为美食领域，原创微博数为 615，转发微博数为 12 650，用户 Hub 值为 0.719 3；

② 专家用户 3606455372，微博主题领域为情感生活领域，原创微博数为 12 675，转发微博数为 3 009，用户 Authority 值为 0.672 3；

③ 普通用户 1863606222，微博主题领域为科技领域，用户原创微博数为 488，用户转发微博数为 576。

对于不同类型微博用户的推荐结果如表 3～表 5 所示。

表 3～表 5 中的数据表明：在三种类型用户的 Top-10 推荐列表中，对于同一个用户 ID 号，应用本文算法所推荐的主题相似度值均要高于 MISUR 算法和 TCF 算法。

表 3　专家用户推荐结果列表

用户 ID 号	不同算法的主题相似度			用户 Authority 值
	MISUR	TCF	本文	
ZJ2268603763	0.69	0.63	0.83	0.83
ZJ3171567737	0.65	0.61	0.80	0.79
ZJ2131423263	0.64	0.56	0.79	0.76
ZJ2376786824	0.61	0.54	0.78	0.75
ZJ2602452372	0.60	0.52	0.77	0.71
ZJ3863142166	0.57	0.49	0.74	0.70
ZJ1987582672	0.56	0.46	0.73	0.67
ZJ2265522924	0.55	0.45	0.70	0.65
ZJ3653514533	0.53	0.41	0.69	0.63
ZJ1684863852	0.52	0.40	0.67	0.63

表 4　中枢用户推荐结果列表

用户 ID 号	不同算法的主题相似度			用户 Authority 值
	MISUR	TCF	本文	
ZS1762264573	0.64	0.55	0.78	0.70
ZS1408935667	0.63	0.54	0.77	0.68
ZS1691703935	0.62	0.51	0.74	0.66
ZS1732935622	0.61	0.49	0.73	0.65
ZS1976498155	0.60	0.46	0.72	0.64
ZS3653515000	0.58	0.44	0.70	0.61
ZS2775437003	0.58	0.41	0.65	0.59
ZS1098775524	0.57	0.41	0.65	0.55
ZS1768367471	0.56	0.39	0.64	0.53
ZS3653510470	0.55	0.38	0.63	0.52

表 5　普通用户推荐结果列表

用户 ID 号	不同算法的主题相似度			用户 Authority 值
	MISUR	TCF	本文	
PT2609110630	0.69	0.67	0.79	0.77
PT1695018985	0.68	0.66	0.78	0.74
PT2206047554	0.67	0.63	0.76	0.74
PT2041182062	0.65	0.62	0.75	0.71
PT1179931324	0.64	0.61	0.73	0.67
PT1216969484	0.63	0.58	0.71	0.64
PT1864274385	0.61	0.57	0.68	0.62
PT1840113813	0.60	0.56	0.65	0.60
PT1795923673	0.59	0.54	0.64	0.53
PT1784457002	0.57	0.52	0.62	0.52

为了更好地验证本文算法的优势，本文选取了传统的微博用户推荐算法与本文算法进行对比：①TCF[18]算法将协同过滤和标签进行结合：算法在计算资源特征相似性和用户偏好度时融入资源标签，并应用基于资源的协同过滤推荐算法来完成资源的 TOP-N 个性化推荐；MISUR[5]算法在进行微博用户推荐时，考虑用户的社交信息、交互关系以及微博内容等多源信息，在计算总相似度时融入了时间权重因子及丰富度权重因子，最后根据计算所得的相似度值向用户进行 Top-N 推荐；本文算法与 TCF 算法和 MISUR 算法在推荐好友个数不同情况下的准确率、召回率和 $F1$ 度量值[19]对比结果如图 2～图 4 所示。

