描述性创新，真实创新活动和盈余持续性

—基于文本分析和机器学习[1]

胡楠　张婷 李效宇

（西安交通大学管理学院 陕西 西安　710049）

摘要：本文运用中文文本分析和 Word2Vec 机器学习方法提取并分析了沪深 A 股上市公司年报中的描述性创新信息，研究发现：（1）描述性创新披露与当期真实创新活动水平（研发强度和专利数目）和当期盈余水平均正相关。（2）描述性创新可以正向预测未来的真实创新活动水平和盈余持续性。（3）进一步地，描述性创新内容中的非前瞻性信息含量和语调积极程度越高，未来真实创新活动和盈余持续性也越高。（4）2012 年公司年报披露准则的实施减弱了描述性创新披露对未来真实创新活动和盈余持续性的正向预测作用。本研究首次基于大样本深度文本分析，不仅为中国上市公司描述性创新披露的有效信息含量提供了经验证据，还为企业创新水平衡量体系的完善提供了全新的参考。

关键词：描述性创新　真实创新活动　盈余持续性　信息有用性　机器学习

[1] 本文为第十七届中国实证会计国际研讨会（2018 年 12 月）会议报告论文，文章题目和正文在会议报告版本的基础上略有改动。

Narrative Innovation, Real Innovative Activities and Earnings Persistence —Evidence from Text Analysis and Machine Learning

Nan Hu, Ting Zhang, Xiaoyu Li

(Xi'an Jiaotong University, Xi'an, 710049, China)

Abstract：Assisted with Natural Language Processing tool and a neural network algorithm termed as Word2Vec, we extract and analyze the narrative innovation disclosure from annual financial reports of Chinese listed companies (A share). We find that narrative innovation disclosure is positively related to the level of current real innovative activities and current earnings. More, narrative innovation disclosure can positively predict future real innovation activities and earnings persistence. In addition, we also find that the non-forward-looking statement intensity and tone of narrative innovation disclosure positively predicts future real innovation activities and earrings persistence. Finally, the implementation of revised standards on financial report disclosure (2012) weakens the predictive effect of narrative innovation on future real innovation activities and earnings persistence. Based on advanced large-scale text analysis, we not only provide effective empirical evidence for the information usefulness of narrative innovation in corporate disclosure, but also offer a potential novel index to the evaluation framework on enterprise innovation capability.

Keywords: Narrative innovation, Real innovative activities, Earnings persistence, Information usefulness, Machine learning

一、引言

创新活动作为企业价值增长的重要途径和可持续发展的核心动力，近年来受到越来越多学者的关注 (Chang et al., 2017 , Gennaioli et al., 2012 , Manso, 2011 , Mukherjee et al., 2017 , Tian and Wang, 2011 , 陈怡欣 et al., 2018 , 冯根福 and 温军, 2008 , 王红建 et al., 2017 , 张兆国 et al., 2014)。现有国内外研究主要使用研发投入和专利等真实创新活动指标来衡量企业的创新水平，然而这些定量指标由于自身特征存在着一定的局限性。研发投入经常用于衡量创新活动的投入，但研发投入只能捕捉到部分能够被货币化的研发活动，并且其披露数值受到会计准则的影响较为显著，因此研发投入不能十分准确地反映公司创新投入情况 (Acharya et al., 2013 , Aghion et al., 2013)。而专利数据常用于衡量创新产出水平，但不是所有的创新产出最终都付诸以专利的形式。一则专利申请本身需要一定条件，二则企业由于自身利益考量有时选择以商业机密而非专利的形式保存包括过程创新等在内的创新成果 (He and Tian, 2018)。

而事实上，作为公司信息披露最主要的载体，非结构化的文本信息是传统定量的数值信息之外又一丰富的价值矿藏 (Loughran and McDonald, 2016)。相比于研发投入和专利等真实创新活动信息，公司披露文本中的描述性创新内容向投资者传递了更多关于公司创新活动的信息。那么，这些描述性创新信息是否有真正的信息含量？它和企业的真实创新活动之间的关系有多大？是否可以在真实创新活动之外增量反映企业创新活动水平？而描述性创新的披露与企业的盈余质量是否也相关呢？

目前国内针对描述性创新信息披露的相关研究较少 (韩鹏 and 岳园园, 2016 , 赵武阳 and 陈超, 2011 , 韩鹏 and 彭韶兵, 2012 , 薛云奎 and 王志台, 2001)。其中，韩鹏 and 岳园园 (2016) 采用人工评分法的方式研究了创业板上市公司 2012-2014 年的创新行为信息披露的相关经济后果。然而，人工评分法往往由于需要较大的人力成本而导致数据样本有限，研究结果的普适性较低。并且，由于人工评分结果高度依赖研究人员的经验判断，使得研究结果的可复制性较弱。

近年来，随着自然语言处理技术的飞速发展，国外越来越多经济管理领域的学者开始利用文本分析技术探索上市公司披露文本中所蕴藏的价值信息 (Bonsall IV et al., 2017 , Bushee et al., 2018 , Davis et al., 2015 , Li, 2008 , Li, 2010(a) , Loughran and McDonald, 2016 , Muslu et al., 2014)，而国内目前关于公司文本信息披露相关研究方兴未艾 (蒋艳辉 and 冯楚建, 2014 , 李常青 et al., 2008 , 王雄元 and 高曦, 2018 , 谢德仁 and 林乐, 2015 , 阎达五 and 孙蔓莉, 2002 , 姜付秀 et al., 2015)。国内外文本研究目前所涉及的研究对象既包括可读性、情感、文本相似性等与内容无关的文本指标 (Bonsall IV et al., 2017 , Bushee et al., 2018 , Li, 2008 , Lehavy et al., 2011 , Loughran and McDonald, 2011 , 蒋艳辉 et al., 2014 , 阎达五 and 孙蔓莉, 2002)，也包括风险、虚假性、前瞻性、诚信文化等与内容有关的文本指标 (Larcker and Zakolyukina, 2012 , Muslu et al., 2014 , 李常青 et al., 2008 , 王雄元 and 高曦, 2018 , 王雄元 et al., 2017 , 姜付秀 et al., 2015)。而本文的主要研究对象是与内容有关的描述性创新披露指标，那么，如何利用新兴的文本分析技术从公司披露文本提取出准确有效的描述性创新信息便成为我们的研究重点之一。

本研究使用“种子词集+Word2Vec 相似词扩充”的词频分析法通过以下三个步骤从沪深 A 股上市公司 2007~2016 年的年报中提取描述性创新文本指标：首先，查阅国外相关权威文献，确定描述性创新的种子词集。然后，运用 Word2Vec 神经网络语言模型构建基于财报语料专用的语义相似词“词典”，并利用该算法所生成的“词典”对种子词集进行相似词扩充，获得描述性创新关键词词集。最后用关键词总词频在财报全文中的占比表示描述性创新披露水平。

通过对描述性创新披露的深入分析，我们发现：第一，描述性创新披露受到当期真实创新活动（研发强度和专利）和当期盈余水平的正向影响；第二，描述性创新披露可以正向预测企业未来的真实创新活动水平和盈余持续性。描述性创新对未来真实创新活动和盈余持续性的预测效应即使在控制了当期研发强度和专利水平的条件下依然十分显著，这不仅说明了描述性创新披露的有效信息含量，还说明了描述性创新指标在传统的真实创新活动水平之外对企业的创新水平具有增量的表示作用。

第三，通过对描述性创新披露相关细节的进一步研究发现：当描述性创新信息中的非前瞻性信息含量越高，语调积极程度越大时，其对企业未来真实创新活动和盈余持续性的正向预测作用也越强。这表明企业的描述性创新披露在详细内容和语言风格上均具有较高的可信度。第四，证监会于 2012 年规定上市公司在年报中不仅要说明报告期内研发项目的目的、进展、拟达到目标和预计对未来发展的影响，还应当披露新年度包括研发计划在内的经营计划以及为达到目标拟采取的策略和行动等。研究发现：该准则实施之后，年报的描述性创新披露水平应声上涨，但描述性创新的非前瞻性信息密度和语调积极程度却有所降低。同时，描述性创新披露对未来真实创新活动和盈余持续性的正向预测作用也有一定程度的下降，我们认为这可能与 2012 年之后年报描述性创新披露中说明未来计划、具有较强不确定性的前瞻性信息含量的增加有关。

本研究的潜在贡献有以下三点：第一，研究首次通过大样本实证分析证明了中国上市公司描述性创新及其披露细节的有效信息含量，为公司文本型信息披露的有用性提供了经验证据；第二，不限于以往传统的研发投入和专利等测度方式，本文构建的描述性创新指标为企业创新水平衡量体系的健全和完善提供了全新的参考；第三，文本首次引入基于深度学习思想的 Word2Vec 神经网络模型成功构建了适用于中文年报等于财经专用语料的文本指标，采用新兴机器学习技术与经典计量经济学研究范式相结合的方法，对近年来学术界关于“管理研究+人工智能”话题的深入探讨具有一定的参考价值。本文后续内容结构如下：第二部分总结相关研究并提出研究假设，第三部分介绍研究样本和模型设计，第四部分报告实证分析结果和后续稳健性检验，第五部分是研究结论和相关启示。

二、文献综述与研究假设

（一）描述性创新披露研究

企业的描述性创新信息是企业在定量的真实创新活动信息之外所披露的与创新相关的其他信息，其存在形式以文本为主。企业创新广义上可分为管理创新、制度创新、技术创新和组织创新等，而狭义的企业创新则一般单指企业的技术创新 (魏江 and 寒午, 1998)。

