目录

[基于PageRank和用户动态分析的知乎大V影响力评估 2](#_Toc459316155)

[1. 绪论 2](#_Toc459316156)

[1.1. 研究背景 2](#_Toc459316157)

[1.2. 研究现状分析 3](#_Toc459316158)

[2. 知乎大V影响力评估 5](#_Toc459316159)

[2.1. PageRank算法[3] 5](#_Toc459316161)

[2.2. 基于用户信息的PageRank改进算法 6](#_Toc459316162)

[2.3. 基于用户信息、问题动态的后验InfoActRank影响力算法 6](#_Toc459316163)

[3. 实验 6](#_Toc459316164)

[3.1. 数据获取 6](#_Toc459316168)

[3.2. 数据整理 10](#_Toc459316169)

[4. 最优子集选取 11](#_Toc459316170)

基于PageRank和用户动态分析的知乎大V影响力评估

# 绪论

## 研究背景

随着互联网的迅猛发展，我国网名数量激增，CNNIC2016年第38次互联网报告统计：“截至 2016年 6月，我国网民规模达 7.10 亿，上半年新增网民 2132万人，增长率为 3.1%。我国互联网普及率达到 51.7%， 与 2015 年底相比提高 1.3 个百分点。”

社交网络服务（Social Network Service, 简称SNS）作为互联网应用发展的必备要素，不再局限于信息传递，而是与沟通交流、商务交易类应用融合，借助其他应用的用户基础，形成更强大的关系链，从而实现对信息的广泛、快速传播。如Twitter、Facebook、微博、微信在过去几年人们的人际交往中发挥着重要的作用。

影响力评估是社交网络分析中的重要内容之一，社交网络用户影响力可以通过其与其他用户之间的社交活动以及其自身关键信息体现出来。自社交影响力在社会生活和思考决策等方面发挥重要作用之后，影响力评估便在多个邻域得到了广泛的关注，比如社交网络信息传播、兴趣模式挖掘、用户或信息推荐、病毒式营销、领袖意见发现、突发事件检测和广告投放[1]等。对于特定社交网络的分析亦可以帮助其网站发现存在的问题。

“知乎”是社区氛围友好与理性，连接各行各业的精英的一个真实的网络问答社区。用户分享着彼此的专业知识、经验和见解，为中文互联网源源不断地提供高质量的信息。可以用知乎用户谌斌的一句话形象的描述知乎的社交网络氛围：“谈笑有鸿儒，往来无白丁，无微博之乱耳，无SNS之劳形。”通过对知乎用户的影响力评估，不仅可以支撑以上受关注领域资源基础；而且可以预测甚至控制某些问题和价值观的走向，并且通过社交网络中的重要人物（大V）之间的联系，确定这个社交网络的基本结构。因此，在此社交环境下，用户影响力评估有着重要的意义和价值。

## 研究现状分析

国内对于知乎社区的影响力分析评估这方面的研究很少几乎为零，因此以下大多从微博用户影响力分析或其他社区网络影响力分析说明现阶段的研究现状。

常用的影响力度量方法大致可以划分为基于网络拓扑结构、基于用户行为和基于用户交互信息的度量等类型。基于网络拓扑结构的度量，顾名思义就是通过用户与其关注者之间的拓扑网络结构关系来体现用户的影响力，且网络拓扑结构的数据获取相对比较简单，度量方法相对成熟，由于需要迭代运算所以计算量不一而论。 但是，网络拓扑结构无法呈现用户行为和用户与其他用户之间的交互活动。因此基于用户行为和基于用户交互信息的度量方法可以弥补这一缺点，将影响力的变化通过用户动态信息精准化、细节化。

Freeman提出的节点出入度度量法[2]，通过节点的出度和入度来衡量社交网络中节点用户与影响力有关的指标（如关注数、推荐数、跟帖数等）。这种算法在一定程度上可以通过用户及其邻居之间的联系反映用户的影响力，但也是因为这个原因，这只能是对用户局部影响力的度量，无法衡量用户在整个社交网络中的影响力。

