基于文本大数据分析的会计和金融研究综述①

马长峰[1]，陈志娟[2]，张顺明[3*]
(1. 上海国家会计学院数字金融研究中心，上海 201702; 2. 浙江工商大学金融学院，杭州 310018;
3. 中国人民大学财政金融学院，北京 100872)

摘要: 作为一种非结构化②数据，文本大数据最近十年深刻影响会计学和金融学研究. 这种影响体现在两类文献: 第一类以信息为中心，将文本分析技术用于信息的品质(可读性)和数量(文本信息含量)、信息披露和市场异象等方面的研究; 第二类与信息无关，主要是利用文本大数据分析技术构建全新指标，例如基于文本分析的公司竞争力、创新和经济政策不确定性等新变量. 梳理上述文献研究脉络，揭示文本分析技术的优缺点，并且指出在会计和金融领域应用文本大数据技术的研究面临的挑战和机遇.

关键词: 可读性; 信息; 欺诈; 创新; 经济政策不确定性
中图分类号: F713.36 **文献标识码**: A **文章编号**: 1007 - 9807(2020)09 - 0019 - 12

0 引 言

中共十九大报告提出“推动互联网、大数据、人工智能和实体经济深度融合”，说明大数据研究契合国家经济发展战略，具有重大意义. 大数据③多表现出非结构化特征，要求新的数据处理技术(例如机器学习)，能够产生新的发现. 大数据包括文本、音频、图像和视频等类型. 过去十年，文本大数据对金融和会计研究产生巨大影响. 对此梳理相关研究脉络，揭示最新研究动态; 比较文本分析的优缺点并展望未来研究方向，提供新的研究视角并激发研究思路.

文本分析是计算语言学、自然语言处理、信息恢复、内容分析等领域的交叉学科. 应用文本分析进行会计和金融研究的主要步骤如下: 采集文档，解析文档，文本挖掘，指标构建，计量分析. 由于通常没有现成的文本大数据可用，研究者一般需要开发网络爬虫采集原始文档. 解析文档之前可能需要对原始文档进行格式转换，例如中国上市公司年报是 PDF 格式，须转换为容易被计算机处理的文本格式. 解析文档主要是删除图形、表格、标签(例如 HTML 标签)和冗余标点符号等噪音从而提供干净文本. 正如 Loughran 和 McDonald [5] 指出，解析文档难度很大，是关键环节. 基于干净文本，采用词袋技术等文本挖掘工具，即可构建可读性(readability)、语调、文档相似性等指标，进而进行计量分析.

随着文本大数据的出现和计算语言学的发展，文本大数据分析成为会计和金融研究的重要

① 收稿日期: 2018 - 09 - 16; 修订日期: 2019 - 06 - 30.
基金项目: 上海市社科规划一般课题资助项目(2018BGL013); 国家自然科学基金资助项目(71401155; 71573220; 71773123); 上海市教育委员会“人工智能时代会计硕士教育创新资助项目(2019)”.
通讯作者: 张顺明(1966—)，男，湖北广水人，博士，教授，博士生导师. Email: szhang@ruc.edu.cn
② 按照 Li [1]，非结构化是指数据没有预先定义的数据模型，或(且)不能填入关系型表格. 非结构化文本天然是高维数据，通常模糊且不规则，难以直接被计算机处理，因而不能直接用于计量分析.
③ Gepp 等 [2] 指出大数据特征表现为 4V: Volume、Velocity、Variety、Veracity，分别指数据量大、产生速度快、格式或类型多、数据品质和相关性时变. 陈国青等 [3] 是大数据对管理的影响的综述，而 Zhu [4] 证明大数据能够起到公司治理的作用.

工具④. 根据研究内容,可将相关文献分为两类:第一类将文本分析用于信息品质和数量、信息披露和市场异象等问题的研究,第二类用于构建公司创新、竞争力等新指标.

Li[1] 是较早关于文本分析方面的研究综述,但该文限于公司信息披露方面的研究. Lourghran 和 McDonald[7] 范围更广,但该文以研究方法为主线,且未包含许多金融和会计重要领域的研究,特别地,该文并没有包含中文文本分析的研究. 沈艳等[8] 虽然包含了中文文献,但按照学科梳理文献,与以下按照研究内容的视角完全不同. 根据是否与信息相关来梳理文献,这一新视角是对此前文献综述的发展. 同时,相比此前文本分析的文献综述多以英文文献为主,此处兼顾中英文文献的综述更全面.

学术贡献为:1) 总结了应用文本分析研究会计和金融问题的一般步骤;2) 理清了应用文本分析进行会计和金融研究的脉络:按照是否和信息相关,将文献归结为两类;3) 对比了文本分析的优势和缺点,有助于研究者全面认识文本分析并规避方法缺陷;4) 指出未来应用文本分析技术进行会计和金融研究的可能方向,为后续研究提供参考;5) 分析中文文本分析的难度和前景,有利于形成中国特色的会计和金融研究体系.

1 信息相关研究

金融和会计都与信息密切相关,因此有大量文献应用文本分析在信息品质、信息数量、信息披露和与信息相关的市场异象等方面开展研究.

1.1 信息品质——财务文档可读性

作为信息品质的一种度量,可读性反应了投资者获取文档中信息的难易程度,而这与公司信息披露、信息环境和市场对信息的反应密切相关. 文献中的可读性指标分为以下三类.

1.1.1 基于词句难度的可读性指标

Fog index 是语言学中度量文档可读性的指标,最先被 Li[9] 引入会计和金融研究. Fog index 是句子难度(一句话含词越多越难)和词语难度(一个词音节越多越难)之和,具体公式为

$$Fog = 0.4 \times (\text{平均单词个数} / \text{句} + \text{多于 2 个音节单词占比百分数}) \quad (1)$$

Fog index 数值代表第一遍阅读一篇文档时,一个人需要几年的教育才能读懂该文. 例如,如果一篇文档的 Fog index 是 5,意味着至少 5 年的教育才能使一个人在第一遍阅读时能读懂该文. 由于这个指标提出较早,并且适用于大规模文本分析,因此这个指标在财务文档可读性研究中应用广泛. 和 Fog index 类似,Flesch index 将 Fog index 中的第二项换成单词的音节数,而 Flesch-Kincaid index 则是将指标变化范围调整到 0 ~ 100.

