

研究报告

2017 年第 65 期

2017.05.31

执笔：宋丹 黄旭

dan.song@icbc.com.cn

huangxu@icbc.com.cn

社交大数据在金融行业的应用 现状与启示

要点

- 随着社交网络成为互联网最大的流量入口，第三方支付、基金、小额贷款等金融业务大规模植入，并催生了社交金融这种新型互联网金融模式。社交金融除了将社交网络当作为其引入流量的业务渠道，同时也是用户行为、社交关系等社交数据的重要来源。社交数据目前已经在量化投资、银行零售、征信和电商等领域取得了成功应用。
- 互联网金融企业一直是社交网络商业化进程中最大的受益者，借助强大的科技实力，其对社交数据价值的挖掘和变现可谓占尽先机。通过对互联网金融企业相关案例的分析可以得到启示：有效的应用策略、广泛的数据来源和强大的数据分析能力是企业成功应用社交数据的关键。
- 面对互联网金融的冲击和自身转型的需求，传统金融企业也在试图搭乘社交网络的“快车”。在传统金融企业数据资源趋同的形势下，社交数据的引入可为业务能力提升和转型提供更多维度的帮助。传统金融企业可充分借鉴互联网企业的成功经验，在开拓社交数据积累和应用渠道、提升非结构化数据处理能力等方面加大投入。

重要声明：本报告中的原始数据来源于官方统计机构和市场研究机构已公开的资料，但不保证所载信息的准确性和完整性。本报告不代表研究人员所在机构的观点和意见，不构成对阅读者的任何投资建议。本报告（含标识和宣传语）的版权为中国工商银行城市金融研究所所有，仅供内部参阅，未经作者书面许可，任何机构和个人不得以任何形式翻版、复制、刊登、上网、引用或向其他人分发。

社交大数据在金融行业的应用现状与启示

随着以 Facebook 和微信为代表的社交网络平台迅速崛起，社交网络已经成为互联网最大的流量入口，第三方支付、基金、众筹和 O2O 等金融产品的植入，使得社交网络逐渐带有金融属性，并催生了社交金融的产生。社交金融这种以高频社交应用带动中低频金融活动的商业模式除了将社交网络平台用于业务部署外，也将其作为收集社交数据的重要渠道。社交数据作为大数据的重要组成部分，包含了海量的客户行为、社交话题和网络关系。通过对社交数据的挖掘与分析，从中获取用户的行为偏好、意见诉求和心理活动，弥补了金融活动过程中只保留结果数据而缺少过程数据的不足，并将用户的社交属性与金融机构的信用属性、流动属性、风险规避属性和杠杆属性相匹配，能够促使社交金融不断进化。

一、社交数据在金融行业的应用现状

互联网金融企业¹是将社交网络与金融领域结合的先行者，通过其先天的科技优势在社交数据分析与应用场景创新方面不断地推陈出新。而很多传统金融企业在互联网金融时代面临外部冲击、渠道抢夺和优质客户流失时，也纷纷依托社交网络向互联网金融转型。可以说，无论是互联网企业还是传统的金融企业都已看到在线社交网络带来的巨大红利，并将在线社交作为企业发展战略中的一步大棋。社交数据也因此引起了金融企业的重视，在量化投资、零售、征信、电商等领域都有着广泛的应用。

（一）应用领域

1. 基于社交数据的量化投资。股票投资实质上是将影响股价的因子不断量化的过程。上世纪 70 年代以前，股票投资仅仅是一种定性的分析，而没有数据方面应用。随着计算机的普及，很多人开始研究驱动股价变化的规律，把传统基本面研究方法用模型代替，随着计算机信息处理能力的提高，金融工具的丰

¹ 本文中互联网金融企业特指涉足金融业的互联网企业，不包括通过互联网发展业务的传统金融企业。



富及交易成本的降低，量化投资由此快速兴起。

随着投资机构在量化投资领域的投入，量化投资的同质化竞争愈演愈烈，各家机构的量化模型越来越趋同，导致投资结果同涨同跌，量化投资专家们希望通过更大规模的数据来寻找规律，于是席勒理论的第三层变量——市场情绪开始进入到量化投资模型之中。传统的市场情绪量化主要通过计算机分析新闻、研究报告和搜索行为等，借助自然语言处理方法，提取并量化市场情绪信息，而社交媒体作为重要的用户情绪发布平台，其数据的引入极大丰富了市场情绪信息的来源，通过监测市场舆论和用户行为，能够对投资做出秒速判断。

