证券行业文本挖掘技术应用现状与探讨

白雪，熊昊
上海证券交易所 发展研究中心，上海 200120
E-mail：xbai@sse.com.cn

摘　要：证券行业的海量信息由结构化数据和非结构化数据构成。在当今大数据背景下，越来越多的有价值信息隐藏在海量文本数据中，从而加大了对自动快速的从大规模文本数据中提取信息、发现知识的需求。文本挖掘是自动从文本数据中挖掘潜在的事先未知的新知识的过程，其挖掘算法的发展与积累为证券行业文本数据分析与信息获取打下了坚实的技术基础。基于蕴藏着宝贵信息的大数据，如何结合证券行业的特点和需求，借助挖掘算法与模型进行服务创新和交易方式创新，是本文所关注的重点。为此，文本对国内外证券市场的文本数据服务进行了一系列调研，并对我国证券市场基于文本挖掘的应用与服务进行了分析与讨论。

关键词：数据挖掘；文本挖掘；情感分析；大数据；创新服务

1 引言

近年来，互联网在线文本数据的爆炸式增长大大增加了各行各业的相关信息阅读量，如何从充斥着噪音及各类繁杂信息的数据海洋中自动提取高度相关的有价值的信息已成为企业界与研究界共同关注的热点。作为一个跨学科交叉领域，文本数据挖掘涉及了多个研究方向，如数据挖掘，自然语言处理，信息检索，机器学习等等，正越来越多地应用于现实生活中各类应用场景。

证券行业常见的数据包括了股票价格、成交量等等结构化数据，和包含了各类公司信息、新闻等非结构化数据。其中，结构化数据通常以数值形式存放于标准数据库中，这类数据是各种策略设计、趋势判断的基础。然而结构化数据仅占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一种非结构化的数据，如上市公司公告、财报、财经新闻、股吧、微博、社交网络等等。这类海量数据中隐含了很多重要信息，例如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，都密切影响着未来市场的趋势。因此，在大数据时代背景下，如何结合证券行业的业务需求，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，给证券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，促进了一批基于证券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证券市场信息服务的创新将有助于减小证券市场信息不对称性，增加信息透明度，加快信息的传播，促进证券市场长期健康稳定发展。

文本将调研并探讨证券行业文本挖掘服务现状。首先在第二部分对文本挖掘的基本概念、挖掘步骤、常用开源工具和常见的几类挖掘算法展开简要介绍。在第三部分对国内外证券行业文本信息服务进行了调研，将已有的服务分为三大类，分别为投资综合性社区、文本信息资讯和专业文本挖掘。第四部分讨论了我国证券行业文本挖掘的应用现状和面临的问题。最后，对全文进行了总结并展望。

2 文本挖掘概述

2.1 文本挖掘简介

文本挖掘是自动从文本数据中挖掘潜在的事先未知的新信息的过程，与自然语言处理，信息检索，信

息提取，知识发现，数据挖掘，机器学习，统计学等研究领域密切相关[1]。文本数据具有高维、稀疏等特点，可以基于不同层次的表示法展开分析。例如词袋法(bag-of-words)，或词串法(string of words)。目前大多数文本挖掘方法都基于词袋法，与基于语义及自然语言处理的词串法的相比，词袋法相对较简单，处理较为方便。

文本挖掘通常可分为两大步骤，首先是文本数据准备，包括文本获取，预处理，分词，词性标注，文本表示等等；第二步是文本数据挖掘，如文本分类，主题挖掘，情感分析等多种基于各类应用与需求的分析挖掘。近年来，学术界已涌现出很多经典的文本挖掘综述文章，如 A.Hotho 等人[1]的《文本挖掘综述》(A Brief Survey of Text Mining)，C.C.Aggarwal 等人[2]编辑的书籍《挖掘文本数据》(Mining Text Data)，对文本挖掘及其相关方向进行了系统的介绍、总结与综述。下面对文本挖掘两大步骤及其常见技术、算法模型展开简要介绍。

2.2 文本数据准备

文本数据准备主要进行数据获取、预处理和分词，为进一步文本数据挖掘做准备。关于文本数据获取、中文分词、自然语言处理等细分方向，已涌现出众多论文综述和开源系统。本节从实际应用角度出发，简要概述文本数据准备的几个主要步骤和每个步骤目前较为流行的开源工具。

