doi:10.3772/j.issn.2095-915x.2017.02.008

面向舆情监管的大学生微信公众号关注情况分析

大连海事大学交通运输管理学院　大连　116026

于卫红　李杰

摘 要 近年来，微信公众平台迅猛发展，在大学校园内，微信公众平台已成为舆论的生成地和重要的扩散渠道，因此，应全面分析大学生对微信公众号的关注情况，为建立及时有效的舆情引导和监管机制奠定基础。本文收集了近 2000 名在校生的基本信息及其所关注的公众号信息，分析了样本数据的特征，构建了大学生微信公众号关注情况分析模型，并给出了一个基于 Hadoop MapReduce 模式与 Eclat 算法的分析实例。

关键词： 校园舆情，微信公众号，Hadoop 大数据平台，大数据挖掘

中图分类号： G35

Analysis of College Students' Attention to WeChat Official Accounts Oriented to Public Opinion Supervision of Campus

Transportation Management College, Dalian Maritime University, Dalian 116026, China

YU WeiHong　LI Jie

Abstract In recent years, with the rapid development of WeChat official accounts, they have become the generation of public opinion and important diffusion channels in university campuses. Therefore, in order to establish a timely and effective public opinion guidance and supervision mechanism, a comprehensive analysis of college students' attention to WeChat official accounts need to be done. In this paper, the basic information of nearly 2000 students, and the information of official accounts which the students subscribed

基金项目： 本文受2015年度教育部人文社会科学研究规划基金项目：微信环境下基于大数据的高校舆情监管机制研究（15YJAZH102）的资助。

作者简介： 于卫红（1972–），博士，副教授，研究生导师，研究方向：智能信息处理；李杰（1992–），硕士研究生，研究方向：智能信息处理。

were collected. The characteristics of sample data were analyzed. Moreover, a model for analyzing college students' attention to WeChat official accounts were constructed, and an analysis case based on Hadoop MapReduce and Eclat algorithm were presented.

Keywords: Campus public opinion, WeChat official accounts, hadoop big data platform, big data mining

1 引言

据微信官方最新数据显示，目前，微信公众号的数量已经突破 1000 万，每天还在以 1.5 万的速度增加[1-2]。微信公众号所具有的“操作便捷、人际交流高时效、发布内容丰富”等特点，极大地迎合了青年人追求时尚、猎奇八卦、关注时事等心理。因此，作为微信用户的主体之一，大学生群体关注的微信公众号更是数量众多、五花八门。由于发布者水平良莠不齐，传播又相对封闭、纠错机制弱，很多微信公众号频繁发布旧文章、老谣言、伪科学等垃圾信息，使得微信公众平台成为了舆情易发难管的地带。大学生群体知识层次高、思维活跃、参与意识强但尚未完全成熟，部分大学生更是主流意识模糊、价值观念异化[3]，很容易受不良舆论的影响改变其心理、人生观、社交方式等并导致校园内舆论传播的“蝴蝶效应”。

因此，全面掌握大学生对微信公众号的关注情况，可以从中分析出大学生的兴趣爱好、就业意向、心理特征、价值取向等，并进一步有针对性地帮助大学生提高信息甄别的意识和能力，及时有效地进行舆情引导和监管。

目前，很多学者开展了微信对大学生思想政治教育、人际交往、教育管理等方面的影响的研究。如，全永丽、李烽、李雯静等研究了微信与思想政治工作的联系，提出了利用微信有效开展大学生思想政治工作的对策[4-6]。温如燕、史新权等分析了微信对大学生社交行为的积极和消极影响，探究了如何利用微信提升大学生社会交往的质量[7-8]。郑天炜、韩雅嫩等分析了微信给高校学生教育管理带来的机遇和挑战，提出了利用微信提高高校学生教育管理实效性的对策[9-10]。上述成果从不同的角度分析了分析对大学生的影响，为其他学者提供了可借鉴的思路，但这些研究大都以定性分析为主，缺乏从大量数据中寻找规律、趋势等的量化过程。

