

**课 程 论 文**

题目（中文）： 知乎网络中核心用户的发现

（英文）： The discovery of important user in ZhiHu Social Net

学 院 信息与机电工程学院

年级专业 2015级计算机科学与技术

学生姓名

学 号

课程名称 复杂网络与社会计算概论

**完 成 日 期 2016 年 12 月**

## 摘要：

进入Web2.0时代之后，各大网站纷纷去中心化，淘汰了以往以网站管理员为核心发布内容的体系，转为以用户为中心强调用户与网站之间的交互，将用户作为网站内容的发布者与维护者。同时这也为社交网站的兴起作了完美的铺垫，国内外社交网站纷纷兴起，例如国外的Facebook, Twitter亦或是国内新浪微博、豆瓣等。

随着社交网站用户量的渐渐增加，如何发现网站的核心用户是网站必须解决的问题。只有找到了网站所面向的核心用户，才能更好的针对核心用户进行网站内容架构的升级，以及着手研究如何吸引更多的新用户。

本文采用数据挖掘这一方法，以知乎这一新兴的社交网络平台为例，在爬取的大量用户数据的基础上，进行数据可视化分析，以得到核心用户所具有的特征，并将各特征总结为用户重要性的预测公式。

**关键词：**知乎；数据挖掘；核心用户发现；社交网站；

目录

**1 绪论**1

1.1 研究背景1

1.2 研究对象1

1.3 研究思路2

1.4 研究意义2

**2 核心用户阐释**2

**3 知乎用户维度与社交属性的分析**3

3.1维度介绍3

3.2 数据来源3

3.3 用户维度3

3.3.1 地域3

3.3.2 行业5

3.3.3 学历6

3.4 社交属性7

3.4.1 回答7

3.4.2 赞同8

3.4.3 关注9

**4 核心用户的发现与预测**11

**5 结论**12

**参考文献**13

1. 绪论
   1. **研究背景**

数据挖掘（Data mining），又译为资料探勘、数据采矿。它是数据库知识发现（Knowledge-Discovery in Databases，简称：KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

进入Web2.0的互联网时代后，用户在互联网上的行为不局限于原来的搜索获取信息，而是扩展为更广阔的生活娱乐社交等方面。网页网站的设计也不再同于Web1.0模式中的单向交流，更注重于网站和用户的交互，以用户为核心，使得用户成为网站内容的制造者与传播者而非以往的单一的阅读者。这种核心思想的进步与技术的发展带动了新一代在线社交网络平台的兴起与发展，微博，博客，论坛等社交网络平台成为了人们网络生活的重要组成部分。

中国的社交网络平台基本以新浪微博，QQ空间，豆瓣网，人人网为核心并趋于稳定，但在近两年以知乎为代表的社会化问答社区却在快速兴起。

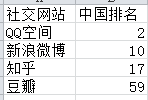


图 1网站排名详情（数据来源ALEX）

* 1. **研究对象**

本文选择社会化问答网站---知乎为主要研究对象，研究知乎网站核心用户的发现。

知乎于2011年1月26日正式上线，截止至2016年5月，知乎具有1300万日活跃用户，50亿月浏览器，33分钟人均访问时长。

知乎的官方定义：知乎是一个真实的网络问答社区，社区氛围友好与理性，连接各行各业的精英。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。

* 1. **研究思路**

1.分析知乎的用户维度与社交属性。

2.利用爬虫爬取知乎用户资料并储存至MySQL数据库

3.导出数据至EXCEL进行数据筛选。

4.利用数据分析软件tableau进行聚类分析

5.从用户维度与社交属性分析核心用户所具有的特征

6.根据特征进行数学分析得到评定用户重要性的模型或者公式

* 1. **研究意义**

在Web2.0时代，网站保持其活力与用户的参与密不可分，其中除去基数较大的普通用户，VIP用户是网站正常发展所不可缺少的。

对于社交网络而言，常常是以某些重要用户为核心，达到传播信息，发布活动，吸引更多用户的目的，所以VIP用户的发现对其是尤为重要的。本研究通过研究知乎网络的各方面维度，以达到发现核心用户的目的，使得网站能更好针对核心用户进行运营，增长用户流量。

