



融合 Spark 与隐性兴趣的用户综合影响力度量

童曼琪^{1a}, 黄江升², 郭 昆^{1b}

(1. 福州大学 a. 福建省空间数据挖掘与信息共享教育部重点实验室;
b. 福建省网络计算与智能信息处理重点实验室, 福州 350002;
2. 国网信通亿力科技有限责任公司, 福州 350003)

摘 要: 为解决传统用户影响力度量算法面向海量数据处理时运行速度下降的问题, 提出一种基于隐性兴趣的用户综合影响力度量算法。通过隐含狄利克雷分配模型得到用户隐性兴趣偏好, 根据困惑度和平均话题相似度综合确定最优兴趣话题数, 并改进 PageRank 算法的用户兴趣传播转移率获得用户隐性兴趣传播影响力。在 Spark 计算框架的基础上, 采用层次分析法且结合用户自身影响力和用户隐性兴趣传播影响力, 计算得到最终用户影响力。实验结果表明, 该算法综合考虑用户兴趣和用户自身影响因素, 能够更客观高效地评估用户的真实影响力。

关键词: 用户影响力; 用户兴趣相似度; PageRank 算法; Spark 计算框架; 隐含狄利克雷分配模型

开放科学(资源服务)标志码(OSID):



中文引用格式: 童曼琪, 黄江升, 郭昆. 融合 Spark 与隐性兴趣的用户综合影响力度量[J]. 计算机工程, 2020, 46(11): 61-69.

英文引用格式: TONG Manqi, HUANG Jiangsheng, GUO Kun. Comprehensive user influence measurement combining Spark and recessive interest[J]. Computer Engineering, 2020, 46(11): 61-69.

Comprehensive User Influence Measurement Combining Spark and Recessive Interest

TONG Manqi^{1a}, HUANG Jiangsheng², GUO Kun^{1b}

(1a. Fujian Provincial Key Laboratory of Spatial Data Mining and Information Sharing, Ministry of Education;

1b. Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou 350002, China;

2. State Grid Info-Telecom Great Power Science and Technology Co., Ltd., Fuzhou 350003, China)

【Abstract】 The speed of traditional user influence measurement algorithms is reduced when dealing with massive data. To address the problem, this paper proposes a comprehensive user influence measurement algorithm based on recessive interest. The Latent Dirichlet Allocation (LDA) model is used to obtain the recessive interests of the user, and the number of the optimal interest topics is determined based on the perplexity and the average topic similarity. Then, the transmission rate of user interests in the PageRank algorithm is improved to obtain the User Interest Factor (UIF). Finally, based on the Spark computing framework, the Analytic Hierarchy Process (AHP) is used to calculate the ultimate user influence by combining the influence of the user and UIF. Experimental results show that the proposed algorithm has a holistic consideration on user interests and the influence factors of the user, which enables it to provide more efficient and reasonable evaluation of the real influence of the user.

【Key words】 user influence; user interest similarity; PageRank algorithm; Spark computing framework; Latent Dirichlet Allocation (LDA) model

DOI: 10.19678/j.issn.1000-3428.0056187

0 概述

据中国互联网络信息中心于 2019 年 2 月 28 日在

北京发布的《第 43 次中国互联网络发展状况统计报告》^[1] 可知, 截至 2018 年 12 月微博使用率达到 42.3%, 较 2017 年底上升 1.4 个百分点。在微博、

基金项目: 国家自然科学基金(61300104); 福建省高等学校新世纪优秀人才支持计划(JA13021); 福建省杰出青年科学基金(2015J06014); 福建省高校产学研合作项目(2017H6008)。

作者简介: 童曼琪(1993—), 女, 硕士研究生, 主研方向为数据挖掘; 黄江升, 工程师; 郭 昆(通信作者), 副教授、博士。

收稿日期: 2019-10-08 **修回日期:** 2019-11-08 **E-mail:** gukn123@163.com

Twitter、YELP 和大众点评等社交应用中,其社交属性决定了被分享的话题多数为社交关系圈内热点或共同关注、感兴趣的话题。而社交网络在信息传播过程中通常会存在一些影响力大的用户,他们通过评论可以在短时间内使信息得到广泛传播,甚至会引导舆论走向。因此,用户影响力度量^[2]对于信息传播具有重要作用。

目前,国内外学者对于社交网络中用户影响力度量的研究包括基于用户自身属性、用户动态行为和用户特性综合考虑的 3 类用户影响力度量方法。

第 1 类方法通过发博数、好友数、评论数和转发数等用户自身属性度量用户影响力。文献[3]定义用户直接影响力及级联影响力,提出基于用户消息传播范围的用户影响力度量化方法,并给出用户影响力计算方法。文献[4]基于微博数据得到传播影响力、信息完整度、活跃度和认证指数 4 项评价指标,构建用户权威性定量计算模型。文献[5]基于用户节点度计算用户影响力。但此类算法仅考虑了用户自身属性,未考虑其他影响因素,不能排除沉默用户或者僵尸用户对网络节点影响力的干扰。

第 2 类方法通过转发、回复等用户动态行为度量用户影响力。文献[6]定义用户影响力分类概念,并且综合考虑微博中的转发、回复、复制和阅读 4 种关系,提出基于多关系网络的遍历所有话题的随机游走模型。文献[7]在 Twitter 数据集中,使用种子节点扩散范围衡量每个种子节点的影响力。为改进回复关系链接稀疏的问题,文献[8]引入帖子作为节点的间接回复网络,通过用户回复帖子的情感倾向性来度量用户节点之间的影响,提出基于倾向性转变的 TTRank 算法。为衡量消息传播过程的影响力,文献[9]采用幂率衰减函数估计用户初始影响力、信息传播衰减系数以及传播持久性指标,综合度量节点影响力。文献[10]通过考虑用户阅读行为特征和博文转发情况来综合度量用户影响力。此类方法能更全面地描述用户传播影响力,但未考虑用户认证情况、好友数等用户自身属性对影响力的贡献。

