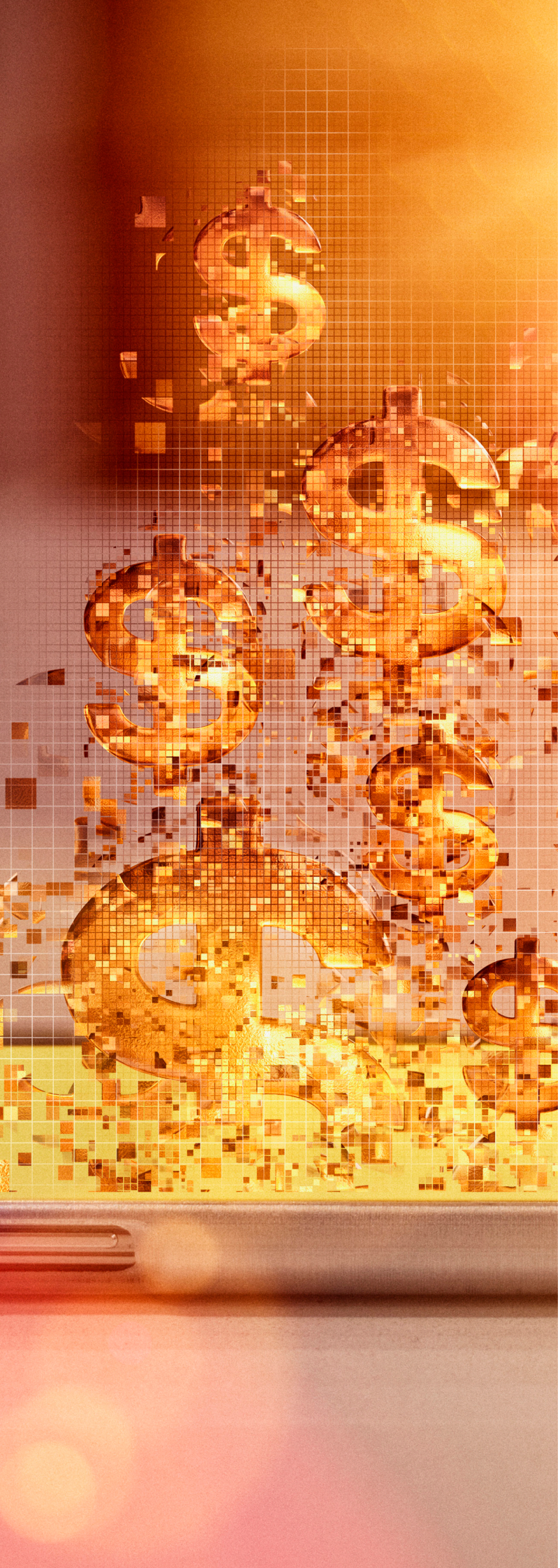


专家洞察

匠心独运 厚积薄发

详解银行非结构化
文本数据背后的价值

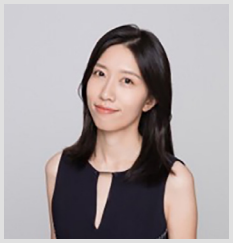
IBM 商业价值研究院

IBM

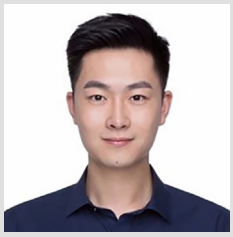
主题专家

吴大维
IBM GBS CBDS 团队
副合伙人
wudavid@cn.ibm.com

杨杭
IBM GBS CBDS 团队
人工智能解决方案负责人
首席业务咨询顾问
bjyhang@cn.ibm.com



郝希蓓
IBM GBS CBDS 团队
资深数据科学家
haoxish@cn.ibm.com



高康睿
IBM GBS CBDS 团队
高级数据科学家
gaokrui@cn.ibm.com

何佳惠
IBM GBS CBDS 团队
高级数据科学家
hjhjhui@cn.ibm.com

宋鹤
IBM GBS CBDS 团队
高级数据科学家
songhe@cn.ibm.com

王莉
IBM 商业价值研究院
高级咨询经理
gbswangl@cn.ibm.com

扫码关注 IBM 商业价值研究院

官网　微博　微信公众号　微信小程序

谈话要点

掘金非结构化数据

面对银行内非结构化数据体量大、增速快、利用不足的现状，建议深挖非结构化数据，加速释放大数据价值。

巧用 NLP（自然语言处理）技术

形式多样的中文表述需要灵活运用各种 NLP 技术与创新性思维。

业务应用为本

基于银行内常见的非结构化数据来源，探索非结构化数据在银行业的应用场景。

—

深挖非结构化数据宝藏助力银行精细化运营

IDC 预测，全球数据圈将从 2018 年的 33ZB 增至 2025 年的 175ZB。在全球数据圈扩张进程中，中国数据圈增速最快，预计到 2025 年将增长至 48.6ZB，占全球数据圈的 27.8%，将成为全球最大的数据圈。[1] 而在这些数据中，结构化数据仅占 20%，其余 80% 都是以文件、语音等形式存在的非结构化和半结构化数据，这些非结构化数据平均每年增加约 60%。[2]

鉴于当前各大银行对于结构化数据的利用和挖掘已进入瓶颈和饱和状态，通过挖掘非结构化数据来为营销或运营等场景提供智能化的决策支持已成为银行数字化转型过程中的热门话题。

非结构化数据的价值主要体现在以下三个方面：

数据量大：非结构化数据已经占到大部分银行信息总量的 80% 甚至更高，是银行非常宝贵的数据资产。[3]

