

研究报告

2016 年第 124 期

2016.12.22

执笔：宋丹 冯乾

dan.song@icbc.com.cn

qian.feng@icbc.com.cn

非结构化文本数据发展现状及其在 商业银行的应用模式探索

要点

- 目前，部分行业已经开始关注非结构化文本数据所带来的商业价值，然而，非结构化数据分析成果转化却出现与非结构化数据的规模和增长速度相背离的现象，文本挖掘商业化过程中存在的问题主要源自应用场景的缺乏、文本数据分析的复杂程度、文本数据监管不力以及对文本数据重视程度不高。
- 随着金融行业的 IT 投资规模逐渐增大，对文本数据的分析挖掘和应用需求也不断增加。目前，根据金融行业的文本数据特点主要衍生出金融文本资讯类服务和基于文本数据的金融信息挖掘与决策两种数据应用模式，其中以证券行业应用最为广泛。
- 文本数据的挖掘与应用目前在商业银行处于萌芽阶段，主要应用场景包括意见挖掘、舆情分析、客户画像、客户经理服务效率提升和个性化内容推荐等等，商业银行发展文本数据的挖掘和应用还要注重文本数据的收集和治理、强化软硬件资源、文本数据与结构化数据的融合、自身禀赋挖掘与对外协作以及数据安全等问题。

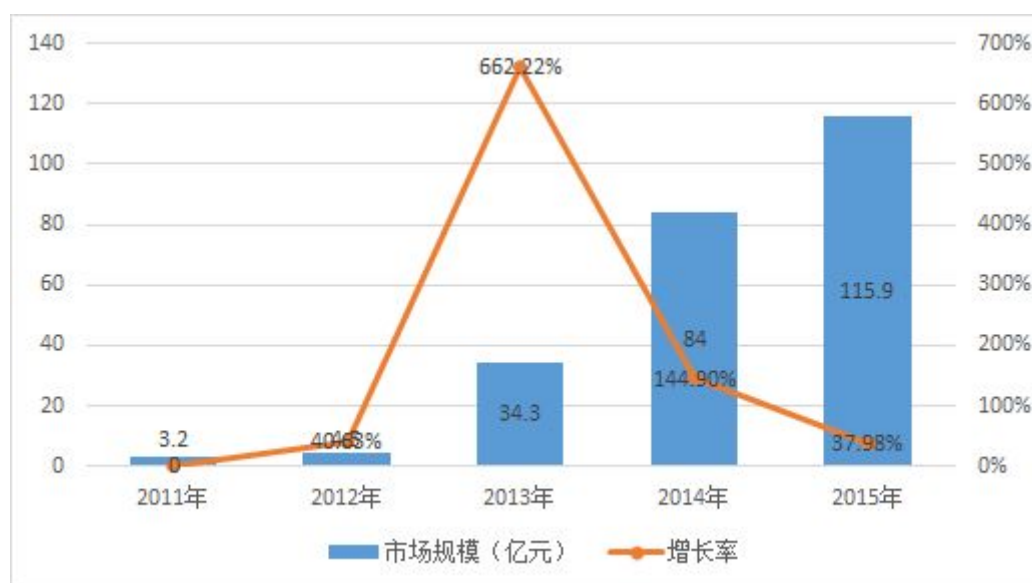
重要声明：本报告中的原始数据来源于官方统计机构和市场研究机构已公开的资料，但不保证所载信息的准确性和完整性。本报告不代表研究人员所在机构的观点和意见，不构成对阅读者的任何投资建议。本报告（含标识和宣传语）的版权为中国工商银行城市金融研究所所有，仅供内部参阅，未经作者书面许可，任何机构和个人不得以任何形式翻版、复制、刊登、上网、引用或向其他人分发。

非结构化文本数据发展现状及其在商业银行的 应用模式探索

一、非结构化文本数据应用发展现状及难点

（一）非结构化数据应用发展现状

随着互联网、多媒体、传感器和社交网络的应用，每时每刻都有大量的数据产生，有资料表明 2015 年全球数据总量达到 8.6ZB，并且以年复合增长率 50% 的速度增长。数据增长的同时，数据的使用和分析市场迅猛扩张，据易观数据统计，2015 年我国大数据市场规模 115.9 亿元，比 2014 年增长 38%（见图 1）。