第 3 类方法基于改进 PageRank 算法并综合考虑粉丝和追随者数量等用户特性来度量用户影响力。文献[11]考虑了用户好友拓扑并分析博文的主题相似性,得出用户综合影响力是每个主题下的影响力与相应权重的乘积之和,但是该方法未考虑用户活跃度和权威性等因素。文献[12]指出用户影响力由其自身属性及其粉丝共同决定,但是在量化用户自身影响力和粉丝对其影响力时,特征均采用均一化处理方式,从而导致计算结果与实际情况不太符合,用户影响力评价客观性较差。文献[13]对用

户自身影响因素进行量化,通过设置不同行为的权重值,解决了文献[12]算法中追随者影响力等值传递的问题,但未考虑兴趣对用户影响力的贡献。此外,上述算法在处理海量数据时运行速度均有所下降。

本文在 PageRank 算法的基础上,提出融合隐性兴趣的用户综合影响力度量算法 IBPR。利用隐含狄利克雷分配(Latent Dirichlet Allocation, LDA)模型得到用户隐性兴趣偏好,通过困惑度^[14]和平均话题相似度^[15]确定最优兴趣话题数,并建立用户好友兴趣拓补网络,扩展用户之间的隐性兴趣关联关系,同时综合用户自身影响力和隐性兴趣传播影响力,过滤大部分僵尸用户,使用户影响力评估更全面客观。

1 Spark 计算框架

Spark^[16]是加州大学伯克利分校 AMP 实验室开发的通用内存并行计算框架,基于有向无环图(Directed Acyclic Graph, DAG)的任务调度执行机制,支持在内存中对数据进行效率更高的迭代计算。Spark 生态圈即伯克利数据分析栈(BDAS),其包含了 Spark Core、Spark SQL、Spark Streaming、MLLib 和 GraphX 等组件,是实现无缝集成并提供一站式解决方案的平台。官方数据表明,如果数据基于 Spark 磁盘读取,速度是 Hadoop 的 10 倍以上,如果数据基于 Spark 内存读取,速度是 Hadoop 的 100 倍以上,如图 1 所示。

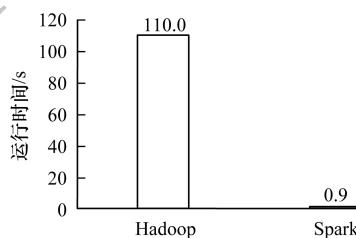


图 1 Hadoop 与 Spark 逻辑回归运行时间对比

Fig. 1 Comparison of logistic regression running time for Hadoop and Spark

2 融合隐性兴趣的用户综合影响力度量

本文通过综合分析用户的个人信息和动态行为数据,设计基于用户隐性兴趣传播影响力(User Interest Factor, UIF)、用户认证权威性(User Identity Authority, UIA)及用户活跃度(User Activity Degree, UAD)3 个维度的 IBPR 算法,如图 2 所示,用户影响力包括用户隐性兴趣传播影响力和用户自身影响力,而用户自身影响力又包括用户认证权威性以及用户活跃度。

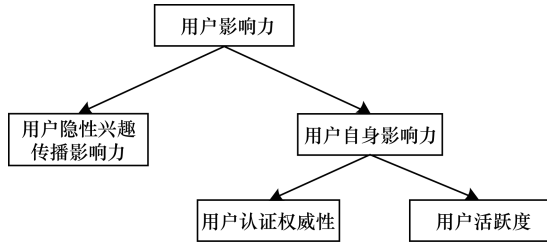


图 2 IBPR 算法层次结构

Fig. 2 IBPR algorithm hierarchy

2.1 用户认证权威性

定义 1 (用户认证权威性) 用户认证权威性包括用户是否被认证过精英、最近精英年份贡献率、精英年份贡献率,三部分共同构成用户认证权威性三元组 (IA, AR, AC), 计算公式如式 (1) 所示:

$$\begin{aligned} \text{UIA}(u_i) &= \delta \times \text{IA}(u_i) + \nu \times \text{AR}(u_i) + \theta \times \text{AC}(u_i) \\ \delta + \nu + \theta &= 1 \end{aligned} \quad (1)$$

1) $\text{IA}(u_i)$ 表示用户是否被认证过精英的度量。YELP 官网每年会评选出精英用户, 评选经过官方审核, 可信度较高。经过官网认证的精英用户更容易受到关注和重视, 计算公式如式 (2) 所示:

$$\text{IA}(u_i) = \begin{cases} 1, & u_i \text{ 为精英用户} \\ 0, & \text{其他} \end{cases} \quad (2)$$

2) $\text{AR}(u_i)$ 表示在 u_i 为精英用户情况下的最近精英年份贡献率度量。一些精英用户虽然注册时间较晚, 但消息发布活跃且消息内容吸引人也会吸引较多关注, 计算公式如式 (3) 所示:

$$\text{AR}(u_i) = \frac{\text{lastTime}(u_i) - \text{signTime}(u_i)}{\text{maxTime} - \text{signTime}(u_i) + \tau} \quad (3)$$

其中, $\text{lastTime}(u_i)$ 表示用户 u_i 被评选为精英的最近年份, maxTime 表示所有用户评选为精英的最近年份集合的最大值, $\text{signTime}(u_i)$ 表示用户 u_i 的注册时间, τ 设置为 1, 以避免式 (3) 的分母为 0。

3) $\text{AC}(u_i)$ 表示在用户 u_i 为精英情况下的精英年份贡献率度量。用户精英年份贡献率越高, 用户被评选为精英的次数占比越大, 越容易受到关注, 对其他用户影响也越大, 计算公式如式 (4) 所示:

$$\text{AC}(u_i) = \frac{\text{count}(u_i) - \text{minCount}}{\text{maxCount} - \text{minCount}} \quad (4)$$

其中, $\text{count}(u_i)$ 表示用户 u_i 被评选为精英的年份数, minCount 表示所有用户中最少的精英年份数, maxCount 表示所有用户中最多的精英年份数。

在上述 3 个度量因素中, 笔者认为决定用户认证权威性的首要因素为用户 u_i 被认证过精英, 其次是精英年份贡献率, 最后是最近精英年份贡献率。由于精英评选经过官方审核, 精英年份贡献率越高表示用户多次被评选为精英, 相对于一些注册时间较晚只被评过较少精英次数的用户而言, 前者更可