通常来说，文本数据的获取可以通过选择若干知名财经门户网站、股吧网站、微博等作为目标源，搭建网络爬虫抓取相关文本数据。网页抓取策略可以分为深度优先、广度优先和最佳优先三种。常用的开源爬虫工具有 Heritrix，Nutch，Larbin 等。Heritrix 基于 Java 语言开发，是一个开源、可扩展的网页爬虫框架，支持网页镜像保存，适用于 Linux 系统和 Windows 系统。在 Heritrix 的配置页面，用户可以进行详尽的设置，包括网页抓取范围，抓取到的信息是以压缩还是镜像的方式写入磁盘，抓取线程个数，抓取间隔时间等等。用户可以通过配置参数或修改扩充源代码，面向特定主题搜索数据。

爬虫工具所抓取的网页通常包含很多乱码、链接、图片等噪音，需要进行文本提取，去噪，去重复等操作进行清洗处理。经过清洗后的干净的文本文档，再进行下一步的操作。英文文本的数据预处理通常包括过滤、词性还原、词干提取、关键词提取、句法解析等步骤。由于英文中文语法、文法的差异性，中文文本的预处理并不需要进行词性还原、词干提取等过程，而中文分词则为文本数据准备中最为关键的一个重要步骤。

中文分词是将中文字符序列切分成一个个单独的词的过程，是文本挖掘的基础。中文分词的方法有的基于统计，有的基于字符串匹配，有的基于句法语义来进行分词。中文分词的学术性中国科学院计算技术研究所研究的汉语词法分析系统 ICTCLAS(Institute of computing Technology, Chinese Lexical Analysis System)，可以进行分词、词性标注、命名实体识别，且支持用户自定义词典，是目前最为热门的中文分词系统。ICTCLAS 基于 C/C++ 语言开发，后期也推出了支持 Java 等开发语言的版本。其它的分词系统还有 IKAnalyzer，LibMMSeg 等等。

2.3 文本数据挖掘

文本数据挖掘包含了文本分类、文本主题挖掘、文本情感分析、文本聚类、生物文本挖掘等等各种细分研究领域和各类应用场景。通常来说，文本挖掘算法模型的选择和设计与具体应用密切相关。例如，若需要调查证券行业在线股评新闻的褒贬态度，以了解和跟踪市场投资者情绪，则需要用到文本情感分析相关的模型和算法。下面简要介绍文本分类和文本情感分析相关概念和常用模型。

2.3.1 文本分类

文本分类的问题定义如下[2]：已知一个训练数据集 $D=\{X_1,...,X_N\}$，其中每条记录打有一个类标签，类标签的值取自由 k 个不同离散值组成的集合 $\{1,...,k\}$。由训练集构造一个分类模型，用于给新的文本记录分类，预测相应的类标签。

文本数据具有高维、稀疏等特点，若直接采用词汇向量作为特征进行计算，容易造成维数灾难（Curse of Dimensionality）。因此文本分类常常首先进行文本特征选择(Feature Selection)，以决定哪些是最为相关的、重要的特征，以提高分类的效率。常见的文本特征提取方法有基尼系数(Gini Index)，信息增益(Information Gain)[3]，互信息(Mutual Information)[4]，

LSI(Latent Semantic Indexing)[5] 等等。

然后，选择分类模型构建文本分类器。决策树 [6]，支持向量机 (SVM)[7]，神经网络，贝叶斯分类等等均为经典的分类算法，其中 SVM 模型为近年来较为流行的算法。

支持向量机 (SVM) 是 90 年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化。它的基本思想是在样本输入空间或特征空间构造出一个最优超平面，使得超平面到两类样本集之间的距离达到最大，从而取得最好的泛化能力。支持向量机在解决小样本、非线性和高维等问题中表现出了很多特有优势，且分类准确性较高，稳定性较好，已成为众多文本分类研究首选的分类模型。

SVM 的分类基本思想如下图所示，原点与方块分别代表两类样本，H 为划分超平面 ,H1，H2 分别为过这两类中距离超平面最近的样本且平行于超平面的平面，它们之间的距离叫做分类间隔 (margin)。所谓最优超平面就是要求此平面不但能将两类正确分开 , 而且使分类间隔最大。支持向量机就是要寻找这个最优超平面，而那些边缘分类面 H1，H2 上的点 (蓝色点) 就是支持向量。

SVM 的最终决策函数只由少数的支持向量所确定 , 计算的复杂性取决于支持向量的数目 , 而不是样本空间的维数 , 这在某种意义上避免了“维数灾难”。少数支持向量决定了最终结果 , 这不但可以抓住关键样本、“剔除”大量冗余样本 , 且使得该方法不但算法简单 , 并具有较好的鲁棒性。