印第安纳大学计算机科学家 Johan Bollen 在其研究报告中指出，海量 Twitter 信息中的情绪状态指标与道琼斯工业标准指数之间存在联动关系，国外新闻分析领导者 Raven Pack 公司通过对 Twitter 等社交网站的情绪数据采集、进行量化分析并发现投资市场波动趋势，他们发现大众不同的情绪将对股票市场产生不同的显著影响，而且关联性的准确度能够达到 80%。MarkerPsych 和路透社合作了 119 个国家的 18000 多个独立指数，如每分钟的心情状态，包括乐观、忧郁、快乐、恐惧和生气等，并将其作为全球股市投资的信号。

与欧美等成熟资本市场主要由理性机构投资者构成相比，东亚尤其是中国的股票类证券投资市场仍以散户为主，因此市场受投资者情绪和宏观政策性因素影响很大。而个人投资者行为可以更多地反映在互联网用户行为大数据上，所以社交数据对中国证券投资更有借鉴意义。例如腾讯自选股 APP 是以腾讯自选股大数据为基础，结合行为金融模型并通过量化手段分析互联网用户行为与二级市场股票价格表现之间的关联性，选取对股价波动具有较强解释能力的用户行为指标建立大数据量化投资策略。

总而言之，利用社交数据量化投资，比传统量化投资构建的模型更加精准的原因在于，社交数据在原有的结构化金融数据基础上增加了非结构化的社交情感数据，通过非结构化数据分析方法挖掘市场情绪，并将其量化成为投资选股策略。

2. 基于社交数据的银行零售。BCG 在其《完美零售银行 2020》中阐述：过

去客户获取信息、比价、购买和互动等行为都非常依赖银行的网点及其员工，但时至今日，银行客户有更多元化的渠道来获取信息、比价、购买和互动，可以通过互联网、智能终端、社交媒体和朋友圈等来实现。美国著名的互联网市场调研公司 ComScore 发布的在线和移动银行报告显示，社会化媒体正在成为银行新的展示渠道，社交银行已日趋成熟。

数据显示，2016 年，访问银行 Facebook 主页的用户数量增长近 25%，而 Twitter、Linkedin 的访问数量也有一定程度的增长。美国的富国和美国运通等银行很早就开始推广和运营社交网络，并不断加大在社交网络的品牌影响力和部署各种营销活动。早在 2010 年，富国银行就在 Facebook 推出了“小型星期六”的商业活动。该活动旨在鼓励消费者在本地商店里多消费，推出 3 个星期增加了 100 多万的粉丝。美国运通银行也认识到社交网络的营销潜力，不断加大在 Facebook、Twitter、Youtube 上的投入，拓展了新的沟通渠道。此外，美国运通银行还针对小企业建立了专属社交平台，为小企业主讨论共同遇到的问题提供平台，而且还建立了“社员计划”，即为持卡人创建独特的在线社区，包括商家数据库、智能产品资源库等，利用运通银行系统中的数据向持卡人提供相关服务。

我国作为社交网络用户最多的国家之一，社交网络平台已经逐渐成为各大商业银行推广和运营的重要渠道。目前我国大多数银行基于微博和微信公众平台影响力、用户粘性高的特点，纷纷接入第三方社交平台，并通过开展品牌宣传、在线社交和举办营销活动等方式增加自身获客能力。

3. 基于社交数据的征信。根据调查，80%左右的信贷风险来自信贷审批环节，一旦消费者获得贷款，后续的管理只能控制 20%的风险，由此可见科学的信贷审批管理十分重要。开发高质量的信贷审批评分模型，进行科学的审批风险管理，可以大幅降低坏账率，并取得比较好的经济效益。

在美国，传统的征信评分方法（如 FICO）主要服务于 85%有信贷记录的客群，该方法从银行客户数据中抽取 15-50 个变量，通过 logistic 回归方法建立征信模型。传统方法将每个人初始分值定为 850 分，信用评分模型利用征信数据



从多个评分因素考察消费者的信用风险，从 850 分中减分。其中，美国个人消费者的平均 FICO 评分为 678，还有 15% 人群远低于平均的 678 分，根据 FICO 的标准，如果这部分人未能如期还款，或者缺乏借贷经历，无论是否事出有因（比如遇到了医疗紧急事故，或者最近刚刚移民美国），他们就会自动被视为风险人士，他们的贷款也就会被惩罚性地给以更高的利率，或者贷款申请被拒。