能被其他用户查看并传播影响力。

本文采用层次分析法 (Analytic Hierarchy Process, AHP) [4] 确定用户认证权威性评价特征权值。层次分析法主要用于解决复杂的多因素决策问题, 是一种层次权重决策分析方法。对于参数 δ 、 ν 及 θ , 构建评价特征的判断矩阵 A_{UIA} , 根据变量相对重要性等级表 [17] 并结合 3 个度量因素的相对重要关系对矩阵元素 a_{ij} 赋值, 例如用户是否被认证过精英相对精英年份贡献率更重要, 因此 a_{12} 取值为 7, 计算公式如式 (5) 所示:

$$A_{\text{UIA}} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 7 & 4 \\ 1/7 & 1 & 1/2 \\ 1/4 & 2 & 1 \end{bmatrix} \quad (5)$$

计算得到判断矩阵 A_{UIA} 的最大特征值 $\lambda_{\text{max, UIA}}$ 为 3.002, 该特征值对应的特征向量 $W_{\text{UIA}} = [2.147 \ 3, 0.293 \ 4, 0.561 \ 3]^T$, 其一致性比率 $\text{CR} = 0.001 \ 7$, 远小于 0.1, 因此满足一致性检验相关性要求, 从而判定矩阵 A_{UIA} 合理, 并对特征向量 W_{UIA} 进行归一化处理得到最终权重比例, 即 $(\delta, \nu, \theta) = (0.715 \ 3, 0.097 \ 7, 0.187 \ 0)$ 。

2.2 用户活跃度

定义 2 (用户活跃度) 用户活跃度即度量用户的动态交互行为频繁程度, 采用平均评论数和被评论数对其进行综合度量, 计算公式如式 (6) 所示:

$$\text{UAD}(u_i) = \frac{\sum_{k=1}^n \eta \times \text{RI}(u_{i,k}) + (1 - \eta) \times \text{RO}(u_{i,k})}{n} \quad (6)$$

其中, $\text{RI}(u_{i,k})$ 表示用户 u_i 第 k 年收到别人的评论数, $\text{RO}(u_{i,k})$ 表示第 k 年评论别人的博文评论数, η 表示权重, n 表示从用户注册一直到该用户产生最新评论时所经过的年份。

2.3 用户隐性兴趣传播影响力

本文通过 LDA [15, 18] 模型得到隐性用户兴趣偏好, 将兴趣话题间的相似度与用户好友拓扑相结合, 改进 PageRank 算法的转移率得到用户隐性兴趣传播影响力。用户隐性兴趣传播影响力计算过程具体如下:

1) 博文数据预处理。对每个用户的博文数据汇总并进行去噪、分词、去停用词等预处理操作。

2) 兴趣话题数确定。通过 LDA 模型得到用户兴趣话题偏好, 并综合确定话题的最优兴趣话题数。

3) 用户隐性兴趣传播影响力计算。基于 PageRank 算法, 结合用户好友兴趣拓网络计算兴趣传播影响力。

2.3.1 博文数据预处理

为减少数据噪声并避免噪声干扰, 需对用户博文数据进行预处理, 主要包括去除噪声、文本分词和词性标注、去停用词等步骤 [19]。噪声数据会影响兴

趣话题的发现,继而降低话题质量。一般噪声数据是指对于其他用户贡献小的用户数据,例如沉默用户或者僵尸用户。根据式(6)计算每个用户的活跃度,将活跃度低于阈值的用户数据标记为噪声数据并剔除。文本分词和词性标注使用 Stanford CoreNLP^[20]开源工具实现。去停用词之前将所有分词均转化为小写形式。去停用词的操作包括去除意义相对较小的词、将数字替换为字符以及去除中文字符的词,保留名词、动词、形容词用于话题发现。

2.3.2 兴趣话题数确定

本文通过 LDA 模型得到用户兴趣话题偏好,然后根据困惑度和平均话题相似度综合确定最优兴趣话题数。

定义 3(余弦相似度) 将向量 a 和向量 b 的相似度 $\text{Sim}_{a,b}$ 定义为两个向量间的余弦相似度,计算公式如下:

$$\text{Sim}_{a,b} = \frac{\sum_{i=1}^n a_i \times b_i}{\sqrt{\sum_{i=1}^n a_i^2} \times \sqrt{\sum_{i=1}^n b_i^2}} \quad (7)$$

其中, a_i, b_i 表示向量 a, b 对应第 i 维的数值。

定义 4(平均话题相似度) 平均话题相似度为所有两两话题向量之间的相似度均值。话题之间相似度越低,说明该话题模型性能越好,计算公式如下:

$$\text{Sim}_{\text{avg}} = \frac{2 \times \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{Sim}_{i,j}}{n \times (n-1)} \quad (8)$$

其中, n 表示话题数, $\text{Sim}_{i,j}$ 表示第 i 个话题和第 j 个话题之间的相似度。

定义 5(困惑度) 对于一篇文档 d ,所训练出的模型对文档 d 属于哪个主题具有不确定性,该不确定性即困惑度。困惑度越低,说明聚类效果越好,计算公式如下:

$$p(D) = \exp \left(- \frac{\sum_{i=1}^Z \log_a(p(w_i))}{\sum_{i=1}^Z N_i} \right) \quad (9)$$

其中, Z 表示文档数, N_i 表示文档集合 D 中的文档 d 经分词处理后的单词数。

算法 1 最优兴趣话题数确定算法

输入 分词后的文本数据 text 、兴趣话题数 x

输出 文本数据集的困惑度 $p(D)$ 、平均话题相似度 Sim_{avg} 、最优兴趣话题数 x_{out}

1) 随机初始化兴趣话题数 $x, x \in (20, 90)$ 。

2) 利用 LDA 模型生成话题和话题词,根据式(8)和式(9)分别计算平均话题相似度和困惑度。

3) 循环执行步骤 2,保存每次计算得到的 $p(D)$

和 Sim_{avg} 。

4) 选择平均话题相似度和困惑度结果最低的兴趣话题数 x_{out} 。

2.3.3 用户隐性兴趣传播影响力计算

影响力大的用户博文通常会受到较多的关注,而用户之间也通过兴趣而产生吸引力并相互关注,即同质性^[11]。对于用户兴趣相似度的计算,目前主要利用 LDA 模型发现兴趣话题,再使用 KL 散度计算用户兴趣话题的相似度,但 KL 散度具有不对称性,即 $\text{KL}(P \| Q) \neq \text{KL}(Q \| P)$ (两个用户的概率分布为 P, Q),因此一般利用取平均值的倒数来近似表示用户相似度。考虑到上述情况,本文采用皮尔逊相关系数计算用户相似度。