1. 核心用户阐释

核心用户是目标用户的一个子集，因此核心用户的特征可以在已设定的目标用户中寻找，而这些特征的选取则是依据当前阶段产品的运营需求所决定的。归根到底，核心用户的定义是为产品运营服务的。准确地定义核心用户，有利于集中有限的资源攻克产品所面临的发展难题。

本文选择了新浪微博，豆瓣，百度知道，知乎四个社交网络平台进行对比，分析其VIP用户的定义。

1.新浪微博：拥有庞大粉丝量的用户，有一定知名度和影响力，是信息传播的重要节点。核心用户多为明星大腕。

2.豆瓣：线下线上活动的积极参与者与发起者，有较好的文艺素养（较多的书评、影评输出）且具有高活跃度，能及时回复他人私信与评论，核心用户多为文艺青年。

3.百度知道：答题数量多或者为特定领域下的行业专家。

由于网站其本身的定位不同所以致使所追求的核心用户也不同，从知乎官方对其的定义: 高质量社交问答平台，通过高质量的问题和答案导向，提升品牌质量，从而引入更多用户，以“精英人才”为主要领路人，带动产品热度。

我们可以推测出知乎核心用户的大致定义：有高质量回答输出，积极参与知乎社区活动（盐CLUB、知乎Live），乐于传播和分享知识，有一定社会影响力的用户。

1. 知乎用户维度与社交属性的分析
   1. **维度介绍**

知乎用户维度：地域，行业，公司，职位，学历

知乎社交属性：提问，回答，文章，赞同，感谢，被关注数，关注数

分析得到了用户维度和社交属性，故可以确定爬取的关键词与所要进行分析的内容。

* 1. **数据来源**

从知乎进行数据分析的用户苏莉安处分享得到了相应的SQL文件，导入MySQL数据库。经查询有127366条记录，收集日期从2013年11月至2016年10月。由于硬件设施的不完善，只能筛选出其中的1万条数据导入EXFCEL进行数据分析与探索。

