基于微博的大数据用户画像与精准营销

曾　鸿　吴苏倪　成都信息工程大学统计学院

摘要：在大数据时代，通过有关技术手段对新浪微博数据进行采集分析，构建用户画像模型，描述企业用户群体行为特征，为精准营销带来了可能。用户画像系统为企业提供全方位的掌握客户群体的信息标签，使企业了解、认知自己的客户。同时在品牌的传播与建设中，用户画像也是一个不错的思路。这为企业制定科学准确的营销方案打下了良好的基础。

关键词：大数据；用户画像；精准营销

中图分类号：TS941.1　　文献识别码：A　　文章编号：1001-828X(2016)024-000306-03

User image and precision marketing on account of big data in Weibo

School of Statistics, Chengdu University of Information Technology

Zeng Hong　Wu Su Ni　Chengdu 610103

In the era of big data,collect and analyseWeibo's data by relevant technology、structure userimage model and describe company user group's behavior characteristics make precision marketing possible. User image system let company obtain user group's information label comprehensively,then company can know its own customers. At the same time, user image is a good thinking in brand's diffusion and development.It makes a good base for company to make the scientific and accurate marketing programme.

Keyword:Big Data、User Image、Precision Marketing

大数据 (big data)，是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

IDC 调查公布人类在 2003 年全部的数据只有 5EB，只相当于现在两天的数据，到 2020 年全球的数据将达到 35ZB。丰富的数据资源蕴含了大量的客户信息，如何利用大数据为企业营销服务是当今企业正在探索的问题，也是未来企业致胜的法宝。企业可以运用大数据，对广大的人群进行筛选，得到目标用户，然后通过对目标用户的分析，得到目标用户的需求和特点，从而进行精准化营销。

大数据在企业市场营销中的应用，最先开始的是淘宝、腾讯、京东等几大电商的广告推荐系统。当浏览、购买或收藏了某些商品后，网民总能在商城的某个位置看到类似于“猜你喜欢”的内容推荐，或者说当网民在网上阅读一篇文章、观看一个视频或者和他人聊天时，总是能弹出一个广告推荐窗口，推荐的内容恰好是最近自己关注过的相关商品的信息，这便是典型的大数据精准营销的应用之一。

在大数据精准营销中，企业拥有自身的数据是远远不够的，业务本身带来的数据，往往容易形成信息孤岛。随着大数据技术的发展与应用，通过网络蜘蛛或爬虫，对消费者在互联网上留下的消费痕迹进行数据爬取，形成数据仓库，对每一个消费者进行用户画像，从而实现广告的精准推送，为企业抓取产品潜在的用户群体，提升产品的知名度和曝光度，使企业在市场占有一定份额，最终给企业带来盈利。

一、用户画像

在互联网逐渐步入大数据时代后，不可避免的给企业和消费者行为带来一系列改变和重塑。其中最大的改变莫过于消费者的行为在企业面前将是“可视化”。随着大数据技术的深入研究和应用，企业的专注点日益聚焦于怎样利用大数据来为精准营销而服务，于是“用户画像”的概念由然而生。

用户画像，即用户信息的标签化，是真实用户的虚拟代表，是建立在一系列数据之上的目标用户模型。用户画像根据用户的社会属性、生活习惯和消费行为等信息，抽取出一个或一类用户的标签，给用户信息进行结构化处理。用户画像的意义在于了解用户，猜测用户的真实需求和潜在需求，精细化地定位人群特征，挖掘潜在的用户群体，为媒体网站、广告主、企业及广告公司充分认知群体用户的差异化特征，帮助客户找到营销机会、运营方向，全面提升企业的核心影响力。

