基于 MapReduce 的新型微博用户影响力排名算法研究

徐文涛 刘 锋 朱二周
(安徽大学计算机科学与技术学院 合肥 230601)

摘 要 微博凭借其即时发布、实时传播、简便易用的特点逐渐成为最为主流的自媒体平台。用户影响力评价是微博社交网络中基本而又重要的问题,它对于优化与推动社会信息传播来说有着重要意义。以新浪微博为实验对象,通过综合考虑微博用户关系网络特性和用户行为,结合 MapReduce 编程计算模型,提出了一种基于 MapReduce 的新型用户影响力排名算法——QRank。在 Hadoop 平台上的实验结果表明,QRank 算法具有良好的可扩展性,能够有效结合微博用户关系网络与行为特性,从而更加真实与充分地反映用户的实际影响力。

关键词 PageRank 算法,MapReduce,用户影响力,Hadoop 平台
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.012

Research on Novel Ranking Algorithm of Microblog User's Influence Based on MapReduce

XU Wen-tao LIU Feng ZHU Er-zhou
(School of Computer Science and Technology, Anhui University, Hefei 230601, China)

Abstract Featured by instant release, real-time transmission and easy to use, microblog has gradually stepped into the rank of the most popular self-media information platform. User's influence, which is of great importance to optimize and motivate social information transmission, plays a basic as well as important role in microblog social network. Taking into account the network features of microblog users' relationship as well as their behaviors, taking Sina microblog as the experimental subject, this paper aimed to introduce the QRank algorithm, a new ranking algorithm based on MapReduce to judge user's influence. An experiment on the Hadoop platform shows that, with great scalability, QRank algorithm can effectively combine the relationship and behavior features of microblog users and reflect the real influence of users in a more convincing and sufficient way.

Keywords PageRank algorithm, MapReduce, User's influence, Hadoop platform

1 引言

随着网络技术的迅速发展,社交网络在推动信息传播中起到了重要的引导作用[1,2],它将传统的以内容和话题为基础的信息传播方式发展为以人际关系为基础。作为社交网络的一种重要形式,微博有着即时发布、互动性强、简便易用等特点,并迅速发展为最为主流的社交平台。据 CNNIC 发布的第 35 次《中国互联网络发展状况统计报告》数据显示[3],截至 2014 年 12 月,我国微博客用户规模为 2.49 亿,而新浪微博在微博社交网络中处于绝对领先地位。用户通过发布微博、分享转发以及参与讨论使得网络信息传播得更加迅速。

面对大规模的微博用户群体,微博用户的影响力作为其基本特征吸引了广大学者对此进行研究。在微博网络中,影响力较大的用户拥有更多的粉丝数和关注度,他们所发布或转载的信息会得到更多关注,在较大程度上影响了信息的传播[2]。但在实际环境中,很多用户通过人为增加粉丝的作弊手段提高自身的影响力[4],而这些人为的粉丝("僵尸粉"、"水军")并不能代表真实的用户关系。若不法分子利用该手段加速不良信息的传播,则容易造成舆论恐慌。另一方面,微博用户数据规模越来越大,通过传统的单机环境计算用户影响力排名对内存的需求很高,在海量数据盛行的今日,大规模数据的并行化处理成为必然。

在数据挖掘和机器学习领域中,存在着众多并行处理方法,如 MPI 并行编程工具、GraphLab 并行框架、MapReduce 编程模型等。对于不同的机器学习算法,MPI 方式的代码复用率低,需重写其数据分配、通信等实现细节,对编程人员要求高,而面向机器学习的 GraphLab 并行框架没有提供容错功能。对比来看,MapReduce 编程模型不要求编程人员具备很强的并行算法设计能力,Map 和 Reduce 两个函数简化了并行算法的设计难度,集群中的每台机器各自负载计算过程,数据的并行程度较高,对数据统计类算法并行化效果明显,学术界和工业界也较多地使用 MapReduce 编程模型对大规模数据进行并行化处理。

当前存在一些微博影响力排名算法,如考虑博文相关因素的博主影响力的评估方法[5]、使用 MapReduce 框架计算用户排名的方法[6]、使用基于用户质量的 UIR 排序算法[7]以及

到稿日期:2015-07-16 返修日期:2015-09-03 本文受国家自然科学基金(61300169)资助。
徐文涛(1992-),男,硕士生,主要研究方向为大数据、数据挖掘,E-mail:994886058@qq.com;刘 锋(1962-),男,教授,硕士生导师,主要研究方向为高性能计算、并行计算;朱二周(1981-),男,硕士生导师,主要研究方向为虚拟化、移动云计算,E-mail:ezzhu@ahu.edu.cn(通信作者)。