定义 6(用户相似度) 用户相似度包括用户间的相似度和兴趣话题间的相似度,采用皮尔逊相关系数进行计算,计算公式如下:

$$\text{US}(u_i, u_j) = \frac{\sum_{k=1}^n (r_{u_i, o_k} - r_i^-)(r_{u_j, o_k} - r_j^-) \text{Sim}(o_k, o_{k'})}{\sqrt{\sum_{k=1}^n (r_{u_i, o_k} - r_i^-)^2} \sqrt{\sum_{k=1}^n (r_{u_j, o_k} - r_j^-)^2}} \quad (10)$$

其中: $\text{Sim}(o_k, o_{k'})$ 表示兴趣话题 k 和 k' 之间的相似度,通过计算两个主题向量的余弦相似度得到兴趣话题的相似度; r_i^- 表示用户 u_i 的平均兴趣话题; r_{u_i, o_k} 表示用户 u_i 的第 k 个兴趣话题, n 表示话题数。

定义 7(用户兴趣传播转移率) 用户兴趣传播转移率即用户间兴趣传播的概率,基于好友拓补网络得到基于兴趣相似度的用户兴趣传播转移率,计算公式如下:

$$W(u_i, u_j) = \frac{\mu \times \text{US}(u_i, u_j)}{\sum_{u_k \in F_{u_i}} \text{US}(u_i, u_k)} + \frac{1 - \mu}{\sum_{u_k \in F_{u_i}} |\text{UF}(u_k)|} \quad (11)$$

其中, F_{u_i} 表示用户 u_i 的粉丝集合, u_k 为 u_i 的任意粉丝,

$\frac{\text{US}(u_i, u_j)}{\sum_{u_k \in F_{u_i}} \text{US}(u_i, u_k)}$ 表示用户间的兴趣转移率, $|\text{UF}(u_k)|$

表示用户 u_k 的出度数。

定义 8(用户隐性兴趣传播影响力) 用户隐性兴趣传播影响力即用户隐性兴趣产生的传播影响力,基于 PageRank 算法改进得到用户隐性兴趣传播影响力,计算公式如下:

$$\text{UIF}(u_i) = (1 - d) + d \times \sum_{u_j \in F_{u_i}} W(u_i, u_j) \times \text{UIF}(u_j) \quad (12)$$

其中, d 表示阻尼系数。

2.4 用户影响力度量

定义 9 (用户影响力) 用户影响力包括用户认证权威性、用户活跃度、用户隐性兴趣传播影响力三部分, 计算公式如下:

$$\begin{aligned} UI(u_i) &= \alpha \times UIF(u_i) + \beta \times UAD(u_i) + \gamma \times UIA(u_i) \\ \alpha + \beta + \gamma &= 1 \end{aligned} \quad (13)$$

其中, $UIF(u_i)$ 为用户 u_i 的隐性兴趣传播影响力, $UAD(u_i)$ 为用户 u_i 的活跃度, $UIA(u_i)$ 为用户 u_i 的认证权威性。

对于参数 α, β 及 γ , 构建评价特征的判断矩阵 A_{UI} , 计算公式如下:

$$A_{UI} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 7 & 2 \\ 1/7 & 1 & 1/3 \\ 1/2 & 3 & 1 \end{bmatrix} \quad (14)$$

计算得到判断矩阵 A_{UI} 的最大特征值 $\lambda_{\max, UI}$ 为 3.002 7, 该特征值对应的特征向量 $W_{UI} = [1.847 4, 0.277 8, 0.877 4]^T$, 其一致性比率 $CR = 0.002 3$, 远小于 0.1, 因此满足一致性检验相关性要求, 从而判定矩阵 A_{UI} 合理, 并对特征向量 W_{UI} 进行归一化处理得到最终权重比例, 即 $(\alpha, \beta, \gamma) = (0.615 3, 0.092 5, 0.292 2)$ 。

算法 2 融合隐性兴趣的用户综合影响力度量算法

输入 用户相似度数据集 $S(US(u_i, u_j))$ 、用户好友关系数据集 $S(F(u_i, u_j))$ 、最大迭代次数 $iter_{\max}$ 、节点影响力迭代阈值

输出 用户隐性兴趣传播影响力 $UIF(u_i)$

1) 根据算法 1 的最优兴趣话题数计算得到 $S(US(u_i, u_j))$ 。

2) 若 $iter < iter_{\max}$, 则执行步骤 3; 否则执行步骤 6。

3) 根据式 (10) 获取用户相似度。

4) 根据式 (11) 计算用户兴趣传播转移率。

5) 根据式 (12) 计算用户隐性兴趣传播影响力 UIF 。

6) 遍历判断每个用户节点, 若对于所有用户节点 $|UIF(u_i) - UIF| < \varepsilon$ 均成立, 则执行步骤 8; 否则执行步骤 7。

7) 将每个用户的兴趣传播影响力 UIF 赋值给上一轮计算的用户影响力 UIF_{tmp} 并累加 $iter$ 迭代次数, 返回步骤 2。

8) 迭代结束, 求得每个用户的兴趣传播影响力 $UIF(u_i)$ 。

9) 根据定义 1 计算用户 u_i 的认证权威性 $UIA(u_i)$ 。

10) 根据定义 2 计算用户 u_i 的活跃度 $UAD(u_i)$ 。

11) 根据式 (13) 计算用户影响力 $UI(u_i)$ 并按从大到小的顺序输出。

考虑到实验数据量较大以及 IBPR 算法迭代计算耗费时间较多, 对算法 2 迭代过程进行基于 Spark 的并行化计算。算法 2 迭代过程 (步骤 2 ~ 步骤 8) 的伪代码具体如下:

```
1. for (i < -1 to iters) {
2.   var oldFinalRanks = finalRanks
3.   val oldRanks = ranks
4.   val oldInterSetRanks = interestRanks
5.   val contribs = links.join(oldRanks).values.flatMap {
6.     case (urls, rank) => val size = urls.size => urls.map(url => (url, rank / size)) }.repartition(5 000)
7.   val interSetContribs = interestLinks.join(oldInterSetRanks).values.flatMap {
8.     case (urls, rank) => urls.map(url => (url._1, url._2 * rank)) }.repartition(5 000)
9.   loop = i
10.  ranks = contribs.mapValues(_ * _)
11.  interestRanks = interSetContribs.mapValues((1 - mu) * _)
12.  finalRanks = (interestRanks).++(ranks).reduceByKey(_ + _).mapValues(1 - d + d * _)
13.  if (delta(oldFinalRanks, finalRanks, min_delta) == true) {
14.    break()
15.  }
16. }
```