产生速度快：短时间可产生大量数据，如某股份制商业银行日产生录音文件量在 200G 以上。[4]

数据来源丰富：既包括银行内部数据源，如客服、邮件等，也包括外部数据源，如社交媒体、企业财报等。

同时，无法回避的是，非结构化数据的分析技术难度大、对存储运算要求高、应用难度较大：

技术难度大：相比于结构化数据的机器学习算法，非结构化数据的各种自然语言处理技术和深度学习技术在模型复杂度和方案成熟度等方面的技术难度都更大。

数据量攀升对存储运算要求高：对于大部分银行来说，非结构化数据的存储并不统一，且增长迅速，数量巨大，因此对非结构化数据的存储、治理和分析挖掘的运算能力都提出了更高要求。

应用于实际业务的难度较大：非结构化数据包罗万象，纷繁复杂，如何基于业务理解，聚焦业务场景，充分挖掘数据价值，赋能业务实际，仍是非结构化数据分析应用的一大难点。

IBM 基于多年项目经验，建议银行业可以从以下四类非结构化数据入手，探索非结构化数据的应用：

- **客户沟通数据**：不管是呼叫中心的语音，还是线上渠道的聊天文本，银行内沉淀了大量包含客户身份信息、偏好选择、服务投诉、业务咨询的沟通数据，可应用于客户营销与优化服务运营。
- **银行内部工单数据**：通过对工单的自动化分类及自动化摘要，减轻内部人工压力，提升运营效率；同时基于对工单内容的预警监控，可有效预防投诉风险或群体性事件发生。
- **商户数据**：针对商户名称、地址和主要业务的解析，可加深前台业务人员对商圈及潜在商户的了解；而分析行内对公企业的交易流水文本，不仅可获得客户的部分资金分配及使用信息，亦可丰富商户的上下游供应链信息。
- **外部舆情数据**：行业报告、券商研报、公司财报 / 公告、社交媒体信息等非结构数据能有效补充银行内部持有的企业局部资金数据，为银行有效评估企业的价值与风险提供更为全面的视图。

海量用户之声信息中隐藏的关键信息

当银行为应当如何优化客户营销和服务运营而苦恼时，殊不知，答案已经存在于银行内，只是缺乏挖掘宝藏的方法。通过分析海量的用户之声信息，银行能够实现投诉升级的快速响应、潜在商机的挖掘和营销成功率的提升。

AI 助力投诉升级快速响应

日益增加的客户数量为人工坐席带来了越来越大的处理压力。客户需求量的增加导致了投诉数量激增和投诉问题的升级，甚至可能升级到外部银监会。为避免客户投诉升级到银监会，并在基层对问题进行合理处理，需开发一款对客户的投诉升级倾向敏锐的模型，及时提醒业务部门，在源头上采取缓解甚至断流的措施。