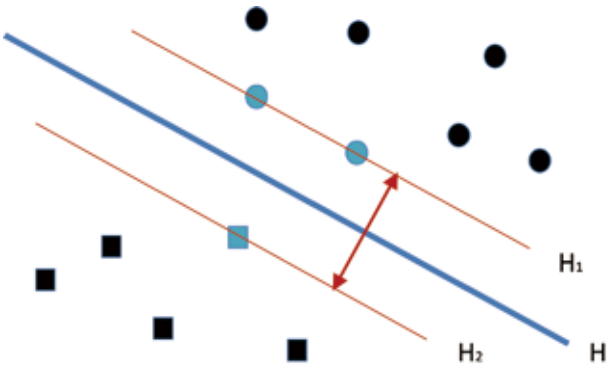

图 1 SVM 最优超平面

2.3.2 情感分析

情感分析 (Sentiment Analysis)，又称为观点挖掘 (Opinion Mining)，主要研究从文本数据中识别和发现主观性情感信息，并对情感倾向性进行深入分析。一个观点可以定义为一个五元组，即 $(e_i, a_j, o_{ijkl}, h_k, t_l)$ [8]，其中 e_i 为一个实体 (Entity)，a_{ij} 为 e_i 的一个方面 (aspect)，o_{ijkl} 为针对实体 e_i 的属性 a_{ij} 的情感观点倾向，h_k 为观点持有人，t_l 为观点发表的时间。其中观点倾向可以是正面的、负面的或是中性的，还可以伴有不同的情感强度。给定一组含有观点的文档集合，情感分析旨在找到集合中所有的观点五元组。因此，情感分析的任务为，给定一个包含对某一对象情感的评价文章集合，从中抽取每一篇文档中所评论对象实体及其对应的属性和组成元素，以及相对应的观点，提取观点持有人及发表时间，并判断观点是正面的、负面的还是中性的。

情感分析方法根据研究层次的不同，可以大致分为基于语料库级别、基于文档级别、基于句子级别、基于属性级别四大类。其中，语料库级别情感分析也被称为整体倾向性挖掘，是以海量文档数据作为分析对象，对情感倾向性信息进行统一的集成和分析，得到整体倾向性特点。基于文档级别情感分析假设一篇文档包含针对某实体的一个观点 [9]，以整篇文档为基本单位进行情感挖掘。基于句子级别情感分析则更为细粒度，以语句作为基本处理单位来判断其情感倾向性。基于属性的情感分析针对实体对象的不同属性 (方面) 分别进行情感倾向性挖掘，这种级别的分析方法将细至词汇级的粒度，提取出实体的多种属性并分别进行倾向性判别 [9]。文献 [10] 对亚马逊电子商务上的评论进行了产品属性提取和情感挖掘。Mei 等人 [11] 提出一个挖掘主题和情感极性的混合模型，可以分别得到针对每一个方面的评价极性。文献 [12] 将主题和情感极性的挖掘分成两个步骤，先寻找主题，再判别极性。

情感分析方法主要可以分为两大类：基于字典的方法和机器学习方法。基于情感字典的情感分析方法常常首先研究如何建立一个情感词典，或利用已有语义词典中的词语来扩展和构建规模更大的新情感词典；然后基于情感词的极性构造若干函数来计算文本的正面或负面的程度。由于本文主要针对中国证券市场，

将主要研究中文文本词典的构造。现有的很多文献基于董振东的知网 (HowNet) 进行改进。比如文献 [13] 基于 HowNet 提出了一种中文词汇情感量化的方法。文献 [14] 基于多个情感词典构造了一个统一的中文情感词典等等。基于机器学习方法主要研究情感倾向性的判别问题，通常来说，将其看作是一个分类问题，常用的分类方法比如支持向量机模型 (SVM)，朴素贝叶斯、最大熵等等，参见 2.3.1。Pang 等人 [15] 比较了朴素贝叶斯模型，支持向量机，最大熵等分类模型用于给电影评论的情感极性做分类时的效果，其中支持向量机方法取得了相对较好的结果。

情感分析具有较强的领域性特点，即分析的准确性会受到领域不同的影响。这是由于同一个词，在不同的领域可能代表了不同的情感极性所导致。领域不同，分析模型的性能可能相差甚远。常见的情感分析应用方向有电子商务领域的消费者产品在线评论的倾向性挖掘，微博或网上社区的舆情分析，股票评论分析与投资者信心预测等等。