PageRank[3]算法是由Google的创始人发明的，现在基于这一思想的各种变体已被所有大型搜索引擎采用。该算法为每个网页都赋予了一个只是网页重要程度的评价值。网页的重要性是依据指向该网页的所有其他网页的重要性，以及这些网页中所包含的链接数求得的。

Haveliwala, T.H. 等人对PageRank算法做了进一步的改进，在使用算法进行网页排序时不仅要考虑网页之间的链接，还要考虑到正文上下文和主题等因素[4]。

Xiaodan Song等人提出InfluenceRank算法[5]，不仅仅考虑到自身博客对其他博客的影响，而且考虑了博客自身的重要度和新颖性，并将此算法用于博客社交模型中，来甄别确定领袖意见。

王彪在PageRank的核心思想的基础上提出了PeopleRank算法[6]，即被越多影响力大的人所关注，则他的影响力大的概率就越大。公式如下：



其中，表示用户的影响力，指所有关注了用户的用户集合。后根据实际情况对公式做出修改：



其中，新增的为阻尼系数，为用户的关注数目。

齐超等人提出通过提取用户转发、评论和提及三种行为建立传播影响力评估模型[7]，运用统计分析的方法就用户不同行为对影响力传播做出的贡献进行度量。



其中，转发、评论、提及三种行为分别由 表示对用户传播影响力的贡献度， 则表示用户对用户的贡献权值，分别表示转发人次、评论人次和提及人次。值得一提的是，就他们的研究发现，在微博网络社区环境中最能反映用户传播影响力的用户特征是用户转发情况。

张玥提出用户的影响力来自于他的文章的广度和深度[8]，广度可以通过回复人数来度量，深度可以通过回复的频率和回复的质量来度量。结合广度、深度和PageRank算法对用户影响力做出评估。将主题内用户间的交互信息抽象为有向图，V是用户集，E是由回复关系构成的边集，。定义函数如下：



其中表示用户在主题中的深度影响力，表示时间中回复 的回复长度，表示在时间内用户获得的广度和深度影响力总和，则为多主题的影响力之和。值得一提的是，张玥同学而后又对影响力细化，分成长期影响、中长期影响、短期影响并对其分别论述。

。。。。。。

。。。。。。

。。。。。。

。。。。。。

以上算法在一定程度上都可以对用户的影响力做出相对有效的评估，但是有些算法使用指标过于简单化，或只用用户的粉丝数和关注数，或分析用户行为，或将二者相结合。但是，都忽略了用户的文章或者行为在未来几天或十几天里所带来的实际上的影响。因此，为了提高对用户影响力评估的准确性，本文将通过对知乎网络问答社区大V的自身行为、关注者和被关注者之间的联系以及大V的行为为文章或问题在未来一段时间内所带来的实际影响相结合，提出了一种基于PageRank、用户行为和真实影响用户影响力评估算法。

# 知乎大V影响力评估



## PageRank算法[3]

PageRank的初衷指的是计算某个人在任意次点击链接之后到达某一网页的可能性。如果一个网页拥有来自其他热门网页的指向链接越多，那么人们访问到这个网页的可能性也就越大，即如果网页BCD都指向网页A，那么网页A被随机访问的可能性会很高；如果网页A指向网页B，且网页A被随机访问的可能性很高，那么网页B被随机访问的可能性也将很高；如果网页A同时指向网页B、C、D，那么网页A的可能性将会平分给网页B、C、D。

如果，用户始终不停地点击，那么每一个网页终将都会被访问到，但是并不符合实际，大多数人在浏览一段时间后都会停止点击。因此，为了切合这一实际，PageRank还使用了一个值为0.85的阻尼因子，用来表示用户点击每个网页中的链接的概率为85%。

通过不断的迭代，直到网页的PageRank值到达一种稳定的收敛的状态。我们就可以将每个用户看成是网页，用户关注了谁即类似当前网页所指向的链接。以下以几名用户模拟PageRank算法。

 用户BCD都关注了A，且他们的PageRank值都已经计算得出，记为Pr(B)。其中用户B还关注了其他两位用户，用户C还关注了其他三位用户，用户D只关注了A。**（公式）**