目前，银行间的普遍做法是利用结构化数据信息，包括客户工本日志，人工座机来电记录，过往投诉处理历史数据，客户在本行支配业务的多模块信息以及客户的基本信息进行归纳总结。

该种方法有效的前提是具有共同特点的群体用户对服务质量等要求苛刻。通过学历等指标能够一定程度发现高投诉倾向的群体。但客户的高投诉倾向亦可能是来自于对服务质量差的真实表达，或与坐席的交涉中对解决方案不满意，突然产生。同时，通过传统指标建模，响应时间较长，无法在有效时间内快速回应用户疑问，可能导致事态的进一步恶化。

针对此种问题，利用 NLP（自然语言处理）手段构建近实时预警模型，能够更好的解决传统指标无法解决的痛点。在呼叫中心电话端实时产生语音文本，出现问题时进行预警。在面对有高投诉倾向的客户时，可以第一时间派遣专业处理人员进行安抚。

然而，由于行方能够提供的投诉语音样本较少，常常导致无法建模。一是因为投诉升级本就是小概率事件，二是语音转文本普遍字数较多，而且由于口音、环境噪音等因素，文本往往质量较差，人工阅读较为费力。最终导致大量的行内投诉没有被明确标签化，淹没在海量语音数据中。利用多种 AI 技术手段，可以有效解决这些挑战，进行建模。

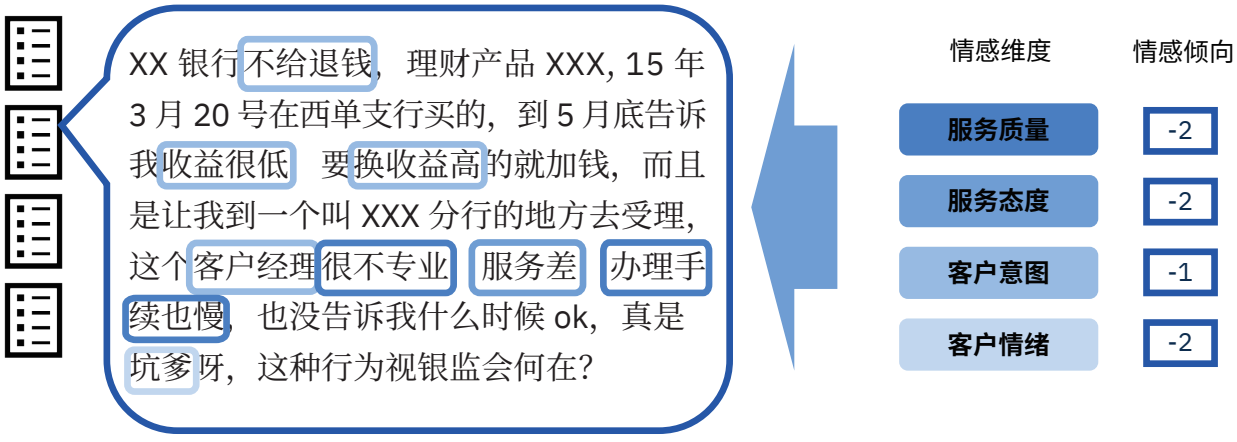
例如，在某大型国有银行，IBM 利用多种 AI 技术手段，从一千万通语音对话数据中挖掘出 7000 余条疑似投诉语音样本。这些样本因为各种原因（如被客服主观忽视）而没有上报。

IBM 将疑似投诉样本返回给业务部门进行快速核实确认。业务部门在一天之内完成确认，确认 5000 余条投诉样本、1500 余条非投诉样本以及 500 余条无法确认的疑似投诉样本。IBM 利用确认数据，形成建模条件。

该预警模型在广东省跨期验证期间精确率和召回率都达到了 85% 以上，针对有不满情绪的用户，快速响应，积极安抚，防止了多起银行内投诉事件升级到外部银监会（见图 1）。