在大数据时代，用户的信息变得更加多样化，尤其是在线社交数据的引入极大丰富了征信模型的变量。美国 Zestfinance 公司以大数据技术为基础采集多源数据，一方面继承了传统征信体系的决策变量，重视深度挖掘授信对象的信贷历史；另一方面，将能够影响用户信贷水平的其他因素也考虑在内，如社交网络信息、用户申请信息等。ZestFinance 的数据来源十分丰富，依赖于结构化数据的同时也引入了大量的非结构化数据。另外，它还包括大量的非传统数据，如借款人的房租缴纳记录、典当行记录、网络数据信息等，甚至将借款人填写表格时使用大小写的习惯、在线提交申请之前是否阅读文字说明等边缘信息作为信用评价的考量因素。可以说 ZestFinance 开创了通过非常规数据（社交数据、非结构化数据）进行信用评估的先河，通过大数据手段覆盖了传统的信用评估服务所无法覆盖人群，特别是弱势群体。

随着社交网络中用户信息价值越来越大，许多美国贷款公司开始深入挖掘 Facebook、Twitter 等社交媒体的信息，以此来决定客户的信用等级。例如部分贷款公司会关注申请人递交的工作信息是否与 LinkedIn 上一致，申请人是否在 Facebook 上分享过工作、生活和被解雇的经历等等。旧金山的新兴贷款公司 LendUp 采用信用机构和社交网络的综合数据来对借款人进行综合评判。借款申请者可以自愿分享 Facebook、Twitter 等社交信息，他们提供的数据越多，获得贷款的可能性就越大。而在墨西哥、哥伦比亚和菲律宾等国家同样已经有公司通过分析用户的 Facebook、LinkedIn 和 Twitter 账户来评估他们的信用情况。2015 年 8 月，Facebook 在美国申请专利，通过其好友关系来评估个人信用。当用户申请贷款时，贷款方会审查该用户社交网络好友的信用等级。只有这些好友的平均信用等级达到了最低的信用分要求，贷款方才会继续处理贷款申请，否则

该申请即被拒绝。Facebook 能够成功地申请专利说明了在线社交网络对于线下人脉和社交关系的影响，充分体现了“物以类聚，人以群分”的思想理念。

在我国，一些互联网金融公司同样应用社交关系和社交行为建立了客户信用评分体系，例如，阿里巴巴的芝麻信用依据个人网上购买记录和好友的购买记录来建立客户信用评分；基于腾讯信用的微粒贷依靠对 QQ 用户的说说、日志等文本信息进行挖掘，并结合操作行为和 LBS 地理位置信息综合构建客户画像，评定客户信用得分；2016 年 7 月 15 日立木征信上线，该征信系统专门服务于 P2P 网贷，应用数据涉及 13 大数据源，涵盖电商、运营商、央行征信和重要的社交媒体，其中包括脉脉、LinkedIn 等，能够精准地为网贷企业筛选真实有效的高质量借款人，降低整体风险。

部分信贷评级业务机构表示，虽然目前基于社交数据的征信还只是服务于从事小规模贷款的新兴贷款公司，但这种征信方式很有可能成为主流。

4. 基于社交数据的电商。社交网络用户的激增、移动支付便捷性的提升以及社交网络为企业和个人提供的低成本营销平台，都加速了电子商务与社交网络的结合。国外的咨询机构曾做过调查，87%的电商企业使用社交渠道，79%用来提升品牌认知，74%用来获取客户。

当前社交电商在国内发展趋势如火如荼，主要分为社交网站独立发展电商、电商借助社交平台以及电商与社交平台共同合作的三种模式。传统的电商绝大多数的数据来自消费者的消费信息，而很难获取社交平台上所拥有的用户兴趣和热点话题等数据。而社交电商可以通过挖掘电商客户的消费数据及消费兴趣，并结合社交网络平台的用户社交数据，分析出用户潜在的消费动机，让社交电商能为客户进行更加精准和智能化的推荐。社交电商最大的优势在于它共享并整合了散落在社交网络、电商网站和物联网等网络平台中海量的用户信息、产品偏好、消费信息和行为习惯的数据，并能够使这些数据价值最大化。