本研究从校园舆情监管的角度出发，在我校范围内收集了近 2000 名在校生的基本信息及其所关注的公众号信息，分析了这些样本数据的特征，构建了基于 Hadoop 大数据平台与挖掘算法的大学生微信公众号关注情况分析模型。本文给出了使用 Eclat 算法在 Hadoop 大数据平台下对样本数据进行关联规则挖掘的实例。关联规则挖掘的目的是发现学生普遍关注的微信公众号及其关联模式，利用这一结果可选取关注度较高的微信公众号进行舆情信息智能采集的研究或利用关联规则开展公众号个性化推荐服务，为后续校园舆情的针对性引导奠定基础。定量分析之后，本文根据挖掘结果，做了定性的思考，提出了利用微信进行校园舆情监管的相关建议。

doi:10.3772/j.issn.2095-915x.2017.02.008

2 样本数据的特征分析

根据研究需求，项目组成员按专业分头行动，采用面对面访谈、问卷调查等方法，在我校范围内收集了近 2000 名学生的数据，数据内容主要包括：

（1）学生基本信息：年级、专业、性别、政治面貌、是否为学生干部。

（2）学生所关注的公众号：公众号名称、公众号功能简述、公众号的地理位置信息。

（3）学生对公众号所发布的文章的阅读情况：平均阅读率、平均点赞率、平均转发率、对公众号的关注持续时间。

首先，这些数据在分析过程中又会产生大量中间数据，综合来看具有大数据的 4V 特征：

（1）Volume（数据体量大）：很多学生都关注了近百个公众号，形成了庞大的公众号集合。公众号之间在领域、功能、发布内容等方面可能会存在一定的关联性，当用 Apriori 算法进行关联规则挖掘时，需要多次扫描原始数据，生成大量备选项集，计数工作量将会变得非常大[11]；而且，如果将支持度和置信度设置得较小，在单机上使用 Apriori 算法进行近两千名学生所关注的公众号之间的关联规则挖掘必然导致堆内存的溢出。

（2）Variety（数据具有多样性）：样本数据不仅包括“学生基本信息”这样的结构化数据，在对公众号进行内容挖掘时还涉及到图片、视频、音频、地理位置等非结构化数据，数据具有明显的多样性，对处理能力提出了很高的要求。

（3）Value（数据的价值密度低）：从内容挖掘的角度来看，尽管我们收集来的学生关注的公众号数量众多，但同一个领域的公众号，很多人都是关注了好几个，这些公众号内容同质化相当严重，从中能提取出的有价值信息却极为有限。

（4）Velocity（数据的变化速度快）：一方面，微信公众号的数量增加速度快，学生对公众号的关注也在随时发生变化（增加或取消关注）。另一方面，公众号推送的内容更新速度快。

鉴于样本的上述大数据特征，应考虑选择 Hadoop 之类的大数据平台，使用分布式处理模式构建数据智能分析与挖掘项目。

其次，采集到的原始数据还存在大量的“脏数据”现象：

（1）数据不一致：被调研对象在填写所关注的公众号信息时，有的提供公众号名称，而有的则提供公众号的号码；在填写专业信息时，有人填写专业的全称，而有人则填写简称。因此，需要检查数据的一致性，在数据存储前进行标准化处理。

（2）数据残缺或错误：此类问题通常是由被调研对象的疏忽造成的，如公众号的名称中遗漏了汉字或英文字母、某些信息存在错别字。

（3）数据无效：部分数据通过网络问卷的形式获取，存在胡乱填写的无效答卷，必须将此类数据删除掉。

鉴于上述“脏数据”现象的存在，必须对原始数据进行有效清洗[12]，防止大数据“垃圾进垃圾出”，才能达到“去粗取精、去伪存真、化零为整、见微知著”的数据挖掘效果。

3 大学生微信公众号关注情况分析模型

明确了样本数据的特征之后，遵循数据清洗、数据挖掘等步骤对大学生微信公众号的关注情况进行分析，其结果可应用于大学生兴趣关注点的发现、校园舆情意见领袖的发现、大学生普遍关注的微信公众号发现、校园舆情的监测与引导等。本研究提出了如图 1 所示的分析模型。下面对模型中的数据清洗与数据分析与挖掘环节进行阐述。