与 Fog index 类似,丘心颖等[10] 用笔画数刻画汉字的难度(年报汉字平均笔画数越大,可读性越差),结合句子难度构造了中文年报可读性(复杂性),发现年报可读性越差的公司随后被更多分析师跟踪. 王克敏等[11] 从文本逻辑和字词的复杂性两个角度刻画中文年报文本信息复杂性,发现管理者会操纵年报文本信息复杂性.

作为最早被引入会计研究的可读性指标,Fog index 为基础的研究延伸到很多领域. Dyer 等[12] 用 Fog index 度量可读性,揭示了 1996 年 ~ 2013 年间美国上市公司年报可读性下降的趋势. Bozanic 等[13] 用 Fog index 研究律师对信息披露的影响.

Li[9] 用 Fog index 度量年报可读性,检验“管理层混淆假说”(managers obfuscation hypothesis),发现盈余越低的公司年报可读性越差(Fog index 越大),这是因为管理层为降低市场反应而故意模糊信息. 而业绩好的公司年报不存在这样的情况,因此其盈余容易持续. Lo 等[14] 发现操纵当年盈余高于上一年的公司年报的管理层讨论和分析部分(MD&A)可读性变差. Lo 等[14] 表明盈余水平和 MD&A 部分的可读性负相关,而 Li[9] 表明盈余水平和年报全文的可读性正相关,这说明年报全文和其中 MD&A 部分的可读性受到管理层区别对待,也表明区分年报不同部分的可读性值得研究,可能的原因是管理层对年报不同部分的关注程度不同.

同样是结合语言学和会计金融的研究,Kim 等[15] 发现,体现将来时态程度越高的语言的国家

④ Teoh[6] 对会计研究可用的新数据(不限于文本)进行了分类总结,提出了数据和计量方法的挑战和机遇.

更容易出现盈余操纵,其逻辑是体现将来时态程度这种语法特征越明显,语言中越容易明确表明事件发生的未来时间,则公司高管越能察觉到盈余操纵的后果. 这是第一个语言时态和盈余操纵方面的研究.

可读性影响投资者交易和市场行为方面,Miller[16] 采用 Fog index 度量可读性,发现年报可读性会导致年报发布期间更多的小额投资者交易活动,Lawrence[17] 也发现散户更可能投资年报可读性好、年报短的公司. Rennkamp[18] 在实验中研究可读性,发现可读性会导致散户反应过度,不精明的投资者尤其如此. You 和 Zhang[19] 发现年报词数越多,年报发布后 12 个月内市场反应不足越明显. Lehavy 等[20] 发现年报可读性越差,随后跟踪该公司的分析师人数越多,分析师预测分歧程度越高,分析师预测准确度越低. De Franco 等[21] 证明,采用 Fog index、Flesch-Kincaid、Flesch Reading Ease 度量的分析师研究报告的可读性和交易量正相关.

上述研究表明,强制性财务信息披露的复杂性(可读性的反义词)会恶化公司的信息环境(交易量降低、分析师预测分歧增加和分析师预测准确度降低等). 如果公司高管在信息披露时故意降低可读性来隐藏信息,那么可读性差的公司减少(或者不改变)自愿信息披露. 然而,如果较低可读性源于公司自身的商业复杂性或者信息披露准则,那么公司高管可能会增加自愿信息披露来减弱信息环境恶化. Guay 等[22] 证明,为了降低强制性信息披露带来的信息环境恶化,公司高管确实会增加自愿信息披露. Lundholm 等[23] 发现,在美国交叉上市的公司倾向于发布可读性更好的文档. Li[24] 发现高管对未来越乐观,MD&A 的可读性越好. Biddle 等[25] 用 Fog index 度量财务信息披露品质,发现信息披露品质能够提高资本投资效率.

1.1.2　文件大小作为可读性指标

虽然 Fog index 被大量文献用于度量财务文档可读性,但是 Loughran 和 McDonald[5] 认为,该指标的第二项,也就是单词难度用多音节来度量并不合理,因为多音节单词对于商业领域并不一定很难. 例如,"company"虽然是个多音节单词,但是这个词对大多数市场参与者来说却很熟悉. 而且他们发现 Fog index 并不能解释分析师预测分歧程度和意外盈余. 因此 Loughran 和 McDonald[5] 将可读性定义为投资者或者分析师从一篇财务披露文档中吸收价值相关信息的难度,并且提出文件大小(file size)作为可读性指标. 文件大小指的是一篇财务披露文档所占字节数. 文件大小作为可读性度量最大的好处是不需要解析财务文档,因此该指标客观,可重复. 更重要的是,Loughran 和 McDonald[5] 发现,年报文件越大(可读性越差),随后公司特质波动率越高,分析师对盈余预测分歧程度越高,意外盈余绝对值越大,也就是说文件大小能够度量可读性. Ertugrul 等[26] 也发现公司年报文件大小度量的可读性越差,外部融资成本越高. Dyer 等[12] 用 LDA(latent dirichlet allocation)技术⑤发现 FASB 和 SEC 要求导致的公允价值、内部控制和风险因素这三个方面是美国上市公司年报长度增加的主要原因. 马长峰等[30] 发现中文年报文件大小不能预测波动率,但摘要的文件大小能负向预测波动率.

1.1.3　基于平实英语的可读性指标

虽然 Loughran 和 McDonald[5] 认为文件大小作为可读性指标易得、准确客观且重复性高,但是 Bonsall 等[31] 却指出,文件大小的时变经常是因为和文本无关的因素(例如 HTML、XML 和 PDF 等相关内容的加入),进而提出了一种基于平实英语(plain English)的可读性指标,也就是 Bog index. 该文认为,可读性应该从美国证监会在 1998 年推出的平实英语要求出发,度量信息披露文档表达是否清楚. 而 Fog index 指标蕴含的句子长度和单词难度仅仅是平实英语的一部分(其他还有语态、动词、俚语、专业术语、抽象词汇、冗余词汇和过度细节等方面),且词语难度用音节数度量并不合理. Bog index 优势在于: 第一,根据几乎全部平实英语特征构造指标; 第二,以一个 20 万词汇列表为基础根据每个单词的熟悉程度打分,从而更真实度量词语难度,也改进了 Fog index 中词语难度度量的问题. 的确,该文发现只有 Bog index

⑤　Jegadeesh 和 Wu[27]、Hoberg 和 Lewis[28] 和 Ganglmair 和 Wardlaw[29] 等文献也使用了 LDA 技术.

