

# 基于 PageRank 的热点发现混合算法研究

应毅,黄慧,刘定一

(三江学院 计算机科学与工程学院,江苏 南京 210012)

**摘要:** 社交网络下的热点话题发现技术是当前舆情分析与预测的基础性研究问题。传统的基于聚类、分类的文本分析方法不适用于网络舆情挖掘,经典的 PageRank 算法仅考虑网页间的链接结构,为了更加准确和全面地多角度综合评价舆情热点,文中综合考虑用户社会地位、博文相似度指数和热度指数三个热点发现的重要指标,提出了基于 PageRank 和相似度计算的热点发现混合算法(HDH-PRSC)。其中基于 PageRank 算法与微博用户粉丝间的链接结构图获取用户的社会地位值;结合 TF-IDF 算法与余弦相似性算法计算博文的相似度指数;利用转发数、评论数和点赞数获得博文的热度指数。博文的热度评分由用户社会地位值、博文相似度指数和热度指数三项分值相加获得。依托新浪微博数据的实验表明,HDH-PRSC 算法能够更为合理地发现热点话题。

**关键词:** PageRank;用户社会地位;相似度指数;热度指数

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1673-629X(2019)09-0081-05

doi:10.3969/j.issn.1673-629X.2019.09.016

## Research on Hotspot Detection Hybrid Algorithm Based on PageRank

YING Yi, HUANG Hui, LIU Ding-yi

(School of Computer Science and Technology, Sanjiang University, Nanjing 210012, China)

**Abstract:** Hot topic discovery technology in social networks is a fundamental research issue in current public opinion analysis and prediction. However, the traditional text analysis method based on clustering and classification is not suitable for network public opinion mining, and the classical PageRank algorithm only considers the link structure between web pages. In order to evaluate public opinion hotspots more accurately and comprehensively from different angles, considering user social status, blog similarity index and heat index as three important indicators, we propose a hotspot detection hybrid algorithm based on PageRank and similarity calculation (HDH-PRSC), among which the social status value of users is obtained by the link structure map of micro-blog followers of each certain user, the similarity index of blog text is calculated based on TF-IDF algorithm and cosine similarity algorithm, while the heat index is got by forwarding numbers, comment numbers and point of praise. Finally, the heat score of blog can be obtained by adding the three scores of social status value, similarity index and heat index together. Experiments based on Sina micro-blog data shows that the HDH-PRSC algorithm can find hot topics more reasonably and effectively.

**Key words:** PageRank; user social status; similarity index; heat index

## 0 引言

近年来,随着互联网技术的发展,微博、微信等新兴微媒体受到大众广泛关注,被作为获取时事要闻、发表个人意见的主要途径。由于网络热点事件具有传播速度快、范围广的特征,使得微媒体平台在事件爆发后,能迅速积累大量的热点信息。从微媒体的海量信息中挖掘热点话题,有助于政府机关和相关部门及时

掌握互联网事件的主导权,积极传播正能量,是当前微媒体热点引导工作面临的重要任务。

对于热点问题的发现,除了传统的针对文本信息的聚类、分类等数据挖掘技术外,近年来也广泛采用基于联系的链接分析技术。PageRank<sup>[1-2]</sup> 是进行网页权重计算的链接分析算法,起源于文献引文分析方法<sup>[3]</sup>,主要思想是依次统计每个网页的入度和出度,入度为

收稿日期:2018-10-30

修回日期:2019-03-04

网络出版时间:2019-04-24

**基金项目:** 江苏省高校哲学社会科学研究基金项目(2018SJA0506);江苏高校“青蓝工程”资助(苏教师[2018]12号);江苏省高等学校自然科学研究项目(18KJB520042)

**作者简介:** 应毅(1979-),男,硕士,副教授,研究方向为大数据处理与数据库。

**网络出版地址:** <http://kns.cnki.net/kcms/detail/61.1450.TP.20190424.1051.048.html>

当前网页被其他网页引用的次数,且次数越多,说明该网页的质量越高;出度为当前网页引用其他网页的次数,算法通过迭代的方式反复计算每个网页的 PageRank 值<sup>[4]</sup>,直至达到平稳分布为止。搜索引擎即通过 PageRank 算法分析互联网上网页间的相互引用关系,进而衡量网页的重要程度。