* 1. **用户维度**
     1. **地域**

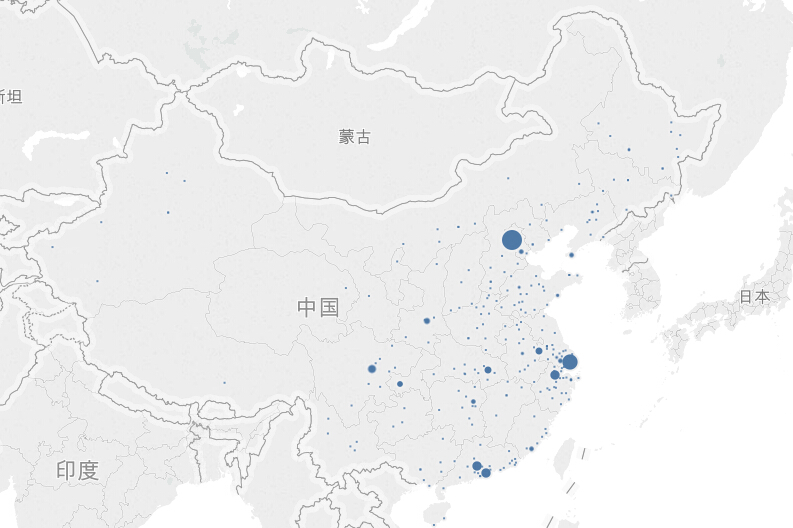
Tableau软件进行数据可视化分析，如图

图 2 关注总数与地域分布

其中点的大小代表该地域关注数目的多少

观察分析可以发现关注大量分布于北上广深等超级城市，散点状分布于沿海城市与长三角地区。由于互联网基础建设以及经济发展水平的差异，致使分布状况差别极大。

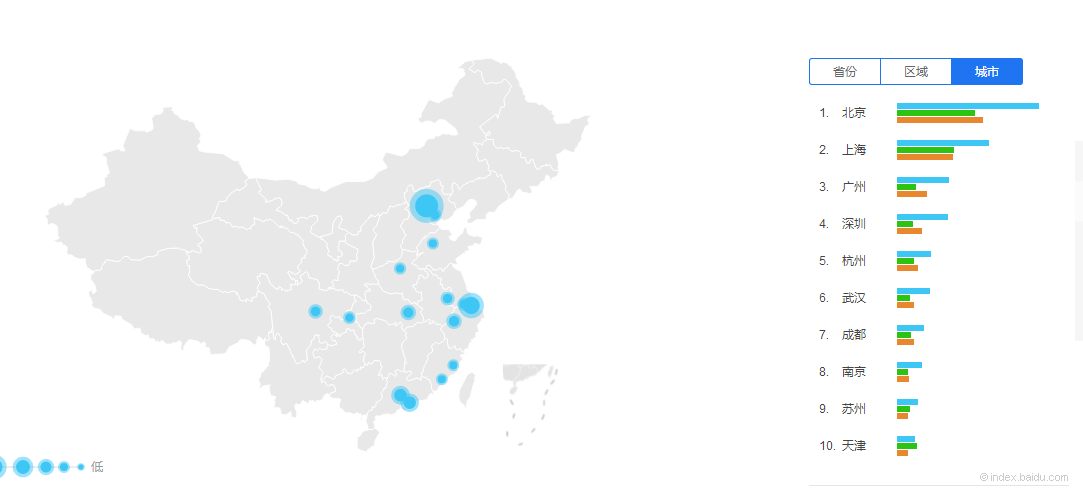


图 3 知乎用户分布情况（数据来源于百度指数）

由图3可知知乎用户数量的分布与关注分布相似，所以关注度数目分布在这些地域是有广大的用户基数为基础的。