（二）典型案例

1. 百度股市通。百度股市通是百度公司开发的一款手机股票软件。该软件定位为股民选股的辅助工具，弥补了市面上股票软件在消息实时性、全面性、



关联性和智能性上的不足。它的核心思路是利用大数据手段，聚合有价值的新闻信息、用户言论和热搜数据，寻找股票与信息以及股票与股票之间的关联，并通过整合第三方金融数据，给股民最有价值的投资信息。

下图为百度股市通 APP 界面。



图 1 百度股市通 APP 功能界面

资料来源：百度股市通 APP 截图

其中，红色方框的部分表示百度股市通具备个股舆情指数（舆情利好、舆情利空排名）、概念搜索热度和用户对个股关注程度的功能。百度股市通之所以能够生成独家热搜指数，准确预判热点的趋势发展，很重要的原因是该 APP 除了运用传统的结构化数据进行趋势预测之外，还运用了来自百度搜索、股吧、百度贴吧、雪球、微博等搜索引擎、投资社区和社交网络的非结构化数据和社交网络数据。这些数据包括了股票发行机构的舆情以及大量的投资者评论、意见和情感，对于挖掘投资者情绪和市场信息、分析消息对股票的影响趋势、向投资者进行个股推荐发挥了巨大的作用。

百度股市通主要将社交数据用于个股分析和信息推荐，其核心是建立了一个集数据采集、数据分析、资源整合与数据应用为一体的知识图谱体系（见图 2）。

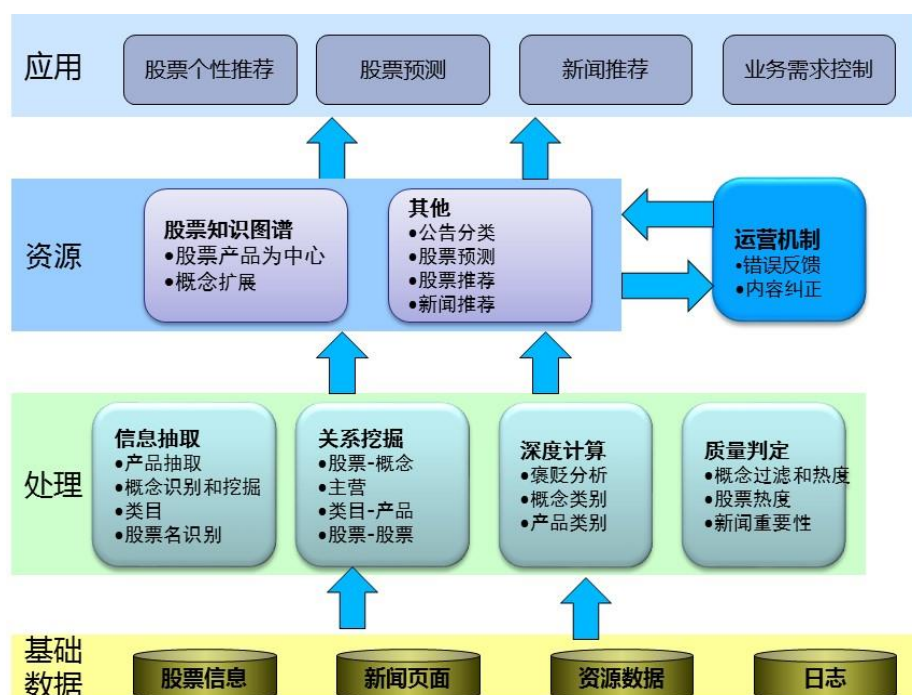


图 2 百度股市通知识图谱挖掘体系

资料来源：百度技术报告

知识图谱是百度股市通的“大脑”，依托百度服务器每秒处理上亿比特数据，能够将用户热搜概念和产品与股票进行关联，同时也将行业信息与相关股票进行关联，挖掘热点信息与个股之间的关系，并分析哪些热点事件能够影响个股价格。其任务主要包括如下几个方面：

(1) 概念的分类与识别：互联网上存在着大量的与个股有关的信息，如何量化出这些信息所蕴含的影响以及投资者的情绪，并将此信息与相应股票进行匹配是分析事件命中股的关键。其中很重要的步骤就是对消息中的概念进行识别，以建立消息与股票之间的关联关系。挖掘概念的消息来源来自于百度搜索信息、股吧、新闻信息和其他数据源，通过文本挖掘的方法识别出概念实体，并将概念分为产品服务类、政策地域类和其他类三个类别，类似于股票概念板块。