基于这一思想,学者们针对热点话题的挖掘展开了广泛的研究。文献[5]将主题特征和时间因子引入 PageRank 算法进行热点分析,与传统方法相比,获得了更为可靠的分析结果。文献[6]提出网页时间权值的概念对热点问题分析,通过新方法弥补了传统算法中新网页总是处于“弱势”地位的不足,强调最新发布网页的重要性,对网页的排序搜索起到了优化的作用。文献[7]引入时间维和空间维的主题因子,为网络舆情处理提供主题样本。

此外,学者们还从词频热度、事件要素、情感趋势等多个角度结合 PageRank 算法来分析热点问题,都取得了一定的进展<sup>[8-12]</sup>。

## 1 热点发现研究现状及文中工作

微博作为大众获取信息的重要来源,在热点发现和舆情把控方面发挥了重要作用。使用传统的 PageRank 算法对微博数据进行热点挖掘时,会产生以下几个问题:

(1) 微博不同于网页,没有频繁的入度与出度,因此无法直接使用 PageRank 算法获取每篇博文的重要程度。

(2) 传统算法未考虑文本之间的相似度,一般而言,相应时间段内某个主题内容被提出的频次越高,说明该主题内容被关注的大众越多,越易成为热点话题。

(3) 微博通常存储了博文的转发数、评论数和点赞数等数据,这些数据也应被充分利用以发现热点话题。

针对以上问题,文中通过网络爬虫工具对新浪微博的数据进行爬取,并从多角度对数据加以分析,获取更为可靠的热点话题。文中的贡献如下:

(1) 引入微博用户的社会地位,通常用户的粉丝数越多,其社会地位越高,该用户发表的博文影响力也越大,越易被传播。文中将用户的社会地位融入 PageRank 算法,将用户之间拥有的粉丝信息形成链接结构,通过计算每位用户的 PageRank 值来衡量其社会地位,PageRank 值越高,其社会地位也越高。

(2) 将 TF-IDF 算法与余弦相似性算法结合进行博文之间的相似度计算,若时间段内某篇博文的相似文章数量较多,说明该时间段内此话题被多个用户关注,是热点话题。

(3) 充分利用博文的转发数、评论数和点赞数等数据形成热度指数,热度指数越高,说明博文的影响力越大,越易形成热点。

文中在 PageRank 算法和相似度计算技术的基础上进行改进,提出热点发现混合算法(hotspot detection hybrid algorithm based on PageRank and similarity computation, HDH-PRSC 算法)。

## 2 HDH-PRSC 算法

HDH-PRSC 算法思想是通过计算每篇博文的评价获取热点信息,其中博文评分由三部分组成:微博用户的社会地位、文本相似度分值和热度指数。热点分析时,将算法产生的三项值做加法运算,得到的总评分越高,则博文成为热点的可能性越大。

### 2.1 微博用户的社会地位

基于 PageRank 算法的基本思想计算微博用户的社会地位。

定义 1(合法用户):  $U = \{U_1, U_2, \dots, U_n\}$  为所有微博用户集合。如果满足  $U_i \in U$  且  $1 \leq i \leq n$ , 则称  $U_i$  为集合中的合法用户,简称用户。

定义 2(合法粉丝):  $V = \{V_{i1}, V_{i2}, \dots, V_{im}\}$  为用户  $U_i$  的粉丝集合。如果满足  $V_{ij} \in V$  且  $1 \leq j \leq m$ , 则用户  $U_j$  是用户  $U_i$  的合法粉丝,简称粉丝,记作  $V_{ij}$ 。

定义 3(链接结构图): 每个用户形成图中的一个节点,如果用户  $U_j$  是用户  $U_i$  的粉丝,则形成  $U_j$  节点至  $U_i$  节点的一条有向边,指向  $U_i$  的边数称为  $U_i$  的入度。同理,  $U_i$  也可作为其他用户的粉丝,  $U_i$  节点指向其他用户的边数称为  $U_i$  的出度。由节点和有向边形成的图称为链接结构图。

基于 PageRank 算法的基本原理,微博用户的社会地位计算方法如式 1 所示。

$$\text{PR}_{\text{status}}(U_i) = \frac{1-d}{U_{\text{total}}} + d \sum_{j=1}^m \frac{\text{PR}_{\text{status}}(V_{ij})}{N(V_{ij})} \quad (1)$$