在现有文献中，学者们对企业技术创新的衡量主要表现为企业所披露的定量创新信息，也就是真实创新活动信息。真实创新活动信息可分为创新投入和创新产出两类。其中，创新投入主要指研发费用、研发人员数量等，目前已有大量文献针对创新投入方面（以研发费用为主）进行了深入研究 (Bushee, 1998 , de la Potterie, 2008 , Kothari et al., 2002 , Xu and Yan, 2014 , 王红建 et al., 2017 , 潘越 et al., 2015 , 温军 and 冯根福, 2012)。而创新产出主要指专利申请或授权数量、开发新产品数量及其收入等，国内外研究针对创新产出的研究也十分广泛 (Balsmeier et al., 2017 , Fang et al., 2014 , He and Tian, 2013 , Hendricks and Singhal, 1997 , Hitt et al., 1996 , 温军 and 冯根福, 2012 , 陈怡欣 et al., 2018)。

而事实上，除了上述以定量信息为主的真实创新活动之外，企业还会以文本的形式披露大量关于技术创新的信息，也就是描述性创新信息。我国企业披露的描述性创新信息主要集中于企业财务报告等文件中，涵盖了报告期已开展研发活动的一般性描述、研发战略及研发计划的描述、研发机构或研发中心及具体研发形式的描述等内容 (韩鹏 and 岳园园, 2016)。现有研究中描述性创新信息披露的衡量方式主要以人工评分法为主 (Gu and Li, 2003 , James, 2011 , Jones, 2007 , Nekhili et al., 2012 , Nekhili et al., 2016 , 韩鹏 and 岳园园, 2016)。例如，韩鹏 and 岳园园 (2016) 采用二分法对描述性创新的具体指标进行评分，并运用变异系数法确定样本的描述性创新披露指数。而人工评分法由于需要较大的人力成本且高度依赖研究人员的经验判断，往往面临小样本困境、可复制性较弱和普适性较低的情形。因此如何利用现有技术突破以往研究局限，构建真正适用于我国上市公司的描述性创新指标，便成为亟待解决的问题。

（二）描述性创新披露的影响因素

1. 当期真实创新活动

真实创新活动水平作为企业所披露的定量创新相关信息，主要分为创新投入和创新产出两部分。在创新投入方面，企业的研发活动往往伴随着较大的不确定性，而研发活动的成功与否对企业未来经营具有重大影响，此时研发活动所带来的不确定性使得管理层与外部投资者之间的信息不对称程度显著增大，投资者对企业研发信息的需求随之增加。因此，管理层有更强的动机在公司披露文本中展现更多研发活动相关的信息。国外已有研究表明企业研发活动投入强度与描述性创新披露之间呈现显著的正相关关系 (Entwistle, 1999 , Nekhili et al., 2012 , Merkley, 2013)。而在创新产出方面，由于创新产出代表创新活动的成果，有利于提升企业盈利水平，为企业实现持续的价值创造，因此管理层倾向于通过描述性创新的披露来向投资者传递价值相关信息。本文中我们使用企业研发强度和授权专利水平代表真实创新活动。基于此，我们提出以下假设：

H1：描述性创新披露与当期真实创新活动水平（研发强度和专利）正相关。

2. 当期盈余水平

当期盈余水平是企业对外界报告的最重要的指标之一，而当期盈余的大小对企业描述性信息披露水平也有着显著的联系 (Merkley, 2013 , Li, 2008)。已有研究表明，企业盈余绩效较好时，管理层会增加盈余信息披露的频次，向投资者强调企业良好的经营状况 (Chen et al., 2011 ,

Houston et al., 2010 , Miller, 2010)。而创新信息的披露由于其自身特点，与当期盈余的关系可能相较于盈余信息披露存在差异，因此我们针对描述性创新披露与当期盈余的关系提出如下竞争性假设分析。

一方面，描述性创新信息披露具有一定的专有成本，披露信息越详细，竞争对手可能会获得越多有用信息，从而对企业自身产生不利影响 (Dontoh, 1989 , Sadka, 2004 , Segerstrom, 1991)。尤其是当企业盈余情况较好时，描述性创新信息的披露可能导致较大的专有成本。Merkley (2013) 发现研发投入相关的描述性信息与企业当期盈余水平负相关，该研究认为企业盈余表现不佳可能会给投资者传递企业价值的负面信号，此时外部投资者需要企业释放更多价值相关信息来支持他们的投资决策行为。这使得企业不得不展示更多研发相关的信息，向外界释放企业盈余表现欠佳并非持续现象的积极信号。因此，描述性创新披露与当期盈余水平可能存在负相关关系。

而另一方面，当期盈余水平较好时，企业有动机披露更多的描述性创新活动，这样不仅可以用于分析其盈余表现良好的原因，还可以向投资者传递公司通过创新塑造核心竞争力，实现未来持续价值创造的信心。值得注意的是，不同于 Merkley (2013)，本研究的描述性创新指标不仅包含了研发投入的信息，还包含创新成果相关的描述。也就是说良好的当期盈余一定程度上可能代表创新成果向实际价值的有效转化，而披露较多的描述性创新信息则会更加有助于市场对于企业价值的正面判断。因此描述性创新披露可能受到当期盈余水平的正向影响。

基于上述两种不同的视角，我们对描述性创新披露与当期盈余水平之间的关系提出竞争性假设：

H2a：描述性创新披露与当期盈余水平正相关。

H2b：描述性创新披露与当期盈余水平负相关。

（三）描述性创新与未来真实创新活动

描述性创新的披露是否可以预测企业未来的真实创新活动水平？Bellstam et al. (2017) 通过研究英文分析师报告发现，报告中关于企业的创新主题含量与未来的专利水平正相关。我们认为，企业年报中的描述性创新披露可以捕捉到传统真实创新活动之外与企业创新行为相关的增量信息。而企业的创新行为作为企业实现差异化战略的重要途径，往往具有较强的持续性。那么，即使在控制了真实创新活动水平的情况下，描述性创新也应对未来真实活动水平具有增量的正向预测作用。因此，我们提出以下假设:

H3：描述性创新披露可以正向预测未来的真实创新活动水平（研发强度和专利）。

（四）描述性创新与盈余持续性

盈余持续性作为衡量企业盈余质量的重要标准之一，表示当期盈余对未来盈余的预测程度，较高的预测程度代表着较好的盈余持续性 (Krishnan and Parsons, 2008 , Richardson et al., 2005 , Sloan, 1996)，而较好的盈余持续性往往意味着企业拥有较为平稳的盈余情况和较高的经营管理质量 (肖华 and 张国清, 2013)。因此盈余持续性作为决策有用信息，是外部投资者进行投资决策的重要参考(Collins and Kothari, 1989 , Francis et al., 2004 , Li, 2008 , 窦欢 and 陆正飞, 2017 , 方

红星 and 张志平, 2013)。

本文从资源基础观（RBV）的视角来分析描述性创新与盈余持续性之间的关系。资源基础观认为企业主要通过获取有价值的、难以被模仿和替代的异质性资源来保持竞争优势，而创新能力则是产生异质性企业资源，实现企业差异化竞争战略的重要途径 (Barney, 1991 , Wernerfelt, 1984)。创新能力越强的企业，更有可能通过不断获取异质性资源来保持竞争优势，进行更加持久的价值创造，从而拥有良好的盈余质量 (D. Banker et al., 2014)。Asthana and Zhang (2006)发现 R&D 投入越多的公司其盈余持续性较高。那么，描述性创新披露作为真实创新活动之外的创新水平的有效测度，应当和真实创新活动具有类似的效应，并且这种效应在控制了真实创新活动水平的情况下依然成立。基于此，我们对描述性创新披露与盈余持续性的关系提出以下假设：

H4：描述性创新披露可以正向预测未来的盈余持续性。

三、样本数据与研究设计

（一）研究样本与数据来源

由于企业研发费用可获取的数据区间始于 2007 年，而企业专利数目可获取的数据区间目前截至 2016 年，本文选取 2007-2016 年沪深 A 股上市公司为研究样本。本文描述性创新指标构建的相关文本数据来源于 WinGo 财经文本数据平台，机构投资者持股比例数据来自锐思数据库，其他数据则来自国泰安数据库。

本文对样本数据进行了以下处理：（1）研发费用和专利数目缺失的样本用 0 替代；（2）剔除金融业公司的观测值；（3）剔除 ST 类特殊处理公司的观测值；（4）剔除表 6（描述性创新披露的影响因素）的变量含有缺失的观测值。为了消除样本离群值的影响，本文对所有连续变量按照 1%的标准进行缩尾处理，最终获得 16405 个观测值。

（二）描述性创新的测度

描述性创新信息披露是企业在定量的真实创新活动信息之外披露的与创新相关的文本信息。这里的创新主要指狭义的技术创新 (魏江 and 寒午, 1998)，包括与技术创新相关的投入和产出的描述性信息披露。

本文采用关键词词频的大样本文本分析法构建描述性创新指标。词频分析法作为目前最为常用的文本指标构建方法，因其可理解性强，易于复制等特点而受到广泛关注与应用 (Fiordelisi and Ricci, 2014 , Loughran and McDonald, 2011 , Loughran and McDonald, 2016 , Bellstam et al., 2017)。不同于传统的词频分析法，本文采用“种子词集+Word2Vec 相似词扩充”的方法构建描述性创新指标。首先，我们通过阅读相关文献，收集描述性创新的种子词集 (Merkley, 2013 , Entwistle, 1999 , Jones, 2007 , Uotila et al., 2009)；然后使用 Word2Vec 神经网络相似词算法，在种子词集的基础上进行词汇扩充，确定描述性创新关键词集；最后将年报中描述性创新关键词集的词频之和在年报全文总词数中所占比例作为描述性创新指标。创新关键词词集的词频、年报全文总词数等信息来自 WinGo 财经文本数据平台的财务报告词频数据库，该平台的财报文本处理主要流程可参见附录 1。具体来说，我们的描述性创新指标构建过程如下：