指标在平实英语监管要求出台前后显著变化，而其他可读性指标都不能捕捉这一监管规则变化. Bonsall 和 Miller[32] 采用 Bog index 度量可读性，发现财务信息披露文档可读性越差，导致公司债券评级越低(违约风险越高)，评级机构之间分歧程度越高，债务资本成本越高. Bonsall 等[31] 也发现年报公布之前操纵盈余的公司 MD&A 可读性变差. Li[9] 是用可读性(Fog index)预测盈余品质，而 Bonsall 等[31] 却发现盈余品质同样预测年报 MD&A 部分的可读性(Bog index). Asay 等[33] 用基于平实英语的可读性证明高管会操纵信息披露中的表达方式.

需要指出，文档可读性(所含信息被理解的难度)很难和公司从事商业活动本身的复杂性分离，因此可读性的某些影响可能是可读性和复杂性共同起作用的结果.

1.2 文本的信息含量

可读性借鉴计算语言学，刻画了文档的一种语言特征，而这种特征影响公司与市场参与者之间的信息交换. 那么，除语言特征之外，文档中的文本是否含有信息呢？许多文献发现，除了数字报表，文本也含有信息，能够对市场和投资者产生影响. 因此文本信息是相对于财务报表的增量信息. Henry[34] 发现盈余新闻稿语调影响投资者反应.

Li 和 Ramesh[35] 发现季报的市场反应限于首次发布，而年报的市场反应则限于季. Loughran 和 McDonald[36] 改进了用于构造语调的负面词汇列表，发现年报全文而不是 MD&A 部分的负面词汇出现频率能预测年报公布后股票的价格、交易量、波动率等变量，并且能预测高管欺诈. 只分析 MD&A 部分的文本的文献也发现了文本含有信息的有力证据. 例如，Feldman 等[37] 发现年报和季报 MD&A 部分语调变化引起短期市场反应. Li[24] 使用朴素贝叶斯机器学习方法，发现 MD&A 中高管陈述越乐观，公司未来的盈余和流动性就越好. Brown 和 Tucker[38] 发现 MD&A 变化引起市场反应，但并不影响分析师预测的修改，因此分析师并不用 MD&A 的信息. Kothari 等[39] 研究了公司报告、分析师报告和商业新闻等方面的披露文本，发现信息披露语调乐观(悲观)伴随着公司风险降低(升高)，其中风险包括资本成本、波动率和分析师预测分歧程度.

中文文本信息含量的研究成果非常丰富. 赵子夜等[40] 研究管理层报告的样板化及其经济后果. 谢德仁和林乐[41] 发现管理层净正面语调与公司来年业绩显著正相关. 林乐和谢德仁[42] 证明业绩说明会管理层语调能预测分析师行为. 孟庆斌等[43] 发现 MD&A 信息含量越高未来股价崩盘风险越低. 薛爽等[44] 发现亏损公司 MD&A 中提及的外部或内部原因越多，下一年扭亏的可能性越小；经营计划中提及的战略性改进措施越多，则下一年度扭亏的可能性越大；当下一年度计划增加研发支出时，会提高扭亏的概率.

De Franco 等[45] 研究了分析师报告中的股东和债权人利益冲突事件，发现损害债权人利益事件的分析引起信用利差增大，债券交易量增加. Huang 等[46] 发现分析师的正面文本比负面文本更强烈引起投资者反应，分析师报告文本能预测未来 5 年的盈余增长. 林乐和谢德仁[42] 发现管理层净正面语调提高了分析师更新其荐股报告的可能性及更新人数比例，并会提高分析师荐股评级水平及其变动.

Price 等[47] 发现季度业绩说明会中高管在问答中的语调为正(负)，则随后 3 天和 2 个月的股价上升(下降). Blau 等[48] 发现业绩说明会语调为正则卖空交易活动会降低，卖空对收益率预测能力变强. Borochin 等[49] 发现季度业绩说明会语调和期权市场蕴含的不确定性负相关. Doran 等[50] 研究了 REITs 的季度业绩说明会，发现业绩说明会语调能很好地解释季度盈余公告当日和随后的超额收益率. Doran 等[50] 分析了盈余公告相关的业绩说明会，发现公司高管个人(而非集体)的乐观能解释业绩说明会乐观语调(在控制了公司商业因素之后).

Tetlock[51] 发现新闻媒体的悲观引发股价下跌，但随后翻转，同时过高或过低的悲观导致交易量增大. 这表明新闻媒体并未体现基本面信息而是伴随着流动性交易和噪音交易. Tetlock 等[52] 发现公司新闻中负面词汇出现频率较高能预测未来盈余较低，虽然股价短期对此反应不足，但是很快就将文本包含的基本面信息反应出来.

1.3 公司高管策略性信息披露行为

Bozanic 等[53] 发现季度盈余披露中关于盈余的量化陈述和其他陈述都引起投资者和分析师反

应,但高不确定性使高管增加其他陈述. Arslan-Ayaydin 等[54] 发现更多的股权激励会让高管在盈余披露新闻稿中语调变得更积极. 程新生等[55] 证明进行盈余重述的公司会在 MD&A 中披露更多的非财务信息.

也有文献研究了文本信息和诉讼风险之间的关系. Levy 等[56] 发现非董事首席财务官(CFO)在业绩说明会中的语调比董事 CFO 更悲观,会更早且更保守披露坏消息. 这种现象的原因是 CFO 想规避自己(而不是公司)被诉讼的风险. Rogers 等[57] 发现乐观的陈述容易引起股东被控告,并且被告公司的盈余公告更乐观.

Hanley 和 Hoberg[58] 研究美国 1933 年证券法第 11 款规定的诉讼风险和 IPO 折价以及自愿信息披露之间的关系⑥. 在 IPO 之前的询价和路演过程中,出现好消息时,信息披露成本高(可能为竞争对手提供信息),上市公司倾向于用折价来规避诉讼风险; 出现坏消息时,信息披露成本低,不披露风险高,公司通过全面披露规避诉讼风险. 该文通过文本分析技术揭示,除折价之外,策略性信息披露是上市公司规避诉讼风险的另一个策略,并且 IPO 信息披露策略和正常时期相反.