代码中的第 5 行和第 6 行分别根据用户好友关系和用户兴趣计算好友转移率和兴趣转移率, 第 8 行 ~ 第 10 行计算得到用户综合影响力, 第 11 行为判断是否达到终止迭代阈值。

3 实验结果与分析

3.1 评价方法与指标

实验使用 M 折交叉验证方法^[6]衡量 IBPR 算法的有效性, 同时选取 4 种对比算法的 Top-10 用户来验证 IBPR 算法的客观性。

1) 采用 4 种常用的用户影响力算法作为对比算法, 即共 5 种算法参与实验。对于每种算法分别计算出 Top- K 的用户及其对应影响力。

2) 构造数据集 I_M 表示任意 M 种算法均投票认为正确的结果, 计算公式如式 (15) 所示:

$$I_M = \bigcup (I_{M_i}), 1 \leq i \leq C_5^M \quad (15)$$

其中, M 表示交叉折数, I_{M_i} 表示在 5 种算法中随机选取 M 种算法得到的第 i 个交集, i 的取值有 C_5^M 种, 对所有 I_{M_i} 取并集得到数据集 I_M 。

假设算法 A 的准确率 (P_A)、召回率 (R_A) 和 F 值 (F_A) 计算公式如式 (16) ~ 式 (18) 所示:

$$P_A = \frac{|I_A \cap I_M|}{|I_A|} \quad (16)$$

$$R_A = \frac{|I_A \cap I_M|}{|I_M|} \quad (17)$$

$$F_A = 2 \times \frac{P_A \times R_A}{P_A + R_A} \quad (18)$$

其中, I_A 为算法 A 计算得到的用户影响力 Top-K 用户集合。

3.2 数据集

实验数据采用餐厅点评网站 YELP 提供的公开数据集,其主要为用户对餐厅的评论信息,在过滤活跃度小于 10 的用户后,筛选出的相关数据如表 1 所示。

表 1 实验数据集设置

Table 1 Setting of experimental dataset

参数	参数值
用户数	162 652
总边数	3 823 074
总博文数	3 303 172
人均发博数	20.31

3.3 参数设置

根据文献[21]设置,LDA 模型的兴趣话题数为 X 、 $\alpha = 50/X$ 、 $\beta = 0.01$ 、迭代次数为 2 000。IBPR 算法的最大迭代次数 $iter_{max} = 2\ 000$,迭代阈值 $\varepsilon = 10^{-9}$ 。考虑到平均评论数和平均被评论数能部分反映用户活跃度,因此设置 $\eta = 0.5$ 。考虑到用户兴趣转移率更能体现用户的隐性联系且不受僵尸粉的影响,因此设置用户兴趣传播转移率权值 $\mu = 0.6$,用户隐性兴趣传播影响力的阻尼系数 $d = 0.85$ 。

3.4 实验环境

实验使用 4 台虚拟机搭建 Hadoop 和 Spark 集群,每台虚拟机配置为双核 CPU 2.60 GHz、16 GB 内存、500 GB 硬盘,操作系统为 Ubuntu 16.04.3,实验集群设置如表 2 所示。

表 2 实验集群设置

Table 2 Setting of experimental cluster

主机 IP	主机名	集群角色
172.27.53.181	Master	NameNode 和 Master
172.27.53.182	Slave1	DataNode 和 Worker
172.27.53.183	Slave2	DataNode 和 Worker
172.27.53.184	Slave3	DataNode 和 Worker

3.5 结果对比

本文选取了目前较主流的 4 种用户影响力度量算法进行兴趣话题数分析、IBPR 算法有效性及客观性验证实验:1) PageRank 算法,由于 IBPR 算法是基于 PageRank 的改进算法,因此将 PageRank 算法作对比可以突出隐性兴趣因素,使结果更客观;2) TwitterRank 算法^[11],该算法融合了用户隐性兴趣,其作为对比算法用于验证用户综合影响力结果的合理性;3) 基于用户粉丝数与发博数的排名算法:FollowerRank 和 BlogRank^[22]。

3.5.1 兴趣话题数分析

兴趣话题数的确定考虑困惑度和平均话题相似度两个评价指标。困惑度倾向于选择大的主题数,容易造成话题间相似度较高,因此将两者综合考虑可以得到最优兴趣话题数。图 3 是 LDA 模型在不同兴趣话题数下的平均话题相似度和困惑度曲线。可以看出,当兴趣话题数取 55 时的平均话题相似度和困惑度值最低,因此本文确定最优兴趣话题数为 55。

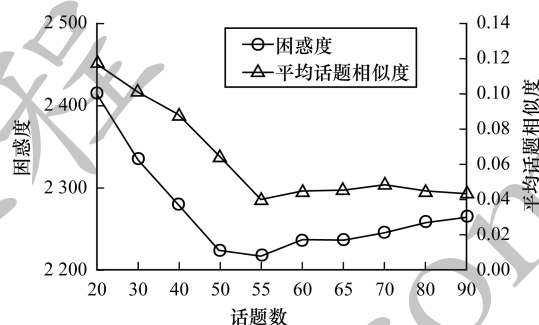


图 3 兴趣话题数与平均话题相似度和困惑度的关系

Fig. 3 The relationship between the number of topics of interest and the average topic similarity and perplexity

3.5.2 算法有效性验证

本文针对 M 取 2 和 3 情况下对 5 种算法进行交叉验证,比较 Top-K 用户 ($K = \{100, 200, \dots, 1\ 000\}$) 的准确率、召回率和 F 值。