数据来源：易观国际

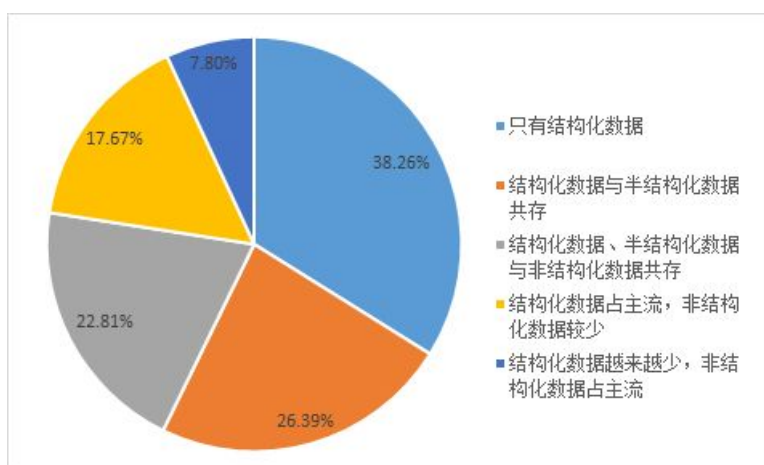
图 1 我国大数据市场规模（2011-2015）

大数据主要包含了结构化数据、半结构化数据和非结构化数据，其中非结构化数据占据了数据总量的 80%。非结构化数据是指其字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库，主要包括文本、图像、声音、影视、超媒体等信息。除了数据结构不同之外，与结构化数据相比，非结构化数据还具有以下特点：首先，数据总量更大，如前文所述非结构化数据已经占据了整个数据规模的 80%；其次，数据产生速度更快，随着社交网



络和智能终端的发展，非结构化数据的增速要远远大于结构化数据，并且有数据统计在非结构化数据中，有75%都来自于人与人之间的交互；再次，数据类型更丰富，非结构化数据主要包含了文本、图像、音频、视频和超媒体等数据类型，但实际上每个类别都可以进行细分，例如图像根据格式不同可分为灰度、彩色、红外和高光谱等等；最后，价值丰富但价值密度更稀疏，虽然非结构化数据拥有丰富的价值，但更多的数据是噪声，因此从海量数据中挖掘价值犹如海底捞针。

随着大数据市场规模不断膨胀和非结构化数据的不断增长，非结构化数据分析成果的转化却出现与非结构化数据的规模和增长速度相背离的趋势。也就是说，非结构化数据增长的速度和现有体量大于结构化数据，但对于非结构化数据的挖掘方法和成果转化还少之甚少，远不及结构化数据的应用规模。



数据来源：中国产业信息网

图2 目前我国企业数据类型分布的抽样调查

中国产业信息网对目前我国企业数据中存在的类型进行了抽样调查（见图2）。调查结果显示，目前，只有结构化数据的企业占据了38.26%，拥有结构化和半结构化数据的企业占据26.39%，存储非结构化数据的企业尚不足调研企业的一半。而且在这些企业中，大部分是结构化数据占据主流，非结构化数据较少，非结构化数据占主流并不断增长的企业仅有7.8%。

可见，虽然目前非结构化数据的规模及增速远大于结构化数据，但就其企业存储、分析和应用程度而言，其规模和应用程度呈现背离趋势，远不及结构化数据的应用程度。因此，非结构化数据的分析和应用具有很大的上升空间。

（二）非结构化文本数据及其分析应用过程

文本数据可以说是自然界中存在最早的非结构化数据，也是目前结构化程度最低、信息表达最抽象、数据规模最大和最常见的数据源。电子邮件、短信、微博、社交媒体网站的帖子、即时通信、实时会议以及可以转换成文本的录音信息都属于非结构化文本数据的范畴。这些数据通常具有以下共同特点：第一，数据取值通常是文本或字符串；第二，文本长度不一致，其长度取值从1字节起且上限不定；第三，数据通常无明确值域范围。

