基于回归分析的高校图书馆微博知识推荐影响力分析研究

马丹琳[1]，程秀峰[2]

（1.北京邮电大学经济管理学院，北京 100876；2.华中师范大学信息管理学院，武汉 430079）

摘要：在运用非概率抽样的滚雪球式抽样方法的基础上，采用 python 编写网络爬虫程序获取百所高校图书馆两个月内的微博相关数据，通过回归分析探究微博发布过程中各因素对知识推荐影响力的线性关系，同时依据分析结果，归纳出微博知识推荐过程中影响推荐质量与传播质量的因素，并提出提升高校图书馆微博账户知识推荐影响力的对策与建议。

关键词：高校图书馆，知识推荐，影响力，线性回归，微博

引言

伴随着社交媒体的发展，我国高校图书馆业务重心逐渐由“优化馆藏建设”向“优化用户服务”转移，深化用户服务内容与方式的一个重要手段就是图书馆官方微博的开通。重庆大学于 2009 年即开通了官方微博[1]，到 2015 年，绝大部分高校都已经开通了自己的图书馆微博，但从质量上看，图书馆微博的知识传播水平始终与图书馆本身的馆藏水平与文献利用水平有很大差距，例如，北京大学图书馆 2015 年 11 和 12 月的博文的平均转发量只有 3.93 次。因此，如何用好微博，使其作为知识传播与知识推荐手段的效用发挥到最大，是高校图书馆工作者必须面对的问题。笔者于 2014 年发表的《基于关联规则的高校图书馆微博关注趋势分析》一文中，介绍了微博数据的获取方法与流程，通过对高校图书馆关注数据进行分析，得出如何在高校图书馆官方服务账户中发现高质量的社区和核心用户。在此基础上，本文将微博账户的粉丝数、转发量、关注数作为衡量知识推荐影响力大小的重要指标。通过持续抓取高校图书馆微博账户的数据，采用线性回归的方法，探究三种显性影响因素（粉丝数、博文数、关注数），与三种隐性影响因素（原创率、类型、载体形式与转发量）之间的多维度线性关系，从不同维度分析各因素对微博知识推荐的影响过程与效用，以期针对分析结果提出有利于高校图书馆微博知识推荐发展的一些建议。

1 数据的获取方式及预处理

1.1 数据获取的抽样方式及其优点

本文依据《基于关联规则的高校图书馆微博关注趋势分析》一文中论述的依据关联规则发现关注关联性的方法，即“如果一个图书馆微博关注了某个高校图书馆微博，那么它有很大可能关注其他高校图书馆的微博”[2]，采用统计学中非概率抽样的滚雪球式抽样方法[3]，选取了 2013 年 9 月—10 月间高校图书馆官方微博账号与相关微博数据。滚雪球抽样式方法对用于稀少群体调查时具有明显优点，容易找到属于特定群体的被调查者，调查方式简单易行[4]。由于用爬虫进行数据挖掘之前需要找到大量图书馆官方账户，采用此类数据抽样的方法十分适合本次的研究。

1.2 数据的预处理

我们首先以普通高等本科院校为实例，分别以“大学图书馆”、“大图书馆”、“学院图书馆”为检索词进行图书馆微博用户的筛选，一共得到 943 条结果，其中经过新浪机构认证的用户占 21%，未经过认证（无“V”认证）的占 79%(如图 1)。对比认证用户和未认证用户的注册信息发现，认证用户大多为国内“985”及“211”工程等综合实力较强的大学及学院图书馆所开设的微博客，其在微博内容更新和内容管理上更加具有条理性、系统性，更具有影响力研究分析价值[5]。

接着，我们在得到的 198 所“加 V 认证”用户中，综合考虑某段时间内用户活跃度和高校综合影响力这两个因素，并数据进行人工排查、去重，最终遴选出 100 所高校作为本次研究的主要数据来源。[6]