准确率是衡量算法正确计算出 Top-K 用户占所有用户数量 K 的百分比。如图 4 所示,IBPR 算法在 M 和 K 取不同值时准确率均优于对比算法,其中 $M = 2$ 时各算法准确率相对 $M = 3$ 时要高约 5%,其主要原因为交叉折数为 2 时的集合 I_M 比交叉折数为 3 时的集合 I_M 多。

召回率表示算法正确识别影响力排名 Top-K 的用户占标准集合 I_M 的用户比例。图 5 表示 Top-K 影响力用户的召回率分布,从 $M = 2$ 和 $M = 3$ 两组实验结果可以看出,IBPR 算法在不同用户规模与交叉折数下准确率均优于对比算法,其中 $M = 3$ 时的召回率较高。

F 值是正确率和召回率的调和平均值,其综合考虑了准确率和召回率。图 6 表示各算法的 F 值比较结果,可以看出由于 TwitterRank 算法只考虑了与用户兴趣相关的影响力而忽略了其他因素,因此评估效果一般,而 PageRank 算法是基于用户好友关系,容易受到粉丝数目的影响以及僵尸粉的干扰,导致评估精度降低。FollowerRank 和 BlogRank 算法由于只考虑了用户自身属性的影响力,因此评估效

果也不理想。IBPR 算法相对对比算法具有明显优势, 主要因为其综合考虑了用户认证权威性、用户

活跃度、用户兴趣等因素, 能够更全面地评估用户影响力。

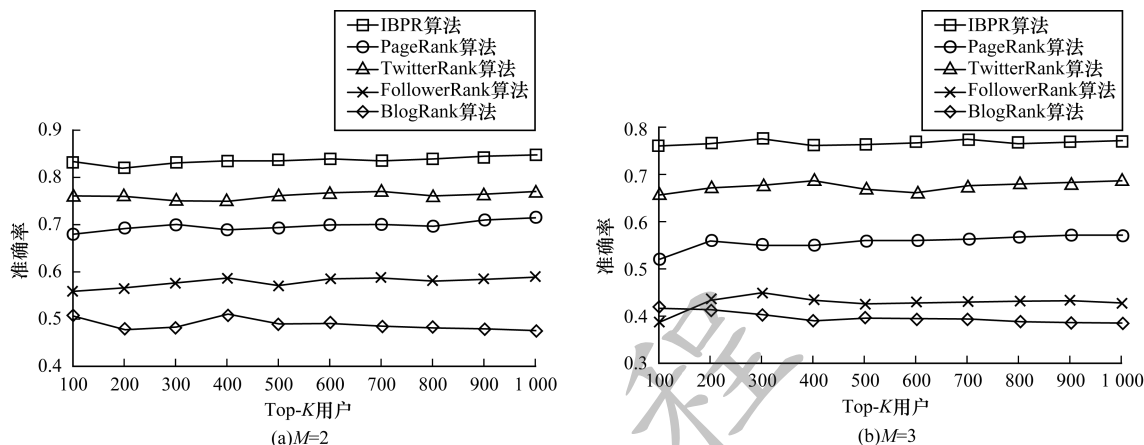


图 4 5 种算法在交叉验证中的准确率比较

Fig. 4 Comparison of the accuracy of five algorithms in cross-validation

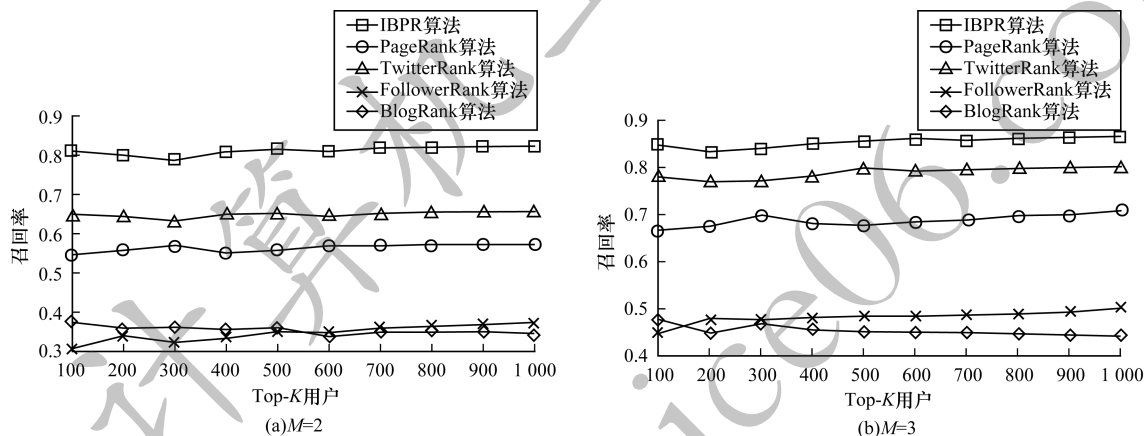


图 5 5 种算法在交叉验证中的召回率比较

Fig. 5 Comparison of the recall of five algorithms in cross-validation

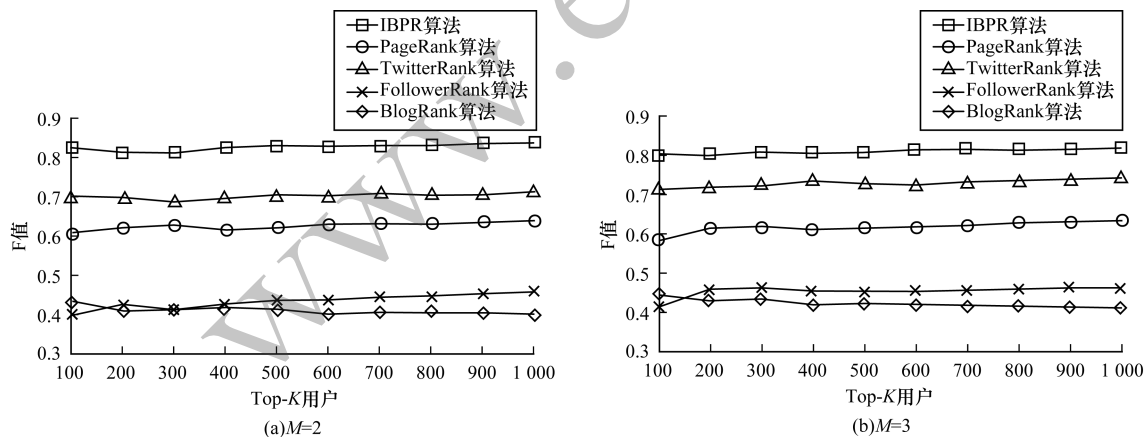


图 6 5 种算法在交叉验证中的 F 值比较

Fig. 6 Comparison of the F-value of five algorithms in cross-validation

3.5.3 算法客观性验证

本文选择 IBPR 算法计算的 Top-10 用户集合, 根据其对比算法中这 10 个用户排名位置变化进

行客观性分析。表 3 给出了 IBPR 算法的 Top-10 用户在对 PageRank 算法中的排名。实验结果表明, 两种算法的排名基本一致, 这是因为 IBPR 算法是基

于 PageRank 算法进行改进,但是 PageRank 算法仅考虑用户好友关系产生的影响,未考虑其他因素对于用户影响力的贡献。例如在 PageRank 算法分别排名为第 9 名与第 7 名的用户在 IBPR 算法中的排名为第 8 名与第 9 名,其原因主要为在 IBPR 算法排名第 8 名的用户认证权威影响力远大于排名第 9 名的用户,这一结果说明 IBPR 算法考虑了用户认证权威性对用户影响力的贡献,相对 PageRank 算法更全面客观。

表 3 IBPR 算法与 PageRank 算法排名对比
Table 3 Ranking comparison of IBPR algorithm and PageRank algorithm

IBPR 排名	用户名	PageRank 值	PageRank 排名
1	NhgU7RhuYYF	45.631	1
2	dIIKEfOgo0KqU	33.685	2
3	DkbTJFNSW4P	32.408	3
4	E43QxgV87Ij6	28.774	4
5	UYcmGbelzRa0	24.640	6
6	vRjVhl3ONG2G	24.829	5
7	UUqGHQFu2tQ	23.945	8
8	rMsB82tk9uOB6	23.880	9
9	jJDEwznWHQIa	24.242	7
10	AyYKTOCL5qM	22.761	10

TwitterRank 算法主要基于用户相似度来计算用户兴趣影响力,但是未考虑用户自身影响力。表 4 给出了 IBPR 算法的 Top-10 用户在对应 TwitterRank 算法中的排名。实验结果表明,用户隐性兴趣传播影响力虽然可以衡量用户影响力,但是不够全面,例如用户在 TwitterRank 算法中分别排名为第 9 名与第 8 名而在 IBPR 算法中的排名为第 8 名和第 9 名,其原因主要为 IBPR 算法中排名第 8 名的用户活跃度大于排名第 9 名的用户,这一结果说明 IBPR 算法考虑用户活跃度对用户影响力的贡献,相对 TwitterRank 算法更全面客观。