1.4　文本分析和财务欺诈

财务欺诈指公司未如实披露法定信息而欺骗股东的行为,是监管者、投资者和审计师都重视的问题. 而 Amani 和 Fadlalla[59] 指出,财务欺诈是最受益于数据挖掘等大数据技术的财会领域. Loughran 和 McDonald[36] 用财务诉讼词汇占比、负面词汇占比和不确定性词汇占比预测财务报告欺诈. 而 Purda 和 Skillicorn[60] 则用决策树模型找出最能区分欺诈和真实报告的有序单词列表,基于上述列表中的前 200 个单词,采用支持向量机(SVM)技术,对每一篇财务报告文档标注真实概率,从而预测财务报告的欺诈可能性.

Gray 和 Debreceny[61] 对数据挖掘(包括文本挖掘)在财务欺诈方面的研究进行了综述,并且对欺诈类型提出了一个分类方法. Glancy 和 Yadav[62] 发展了一种财务报告欺诈探测计算模型. Dilla 和 Raschke[63] 从理论上分析了文本等数据可视化在欺诈交易中的应用. West 和 Bhattacharya[64] 对商业智能为基础的财务欺诈探测技术的探测算法、欺诈类型和效果等方面进行了综述,其中包括文本挖掘技术. Lin 等[65] 比较了不同技术在探测财务欺诈方面的效果,发现人工神经网络、决策树这两种方法优于 Logistics 回归. Cecchini 等[66] 则发现结合文本分析和财务数字预测欺诈能力比其中任何单一技术效果都好.

还有文献分析了欺诈信息披露的特征. Goel 等[67] 发现欺诈年报比非欺诈年报使用更多被动语态句子、不确定性词汇和词典. Humpherys 等[68] 发现欺诈性披露比非欺诈披露使用更多煽动性语言和词汇,貌似可信实则没有实质性内容. Hoberg 和 Lewis[28] 发现欺诈公司年报的 MD&A 过少解释公司绩效来源、过多披露美化公司绩效的信息.

1.5　文本信息和市场异象

许多金融市场异象和信息有关. 由于文本信息比传统金融和会计数据更不明确,因此更难被投资者处理. 如果市场异象来自信息处理,那么研究者应该考虑用文本信息解释市场异象.

You 和 Zhang[19] 发现年报单词个数过多导致市场反应不足,探讨市场对文本信息反应不足是否 PEAD 这一异象的成因. Feldman 等[37] 则分析了年报和季报 MD&A 中的语调,发现语调变化能预测随后季度的意外盈余和价格漂移. Lee[69] 分析了季报可读性是否影响了股价的有效性,发现季报越长(可读性越差),季报公告后三天股价反应的盈余相关信息越少,同时发现季报可读性差伴随着信息不对称.

对应计异象(accrual anomaly)这个问题,Li[24] 发现,如果应计项为正(负)而公司高管在 MD&A 中对于应计项的语调却是负(正),那么应计异象消失,就是说应计项和未来收益率不再相关.

文本分析也被用于 IPO 定价研究. Hanley 和 Hoberg[70] 利用文本分析技术,将 IPO 招股说明书能被刚刚发生的 IPO 或者同行业 IPO 解释的信息作为标准分量,不能被解释的部分作为信息分量.

⑥ 美国 1933 年证券法第 11 款规定,由于 IPO 招股说明书材料披露不足导致股价低于发行价产生损失的情况下,投资者可以起诉承销商和发行者. 因此,发行者和承销商只要能避免股价低于发行价或者全面披露中的一条即可规避这种诉讼风险.

该文发现信息分量(标准分量)越大,定价准确度越高(低),折价越低(高).其原因在于信息分量减少了投资者在询价中生产信息的成本. Arnold 等[71]研究 IPO 招股说明书⑦中的风险因素部分,将风险因素中的词数相对于总词数(或者特定内容词数)之比作为不确定性,发现不确定性和 IPO 首日收益率正相关. Loughran 和 McDonald[72]研究美国 IPO 过程中的 S-1 表格,发现这个文件中不确定性词汇占比越高,首日收益率越高、发行价修正绝对值越大,随后波动率越大. Bajo 和 Raimondo[73]发现公司在 IPO 之前的新闻报道正面语调伴随 IPO 折价,且这种效应随着临近 IPO 日期更加明显.

2 文本分析产生新指标和新变量

文本分析将研究对象从结构化数据拓展到非结构化文本数据,因此可以对原有变量构建新的度量指标,或者直接构造新的变量.

2.1 财务约束的度量

Bodnaruk 等[74]用上市公司年报中和"约束"相关的词汇频率作为财务约束的度量,发现这个基于文本分析的财务约束指标能预测股利缺失或增加、股权回收(equity recycling)和养老金不足等流动性事件,优于传统的财务约束指标.借助于年报中的负面词汇度量财务约束,发现财务约束导致公司追求更为激进的税务策略,包括更高的未确认税收抵扣、更低的有效税率、税收天堂利用的增加和更高的审计调整. Hoberg 和 Maksimovic[75]也用文本分析构建财务约束指标,Buehlmaier 和 Whited[76]用机器学习构造财务约束指标.

2.2 创新的度量

Hoberg 等[77]采用 t 年上市公司 i 年报中"商业描述"的名词构造名词向量 $NV(i,t)$,将 $NV(i,t)$ 和 $NV(i,t-1)$ 之间的差异作为 t 年公司 i 的创新的度量,构造了一种基于文本的创新度量指标.同时,将 $\sum_{i=1}^{N_t} NV(i,t) - \sum_{i=1}^{N_{t-1}} NV(i,t-1)$ 作为市场整体指标 $D(t)$,其中 N_t 和 N_{t-1} 分别表示在 t 年和 $t-1$ 年的公司个数. $NV(i,t)$ 和 $D(t)$ 之间的相似度就是 t 年上市公司 i 的另一种创新指标.

Bellstam 等[78]收集分析师关于上市公司的研究报告,利用 LDA 技术将所有上市公司研究报告的词汇分为 15 个主题,从中选出和主流创新教科书词频最接近的一个主题作为创新主题,然后通过个股研究报告中含有的创新主题词汇的强度来度量公司创新.