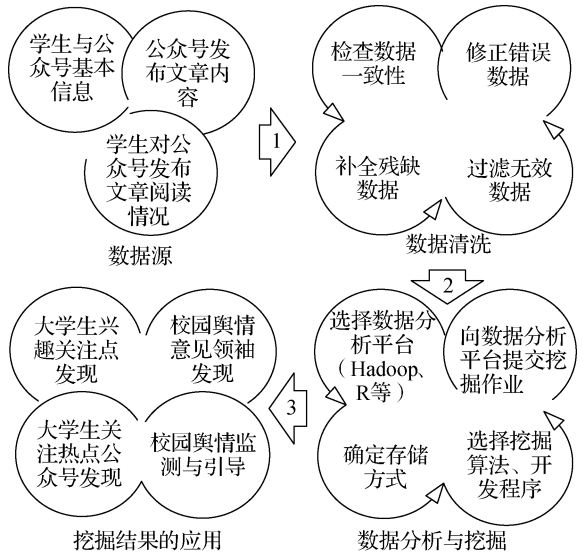

图1 大学生微信公众号关注情况分析模型

3.1 数据清洗

研究中，按照定义错误类型 → 搜索错误记录 → 修正错误的步骤，采用人工检测与统计算法相结合的方法进行数据清洗：

（1）对于数据不一致现象，通过建立数据字典、分析数据字典和元数据来梳理数据间的关系并进行修正。

（2）对于错误数据，在采用偏差分析、正态分布等统计方法的基础上，建立规则库，通过检查数值范围或基于属性的约束关系来识别错误。

3.2 数据分析与挖掘

3.2.1 Hadoop 大数据分析平台

本研究选用 Hadoop 作为数据分析平台，Hadoop 主要由 HDFS（分布式文件系统）、MapReduce（分布式计算）和 HBase（分布式数据库）组成，是一个处理、存储和分析海量的分布式、非结构化数据的开源框架 [12-13]，它在可伸缩性、健壮性、计算性能和成本上具有无可替代的优势，用户可以在不了解分布式底层细节的情况下开发分布式程序来进行数据分析工作。

3.2.2 基于 HDFS 的分布式数据存储

Hadoop 平台使用 HDFS 作为文件存储系统，能够在一个大集群中跨机器可靠地存储超大文件。HDFS 对存储的数据格式并无苛刻的要求，数据可以是非结构化或其它类别，其核心存储机制可概括为如下两点：

（1）主从式的存储结构 [13]：HDFS 集群以主从模式运行，主要有两类节点：一个 NameNode（即主节点）和多个 DataNode（即从节点）。NameNode 管理文件系统的命名空间、维护文件系统树以及文件树中所有的文件和文件夹的元数据。DataNode 是文件系统的工作节点，它们根据客户端或 NameNode 的调度存储和检索数据，并且定期向 NameNode 发送它们所存储的文件块列表。

（2）数据分块 + 副本的存储策略 [13]：HDFS 中的实际数据由 DataNode 负责存储和

doi:10.3772/j.issn.2095-915x.2017.02.008

维护，DataNode 的首要任务是 K-V 存储，K 代表 Key（键），V 代表 Value（值），也就是将数据组织成键值对的形式，一个独立的键值对应一条数据。把文件上传到 HDFS 上，HDFS 会根据设定的块的大小（默认是 64M），来分块存放文件，并且在不同的机架上放置多个数据块的副本（默认存 3 份），同时提供容错机制，当副本丢失或宕机时会自动恢复。

3.2.3 MapReduce 与智能算法相结合的分析模式

为了满足大数据处理的需要，在 Hadoop 平台下采用 MapReduce 计算模式，选取 Apriori、Eclat、K-Means、神经网络等智能算法开发分布式应用程序。MapReduce 是在 HDFS 的基础上实现的并行编程模式[13]，它采用“分而治之”的思想，把对大规模数据集的操作，分发给一个主节点管理下的各个分节点共同完成，然后通过整合各个节点的中间结果，得到最终结果。用于执行 MapReduce 任务的机器角色有两个：一个是 JobTracker；另一个是 TaskTracker，JobTracker 用于调度工作，TaskTracker 用于执行任务。一个 Hadoop 集群中只有一台 JobTracker。每个 MapReduce 任务都被初始化为一个 Job，每个 Job 又可以分为两种阶段：Map（映射）阶段和 Reduce（归约）阶段。首先，Map 函数把一组（Key，Value）输入，映射为一组中间结果（Key，Value），然后通过 Reduce 函数把具有相同 Key 值的中间结果，进行合并化简。简言之，Map 负责把任务分解成多个任务，Reduce 负责把分解后多任务处理的结果汇总起来。