一个通用的非结构化文本数据分析与应用流程如图3所示：

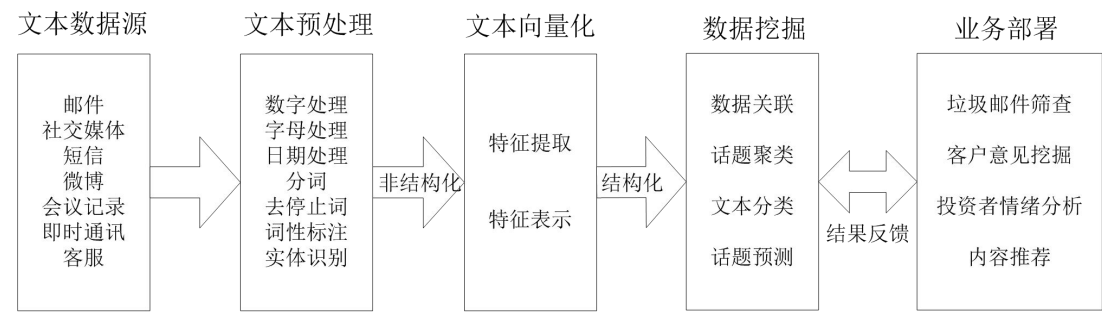


图3 非结构化文本数据分析与应用流程

1. 文本预处理。文本预处理是将文本转化为词汇的过程。由于计算机自身无法理解人类语言，因此做文本分析的第一个步骤就是将文本分解为文本挖掘所需的最小粒度：词汇。在做分词之前，首先需要对词汇中的数字、字母以及日期进行处理；分词之后，还需要根据语义习惯去除一些连词和象声词之类的无意义词汇。如果有特殊需求，还需要对分词词性和实体进行标注。需要注意的是，分词实际上类似一个有损压缩的过程，文本在转化为词汇的过程中损失了原有的语义语境关系（词汇排列无先后顺序），而且在分词过程中，部分词汇可能存在一词多义或出现生僻词汇的现象，这都对分词的准确性产生了影响。

2. 文本向量化。文本向量化是将非结构化文本词汇转化为结构化特征向量



的过程，这个步骤和其他非结构化数据处理的方法相类似，都可以看做是由非结构化数据向结构化数据转化的过程。因为计算机自身能够处理的都是结构化的数据，因此，非结构化数据向结构化数据的转化就变得尤为重要。常用的文本向量化方法主要可以分为特征提取和特征表示两个步骤，涉及的算法包括 word2vec、TF-IDF、VSM 和词频矩阵等。

3. 数据挖掘。数据挖掘是通过数据挖掘手段对文本向量进行建模，并且提取模式的过程。文本数据挖掘基本上采用了通用的数据挖掘方法（关联分析、聚类、分类和预测等），根据不同的业务场景需选用不同数据挖掘方法或选择多种方法结合使用，并且要通过准确率、召回率以及模型的可解释程度来度量模型的有效性。

4. 业务部署。业务部署是将文本数据挖掘模型最终应用在解决具体的业务问题上。具体的业务问题都可以转化为数据挖掘的模型，例如在垃圾邮件筛查过程中，将会用到分类和聚类的模型，客户意见挖掘实际上可以转化为话题分类模型，投资者情绪分析是一个情感分类问题，内容推荐可以转化为数据关联问题。数据模型应用于具体的业务问题之后，还需要根据业务的具体应用情况进行反馈，反馈数据可以用于模型的迭代优化。

（三）文本数据在行业应用过程中面临的问题

虽然部分行业在挖掘文本数据、利用文本数据来更好地做出商业决策方面已经做了很多工作，但不可否认的是，与结构化数据分析和成果应用程度相比，非结构化文本数据的分析和商业应用还有很长的道路要走。