基于用户行为的评价方法[8]等。但这些方法均存在一定的局限性，如文献[5]提出的 Influence Rank 算法加入了博文相关评价指标来弥补单纯依靠用户关系的不足，但其在时间参量的选择等方面上具有局限性；文献[6]中的评价方法过于单一，只是将单纯地将 PageRank 算法应用于用户关系网络，没有考虑微博网络及用户的基本特征；文献[7]没有考虑到把用户行为作为影响因素，且在单机环境下实现算法具有局限性；文献[8]考虑并分析了用户的行为特性，但没有考虑算法的并行实现。

本文利用 PageRank 网页排名算法，以新浪微博为例，综合考虑微博用户的关系网络特性与用户行为，提出了基于 MapReduce 编程模型的新型用户影响力 QRank 排名算法。第 2 节介绍了相关知识；第 3 节为提出的 QRank 算法的具体设计与实现；第 4 节是实验结果与分析；最后对全文进行总结，并对下一步工作进行展望。

2 相关知识介绍

2.1 PageRank 算法

PageRank 算法[9]是一种基于链接分析的网页排序算法，也是 Google 搜索引擎的核心算法。它建立在一个随机浏览者模型上，利用 Web 网络的庞大链接结构来作为单个网页质量的参考。本质上，PageRank 算法将网页 V 到网页 U 的链接当作是一种投票行为，由网页 V 投票给网页 U[10]。网页的 PageRank 值(PR 值)由链入到它的网页数目和这些入链网页的质量共同决定。

将整个 Web 网络看作是一个有向图 $G=(V,E)$，其中 V 是所有网页的集合，E 是所有有向边(即超链接)的集合，则该 Web 网络的 PageRank 计算公式可以表示为：

$$PR(u)=(1-d)+d\times\sum_{i=1}^{n}\frac{PR(v_i)}{C(v_i)} \tag{1}$$

其中，$PR(u)$ 表示网页 u 的 PR 值，v_i 是链接到 u 的一个网页，$PR(v_i)$ 是该页面的 PR 值，$C(v_i)$ 是网页 v_i 向外链接的网页总数，d 表示衰减系数(通常取 0.85)。在该模型中的任何一个网页上，一个随机浏览者随机选择一个链出链接进而继续浏览的概率是 d，而不通过点击链接直接跳到另一个随机网页的概率是 $1-d$。利用式(1)进行迭代计算，当网页的 PR 值不再明显变化或收敛时停止迭代。

微博用户相互关注关系与网页链接结构极其相似。通过“把用户的粉丝看作是用户的入链，用户的关注看作用户的出链”的方式便可以将 PageRank 算法运用到用户影响力排名中。

然而，随着微博的迅速发展，用户大量购买粉丝等作弊现象日益严重，微博网络中用户质量也存在差异。这使得PageRank算法在迭代过程中平均分配用户权值的方法显得不够合理[11]。

例如，在微博用户网络中，可能会出现如下情况：两个用户的 PR 值是相同的，但是不能绝对地理解为这两个用户的影响力是等效的，还需要考虑用户的粉丝集合。

如图 1 所示，图中每个图形都代表一个微博用户，图片的大小表示用户 PR 值的高低。用户 A 的粉丝贡献值源于他的 3 个粉丝，用户 B 的粉丝贡献值源于他的 6 个粉丝，用户 A 与 B 的 PR 值是相同的。然而在现实情景中，用户 A 的影响力应该是大于用户 B 的。因为用户 A 的粉丝活跃度高，影响力较大；而 B 的粉丝活跃度低，有可能是通过作弊手段购买的大量“水军”和“僵尸粉”，所以在粉丝投票过程中用户 A 应该被赋予更高的权值。

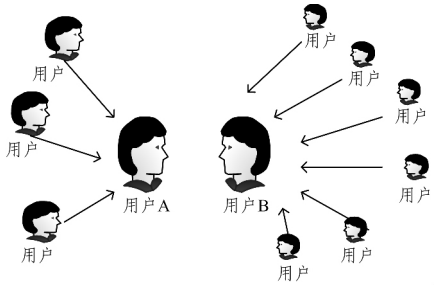

图 1 用户质量比较

考虑到微博用户实际行为特征，一个微博用户的影响力不仅与其入度(粉丝数)有关，还与用户认证、微博发布频率以及转发次数等有关。因此，直接利用 *PageRank* 算法来评估微博用户的影响力，对其粉丝贡献的 PR 值进行迭代叠加，并不能客观地反映出用户的真实影响力，必须综合考虑微博用户的粉丝数量、微博的转发率、评论率等特征。由于每个用户在微博网络中的影响力不同，且用户的影响力又直接影响了网络中其他微博用户的影响力，因此在粉丝投票的迭代过程中需要采用不均匀分配权值的方法。