其中,  $U_i$  为某个待评价的用户;  $j$  为用户  $i$  的粉丝数;  $U$  为用户总数;  $N(V_{ij})$  为用户  $U_i$  的第  $j$  个粉丝节点的出度;  $\text{PR}_{\text{status}}(U_i)$  为用户  $U_i$  的社会地位值;  $\text{PR}_{\text{status}}(V_{ij})$  为用户  $U_i$  的粉丝的社会地位值;  $d$  为用户之间彼此链接的阻尼系数,依据经验,通常取值为 0.85。

按照式 1 对用户的社会地位值进行迭代运算,由于该算法是收敛的,直到用户的社会地位值趋于平稳,则计算结束。

通过算法获取每位用户的社会地位 PageRank 值并对 PageRank 值从高至低排序,同时计算所有用户的 PageRank 值的平均值,记为  $\text{AVG}(\text{PR})$ 。为提高算法运行效率,依据  $\text{AVG}(\text{PR})$ ,将用户的社会地位划分为三类:

$$\begin{cases} \text{积极户用, 记作 ActiveUser, 当 } PR_{status}(U_i) \geq AVG(PR) + e \text{ 时} \\ \text{普通用户, 记作 NormalUser, 当 } AVG(PR) - e \leq PR_{status}(U_i) \leq AVG(PR) + e \text{ 时} \\ \text{消极用户, 记作 InactiveUser, 当 } PR_{status}(U_i) \leq AVG(PR) - e \text{ 时} \end{cases}$$

$e$  为用户定义的阈值, 通常情况下, 消极用户的发文成为热点的概率较低, 因此计算热度时, 不考虑消极用户的发文, 以此提高算法效率。

算法 1: 计算微博用户社会地位的算法为 GetUserStatus, 其伪代码如下:

```
输入: 链接结构图,  $e$  值;  
输出:  $PR_{status}(U_i)$  ( $U_i \in \text{ActiveUser} \ \&\& \ U_i \in \text{NormalUser}$ )。  
FOREACH  $U_i \in U$   
BEGIN  
  Array[  $i, 0$  ] = GetPRstatus(  $U_i$  ) // 获取每个用户的社会地位的 PageRank 值  
  Array[  $i, 1$  ] = UserID(  $U_i$  ) // 获取用户的 UserID 赋值给数组  
  Array[  $i, 2$  ] = NULL // 用户初始的状态为空  
END  
SORT( Array ) // 对用户的社会地位从高到低排序  
PRAvgStatus = AVG( PR ) // 计算所有用户的社会地位均值  
赋值变量 PRAvgStatus  
FOREACH  $i$  in Array  
BEGIN  
  IF( Array[  $i, 0$  ] >= PRAvgStatus +  $e$  )  
    Array[  $i, 2$  ] = ActiveUser  
  ELSE IF( Array[  $i, 0$  ] >= PRAvgStatus -  $e$  && Array[  $i, 0$  ] <= PRAvgStatus +  $e$  )  
    Array[  $i, 2$  ] = NormalUser  
  ELSE  
    Array[  $i, 2$  ] = InactiveUser  
END  
DELETE( InactiveUser )
```

## 2.2 博文相似度指数

当某篇博文具有较多的与其相似度较高的其他博文时, 说明该博文可能是某时间段内的热点话题, 被较多的用户关注。因此, 统计博文的相似度指数, 也是热点发现研究中的重要指标。文中通过结合 TF-IDF<sup>[13]</sup>方法和余弦相似性算法<sup>[14]</sup>获得博文的相似度指数。

定义 4(TF-IDF): TF-IDF 是用于评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。其中 TF(term frequency) 表示词条在文章中出现的频率; IDF(inverse document frequency) 表示包含某个词的文档越少, 则这个词的区分度就越大, 即 IDF 越大。

TF=某个词在文章中出现的次数/文章总词数

(2)

IDF=log[ 词料库的文档总数/( 包含该词的文档数+1) ]

(3)

TF-IDF=TF×IDF

(4)