1. 用户画像的数据源获取

用户画像首先要获得的是用户的行为数据，在互联网上，用户的行为数据大多来自网站的访问日志，日志里记录了用户在网站的浏览、搜索、点击、跳转等一系列用户行为轨迹，这些不断变化的行为信息，归属于用户的动态数据。在大数据的技术运用中，除了网站的访问日志外，还可以利用网络爬虫技术，追踪用户在全网的行为信息，根据关联规则，联系到用户的行为偏好，将完善用户画像的模型，提高预测用户需求的概率，这需要确保用户 IP 的一致性，保证追踪到的用户信息的前后是一致的。

除了用户的行为信息外，用户画像还需要用户的个人信息，例如：用户的姓名、年龄、性别、职业、收入、地区、手机号等人口基本属性，这些信息相对稳定，可以归属为用户的静态信息。

对收集得到的用户画像数据进行训练，利用数据挖掘算法模型，抽取出用户画像标签，构建用户画像标签体系。

2. 用户画像标签建模

用户画像的焦点工作就是为用户打“标签”，而一个标签通常是人为规定的高度精炼的特征标识，如年龄、性别、地域、用户偏好等，最后将所有标签综合起来，就可以勾勒出该用户的“画像”了。用户画像的核心工作是为用户打标签，标签提供了一种便捷的方式，使得计算机能够程序化处理与人相关的信息。用户标签要求呈现出两个主要的特征，一是语义化，人能很快理解每个标签的含义，二是短文本，每个标签通常只表示一种含义。

用户画像标签建模主要包括四个步骤，首先得取得原始数据，这里的原始数据主要包括企业历史交易数据和用户的基本信息数据，另外一部分是互联网数据，这部分数据主要通过网络爬虫等技术，对用户行为数据进行爬取；其次对原始数据进行统计分析得到事实标签，例如年龄分布、性别比例、购买频率等；然后对事实标签进行建模分析，得到模型标签，例如人口属性、产品购买偏好和用户关联关系等；最后进行模型预测，得到预测标签，主要是对未来数据的一种用户行为预测。

二、基于微博的话题聚合和用户画像

新浪微博拥有超 5 亿级的用户数，这海量的数据后面隐藏着巨大的商业价值。新浪微博在用户注册的过程中，已有一些用户的基础信息，诸如年龄、地域、性别、关注数、粉丝数、兴趣标签等，但这些弱关系数据信息还不足以给定一个人或一群人的用户画像，为了使用户画像描述更加精确，还缺少相应的兴趣图谱。在新浪微博海量的数据中，不可能分析每一条用户微博后面的兴趣倾向，为此新浪微博通过兴趣话题，把对话题同样感兴趣的一类人聚合到一起，参与话题讨论，这样通过话题聚合，就能获取这类人群的信息，提取该类人群标签，构建人群用户画像，这样作为商家、广告商，就能对该类人群进行微博广告投放，达到精准营销的目的。

1. 数据源获取

本文数据来源于新浪微博热门话题数据，通过网络爬虫和微博指数，对数据进行采集、统计与分析。

通过利用爬虫软件，爬取了微博热门话题榜的数据。首先提取热门话题榜的话题和分类标签，共爬取了 700 条数据，然后进行数据筛选、清洗，得到有效数据 614 条。通过数据预处理后的数据，集成数据仓库，初步对数据进行描述性统计分析，本文主要进行的是分类数据统计，对于统计结果进行排序，挑选话题讨论最多的标签进行接下来的数据挖掘工作。

新浪微博为用户提供微博数据统计，微博指数是基于海量用户行为数据、博文数据，采用科学的统计方法得到反映不同事件领域发展状况的指数。本文将通过提取关键字，对用户群体进行标签化，最终对微博话题人群聚类画像。

2. 数据描述

新浪微博提供的话题分类标签共 37 个，具体包括明星电视剧、社会、综艺、电影、音乐、动漫、情感、美食、时尚美妆、旅游、读书、公益、美图、生活记录、教育、体育、文化艺术、笑话、运动健身、创意征集、科技、汽车、电视节目、财经、萌宠、游戏、健康、情感两性、星座、搞笑幽默、化妆造型、数码、投资理财、网络文学、休闲娱乐、政务。通过网络爬虫，对新浪微博话题榜的数据进行爬取，包含话题名称和话题标签，爬取结果部分内容截图如图。