图 1

AI 助力投诉升级快速响应



越来越多的银行用户倾向于询问机器人关于银行产品的相关问题。

机器人相关渠道中的潜在商机挖掘

目前，各大银行都通过机器人客服替代人工，对接官网、微信公众号等沟通渠道。银行客服中心大量运维人员需要维护机器人客服背后庞大的知识库，使机器人能够应对绝对大多数情况下用户的问题。但银行业务繁多、用户问题多种多样。客服中心工作人员维护知识库力不从心。

据某大型国有银行统计，知识库中知识数量级在一百万左右。机器人对客户问题的未回答率虽然只有约 1%，但相当于每月 25 万条，这对人工坐席造成非常大的压力，且严重影响了用户体验。同时，IBM 发现，越来越多的银行用户倾向于询问机器人关于银行产品的相关问题。此类询问往往含有巨大的商机，但由于机器人的技术限制，无法应对用户问题，导致错失商机。

在与某大型银行的合作中，IBM 利用多种 AI 技术手段，从海量未回答的用户问题中，挖掘热点话题，总结归纳，提出针对机器人的优化建议，将未回答问题从 25 万条 / 月降到 18 万条 / 月。

同时，IBM 发现，在机器人与用户的交互记录中隐藏着大量的潜在商机。如某用户询问是否有支付宝上某理财产品的相似产品，但机器人面对此类问题往往不能较好回应。个人金融类产品如贷款或理财，作为银行产品中非常重要的一部分，其营销一直是各家银行的重中之重。用户对于贷款等产品的需求往往时间较紧迫，错过与用户的沟通时机，可能直接失去一个潜在的拓客机会。

此外，IBM 在项目中，针对用户与机器人的交互文本数据，挖掘出近期存在个人快贷相关需求的潜在用户 1000 多名，提供给客服中心进行电话营销，并建立模型定期生成营销名单，利用线上多种渠道营销推介。

人工电话营销中交谈数据的价值与应用

呼叫中心作为银行在线上渠道服务客户的重要窗口，不仅承载着打造完美客户体验的责任，更成为银行实现精准营销的重要阵地。但现实中，电话营销一直存在两大痛点：一是营销成功率偏低，过度的致电推荐不适合的产品，会严重影响客户满意度；二是在以“劳动力密集型”为典型特征的呼叫中心，人工坐席的流失率高人员流动性大，导致服务质量不一致。

近些年，银行呼叫中心的规模与日俱增，语音数据体量巨大，是典型的非结构化“大数据”。这些数据内含客户身份信息、偏好选择、服务投诉、业务咨询等重要信息，是银行优化服务质量、提高运营效率的重要参考。如果可以充分挖掘数据价值，可以在一定程度上缓解以上两大难题。

针对营销成功率偏低的问题，不仅可利用语音数据中的关键信息优化外呼营销清单，还可通过分析营销成功的对话数据，提取打动客户的产品有效卖点，挖掘针对特定产品的最优的电话营销对话流程。该对话流程既可作为人工坐席的培训范本，使新入职的人工坐席快速上手，保障外呼质量，也可指导智能外呼机器人的外呼对话流程配置。此外，也建议定期提取全员对话流程，以确保人工坐席的外呼过程是符合业务要求的，实现对话流程的智能审核。

IBM 在某国有四大行外呼挽留对话流程挖掘项目中，通过分析 1.44 万通信用卡注销的挽留对话，提取出 50 多个客服及客户热点话题标签，识别出客户与人工坐席互动较多的三个话题（积分、年费、额度），以及基于不同注销原因的最佳挽留对话流程。同时，也发现了部分人工坐席存在核心对话环节缺失的问题（比如未进行身份核实）。该项目针对某外呼营销场景提供了 6 条优化策略，提高了外呼营销成功率，得到了业务部门的高度认可。

智能工单处理助力客户诉求快速预警检测

银行客服处每天会产生大量工单反映客户诉求。通过对工单进行预警监控，可有效预防投诉风险或群体性事件发生，减轻线上坐席压力。随着 AI 技术的不断成熟，工单的处理手段日益高效便捷（见图 2）。