3 证券行业文本信息服务

文本信息在证券行业中扮演着尤为重要的角色。专业资讯供应商如彭博、路透、万得终端在信息的快速搜集和推递方面依旧保持着他们强大的优势，而其高昂的服务价格常使得广大个人投资者望而却步。传统的股吧、财经论坛、财经门户新闻一直以来都是散户们获取信息的途径与交流平台。在信息爆炸的今天，这些传统论坛简单的呈现方式使得个人投资者从充斥了大量噪音的文本数据中获取有价值信息已越来越难；同时，对专业资讯供应商而言，如何更好的利用海量文本数据为高端用户提供更专业更有效率的文本服务，亦为一项重要的创新与研发方向。

近年来，涌现出一批面向广大个人投资者的新型投资综合性社区，这类社区与传统的股吧相比，在处理海量信息方法和活跃投资者讨论方面做了大量的革新，大大增加了人们之间的讨论与交流以及相关信息获取的便捷性。并且，这类投资综合性社区积累了个人投资者的发布、讨论、关注等各种海量行为数据和文本数据，基于此类数据的分析与挖掘将展现出这类交互社区独有的群体热点、群体观点、群体智慧汇集等优势，这使得投资社交网络逐渐成为行业中尤为重要的一类信息获取通道。3.1 节将首先介绍投资综合性社区及其代表性企业。

在专业资讯供应商方面，传统的行业巨头也在加大信息分析、整合的力度与深度，以求在信息爆炸时代更好的提取有用信息。同时，也出现了基于文本信息整合与分析的细分领域资讯供应商。将在 3.2 节展开介绍。

在文本数据挖掘服务的细分领域，开始涌现出一批专业性、技术性较强的公司，基于文本挖掘复杂算法与模型为用户提供智能文本挖掘服务，将在 3.3 节进行阐述。

3.1 投资综合性社区

投资综合性社区是近几年来兴起的一类专业型社交网络平台。与传统的股吧、论坛不同，投资社交网络以“投资”这一共同的兴趣爱好聚集了大量的用户群体，专业性较强，并且逐渐建立起人与人之间、人与股票之间、股票与题材之间等等各类较为稳固的联系，形成了各种重要消息发布、频繁互动、思想碰撞的交流平台。

目前最著名的投资综合性社区为国外的“StockTwits”和国内的“雪球”。这类论坛围绕“投资”主题，提供了各种便利服务，例如关注股票和话题，订阅股票基金 ETF，收取新闻公告，参与用户讨论实时交流互动等等。投资综合性社区目前已经聚集了大批投资者和证券行业人士，他们在证券相关主题上发表各类言论，提供专业资讯与见解，并参与话题讨论。

投资综合性社区积累了大量的用户发文、评论、关注以及各类行为的文本数据，为基于该社区平台的的各种文本挖掘服务提供了稳定的数据源。这些文本数据通常带有主观性情感，有较为明确的评论对象，相对其他论坛而言信息含金量较高，这些特点均为细分文本分析服务提供了高质量的原始数据。

随着这类数据大量累积，投资综合性社区将拓展开发出越来越丰富的信息服务与个性化服务。如，可以基于积累的大数据做进一步深层次的分析与挖掘，整合社区群体信息提取加工成有价值的情报，进一步提高用户体验和用户粘性。这类面向终端投资者的文本数据挖掘服务，既可以是投资社区自主研发，也可以与第三方 IT 公司合作基于该平台数据提供更为细分且丰富的服务，从而形成基于投资主题的面向终端用户的全新的信息服务产业链。

3.1.1 StockTwits

美股投资社区 StockTwits 创办于 2008 年，是一个投资行业人士发布和关注金融新闻，并对股票和题材进行讨论交流的平台。StockTwits 受 Twits 风格影响，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。

为了更好的服务投资者，StockTwits 对投资相关的各类数据进行整合、分析与展示。以 Google 股票为例，当用户选择关注 $GOOG 后，主页面将展示所有关于 Google 的讨论，以时间先后排列；页面右上方展示关注用户数，以及价格、讨论量、情感倾向性的随时间变化走势图，如图 3.1 所示。其中第一项为投资者最为熟悉的量价 K 线图；第二项为评论热度统计的时间序列展示，可反映出该股票被关注、评论的随时间变化的热烈程度；而第三项 $GOOG Sentiment 为基于海量文本数据的情感分析 (看涨看跌倾向性) 变化图。截至 2014 年 1 月 26 日，约 88% 的评论看涨，12% 的评论看跌，且近几日来看涨的评论比例处于缓慢增加趋势。

除了针对个股的文本统计与分析，StockTwits 还有热点地图、讨论最多的行业、热门公司等服务。通过这些应用服务，StockTwits 将大量经用户思想碰撞所产生的有价值的信息从海量文本数据中提取挖掘并展示给用户，这在信息大爆炸的今天尤为重要，因为只有借助文本挖掘手段自动从海量且包含了各种噪音的大数据中及时发现隐含的有价值的信息，才能大大提高有价值的信息的利用率并加速其流动与传播。