	话题	标签
2	#盗墓笔记#	财经
3	#来吧冠军#	财经
4	#PressYourNumber#	财经
6	#杨怡罗仲谦结婚#	创意征集
7	#同道大叔为处女座...	创意征集
8	#UNIQ王一博#	创意征集
9	#搞笑精选#	创意征集
10	#我是歌手4尽在虾...	创意征集
11	#来吧冠军#	创意征集
12	#值得去#	创意征集
14	#天天向上#	电视节目
15	#超能星学园#	电视节目
16	#GFriend#	电视节目
17	#蜡笔小新#	电视节目
19	#鹿晗Reloaded演唱会#	电视剧
20	#恐怖将映#	电视剧

	标签	提及数
2	明星	64
3	电视剧	54
4	社会	53
5	综艺	53
6	电影	37
7	音乐	30
8	动漫	26
9	情感	21
10	美食	19
11	时尚美妆	16
12	旅游	14
13	读书	13

图 1 分类标签统计

对爬取的结果，按标签进行排序分类统计描述，得到热门话题榜里，37 个标签提及数，通过排序可以看出排名前 10 的标签依次是明星、电视剧、社会、综艺、电影、音乐、动漫、情感和美食。

3. 用户画像构建

新浪微博提供了一个明星和粉丝对话交流的平台，粉丝可以在微博关注自己喜欢的明星，与明星互动，拉动了明星和粉丝之间的交互作用；而明星的一些商业活动，对于粉丝来说便是一种刺激消费，这里的商业活动包括明星代言的广告、主演的电影、电视剧、出席的活动等，相应而产生的便是叫做“粉丝经济”的经济模式。对于企业来说，将企业营销决策与明星效应相结合，聚合该类粉丝人群，通过大数据对人群进行用户画像，掌握该类人群特征，对这些用户粉丝进行广告投放，有的放矢，避免了过多的广告输出造成的消费者反感。

在新浪微博的明星势力榜里，选取两位明星，对其粉丝人群进行人群画像。微博明星势力榜将对在一段时间内排名前 50 的明星进行打分，评分维度有提及量、互动量、搜索量和粉丝爱慕值四个维度。

选取 2016 年 3 月份微博明星势力榜排名前 10 明星榜单，按综合得分大小降序排序如图 3–5 所示，分别有 TFBOYS– 易烊千玺 97.41、TFBOYS 王俊凯 97.25、许魏洲 90.72、黄景瑜 90.69、王凯 87.57、TFBOYS– 王源 87.43、鹿晗 86.82、李易峰 85.84、杨洋 84.63、吴亦凡 83.84。

明星	提及量	互动量	搜索量	爱慕值	综合得分
TFBOYS-易烊千玺	1746806	3120830	8975471	565304	97.41
TFBOYS-王俊凯	2415918	3061212	6529074	531304	97.25
许魏洲	670900	1531332	7637932	135904	90.72
黄景瑜	410035	2001700	14492435	116840	90.69
王凯	331301	1410036	5870151	75678	87.57
TFBOYS-王源	1433937	2725675	3520894	34856	87.43
鹿晗	1932469	7080909	18082575	8764	86.82
李易峰	1067082	2857730	6907010	17658	85.84
杨洋	687006	2005314	3137126	22812	84.63
吴亦凡	741139	3120343	10999307	8588	83.84

图 2 三月份微博明星势力榜 TOP10

新浪微博每天都能产生大量的数据，基于海量的用户行为数据、博文数据，新浪微博为广大用户提供关键词微博指数，采用科学的计算方法统计得出的反映不同事件领域的发展状况。新浪微博指数包括热词趋势、实时趋势、地域解读和属性分析四个部分。基于微博数据，就可对明星类标签进行微博话题人群用户画像建模。