无人工干预的 AI 减轻筛查压力

目前很多客服处仍采用人工查阅文本的方式对工单进行分类。一方面耗时耗力；另一方面，虽然工单分类体系相对来说比较完善，但仍有部分工单因无法放到某个有意义的类别中而被分到了 “其他”类别中，当这部分工单需要处理时仍需要人力二次细分，造成了资源的进一步占用。

IBM 通过 NLP 算法对工单文本进行语义分析，并结合机器学习算法进行无监督聚类，完美地解决了上述两个问题。一方面避免了大量人力参与分类过程，节约了人力资源；另一方面可以应对新类别工单产生后无法放入现有工单体系的困扰。

智能 AI 工单分类，大幅提升效率和准确性

某国有四大行信用卡部在进行投诉工单分类时发现，已有的工单体系不能及时、准确的覆盖来自全行的繁杂工单种类，致使业务人员不得不投入大量的时间去一篇篇点读、理解各篇工单的内容并进行处置。

由于工单种类每日均不相同，而且会随着时间生成新的问题工单，IBM 认为，此类问题应优先考虑技术难度较高的自动化动态归类设计思想。根据工单内容，结合外部公开数据，对工单内部非结构化的文本信息进行建模学习，并根据动态聚类的算法，实现高频问题工单及时预警和低频问题工单及时发现的功能。

该项目首先将 2020 年 1 月至 8 月累计的 6 万余篇非正常工单进行动态聚合。在无任何人工干预的情况下，成功将 6 万篇工单自动归类出 113 个子类，解决了大量历史遗留问题工单。同时，按照月度跑批流程，将 9、10 月非正常工单共计 6 千余篇自动化归类为 21 个子类别。

该项目大大降低了业务人员每日处理非正常工单的工作时间，同时也提高了业务人员处理非正常工单时的业务准确性。

图 2

工单处理手段的演进

借助摘要提取手段获取工单核心信息

工单文本分类后，客服人员需要了解每类工单反映的具体问题才能进行进一步处理。原工单文本长短不一，对于长文本工单，阅读花费时间较多，难以快速定位问题。为此，有银行尝试通过关联算法的方式对工单进行多元热点词组挖掘，但是词组展示的形式存在一定的局限性，需要较高的“想象力”将词语连接成句，不利于客服人员准确把握问题。

因此，用高效简洁的语句作为工单摘要将工单文本信息传达给客服人员，是一种较为友好的展现形式。工单摘要提取分为两种：抽取式和生成式。

抽取式：顾名思义，是一种直接从原文中选择若干条重要的句子，并对它们进行排序和重组而形成摘要的方法。这种方法目前已经比较成熟，有语法通顺、适应性广、速度快的优点，但缺点是灵活性较差。

生成式：是计算机通读原文后，在理解整篇文章意思的基础上，按自己的话生成流畅的翻译。这种方法灵活度高，并且伴随深度学习，生成式摘要的质量和流畅度都将有很大的提升，但目前也存在原文本长度过长、抽取内容不佳的限制。

这两种方法各有优劣（见表 1），银行可根据自身需求进行选择。

表 1

两种工单摘要提取方式对比

	优势	劣势
生成式	- 灵活度高 - 流畅性高	- 速度较慢 - 技术难度大 - 原文本过长时效果欠佳
抽取式	- 语法、句法错误率低 - 适应性广 - 速度快	- 抽取内容质量有限 - 灵活度低 - 连贯性差

智能工单摘要提取，准确获取核心信息

某国有四大行之一携手 IBM，开展了智能工单摘要提取项目。由于有些工单涉及内容过多且逻辑复杂，导致业务人员需要花费大量时间用来阅读。IBM 项目组利用工单摘要方案，结合业务需求，顺利完成工单摘要生成工作。

例如，某篇工单内容示例如下：“客户来电表示，从今年 1 月底开始，受国内疫情影响，客户所在城市进行了大规模的封城行动，并限制了客户所在小区的装修活动，导致小区装修受到了严重影响。客户从我行信用卡部申请的专项贷款因不可抗拒原因不能按照计划进行使用，现申请延长专项贷款额度有效期。由于客户所在小区被封禁，导致无法外出进行信用卡偿还而产生信用卡逾期，同时申请减免违约金及利息并删除客户个人征信记录。望我行尽快办理此业务。”