随后，经简单统计发现，100 所高校图书馆官方微博认证账户中，985,211 高校的图书馆微博占 30%，非 985,211 高校的图书馆微博占 70%。

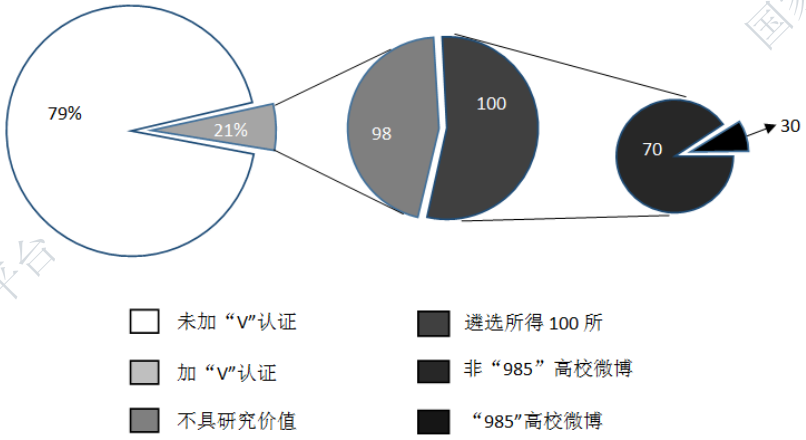

图 1 高校图书馆微博账户认证状况的数量分布

2 显性影响因素线性关系（粉丝数、关注数、微博数）

高校图书馆微博账户的粉丝数是体现其知识推荐影响力的重要指标；微博账户的关注数反映了微博账户在网络中传播知识的能力。同时，通过关注与被关注这一链接形态，微博实际上形成了其特有的知识传播模式；微博账户一段时间发布的微博数体现了其在该段时间内进行知识推荐的强度[7]。从认知层面上看，微博账户的粉丝数应与其一段时间内发布的微博数量、该微博账户的关注量存在某种函数关系，然而，这种函数关系是随时间微博内容等第三类因素变化而变化的，这就为它们之间的相关性提供了不确定性，我们只能通过时间片采样等方式尽量减少，而不能避免这种不确定性，这也就是为何在获得影响因素之间的线性关系后，需要对分析结果对预期进行主观判断的原因。

2.1 相关性的描述与测度

在进行多元线性回归分析之前需要判定变量之间的关系形态，即是否是线性相关。笔者通过将获取的 100 所高校图书馆官方微博的粉丝数量、关注数、微博数量导入 SPSS 统计分析软件，将关注数、微博数量作为自变量，粉丝数量作为因变量，输出结果如下：

$$R^2=\frac{\sum\left(\hat{y}_i-\bar{y}\right)^2}{\sum\left(y_i-\bar{y}\right)^2}=0.884 \quad , \quad s_e=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-k-1}}=298.1$$

经判定系数 R^2 检验：在高校官方微博账户粉丝数量的变差中，能被粉丝数与关注数、微博数的多元回归方差所解释的比例为 88.4%，具有良好的回归方程的拟合优度。因此，可以用多元线性回归模型对粉丝数与关注数、微博数进行分析。从随后所建立的多元回归方程来看，以关注数、微博数量作为自变量来预测粉丝数时，平均预测误差为 298.1 个粉丝。

2.2 线性关系及显著性检验

设自变量 x_1 为关注数、自变量 x_2 为微博数量、 y 为因变量粉丝数，采用最小二乘法估计多元回归方程，经计算得出：

$$\hat{y} = -3.471x_1 + 5.357x_2 + 533.357$$

线性关系检验 $F = 104.815 > \mathrm{F}_{0.05}$，证明粉丝数与关注量、微博数量的线性关系显著；$|t_1| = 2.114 > \mathrm{t}_{\alpha/2}$ 且 $|t_2| = 13.447 > \mathrm{t}_{\alpha/2}$，证明线性关系能通过回归系数检验。观测量累积概率 P-P 图（见图 2），得出残差分布服从正态性，证明线性回归正确。