1. 种子词集的选择

我们主要通过翻译筛选现有英文文献中与描述性创新相关的词汇来确定种子词集。现有文献主要分为两类：一是与研发活动相关的研究 (Merkley, 2013 , Entwistle, 1999 , Jones, 2007)，二是广义创新的相关研究 (Bellstam et al., 2017 , Uotila et al., 2009)。考虑到我们的描述性创新以技术创新为主，因此我们以第一类研究为主要参考，辅之以第二类研究，同时认真研读财报样本进行校验。经过翻译和筛选，最终确定了 6 个种子词汇（技术创新、研究、开发、研发、专利、发明）。种子词汇作为描述性创新指标构建的基准，其词汇准确性的要求需十分严格。为减少第一类统计错误 (即某词汇在特定语境中本没有描述创新相关信息但其词频依然被计入描述性创新指标中的情况) 的发生概率，在人工筛选的过程中，我们剔除了诸如“突破”、“创新”等词汇，理由如下：（1）“突破”等词汇在财报文本中除了与技术创新相关以外，还和盈余水平紧密相关，这类词汇的存在会让描述性创新指标因携带过多盈余信息而失去准确性；（2）“创新”等词汇其外延远远超出了技术创新的范畴，因此也未被纳入最终种子词集。

2. Word2Vec 相似词扩充

针对同一概念或事物，表达者往往会使用多个语义相似的词汇进行描述。因此在选定种子词集之后，我们需要对词集进行相似词扩充。现有文献中使用的相似词扩充方法多为人工扩充法和通用同近义词典扩充法 (Fiordelisi and Ricci, 2014 , Merkley, 2013)。人工扩充法主要受限于研究人员的主观经验判断，可靠性和可复制性较弱；而通用同近义词词典扩充法（如 Harvard IV-4 Psychosocial Dictionary，WordNet，HowNet 等）对普通语料较为适用，但对年报等财经专业语料的适用性较弱 (Loughran and McDonald, 2011)。基于此，选择一个财经语料专用的相似词扩充方法无疑更加适合本研究。经过对目前自然语言处理技术相关文献的研究和梳理，本文使用 Word2Vec 神经网络模型实现基于财经专用语料的相似词扩充。

Word2Vec 神经网络模型由 Mikolov et al. (2013) 提出，是近年来深度学习领域的里程碑式成果 (LeCun et al., 2015)。Word2Vec 模型根据上下文内容将词汇表征为多维向量，并通过计算向量的相似度得到词汇间的语义相似性 (Bengio et al., 2003)。Word2Vec 模型的核心思想基于一种朴素的语言学原理，即对于语义相似的词汇，其邻近的词汇往往也较为类似，因此可以用邻近词汇的出现情况来表示词汇本身 (Harris, 1954)。我们通过以下例子来阐述该语言学原理：当计算“研发”和“危机”两个词汇的相似度时，我们可以计算文本中两者相邻词汇出现的次数来表示这两个词汇。假设相邻词共有 5 个词汇（成果，攻克，进行，遭遇，度过），我们用这 5 个词汇作为相邻词出现次数的对应 5 维向量来表示“研发”和“危机”。如果在“研发”的相邻词中上述五个词汇出现次数分别为：“成果”5 次，“攻克”3 次，“进行”5 次，“遭遇”0 次，“度过”0 次，那么“研发”则可以表示为向量 A：(5,3,5,0,0)。类似地，“危机”可以表示为向量 B：(0,0,2,5,5)。因此，通过计算向量 A, B 的余弦相似度得到“研发”和“危机”的相似度为 0.18。当然，实际应用中 Word2Vec 的词汇向量维度一般远大于上述示例，整个计算过程和向量维度含义也更为复杂，本文的附录 2 介绍了 Word2Vec 模型的相关原理。

Word2Vec 模型的使用让我们可以直接获取描述性创新关键词在财报等财经专业语料的相似词候选集。接下来，我们首先找出种子词集中每个词汇的前 200 个相似词，经过去除重复词汇和部分低频词汇之后，由两名专业研究人员各自进行词汇筛选工作，然后将两名研究人员均认可的词汇添加至关键词词集。相似词扩充完成后，我们总共得到了 401 个描述性创新关键词，词集构成如表 1 所示。

3. 最终指标验证

在得到描述性创新关键词词集之后，我们通过邀请行业专家进行核验以及对比财报文本样例的方式对关键词词集进行再次确认，最终通过计算关键词在年报文本中出现的词频比例构建出描述性创新指标。有关描述性创新指标的相关统计分析见 4.1 节。

表 1 描述性创新关键词词集示例

种子词汇	扩充词汇
技术创新、研究、开发、研发、专利、发明	科技创新、技术革新、核心技术、实验、试验、原创、独创、版权、知识产权、专利技术、发明专利、实用新型、外观专利、孵化、产学研、前沿、尖端、绿色技术、信息化、人工智能、云计算、3D打印、智能制造、生物育种、培植、新品种、临床、新药、药学研究……

（三）其他变量和模型设定

为了检验描述性创新与当期真实创新活动和当期盈余水平的关系（假设 1、2），我们提出以下模型（1）：

$$TEXT\_INNO\_W_{i,t}=\alpha+\beta_1 RD_{i,t}+\beta_2 PATENT_{i,t}+\beta_3 ROA_{i,t}+\beta_4 BM_{i,t}+\beta_5 CAPINT_{i,t}+\beta_6 SIZE_{i,t}+\beta_7 ANALYSTS_{i,t-1}+\beta_8 INST\_OWN_{i,t-1}+\beta_9 MF\_COUNT_{i,t}+\beta_{10} AGE_{i,t}+\beta_{11} RETVOL_{i,t}+\beta_{12} ROAVOL_{i,t}+\beta_{13} LEV_{i,t}+\beta_{14} STOCK\_ISS_{i,t}+\beta_{15} HHI_{i,t}+\beta_{16} FILE\_LENGTH_{i,t}+INDUSTRY_i+YEAR_t+\varepsilon_{i,t} \quad (1)$$

其中，因变量 TEXT_INNO_W 表示企业的描述性创新披露的水平。参考虞义华 et al. (2018)，本文使用研发费用与营业收入之比作为企业的研发强度（RD）。参考陈怡欣 et al. (2018)，使用当年申请并截至统计年已获得授权的专利自然对数（PATENT）衡量企业的创新成果。而资产收益率（ROA）则用于衡量企业的盈余水平。

为了控制其他公司层面特征对研究变量关系可能产生的影响，参照以往研究(Merkley, 2013 , Nagar et al., 2003)，本文将以下变量作为控制变量：衡量企业投资结构的账面市值比和有形资产占比；表示公司信息披露环境的公司规模、分析师跟踪数目、机构持股比例、管理层预测次数和公司年龄；衡量公司不确定性的股票收益波动性、盈余波动性；代表公司融资活动的资产负债率和股票发行和衡量企业专有成本的赫芬达尔指数。为了控制年报信息含量对企业创新信息的披露，本文在模型中还加入了年报字数大小变量即年报长度。最后，本研究包括模型（1）在内的所有模型均控制了年份和行业[2]固定效应，以及公司的聚类效应。

为了验证描述性创新与未来真实创新活动的关系（假设 3），本文提出实证模型（2）。因变

[2] 行业分类遵循证监会 2012 版上市公司行业分类指引，其中制造业行业按二级行业代码进行分类，其余行业按一级行业代码分类。

量分别使用未来一期到未来三期（s=1,2,3）的研发强度或者专利水平来衡量未来真实创新活动。参考以往研究 (陈怡欣 et al., 2018 , 袁建国 et al., 2015 , Chemmanur et al., 2014)，本文在模型中控制了资产收益率、公司规模、资产负债率、公司年龄、有形资产占比、账面市值比、托宾 Q 值、赫芬达尔指数、机构持股比例和现金资产比率。此外，模型还对当期的真实创新活动水平（专利水平和研发强度）进行了控制。

$$RD_{i,t+s}(PATENT_{i,t+s})=\alpha+\beta_1 TEXT\_INNO\_W_{i,t}+\beta_2 PATENT_{i,t}+\beta_3 RD_{i,t}+\beta_4 ROA_{i,t}+\beta_5 SIZE_{i,t}+\beta_6 LEV_{i,t}+\beta_7 AGE_{i,t}+\beta_8 CAPINT_{i,t}+\beta_9 BM_{i,t}+\beta_{10} TOBINQ_{i,t}+\beta_{11} HHI_{i,t}+\beta_{12} INST\_OWN_{i,t-1}+\beta_{13} CASH\_ASSET_{i,t}+INDUSTRY_i+YEAR_t+\varepsilon_{i,t} \quad (2)$$

为了验证描述性创新与盈余持续性的关系（假设 4），我们提出实证模型（3）。该模型的因变量为未来一期到未来三期（s=1,2,3）的资产收益率。参考以往研究 (D. Banker et al., 2014 , Li, 2008 , 肖华 and 张国清, 2013)，我们在模型中控制了真实创新活动水平（专利水平和研发强度）、公司规模、资产负债率、账面市值比、公司年龄、有形资产占比、股票收益波动性、盈余波动性和股票发行等变量。盈余持续性代表了当期盈余对未来盈余水平的预测程度。由于我们关注的是描述性创新披露水平与盈余持续性间的关系，因此我们通过观察系数 β_3 是否显著来验证假设[3]。

$$ROA_{i,t+s}=\alpha+\beta_1 TEXT\_INNO\_W_{i,t}+\beta_2 ROA_{i,t}+\beta_3 C(TEXT\_INNO\_W_{i,t})\times C(ROA_{i,t})+\beta_4 PATENT_{i,t}+\beta_5 RD_{i,t}+\beta_6 SIZE_{i,t}++\beta_7 LEV_{i,t}+\beta_8 BM_{i,t}+\beta_9 AGE_{i,t}+\beta_{10} CAPINT_{i,t}+\beta_{11} RETVOL_{i,t}+\beta_{12} ROAVOL_{i,t}+\beta_{13} STOCK\_ISS_{i,t}+INDUSTRY_i+YEAR_t+\varepsilon_{i,t} \quad (3)$$