通过统计 TF-IDF 的值, 将每篇博文取值较高的前四个词作为该篇博文的关键字。

定义 5(余弦相似度): 余弦相似度利用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。文中将两篇博文的关键字分别作为空间向量, 计算相似度, 公式如下:

$$Sim(X, Y) = \frac{\sum_{i=1}^n (X_i \times Y_i)}{\sqrt{\sum_{i=1}^n X_i^2} \times \sqrt{\sum_{i=1}^n Y_i^2}}$$

(5)

其中,  $X_i$  为  $X$  文档中第  $i$  个关键词出现的次数;  $Y_i$  为  $Y$  文档中第  $i$  个关键词出现的次数。通过式 5 计算向量的余弦值, 余弦值越接近 1, 表明夹角越接近 0, 则两篇文档越相似。

统计博文相似度指数的过程分为三个步骤:

步骤一: 利用 TF-IDF 方法获取每篇博文的关键词, 并对关键词出现的次数进行排序, 获取 Top-K 个关键词;

步骤二: 依次遍历博文, 删除不包含关键词的博文;

步骤三: 将剩余的博文对应的关键词按照余弦相似性的方法比较相似度, 同时利用 DBSCAN 算法<sup>[15]</sup>对博文进行聚类, 统计相似博文的篇数。博文的相似度指数为每篇博文的相似篇数与总篇数的比值。

其中 DBSCAN 是一种基于密度的聚类算法, 基本思想是在样本区域内任意选择一个核心对象, 以核心对象为中心计算与该核心对象的距离小于  $\varepsilon$  的其他所有对象, 并将这些对象组合成一个类簇, 将剩余对象作为输入进入下一轮算法的迭代, 直到样本空间中的所有对象都划分为某个类簇, 算法结束。文中的核心对象为算法输入中的任意一篇博文,  $\varepsilon$  为  $Sim(X, Y)$  的取值, 即博文相似度。

算法 2: 统计博文相似度指数的算法为 GetSimScore, 其伪代码如下:

```
输入: 活跃用户与一般用户的博文文章,  $k$  值, 相似度阈值  $m$ ;  
输出: 每篇博文的相似度指数。  
FOREACH  $T_i \in \text{Title}$   
BEGIN  
  Array[  $i, 0$  ] = GetKeyWords() // 通过 TF-IDF 方法获取文章关键字  
  Array[  $i, 1$  ] = TitleID(  $T_i$  ) // 获取文章 ID 号  
END
```

```
FOREACH  $j$  in KEYWORDS
  ArrKeyWords[  $j$  ] = COUNT( Array ) //统计每个关键字出现的次数
  string[ ] ArrImpKeyWords = SORT( ArrKeyWords,  $k$  ) //获取前  $k$  个重要关键字
  FOREACH  $T_i \in \text{Title}$ 
  BEGIN
    IF !  $T_i$ . Contains( ArrImpKeyWords ) //判断博文  $T_i$  是否包含重要关键字
    DELETE(  $T_i$  ) //如果不包含重要关键字,则删除博文
  END
  FOREACH  $T_i \in \text{Title}$ 
  FOREACH  $T_j \in \text{Title}$ 
  BEGIN
    IF Sim(  $T_i$  ,  $T_j$  ) >=  $m$ 
    DBSCAN() //如果两篇博文相似度达到阈值,则调用 DBSCAN 算法进行聚类
  END
  FOREACH  $T_i \in \text{Title}$ 
  SimScore(  $T_i$  ) //获取每篇博文的相似度指数
```

2.3 热度指数

定义 6(热度指数):热度指数由转发数、评论数和点赞数的总和得到,热度指数越高,说明此博文成为热点的概率越大。

热度指数的计算公式如下:

$\text{HotScore}(T_i) = (\text{转发数} + \text{评论数} + \text{点赞数}) / (\text{总转发数} + \text{总评论数} + \text{总点赞数})$  (6)

2.4 HDH-PRSC 算法的实现

博文的最终热度综合评分由微博用户的社会地位、博文相似度指数和热度指数三项分值获得,如下:

$$\text{Score}(T_i) = \alpha \text{UserScore}(U_i) + \beta \text{SimScore}(T_i) + \gamma \text{HotScore}(T_i)$$
 (7)

其中,  $\alpha$ 、 $\beta$  和  $\gamma$  分别为微博用户的社会地位、博文相似度指数和热度指数的评分系数,且  $\alpha + \beta + \gamma = 1$ 。