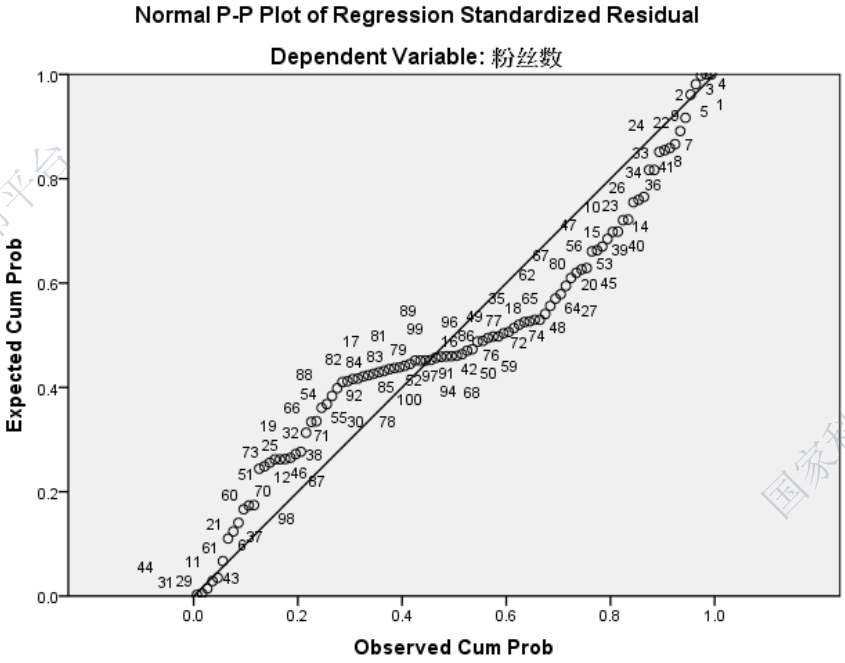

图 2 观测量累积概率 P-P 图

2.3 结果分析

统计结果说明，高校图书馆微博账户的关注数、微博数与粉丝数三者之间存在显著的二元线性关系。

第一，关注数（微博账户所关注的微博数量）与粉丝数量存在负相关，即随着该账户关注数量的增加粉丝数逐渐减少。

第二，高校图书馆一定时期的微博数量与粉丝数量存在正相关，即随着微博数量的增加粉丝数也会增加。

我们对结果进行分析解释：当高校图书馆微博账户的关注量达到一定的数量，且关注量与粉丝数比例高于一定程度时，该高校图书馆在网络中会失去其核心用户的地位，读者或已关注粉丝会通过关注其他核心用户获取核心知识推荐源，而有一定知名度的微博账户普遍存在关注微博数量较少的现象，而越是粉丝多的微博，其独特性越高，关注的其它微博的意愿就会越低；高校图书馆微博需要实时发布知识推荐信息，及时满足用户对知识的需求，并且高校图书馆微博数量在一定程度上体现了图书馆知识推送的及时性。因此，高校图书馆关注数与粉丝数量存在负相关。而一定时期的微博数量与粉丝数量存在正相关。

3 原创性与粉丝数之间的线性关系

3.1 内容原创比与粉丝数的回归分析

除了 2 中所述三种显性影响因素外，微博内容的原创性也体现了图书馆在信息生态链层

次中所处的位置与状态[8]：处于高层次“节点”的账户可以吸引更多的粉丝，且粉丝黏合度较高；处于低层次的账户会逐渐失去粉丝的忠诚度。为验证高校图书馆微博账户原创性与粉丝数量的关系，笔者通过高校图书馆微博内容的原创比率来体现微博内容的原创性，采用一元线性回归分析进行验证。

经检验 $R^2=0.831$, $s_e=87.6$。证明在高校官方微博账户粉丝数量的变差中，能被粉丝数与微博内容原创比的多元回归方差所解释的比例为 83.1%，具有良好的回归方程的拟合优度；并得出一元线性回归方程为：

$$\hat{y}=12.56x+51.7$$

经线性关系检验 $F=13.46>\mathrm{F}_{0.05}$，证明粉丝数与微博内容原创比线性关系存在显著的线性关系；且 $|t|=10.208>\mathrm{t}_{\alpha/2}$ 通过回归系数检验。

3.2 结果分析

统计结果说明，高校图书馆微博账户的原创性与粉丝数量之间存在显著的正相关。

我们对结果进行分析解释：微博平台中的信息传播并不像传统信息生态链那样自上而下、点对面的传播。[9]每一个高校图书馆微博账户在社会网络的信息生态链中均是一个“节点”，但不同“节点”之间存在着明显的强弱关系。信息生态链中的这一理论可以很好的解释高校图书馆内容原创比与粉丝数的线性关系。在高校图书馆中，往往微博内容含有较高比例原创性的图书馆微博可以形成“强关系”，而经常转发微博的图书馆微博会占据“弱关系”或逐渐走向“弱关系”。高校图书馆微博处在的“节点”关系越强，进行知识推荐时的影响力就越大；微博内容的原创性在一定程度上加强了“节点”关系的强度，这也恰恰符合了一元线性回归方式中微博内容原创比与微博账户粉丝数具有正相关性的统计结果。