以上三个模型所涉及的变量定义详情可参见表 2。

表 2 变量定义

变量名称	变量标记	变量说明
描述性创新文本变量		
TEXT_INNO_W	描述性创新	描述性创新关键词词集总词频占财报总词数的比例乘以100
NONFORWARD	描述性创新的非前瞻性信息密度	1减去描述性创新内容中包含前瞻性信息的句子所占比例
TONE	描述性创新的语调	描述性创新内容中积极词汇词频与消极词汇词频之差除以积极词汇词频与消极词汇词频加1之和
真实创新活动和盈余水平变量		
RD	研发强度	研发费用除以营业收入，代表真实创新活动水平
PATENT_NUM	专利数目	当年申请并截至统计年已获得授权的发明、实用新型、外观设计专利数目的总和，代表真实创新活动水平
PATENT	专利水平	专利数目加1的自然对数，代表真实创新活动水平
ROA	资产收益率	净利润除以资产总额,代表盈余水平
其他变量		
BM	账面市值比	股东权益除以总市值

[3] 模型（3）的关键交乘项为 C(TEXT_INNO_W)×C(ROA)，其中 C(TEXT_INNO_W)和 C(ROA)分别代表描述性创新指标和资产收益率去均值之后的变量。该项回归结果的系数 β_3 大小与显著性与交乘项为 TEXT_INNO_W× ROA 时一致，不同的是 β_1 的解释在模型（3）中是当 ROA 等于样本均值时，描述性创新与未来 ROA 的回归系数。

CAPINT	有形资产占比	固定资产净额与存货净额之和除以资产总额
SIZE	公司规模	权益市值的自然对数
ANALYSTS	分析师跟踪数目	上期分析师（团队）跟踪的数目加1的自然对数
INST_OWN	机构持股比例	上期机构投资者的持股比例
MF_COUNT	管理层预测次数	当年业绩预告发布的次数
AGE	公司年龄	公司成立年份至统计年份年数加1的自然对数
RETVOL	股票收益波动性	前三年股票收益率的标准差
ROAVOL	盈余波动性	前三年资产收益率的标准差
LEV	资产负债率	负债总额除以资产总额
STOCK_ISS	股票发行	本年度是否存在股票发行，若有取值为1，否则为0
HHI	赫芬达尔指数	公司所在行业产品市场竞争程度，根据 2012 版证监会行业划分计算的公司主营业务收入占行业总收入的赫芬达尔指数
FILE_LENGTH	年报长度	年报文本的总字数加1的自然对数
TOBINQ	托宾Q值	总股数乘以期末收盘价的积与负债总额之和除以资产总额
CASH_ASSET	现金资产比率	现金及现金等价物除以资产总额

四、实证分析

（一）描述性统计

1. 变量的描述性统计

表 3 变量描述性统计

Variable	N	Mean	S.D.	Min	P25	P50	P75	Max
Panel A:描述性创新文本变量								
TEXT_INNO_W	16,405	0.653	0.331	0.150	0.408	0.585	0.841	1.723
NONFORWARD	16,405	0.835	0.059	0.660	0.800	0.841	0.877	0.950
TONE	16,405	0.304	0.096	0.058	0.239	0.305	0.369	0.522
Panel B:真实创新活动和盈余变量								
RD	16,405	0.016	0.027	0.000	0.000	0.000	0.030	0.145
PATENT_NUM	16,405	25.993	74.526	0.000	0.000	3.000	18.000	549.000
PATENT	16,405	1.671	1.699	0.000	0.000	1.386	2.944	6.310
ROA	16,405	0.034	0.059	-0.220	0.011	0.031	0.060	0.211
Panel C:控制变量								
BM	16,405	0.390	0.263	-0.027	0.203	0.329	0.514	1.313
CAPINT	16,405	0.425	0.184	0.032	0.291	0.419	0.559	0.838
SIZE	16,405	22.128	1.062	19.796	21.387	22.093	22.787	25.021
ANALYSTS	16,405	1.332	1.119	0.000	0.000	1.386	2.303	3.611
INST_OWN	16,405	0.176	0.183	0.000	0.033	0.113	0.264	0.762
MF_COUNT	16,405	0.604	1.052	0.000	0.000	0.000	1.000	4.000
AGE	16,405	2.736	0.342	1.792	2.565	2.773	2.996	3.367
RETVOL	16,405	0.146	0.064	0.056	0.102	0.132	0.176	0.403
ROAVOL	16,405	0.033	0.061	0.001	0.007	0.015	0.033	0.461
LEV	16,405	0.491	0.213	0.069	0.331	0.493	0.645	1.093

STOCK_ISS	16,405	0.143	0.350	0.000	0.000	0.000	0.000	1.000
HHI	16,405	0.108	0.099	0.018	0.046	0.073	0.127	0.480
FILE_LENGTH	16,405	10.696	0.255	10.073	10.525	10.694	10.867	11.315
TOBINQ	16,405	2.647	2.052	0.908	1.413	2.009	3.056	13.648
CASH_ASSET	16,391	0.149	0.113	0.005	0.068	0.118	0.197	0.550

表 3 报告了变量的描述性统计结果。其中按描述性创新文本指标均值为 0.653，表示年报全文平均每 100 个词中含 0.653 个描述性创新关键词。研发强度平均为 0.016，每年得到授权的专利数量平均约为 26，而资产收益率的均值 0.034。从平均值和中位数的比较来看，上述主要研究变量均呈右偏分布。

2. 主要变量的相关系数

表 4 报告了描述性创新指标、研发强度、专利水平和资产收益率的 Pearson 相关系数和 Spearman 相关系数。以 Pearson 相关系数为例，描述性创新指标与研发强度和专利的相关系数分别为 0.52 和 0.43，且均在 1%水平上显著，说明描述性创新指标与研发强度和专利数目均存在较强的相关性。资产收益率与描述性创新指标、研发强度和专利数目的相关系数均在 0.1 左右，且在 1%水平上显著。

表 4 主要变量的相关系数

	TEXT_INNO_W	RD	PATENT	ROA
TEXT_INNO_W		0.58*	0.49*	0.10*
RD	0.52*		0.48*	0.07*
PATENT	0.43*	0.35*		0.10*
ROA	0.10*	0.07*	0.11*	

注：下三角为 Pearson 相关系数，上三角为 Spearman 相关系数。*表示 p<0.01。

3. 描述性创新的分析

（1）描述性创新的时间趋势

图 1 展示了描述性创新指标和真实创新活动水平（研发强度和专利水平）随时间变化的趋势图，总体上两者都随时间呈现上升趋势，说明近年来我国企业的总体创新行为处于增长趋势。从图中可以看出描述性创新和研发强度的时间趋势一致性相对较高，而描述性创新与专利水平的一致性则略低于前者。2007 年至 2012 年期间，专利数目保持着稳定的增长趋势，描述性创新指标也呈增长趋势，不过其逐年增长幅度差异相对较为明显。2015 年和 2016 年专利数目均值有所下滑，可能的原因是由于专利数目为当年申请并截至 2016 年已获得授权的专利数目，专利从申请到获得授权往往需要一定的时间，因此取值尚不能完全代表最终的申请授权数目。为减少这部分专利数据不准确的样本对实证分析的影响，我们在后续将剔除 2015-2016 年的样本对结果进行稳健性检验。

（2）描述性创新的行业分布

图 2 展示了研发强度和专利数目的行业均值水平（按描述性创新指标大小降序排序）。从图中可以看出描述性创新指标均值最高的三个行业是信息传输、软件和信息技术服务业，科学研究

和技术服务业，制造业，这三个行业同样也是研发强度最高的三个行业。描述性创新指标和研发强度最低的三个行业也大致相同，都有住宿和餐饮业，交通运输、仓储和邮政业。根据国家统计局公布的《高技术产业分类》文件[4]，信息传输、软件和信息技术服务业，科学研究和技术服务业，制造业，特别是制造业中的计算机、通信和其他电子设备制造业，仪器仪表制造业以及医药制造业等都是技术密集型产业，因此描述性创新指标的行业分布与我国行业特性相符。图 2 中专利数目均值最高的三个行业是建筑业，采矿业和制造业，而最低的三个行业为住宿和餐饮业，卫生和社会工作和房地产业。总体来说，描述性创新指标与研发强度的行业分布一致性较高，而与专利数目的行业分布情况一致性则相对较低。