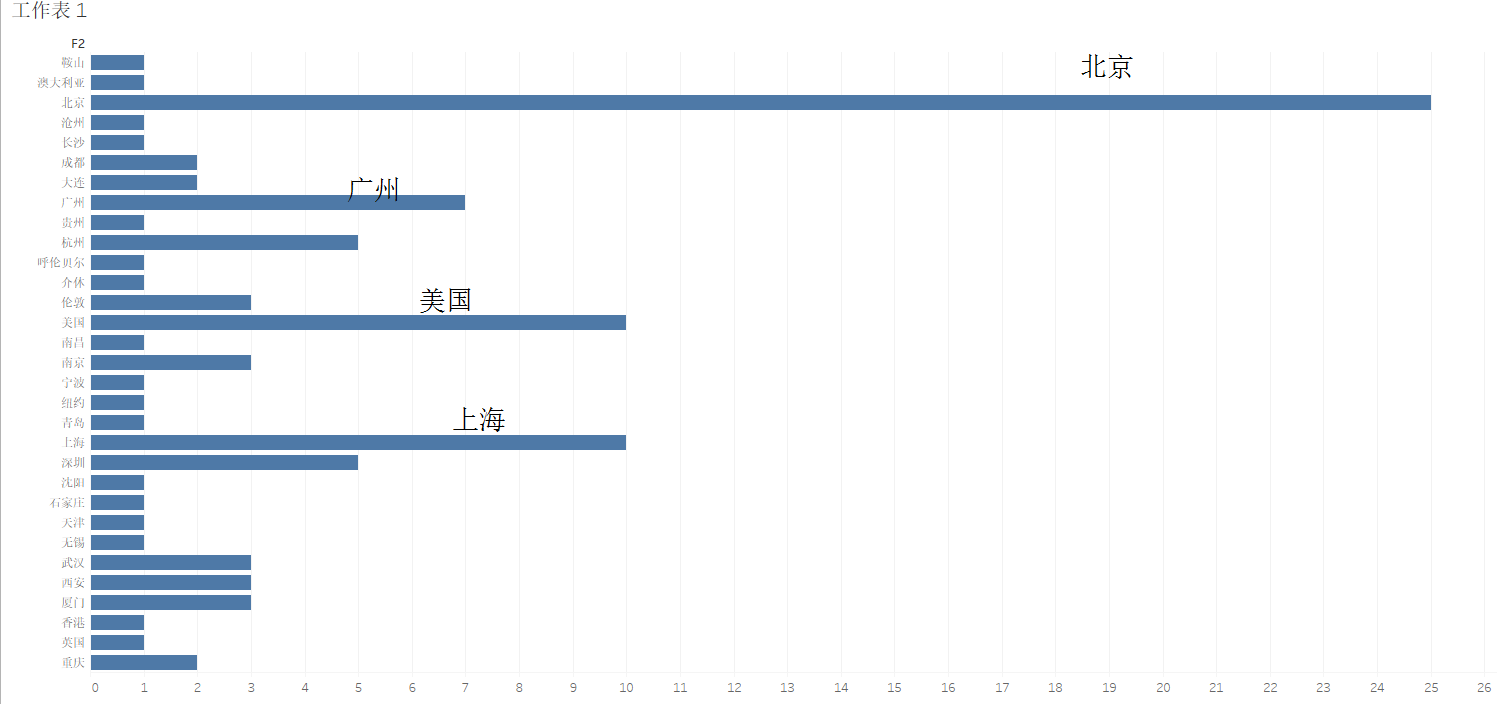


图 4 被关注数TOP100用户地域分布

从图4可以看出表中粉丝量排前100位的用户主要还是分布在北京、广州和上海这些超级城市之中，故可以推断用户如果居住在超级城市或者有国外留学经历，那么对其成为核心用户有一定的帮助。

* + 1. **行业**

由图5可以发现，知乎社交平台中从事计算机与互联网行业的用户最多，其次就是正在接受高等教育的学生或者从事高等教育的教育工作者。金融，艺术，高新科技，法律，科研等热门行业位于第二梯队，符合知乎的高精英人才定位。