1. 应用场景角度。“是否有用”和“怎么用”是文本数据在行业应用过程中最为常见的两个问题。“是否有用”说明在行业应用中没有找到文本数据所能服务的具体业务场景，“怎么用”则表明当前已有具体的业务需求，但是无法将文本数据与业务场景进行关联。因此，应用场景成为文本数据应用落地的最重要因素。如果没有明确的应用场景，则文本数据的发掘和应用将变得毫无意义。

2. 数据分析角度。从数据应用角度来看，非结构化文本数据的 4V 特征比结构化数据更加明显，更加贴近大数据概念的本源，但是文本数据的 4V 特征也是

导致非结构化数据存储、挖掘和应用困难的因素之一。首先，在数据存储方面，大部分企业都建立了能够存储结构化数据的关系型数据仓库，非结构化数据的存储模式仍未成熟，非结构化文本数据治理也缺乏统一的业界标准；其次，非结构化文本数据规模较大、数据类型复杂，无法应用传统的数据挖掘工具，其数据挖掘平台往往需要结合云计算、hadoop、流式计算等技术共同构建；还有，与结构化数据相比，非结构化文本数据的分析结果比较抽象，其决策结果需要大量专家的参与，在具体业务应用方面往往是仁者见仁智者见智，在文本分析过程中总会存在一定的不确定性（例如分类错误和含义模糊的问题），因此需要对错误的结果具备一定的容忍度。

3. 数据监管角度。从数据监管角度看，首先与结构化数据相比，非结构化文本数据虽然结构复杂，但表示内容更加直观、信息更加丰富，因此非结构化文本数据的信息安全问题一直以来都得到了企业的重视。首先，由于非结构化文本数据没有严格的数据结构，因此非结构化数据的信息安全管理更加复杂。其次，非结构化文本数据的应用很可能会涉及客户隐私，这需要引入适当的非结构化数据版权监管制度，以防发生隐私泄露。

4. 数据认知角度。从数据认知角度看，非结构化文本数据挖掘的理论发展已有很长历史，但是商业的广泛应用与结构化数据相比相距甚远，这主要与早期企业不重视非结构化数据有关。当前企业对大数据往往持支持的态度，积极拥抱大数据，但是大多数企业依旧停留在结构化数据的层面，而对于非结构化数据是否有用，以及如何使用持怀疑的态度。认为非结构化数据数据价值过于稀疏，且对于软硬件要求较高，获得的利益不如结构化数据大。这种观念也造成了非结构化数据成果向企业实际应用转化困难的局面：

二、文本数据在商业银行应用模式及发展趋势

（一）金融行业文本数据的来源与组成

金融行业一直以来都是大数据分析与应用的主战场。我国金融行业发展大



数据的原因主要来自于中国经济由高速增长向中高速回落以及互联网金融模式蓬勃发展所带来的挑战。近年来，金融行业的 IT 投资规模也逐渐增大，2015 年我国金融行业 IT 投资规模已达到 1500 亿，比 2014 年增长了 14.9%。

金融行业常见的数据包括了客户信息、交易数据、价格数据、成交量等等结构化数据，这些数据基本都存储在企业的数据库仓库中，是企业策略设计、趋势判断的基础。然而结构化数据仅占有所有金融信息中的一小部分，金融信息中绝大部分的数据均是以文本形式存在的非结构化数据，这同样也是一类非常重要的数据资产。

按照信息来源分类，金融行业文本主要可分为内部文本数据和外部文本数据。

1. 内部文本数据。金融行业的内部文本数据主要是由客户交互数据和企业内生数据构成。

客户交互数据主要包括客户评论、投诉、客户与客户经理的沟通等等，这类数据能够反映客户的行为、态度、意见、偏好和购买倾向。随着线上平台的发展，客户有更多的渠道去表达自己的行为、态度、意见和购买倾向等等，这部分数据也是企业内部文本数据最主要的数据来源之一。