2.2 MapReduce 编程模型

面对大规模的用户数据，合理有效的并行化处理显得尤为重要。近年来，Hadoop 框架[12]成为大数据的重要标签之一。Hadoop 起源于开源的网络搜索引擎 Nutch，其最主要的两个组成部分是分布式文件系统 HDFS 和并行编程模型 MapReduce。

MapReduce 是在 Google MapReduce 的基础上实现的并行编程模型[13,14]，它将复杂的并行计算过程抽象化为 Map 和 Reduce 过程，编程人员只需利用 Hadoop 所提供的简单易用的编程接口就能实现这两个函数，极大地降低了分布式编程的门槛，使更多的编程人员能够实现并行算法。而分布式文件系统 HDFS 利用文件块 Block 将数据分布存储在各个数据节点上，并且通过备份冗余机制为整个框架提供容错能力。

MapReduce 通过〈*Key*，*Value*〉键值对来处理数据，处理过程描述如表 1 所列。MapReduce[15]会对输入的数据集进行 Split 逻辑分片，由用户自定义的 Map 函数读取分片数据并将其转换为一个〈*K*1，*V*1〉键值对列表，框架将 Map 函数的输出进行 Shuffle 和 Sort 处理后，再将结果以〈*K*2，*List*(*V*2)〉形式输入给 Reduce 函数。Reduce 函数接收〈*K*2，*List*(*V*2)〉键值对数据，根据相应的功能对这些 *Value* 进行处理，最后以〈*K*3，*V*3〉形式将计算结果输出到分布式文件系统 HDFS 上。

表 1 Map 和 Reduce 函数

函数	输入	输出
Map	〈K1，V1〉	List(〈K2，V2〉)
Reduce	〈K2，List(V2)〉	〈K3，V3〉

MapReduce 执行流程如图 2 所示。

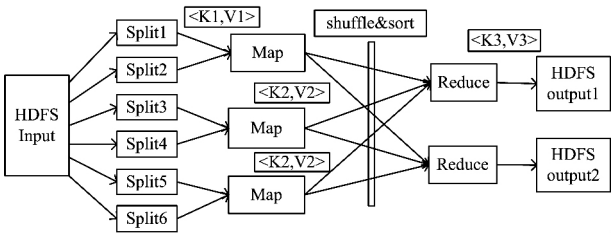

图 2　MapReduce 流程图

2.3　Influence Rank 算法

本文以 PageRank 算法以及文献[5]中的 Influence Rank 算法(简称 IR)为主要对比算法。IR 算法是基于 Page-Rank 的改进算法。该算法分别给出了博文质量系数、用户近期活跃度、博主传播能力的相应定义。

博文质量系数：

$$q_u=\frac{R_u+C_u}{N_u} \tag{2}$$

用户近期活跃度：

$$a_u=\frac{n_u}{t} \tag{3}$$

博主传播能力：

$$s_u=q_u\times a_u \tag{4}$$

其中，R_u 表示用户 u 总的微博被转发次数，C_u 表示用户 u 总的微博被评论数，N_u 表示用户 u 所发布的微博总数，q_u 表示用户 u 所发微博的质量指数；n_u 表示用户 u 在时间 t 内发布微博的数量，a_u 表示用户 u 的近期活跃度；s_u 表示博主 u 的传播能力，由博文质量系数和用户活跃度两个参数定义。

该算法的计算公式如下：

$$IR(I)=1-d+d\times\sum_{V\in K(I)}S(V,I)IR(V) \tag{5}$$

$$S(V,I)=\frac{s_I}{\sum_{u=1}^{N}s_u} \tag{6}$$

其中，N 表示用户 V 的好友数，$S(V,I)$ 表示用户 V 分配给 I 的 IR 值比例，$K(I)$ 表示用户 I 的粉丝集合。

IR 算法考虑了用户的微博发布数目和博文传播能力，并引入了多个相关评价指标，一定程度上弥补了单纯地依靠用户关系来计算用户排名的不足。但该方法在计算用户活跃度时对时间 t 的考虑不尽合理，具有一定的局限性。此外，定义博文质量考虑的是博文的整体质量，有一定的片面性，也没有考虑算法的并行设计方法，不能满足处理大规模用户数据的需要。