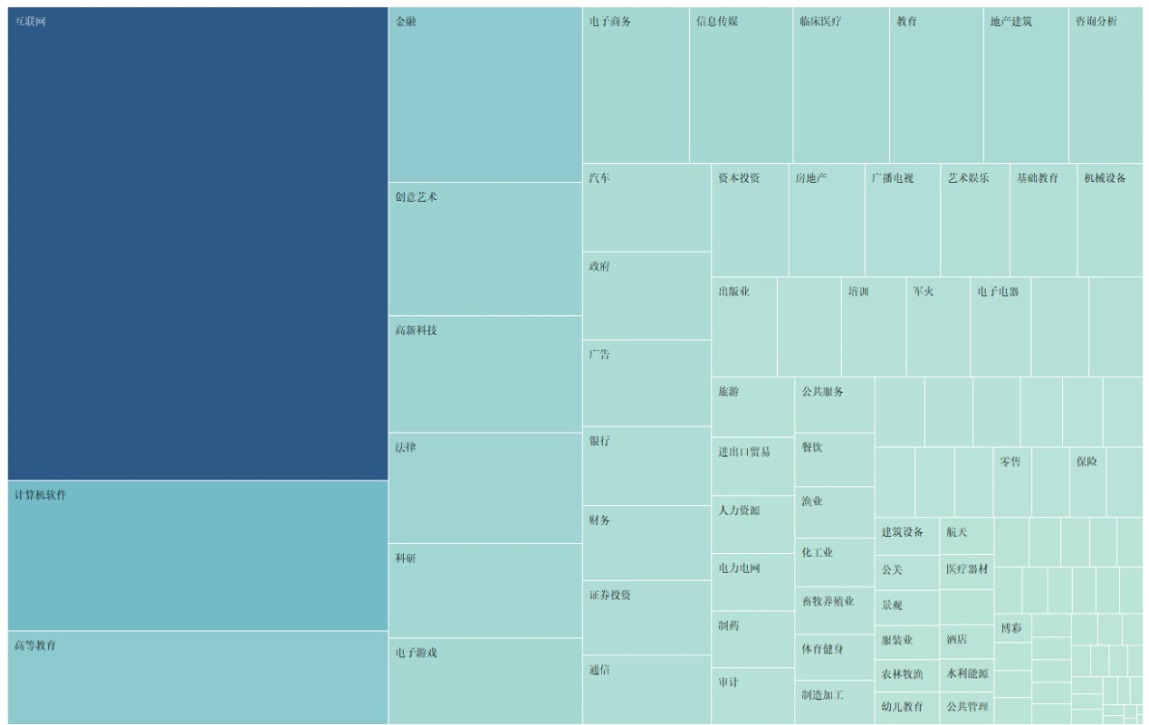


图 5 知乎用户行业分布情况

由于各行业用户基数有很大区别，某以行业用户越多越有可能带动该行业话题走向热门，从而涌现出一批具有高质量回答的用户，所以可以推测处于热门行业的用户更有可能成为核心用户。

同样从粉丝数前100的用户中调查其所处行业，分析结果仍是互联网、金融等行业占主导地位，所以可以较为正确的证明上述判断。

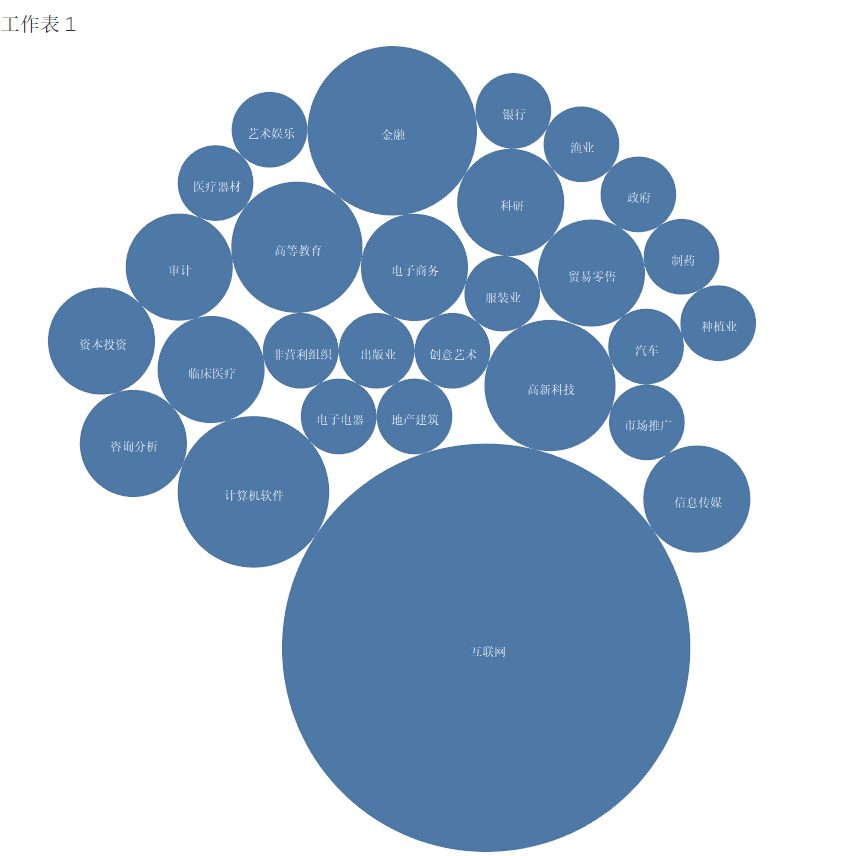


图 6被关注数TOP100用户行业分布

* + 1. **学历**

对于知乎的定位：精英人才的问答社区，那么高学历自然是精英人才所不可缺少的。如果没有足够学历的知识支撑，在知乎此类以问答为主的社区是无法长远的获得高关注度的。所以从EXCEL进行聚类分析，得出知乎用户分布较多的前三高校为：北京大学，清华大学，浙江大学（若将国外知名高校归为一类，那么也将位于前三）。