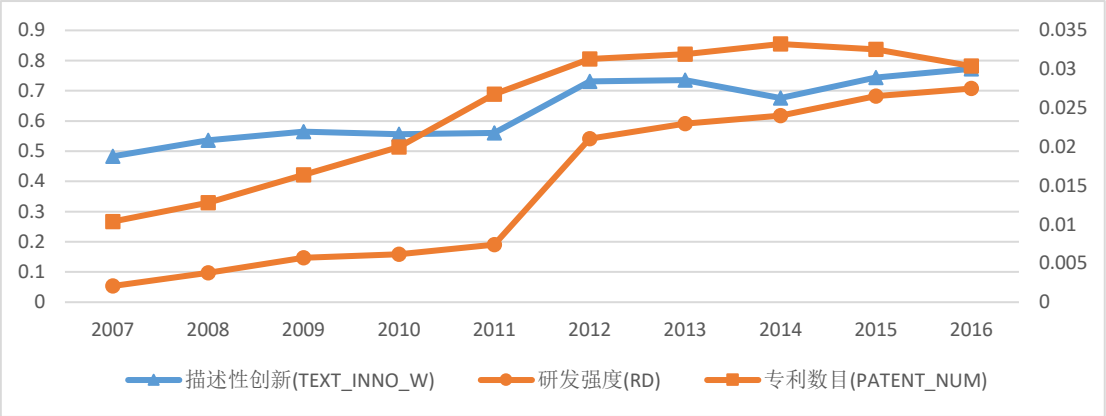
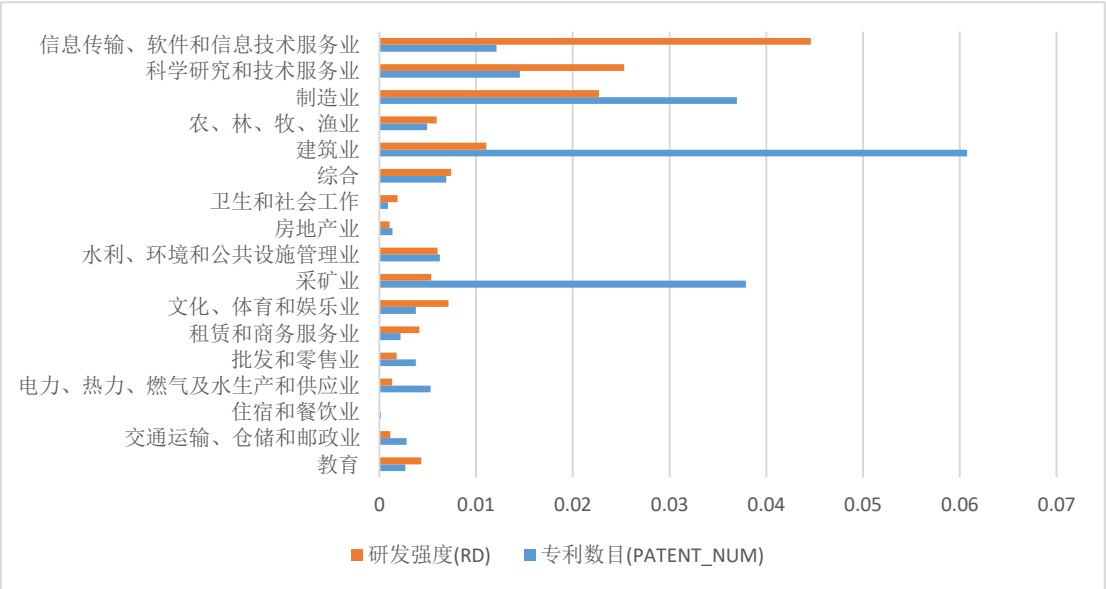


图 1 描述性创新与真实创新活动的时间趋势

注：描述性创新指标为左侧坐标轴，真实创新活动指标（研发强度、专利数目）为右侧坐标轴。为使真实创新活动指标量纲一致，将专利数目指标缩小 1000 倍予以呈现。



[4]（1）高技术产业（服务业）分类（2018）：
http://www.stats.gov.cn/tjsj/tjbz/201805/t20180509_1598315.html
（2）高技术产业（制造业）分类（2013）：
http://www.stats.gov.cn/tjsj/tjbz/201310/P020131021347576415205.pdf

图 2 真实创新活动的行业分布（按描述性创新降序排序）

注：本图先按行业的描述性创新指标降序排序，然后做出对应研发强度和专利数目的行业分布图。为使真实创新活动指标量纲一致，将专利数目指标缩小 1000 倍予以呈现。

（3）描述性创新的持续性特征

本文采用两种方法来检验每家公司描述性创新指标在时间上的持续性水平。首先，我们计算了样本的 Cronbach's alpha 系数。该系数常用于检验数据内部的一致性，当系数得分超过 0.9 时通常认为数据具有高度一致性。我们所研究样本的 Cronbach's alpha 系数为 0.95，因此总体来看，描述性创新指标随时间的持续性较高。除此之外，我们还将根据描述性创新指标将每年的样本分为五等份，观察上五分位区间和下五分位区间的样本在随后三年的样本中所位于五分位的区间情况。如果描述性创新指标持续性较高，那么后续年份其所位于的分位数区间应该和原来一致或相近。如表 5 所示，第 0 年描述性创新指标位于上五分位区间的公司在随后三年的里仍处于上五分位的比例分别为 71.73%，66.24%和 57.85%。从表中可以看出，描述性创新指标具有较高的持续性。

表 5 描述性创新的持续性特征

年份	下五分位区间	第二五分位区间	第三五分位区间	第四五分位区间	上五分位区间
第0年					100.00%
第1年	0.00%	0.43%	3.64%	24.20%	71.73%
第2年	0.22%	0.86%	6.45%	26.24%	66.24%
第3年	1.08%	3.01%	10.32%	27.74%	57.85%

注：表结果为描述性创新指标位于上五分位区间的企业在后续三年中描述性创新指标处于各五分位区间的情况。样本数据未经任何剔除和缩尾处理。

（4）描述性创新代表性企业示例

本文通过按照描述性创新指标对样本企业排序，在附录 3 中示例了不同行业描述性创新指标最高和最低的五家企业，在描述性创新指标最高的企业中，既有育种享誉业界的隆平高科和登海种业，也有中国高性能计算、服务器、云计算、大数据领域的领军企业中科曙光，这些企业的创新水平在业界都十分突出。这使得我们从行业内部的视角检验了描述性创新指标的合理性。描述性创新代表性企业示例的具体情况参见附录 3。

（二）描述性创新的影响因素和预测作用

1. 描述性创新披露的影响因素

表 6 列示了描述性创新影响因素的回归结果，列（1）为未加入控制变量的回归结果，列（2）为加入控制变量的回归结果。由列（2）可知，研发强度（RD）系数为 3.257，t 值为 15.06。专利水平（PATENT）的系数为 0.036，t 值为 11.23。两者均在 1%的显著性水平上显著，说明研发强度越大和专利水平越多的公司会披露更多的描述性创新披露信息，回归结果支持假设 1，即企业的描述性创新水平与真实创新活动正相关。

除了真实创新活动水平以外，从列（2）中还可以看出当期盈余水平（ROA）的系数在 1%的水平下显著（0.179，t=3.18）。本文的描述性创新特征与当期盈余水平正相关，说明企业盈余水

平较好时，企业倾向于披露更多的描述性创新信息。结果支持了假设 2a，即企业通过披露更多的创新相关信息向投资者传递公司通过创新塑造核心竞争力，实现持续价值创造的企业形象。值得注意的是，不同于 Merkley (2013)，我们的描述性创新不仅关注研发投入活动的披露，还关注创新产出的描述。此时，良好的当期盈余一定程度上可能已经表示创新成果向实际价值的有效转化，披露更多的描述性创新信息会更加有助于市场对于企业长期价值的正面判断。

表 6 描述性创新披露的影响因素

		(1)	(2)
		TEXT_INNO_W	TEXT_INNO_W
真实创新活动	RD	3.695***	3.257***
		(17.35)	(15.06)
	PATENT	0.041***	0.036***
		(12.99)	(11.23)
盈余水平	ROA	0.341***	0.179***
		(5.82)	(3.18)
投资结构	BM		-0.054***
			(-3.35)
	CAPINT		0.042*
			(1.80)
信息披露环境	SIZE		-0.002
			(-0.36)
	ANALYSTS		0.011***
			(2.76)
	INST_OWN		-0.005
			(-0.23)
	MF_COUNT		-0.009***
			(-3.03)
	AGE		-0.053***
			(-3.69)
不确定性	RETVOL		-0.033
			(-0.73)
	ROAVOL		-0.421***
			(-8.56)
融资活动	LEV		-0.043*
			(-1.96)
	STOCK_ISS		0.007
			(1.28)
专有成本	HHI		0.035
			(0.45)
年报信息含量	FILE_LENGTH		0.077***
			(4.55)
	Cons	0.430***	-0.176
		(61.56)	(-0.85)

	Year Fixed Effect	Yes	Yes
	Industry Fixed Effect	Yes	Yes
	N	16405	16405
	Adj. R-Squared	0.414	0.431

2. 描述性创新与未来真实创新活动

在研究了描述性创新披露的影响因素之后，我们继续关注描述性创新披露的经济后果。表 7 的列（1）-（3）和列（4）-（6）分别列示未来一期至未来三期研发强度和专利水平与描述性创新指标的回归结果。结果发现，即使是在控制了当期研发强度和当期专利的情况下，描述性创新指标与未来三期的研发强度和专利水平之间依然具有显著的正相关关系。这说明未来真实创新活动水平（研发强度和专利）与当期的描述性创新披露正相关。回归结果支持假设 3，即描述性创新披露可以正向预测未来的真实创新活动水平。

表 7 描述性创新与未来真实创新活动

	未来研发强度			未来专利水平		
	(1)	(2)	(3)	(4)	(5)	(6)
	RD(t+1)	RD(t+2)	RD(t+3)	PATENT(t+1)	PATENT(t+2)	PATENT(t+3)
TEXT_INNO_W	0.003***	0.003***	0.006***	0.213***	0.193***	0.143**
	(5.85)	(4.86)	(6.15)	(7.47)	(4.57)	(2.56)
PATENT	0.001***	0.001***	0.001***	0.744***	0.684***	0.624***
	(9.09)	(9.13)	(7.75)	(87.15)	(62.21)	(44.13)
RD	0.598***	0.477***	0.354***	-0.794**	-1.737***	-1.368*
	(52.57)	(36.32)	(21.24)	(-2.27)	(-3.12)	(-1.85)
ROA	0.009***	0.011***	0.006	0.938***	1.077***	0.864***
	(3.46)	(3.63)	(1.57)	(6.80)	(5.96)	(4.01)
SIZE	-0.000***	-0.001***	-0.001**	0.117***	0.143***	0.188***
	(-3.86)	(-3.98)	(-2.39)	(12.19)	(10.53)	(10.55)
LEV	-0.005***	-0.007***	-0.008***	0.194***	0.246***	0.261***
	(-7.26)	(-7.70)	(-6.91)	(4.98)	(4.16)	(3.43)
AGE	-0.002***	-0.003***	-0.003***	-0.127***	-0.208***	-0.262***
	(-6.76)	(-5.07)	(-4.23)	(-5.38)	(-5.75)	(-5.39)
CAPINT	-0.001	-0.001	-0.001	-0.079*	-0.048	-0.020
	(-1.06)	(-1.43)	(-0.50)	(-1.80)	(-0.75)	(-0.23)
BM	-0.001*	-0.005***	-0.005***	0.147***	0.151***	0.135**
	(-1.85)	(-7.03)	(-5.69)	(3.85)	(2.89)	(2.08)
TOBINQ	0.000***	-0.000	0.000*	-0.009**	-0.014**	-0.012
	(3.44)	(-1.24)	(1.69)	(-2.01)	(-2.17)	(-1.45)
HHI	0.037***	0.046***	0.045***	-0.433*	-0.730	-0.899
	(7.21)	(6.71)	(5.69)	(-1.81)	(-1.53)	(-1.47)
INST_OWN	0.001**	0.001	0.002**	0.044	0.043	0.078
	(2.12)	(1.61)	(2.05)	(1.18)	(0.77)	(1.04)