企业内生数据主要包括电子邮件、内部文档等相关资料，这类数据通常能够反映企业内部经营决策流程和相关的行业信息等等。

2. 外部文本数据。外部数据的种类较多、规模庞大、数据分布也比较广泛。典型数据来源既有彭博、路透、万得这类专业资讯终端，又有如上市公司公告、财报、财经新闻这类官方信息，同时也有来自股吧、微博和社交网络等渠道的用户评论和意见信息。这类海量数据中蕴藏了很多重要信息，例如行业资讯信息、大众对股票的评价和喜好程度、对突发事件的褒贬态度和解读、对社交网络或微博中的热点事件进行分析，这类数据都密切影响着未来市场的趋势并且有助于金融行业找到营销方向和风险规避的方法。

（二）文本数据在金融行业中的主要应用模式

在大数据时代背景下，如何结合金融行业的业务需求，基于人工智能、数

据挖掘、文本挖掘等前沿技术自动分析海量文本数据并从中提取相关有价值信息，给金融行业各层次的企业均提出了挑战，同时带来了互联网商业智能方向的新机遇，促进了一批基于金融行业文本信息服务的创新产业的兴起。文本挖掘技术的发展与金融市场信息服务的创新将有助于减小金融市场信息不对称性，增加信息透明度，加快信息的传播，促进金融市场长期健康稳定发展。对于银行、证券、基金、保险等金融领域的细分行业来说，文本数据会根据不同的应用场景进行具体的功能设计与模型构建，当前，文本数据在金融行业的应用模式总体来说主要分为以下几类：

1. 金融类文本资讯服务。文本资讯在金融行业中扮演着尤为重要的角色。根据文本内容、风格和服务形式的不同，主要可以分为几种模型：

模式一：专业的文本信息资讯服务

随着目前情感分析、热点追踪等文本挖掘技术的兴起，基于机器可读新闻的文本应用服务受到了彭博、路透等专业信息咨询巨头的追捧。机器可读新闻主要指标包括 MRI 市场反应指标和 SIS 情绪指数等等，其中市场反应指标可以实时测量突发新闻在一个特定的证券或指数的价格方向、交易量以及波动率上的影响；而情感指数可以根据给定的公司、指数或特定主题实时展示出大众情感是怎样随着时间演化的滚动均值，可以通过提供新闻褒贬度来评估突发新闻对股票价格的影响，亦可以作为量化交易中策略参数的输入值之一。其中，路透基于机器可读新闻(machine readable news)开发出各类倾向性指数、情感追踪等服务，在新闻被发布的若干秒时间内快速生成情感正负分数，并在门户网站或图表应用程序中绘制展示给用户；瞬时量化实时新闻的影响，帮助用户快速进行高概率方向的交易，并追踪投资组合的风险。其新闻分析覆盖了超过 5000 只美国股票和 1877 只加拿大股票，分析统计多至 50000 新闻网站和 4 百万社交媒体。国内的金融资讯服务商如万德，也提供新闻与研报的关键字搜索服务，并对新闻进行了粗略的主题分类和正面负面分类。

模式二：金融类门户网站与社区



传统的金融类门户网站和社区主要包括股吧、财经论坛、财经门户新闻和微博等，这些网站或社区通常都是金融机构或客户获取和交互信息的途径与平台。但是，传统的金融门户网站和社区论坛并没有对信息进行加工和治理，随着信息容量的不断扩大，往往会存在大量的噪声文本数据。如何更好的利用这些网站和社区的海量文本数据，从中挖掘更多的价值，也是金融行业文本挖掘的热点方向。

近年来，涌现出一批新型互联网金融社区，这类社区与传统的股吧、微博和门户网站相比，在信息治理和海量信息挖掘方面进行了大量的革新，大大增加了相关信息获取的便捷性。并且这类社区积累了各种海量金融文本数据，基于此类数据的分析与挖掘将展现出这类交互社区特有的群体热点和群体观点，这对于行业研究和投资决策来说都是一类重要的信息获取通道。国内比较著名的新型金融社区如雪球。雪球建立了社区热度指标——“雪球指标”，根据关注度及增长率，讨论次数及增长率，分享次数及增长率筛选最热门的股票。这是雪球基于社区论坛积累的海量信息统计分析后的推荐小工具。国外 StockTwits 也是关于股票题材的社交平台，主要以简短讨论为主，为用户提供资讯服务。由于股票市场与投资者群体的信心和看法关联较大，股票市场可能受到群体信心、民意的影响；同时股票的走势以及投资社区上的讨论褒贬情况也在一定程度上反应了公众对经济、行业的各种预期和信心。StockTwits 搜集了投资领域各类用户的舆情和民意信息，经过分析整合以后，可以反映出金融市场的舆情趋势。