算法 3: HDH-PRSC 算法。

输入:  $\alpha$  值,  $\beta$  值,  $\gamma$  值;  
输出: 博文综合评分。

```
FOREACH  $T_i \in \text{Title}$ 
BEGIN
   $U_i = \text{GetUserByTitle}(T_i)$  //通过  $T_i$  获取博文的用户  $U_i$ 
   $\text{UserScore} = \text{GetPR}_{\text{status}}(U_i)$ 
   $\text{SimScore} = \text{SimScore}(T_i)$ 
   $\text{HotScore} = \text{HotScore}(T_i)$ 
  RETURN  $\alpha * \text{UserScore} + \beta * \text{SimScore} + \gamma * \text{HotScore}$ 
END
```

3 实验结果及分析

采用网络爬虫对新浪微博数据进行采集,数据集包括 2018 年 8 月 25 日-2018 年 8 月 30 日期间 4 000 个微博用户共 864 000 条博文。删除明星等名人的博文信息,随机抽取部分数据作为实验基础,数据字段包含用户 UserID、性别、用户粉丝信息、微博数、微博 ID、微博内容、微博发布时间、阅读数、转发数、评论数和点赞数。

实验操作时,用来给用户分类的阈值  $e$  设置为  $\text{AVG}(\text{PR}) \times 0.2$ ; 计算博文相似度指数时,相似度阈值设置为 0.6; 式 7 中的  $\alpha$ 、 $\beta$  和  $\gamma$  均设置为 1/3。由于实验数据规模较大,最终计算得到的用户地位评分、相似度指数和热度指数值均介于 0.000 1 ~ 0.000 4 之间。实验搜索到的热度文章以及各项算法的评分结果如表 1 所示。

表 1 热度博文及评分

ID	博文标题	用户地位评分	相似度指数	热度指数	综合评分
E <sub>1</sub>	三问滴滴,以生命的名义!	0.000 27	0.000 13	0.000 17	0.000 19
E <sub>2</sub>	个税法修改	0.000 31	0.000 03	0.000 07	0.000 14
E <sub>3</sub>	刑法专家谈“砍人反被杀”:构成特殊防卫 担心反扑反击合理	0.000 25	0.000 08	0.000 06	0.000 13
E <sub>4</sub>	哈尔滨一温泉酒店发生火灾已致 19 人死亡	0.000 29	0.000 02	0.000 01	0.000 11
E <sub>5</sub>	24 岁员工全身 98% 烧伤身亡	0.000 14	0.000 02	0.000 15	0.000 10
E <sub>6</sub>	苏炳添打破赛会纪录夺冠! 为中国速度点赞!	0.000 21	0.000 02	0.000 05	0.000 09
E <sub>7</sub>	路中开锅煮面致塞车 涉事男子被公安机关行政拘留	0.000 19	0.000 02	0.000 04	0.000 08
E <sub>8</sub>	致所有关注及热爱电子体育运动的你们	0.000 04	0.000 03	0.000 04	0.000 03

通过实验发现,单一算法(如:基于 PageRank 的用户社会地位计算、结合 TF-IDF 和余弦相似性的博文相似度计算)得到的结果和 HDH-PRSC 算法得到的

热点博文排序不尽相同,E<sub>2</sub>事件在 PageRank 算法中排序第 1,但在 HDH-PRSC 算法中排序第 2;E<sub>3</sub>事件在相似度计算中排序第 2,但在 HDH-PRSC 算法中排序第



3。如图 1 所示,单一算法所得结果相对比较片面,如 PageRank 算法将  $E_2$  事件排名第 1 是因为该博主的发布者是央视财经,算法认为只要是央视财经发布的博文一定都是热点,而实际情况并非如此;同理, $E_3$  事件在相似度计算中排名第 2 是因为其类似文章的出现次数较多,这有可能是因为一个小团体内多次转发了某

篇广告,仅凭此项因素判断热点缺乏充足依据。HDH-PRSC 算法综合了发文用户的社会地位、博文相似度指数、热度指数三个因素,更加全面地从多个角度综合评价博文的热度,思想更为合理,其计算结果也要优于单一算法。