(2) 产品经营关系挖掘：“产品经营”专指股票所在企业直接生产的产品及衍生产品，并将其分为主营和非主营产品。通过产品经营关系挖掘可以将产品信息与股票进行匹配，类似于股票行业板块。



(3) 概念热度计算：根据股吧帖子和百度搜索中的概念信息出现次数（如一带一路、美丽中国等）计算当前概念的热度，推荐热门概念给用户。

(4) 基于语义的关系计算：计算概念与产品之间的紧密度、概念与股票之间的紧密度，并通过紧密度打分机制建立概念与产品、产品与股票之间的匹配关系。

2. 腾讯信用。目前，国内金融机构评估企业和个人信用情况，主要使用央行征信系统的数据。但这一系统的缺陷在于覆盖面不广，仅覆盖 3 亿有征信记录人群，大量农民工、个体户和学生未被纳入，金融机构要想了解这些人群的信用记录，成本高昂，因此往往会忽略其金融需求。2015 年 1 月，央行发文要求腾讯信用、芝麻信用等八家机构在六个月内做好个人征信业务准备工作，建立健全社会征信体系。

目前，腾讯推出了个人征信管理平台——腾讯信用，已经过央行的正式验收，并获得央行颁发的牌照，这也标志着我国开设民间市场化征信机构已走向实践阶段。腾讯信用主要通过“履约、安全、财富、消费、社交”五大指数，基于用户的历史行为，通过大数据分析方法为用户信用评分，评分结果介于 300~850 之间。在数据选择方面，腾讯选用了 QQ 用户的社交数据作为信用评分依据之一，腾讯 QQ 用户超过 10 亿，月活跃客户约 8.6 亿，QQ 空间月活跃用户 6.53 亿，最高同时在线客户 2.39 亿，能够覆盖我国大量无征信记录的“长尾”用户，因此，腾讯 QQ 社交数据比较适合用于评估“长尾”用户的征信情况。腾讯 QQ 数据包含多个维度，资源十分丰富，下表列出了腾讯 QQ 主要的数据类型。

表 1 腾讯 QQ 数据类型

| 数据类别 | 数据细分 |
|-------|-------------------|
| 人口属性 | 年龄、性别、地域、家乡 |
| 关系链 | QQ 群、QQ 关系链 |
| 社交&音乐 | 说说、相册、QQ 音乐 |
| 移动互联网 | LBS、手机 APP、移动设备 |
| 游戏 | 端游、页游、手游 |
| 增值业务 | QQ 会员、Q 币、黄钻、QQ 秀 |

腾讯凭借 QQ 在人群覆盖面广、用户活跃度高等显著优势，依托社交、支付、

金融和社会等多维度数据综合评估，通过海量数据挖掘和分析技术计算用户行为属性，构建用户信用画像，以次来预测用户的风险表现和信用价值，为其建立个人信用评分。

腾讯信用用户画像系统架构如下图所示：



图 3 腾讯信用用户画像系统架构

资料来源：《腾讯：社交数据在征信领域的应用探索》

2015 年 5 月 15 日，腾讯投资的微众银行上线第一款产品——“微粒贷”，该产品基于腾讯信用筛选出预授信客户，并通过 QQ 钱包和微信两个渠道主动向目标客户推送。截至 2016 年 11 月底，“微粒贷”预授信客户数约 5000 万户，累计发放贷款总金额超 1600 亿元，总笔数超 2000 万。

二、社交数据在互联网金融企业的应用启示

1. 企业在构建社交数据分析与应用体系时应将其整合在一般意义的大数据体系中，形成了统一框架、多种方法和不同场景的应用策略。社交数据是大数据的重要分支，二者在数据分析与应用的方法论方面趋同。因此，社交数据分析与应用体系不应独立于企业内部已有的大数据分析与应用体系，而应将二者进行有效整合，形成统一的分析与应用框架。同时针对社交数据类型复杂的特点，应采用不同的数据处理方法，并构建不同的应用场景。