CASH_ASSET	0.003**	0.003*	0.003	0.002	0.148	0.158
	(2.28)	(1.79)	(1.45)	(0.04)	(1.44)	(1.17)
Cons	0.012***	0.021***	0.016***	-1.950***	-2.084***	-2.767***
	(4.03)	(4.79)	(2.73)	(-9.23)	(-6.87)	(-6.98)
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Industry Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
N	16349	14302	12321	14370	12401	10453
Adj. R-Squared	0.672	0.587	0.528	0.761	0.704	0.652

3. 描述性创新与盈余持续性

表 8 列示了盈余持续性与描述性创新指标的回归结果。在控制了当期盈余水平的条件下，描述性创新指标（去均值）和当期盈余水平（去均值）的交乘项 C(TEXT_INNO_W) × C(ROA)的系数均正向显著，且显著性水平均达到 1%。在控制了研发强度和专利的情况下，描述性创新披露水平越高的企业具有更好的盈余持续性，回归结果支持假设 4。这表明描述性创新披露信息对企业创新水平有着增量反映作用，在保持其他因素不变时，描述性创新披露越多的公司，其创新能力也越强，从而拥有更加优良的盈余质量。

表 8 描述性创新与未来盈余持续性

	(1)	(2)	(3)
	ROA(t+1)	ROA(t+2)	ROA(t+3)
TEXT_INNO_W	0.457***	0.441***	0.376***
	(8.76)	(7.34)	(6.13)
ROA	0.411***	0.307***	0.276***
	(30.66)	(19.83)	(16.27)
C(TEXT_INNO_W)×C(ROA)	0.308***	0.298***	0.254***
	(8.80)	(7.35)	(6.12)
PATENT	0.002***	0.003***	0.003***
	(5.33)	(6.31)	(6.20)
RD	0.001	-0.008	-0.043
	(0.03)	(-0.30)	(-1.50)
SIZE	0.006***	0.004***	0.002**
	(11.61)	(6.29)	(2.07)
LEV	-0.029***	-0.029***	-0.023***
	(-9.25)	(-8.33)	(-5.86)
BM	-0.028***	-0.027***	-0.022***
	(-13.33)	(-10.97)	(-8.26)
AGE	-0.003**	-0.002	-0.001
	(-2.07)	(-1.05)	(-0.52)
CAPINT	-0.005	-0.003	-0.002
	(-1.62)	(-1.00)	(-0.62)
RETVOL	-0.047***	-0.060***	-0.059***
	(-4.83)	(-5.37)	(-4.95)

ROAVOL	-0.020* (-1.75)	-0.017 (-1.29)	-0.007 (-0.58)
STOCK_ISS	0.005*** (4.62)	0.005*** (4.42)	-0.001 (-0.50)
Cons	-0.395*** (-10.83)	-0.326*** (-7.58)	-0.220*** (-5.07)
Year Fixed Effect	Yes	Yes	Yes
Industry Fixed Effect	Yes	Yes	Yes
N	16379	14332	12352
Adj. R-Squared	0.304	0.223	0.188

注：C(TEXT_INNO_W)×C(ROA)为 TEXT_INNO_W 变量和 ROA 变量去均值之后的交乘项。该项回归结果的系数大小与显著性和交乘项为 TEXT_INNO_W×ROA 时一致，不同的是 TEXT_INNO_W 项系数的解释在这里为当 ROA 等于样本均值时，描述性创新与未来 ROA 的回归系数。

（三）描述性创新的细节特征

从描述性创新披露的总体水平来看，描述性创新受到当期真实创新活动和当期盈余水平的显著正向影响，并且对未来的真实创新活动和盈余持续性具有正向预测作用，支持企业披露描述性创新的信息有用性动机，而非刻意隐藏或是粉饰业绩。为了进一步验证描述性创新信息披露的动机，我们对描述性创新文本的细节特征做了深入分析，包括描述性创新文本中的非前瞻性信息密度以及语调特征。

描述性创新文本中不仅包含企业对历史和当前创新投入和产出活动的描述（即非前瞻的描述性创新信息），还包含企业未来的创新活动相关的计划性说明（即前瞻的描述性创新信息）。前瞻描述性创新信息虽然也可反映企业对创新活动的重视水平，但由于尚未发生，具有一定的不确定性。而非前瞻描述性创新信息，属于已经发生的企业创新活动描述，属于确定的可供外界利益相关者参考的信息。因此我们对非前瞻描述性创新信息的含量做了进一步研究。本文先通过“种子词集+Word2Vec 相似词扩充” 的方法确定前瞻性关键词词集[5]，然后算出描述性创新信息句子中包含前瞻性词汇的句子比例，进而用 1 减去该比例得到描述性创新的非前瞻性信息密度指标。

表 9 报告了描述性创新的非前瞻性信息密度回归结果。从 Panel A 可以发现，描述性创新中非前瞻性信息密度与当期的研发强度和专利水平均显著正相关，且显著性水平达到 1%，这表明真实创新活动越多，企业对已发生的创新行为描述的含量也随之增加。而它与当期盈余水平的系数不再显著，这说明当期的盈余水平对描述性创新的非前瞻性含量披露并无直接影响。从 Panel B 和 Panel C 可以看出，描述性创新的非前瞻性信息密度与未来三期的研发强度和专利水平均正相关，同时与未来两期的盈余水平也正相关。这表明平均而言描述性创新信息中的非前瞻性信息越多，其对未来真实创新活动和盈余持续性的正向预测作用也越明显。

表 9 描述性创新的非前瞻性信息密度

[5] 前瞻性词集的构建过程为：首先参考 Li (2010b)和 Muslu et al. (2014)的前瞻性信息词集进行中文翻译形成前瞻性种子词集，然后对种子词集进行 Word2Vec 相似词扩充。最终前瞻性词集包含了计划、预计、未来、目标、如果、预期、预测、今后、前景、希望、相信、愿景、期待、明年、打算等 120 个词汇。

Panel A: 描述性创新（非前瞻性信息密度）的影响因素

		(1)	(2)
		NONFORWARD	NONFORWARD
真实创新活动	RD	0.068**	0.083***
		(2.32)	(2.71)
	PATENT	0.003***	0.003***
		(6.51)	(4.99)
盈余水平	ROA	-0.017	0.005
		(-1.46)	(0.45)
	Cons	0.835***	0.748***
		(430.20)	(17.93)
其他影响因素	Controls	No	Yes
	N	16405	16405
	Adj. R-Squared	0.036	0.043

Panel B: 描述性创新（非前瞻性信息密度）与未来真实创新活动

	未来研发强度			未来专利水平		
	(1)	(2)	(3)	(4)	(5)	(6)
	RD(t+1)	RD(t+2)	RD(t+3)	PATENT(t+1)	PATENT(t+2)	PATENT(t+3)
NONFORWARD	0.004**	0.005**	0.007**	0.374***	0.449***	0.575***
	(2.13)	(2.19)	(2.50)	(3.24)	(2.89)	(2.89)
Cons	0.011***	0.019***	0.012**	-2.179***	-2.375***	-3.178***
	(3.14)	(3.87)	(1.97)	(-9.32)	(-7.23)	(-7.49)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	16349	14302	12321	14370	12401	10453
Adj. R-Squared	0.671	0.586	0.525	0.761	0.704	0.652

Panel C: 描述性创新（非前瞻性信息密度）与盈余持续性

	(1)	(2)	(3)
	ROA(t+1)	ROA(t+2)	ROA(t+3)
NONFORWARD	0.714**	0.567*	0.031
	(2.18)	(1.65)	(0.11)
ROA	0.380***	0.272***	0.239***
	(24.57)	(16.02)	(13.88)
C(NONFORWARD) ×C(ROA)	0.487**	0.387*	0.020
	(2.20)	(1.68)	(0.11)
Cons	-0.697**	-0.516*	-0.005
	(-2.54)	(-1.79)	(-0.02)
Controls	Yes	Yes	Yes
N	16379	14332	12352
Adj. R-Squared	0.296	0.214	0.181

注：以上所有回归均控制了年度和行业固定效应，同时控制了公司层面的聚类效应。C(NONFORWARD)×C(ROA) 为 NONFORWARD 变量和 ROA 变量去均值之后的交乘项。该项回归结果的系数大小与显著性和交乘项为 NONFORWARD×ROA 时一致，不同的是 NONFORWARD 项系数的解释在这里为当 ROA 等于样本均值时，描述性创新的非前瞻信息密度与未来 ROA 的回归系数。

除了非前瞻性信息密度，我们还研究了描述性创新信息的语调，即年报的描述性创新相关内容中积极和消极情绪词汇[6]的相对使用情况。参考 Price et al. (2012)和谢德仁 and 林乐 (2015)，语调的定义如下所示：