2.3 竞争力指标

Li 等[79]通过文本分析技术构造了一个新的公司层面的竞争指标.之前的竞争性指标大多采用 Herfendahl index 和四企业集中度(four-firm concentration ratio),是行业层面的竞争性指标.然而,同一行业内的公司之间的竞争性必然存在差异,而行业竞争性指标显然不能度量公司层面竞争力.该文通过公司年报中提到竞争对手的频率来度量公司的竞争力.

2.4 知识的度量

Li 等[80]分析上市公司业绩说明会的文本记录,通过发言内容来揭示高管发言人熟悉公司的哪些情况(知识),并且发现知识能够影响高管薪酬.

2.5 经济政策不确定性(EPU)

Baker 等[81]将主流报纸包含经济、政策和不确定性词根的文章数对文章总数的占比作为经济政策不确定性指标,发现 EPU 提高股市波动率,抑制投资.陈国进等[82]和雷立坤等[83]采用 EPU 研究中国市场.

2.6 产品相似性

Hoberg 和 Phillips[84]基于上市公司年报产品描述部分的词汇,构建了不同公司之间的产品相似性(差异性)指标,发现产品相似性促成并购,并且能够提升并购后的公司产品独特性(新产品).基于文本分析的相似性可以跨行业比较两个公司的产品相似程度,而行业代码却不能实现这一点.

3 机遇和挑战

3.1 文本分析的优势和缺点

文本分析的优势在于,第一,提供了文本形式

⑦ 美国 IPO 招股说明书包括:概要、风险因素、募资用途和 MD&A.

的非结构化数据,丰富了数据类型,从而拓展了研究对象和研究范围; 第二,文本大数据拓展了原来的研究边界,例如引入语言学开展可读性研究; 第三,提供新的工具、变量和指标; 第四,提供新的研究视角,例如用文本信息研究 IPO 定价.

同时,文本分析也有明显的缺点: 第一,文本信息本身并不明确,必须经过研究者加工处理才能用于计量分析,而这会引入噪音甚至错误,同时数据处理的可重复性存疑; 第二,文本数据大多数缺乏权威来源,数据来源存疑; 第三,文本数据量很大,现有的计量分析方法并不一定适用; 第四、应用文本大数据分析技术研究会计和金融问题,对研究者的综合能力尤其是编程和数量分析能力提出了挑战. 最后,文本既可能含有传统财务数字没有的信息,也可能是管理层操纵文本的表现,这是应用文本分析研究会计和金融学问题的一大挑战.

3.2　研究方向展望

第一,既要重视数据,也要重视算法. 未来文本数据量进一步加大,这就要求研究者必须加强两个方面的技能: 1) 人工智能算法及其实现,尤其是机器学习和深度学习; 2) 基于大数据的计量分析方法.

第二,中文文本分析的研究空间很大. 现有研究大部分针对英文文本,一个自然的借鉴是将英文研究方法用于中文文档. 但是,英文和中文是两种不同的语言: 英文天生用空格分隔词汇,而中文则没有词汇分隔符. 这就导致中文分词比英文分词难得多. 幸运的是,Python 中已经有结巴这一模块,可以进行中文分词、词性标注等. 虽然如此,中文分词仍然不如英文分词准确,因为中文分词基于自然语言处理技术,本质上并不精确. 同时,大部分中文财务文档是 PDF 格式,不能直接用计算机处理. 虽然如此,应用中文文本分析的金融和会计研究仍然大有前途. 原因有二: 1) 相对于丰富的英文文本分析的研究成果,中文相关研究成果明显不足; 2) 也许更重要的是,这对于形成有中国特色的会计和金融研究体系极具价值.

第三,结合中国特有的信息披露规则,利用文本分析挖掘具有中国特色的会计和金融问题. 虽然国际会计准则已经存在,但不同国家的信息披露规则并不相同. 从不同的法规环境出发,可能找到中国特色的研究问题,甚至能够研究其他国家地区不能研究的议题,例如马长峰等[50] 对年报摘要文本的分析.

第四,内生性问题. 近年来,寻找天然实验是一种解决内生性问题的好方法. 而中国作为发展中国家,法律法规经常变动,因此在这些变动中发掘天然实验,有利于解决经验分析中的识别问题. 例如,公司信息透明性是否降低资本成本? 另外,现场实验、工具变量、断点回归等方法也需要关注.

第五,词典的构建. 现有文献表明,在商科研究中直接采用语言学通用词典并不合适,因此需要商科研究中的专用词典. 不论对于公司年报可读性还是情绪分析都是如此. 同时,中文不同于英文,因此有必要构建中文专用词典.

第六,分析师受到可读性影响的原因探讨. 作为专业信息解读者,分析师被指出未能发挥专业信息解读能力. 这种现象的原因可从行为偏差、制度安排和激励机制等方面分析.

第七,文本分析是语言经济学的重要工具. 近年来语言经济学指出语言特征影响人的经济行为⑧,而文字包含了语言特征,因此通过文本分析技术挖掘语言特征是发展语言经济学研究的重要手段.

4　结束语

过去十年,学术界应用文本大数据分析技术在信息品质、信息数量的度量,信息披露,市场异象和资产定价等方面取得了大量研究成果; 产生了创新、竞争力、实际权力、经济政策不确定性和产品相似性等新指标和新变量. 结合大数据处理方法和大样本计量分析方法,未来文本大数据在会计和金融领域将会取得更大研究进展.

一方面,基于英文的文本分析必将进一步深化金融和会计问题的研究; 另一方面,基于中文文

⑧　程博和潘飞[85] 分析了语言多样性对分析师盈余预测质量的影响.

本分析的会计和金融研究才刚刚起步. 更重要的是,加强基于中文文本分析的会计和金融研究对于形成有中国特色的会计和金融研究体系具有重要意义,这正是中国学者的使命.

参 考 文 献:

[1] Li Feng. Textual analysis of corporate disclosure: A survey of the literature [J]. Journal of Accounting Literature, 2010, 29: 143 – 165.

[2] Gepp A, Linnenluecke M K, Smith T. Big data techniques in auditing research and practice: Current trends and future opportunities [J]. Journal of Accounting and Literature, 2018, 49: 102 – 115.

[3] 陈国青, 吴 刚, 顾远东, 等. 管理决策情境下大数据驱动的研究和应用挑战——范式转变与研究方向 [J]. 管理科学学报, 2018, 21(7): 1 – 10.
Chen Guoqing, Wu Gang, Gu Yuandong, et al. The challenges for big data driven research and applications in the context of managerial decision-making: Paradigm shift and research directions [J]. Journal of Management Sciences in China, 2018, 21(7): 1 – 10. (in Chinese)

[4] Zhu C. Big data as a governance mechanism [J]. The Review of Financial Studies, 2019, 32(5): 2021 – 2061.