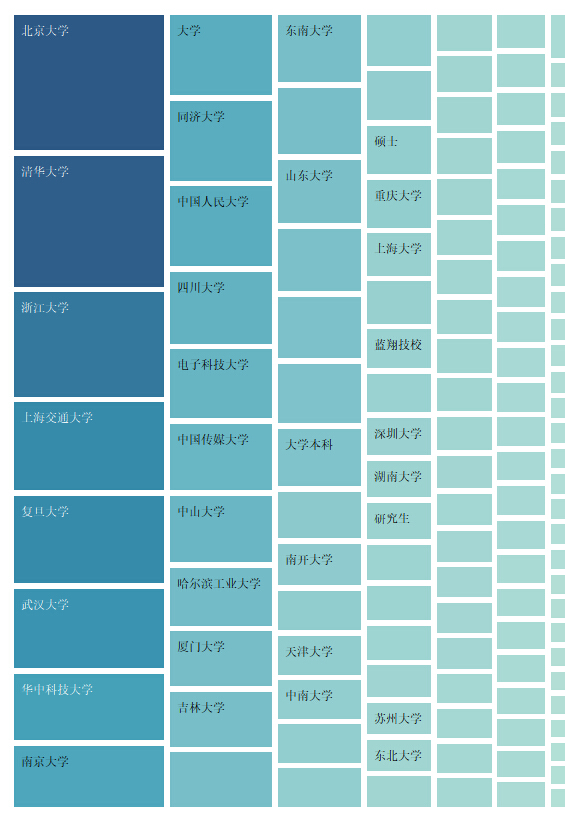


图 7 被关注数TOP100用户毕业院校分布

从粉丝数前100的用户中分析不难发现，大部分用户还是来源于各知名985,211高校。故高学历能为成为核心用户提供助力。

* 1. **用户行为**
     1. **回答**

知乎的定位为一个问答型社区，所以回答是社区的核心内容。

一般而言，一个普通用户在知乎从零起步的过程应该为：

1.回答问题

2.因为优质答案而获得赞同

3.引发别人兴趣，进而获得关注者

所以回答的数量和优质与否是取得大量关注者的重要环节。

但是即使有大量的答案输出，在知乎环境下，如果没有输出优质的答案，他的关注者也不一定会很多。根据回答数降序，随机抽取了一名用户查看，其回答数有4000左右，但关注数和赞同数极少。故高质量答案的输出对于成为核心用户尤为重要。

1. 判断优质答案

优质答案的判断可以从较多方面判断，例如插入图片数量，参考文献数目，引用数目等等。本文采用以字数为主要判断依据，进行判断该答案是否优质。

首先对所有用户回答的点赞数与字数绘制散点图，由图可以知道，除了少部分言简意赅的精简答案可以成为优质内容，获得赞同数目较多的还是千字左右的答案

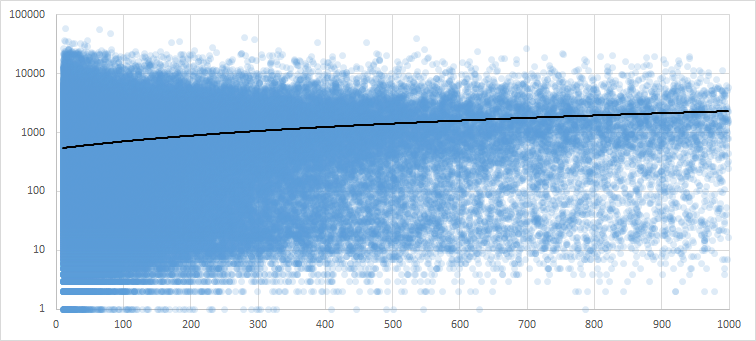


图 8用户点赞数与字数散点图（图片来源于知乎用户苏莉安）

* + 1. **赞同**

赞同是评定某答案是否优质的重要指标，它来源于多位用户对于该答案的认可，赞同数目越多越说明该答案被用户的认可度越高。故用户对答案的认可度决定了用户是否会有点赞这一行为