语调 =（积极词汇词频-消极词汇词频）/（积极词汇词频+消极词汇词频+1），

语调取值越大，则说明描述性创新披露的情感倾向越积极。表 10 报告了描述性创新的语调回归结果。Panel A 中描述性创新语调与当期的真实创新活动水平（研发投入、专利）和当期的盈余水平都有着显著的正相关关系，而从 Panel B 和 Panel C 可以发现，描述性创新的语调同时对未来的真实创新活动水平和盈余持续性具有显著的正向预测作用。以往研究 (林乐 and 谢德仁, 2017，谢德仁 and 林乐, 2015) 通过对业绩说明会文本的研究发现管理层语调具有一定的有效信息含量，我们从年报的描述性创新披露风格的视角也证实了语调具有良好的可信度。

表 10 描述性创新的语调

Panel A: 描述性创新（语调）的影响因素

		(1)	(2)
		TONE	TONE
真实创新活动	RD	0.374***	0.205***
		(7.74)	(4.31)
	PATENT	0.009***	0.004***
		(10.52)	(4.89)
盈余水平	ROA	0.368***	0.238***
		(21.36)	(13.15)
	Cons	0.299***	0.072
		(98.37)	(1.10)
其他影响因素	Controls	No	Yes
	N	16405	16405
	Adj. R-Squared	0.261	0.326

Panel B: 描述性创新（语调）与未来真实创新活动

	未来研发强度			未来专利水平		
	(1)	(2)	(3)	(4)	(5)	(6)
	RD(t+1)	RD(t+2)	RD(t+3)	PATENT(t+1)	PATENT(t+2)	PATENT(t+3)
TONE	0.002	0.005***	0.008***	0.320***	0.481***	0.565***
	(1.39)	(2.67)	(3.61)	(3.90)	(4.04)	(3.67)
Cons	0.013***	0.022***	0.017***	-1.896***	-2.046***	-2.753***
	(4.35)	(4.97)	(3.03)	(-8.95)	(-6.73)	(-6.93)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	16349	14302	12321	14370	12401	10453
Adj. R-Squared	0.671	0.586	0.525	0.761	0.704	0.652

Panel C: 描述性创新（语调）与盈余持续性

[6] 语调指标的构建过程为：首先将 L&M 情感词典(Loughran and McDonald, 2011)使用百度词典、金山词霸、有道词典等进行翻译形成种子词集，然后使用同义词词林工具对种子词集进行扩充，最终得到 3823 个积极词汇和 7697 个消极词汇。

	(1)	(2)	(3)
	ROA(t+1)	ROA(t+2)	ROA(t+3)
TONE	2.387***	2.213***	1.926***
	(15.09)	(12.12)	(11.02)
ROA	0.457***	0.344***	0.295***
	(33.38)	(22.26)	(18.32)
C(TONE)×C(ROA)	1.589***	1.480***	1.288***
	(14.96)	(12.02)	(10.94)
Cons	-0.799***	-0.690***	-0.542***
	(-16.43)	(-12.19)	(-9.85)
Controls	Yes	Yes	Yes
N	16379	14332	12352
Adj. R-Squared	0.325	0.243	0.206

注：以上所有回归均控制了年度和行业固定效应，同时控制了公司层面的聚类效应。C(TONE)×C(ROA) 为 TONE 变量和 ROA 变量去均值之后的交乘项。该项回归结果的系数大小与显著性和交乘项为 TONE×ROA 时一致，不同的是 TONE 项系数的解释在这里为当 ROA 等于样本均值时，描述性创新的语调与未来 ROA 的回归系数。

（四）年报披露政策的影响

2012 年证监会发布了第 22 号公告《公开发行证券的公司信息披露内容与格式准则第 2 号——年度报告的内容与格式》，规定上市公司不仅要说明报告期内研发项目的目的、进展、拟达到目标和预计对未来发展的影响，还应当披露新年度包括研发计划在内的经营计划以及为达到目标拟采取的策略和行动等。相比以往准则，该政策对描述性创新披露的内容做出了更加详细的指引和要求。为了分析该政策对描述性创新披露的具体影响，我们设置了时间变量 YEAR12，2012 年之前取值为 0，2012 年及之后取值为 1，最终得到表 11 中的实证结果。

从表 11 的 Panel A 中可以看出，在 2012 年政策颁布之后，描述性创新整体披露水平有了显著提升，但是描述性创新披露的非前瞻性信息密度和语调积极程度却有所降低。这表明描述性创新整体水平的提升主要来源于前瞻性描述性创新披露内容的增加。与此同时，由于前瞻性的描述性创新信息的不确定性更高，整体语调的积极程度也随之有所下降。该结果再次证实企业描述性创新信息披露不存在明显的语调操纵行为。表 11 的 Panel B 的结果表明， 2012 年之后描述性创新对未来真实创新活动（研发强度和专利水平）仍然具有显著的正向预测作用，但该作用有一定程度的减弱。而 Panel C 中的分组回归结果也得到了类似的结论，即 2012 年披露政策的出台在一定程度上弱化了描述性创新对未来盈余持续性的正向预测作用。我们认为这可能与 2012 年之后描述性创新披露中前瞻性信息含量的增加有关。总体上，不管是在年报披露政策实施之前还是之后，描述性创新披露对未来真实创新活动水平和未来盈余持续性都具有显著的正相关关系。

（五）稳健性检验

1. 高管个人特征与描述性创新披露

考虑到企业高管个人特征可能会同时影响企业的真实创新活动水平、盈利水平和描述性创新披露行为，参考以往研究 (刘运国 and 刘雯, 2007，文芳 and 胡玉明, 2009)，我们在模型中加

入企业 CEO 的年龄、性别、教育水平、职业背景、任期、薪酬水平以及 CEO 是否兼任董事长和 CEO 是否持股的个人特征作为控制变量。结果发现[7]：当 CEO 为技术类（研发、设计和生产）职业背景、CEO 持有公司股票、以及 CEO 的薪酬和教育水平越高时，企业的描述性创新信息的披露水平会有显著增加；但我们并未发现 CEO 的年龄、性别、任期和是否兼任董事长的特征对描述性创新披露有任何显著的影响。更重要的是，即使在控制了 CEO 个人特征之后，描述性创新披露与当期真实创新活动（研发强度、专利）和当期盈余水平依然显著正相关，从而证明了本研究中假设 1 和假设 2 实证结果的稳健性。

表 11 年报披露政策的影响

Panel A: 年报披露政策对描述性创新（及其细节特征）披露的影响

	(1)	(2)	(3)
	TEXT_INNO_W	NONFORWARD	TONE
YEAR12	0.166***	-0.006*	-0.098***
	(11.22)	(-1.91)	(-20.36)
Cons	-0.176	0.748***	0.072
	(-0.85)	(17.93)	(1.10)
Controls	Yes	Yes	Yes
N	16405	16405	16405
Adj. R-Squared	0.431	0.043	0.326

Panel B: 年报披露政策，描述性创新与未来真实创新活动

	未来研发强度			未来专利水平		
	(1)	(2)	(3)	(4)	(5)	(6)
	RD(t+1)	RD(t+2)	RD(t+3)	PATENT(t+1)	PATENT(t+2)	PATENT(t+3)
TEXT_INNO_W	0.006***	0.008***	0.012***	0.360***	0.412***	0.397***
	(6.58)	(7.40)	(8.92)	(9.24)	(7.30)	(5.82)
YEAR12	-0.015***	0.015***	0.022***	0.007	0.140***	0.175***
	(-21.83)	(18.28)	(21.23)	(0.18)	(2.81)	(2.74)
TEXT_INNO_W ×YEAR12	-0.004***	-0.008***	-0.012***	-0.244***	-0.419***	-0.596***
	(-4.11)	(-6.57)	(-8.34)	(-5.99)	(-6.60)	(-7.42)
Cons	0.011***	0.020***	0.013**	-2.008***	-2.159***	-2.826***
	(3.72)	(4.42)	(2.40)	(-9.54)	(-7.15)	(-7.17)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	16349	14302	12321	14370	12401	10453
Adj. R-Squared	0.672	0.589	0.533	0.762	0.705	0.654

Panel C: 年报披露政策，描述性创新与盈余持续性

	Y=ROA(t+1)		Y=ROA(t+2)		Y=ROA(t+3)	
	(1)	(2)	(3)	(4)	(5)	(6)
	YEAR12=0	YEAR12=1	YEAR12=0	YEAR12=1	YEAR12=0	YEAR12=1
TEXT_INNO_W	0.628***	0.388***	0.504***	0.405***	0.477***	0.291***
	(6.71)	(6.49)	(4.55)	(5.59)	(4.99)	(3.67)

[7] 由于篇幅原因，正文中暂未列示稳健性检验的所有表格，如有需要，后续可作为附录予以呈现。

ROA	0.437***	0.403***	0.309***	0.310***	0.289***	0.273***
	(23.64)	(21.77)	(13.34)	(13.97)	(11.74)	(10.83)
C(TEXT_INNO_W)×C(ROA)	0.425***	0.262***	0.340***	0.274***	0.323***	0.197***
	(6.69)	(6.54)	(4.52)	(5.62)	(4.97)	(3.66)
Cons	-0.514***	-0.334***	-0.395***	-0.272***	-0.302***	-0.164***
	(-8.00)	(-7.92)	(-5.21)	(-5.23)	(-4.58)	(-2.95)
F-test (p value)	7.68(0.00)		1.11(0.35)		13.07(0.00)	
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	6825	9554	6806	7526	6797	5555
Adj. R-Squared	0.300	0.316	0.216	0.243	0.192	0.204