4 分析实例 ---- 基于 Hadoop 平台的大学生关注公众号关联规则挖掘

4.1 样本数据

在本实例中，样本数据主要描述了学生的基本信息（年级、专业、性别、政治面貌、是否为学生干部）以及学生所关注的微信公众号名称列表。样本采用非结构化的方式存储，样本示例见图 2。

T1：国贸专业；女；大二；大连海事就业；
大连校园招聘；电影演出票；滴滴打车；
涤生爱电影；分众专享；FT中文网

T2：法学专业；女；大一；学生干部；爱在海大；
大连海事大学；看历史；美容瘦身日记；
新周刊报文周刊；深度好文；短片集；
大连3D错觉艺术馆；大连海事团委；
创意生活；人民日报

图2 样本数据示例

我们选取 Eclat 算法，在 Hadoop MapReduce 模式下对样本数据进行频繁模式和关联规则挖掘，从中发现学生普遍关注的公众号组合模式并提取关联规则。

4.2 基于 Hadoop MapReduce 编程模式与 Eclat 算法的挖掘过程

4.2.1 Eclat 算法与关联规则挖掘

关联规则是一个蕴含式[11]：$R: X \Rightarrow Y$，其中 $X \subset I$，$Y \subset I$，并且 $X \cap Y = \varnothing$，表示项集 X 在某一交易中出现，则导致 Y 以某一概率也会出现。用户关心的关联规则，可以用两个标准来衡量：支持度和置信度。对于关联规则 R，支持度反映了 X、Y 同时出现的概率，可信度是指包含 X 和 Y 的交易数与包含 X 的交易数之比。一般来说，只

有支持度和可信度较高的关联规则才是用户感兴趣的。

传统的数据挖掘算法多采用水平数据表示，在水平数据表示中，数据集合的一条记录由记录标识符（Tid）和一个或多个项目（Items）组成。Apriori 和 FP-Growth 算法都采用这种水平数据表示方法，这两种算法虽然简单易用，但在处理大数据方面都存在明显的缺点：

（1）Apriori 算法在处理过程中需要重复扫描数据集并且会产生大量的频繁集；

（2）FP-Growth 算法虽然只需对数据集进行两次扫描，避免了大量候选集的产生，但是该算法要递归生成条件数据集和条件 FP 树，所以内存开销大，只能用于挖掘单维的布尔关联规则。

因此，本研究采用了基于深度优先算法和垂直数据表示的关联规则挖掘算法—Eclat[14]。它的优势在于：（1）深度优先算法降低了对内存的需求；（2）通常速度比 Apriori 快；（3）不需要重复扫描数据集。Eclat 算法的基本原理为：

（1）引入倒排思想，扫描一次数据集，将水平数据表示转换成垂直数据表示，即：数据集中的每一条记录由一个项目及其所出现过的所有事务记录的列表构成。垂直数据表示的目的是为了求项集间的交集运算。如，用 Tid 表示事务 ID，用 Item 表示事务中的项（公众号列表），由水平数据表示转换成垂直数据表示的示意图如图 3 所示。

Tid	Item
1	大连海事大学就业；大连海事大学团委
2	大连海事大学；大连海事大学团委；大连领先考研
	大连海事大学就业；大连领先考研

水平数据表示

转换

Item	Tids
大连海事大学就业	1，3
大连海事大学团委	1，2
大连海事大学	2
大连领先考研	2，3

垂直数据表示

图3 水平数据表示转换成垂直数据表示的示意图

（2）交集运算：只需扫描一次数据集，然后通过交集运算生成候选频繁项集并计算其支持度。即，由频繁 k 项集求交集，生成候选 k+1 项集。对候选 k+1 项集做裁剪，生成频繁 k+1 项集，再求交集生成候选 k+2 项集。如此迭代，直到项集归一。支持度与置信度的计算公式为：