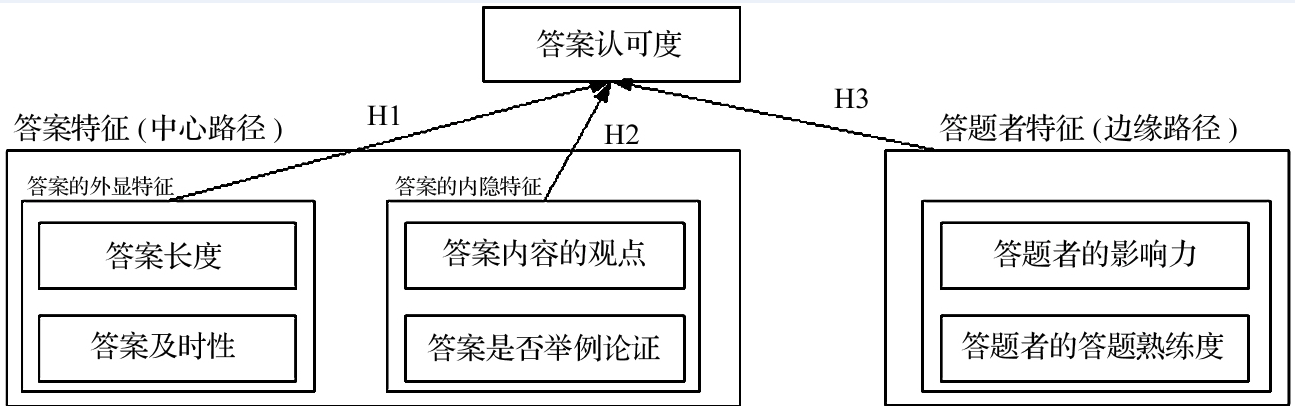
在知乎的问答体系中认可度的具体来源：如图9[1]，答案的认可度受答案特征与答题者特征的双向影响，同时答案本身的特征为中心路径。

图 9 用户对于答案点赞的影响因素

所以知乎以问答的方式进行互动，其答案赞同数是较难提升的。同时获得知乎的赞同与其他社交网站的赞同不太一样（不如新浪微博动辄万赞万转发），知乎很难具有突破千赞万赞的例子。知乎的赞同需要回答者足够的付出



图 10 知乎用户赞同数排行（来源于知乎用户苏莉安）



图 11 微博某用户一条微博的转发点赞详情

从图10与图11可得，与微博对比中知乎的赞同数的含金量较大，有重要的借鉴意义，可以成为判断用户是否成为了核心用户的关键依据。

* + 1. **关注**

一个陌生用户关注的动机往往是因为该用户的多个答案优质或者与该用户有较多的互动且有相同的价值取向，或者受从众心理影响由于该用户本身受关注度很高。

若某位用户关注了另一位用户，那么将会在首页收到关注用户的动态（他的回答，赞同，收藏等），所以被关注者有一种导向作用，帮助关注者更方便的找到优质答案。对于被关注者来说，关注者起着一种激励与监督作用，督促被关注者更多更好的输出优质答案