选取 TFBOYS 组合和王凯两位明星进行关键词的微博话题人群画像，时间自定义选择为 2016 年 3 月 1 日到 2016 年 3 月 31 日。

图 3 为输入关键词“TFBOYS”的整体趋势图，图 4 为移动端和 PC 端趋势图。由图可知，TFBOYS 在整个三月份微博被提及量都比较高，整体趋势当月热议均值达到 931541 次，当月最高达 1245028 次，PC 端当月热议均值达 241611 次，最高达 340919 次，移动端当月热议均值达 689930 次，最高达 904109 次。

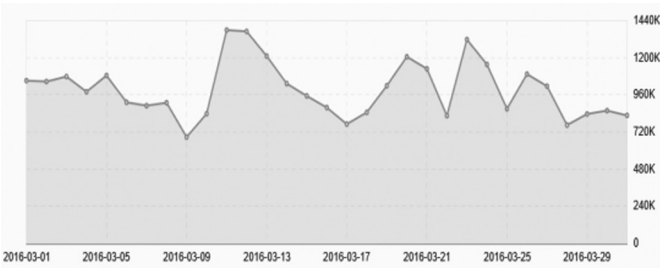


图 3 整体趋势图

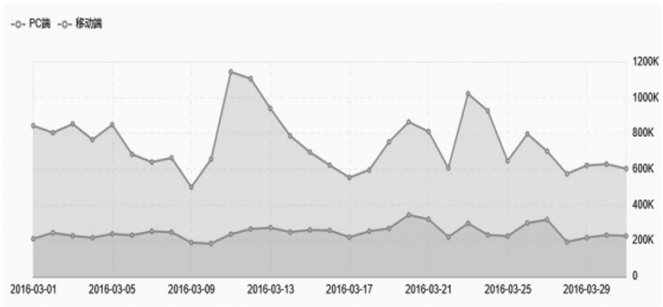


图 4　PC 和移动趋势图

从地域热议度和用户热议度看 (图略)，在排名前十的地域城市中，地域热议度和用户热议度差异并不大，其中排名前三热议度地区分别为广东、重庆和北京。

在属性分布下，主要是提及到 TFBOYS 的人群属性信息。在性别比例中 (图略)，男性占 16.96%，女性占 83.04%。在年龄分布下 (图略)，12 到 18 岁年龄段人数最多，占 38.18%，19 到 24 岁占 35.38%，25 到 34 岁占 13.92%，35 到 50 岁占 7.65%，12 岁以下占 4.58%，51 到 80 岁占 0.28%。在人群标签比例中 (图略)，名人明星标签占比最多为 133415 人。人群星座比例下 (图略)，魔蝎座占比最高为 12.76%，处女座为 11.84%，天蝎座为 10.1%。