关联规则 x→y 的支持度定义为：

$$support(x,y)=\frac{|Tidset(x)\bigcap Tidset(y)|}{|T|}$$ （公式 1）

关联规则 x→y 的置信度定义为：

$$confidence(x,y)=\frac{|Tidset(x)\bigcap Tidset(y)|}{|Tidset(x)|}$$ （公式 2）

4.2.2 基于 Hadoop MapReduce 与 Eclat 的分布式数据挖掘

根据 MapReduce 的基本原理以及 Eclat 算法的核心思想，给出 MapReduce 模式下使用 Eclat 算法进行分布式数据挖掘的核心伪代码如下：

doi:10.3772/j.issn.2095-915x.2017.02.008

```
（1）Map输入：<key,value>形式存储的数据集，key表示记录的标示符，value表示项目集。
（2）Map函数：
    Map(key,value){
      ① 读每一条记录，使用字符串分割算法从value中获取所有项目；
      ②计算各项目的计数，从1-项集中删除小于支持度的项目；
      ③对剩下的项目按字典排序，排序后的结果存入items数组；
      ④对于长度为n的项目集，做如下循环生成2-项集，并输出2-项集和其Tidlist
      for（i=0；i＜n－1；i++）{
           for（j=i+1；j＜n；j++）{
                output（items［i］∩items［j］，key）；}}}
（3）Reduce输入：将Map函数的输出作为Reduce过程的输入，通过Reduce获得所有的2-项频繁项目集以及它的Tid-lists。
（4）Reduce函数：
         Reduce（key，value）｛
         Tid_list=null；
         sum=0；
         for (each Tid in value){
          sum++；
         Tid_list=Tid_list∪Tid；}
         if（(sum/总记录数)≥最小支持度）{ output（key，Tid_list）；//输出2-项频繁项目集}
    ｝
（5）重复上述Map和Reduce过程，生成不为空的频繁K项集。
```

4.2.3 挖掘结果分析与建议

基于上述原理，我们在 Hadoop 平台下使用 Eclipse 环境构建了 MapReduce+Eclat 的 Maven 项目，对样本数据进行了频繁模式和关联规则挖掘。频繁模式包含了学生普遍关注的微信公众号及其组合形式；关联规则揭示了学生信息与所关注的公众号间的内在联系。在实际挖掘中，用户可以根据需要设定支持度和置信度的阈值，以设定支持度为 0.4，置信度为 0.6 为例，部分运行结果如图 4 所示。

分析挖掘结果，给出如下结论及建议：

（1）尽管大学生个体关注的微信公众号五花八门，但是学校官方、团委、学院、专业等开设的主流公众号普遍关注度相对较高。这说明大学生对学校、学院、专业有一定的认同归属感、责任感和参与校园活动的意识，高校应充分保护并进一步提高学生的这种热情，利用微信之类的新媒体提升校园文化的舆论引导力，激发学生参与专业建设、校园建设的积极性。

（2）目前，高校内以学校、院系、社团等名义开设的微信公众号数量众多，高校应充分重视校园内微信公众平台的体系建设和安全管理。一方面，应鼓励高校党政机关、职能部门积极参与微信平台建设，构建以学校官方微信为主，院系、团委、专业等微信为辅，社团、意见领袖等微信为补充的校园舆情传播与引导新格局。另一方面，学校应重视对校园内微信公众平台的及时备案，建立有效的微信公众平台信息管理制度，防微杜渐，遏制假冒学校各职能部门名义发布虚假、不良信息的现象发生。