3　QRank 用户影响力模型

3.1　用户相对质量

为了解决 PageRank 算法在迭代过程中平均分配权值不够合理的问题，本文提出考虑用户质量的 QRank 算法，引入用户相对质量的概念，利用每次迭代后的用户 QRank 值与用户相互关注关系结构来计算用户相对质量，并将其应用到用户 QRank 值的分配中。

首先定义用户相对质量 $Q(I)$：

$$Q(I)=\frac{QRank(I)}{\sum_{V\in K(I)}QRank(V)/MQRank(I)} \tag{7}$$

$$MQRank(I)=\max_{V\in K(I)}QRank(V) \tag{8}$$

其中，$K(I)$ 表示用户 I 的粉丝集合，$MQRank(I)$ 表示 I 的粉丝集合中用户的最大 QRank 值。

通过以上对用户相对质量的分析，结合 PageRank 算法初步给出 QRank 算法公式：

$$QRank(I)=(1-d)+d\sum_{V\in K(I)}QRank(V)\times Pt(V\to I) \tag{9}$$

$$Pt(V\to I)=\frac{Q(I)}{\sum_{t\in O(V)}Q(t)} \tag{10}$$

其中，$K(I)$ 表示用户 I 的粉丝集合，$Pt(V\to I)$ 表示用户 V 投票给用户 I 的 QRank 值比重，$O(V)$ 表示用户 V 所关注的用户集合。

图 3 是一个考虑用户相对质量分配 QRank 值的例子，按照传统的 PageRank 算法，用户 C 分配给用户 A 和 B 的 QRank 值均为 3，而根据改进后的算法得：

$$Q(A)=\frac{10}{(2+3+1+6)/6}=5$$

$$Q(B)=\frac{10}{(4+6)/6}=6$$

$$Pt(C\to A)=\frac{5}{(5+6)}=\frac{5}{11}$$

$$Pt(C\to B)=\frac{6}{(5+6)}=\frac{6}{11}$$

因此按照上述算法思想，用户 C 在投票过程中会多分配 QRank 值给用户 B 来提高用户 B 的影响力。

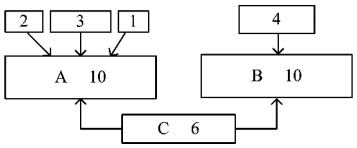

图 3　考虑用户相对质量的 QRank 分配图

3.2　QRank 算法

QRank 算法不仅考虑了用户自身质量及其粉丝质量这两方面因素，还综合考虑了用户微博评论率、转发率、是否微博认证等因素。首先对微博相关属性信息做出定义[16]。

用户的微博转发率：

$$R(I)=\frac{SumR(I)/SumP(I)}{Sum(User)} \tag{11}$$

用户微博评论率：

$$C(I)=\frac{SumC(I)/SumP(I)}{Sum(User)} \tag{12}$$

用户是否微博认证：

$$V(I)=(e,0) \tag{13}$$

其中，$SumR(I)$ 表示用户 I 的微博被转发的总次数，$SumP(I)$ 表示用户 I 的发布的微博总数，$SumC(I)$ 表示用户 I 的微博被评论的总次数，$Sum(User)$ 表示微博总用户数，$V(I)$ 表示用户 I 是否认证；e 是一个设定常数，表示认证用户的加权值(本文中 e 取 0.5)，若用户 I 是微博认证用户，则 $V(I)=e$，反之 $V(I)=0$。

结合上述信息，将 QRank 算法进一步定义为：

$$QRank(I)=QR\_self(I)+QR\_follow(I) \tag{14}$$

其中，

$$QR\_self(I)=R(I)+C(I)+V(I) \tag{15}$$

$$QR\_follow(I)=(1-d)+d\sum_{V\in K(I)}QRank(V)\times Pt(V\to I) \tag{16}$$

其中，I，V 表示微博用户，$K(I)$ 表示用户 I 的粉丝集合，用户

V 是用户 I 的一位粉丝，$QRank(I)$，$QRank(V)$ 表示用户 I，V 的 QRank 影响力值，$QR\_self(I)$ 表示用户 I 的自身质量指数，$QR\_follow(I)$ 表示用户 I 的粉丝贡献的影响力指数，即用户 I 的粉丝质量指数。微博用户影响力由自身质量指数与粉丝质量指数共同组成。