2.社交数据分析与应用需要庞大的数据规模、广泛的数据来源和多样化的数据类型，其中非结构化数据的分析和应用占据了十分重要的地位。2016 年全球 67.4%的网民使用社交网络，Facebook 一分钟产出 350GB 的数据，Twitter 一分钟内用户总共发布 27.8 万条博文，LinkedIn 一分钟内有 11000 次专业搜索。社交网络用户量和数据量的几何增长带来了巨大的商业价值，舍恩伯格在《大数据时代》一书中指出，大数据的简单算法比小数据的复杂算法更有效，拥有大规模复杂数据所能带来的商业利益远远超过少量精确性数据。从百度和腾讯等互联网金融企业的成功案例来看，庞大的社交数据规模是其精准分析和洞察用户需求的基础。这种数据基础首先得益于企业早期构建的腾讯 QQ 和百度贴吧等社交平台，该平台覆盖了海量的用户群体并且具有较高的活跃度，所产生的大规模数据样本能够反映的用户行为、需求和情绪。

除数据规模外，数据来源与数据类型是衡量社交数据价值的另一种方式。数据类型的“混杂性”为应用社交数据预测用户行为提供了更多维度的观测变量。在国内，互联网行业在大数据的积累和应用以百度、腾讯和阿里巴巴最为值得关注，下表以这三家公司为例对其互联网数据进行梳理，以此来说明数据来源和数据类型的重要性。

表 2 百度、阿里和腾讯三家公司的数据来源和数据类型

| 企业名称 数据类型 | 百度 | 阿里巴巴 | 腾讯 |
|--------------|--------------|-----------|------------------|
| 电商数据 | 百度 MALL、百度糯米 | 淘宝、天猫 | 拍拍、京东 |
| 支付数据 | 百度钱包 | 支付宝 | 财付通、微信 |
| 社交数据 | 贴吧 | 旺信、来往、支付宝 | QQ、微信 |
| 新闻资讯数据 | 百度新闻 | 聚宝头条 | 腾讯新闻 |
| 视频数据 | 爱奇艺、百度视频 | 优酷 | 腾讯视频 |
| 浏览器数据 | 百度浏览器 | 淘宝浏览器 | QQ 浏览器、 搜狗浏览器 |
| 搜索数据 | 百度搜索 | 一淘 | 搜狗搜索、SOSO |
| 游戏数据 | 百度游戏 | 阿里游戏 | 腾讯游戏 |
| 音乐数据 | 百度音乐 | 虾米音乐网 | QQ 音乐 |
| 旅游数据 | 百度旅游、去哪儿 | 穷游网 | 携程网 |
| 地图数据 | 百度地图 | 高德地图 | 腾讯地图 |

| | | | |
|--------|------|------|--------|
| 餐饮数据 | 百度外卖 | 口碑 | 大众点评 |
| 人机交互数据 | 度秘 | 我的小蜜 | QQ 机器人 |

从原生业务角度来说，百度侧重搜索、阿里巴巴侧重电商、腾讯侧重社交，三者业务倾向性不同，积累的原生数据也各有侧重。但是这三家企业经过长期业务整合、并购重组不断拓展业务体系，其数据积累也逐渐趋同。由表 2 可以看出，百度、阿里巴巴和腾讯三家企业数据类型已经涵盖电商、支付、社交、新闻资讯、视频、浏览器、搜索、游戏、音乐、旅游、地图、餐饮和人机交互等方面。从广义的数据类型来说，电商、交友、视频（包含弹幕）、搜索、餐饮（包含评论）和人机交互数据都可定义为社交数据，通过分析不同类型的社交数据挖掘社会热点、舆情和用户行为，并将其应用于互联网金融产品。表 3 总结了百度、阿里巴巴和腾讯三家公司的主要金融产品以及社交数据在该产品中的应用。

表 3 百度、阿里和腾讯金融产品及社交数据应用策略

| 企业名称 | 产品名称 | 社交数据应用策略 |
|------|-------|--|
| 百度 | 百度股市通 | 通过百度热搜、网络舆情等数据构建股票知识网络，向用户推荐事件命中股 |
| | 百度理财 | 百发 100 指数以百度海量搜索数据为基础，可评估股票涵盖 2500 多只 A 股股票 |
| 阿里巴巴 | 蚂蚁聚宝 | 通过用户的购买、支付以及对聚宝头条的点击行为，推出“资产+资讯+产品+社区”的解决方案，把用户最关心的内容放在最重要的“首页”位置，降低了大众用户的理财难度 |
| | 芝麻信用 | 通过淘宝依赖程度、支付情况和支付宝好友信用情况等数据为支付宝用户进行信用评分 |
| 腾讯 | 腾讯信用 | 通过 QQ 相册评论、说说、日志、群聊和好友关系网络等数据为腾讯用户进行信用评分 |