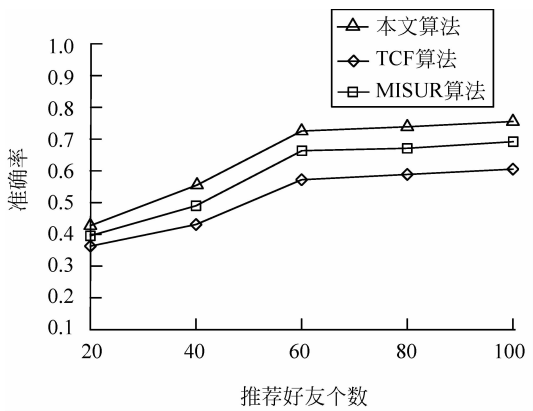

图 2　不同算法推荐准确率对比

由图 2 可以观察到三种算法的推荐准确率随着推荐好友个数增加而不断提升，当推荐好友个数达到 60 时趋于平缓。在推荐好友为 20 时，本文算法的准确率比 TCF 算法提高了 19.4%，比 MISUR 算法提高了 7.5%；在推荐好友为 60 时，本文算法的

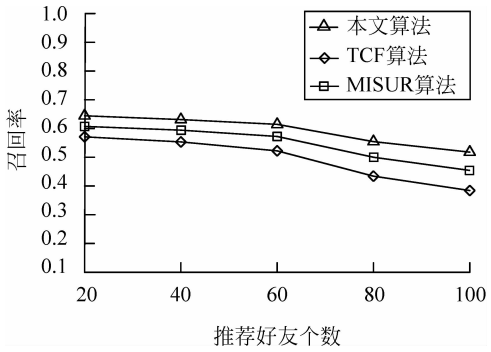

图3 不同算法推荐召回率对比

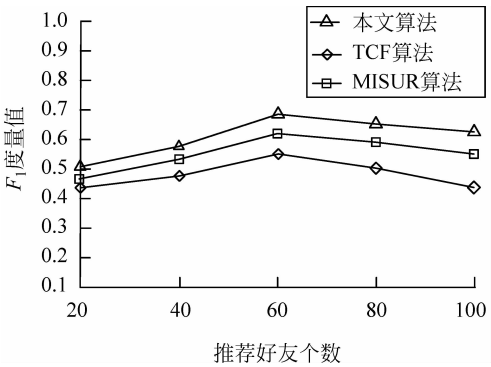

图4 不同算法推荐F1度量值对比

准确率比TCF算法提高了28.6%,比MISUR算法提高了9.1%;在推荐好友为100时,本文算法的准确率比TCF算法提高了25%,比MISUR算法提高了8.7%。

从图3可以看出三种算法的召回率随着推荐好友个数增加而下降,当推荐好友个数小于60时下降较为平缓。当推荐好友个数为20、60、100时,本文算法的召回率与TCF算法相比分别提升了14.3%、17.3%、40.5%,比MISURF算法分别提升了4.9%、7%、15.6%。分析表明,随着推荐好友个数的增加,本文算法的召回率要优于其他两种对比算法。

图4中的数据表明,当推荐好友个数增加时,三种推荐算法的$F1$值都有所提升。当推荐好友数量达到60时,各算法的$F1$度量值均达到峰值;当推荐好友数量超过60后,各算法的$F1$度量值均在下降。产生这种情况的主要原因为:当推荐好友数量偏多时,排名靠后用户的Authority值和Hub值都偏低,应用本文算法时无法体现这些专家/中枢用户的重要度,所以影响了微博主题相似性的计算,导致$F1$度量值下降。

根据图4中三种算法$F1$值的对比可以看出:在推荐结果准确率上本文算法要优于MISUR算法和TCF算法。原因如下:数据稀疏性问题是传统的TCF算法中存在的弊端,也是推荐准确率低的重要原因;MISUR算法虽然通过用户多源信息的融入来缓解数据稀疏问题对推荐结果的影响,但是缺乏对微博用户的社交结构信息进行考量。本文算法首先运用改进的HITS算法分析微博用户的网络拓扑结构,并根据分析的结果将用户划分为专家用户、中枢用户和普通用户,进行相似度计算时结合不同类型用户的行为特点,即本文在进行微博主题相似性计算时既考虑到微博用户的社交网络结构,又引入了Authority值和Hub值,从而使推荐准确率得到有效提升。

5 结束语

本文算法旨在解决传统个性化推荐算法普遍存在的推荐准确率较低的不足。所提算法在使用改进的HITS算法对微博用户的社交网络结构进行分析的基础上,将微博用户划分为专家用户、中枢用户及普通用户3个类型;在计算微博主题相似度时,既结合用户的Authority值和Hub值,又结合用户原创及转发微博信息的数量。通过本文算法与TCF算法和MISUR算法进行对比,结果表明本文算法由于考虑到微博用户的网络结构信息,同时采用改进的HITS算法准确地计算微博用户的Authority值和Hub值,在进行微博主题相似度计算时,融入计算所得的Authority值和Hub值,从而较大程度地提升了推荐的准确率。

参考文献

[1] 徐志明,李栋,刘挺,等.微博用户的相似性度量及其应用[J].计算机学报,2014,37(1):207-218.