下面给出收敛性证明，将式(16)进一步分解得：

$$QR\_follow(I)=(1-d)+Pt(V\rightarrow I)\times d\sum_{V\in K(I)}(QR\_self(V)+QR\_follow(V)) \quad (17)$$

其中，$QR\_self(V)$ 与 $Pt(V\rightarrow I)$ 均为定值，所以根据 PageRank 算法迭代收敛易知式(17)也收敛，结合式(14)、式(15)、式(17)可知 QRank 算法也服从迭代收敛。

由上述分析过程可知，QRank 算法属于数据统计类计算，且需要迭代处理。整个计算过程部署在 Hadoop 平台下，基于 MapReduce 计算模型的 QRank 算法的伪代码描述如下。

```
Map(key,value)
{ //key 为用户 V
//value 为(QRank(V),List[K1,K2,…,Kn])
//list 为用户 V 的关注集合
OutputV→List[K1,K2,…,Kn]//用于下次迭代
foreachKinList
  QR_self(K)=R(K)+C(K)+V(K)
outputK→QR_self(K)
  Pt(V→K)=Q(K)/∑t∈O(V) Q(t)
outputK→QRank(V)Pt(V→K)
}
Reduce(key,value)
{ //key 为用户 K
    //value 为 List of (QRank(V)Pt(V→K)),
List[K1,K2,…,Kn]
QRank(K)=(1-d)+QR_self(K)
foreachV in List//V 为用户 K 的粉丝
    QRank(K)+=QRank(V)Pt(V→K)×d
outputK→QRank(K)
}
```

集群中的各个数据节点实现了本地计算，降低了对网络带宽的需求。同时，数据节点易于扩充，集群所提供的计算能力也随之增长，因此 QRank 算法具有良好的可扩展性。为了减少迭代过程中的时间开销，在一次迭代后会将中间结果与用户关注关系合并作为下一次的迭代输入，避免了两次访问 HDFS，达到预先设定的收敛阈值后，停止迭代。

3.3 算法优势

QRank 算法考虑了微博用户基本特性，如微博转发率、评论率以及粉丝质量，并结合用户相对质量对粉丝贡献值作不平均分配，改进后的算法更全面且更加符合实际，能更好地衡量用户的实际影响力。

QRank 算法与单纯的 PageRank 算法相比，算法时间复杂度有所增加，需要计算用户 I 的自身质量指数 $QR\_self(I)$ 值，以及 $Q(I)$ 值、$Pt(V\rightarrow I)$ 值。但它们计算简单，仅需要用户自身微博信息及相互关注关系，因此复杂度增加幅度很小，且该算法使得高影响力用户的 QRank 值可以更快地积累。而利用 MapReduce 编程模型来实现 QRank 算法的并行化处理也使算法更具扩展性，处理大规模用户数据更加合理有效。

4 实验

4.1 实验数据与环境

本文通过新浪 API 爬取新浪微博实验数据(受 API 接口不断收缩的限制，获取的数据并不完整)。实验数据中包括 5971540 个微博用户数据，其中包含用户微博基本信息、用户相互关注关系等。实验数据并不太完整，对比实验也是基于该数据集进行统计计算，实验结果更具合理性。为了方便做对比实验，本文对 PageRank 算法、IR 算法的计算过程也进行了 MapReduce 并行化设计。

实验使用 4 台机器搭建 Hadoop 集群模式，其中每台机器的配置均为：Pentium(R) Dual-CoreE6700 3.20GHz，内存 DDR3 2GB，硬盘 500GB。在每台主机上安装 Ubuntu14.04、Hadoop-2.6.0、OpenJDK1.7、SSH 等，并进行相关文件配置，HDFS Block 大小采用默认值。具体集群概况如表 2 所列。

表 2 集群信息

IP 地址	主机名	集群角色
172.19.198.52	Master	NameNodeSecondaryName NodeResourceManager
172.19.198.73	Slave1	DataNode NodeManager
172.19.198.117	Slave2	
172.19.198.61	Slave3	

4.2 数据处理

爬取到的实验数据并非形成一个理想的网络连通图，需对数据进行预处理，去除用户链接关系网络外的用户节点后剩余 3574983 个微博用户数据。为了便于计算 QRank 值，结合用户微博基本信息，提取出每个用户的认证信息、用户微博总数、微博转发次数和微博评论次数等有效字段，同时记录微博总条数，从而计算出微博平均转发率和平均评论率。同时，为了减少数据传输开销，将微博用户相互关注关系转换成用户链接结构表，只保留用户 ID 标识号，如表 3 所列。