2. 基于文本数据的金融信息挖掘与决策。文本数据挖掘除具备资讯服务的功能外，近年来，越来越多的机构期望能够通过复杂的文本挖掘算法与模型来提供智能文本挖掘服务，建立基于文本数据的金融信息挖掘和决策系统，并服务于机构和用户。

文本挖掘所用的数据源既可以来自外部（互联网的新闻、博客、微博以及各类社区）也可以来自内部（银行、保险、证券、基金和互联网金融等公司），文本挖掘系统的开发和应用主体既可以是金融机构本身，也可以是专门从事文

本挖掘的第三方机构。

文本挖掘在证券行业的应用比较广泛，最常见的应用场景就是通过对上市公司公告、行业研究报告、网络新闻、论坛、微博等文本数据的挖掘分析投资者情绪，进而预测投资者情绪对大盘的影响；此外，还可以通过对大量文本信息进行加工、分析，识别行业术语并与基本面相关联，探寻其与股价之间的联系，以观察个股是否超预期。例如，美股情感分析服务网站 Stock Sonar 检索、读取和分析来自文章、博客、新闻稿等文本资源，为用户提供即时的美股文本情感分析服务，用于辅助交易决策；国内光大证券启动了“中文云”项目，专门从事金融文本挖掘研究，具有爬虫功能、文本索引、文本统计、文本热度分析、智能选股等一系列功能。

三、商业银行文本数据应用场景与应用建议

（一）商业银行文本数据的应用场景

商业银行在文本挖掘与应用方面与证券行业相比还尚未成体系，而且在理论方面的研究也相对较少，这主要与前期银行文本数据的多样性、复杂性以及缺少业务应用场景有关。银行文本数据来源包括了大量的线上客服记录、客户投诉工单、客户意见与反馈记录、客户经理与客户的销售信息、日志信息、其他类客户行为信息以及邮件和相关行业资料信息等，目前很多商业银行尚无有效的智能化手段对这类数据进行挖掘应用，仍然采用传统的人工方式对文本数据进行分类统计和处理。传统的处理方式无疑会导致处理效率低下和人力资源的浪费，同时也会造成大量有价值的信息由于得不到及时处理而被湮没，导致数据资源的浪费。但是随着新媒体的发展以及银行对文本数据的重视程度加深，商业银行目前已经拓展越来越多的文本数据挖掘与应用场景，典型的应用场景主要有以下几个方面：

第一，客户意见挖掘。对客户意见和投诉工单的处理效率，是检验银行运营状况的一个重要途径。客户意见与投诉信息能够直观地反应客户的诉求和情



绪，通过运用情感分析和文本分类等文本挖掘技术手段，可以将难以量化的非结构化文本进行量化分析，并且能够在短时间内通过传统手段获得客户对银行产品、事件和人物的评论。

第二，舆情分析。近年来，关于银行业的负面新闻报道数量持续增多。一方面，是由于客户维权意识普遍增强；另一方面，客户维权和评论的渠道增加（如商业媒体、网站、微博、朋友圈和公众号等），而且一些商业新闻媒体以及微博“大V”往往在突发性金融事件报道中夸大银行的过失和错误来吸引眼球和粉丝，这也导致银行业的声誉风险。目前，一些互联网公司已经率先建立了舆情分析体系，例如百度根据百度搜索建立了百度舆情，新浪根据微博热度建立了新浪舆情等。

第三，客户画像。客户画像的核心工作是客户信息的标签化，传统银行的客户画像一般是基于结构化数据，通过客户交易类数据以及客户基本信息构建客户的性别、年龄、行龄、资产负债和理财产品购买等标签。但是，仅用结构化数据无法全面反映客户的基本信息，例如最基本的客户风险偏好、厌恶、情感、性格等指标无从量化，这类指标通常包含在客户与银行之间的沟通信息以及客户的投诉工单中。结合文本内容的客户画像可以弥补传统基于结构化数据的客户画像中存在的不足。