3.1.2 雪球

雪球社区创立于 2011 年，为新兴的中文投资者社交网络。与 StockTwits 类似，雪球以“加关注”的方式在人与人之间建立起联系，以“@ 股票”的方式将共同关注某一股票或某题材的用户聚集起来。雪球的用户包括业内人士、个人投资者、公司高管等等各种与投资相关的人，致力于帮助用户快速获得公司公告，相关新闻，和用户讨论。

雪球社区在成立初期，新用户须通过老用户邀请才能注册。2012 年 10 月，雪球正式对外开放注册，用户数稳步增加。目前雪球已覆盖中、港、美三大市场和股票、债券、基金、信托、理财产品等多个投资品种，以及比特币等新兴热门产品，用户可以方便地获取这些投资产品的价格、数据、资讯和讨论。雪球社区经常组织公司代表访谈等活动，方便公司维护投资者关系，有利于投资信息的披露与传播。

2014 年 1 月，雪球新增社区热度指标——“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。对个股而言，其评论热度和评论数可以与股价走势的时间点相叠加，用户可以同时浏览在某个时间点的量、价

图 2 StockTwits $GOOG 数据一览

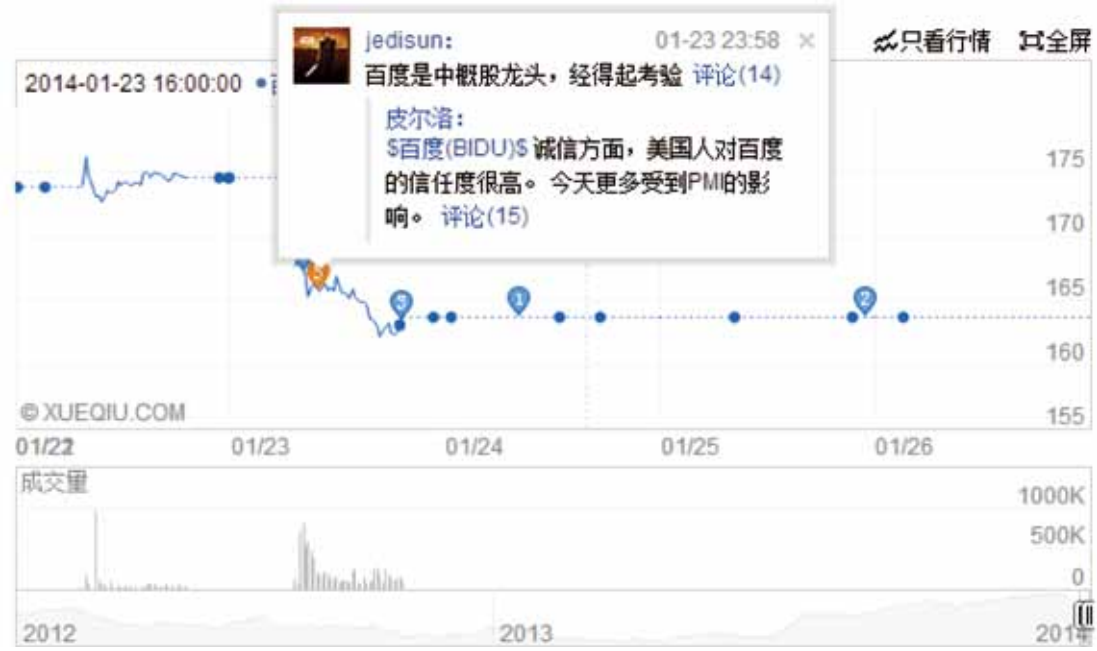

图 3 雪球个股 (BIDU) 讨论热度与内容

和评论，参见图 3.2。该功能有些类似于 StockTwits (图 3.1)，但相对较初步和简单，并未将文字中的褒贬语义进行进一步分析与剖析。

3.2 文本信息资讯服务

受 Twitter 情感分析、热点地图等文本挖掘服务获得广泛好评及热议的影响，彭博社、汤姆森路透等传统金融资讯巨头近年来也开始启动基于财经新闻文本挖掘、分类、情感分析等服务，同时加大了对该方向人才的招聘力度。其中，汤姆森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，大大提高了新闻的解读速度和运作效率。