表 4 IBPR 算法与 TwitterRank 算法排名对比
Table 4 Ranking comparison of IBPR algorithm and TwitterRank algorithm

IBPR 排名	用户名	TwitterRank 值	TwitterRank 排名
1	NhgU7RhuYYF	0.281	1
2	dIIKEfOgo0KqU	0.279	2
3	DkbTJFNSW4P	0.273	3
4	E43QxgV87Ij6	0.297	4
5	UYcmGbelzRa0	0.275	5
6	vRjVhl3ONG2G	0.277	6
7	UUqGHQFu2tQ	0.267	7
8	rMsB82tk9uOB6	0.268	9
9	jJDEwznWHQIa	0.286	8
10	AyYKTOCL5qM	0.276	14

表 5 给出了 IBPR 算法的 Top-10 用户在对应 FollowerRank 算法中的排名。实验结果表明,粉丝数虽然对于衡量用户影响力有一定作用,但是也容易受到僵尸粉的干扰,例如在 IBPR 算法中排名第 4 名和第 9 名的用户在 FollowerRank 算法中分别排名为第 76 名与第 58 名,而 IBPR 算法中排名第 9 名的用户认证权威性远小于排名第 4 名的用户,这一结果说明 IBPR 算法考虑更全面客观,可以依据用户认证权威性来减少僵尸粉的干扰。

表 5 IBPR 算法与 FollowerRank 算法排名对比
Table 5 Ranking comparison of IBPR algorithm and FollowerRank algorithm

IBPR 排名	用户名	FollowerRank 值	FollowerRank 排名
1	NhgU7RhuYYF	5 026	26
2	dIIKEfOgo0KqU	6 187	8
3	DkbTJFNSW4P	3 732	46
4	E43QxgV87Ij6	2 758	76
5	UYcmGbelzRa0	2 916	69
6	vRjVhl3ONG2G	5 567	19
7	UUqGHQFu2tQ	7 854	2
8	rMsB82tk9uOB6	3 978	42
9	jJDEwznWHQIa	3 303	58
10	AyYKTOCL5qM	3 916	43

4 结束语

本文提出一种基于 Spark 与隐性兴趣的用户影响力度量算法,结合兴趣话题相似度重新定义 Pearson 相关系数,改进 PageRank 算法的转移率计算用户隐性兴趣传播影响力,并且采用层次分析法,综合用户自身影响力、用户行为和用户隐性兴趣传播影响力得到最终用户影响力,同时基于 Spark 平台加快用户综合影响力的计算速度。在公开数据集上的实验结果表明,该方法能更全面客观地评估用户影响力。后续将结合地理位置、评论关系等用户信息,进一步提高用户综合影响力的度量准确性。

参考文献

- [1] CNNIC. The 43st statistical report on the development of China's Internet network [EB/OL]. [2019-09-05]. <http://www.199it.com/archives/839540.html>. (in Chinese) CNNIC. 第 43 次中国互联网络发展状况统计报告 [EB/OL]. [2019-09-05]. <http://www.199it.com/archives/839540.html>.
- [2] HAN Zhongming, CHEN Yan, LIU Wen, et al. Research on node influence analysis in social networks [J]. Journal of Software, 2017, 28(1): 84-104. (in Chinese) 韩忠明, 陈炎, 刘雯, 等. 社会网络节点影响力分析研究 [J]. 软件学报, 2017, 28(1): 84-104.
- [3] GUO Hao, LU Yuliang, WANG Yu, et al. Measuring user influence of a microblog based on information diffusion [J]. Journal of Shandong University (Natural