第四，个性化产品推荐与客户经理服务效率提升。客户经理与客户之间的沟通聊天数据一般包括客户向客户经理的咨询信息、客户经理向客户的产品推荐信息、客户对某类产品或某类服务的情感表达、客户行为及其动因等内容。可以说，客户经理和客户的聊天信息是挖掘客户偏好、行为及潜力的有效方式。通过对大量客户经理进行产品推荐时客户与客户经理互动聊天中的博弈关系进行分析，分析成败原因，从而针对不同类型的客户制定不同的营销方案，提升销售成功率；此外，还可以通过文本挖掘技术分析出客户的行为模式、产品偏好，并向客户进行个性化产品推荐。

第五，个性化广告展示。个性化广告展示在互联网中应用得比较成熟，例如基于用户搜索内容和计算广告技术网站个性化广告展示，对不同的登陆用户

推送不同的广告。同样根据客服记录、意见评论和客户与客户经理的沟通信息可以在银行网站或其他渠道向银行客户推送个性化广告。

（二）商业银行文本数据的应用建议

目前，中国银行业基于文本挖掘的应用服务较少，但对于这类信息服务有着大量的需求。为了降低各种不利因素的影响，从海量文本数据中自动快速地提取出有价值的真实信息，更好的推动银行与客户之间的信息流动性，促进客户的发展，我们提出如下建议：

第一，重视文本数据，做好文本数据的收集和治理。在文本挖掘体系中，文本的获取是最重要的步骤之一，因此，商业银行应用文本数据的第一步就是文本数据的收集。对于外部数据，可以采用网络爬虫工具对财经新闻、投资社区和社交网络平台的文本数据进行采集；对于内部数据，推动数据存储规范化和噪声数据的治理也是文本数据价值最大化的前提。

第二，强化文本数据处理硬件和软件资源。作为大数据的一个重要分支，文本数据的数据总量远大于传统的结构化数据，而且文本数据的价值密度要比结构化数据更加稀疏，因此，为在文本数据中提取有效的知识和模式，往往需要一次性处理 GB 级别的文本数据，这会占用极大的存储器和处理器资源。因此，建议为文本数据挖掘配置专用的分布式服务器和专用的文本挖掘软件。

第三，加强文本数据与结构化数据的融合。在商业银行非结构化数据虽然规模庞大，但结构化数据依然是银行数据分析与应用的主体。结构化数据主要包括用户信息、资产负债、资金交易、产品购买等明细数据，而非结构化文本数据则包含了客户行为、个人偏好、情感状态等抽象数据，因此可以说非结构化数据与结构化数据是互为补充的关系。

第四，自身禀赋挖掘与对外协作。商业银行发展文本数据挖掘的优势在于其业务应用场景和宝贵的数据资源，但是在数据挖掘方法方面与 IT 或互联网公司相比还存在一定的差距。目前很多银行都采用与 IT 公司合作的方式开展文本数据挖掘与应用，例如建设银行与神州泰岳联合建立智能文本挖掘系统——新一代核心系统文本分析与互联网信息采集工具项目。



第五，重视文本数据的隐私安全及文本数据使用规范的建立。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，文本数据包含了客户与银行之间沟通的信息，甚至可能包含卡号和密码信息，如果不注重隐私安全，极易产生隐私泄露的问题，甚至产生法律风险。因此，文本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。

第六，突破传统的数据分析与评价体系。文本数据作为一类非结构化数据，其分析方法、应用模式和评估准则与结构化数据都有显著的不同，因此，文本数据挖掘的需要建立不同的数据挖掘体系架构。而且，文本分析本身是一个数据有损压缩的过程，在数据预处理过程中会损失部分短句、数字和字母，在分词过程中会损失语义和上下文关系，这都会造成文本数据价值的损失，因此，不能以结构化数据的准确率、召回率评价模型来评价文本数据挖掘效果。