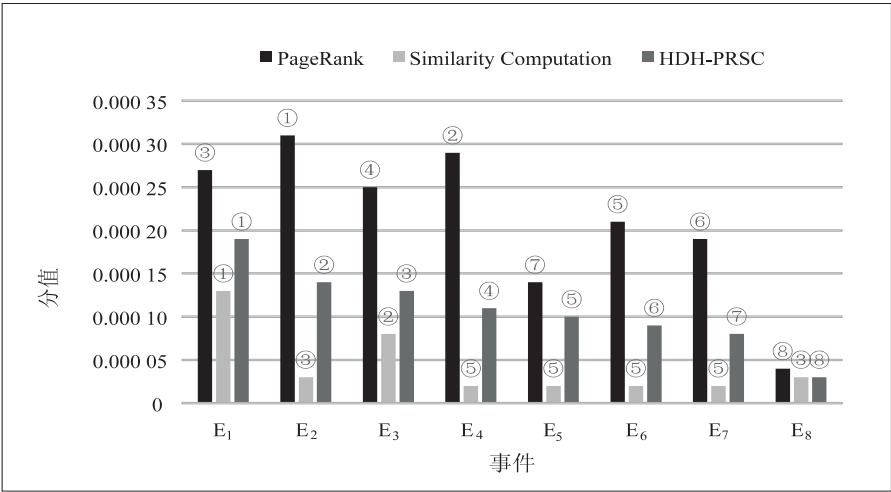


图 1 热点发现算法比较

4 结束语

基于用户社会地位、博文相似度指数以及热度指数对微博数据进行分析,提出一种基于 PageRank 和相似度计算的热点发现混合算法。该算法弥补了传统算法中未考虑用户社会地位和文本相似度的不足。实验结果表明,HDH-PRSC 算法能够更为合理地发现热点话题。下一步将在更大规模的数据集上进行实验分析,并研究如何通过算法获取文中  $\alpha$ 、 $\beta$  和  $\gamma$  评分系数的合理取值,以得到更为理想的实验效果。

参考文献:

[1] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web[M]. Palo Alto, California: Stanford University Press, 1998.

[2] ARASU A, NOVAK J, TOMKINS A, et al. PageRank computation and the structure of the web: experiments and algorithms[C]//Proceedings of the eleventh international world wide web conference. Honolulu, Hawaii: Poster Track, 2002: 107-117.

[3] 耿瑞, 李石君, 尹为民. 基于主题相关性和时间因素的改进 PageRank 算法[J]. 微电子学与计算机, 2015, 32(8): 158-162.

[4] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1): 107-117.

[5] 段淮川, 胡平. 基于主题特征和时间因子的改进 PageRank 算法[J]. 计算机工程与设计, 2010, 31(4): 866-868.

[6] 冯海涛. 基于网页时间权值的 PageRank 算法改进[J]. 西安邮电学院学报, 2013, 18(2): 121-124.

[7] 黄炜, 金雅博, 胡昌龙. 网络舆情主题信息采集研究[J]. 现代图书情报技术, 2012, 33(11): 65-71.

[8] 舒琰, 向阳, 张骐, 等. 基于 PageRank 的微博排名 MapReduce 算法研究[J]. 计算机技术与发展, 2013, 23(2): 73-76.

[9] 李慧, 王丽婷. 基于词项热度的微博热点话题发现研究[J]. 情报科学, 2018, 36(4): 45-50.

[10] 李纲, 徐伟, 王馨平. 基于事件要素的组合模型微博热点事件摘要提取[J]. 图书情报工作, 2018, 62(1): 96-105.

[11] 何跃, 朱灿, 朱婷婷, 等. 微博热点话题情感趋势研究[J]. 情报理论与实践, 2018, 41(7): 155-160.

[12] LING Zhang, ZHENG Qin. The improved PAGERANK in web crawler[C]//1st international conference on information science and engineering. Nanjing, China: IEEE, 2009: 1889-1892.

[13] 任姚鹏, 陈立潮, 张英俊, 等. 结合语义的特征权重计算方法研究[J]. 计算机工程与设计, 2010, 31(10): 2381-2383.

[14] 王冲, 纪仙慧. 基于用户兴趣与主题相关的 PageRank 算法改进研究[J]. 计算机科学, 2016, 43(3): 275-278.

[15] 伏家云, 靖常峰, 杜明义. 空间密度聚类模式挖掘方法 DBSCAN 研究回顾与进展[J]. 测绘科学, 2018, 43(12): 50-57.