从表 3 可以看出，投资理财、资讯推送和征信是百度、阿里巴巴和腾讯在金融行业取得成功应用的具体领域。从其应用策略可以看出，三家企业在构建金融产品时应用了大量的非结构化数据，这也是与传统金融机构在构建金融产品相区别的地方。因此，在传统金融企业也应当建立更加多元化的数据收集渠道，并且增强非结构化数据分析应用的能力。

3.社交数据的分析能力取决于大数据平台的建设。从数据特征来看，社交数



据是大数据的一个重要分支，具有典型的 4V 特征²，而 4V 特征也是导致大数据存储、挖掘和应用困难的重要因素。在数据存储环节，存储结构化数据的关系型数据仓库已不再适用于类型复杂的社交数据；在数据处理和分析阶段，面对庞大的社交数据模庞，如何结合云计算、分布式计算和流式数据处理方法挖掘数据并且相对实时地得到分析结论也是社交数据挖掘过程中不可避免的问题；在数据应用阶段，企业内不同部门之间如何消除数据壁垒，也是企业大数据应用落地的重要保障，由此可见建设海量存储能力、高效运算速度以及各部门互联互通的大数据平台是社交大数据挖掘的必备条件。

除了海量的存储能力、高效的运算速度以及在企业内部互联互通之外，大数据平台还应具备易用性、容错性和稳定性。Facebook 是全球最大的社交网络平台，分析和研究 Facebook 的大数据体系架构对企业存储、分析和应用社交网络数据具有一定的借鉴意义。一份资料显示，早在 2011 年 Facebook 拥有的压缩数据已经达到 25PB，未压缩数据 150PB，每日有 25 亿条新信息发布，45 亿个点赞数量、3 亿照片上传数量、日均产生数据规模 500TB 以上。为存储和处理规模如此庞大的数据，Facebook 先后在美国、瑞典、爱尔兰和丹麦建立了 5 个数据中心，建设了全球最大的分布式文件系统，至少拥有 6 万台服务器（2010 年数据），单个集群中的数据存储量就超过 100PB。而且在 Facebook 不同部门之间并无设立数据壁垒，数据应用者可以跨部门获得数据。在数据处理系统建设方面，Facebook 的主要设计目标是秒级的延迟，每秒钟能够处理上百 GB 的数据。由于单一的编程语言无法满足所有需求，因此 Facebook 先后开发了 Puma、Stylus 和 Swift 三个不同的流式数据处理系统。同时建立了一个持久化消息总线将所有的处理组件连接起来进行数据传输，并对数据的处理和传输解耦，实现容错、可伸缩、易用性和正确性。

三、传统金融企业分析和应用社交数据的策略与建议

² 大数据 4V 特征包括：数据规模体量巨大（Volume）、数据类型繁多（Variety）、数据产生速度快（Velocity）、数据价值密度低（Value）

用户行为模式的改变、在线社交数据规模的爆发式增长以及企业级数据分析平台的完善加速了社交数据的商业化应用进程，无论是互联网金融企业还是传统金融机构都在试图通过社交数据分析用户的行为、动机、意愿和情绪，为用户量身定制金融产品与服务。互联网金融企业由于其先天的技术优势在大数据分析和社交数据应用方面已占得先机，而传统金融企业面临互联网金融企业的冲击、业务渠道的转变、优质客户的流失以及数据资源的趋同，同样希望应用社交数据为业务能力提升和企业转型提供助力。以互联网金融企业社交数据应用案例和启示为基础，对传统金融企业提出如下建议：

1. 开拓社交数据积累渠道，为传统金融机构社交数据分析应用提供广泛的数据来源。当前，传统金融机构发展社交金融主要有两种方式：一是借助微博、微信等社交平台发展社交金融，例如微信银行等。这种方式最为简单有效，依托高流量的社交平台带动自身业务发展，但最大的缺点在于社交数据掌握在第三方社交平台，不受金融企业自身的掌控，对社交数据的获取主要以合作和购买等方式，需要花费较高的数据成本。二是企业自有渠道的建设，部分科技实力较强的金融机构为摆脱对第三方社交平台的依赖，自己筹建了垂直金融领域的社交平台，例如工银融e联和平安天下通。其优点是企业能够摆脱对第三方社交平台的依赖，增强企业客户和数据资源的安全性，并且在提升平台自身用户规模和活跃度的同时能够及时获取到用户的社交关系、言论、行为等数据，但也面临着社交网络开发、管理和运营所需投入较大的问题。因此，无论是借助第三方平台还是建设自有渠道，数据获取都是传统金融企业分析和应用社交数据过程中面临的最重要的问题。