[5] Loughran T, McDonald B. Measuring readability in financial disclosures [J]. The Journal of Finance, 2014, 69(4): 1643 – 1671.

[6] Teoh S H. The promise and challenges of new datasets for accounting research [J]. Accounting, Organizations and Society, 2018, 68 – 69: 109 – 117.

[7] Lourghran T, McDonald B. Textual analysis in accounting and finance: A survey [J]. Journal of Accounting Research, 2016, 54(4): 1187 – 1230.

[8] 沈 艳, 陈 赟, 黄 卓. 文本大数据分析在经济学和金融学中的应用: 一个文献综述 [J]. 经济学(季刊), 2019, 18(4): 1153 – 1186.
Shen Yan, Chen Yun, Huang Zhuo. A literature review of textual analysis in economic and financial research [J]. China Economic Quarterly, 2019, 18(4): 1153 – 1186. (in Chinese)

[9] Li Feng. Annual report readability, current earnings, and earnings persistence [J]. Journal of Accounting and Economics, 2008, 45(2 – 3): 221 – 247.

[10] 丘心颖, 郑小翠, 邓可斌. 分析师能有效发挥专业解读信息的作用吗? ——基于汉字年报复杂性指标的研究 [J]. 经济学(季刊), 2016, 15(4): 1483 – 1506.
Qiu Xinying, Zheng Xiaocui, Deng Kebin. Can analysts play an effecttive role in professional information interpretation: Evidence based on a complexity/readability index of Chinese corporate annual reports [J]. China Economic Quarterly, 2016, 15(4): 1483 – 1506. (in Chinese)

[11] 王克敏, 王华杰, 李栋栋, 等. 年报文本信息复杂性与管理者自利——来自中国上市公司的证据 [J]. 管理世界, 2018, (12): 120 – 132.
Wang Kemin, Wang Huajie, Li Dongdong, et al. The complexity of the annual report text information and self-interest of managers: Evidence from Chinese listed firms [J]. Management World, 2018, (12): 120 – 132. (in Chinese)

[12] Dyer T, Lang M, Stice-Lawrence L. The evolution of 10-K textual disclosure: Evidence from latent dirichlet allocation [J]. Journal of Accounting and Economics, 2017, 64(2): 221 – 245.

[13] Bozanic Z, Choudhary P, Merkley K. Securities law expertise and corporate disclosure [J]. The Accounting Review, 2019, 94(4): 141 – 172.

[14] Lo K, Ramos F, Rogo R. Earnings management and annual report readability [J]. Journal of Accounting and Economics, 2017, 63(1): 1 – 25.

[15] Kim J, Kim Y, Zhou Jian. Languages and earnings management [J]. Journal of Accounting and Economics, 2017, 63(2): 288 – 306.

[16] Miller B P. The effects of reporting complexity on small and large investor trading [J]. The Accounting Review, 2010, 85

(6)：2107 – 2143.

[17] Lawrence A. Individual investors and financial disclosure [J]. Journal of Accounting and Economics，2013，56(1)：130 – 147.

[18] Rennkamp K. Processing fluency and investors' reactions to disclosure readability [J]. Journal of Accounting Research，2012，50(5)：1319 – 1354.

[19] You Haifeng，Zhang Xiaojun. Financial reporting complexity and investor underreaction to 10-K information [J]. Review of Accounting Studies，2009，14(4)：559 – 586.

[20] Lehavy R，Li Feng，Merkley K. The effect of annual report readability on analyst following and the properties of their earnings forecasts [J]. The Accounting Review，2011，86(3)：1087 – 1115.

[21] De Franco G，Hope O K，Vyas D，et al. Analyst report readability [J]. Contemporary Accounting Research，2015，32(1)：76 – 104.

[22] Guay W，Samuels D，Taylor D. Guiding through the fog：Financial statement complexity and voluntary disclosure [J]. Journal of Accounting and Economics，2016，62(2)：234 – 269.

[23] Lundholm R J，Rogo R，Zhang J L. Restoring the tower of babel：How foreign firms communicate with U. S. investors [J]. The Accounting Review，2014，89(4)：1453 – 1485.

[24] Li Feng. The information content of forward-looking statements in corporate filings：A Naïve Bayesian machine learning approach [J]. Journal of Accounting Research，2010，48(5)：1049 – 1102.

[25] Biddle G C，Hilary G，Verdi R S. How does financial reporting quality relate to investment efficiency? [J]. Journal of Accounting and Economics，2009，48(2)：112 – 131.

[26] Ertugrul M，Lei Jin，Qiu Jiaping，et al. Annual report readability，tone ambiguity，and the cost of borrowing [J]. Journal of Financial and Quantitative Analysis，2017，52(2)：811 – 836.

[27] Jegadeesh N，Wu Di. Deciphering Fedspeak：The Information Content of FOMC Meetings [R]. Working Paper，Ann Arbor：University of Michigan，2017.

[28] Hoberg G，Lewis C. Do fraudulent firms produce abnormal disclosure? [J]. Journal of Corporate Finance，2017，43：58 – 85.

[29] Ganglmair B，Wardlaw M. Complexity，Standardization，and the Design of Loan Agreements [R]. Working Paper，Texas：University of Texas at Dallas，2017.

[30] **马长峰，陈志娟，张顺明. 文件大小能度量中文年报可读性吗? [Z]**. Working Paper，2018.
Ma Changfeng，Chen Zhijuan，Zhang Shunming. Can File Size Measure Chinese Firm's Annual Report's Readability? [Z]. Working Paper，2018.(in Chinese)

[31] Bonsall S B，Leone A J，Miller B P，et al. A plain English measure of financial reporting readability [J]. Journal of Accounting and Economics，2017，63(2 – 3)：329 – 357.

[32] Bonsall S B，Miller B P. The impact of narrative disclosure readability on bond ratings and the cost of debt [J]. Review of Accounting Studies，2017，22(2)：608 – 643.

[33] Asay H S，Libby R，Rennekamp K. Firm performance，reporting goals，and language choices in narrative disclosures [J]. Journal of Accounting and Economics，2018，65(2 – 3)：380 – 398.