## 基于用户信息的PageRank改进算法

## 基于用户信息、问题动态的后验InfoActRank影响力算法

# 实验



## 数据获取

本文选取知乎作为数据来源，由于知乎并没有提供API接口，所以利用Python爬虫爬取知乎数据，数据库使用非关系型数据库MongoDB。爬取的数据分为两个阶段、三个部分，前期:知乎用户基本信息（用户唯一标识id，用户名，关注数，回答数等基本信息），用户间的关系；后期：大V动态数据（问题唯一标识qid，回答唯一标识aid，动态时间等），问题或回答动态数据（问题浏览数，回答数，回答赞同数等）。

硬件介绍：

。。。。。。。

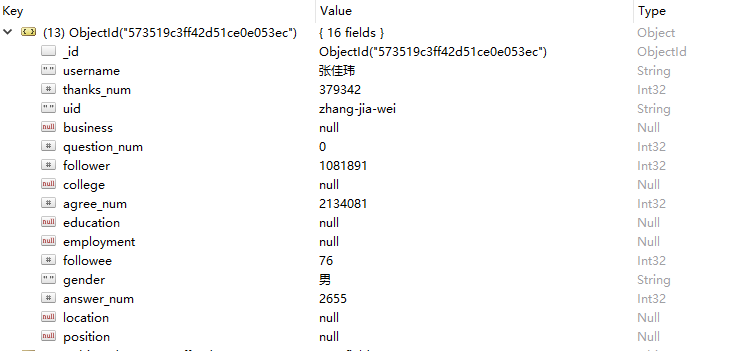
1. **知乎用户基本信息**

本次共爬取13122位知乎用户的基本信息，虽然只是知乎全部用户的一小部分，但用于研究分析已经足够。爬取方式是以一位公认知乎大V张佳玮为种子，爬取他所关注的人（followee）的信息存入数据库，再循环爬取数据库中没有被爬取过的用户所关注的人。为什么不从关注他的人（即追随者，follower）入手施行爬取，主要从以下几个方面考虑：首先，关注张佳玮的用户高达100多万，爬取他一个人的追随者就已经非常困难；其次，因为爬取的都是你所关注的人，所有在一定程度上剔除掉了僵尸用户；再者，爬取你所关注的人大多都是高质量的用户，所以数据更有价值。

用户信息的数据表结构详见下表。

|  |  |  |
| --- | --- | --- |
| 表3-1 知乎用户信息表结构 | | |
| 主键 | 数据类型 | 名称 |
| \_id | ObjectId | MongoDB默认主键 |
| uid | String | 用户唯一标识ID |
| username | String | 用户名 |
| gender | String | 性别 |
| follower | Int32 | 追随者数 |
| followee | Int32 | 用户关注的人数 |
| question\_num | Int32 | 提出问题数 |
| answer\_num | Int32 | 回答问题数 |
| agree\_num | Int32 | 获得赞同数 |
| thanks\_num | Int32 | 获得感谢数 |
| business | String | 行业 |
| education | String | 学校 |
| college | String | 所学专业 |
| employment | String | 工作单位 |
| position | String | 工作所在部门 |
| location | String | 居住地 |

以下数据库截图3-1展示的是知乎用户张佳玮的基本信息，null代表用户没有填写此方面信息。



在13万多用户中，预拟定追随者数大于5000人、回答问题数大于20的1004位知乎用户为知乎大V。

用户关系数据表详见下表。

|  |  |  |
| --- | --- | --- |
| 表3-2 知乎用户关系表结构 | | |
| 主键 | 数据类型 | 名称 |
| \_id | ObjectId | MongoDB默认主键 |
| uid | String | 用户唯一标识ID |
| relationship | List | 关注者列表 |

其中需要说明的是关系字段，类型是List类型，保存用户的关注的所有人的唯一身份标识ID。

。。。。。。

1. **知乎大V动态数据信息**

在2016年7月21日到2016年8月11日期间，对1004位知乎大V动态爬取，获取他们每天更新的动态时间和动态问题或回答的唯一标识ID、动态更新时间。由于，每天要爬取的数据量较大，因此每天只爬取用户首页动态，因此可能会有少量用户动态被遗漏，但这并不影响我们分析数据。