对于知乎而言，由于关注的存在，可以创造一个良性循环，更多的用户可以通过关注某个VIP用户获得所需干货，增强社区活力。所以对于知乎来说，对核心用户的发现尤为重要。

但是如果一个用户拥有了较高的关注量却存在部分答案质量不高的问题，那么高关注量是否也会给答案的赞同造成影响？



图 12 随机抽取黄继新的回答



图 13知乎创始人黄继新的个人资料页

对比图12与图13可知，知乎的创始人黄继新本身具有78万关注者，但却在其部分答案中却出现了个位数赞同的现象，从对比中可以得出这样的结论：

高关注度很难对于赞同的数目产生大的影响，大部分用户依旧按照个人观点进行点赞的用户行为，这在一定程度证明了知乎赞同的有效性和保证了优质答案在社区中的传播。故用户VIP分析中，其赞同数的权重大于关注度的权重

4.核心用户的发现与预测

对于社交网站而言，应当注重的是当前核心用户的研究与未来核心用户特征的发现，故针对不同时期的核心用户，对当前核心用户进行发现查找，对未来可能出现的用户据其特征进行预测。

1.即采用经典的“二八原则”对不同时期用户进行权重分配，并计算其子属性在当前维度所占有权重大小：

当前核心用户：C1权重为0.8，C2权重为0.2；

未来核心用户：C1权重为0.2，C2权重为0.8；

1. 主观分析对子属性进行权重分配：

S1=0.6，S2=0.2，S3=0.2;

S4=0.4，S5=0.3，S6=0.3；

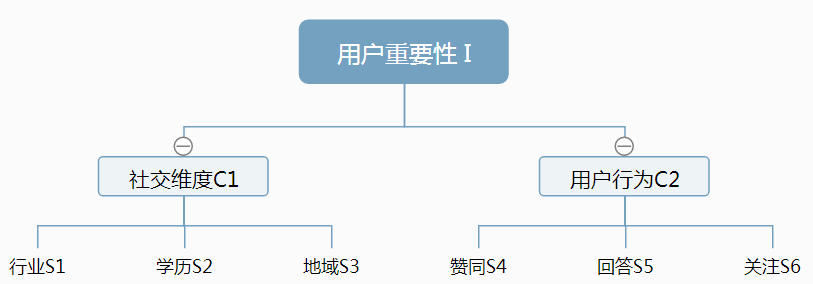


图 14用户重要性判定[2]

4.1当前核心用户的发现

I=（S1\*0.6+S2\*0.2+S3\*0.2）\*0.2+(S4\*0.4+S5\*0.3+S6\*0.3)\*0.8

4.2未来核心用户的预测

I=（S1\*0.6+S2\*0.2+S3\*0.2）\*0.8+(S4\*0.4+S5\*0.3+S6\*0.3)\*0.2

5．结论

为了研究如何在社交网络中发现其核心用户，本文通过分析知乎这一平台，以用户的社交属性与用户行为为主要内容进行核心用户的预测与发现，可以得出以下结论：

1. 社交属性：用户个人所在地域，行业，学历将会影响其成为核心用户。其中位于超级城市，处于热门高技术行业，拥有较高学历的用户有更大的可能成为核心用户。这有助于未来核心用户的预测。
2. 用户行为：点赞，关注，回答是用户在知乎的基本行为，回答为知乎核心内容，关注起着导向作用，点赞对优质内容有着传播筛选作用，故可以以此来衡量已成熟用户在知乎中是否是现阶段的核心用户，其中对于三个维度做了不同角度的剖析，得出其权重关系大小为：点赞>回答≥关注。

本研究也存在一定的不足：例如影响度公式的给出不够严谨，没有足够的数学背景支撑；没有对优质回答内容进一步的筛选；存在其它维度未计入其中等等，在后续研究中需要改进。

**参考文献**

[1] 施国良, 陈旭, 杜璐锋. 社会化问答网站答案认可度的影响因素研究——以知乎为例[J]. 现代情报, 2016, 36(6):41-45.

[2] 李家豪, 刘婧, 吴泽群,等. 基于模糊层次分析法的社交编程网站开发者评估方法[J]. 计算机应用研究, 2016(1):141-146.