国内的金融资讯服务商如万得，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。在金融文本信息资讯方面，朝阳永续为一家专注于文本数据的搜集与处理的公司，在文本资讯细分领域占领了部分市场。然而在文本信息利用、文本快速挖掘方面，国内资讯服务商与国外相比尚有较大的差距。

3.2.1 汤姆森路透——机器可读新闻

机器可读新闻 (Machine Readable News) 为电脑自动生成的满足一定规则的便于给电脑“阅读”的新闻，在公司发布收益报告或政府发布经济统计数据的时候自动提取产生，经过处理转换后直接提供给另外一些根据新闻进行计算或交易的电脑。近年来，路透、彭博 (Bloomberg) 等资讯供应商对于“机器可读新闻”的需求量大幅上升。这些新闻以电脑可读的语言编写，由一连串的字符和数据组成，没有传统新闻中的句子成分。机器可读新闻使得电脑可在收到信息的毫秒级别时间内则可以根据新闻进行相关处理，把解读新闻的工作交给了电脑处理，由电脑自动提取出新闻中的重要信息，其速度远非人类之所能及。

汤姆森路透公司基于机器可读新闻开发出了一系列产品与服务 [16]。例如，基于机器可读新闻可以将所有与股票和衍生品相关的新闻进行情感正负量化打分，其数值在 −1000 到 +1000 之间；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，或是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。公司提供三种封装打包方式：程序化、专业型、移动门户网站，以符合量化交易员、投资组合管理、市场数据管理等各类用户的需求。

基于机器可读新闻的应用服务有例如 MRI 市场反应指标，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。

图 3.3 展示了 EOTPRO 汤姆森路透机器可读新闻门户网站的一个示例，列出前 10 位最受关注的各种正面的或负面的新闻事件和话题。 EOTPRO 网站可实时展示美国股市的社交媒体情感性和新闻的褒贬性，传递大量社交媒体态度和新闻分析，以支持交易，投资和风险管理决策制定。

3.2.2 朝阳永续——研究报告数据库

朝阳永续是一家上市公司盈利预测数据提供商，主要收集的文本数据主要有三类：新闻、研究报告和上市公司公告。公司推出了一款“一致预期”数据库产品，以卖方上市公司研究报告为基础，形成了一套关于个股、行业、指数的未来三年的预期数据库产品。这个数据库包括公告库，研究报告库，可以按照事件分类，也支持关键字查询，如图 3.4 所示，很大程度上帮助了行业内人士对文本信息的获取和分析。

公司主要针对分析师研究报告做整理和分析挖掘，具体来说可以分成五个部分：(1) 卖方原始预测数据的校对和清洗；(2) 一致预期行业基准数据的生成；(3) 报告相关衍生品，例如情绪等；(4) 事件库的建设；(5) 文本挖掘，如分词、重点关键词的提取，分类，主题挖掘等。关于深层次的文本数据挖掘，公司的研究团队尚处于研发阶段，目前已完成对海量新闻数据的采集和清理，分词，预计 2014 年底可以推出新的数据服务和产品。

图 4 EOTPRO 汤姆森路透机器可读新闻门户网站

图 5 朝阳永续研究报告主题筛选页面

3.3 专业文本挖掘服务

近年来，涌现出一批基于文本挖掘复杂算法与模型为用户提供智能文本挖掘服务的互联网公司。这类公司专注于专业文本挖掘服务这一细分市场，为用户提供高端文本挖掘服务。其数据源主要来源于互联网的新闻、博客、微博以及各类社区，也有公司专门基于社交网站的大数据展开深度挖掘与分析，例如 SmogFarm。

3.3.1 美股情感分析服务——Stock Sonar

Stock Sonar 检索、读取和分析来自文章、博客、新闻稿及其它基于对某个文本的意义深入理解的公共信息等广泛的在线资源，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。该系统以网络媒体文本为主要数据源，量化其文本情感的正负及幅度，并实时展示给用户。可以针对个股、指数或者主题板块进行情感度量和比较，即时发现投资机会，为用户订制的投资组合发掘并量化媒体情感，并以可视化方式展示出来。相关的媒体文本将以深色强调显示以便于用户快速阅读。

Stock Sonar 旨在为交易商和投资者提供实时而强大的决策支持工具，快速而清晰的展现与投资决策相关联财经新闻、社交媒体的情感倾向性。正如图 3.5 所示，与 StockTwits 的情感工具相比，Stock Sonar 更为细致专业，展示界面清爽，趋势一目了然。