知乎大V动态信息数据表结构详见下表。

|  |  |  |
| --- | --- | --- |
| 表3-2 知乎大V动态信息表结构 | | |
| 主键 | 数据类型 | 名称 |
| \_id | ObjectId | MongoDB默认主键 |
| uid | String | 用户唯一标识ID |
| qid | String | 问题唯一标识ID |
| aid | String | 回答唯一标识ID |
| time | String | 字符串表示的时间戳 |
| qora | String | 问题或回答标志符 |

其中需要说明的是qora字段，作为问题或回答的标识符，q代表问题，a代表回答。当字段值为q时，qid和aid值相同，都是问题唯一标识，通过访问URL：https://www.zhihu.com/question/qid即可获取该问题的详细；当字段值为a时，qid为问题唯一标识ID，aid为该问题下的一个回答唯一标识ID，通过访问URL：https://www.zhihu.com/question/qid/answer/aid即可获取该回答的相关信息。

。。。。。

1. **问题、回答动态数据信息**

在2016年7月21日到2016年8月11日期间，对动态信息表中的的问题或回答进行动态爬取，获取每个问题或回答随着时间的变化，其相关信息的变化情况。

|  |  |  |
| --- | --- | --- |
| 表3-3 知乎问题信息表结构 | | |
| 主键 | 数据类型 | 名称 |
| \_id | ObjectId | MongoDB默认主键 |
| qid | String | 问题唯一标识ID |
| qname | String | 问题名称 |
| qattention | Int32 | 问题关注数 |
| qread | Int32 | 问题被浏览数 |
| qanswer\_num | Int33 | 问题回答数 |
| time | Date | 即时爬取时间 |

|  |  |  |
| --- | --- | --- |
| 表3-4 知乎回答信息表结构 | | |
| 主键 | 数据类型 | 名称 |
| \_id | ObjectId | MongoDB默认主键 |
| uid | String | 用户唯一标识ID |
| qid | String | 问题唯一标识ID |
| aid | String | 回答唯一标识ID |
| asupport | Int32 | 回答赞同数 |
| atime | Date | 即时爬取时间 |