表 3 用户链接结构表

用户 ID	所关注用户 ID 列表
1	2 3 4
2	3 5
3	4 6

因为用户相互关注关系表数据量较大，所以由用户关系表转换成用户链接结构表的过程也使用 MapReduce 过程来处理[15]。根据用户链接关系结构，计算出用户相对质量及 QRank 值分配比重，最终迭代计算出微博用户的 QRank 值。

4.3 实验结果

本文以实验数据为基础，对实验数据产生的结果进行对比分析。对提出的 QRank 算法计算出的用户影响力进行降序排序，并列出影响力排名前 20 的用户，结果如表 4 所列。

由表 4 可以看出：影视明星等知名人士的用户影响力较高，他们的粉丝众多且活跃度高。同时，公众服务类微博账号的影响力也很大，如“头条新闻”作为实时新闻类账号获得了大量粉丝关注，并且转发、评论率较高。由此可看出如今很多微博用户习惯于在微博中查看社会新闻信息，此类用户对于社会信息传播也具有很强的引导性。大多数影响力大的用户都是微博认证用户，这增加了群众对他们的信任感。

表 4　QRank 算法 Top 20 的用户信息

排名	用户名	粉丝数量	平均转发次数	平均评论次数	认证
1	何炅	81436	2143	609	是
2	小 S	73587	1972	912	是
3	头条新闻	67432	1643	538	是
4	李开复	41472	1256	207	是
5	姚晨	39783	901	412	是
6	微博小秘书	59696	312	104	是
7	文章同学	40057	704	246	是
8	微博客服	52319	204	131	是
9	冷笑话精选	22049	743	89	否
10	谢娜	17359	191	47	是
11	杨幂	14892	347	192	是
12	微博 iphone 客户端	24349	52	17	是
13	赵薇	15486	296	93	是
14	微博 Android 客户端	20197	45	61	是
15	梦想家林志颖	11469	172	148	是
16	微博搞笑排行榜	8037	593	209	否
17	范范范玮琪	9431	165	78	是
18	angelababy	6213	343	123	是
19	王力宏	4982	136	129	是
20	人民日报	7036	32	61	是

QRank 影响力排名前 20 的用户与微博平均转发、评论次数的相关性如图 4 所示。

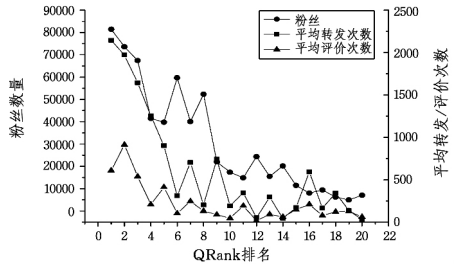

图 4　QRank 影响力排名与用户信息相关性

结合表 4 和图 4 来看，用户影响力与用户粉丝、用户活跃度呈一定的正相关性。同时也可以看到“微博小秘书”、“微博客服”用户的粉丝非常多，但是其影响力排名并没有相应的那么高。分析得知该类账号是微博官方账号，一般用户微博开通后会自动默认推荐其关注该类账号，而大多数人没有去手动取消关注，导致其粉丝数量很大，但其发布的微博转发与评论率相对而言不是很高。综合来看，QRank 算法对用户影响力进行排名时综合考虑了粉丝数量与用户活跃度，一定程度上降低了“僵尸粉”的干扰影响，排名结果也显得更加真实可信。

对比 QRank 算法与 PageRank 算法排名、粉丝数量排名、IR 算法排名，并列出排名前 20 的用户，如表 5、表 6 所列。

从表 5 可以看出，PageRank 算法在计算用户影响力时没有考虑微博用户自身行为，相对而言更多地依赖于粉丝，没有考虑自身活跃度对社交网络信息传播的影响；而粉丝数量排名只是单一对粉丝数进行统计，过于片面，一定程度上会受到作弊手段的干扰。

结合表 4、表 6 可以看出，IR 算法一定程度上降低了“僵尸粉”、“水军”的影响，但是当某账号粉丝数量较多时，其效果并不太明显。例如，“微博小秘书”的活跃度并不高，但其 IR 排名较高，IR 算法在计算博主传播能力时限制较多，过于依赖微博信息，导致在迭代过程中分配用户权值时不尽合理。从表 6 中结果对比来看（如“微博 iphone 客户端”、“杨幂”等账号），IR 算法没有很好地结合粉丝数量与用户活跃度等特性，而 QRank 算法排名结果则相对真实合理。在算法设计上，QRank 算法基于 MapReduce 编程环境，使得处理大规模用户数据变得切实可行。综合对比来看，QRank 算法更加全面灵活，更加注重活跃用户对信息传播的影响，也更具合理性。