2. 提升非结构化数据处理能力。在社交数据中，大量数据是以非结构化的形态存在，文本、图像、语音和视频都是其重要组成部分。从百度股市通、腾讯信用等成功案例可以看出，非结构化数据（尤其是文本数据）的应用在企业大数据体系中的地位愈发重要，在挖掘社会舆情和用户偏好过程中发挥了重要的作用。非结构化数据在信息表达方面比结构化数据更加直观，含有的信息更加丰富，但分析难度更大。首先，非结构化数据的结构复杂，各类分析方法层



出不穷，需要针对不同的实际问题配置不同的分析方法；其次，非结构化数据的分析过程是一个信息损失的过程，例如将文本分割成词汇的过程会损失词汇的上下文关系和语义关系，如何尽量减少信息损失也是一个关键的问题；再次，非结构化数据中无效数据较多，从非结构化数据中分析数据价值往往需要海量的数据，这对于企业数据分析系统的性能带来极大的挑战。目前，传统金融机构在数据存储和分析过程中虽然仍以客户信息、交易等结构化数据为主，但作为企业大数据的重要补充形式，非结构化数据的来源随着信息源的增加而愈发广泛，非结构化数据在企业中的作用和应用价值也愈发重要。

3. 强化大数据平台的处理硬件和软件资源。当前，在线社交数据已经成为互联网最大的数据来源，是大数据的重要组成部分，与一般意义的大数据相比，具有数据规模更大（全球约 23.4 亿人访问社交网络，占全球总人口的 32%）、数据类型更加复杂、数据产生频率更快以及数据价值更加稀疏的特点，因此分析和应用社交数据必须依托更加强大的存储和分析系统。通过数据质量和数据标准管控不断丰富基础数据的来源、扩展处理数据的类型，将社交数据逐步整合纳入大数据基础平台，并依托云服务提供集成型的数据服务和各类分析挖掘工具，实现业务数据的集成与共享，以满足不同时效性的分析需求。

4. 使用多元化的数据分析方法。传统的数据分析方法建立在关系数据模型之上，主要用于分析范围已知并且容易理解的结构化数据，例如统计分析、回归分析、聚类细分和其他一些简单的机器学习方法。对于社交数据而言，复杂的数据类型是其重要特征之一，社交数据既包括结构化数据，也包含海量的非结构化数据（文本、图像、音频和视频等），而且数据之间往往又构成了十分复杂的网络结构。因此，传统的关系型数据模型已不再适用于社交数据的分析，建立多元化的数据分析方法是十分必要的。当前，已经有一些新型数据分析方法在商业领域得到成功应用，社交数据分析所涵盖的新型分析方法主要包含以下几个方面：第一，路径分析。路径分析是一种研究多个变量之间多层因果关系及其相关强度的方法，其目的在于从数据中寻找符合需求的路径并最终找到符合特征的用户，在用户社交影响力分析和精准营销方面有着十分广泛的应用。

第二，文本分析。在社交网络数据中有超过 80% 的数据都是非结构化文本数据，应用文本分析方法挖掘社交网络文本中的语义和模式，从中发现营销机会和用户面临的问题有重要意义。第三，图挖掘。图是最常用的数据结构之一，以描述事物之间错综复杂的关系，对于社交网络中的社区和用户集群的发现都有着重要的作用。第四，大数据可视化。数据可视化是大数据分析过程中一项重要的数据分析辅助手段，在企业智能化运营过程中发挥着重要作用，对于社交数据而言，数据可视化工作尤其重要。

5. 重视社交数据的隐私安全及使用规范。在大数据技术发展过程中，信息安全一直是极为关注的问题，这一点对于社交数据更加重要。社交数据中所包含的用户社交关系网络、用户之间的社交话题以及一些互动行为都属于相对隐私的用户数据。例如，腾讯信用在应用 QQ 用户的社交数据评价用户信用情况时，特意避开了比较敏感的 QQ 用户聊天记录，而是采用了群组标签、QQ 日志、说说等敏感程度相对较低的数据。而且，随着移动地理位置（LBS）信息应用日渐风靡，由地理位置所导致的个人信息资源泄露的问题也时有发生。因此，社交数据在应用之前有必要做好权限控制和数据脱敏工作，建立文本数据应用规范。