（3）从研究结果还可以看出，校园内主流微信公众号的关注度仍有很大的提升空间，各公众平台应积极加强自身建设，吸引更多的关注者，从而更好地充分发挥对校园舆情的引导和监督作用。高校舆情由主体、客体、介体组成，主体指的是高校师生，客体指的是现实社会特别是校园内发生的各种现象、事件等，介体指的是舆情的传播媒介，是连接主体和客体的桥梁，如微博、微信等。作为校园舆情的重要介体之一，微信公众平台应重视对主体的特征了解、对客体的内容建设。如，通过上述频繁模式和关联规则挖掘，能够发现不同年级、不同性别、不同专业、不同

政治面貌的学生的兴趣关注点，微信公众平台可以利用这些挖掘结果对关注者进行灵活地分组，作到信息的定向个性化推送，提高信息的阅读数、转发数、收藏数。在客体的内容建设方面，更应该充分理解大学生的年龄层次、心理特征，掌握他们的信息需求，对社会突发事件、校园生活问题、就业考研指导等的内容发布力争做到在形式上创新、在内容价值上贴近学生的需求。

= 频繁项集 ==========	= 关联规则 ==========

大连海事大学：支持度71.37%
大连海事大学团委：支持度68.28%
大连海事就业：支持度 67.37%
大连海事大学；大连海事大学团委： 支持度46.98%
爱在海大：支持度42.34%
……
……
	女；电商专业->海大电商： 置信度72%
	// 规则解释：电商专业的女生有72%关注了“海大电商”公众号
	爱在海大->大一：置信度 63.2%
	// 规则解释：关注“爱在海大”公众号的学生有63.2%是大一新生
	入党积极分子；国贸；大二；学生干部->党员活动室： 置信度95%
	// 规则解释：国贸专业既是入党积极分子又是学生干部的大二学生有95%关注了“党员活动室”
	……

图4 挖掘结果示意图

5 结语

鉴于微信公众平台对高校校园舆情传播与监管的影响，本文以近 2000 名在校生基本信息及其所关注的公众号信息为研究样本，分析了样本数据的特征，提出了基于大数据思想的大学生微信公众号关注情况分析模型。在给出的分析实例中，以 Hadoop 作为大数据分析平台，阐述了 MapReduce 分布式计算模式与 Eclat 算法相结合的关联规则挖掘方法，并对挖掘结果进行分析，给出了可用于指导校园舆情引导及监管工作的结论和建议。

参考文献

[1] 黄楚新，王丹 . 微信公众号的现状、类型及发展趋势 [J]. 新闻与写作，2015(7):5-9.

[2] 深圳市腾讯计算机系统有限公司 . 腾讯 2015 年第二季度及中期业绩报告 [EB/OL].[2015-05-08].http://www.tencent.com/zh-cn/content/at/2015/attachments/20150812.pdf.

[3] 吕娟 . 高校大学生网络舆情的监管与疏导研究 [D]. 上海 : 华东政法大学，2013.

[4] 全永丽 . 以微信为载体加强大学生思想政治教育研究 [D]. 长春 : 吉林大学，2015.

[5] 李烽 . 微信朋友圈对大学生思想政治教育的影响及对策研究 [D]. 武汉 : 华中师范大学，2015.

[6] 李雯静 . 微信对大学生思想政治教育的影响及对策研究 [D]. 牡丹江 : 牡丹江师范学院，2015.

[7] 温如燕 . 微信对大学生人际交往的影响研究 [D]. 兰州 : 兰州大学，2014.

[8] 史新权 . 微信对大学生社会交往的影响研究 [D]. 石家庄 : 河北师范大学，2016.

[9] 郑天炜 . “微时代”下高校学生教育管理研究 [D]. 西安 : 长安大学，2015.

[10] 韩雅嫩 . 手机微信公众平台下大学生管理探析 [D]. 西安 : 陕西师范大学，2015.

[11] 韩家炜等著，范明等译 . 数据挖掘 : 概念与技术 [M]. 北京 : 机械工业出版社，2012.

[12] 李金海，何有世，熊强 . 基于大数据技术的网络舆情文本挖掘研究 [J]. 情报杂志，2014(10): 1-7.

[13] 刘军 . Hadoop 大数据处理 [M]. 北京 : 人民邮电出版社，2013.

[14] 陈培恩 . 关联规则 Eclat 算法改进研究 [D]. 重庆 : 重庆大学，2010.