在情感倾向性曲线下方，列有详细的文章列表和事件列表，以“打分 – 标题”的方式按照时间先后排列。图 3.6 所示为奇虎 360 股票相关的文章列表，左侧的打分条以颜色区别情感倾向正负性，并标以对应情感程度的数字。同时页面右边将展示原文出处，重点负面关键字段用红色标出，正面字段用绿色标出，帮助人们复查新闻并快速定位到相关段落。

3.3.2 大数据分析——Smog Farm

Smog Farm 是一家进行大数据情绪分析的公司。正如其标语“Harvesting the cloud”所示，公司希望基于数据云“收获”隐藏在海量数据中的价值信息。公司汇集了大数据、语言学、群体心理学等方向的专家，在此交叉领域研究开发各类相关产品与服务。

该公司的首款产品 KredStreet——“社交化股票交易员排名”，主要是根据社交投资网站 StockTwits 的数据进行分析，从而确定交易员整体是看涨或看跌。根据某个时间点某交易员看涨看跌判断记录，与股市

图 6 Stock Sonar 的 Google 正负情感曲线

图 7 奇虎 360 股票相关的新闻列表节选

的真实走势进行对比，对交易员打分并根据准确率进行排名。

图 8 Smog Farm

4 证券行业文本挖掘应用现状

4.1 相关研究报告

近年来，尽管主流的券商研究策略报告仍然基于传统的金融工程、统计方法，已可以观察到越来越多的报告中开始出现数据挖掘、文本挖掘、云平台、大数据等关键词。涉及的领域有金融工程，投资策略建议等等。

光大证券建立了一个中文云文本挖掘系统，并撰写了多份相关研究报告，如 2012 年的《金融文本挖掘前述——系统、数据、指标》，《基于深层次文本挖掘的策略研究》等等。该系统将文本挖掘与金融量化相结合，衍生出多种服务产品，包括概念板块套利，关键词热度，投资者情绪指标择时模型，关注度选股模型等等。

国信证券的金融工程研究团队也就数据挖掘与金融工程交叉领域进行了深入的研究，可搜索到自 2010

年起便有相关研究报告发表。例如报告《国信投资时钟之行业关联网络》借助复杂网络模型研究版块轮动。《基于动态时间规整的择时策略》采用了动态时间规整(Dynamic Time Warping)——一种高灵活性适应性的时间序列相似度度量方法来测量两条时间序列的相似程度，可以大大降低因相邻时间的细微不一致而导致的相异度增加，而从大体形态上捕获时间序列的相似度，进一步聚类以发现模式和规律。《交易性数据挖掘系列报告》等一系列报告也详细阐述了量化投资者情绪指标，基于动态规划做预测，均线型技术跟随等。

此外，海通证券、广发证券、中信证券、国金证券、国泰君安、宏源证券等也就数据挖掘、商业智能、大数据等相关领域进行研究并发布了研究报告。

4.2 行业应用现状与问题探讨

在中国证券市场，目前针对股票领域的文本挖掘的应用与服务较为少见。其原因主要有两点，一是挖掘技术以及数据源面临的难题，二是股票市场自身的复杂性。

股票所涉及的文本种类繁多，例如上市公司公告、行业研究报告、网络新闻、论坛、微博等等。需要对大量文本信息进行加工，分析，识别行业术语，联系基本面，探寻其与股价之间的联系，以观察个股是否超预期。这其中，首先需要对历史文本信息进行收集和积累。目前证券行业内的数据库大多基于结构化数据进行存储，非结构化数据方面，朝阳永续的公告库和研究报告库目前可以进行简单的分类和关键字查询，对于业内人士对文本信息的获取和分析起到一定的帮助。然而由于非结构化数据量大，迫切需要研发自动化挖掘工具和应用辅助业内人士进行信息提取和分析。数据源方面，国外的文本信息量大，且针对个股的文本信息持续性好，历史信息也积累的较为齐全，针对个股做文本挖掘时有大量素材。而在中国，文本信息存在大量冗余，经常一个新闻被转载数十次上百次，且针对个股的文本信息持续性不强，文本积累较少。

信息源的可信程度也是文本挖掘面临的一大难题。比如，如何判断来自网站新闻、股吧、微博等等消息的真假，如何降低噪音对真实信息的影响等等。股票信息源的噪音和虚假新闻众多，难以分辨真伪。信息发布者立场问题也会大大的影响信息的可信度。例如利益相关的新闻发布者发布的信息，将会带有主观倾向性和引导性，极力修饰负面信息而放大正面信息。这些都大大增加了信息的理解和辨别难度。

中国股票市场中，判断信息对于股票的影响是个难题，可因不同个股、板块、事件、政策、环境等各种因素而异。有时候利好消息不一定会导致股票上涨，尤其在流动性不好的股票中，经常会出现相反的情形，这使得信息与股票之间的规律与联系充满了不确定性。