[34] Henry E. Are investors influenced by how earnings press releases are written? [J]. The Journal of Business Communication，2008，45(4)：363 – 407.

[35] Li E X，Ramesh K. Market reaction surrounding the filing of periodic SEC reports [J]. The Accounting Review，2009，84(4)：1171 – 1208.

[36] Loughran T，McDonald B. When is a liability not a liability? Textual analysis，dictionaries，and 10-Ks [J]. The Journal of Finance，2011，66(1)：35 – 65.

[37] Feldman R，Govindaraj S，Livnat J，et al. Management's tone change，post earnings announcement drift and accruals [J]. Review of Accounting Studies，2010，15(4)：915 – 953.

[38] Brown S V，Tucker J W. Large-sample evidence on firms' year-over-year MD&A modifications [J]. Journal of Accounting Research，2010，49(2)：309 – 346.

[39] Kothari S P, Li Xu, Short J E. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis [J]. The Accounting Review, 2009, 84(5): 1639 – 1670.

[40] 赵子夜，杨 庆，杨 楠. 言多必失？管理层报告的样板化及其经济后果 [J]. 管理科学学报，2019，22(3)：53 – 70.

Zhao Ziye, Yang Qing, Yang Nan. The less said the better? Economic consequences of textual similarity in management discussion and analysis [J]. Journal of Management Sciences in China, 2019, 22(3): 53 – 70. (in Chinese)

[41] 谢德仁，林 乐. 管理层语调能预示公司未来业绩吗？——基于我国上市公司年度业绩说明会的文本分析 [J]. 会计研究，2015，(2)：20 – 27.

Xie Deren, Lin Le. Do management tones help to forecast firms' future performance? A textual analysis based on annual earnings communication conferences of listed companies in China [J]. Accounting Research, 2015, (2): 20 – 27. (in Chinese)

[42] 林 乐，谢德仁. 分析师荐股更新利用管理层语调吗？——基于业绩说明会的文本分析 [J]. 管理世界，2017，(11)：125 – 145 + 188.

Lin Le, Xie Deren. Does analysts recommendation updating use the tone of management? A textual analysis based on annual earnings communication conferences [J]. Management World, 2017, (11): 125 – 145 + 188. (in Chinese)

[43] 孟庆斌，杨俊华，鲁 冰. 管理层讨论与分析披露的信息含量与股价崩盘风险——基于文本向量化方法的研究 [J]. 中国工业经济，2017，(12)：132 – 150.

Meng Qingbin, Yang Junhua, Lu Bing. The informative content of management discussion and analysis and stock price crash risk: Based on text vectorization method [J]. China Industry Economy, 2017, (12): 132 – 150. (in Chinese)

[44] 薛 爽，肖泽忠，潘妙丽. 管理层讨论与分析是否提供了有用信息？——基于亏损上市公司的实证探索 [J]. 管理世界，2010，(5)：130 – 140.

Xue Shuang, Xiao Zezhong, Pan Miaoli. Does MD&A provide information? An empirical investigation on losing firms [J]. Management World, 2010, (5): 130 – 140. (in Chinese)

[45] De Franco G, Vasvari F P, Vyas D, et al. Debt analysts' views of debt-equity conflicts of interest [J]. The Accounting Review, 2013, 89(2): 571 – 604.

[46] Huang A H, Zang A Y, Zheng R. Evidence on the information content of text in analyst reports [J]. The Accounting Review, 2014, 89(6): 2151 – 2180.

[47] Price S M, Doran J S, Peterson D R, et al. Earnings conference calls and stock returns: The incremental informativeness of textual tone [J]. Journal of Banking & Finance, 2012, 36(4): 992 – 1011.

[48] Blau B M, DeLisle J R, Price S M. Do sophisticated investors interpret earnings conference call tone differently than investors at large? Evidence from short sales [J]. Journal of Corporate Finance, 2015, 31: 203 – 219.

[49] Borochin P A, Cicon J E, DeLisle R J, et al. The effects of conference call tones on market perceptions of value uncertainty [J]. Journal of Financial Markets, 2018, 40: 75 – 91.

[50] Doran J S, Peterson D R, Price S M. Earnings conference call content and stock price: The case of REITs [J]. The Journal of Real Estate Finance and Economics, 2012, 45(2): 402 – 434.

[51] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market [J]. The Journal of Finance, 2007, 62(3): 1139 – 1168.

[52] Tetlock P C, Saar-Tsechansky M, Macskassy S. More than words: Quantifying language to measure firms' fundamentals [J]. The Journal of Finance, 2008, 63(3): 1437 – 1467.

[53] Bozanic Z, Roulstone D T, Van Buskirk A. Management earnings forecasts and other forward-looking statements [J]. Journal of Accounting and Economics, 2018, 65(1): 1 – 20.

[54] Arslan-Ayaydin Ö, Boudt K, Thewissen J. Managers set the tone: Equity incentives and the tone of earnings press releases [J]. Journal of Banking & Finance, 2016, 72: S132 – S147.

[55] 程新生，刘建梅，程 悦. 相得益彰抑或掩人耳目：盈余操纵与 MD&A 中非财务信息披露 [J]. 会计研究，2015，(8)：11 – 18 + 96.

Cheng Xinsheng, Liu Jianmei, Cheng Yue. A supplement or another lie: Earnings manipulation and non-financial information disclosure in MD&A [J]. Accounting Resarch, 2015, (8): 11 – 18 + 96. (in Chinese)

[56] Levy H, Shalev R, Zur E. The effect of CFO personal litigation risk on firms' disclosure and accounting choices [J]. Contemporary Accounting Research, 2017, 35(1): 434 – 463.

[57] Rogers J L, Van Buskirk A, Zechman S L C. Disclosure tone and shareholder litigation [J]. The Accounting Review, 2011, 86(6): 2155 – 2183.

[58] Hanley K W, Hoberg G. Litigation risk, strategic disclosure and the underpricing of initial public offerings [J]. Journal of Financial Economics, 2012, 103(2): 235 – 254.

[59] Amani F A, Fadlalla A M. Data mining applications in accounting: A review of the literature and organizing framework [J]. International Journal of Accounting Information Systems, 2017, 24: 32 – 58.