[2] Winlaw M, Hynes M B, Caterini A, et al. Algorithmic acceleration of parallel ALS for collaborative filtering: Speeding up distributed big data recommendation in spark[C]//Proceedings of the 2015 IEEE 21st International Conference on Parallel and Distributed Systems. Piscataway, NJ: IEEE, 2015: 682-691.

[3] 毛佳昕,刘奕群,张敏,等.基于用户行为的微博用户社会影响力分析[J].计算机学报,2014,37(4): 791-800.

[4] 仲兆满,管燕,胡云,等.基于背景和内容的微博用户兴趣挖掘[J].软件学报,2017,28(2):278-291.

[5] 姚彬修,倪建成,于苹苹,等.基于多源信息相似度的微博用户推荐算法[J].计算机应用,2017,37(5): 1382-1386.

[6] 彭泽环，孙乐，韩先培，等. 基于排序学习的微博用户推荐[J]. 中文信息学报，2013，27(4)：96-102.
[7] 吴树芳，徐建民. 基于 HITS 算法的微博用户可信度评估[J]. 山东大学学报(工学版)，2016，46(5)：7-12.
[8] 喻依，甘若迅，樊锁海，等. 基于 PageRank 算法和 HITS 算法的期刊评价研究[J]. 计算机科学，2014，41(s1)：110-113.
[9] 苗家，马军，陈竹敏. 一种基于 HITS 算法的 Blog 文摘方法[J]. 中文信息学报，2011，25(1)：104-110.
[10] 安维凯. 基于个性化标签和微博主题的重要用户推荐方法研究[D]. 沈阳：辽宁大学硕士学位论文，2018.
[11] 刘昊，洪宇，姚亮，等. 基于 HITS 算法的双语句对挖掘优化方法[J]. 中文信息学报，2017，31(02)：25-35.
[12] 邸亮，杜永萍. LDA 模型在微博用户推荐中的应用[J]. 计算机工程，2014，40(5)：1-6.
[13] 周小平，梁循，张海燕. 基于 R-C 模型的微博用户社区发现[J]. 软件学报，2014，25(12)：2808-2823.
[14] 祝婷，秦春秀，李祖海. 基于用户分类的协同过滤个性化推荐方法研究[J]. 数据分析与知识发现，2015，31(6)：13-19.
[15] Han S，Xu Y. Friend recommendation of microblog in classification framework：Using multiple social behavior features [C]//Proceedings of International Conference on Behavior，Economic and Social Computing，NJ：IEEE，2015：1-6.
[16] 任星怡，宋美娜，宋俊德. 基于位置社交网络的上下文感知的兴趣点推荐[J]. 计算机学报，2017，40(4)：824-841.
[17] 陈梅梅，薛康杰. 基于标签簇多构面信任关系的个性化推荐算法研究[J]. 数据分析与知识发现，2017，1(5)：94-101.
[18] 蔡强，韩东梅，李海生，等. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学，2014，41(1)：69-71.
[19] 侯银秀，李伟卿，王伟军，等. 基于用户偏好与商品属性情感匹配的图书个性化推荐研究[J]. 数据分析与知识发现，2017，1(8)：9-17.

王嵘冰(1979—)，博士，副教授，主要研究领域为数据挖掘，云计算、大数据技术等。
E-mail：wrb@lnu.edu.cn

冯勇(1973—)，通信作者，博士，教授，主要研究领域为数据挖掘、大数据技术、个性化推荐等。
E-mail：910527270@qq.com

徐红艳(1972—)，硕士，副教授，主要研究领域为数据挖掘，Deep Web 等。
E-mail：xuhongyan@lnu.edu.cn