以上我们通过微博指数对 TFBOYS 的粉丝进行了分析，现在微博指数输入“王凯”关键词，得出总体趋势图如图 5。

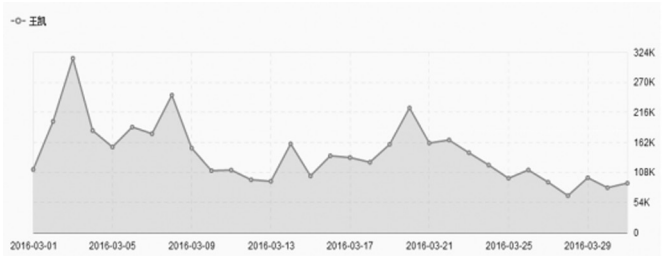


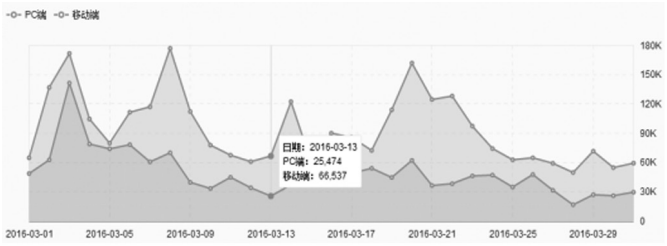
图 5　总体趋势图

从地域分布看 (图略)，其中地域热议度最高的为北京，占比 9.43%，用户热议度最高的也是北京为 8.37%；在人群属性分析下 (图略)，女性人数最多，为 86.94%，男性为 13.06%：在年龄分布下，19 到 24 岁人数最多为 39.59%，其次为 25 到 34 岁，占比 31.71%；从标签比例和星座比例，“名人明星”类标签最多，星座比例中魔蝎座和狮子座最多，分别占比 17.53% 和 14.63%。

通过以上分析，现在得到明星粉丝人群画像如下：

TFBOYS 粉丝群主要集中在广东、重庆和北京，粉丝人群中女性占绝大部分，年龄分布主要集中在 12 到 24 岁，粉丝人群自我标签主要为“名人明星”，这代表这部分用户偏好主要为“名人明星”，星座比例分布较为均匀。

王凯粉丝群主要集中在北京、江苏、上海和广东，粉丝人群中女性占绝大部分，年龄分布主要集中在 19 到 34 岁，粉丝人群中自我标签主要为“名人明星”，喜欢王凯的粉丝中星座分布魔蝎座和狮子座最多。



三、用户画像与精准营销

精准营销依托数据资源和渠道优化的优势，并结合数字化广告的营销传播渠道，将广告内容送达目标人群。精准营销的核心在于直击有需求的用户，然后再提供相应的产品和服务，通过这种针对性的营销，才能够实现更好的流量转化率。当获得了一个话题的用户画像信息后，商家就可以根据用户画像，对这部分群体进行广告投放。

通过以上研究，已获得了明星粉丝的用户画像，企业就可以根据用户画像制定精准营销方案。针对 TFBOYS 和王凯两位明星，明显可以看出两位明星粉丝的人群明显不同，因此挖掘精确的目标人群成了精准营销的关键所在。首先，企业可以在新浪微博建立品牌与明星相关的话题，话题里需要设置“TFBOYS”或“王凯”这样的关键字；其次，通过网络爬虫技术，把参与该话题讨论下的用户信息都收集起来，主要包括该用户的 ID、性别、地域、个人标签和该用户过往所发表微博内容；然后，根据微博的关联关系和该用户所发表的微博内容进行用户偏好分析，其中通过关联关系，可以判断该用户微博所关注的人群是否有该话题明星或和该明星相关的用户，或者通过大数据技术，对用户所发表微博中，是否有该话题明星的关键字，出现的频率是多少。最后通过数据挖掘得到了精准的粉丝群体。那么精准营销的最后一步便是广告推广，企业可以与新浪微博合作，购买新浪微博的广告排位进行广告投放。

在精准营销中，通过定量或定性地描述企业用户群体行为特征，为客户做用户画像给现代数字营销带来了可能。用户画像系统为企业全方位的掌握客户群体的信息标签，使企业了解、认知自己的客户。同时在品牌的传播与建设中，用户画像无疑也是一个不错的选择。

参考文献：

[1] 卞友江 . “大数据”概念考辨 [J]. 新闻研究导刊 ,2013,(5):25-28.

[2] 许瑾 . 精准营销探析 [J]. 信息网络 ,2006,(8):26-27.

[3] 茹仙古力・艾合麦提 . 浅谈文献搜索与毕业论文 [J]. 大科技・科技天地 ,2010,(5):142-143.

[4] 尹启华邓然 . 精准营销研究现状 [J]. 经济研究导刊 ,2010,(9):158-159.

[5] 赵晖 . 互联网思维下的“粉丝经济”[J]. 上海信息化 ,2014,(6):25-27.

作者简介：曾　鸿 (1964-)，成都信息工程大学统计学院教授，研究方向：统计理论、市场调查。

吴苏倪 (1993-)，成都信息工程大学统计学院学生。