4 微博内容与转发量之间的回归分析

4.1 微博内容分类的方法

微博内容是微博的直接体现，内容的质量无法定量判断，然而可以通过对其进行分类，观察各类微博对转发量之间的线性关系，从而判断微博影响力。经排查、去重后筛选出的 100 个大学图书馆微博大致可以按其微博内容划分为通知公告、特色活动、资源动态、书籍推荐和其他五种类型。[10] 笔者认为，图书馆开馆时间安排、馆内日常情况通报等可以归为通知公告类；由本校图书馆举行的展览、培训讲座、真人图书馆、新增信息服务项目以及各种非日常活动归为特色活动类；数字资源购买、更新，纸质资源购买以及特色资源库的构建等归为资源动态类；新到图书的内容简介以及具有阅读价值的书籍推荐归为新书推荐类；非以上内容归为其他类。[11]经归类统计得出高校图书馆一段时间内容的分类统计及转发量。（表 1 为统计之后的部分数据）

表 1 部分高校图书馆微博内容分类统计及转发量统计

高校图书馆	通知公告	被转发	特色活动	被转发	资源动态	被转发	书籍推荐	被转发	其他	被转发
清华大学图书馆	23	55	7	45	17	23	0	0	2	33
武汉大学图书馆	69	181	40	373	12	107	35	129	39	569
厦大图书馆	36	88	3	24	6	33	56	219	7	23

广外图书馆	10	104	27	75	0	0	13	79	34	625
复旦大学图书馆	15	45	22	137	12	127	21	128	22	152
暨大图书馆	37	178	16	66	5	51	13	62	36	91
重庆大学图书馆	23	93	12	43	0	0	6	17	18	77
华东师范大学图书馆	62	163	12	53	18	27	0	0	23	195

4.2 分析结果及结论

多元回归分析计算结果表明：通知公告、特色活动、资源动态、书籍推荐和其他这五类高校图书馆日常微博动态与转发量之间存在明显的线性关系，且通过线性关系检验和回归系数检验。其多元线性回归方程为：

$$\hat{y} = -6.594x_1 + 5.343x_2 + 3.581x_3 + 8.202x_4 + 1.288x_5 + 6.797$$

通过分析不同自变量的回归系数可以得知：除通知公告与转发量之间存在负相关的函数关系外，其余自变量均与转发量之间呈现正相关；书籍推荐与特色活动对转发量的正相关系数最大；资源动态与其他类型的微博发布对转发量的提升显著效果较低。

该结果表明：用户对高校图书馆官方微博发布的通知公告并不感兴趣，过多的发布通知公告会降低用户的兴趣度；与书籍推荐、特色活动相比，发布资源动态和其他类的微博用户兴趣度较低。因此，高校图书馆官方微博应当通过合理安排发布微博的内容结构提升自身在知识推荐过程中的影响力。

5 信息载体与转发量的回归分析

5.1 微博信息载体的类型

微博实际上是文字、图片、视频三种载体的综合，而三种载体则各有特色。文字描述是一种比较传统的信息描述方式，通过简单的文字排列可以构成丰富的信息内容，但篇幅较长的文字容易让读者产生视觉疲劳，不利于信息传递；图片则是较为直观的信息表现方式，图片信息可以通过图形、色彩的排列来传达，同时规避了单一的黑白色调给读者造成的视觉疲劳；视频信息则是将文字、图以及声音结合在一起，从三方面对读者进行神经刺激，更容易让人产生共鸣，使人记忆深刻。

高校图书馆微博内容所依托的载体在很大程度上决定了所发微博的转发量，进而影响其知识推荐涉及的范围。不同载体形式是否对转发量有显著影响，而哪种载体对转发量的影响较大是高校图书馆通过微博进行知识推荐时应当广泛关注的重点。

5.2 分析结果及结论

多元回归分析计算结果表明：文本、图片、视频作为转发数量的影响是明显的，其具有明显的线性相关性；且均通过线性关系检验和回归系数检验。其多元线性回归方程为：

$$\hat{y} = 2.428x_1 + 5.37x_2 + 9.557x_3 + 7.831$$

由方程可知，在不考虑其他因素的情况下，综合发布文本、图片、视频这三种形式的微博对知识推荐影响力的增大均存在正向相关关系；这比之发布纯文本、纯图片或纯视频的微博有较大优势。增加影响力（转发量）的关键点在于三中载体形式的综合与平衡使用。