目前，在量化投资中大量使用的是结构化数据。然而价格一定有不可解释的部分，很多信息隐藏在非结构化数据中。例如,各种新的“概念”和“热点”的挖掘。研究报告上的点评的理解、解读和联系，这些都需要对非结构化数据进行分析。有些时候，一些股价动荡

是因为被动的调整，比如 2013 年 12 月 20 日 QFII 跟踪指数进行调仓，导致建设银行、中信银行、交通银行等相关股票尾盘异动，这类情形也只能从文本获得信息进行解释。

总之，中国证券行业目前基于文本挖掘的应用服务较少，对于这类信息服务有着大量的需求。如何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，以更好的推动证券市场信息流动性，促进市场的健康有序发展，仍是一个需要不断努力的课题。

5 总结与展望

在当今大数据时代，隐含了有价值信息的、与人类行为息息相关的大规模数据将日益成为稀缺资源和宝贵财富。这其中，大量产生于社交媒体，门户网站，微博等等的互联网文本数据，近年来发展迅猛。基于互联网文本数据的分析挖掘与知识发现，已然成为了全球研究界和企业界关注的热点课题。由于中文文本自身的特性，国外基于英文的文本挖掘方法很多无法通用，这需要我们基于中文特点和具体的应用场景进行持续深入的研究。

目前文本挖掘在证券市场的应用较少，且面临着各种难题，如文本数据历史库的收集与处理目前尚未成熟，文本倾向性与股市走势时而相背离，文本信息可信度以及发布者利益相关问题等等。然而结构化数据并不能解释所有证券市场的所有现象，大量的信息隐藏于海量本文数据中。投资者需要从海量文本中获得准确信息来解释市场的行为和现象，并辅助投资决策，因而对文本挖掘以及基于文本数据的各种信息服务有着迫切的需求。

中国证券行业文本数据挖掘服务尚处于起步阶段，尚面临着各种机遇和挑战。可通过借鉴国外成熟应用的例子，广泛深入调研并结合现实应用场景，加大中文文本挖掘的研发力度，以迎来大数据文本挖掘服务交叉领域产业与服务的腾飞。

参考文献：

[1] A. Hotho, A. Nurnberger, and G. Paa. A Brief Survey of Text Mining. in J. for Computational Linguistics and Language Technology, 2005.

[2] C. C. Aggarwal, C.X. Zhai (Eds.) Mining Text Data. Springer ISBN 978-1-4419-8462-3, 2012.

[3] Yang Y Pedersen JO. A comparative study on feature selection in text categorization. In：Fisher DH,ed Proceedings of the 14th International conference on Machine Learning (ICML97)1997：Nashville：Morgan Kaufmann Publishers；1997.412-420.

[4] MACRO Zaffalon,MARCUS Hutter.Robust feature selection by mutual information Distributions.In: Proceedings of the 18th international conference on uncertainty in ratification intelligence,UAI,2002,577-584.

[5] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman. Indexing by Latent Semantic Analysis. JASIS, 41(6), pp. 391–407,1990.

[6] J. R. Quinlan, Induction of Decision Trees, Machine Learning,1(1), pp 81–106, 1986.

[7] C. Cortes, V. Vapnik. Support-vector networks. Machine Learning, 20: pp. 273–297, 1995.

[8] B.Liu and L.Zhang. A Survey of Opinion Mining and Sentiment Analysis.Book Chapter in Mining Text Data, Ed. C. C. Aggarwal, C.X. Zhai, 2012.

[9] Ronen Feldman. Techniques and Applications for Sentiment Analysis. Communications of the ACM, Vol. 56 No. 4, Pages 82-89.

[10] 冯小翼．在线评论的产品属性提取与情感分析研究．华中科技大学硕士学位论文，2011.

[11] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and Chengxiang Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In Proceedings of WWW, Pages 171-180, New York, NY, USA, 2007.

[12] Koji Egchi and Chirag Shah. Opinion retrieval experiments using generative models: Experiments for the TREC 2006 blog track. In Proceedings of TREC, 2006.

[13] 柳位平，朱艳辉等．中文基础情感词词典构建方法研究 [J]. 计算机应用，2009,29(10):2875-2877.

[14] 王素格，杨安娜，李德玉．基于汉语情感词表的句子倾向性分类研究 [J]. 计算机工程与应用，2009,45(24):153-155.

[15] Bo Pang, Lilian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Pages 79-86, 2002.

[16] http://www.machinereadablenews.com/index.php