表 5　QRank 与 PageRank、粉丝数量对比

排名	QRank 排名	PageRank 排名	粉丝数量排名
1	何炅	何炅	何炅
2	小 S	小 S	小 S
3	头条新闻	头条新闻	头条新闻
4	李开复	微博小秘书	微博小秘书
5	姚晨	李开复	微博客服
6	微博小秘书	微博客服	李开复
7	文章同学	姚晨	文章同学
8	微博客服	文章同学	姚晨
9	冷笑话精选	微博 iphone 客户端	微博 iphone 客户端
10	谢娜	冷笑话精选	冷笑话精选
11	杨幂	谢娜	微博 Android 客户端
12	微博 iphone 客户端	微博 Android 客户端	谢娜
13	赵薇	杨幂	赵薇
14	微博 Android 客户端	赵薇	杨幂
15	梦想家林志颖	梦想家林志颖	梦想家林志颖
16	微博搞笑排行榜	范范范玮琪	范范范玮琪
17	范范范玮琪	微博搞笑排行榜	微博搞笑排行榜
18	angelababy	人民日报	人民日报
19	王力宏	angelababy	angelababy
20	人民日报	王力宏	veggieg

表 6　QRank 与 IR 排名结果对比

排名	QRank 排名	IR 排名
1	何炅	何炅
2	小 S	小 S
3	头条新闻	头条新闻
4	李开复	微博小秘书
5	姚晨	姚晨
6	微博小秘书	李开复
7	文章同学	微博客服
8	微博客服	文章同学
9	冷笑话精选	冷笑话精选
10	谢娜	微博 iphone 客户端
11	杨幂	谢娜
12	微博 iphone 客户端	微博 Android 客户端
13	赵薇	赵薇
14	微博 Android 客户端	杨幂
15	梦想家林志颖	梦想家林志颖
16	微博搞笑排行榜	范范范玮琪
17	范范范玮琪	微博搞笑排行榜
18	angelababy	angelababy
19	王力宏	人民日报
20	人民日报	王力宏

结束语　用户影响力作为社交网络数据挖掘领域的研究热点，对社交网络中舆论管理、营销合作起到了重要作用。本文分析了微博社交网络的用户关系网络特征，结合 PageRank 排名算法，以 Hadoop 环境为主要计算平台，提出了 QRank 用户影响力评价算法。该算法综合考虑用户间关注和行为特征，从用户的粉丝质量和微博质量两方面来考量用户的影响力。实验结果表明，QRank 算法模型注重用户互动行为，排名结果更加全面、现实。未来的研究工作当中将把爬取数据、分析数据、计算排名等步骤全部部署在 Hadoop 环境中进行，

（下转第 86 页）

in text categorization[C]// Fourteenth International Conference on Mechine Learning. 1997:412-420

[15] Forman G. An extensive empirical study of feature selection metrics for text classification[J]. The Journal of Machine Learning Research,2003,3(2):1289-1305

[16] Zelikovitz S, Hirsh H. Using LSI for text classification in the presence of background text[C]// Proceedings of the Tenth International Conference on Information and Knowledge Management. ACM,2001:113-118

[17] Seifert C,Ulbrich E,Kern R,et al. Text Representation for Efficient Document Annotation[J]. J. UCS,2013,19(3):383-405

[18] Lewis D D. Feature selection and feature extraction for text categorization[C]// Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, 1992:212-217

[19] Ding C H Q. A similarity-based probability model for latent semantic indexing[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM,1999:58-65

[20] Lee D D,Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature,1999,401(6755):788-791

[21] Tan S B, Wang Y F. Chinese text categorization corps-TanCorp-V1.0[OL]. [2014-4-13]. http://www.searchforum.org.cn/tansongbo/corpus.htm

[22] Zhang H P. The Chinese academy of sciences segmentation kit [OL]. [2014-4-13]. http://www.ictclas.org

[23] He L,Wang Z Y,Jia Y,et al. Category candidate search in large scale hierarchical classification[J]. Chinese Journal of Computers,2014,31(1):41-49

[24] Zhang Yu-fang,Wang Yong,Liu Ming,et al. New feature selection approach for text categorization[J]. Computer Engineering and Applications,2013,49(5):132-135

(上接第 70 页)