同时，我们可以看出，视频对知识推荐影响力的促进作用最大；在一般情况下高校图书馆微博账户发布视频进行知识推荐的影响力最高，而仅以纯文本形式进行知识推送对用户的影响力最有限。

另外，高校图书馆发布微博向读者进行知识推荐，读者则根据自己的兴趣选择是否接收

该信息，那么这个过程可以看做是一个沟通过程。根据 Hovland、Janis 和 Kelly 提出的沟通说服理论，影响沟通效果的因素划分为三大类，即信息来源者的因素、信息本身因素以及信息接受者的因素。信息本身的因素又包括信息的内容、信息的形式、信息的数量等因素。而信息的形式不仅有口头的、文字的，还有图片、动画、视频等视觉信息，即视觉信息线索。薛建儒等学者认为视觉线索（Visual Cues）是对视感知的一种激励[12]，同时对于网络知识推荐领域来说，趣味性是网络信息的重要特征[13]，而丰富的视觉线索激起读者对知识的兴趣，给读者带来更加深刻的印象，并且影响信息的传播效果。

6 提升微博知识推荐影响力的具体建议

高校图书馆微博将微博平台作为知识传播媒介，通过微博的发布与转发实现知识的自创与自播，达到知识的充分共享与充分传递[14]。本文将微博的影响力因素分为显性因素与隐性因素两类：显性因素（微博数量、关注数量与粉丝数量）更加类似于评估指标，通过它们线性相关性分析，可以得出了在某一时间段中三者之间存在着某种动态平衡规律。其中微博数量与关注数量是可调控的，而粉丝数量不可主观调控，但是可以利用前二者与第三者间的相关性规律，解释这三种核心因素与转发量这一主要评价指标的关系。即：①微博数量可以作为转发量的重要影响因素，其与影响力成正相关。②关注其它微博的数量与转发量呈负相关，但这这种负相关不是绝对的的，它受到其它隐性影响因素的影响，在某一时间段这两者可以保持动态平衡，但这种平衡比较脆弱。③粉丝数与转发量之间存在正相关，但不是显性影响因素。若我们将知识本身作为链接源，将微博传播知识的过程看作用户对知识进行充分吸收后的自传播，进而达到知识的再分配过程，则更加便于我们科学利用微博作为手段进行知识推荐。

隐性影响因素在微博的创造、组织、利用与传播中的作用超过显性影响因素，从微博的内容分布、载体形式、主题热度到原创性，无一不是提升微博影响力的重要手段，科学合理观察与利用这些隐性影响因素，能使微博知识推荐工作更加成熟：

第一，高效稳定的更新频率。如果不能及时、有效地更新信息内容，保障推荐知识的时效性，就很难吸引跟随者的持续关注。不固定的更新频率也会让高校图书馆发布的微博被读者端巨大信息量所湮没，进而导致图书馆微博发布的信息以及推荐的知识不能及时有效的传播给用户。[15]

第二，尽量充分的原创内容。由上文分析可知，提高微博内容的原创性是保持微博热度的有效方法。原创程度越高，越能吸引更多用户，越能保持已有粉丝的用户黏性，从而占据信息生态链的顶端。另外，微博的原创性代表着馆藏资源的特殊性与核心竞争力，这种特殊性不仅体现在知识本身，更体现在文字结构、语气以及组织形式上。

第三，动态平衡的知识类型。在推送知识时，应当注重合理分配微博内容形式：既要避免仅仅发布单一类型的知识，也要避免种类过于庞杂。在同一时间段内，宜将图书馆近期需要发布的内容进行预分类，将热点知识与一般知识进行合理搭配、循序渐进的发布相关内容。

第四，灵活友好的用户体验。微博需要结合用户评论与反馈，将知识有针对性的推荐给相关读者。同时，采取措施避免知识推荐的单调、枯燥；在挖掘需求的基础上，以灵活友好的形式发布知识。

第五，科学合理的知识载体。图片、文字、音视频为微博自创与自播提供的载体。由本文第 5 部分的分析结论可知，微博的知识载体的选择在很大程度上影响着读者的阅读感受及兴趣。在进行知识推荐时，直观的图片配以文字可以在一定程度上增加读者的阅读兴趣；而采用视频作为知识推荐的载体更能使用户接受。但是不能长时间运用同一载体。

第六，冷热适度的热点议题。图书馆需要结合时事热点进行知识发布，这样能突出微博的实时性，增加用户兴趣，但不能一味强调时事新闻的重要性，而忽视了图书馆传播知识的本职，需要将知识与时事相互融合，使之相得益彰。