[60] Purda L, Skillicorn D. Accounting variables, deception, and a bag of words: Assessing the tools of fraud detection [J]. Contemporary Accounting Research, 2015, 32(3): 1193 – 1223.

[61] Gray G L, Debreceny R S. A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits [J]. International Journal of Accounting Information Systems, 2014, 15(4): 357 – 380.

[62] Glancy F H, Yadav S B. A computational model for financial reporting fraud detection [J]. Decision Support Systems, 2011, 50(3): 595 – 601.

[63] Dilla W N, Raschke R L. Data visualization for fraud detection: Practice implications and a call for future research [J]. International Journal of Accounting Information Systems, 2015, 16: 1 – 22.

[64] West J, Bhattacharya M. Intelligent financial fraud detection: A comprehensive review [J]. Computers & Security, 2016, 57: 47 – 66.

[65] Lin Chichen, Chiu Anan, Huang Shaioyan, et al. Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments [J]. Knowledge-Based Systems, 2015, 89: 459 – 470.

[66] Cecchini M, Aytug H, Koehler G J, et al. Making words work: Using financial text as a predictor of financial events [J]. Decision Support Systems, 2010, 50(1): 164 – 175.

[67] Goel S, Gangolly J, Faerman S R, et al. Can linguistic predictors detect fraudulent financial filings? [J]. Journal of Emerging Technologies in Accounting, 2010, 7(1): 25 – 46.

[68] Humpherys S L, Moffitt K C, Burns M B, et al. Identification of fraudulent financial statements using linguistic credibility analysis [J]. Decision Support Systems, 2011, 50(3): 585 – 594.

[69] Lee Y J. The effect of quarterly report readability on information efficiency of stock prices [J]. Contemporary Accounting Research, 2012, 29(4): 1137 – 1170.

[70] Hanley K W, Hoberg G. The information content of IPO prospectuses [J]. The Review of Financial Studies, 2010, 23(7): 2821 – 2864.

[71] Arnold T, Fishe R P H, North D. The effects of ambiguous information on initial and subsequent IPO returns [J]. Financial Management, 2010, 39(4): 1497 – 1519.

[72] Loughran T, McDonald B. IPO first-day returns, offer price revisions, volatility, and form S-1 language [J]. Journal of Financial Economics, 2013, 109(2): 307 – 326.

[73] Bajo E, Raimondo C. Media sentiment and IPO underpricing [J]. Journal of Corporate Finance, 2017, 46: 139 – 153.

[74] Bodnaruk A, Loughran T, McDonald B. Using 10-K text to gauge financial constraints [J]. Journal of Financial and Quantitative Analysis, 2015, 50(4): 623 – 646.

[75] Hoberg G, Maksimovic V. Redefining financial constraints: A text-based analysis [J]. Review of Financial Studies, 2015, 28(5): 1312 – 1352.

[76] Buehlmaier M M M, Whited T M. Are financial constraints priced? Evidence from textual analysis [J]. The Review of Financial Studies, 2018, 31(7): 2693 – 2728.

[77] Hoberg G, Phillips G, Prabhala N. Product market threats, payouts, and financial flexibility [J]. The Journal of Finance, 2013, 69(1): 293 – 324.

[78] Bellstam G, Bhagat S, Cookson J A. A Text-Based Analysis of Corporate Innovation [R]. Working Paper, 2019, https://

papers. ssrn. com/sol3/papers. cfm? abstract_id = 2803232.

[79] Li Feng, Lundholm R, Minnis M. A measure of competition based on 10-K filings [J]. Journal of Accounting Research, 2013, 51(2): 399 – 436.

[80] Li Feng, Minnis M, Nagar V, et al. Formal and Real Authority in Organizations: An Empirical Assessment [R]. Working Paper, Ann Arbor: University of Michigan, 2009.

[81] Baker S R, Bloom N, Davis S J. Measuring economic policy uncertainty [J]. The Quarterly Journal of Economics, 2016, 131(4): 1593 – 1636.

[82] 陈国进, 张润泽, 赵向琴. 经济政策不确定性与股票风险特征 [J]. 管理科学学报, 2018, 21(4): 1 – 27.
Chen Guojin, Zhang Runze, Zhao Xiangqin. Economic policy uncertainty and stock risk characteristics [J]. Journal of Management Sciences in China, 2018, 21(4): 1 – 27. (in Chinese)

[83] 雷立坤, 余 江, 魏 宇, 等. 经济政策不确定性与我国股市波动率预测研究 [J]. 管理科学学报, 2018, 21(6): 88 – 98.
Lei Likun, Yu Jiang, Wei Yu, et al. Forecasting volatility of Chinese stock market with economic policy uncertainty [J]. Journal of Management Sciences in China, 2018, 21(6): 88 – 98. (in Chinese)

[84] Hoberg G, Phillips G. Product market synergies and competition in mergers and acquisitions: A text-based analysis [J]. The Review of Financial Studies, 2010, 23(10): 3773 – 3811.

[85] 程 博, 潘 飞. 语言多样性、信息获取与分析师盈余预测质量 [J]. 管理科学学报, 2017, 20(4): 50 – 70.
Cheng Bo, Pan Fei. Linguistic diversity, information acquisition and the quality of analyst's earnings forecast [J]. Journal of Management Sciences in China, 2017, 20(4): 50 – 70. (in Chinese)

A survey on accounting and finance research based on textual big data analysis

MA Chang-feng[1], *CHEN Zhi-juan*[2], *ZHANG Shun-ming*[3*]

1. Digital Finance Research Center, Shanghai National Accounting Institute, Shanghai 201702, China;
2. School of Finance, Zhejiang Gongshang University, Hangzhou 310018, China;
3. School of Finance, Renmin University of China, Beijing 100872, China

Abstract: As a form of non-structural data, textual big data has been deeply influencing accounting and finance research in the recent decade. This influence manifests in two strands of literature. The first centers on information, such as applying textual analysis in measuring the quality (such as readability) and quantity of information (the information content of financial document), information disclosure and anomalies among other related issues. The second strand is mainly on applying textual analysis to constructing new indicators, such as competition, innovation and Economic Policy Uncertainty. This paper first surveys the two strands of literature and then analyzes the advantages and disadvantages of textual analysis. Lastly the opportunities and challenges in applying textual analysis in accounting and finance research are suggested.

Key words: readability; information; fraud; innovation; economic policy uncertainty