并且将用户影响力细化到各个领域,根据领域特点进行相应的算法改进。

参 考 文 献

[1] Lin Jia-li, Li Zhen-yu, Wang Dong, et al. Analysis and Comparison of Interaction Patternsin online Social Network and Social Media[C]// Proc of the 21st International Conference on Computer Communications and Networks. Munich,Germany,2012:1-7

[2] Wu Xin-dong,Li Yi,Li Lei. Influence Analysis of Online Social Networks[J]. Chinese Journal of Computers,2014,37(4):735-752(in Chinese)
吴信东,李毅,李磊. 在线社交网络影响力分析[J]. 计算机学报,2014,37(4):735-752

[3] Statistic Report of the 35th China Internet Developing Situation [R]. Beijing:China Internet Network Information Center,2015 (in Chinese)
第 35 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心,2015

[4] Zhang Qun-yan,Ma Hai-xin,Qian Wei-ning,et al. Duplicate Detection for Identifying Social Spamin Microblogs[C]// Proc of the IEEE International Congress on Big Data. Santa Clara,CA 2013:141-148

[5] Yang Chang-chun,Yu Ke-fei,Ye Shi-ren,et al. New Assessment Method on Influence of Bloggers in Community of Chinese Microblog[J]. Computer Engineering and Applications,2012,48(25):229-233(in Chinese)
杨长春,俞克非,叶施仁,等. 一种新的中文微博社区博主影响力的评估方法[J]. 计算机工程与应用,2012,48(25):229-233

[6] Liang Qiu-shi,Wu Yi-lei,Feng Lei. User Ranking Algorithm for Microblog Search Based on MapReduce[J]. Journal of Computer Applications,2012,32(11):2989-2293(in Chinese)
梁秋实,吴一雷,封磊. 基于 MapReduce 的微博用户搜索排名算法[J]. 计算机应用,2012,32(11):2989-2993

[7] Tang Fei-long, Ye Shi-ren, Xiao Chun. Blogger Influence Ranking Algorithm Based on User Quality in Sina Microblog Community[J]. Computer Engineering and Applications, 2015, 51(4):128-132(in Chinese)
唐飞龙,叶施仁,肖春. 基于用户质量的微博社区博主影响力排序算法[J]. 计算机工程与应用,2015,51(4):128-132

[8] Meeyoung C, Hamed H, Fabricio B, et al. Measuring User Influence in Twitter:the Million Follower Fallacy[C]// Procof the 4th International AAAI Conference on Weblogs and Social Media. Menlo Park:AAAI Press,2010:10-17

[9] Brin S, Page L. The Anatomy of a Large Scale Hypertextual Web Search Engine[C]// Proc of the 7th International World Wide Web Conference. Brisbane:ACM Press,1998:107-117

[10] Cao Shan-shan, Wang Chong. Improved PageRank Algorithm Based on Links and User Feedback[J]. Computer Science,2014,41(12):179-182(in Chinese)
曹珊珊,王冲. 基于网页链接与用户反馈的 PageRank 算法改进研究[J]. 计算机科学,2014,41(12):179-182

[11] Chen Xiao-fei, Wang Yi-tong, Feng Xiao-jun. An Improvement of PageRank Algorithm Based on Page Quality[J]. Journal of Computer Research and Development, 2009, 46(Suppl.): 381-387(in Chinese)
陈小飞,王轶彤,冯小军,一种基于网页质量的 PageRank 算法改进[J]. 计算机研究与发展,2009,46(增刊):381-387

[12] Apache Hadoop[OL]. http://hadoop.apache.org

[13] Lammel R. Google's MapReduce Programming Model Revised [J]. Science of Computer Programming,2007,68(3):208-237

[14] Srirama S N,Jakovits P,Vainikko E. Adapting Scientific Computing Problems to Clouds Using MapReduce[J]. Future Generations Computer Systems,2012,28(1):184-192

[15] Chen Gong, Niu Qin-zhou. Research on PageRank Algorithm Based on MapReduce[J]. Microelectronics & Computer,2012,29(5):81-85(in Chinese)
陈宫,牛秦洲. 基于 MapReduce 的 PageRank 算法的研究[J]. 微电子学与计算机,2012,29(5):81-85

[16] Chen Hao,Die Ge. MicroBlog User Ranking Research Based on Hadoop[D]. Shanghai: East China University of Science and Technology,2014(in Chinese)
陈浩,迭戈. 基于 Hadoop 的微博用户影响力排名算法研究[D]. 上海:华东理工大学,2014