项目组通过丰富的 NLP 手段，将本篇工单提取为两句关键句：1：申请延长专项贷款额度有效期。2: 同时申请减免违约金及利息并删除客户个人征信记录。

通过关键句子摘要，降低了业务人员处理工单的时间，也让计算机帮助业务人员透过复杂的故事逻辑，直击用户的想法。该工单摘要提取解决方案在应用过程中得到了业务人员的一致好评。

多种 NLP 手段助力银行构建商户数字化经营体系

随着我国经济的持续快速发展，居民个人和家庭逐渐成为社会财富的主体，消费也成为我国经济增长的新动能。商户是串联金融生态系统的重要节点和对公对私业务交汇的关键点。商户业务覆盖面广、涉及行业多、价值链长，对带动银行资产、负债、中间业务统筹协调发展、厚植客户基础以及提升资金承接率具有重要作用。基于此，商户业务对于金融机构而言愈发重要。目前开展支付业务的收单机构达千余家，包括商业银行、第三方支付机构、电商平台、证券保险等。

发现市场中新增的潜在商户，是发展商户业务的第一步。但由于现如今市场更新迭代迅速，传统的业务人员“扫街”的营销方式在人力、效率和精准度上已无法应对复杂的市场竞争。“批量获客难”和“商户认知难”已成为当下金融机构发展商户业务的普遍痛点。

基于商户名称分析，挖掘潜在商户

使用解释性较强的数据挖掘手段，大量、精确、分行业的得到潜在商户清单是建设商户数字化经营体系最为重要的一步，也是业务部门的首要痛点。IBM 基于金融行业尤其是收单业务多年的经验与成熟的数据分析方法论，通过商户相关文本信息与行业的关联性的充分解读，使用文本分析的技术和手段，从全量外部工商企业数据中成功识别其中的有效商户，分行业输出数量可观的潜在商户清单。

挖掘潜在商户的难点在于，如何从“鱼龙混杂”的工商数据中将商户提取出来。比如，“国际商业机器有限公司”和“四川 XXX 餐饮股份有限公司”都存在于工商数据中，但“国际商业机器有限公司”显然不是商户。在数千万的工商数据中成功识别有效商户，是运用文本分析技术的主要目的。

分析客户消费流水，开启“财富之门”

银行流水素来以“多”与“繁”著称。但不可否认的是，银行每天产生的上亿条流水中往往蕴含着无穷的“宝藏”。在过往与流水应用相关的实践中，往往更聚焦于流水金额相关的统计学应用，而对于流水文本信息挖掘不足。通过对流水中相关文本进行精细化分析与挖掘，在产生更多的业务价值的同时，也可以帮助业务更好地了解新商户。

客户消费流水在挖掘潜在新商户场景应用中，具有举足轻重的地位。通过流水挖掘的新商户不仅可以得到商户名称，还可以获取商户的多维度衍生信息。但是流水中返回的交易对手信息往往是非标准化的企业名称，或者掺杂着各种无意义的符号与文字，所以，合理的 NLP 技术应用可以被誉为开启“财富之门”的钥匙。

商户营销的难点往往在于客户经理对潜新商户缺乏了解，不知道商户真正需要的是什么，也无法判断商户的潜在价值。基于流水中文本信息的解析，可以了解商户当前使用产品、经营状况等；通过分析市场同类收单产品，可以明确其喜好收单产品的特点；对其经营状况等信息的分析，可以预测其潜在价值，作为营销过程中有效的切入点与“指南针”。

IBM 在与某四大行的合作中，通过对大量消费流水的文本分析，产出超过 50 万潜在商户名单。此外，还从商户经营的角度，针对潜在商户当前现状的多种维度信息，最大程度地描绘商户的经营状况，为客户经理深挖客户在收单与经营上的痛点，形成营销方案提供合理建议，为营销成功提供助益。

分析转账流水，“解码”供应链上下游小微商户

供应链金融作为近几年新兴的金融业务模式，越来越多的银行将打通核心企业及其上下游关系作为对公业务拓展的重要方向。而这种创新的分析策略也同样可以应用于传统商户业务中。通过商户上下游链路挖掘，寻找商户之间的潜在联系，