第六，把握总体平衡。提升高校图书馆微博知识推荐的影响力，更需要在总体把握显性影响因素与隐性影响因素之间的平衡，例如，不能因为粉丝过少而在短时间内盲目提高内容的趣味性；不能因转发量过少而花太多精力在视频的采集与制作上；不能影响力较高而减少对其它图书馆微博的关注。

总结

总之，利用单一手段提升图书馆微博影响力已难以吸引用户，作为高校中传播知识的重要窗口，图书馆微博需要综合利用各种方法，科学合理的改进原有的图书馆馆藏资源推荐模式，深入细致的挖掘本馆特色资源，将更多精力投入到知识推荐的深层挖掘与有效组织上来。使微博成为具备一定影响力知识推荐工具，成为移动用户了解图书馆，获取馆藏资源的重要渠道。微博以它独特的信息传播方式缩短了用户与知识的鸿沟，在知识推荐领域存在着极大的潜力和价值。现代图书馆的建设，应在提高馆藏资源的利用率的同时，充分发挥社交网络在知识服务中的价值。对于有一定影响力的微博，本文数据可以作为保持其影响力的参考资料；对于影响力落后的微博，需要合理调整自己的知识组织模式，学习其它微博的先进经验。

参考文献

[1] 陈琳. 国内图书馆微博应用现状研究[J]. 图书馆学研究, 2011(12):30-33.

[2][6] 程秀峰,李重阳,陈莉玥.基于关联规则的高校图书馆微博关注趋势分析[J]. 图书情报工作. 2014(08):73-78.

[3] Goodman L A. The annals of mathematical statistics [J]. Institute of Mathematical Statistics, 1961, 32 (1): 148-170.

[4] Newman M E. The structure and function of complex networks [J]. SIAM review, 2003, 45 (2): 167-256.

[5] 刘静. 我国高校图书馆认证用户微博调查分析——以新浪微博为平台[J]. 图书馆学研究, 2012(1):90-95.

[7] 祝方林. 大学图书馆微博信息行为分析[J]. 高校图书情报论坛, 2012(6):7-10.

[8] 马捷,孙梦瑶,尹爽,韩朝.微博信息生态链构成要素与形成机理[J]. 图书情报工作. 2012(18):73-75

[9] 杨小溪.网络信息生态链价值管理研究[D].武汉：华中师范大学, 2012：25-27.

[10] 杨玫.公共图书馆微博推广实证研究——以杭州图书馆为例[J].情报资料工作, 2012,(4): 102- 105.

[11] 刘钟美, 张文彦. 高校图书馆的微博新时代[J]. 图书馆理论与实践,2012(4):77-81.

[12] 薛建儒，郑南宁，钟小品，平林江.视感知激励——多视觉线索集成的贝叶斯方法与应

用[J],科学通报，2008, 53(2):172-182.
[13] Chen，Q.，Rodgers，S. Development of instrument to measure web site personality[J].Journal of Interactive Advertising，2006,7(1),47-64.
[14] Mary Hricko. Using Microblogging Tools for Library Services[J]. Journal of Library Administration, 2010(50): 684-692.
[15] 吴桂英等.基于微博客的图书馆信息交流服务模式[J]. 图书馆学刊, 2012(3): 30.

作者简介
马丹琳，女，1993 年生，北京邮电大学硕士研究生，研究方向：大数据、移动互联网， Email：danlinm164_bupt.edu.cn
程秀峰，男，1981 年生，华中师范大学信息管理学院讲师，研究方向：大数据、数字图书馆，

Analysis on Influence of Weibo Knowledge Recommended by College Libraries Based on Linear Regression

Ma DanLin [1], Chen XiuFeng[2]

(1. School of Economics & Management Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. School of Information & Management Central China Normal University, Wuhan 430079, China)

Abstract: On the basis of using snowball sampling—one of the non-probability sampling techniques, grab the relevant date of college libraries microblog within two months by programming web crawler via python, and then explore liner relationship between the factors in the process of microblog release for the influence of recommending knowledge. According to the result of analysis, summarize the factors that influence the quality of recommendation and transmission, and propose countermeasures and advices to promote the knowledge recommendation influence of college libraries microblog account.

Keywords: College Libraries; Knowledge Recommendation; Influence; Linear Regression; Microblog