- Science), 2012, 47(5): 78-83. (in Chinese)
- 郭浩, 陆余良, 王宇, 等. 基于信息传播的微博用户影响力度量[J]. 山东大学学报(理学版), 2012, 47(5): 78-83.
- [4] ZHANG Yangsen, ZHENG Jia, TANG Anjie. A quantitative evaluation method of micro-blog user authority based on multi-feature fusion [J]. Acta Electronica Sinica, 2017, 45(11): 2800-2809. (in Chinese)
- 张仰森, 郑佳, 唐安杰. 基于多特征融合的微博用户权威度定量评价方法[J]. 电子学报, 2017, 45(11): 2800-2809.
- [5] CHA M, HADDADI H, BENEVENUTO F, et al. Measuring user influence in Twitter: the million follower fallacy[C]//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Palo Alto, USA: AAAI Press, 2010: 10-17.
- [6] DING Zhaoyun, ZHOU Bin, JIA Yan, et al. Topical influence analysis based on the multi-relational network in microblogs [J]. Journal of Computer Research and Development, 2013, 50(10): 2155-2175. (in Chinese)
- 丁兆云, 周斌, 贾焰, 等. 微博中基于多关系网络的话题层次影响力分析[J]. 计算机研究与发展, 2013, 50(10): 2155-2175.
- [7] BAKSHY E, HOFMAN J M, MASON W A, et al. Everyone's an influencer: quantifying influence on Twitter [C]//Proceedings of ACM International Conference on Web Search and Data Mining. New York, USA: ACM Press, 2011: 65-74.
- [8] DUAN Songqing, WU Bin, WANG Bai. TTRank: user influence rank based on tendency transformation [J]. Journal of Computer Research and Development, 2014, 51(10): 2225-2238. (in Chinese)
- 段松青, 吴斌, 王柏. TTRank: 基于倾向性转变的用户影响力排序[J]. 计算机研究与发展, 2014, 51(10): 2225-2238.
- [9] WANG Chenxu, GUAN Xiaohong, QIN Tao, et al. Modeling on opinion leader's influence in microblog message propagation and its application [J]. Journal of Software, 2015, 26(6): 1473-1485. (in Chinese)
- 王晨旭, 管晓宏, 秦涛, 等. 微博消息传播中意见领袖影响力建模研究[J]. 软件学报, 2015, 26(6): 1473-1485.
- [10] MAO Jiaxin, LIU Yiqun, ZHANG Min, et al. Social influence analysis for micro-blog user based on user behavior [J]. Chinese Journal of Computers, 2014, 37(4): 61-70. (in Chinese)
- 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): 61-70.
- [11] WENG J S, LIM E P, JIANG J, et al. TwitterRank: finding topic-sensitive influential Twitterers [C]//Proceedings of the 3rd International Conference on Web Search and Web Data Mining. New York, USA: ACM Press, 2010: 261-270.
- [12] ZHANG Hao, LIU Gongshen, SU Bo. A computing method for microblogging users influence [J]. Computer Applications and Software, 2015, 32(3): 41-44. (in Chinese)
- 张昊, 刘功申, 苏波. 一种微博用户影响力的计算方法[J]. 计算机应用与软件, 2015, 32(3): 41-44.
- [13] WANG Ding, XU Jun, DUAN Cunyu, et al. Improved user influence evaluation algorithm based on PageRank [J]. Journal of Harbin Institute of Technology, 2018, 50(5): 60-67. (in Chinese)
- 王顶, 徐军, 段存玉, 等. 基于 PageRank 的用户影响力评价改进算法[J]. 哈尔滨工业大学学报, 2018, 50(5): 60-67.
- [14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
- [15] SUN Yujie, QIN Yongbin. Multi-angle personalized microblog recommendation algorithm based on LDA model [J]. Computer Engineering, 2017, 43(4): 177-182. (in Chinese)
- 孙玉洁, 秦永彬. 基于 LDA 模型的多角度个性化微博推荐算法[J]. 计算机工程, 2017, 43(4): 177-182.
- [16] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark: cluster computing with working sets [C]//Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing. New York, USA: ACM Press, 2010: 1-10.
- [17] SAATY T L. Decision making with the analytic hierarchy process [J]. International Journal of Services Sciences, 2008, 1: 83-98.
- [18] CAO Jie, SU Zhe, LI Xiaoxu. Image annotation method based on Corr-LDA model [J]. Journal of Jilin University (Engineering and Technology Edition), 2018, 48(4): 1237-1243. (in Chinese)
- 曹洁, 苏哲, 李晓旭. 基于 Corr-LDA 模型的图像标注方法[J]. 吉林大学学报(工学版), 2018, 48(4): 1237-1243.
- [19] WANG Ye, ZUO Wanli, WANG Ying. Short-text clustering algorithm based on extension of metaphorical words [J]. Journal of Jilin University (Science Edition), 2018, 56(6): 1447-1452. (in Chinese)
- 王烨, 左万利, 王英. 基于隐喻词扩展的短文本聚类算法[J]. 吉林大学学报(理学版), 2018, 56(6): 1447-1452.
- [20] GAO Chuansong. Design and implementation of question answering system based on SVM text classification [D]. Nanjing: Nanjing University, 2014. (in Chinese)
- 高传嵩. 基于 SVM 文本分类的问答系统的设计与实现[D]. 南京: 南京大学, 2014.
- [21] YAN Xiaohui, GUO Jiafeng, LAN Yanyan, et al. A bitern topic model for short texts [C]//Proceedings of the 22nd International Conference on World Wide Web. New York, USA: ACM Press, 2013: 1445-1456.
- [22] HUANG Xianying, YANG Anzhi, LIU Xiaoyang, et al. An improved algorithm for microblog user influence evaluation [J]. Computer Engineering, 2019, 45(12): 294-299. (in Chinese)
- 黄贤英, 阳安志, 刘小洋, 等. 一种改进的微博用户影响力评估算法[J]. 计算机工程, 2019, 45(12): 294-299.