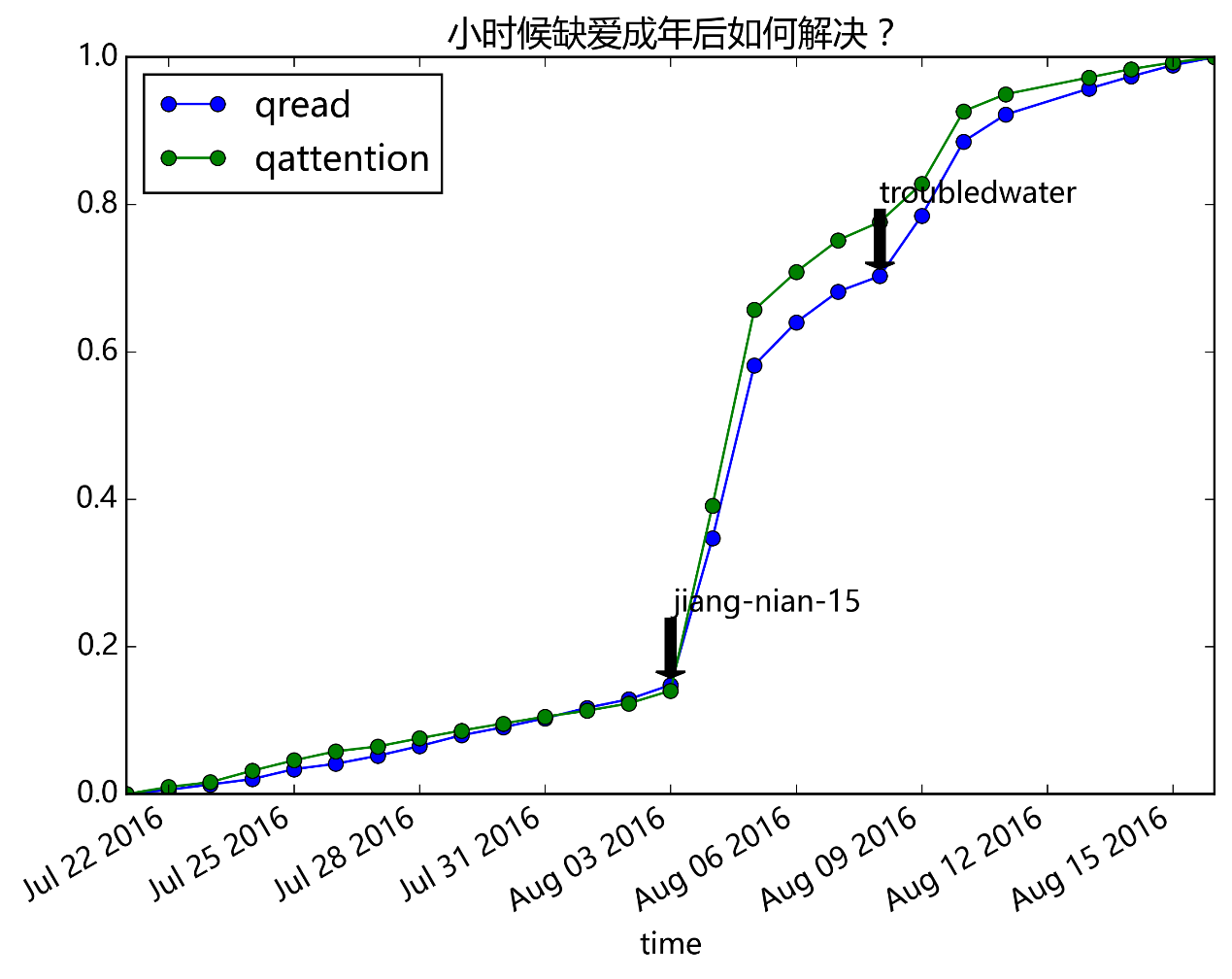
。。。。。。。。。。

## 数据整理

1. **完善用户网络拓扑结构**

因为在前期爬取用户关系时，只是单方面的单向爬取，即只将用户所关注的人存入用户关系结构数据库中。所以需要完善网络结构，将用户的追随者添加到用户关系数据库中，与此同时对之前爬取过程中由于网络问题产生的少许用户关系缺失的情况进行数据修补，针对性的爬取缺失数据。

1. **生成问题趋势变化图**



取某一问题，对阅读数和关注数做归一化处理，然后根据时间序列（单位为1天）画出动态变化图，再从用户动态数据库取出关于这个问题的所有相关动态，主要包括用户ID和动态时间，然后将用户添加到这个时间线上，得到如上图所示的图表。

从上幅图表可以很容易的看出大V所带来的影响，在2016年8月3日的时候ID为jiang-nian-15的用户关注了改问题之后，阅读数和关注数都显著提升，在一段时间后影响力会明显减弱。

# 最优子集选取

1. 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
2. Freeman L C. Centrality in social networks conceptual clarification. Social Networks, 1979, 1(3):215-239
3. Toby Segaram. Programming Collective Intelligence.
4. T.H Haveliwala. Topic-sensitive PageRank: a context-sensitive ranking algorithm for Web search. IEEE Transactions on Knowledge and Data Engineering, 2003, 15(4):784-796.
5. Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng. Identifying opinion leaders in the blogosphere. CIKM ’07, 2007, pages 971-974
6. 王彪. 社交网络中的用户影响力分析[D]. 哈尔滨工业大学，2012.
7. 齐超，陈鸿昶，于洪涛. 基于用户行为中和分析的微博用户影响力评价方法[J]. 计算机应用研究, 2014, 31(7):2004-2007.
8. 张玥. 社会网络中用户影响力分析技术研究[D]. 哈尔滨工业大学，2015.