注：以上所有回归均控制了年度和行业固定效应，同时控制了公司层面的聚类效应。

2. 描述性创新（MD&A）的实证结果

管理层讨论与分析（以下简称 MD&A）作为财务报告中的重要组成部分，近年来受到学术界的广泛关注 (蒋艳辉 and 冯楚建, 2014 , 孟庆斌 et al., 2017 , 薛爽 et al., 2010)。在年报文本中，MD&A 在 2014 年之前大多隶属于董事会报告章节，在 2014 年之后则独立成一个章节。根据中国证监会相关财报披露准则[8]，MD&A 主要包括管理层回顾和展望两部分内容。管理层在回顾部分应当对上市公司报告期内的财务状况、经营成果、所处行业现状和外部环境等进行分析，而在展望部分则当进行公司未来发展战略、发展机遇、所面临风险、行业发展趋势等方面的分析。赵武阳 and 陈超 (2011) 指出 MD&A 部分是公司的研发描述性披露的主要来源之一。为了进一步检验我们描述性创新指标的合理性，我们对年报中的 MD&A 部分也进行了类似的描述性创新指标提取，回归结果中不论是描述性创新的影响因素，还是描述性创新对未来真实创新活动和盈余持续性的预测作用，描述性创新（MD&A）的回归结果与年报全文描述性创新的实证结论均保持一致。

3. 其他稳健性检验

除了上述稳健性检验之外，我们还进行了以下多类稳健性检验：

（1）用描述性创新（句频）表示描述性创新的披露水平：除了主回归中的词频表示方法，我们还使用了描述性创新关键词集所在句子数目占年报总句数的比例重新表示描述性创新，并分析其与真实创新活动和盈余持续性的关系。

（2）标记研发费用缺失或专利缺失的样本：我们对研发费用和专利缺失的样本设立相应缺失标记变量（即缺失取值为 0，否则取值为 1），而后将其作为控制变量加入主回归模型进行分析。

（3）去除 2015-2016 年样本：由于专利数目采用的是企业当年申请并截至 2016 年已获得授

[8]《公开发行证券的公司信息披露内容与格式准则第 2 号——年度报告的内容与格式（2012 年修订）》：http://www.csrc.gov.cn/pub/newsite/flb/flfg/bmgf/xxpl/xxplnr/201310/t20131017_236414.html；
《公开发行证券的公司信息披露编报规则第 13 号——季度报告的内容与格式（2016 年修订）》：http://www.csrc.gov.cn/pub/newsite/flb/flfg/bmgf/xxpl/xxplnr/201701/t20170111_309322.html。

权的专利数目，而专利从申请到获批往往存在一定的时间间隔，这可能导致离统计年份较近的年份专利申请获批数目不准确。因此我们将 2015-2016 年的样本予以剔除之后重新进行回归分析。

从以上检验的结果可以看出，描述性创新的披露不仅与当期的真实创新活动和盈余水平仍然显著正相关，而且同样可以正向预测未来的真实创新活动水平和盈余持续性，再次证明了实证结果的稳健性。

五、研究结论

本研究使用中文自然语言处理技术和基于深度学习思想的 Word2Vec 神经网络语言模型算法从沪深 A 股 2007~2016 年的年报中提取描述性创新文本指标，研究发现：描述性创新披露水平受到当期真实创新活动（研发投入和专利）和当期盈余水平的正向影响。不仅如此，描述性创新披露的水平还可以正向预测企业未来的真实创新活动水平和盈余持续性。这种预测效应即使在控制了研发投入和专利等信息之后依然显著，这说明描述性创新对企业创新水平的测度在传统的评价指标之外具有明显的增量反映作用。

除了描述性创新的总体披露水平，我们还研究了描述性创新披露的细节特征，包括描述性创新披露中的非前瞻性信息密度和语调。结果发现：描述性创新内容中的非前瞻性信息密度和语调积极程度不仅与当期真实创新活动水平显著正相关，还可以正向预测未来真实创新活动水平和盈余持续性。此外，证监会于 2012 年发布了关于年报内容和格式的进一步要求，对公司与创新活动相关披露做出了更加详细的指引和要求。通过分析我们发现：该政策的施行使得年报中描述性创新的披露水平有所增加，但描述性创新披露中的非前瞻性信息含量和语调积极程度却有所下降。与此同时，描述性创新对未来真实创新活动水平和盈余持续性的预测作用也有一定程度的减弱，该现象产生的原因可能与披露政策发布后年报中前瞻性信息含量的增加有关。

总体而言，本研究基于大样本文本分析探索了企业描述性创新信息及其披露细节的有效信息含量，在理论和实践层面具有重要的启示作用。首先，研究从描述性创新信息披露的视角证实了企业信息披露的信息有用性和价值相关性，丰富了目前国内关于文本型非结构化企业信息披露的相关研究。其次，随着大数据和人工智能技术的蓬勃发展，如何应用前沿技术的特点和优势，使其能够更好地服务于我们经济管理领域的科学研究，已经成为近年来学术界热烈讨论的话题。本文采用新兴机器学习技术和传统计量经济学方法相结合的方式探索了企业创新相关主题，或可为该话题的深度探讨带来一定的参考价值。最后，鉴于现有的企业创新水平衡量指标仍存在待改进之处，本研究所构建的描述性创新指标对企业以技术创新为主的创新能力评价体系的完善和发展具有重要借鉴意义，既有助于学术界对企业创新领域的深入研究，又有助于投资者进行相关的投资决策分析，同时还为监管部门制定和健全企业信息披露政策提供了有效的参考。

参考文献

[1] ACHARYA V V, BAGHAI R P, SUBRAMANIAN K V 2013. Labor laws and innovation. The Journal of Law and Economics [J], 56: 997-1037.

[2] AGHION P, VAN REENEN J, ZINGALES L 2013. Innovation and institutional ownership. American economic review [J], 103: 277-304.

[3] ASTHANA S C, ZHANG Y 2006. Effect of R&D investments on persistence of abnormal earnings. Review of Accounting and Finance [J], 5: 124-139.

[4] BALSMEIER B, FLEMING L, MANSO G 2017. Independent boards and innovation. Journal of financial economics [J], 123: 536-557.

[5] BARNEY J 1991. Firm resources and sustained competitive advantage. Journal of Management [J], 17: 99-120.

[6] BELLSTAM G, BHAGAT S, COOKSON J A 2017. A text-based analysis of corporate innovation [M]. SSRN working paper.

[7] BENGIO Y, DUCHARME R, VINCENT P, et al. 2003. A neural probabilistic language model. Journal of machine Learning research [J], 3: 1137-1155.

[8] BONSALL IV S B, LEONE A J, MILLER B P, et al. 2017. A plain English measure of financial reporting readability. Journal of accounting and economics [J], 63: 329-357.

[9] BUSHEE B J 1998. The influence of institutional investors on myopic R&D investment behavior. Accounting Review [J]: 305-333.

[10] BUSHEE B J, GOW I D, TAYLOR D J 2018. Linguistic Complexity in Firm Disclosures: Obfuscation or Information? Journal of Accounting Research [J], 56: 85-121.

[11] CHANG X S, CHEN Y, WANG S Q, et al. 2017. Credit default swaps and corporate innovation.

[12] CHEMMANUR T J, LOUTSKINA E, TIAN X 2014. Corporate Venture Capital, Value Creation, and Innovation. Review of Financial Studies [J], 27: 2434-2473.

[13] CHEN S, MATSUMOTO D, RAJGOPAL S 2011. Is silence golden? An empirical analysis of firms that stop giving quarterly earnings guidance. Journal of accounting and economics [J], 51: 134-150.

[14] COLLINS D W, KOTHARI S 1989. An analysis of intertemporal and cross-sectional determinants of earnings response coefficients. Journal of accounting and economics [J], 11: 143-181.

[15] D. BANKER R, MASHRUWALA R, TRIPATHY A 2014. Does a differentiation strategy lead to more sustainable financial performance than a cost leadership strategy? Management Decision [J], 52: 872-896.

[16] DAVIS A K, GE W, MATSUMOTO D, et al. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. Review of Accounting Studies [J], 20: 639-673.

[17] DE LA POTTERIE B V P 2008. Europe's R&D: missing the wrong targets? Intereconomics [J], 43: 220-225.

[18] DONTOH A 1989. Voluntary disclosure. Journal of Accounting, Auditing & Finance [J], 4: 480-511.

[19] ENTWISTLE G M 1999. Exploring the R&D disclosure environment. Accounting Horizons [J], 13: 323-342.

[20] FANG V W, TIAN X, TICE S 2014. Does stock liquidity enhance or impede firm innovation? The Journal of Finance [J], 69: 2085-2125.

[21] FIORDELISI F, RICCI O 2014. Corporate culture and CEO turnover. Journal of Corporate Finance [J], 28: 66-82.

[22] FRANCIS J, LAFOND R, OLSSON P M, et al. 2004. Costs of equity and earnings attributes. The accounting review [J], 79: 967-1010.

[23] GENNAIOLI N, SHLEIFER A, VISHNY R 2012. Neglected risks, financial innovation, and financial fragility. Journal of financial economics [J], 104: 452-468.

[24] GU F, LI J Q 2003. Disclosure of innovation activities by high-technology firms. Asia-Pacific Journal of Accounting & Economics [J], 10: 143-172.

[25] HARRIS Z S 1954. Distributional structure. Word [J], 10: 146-162.

[26] HE J, TIAN X 2018. Finance and Corporate Innovation: A Survey. Asia‐Pacific Journal of Financial Studies [J], 47: 165-212.

[27] HE J J, TIAN X 2013. The dark side of analyst coverage: The case of innovation. Journal of financial economics [J], 109: 856-878.

[28] HENDRICKS K B, SINGHAL V R 1997. Delays in new product introductions and the market value of the firm: The consequences of being late to the market. Management science [J], 43: 422-436.

[29] HITT M A, HOSKISSON R E, JOHNSON R A, et al. 1996. The market for corporate control and firm innovation. Academy of Management Journal [J], 39: 1084-1119.

[30] HOUSTON J F, LEV B, TUCKER J W 2010. To guide or not to guide? Causes and consequences of stopping quarterly earnings guidance. Contemporary Accounting Research [J], 27: 143-185.

[31] JAMES S D 2011. Strategic R&D disclosure and competition [M]. Working Paper, Ohio State University.

[32] JONES D A 2007. Voluntary disclosure in R&D‐intensive industries. Contemporary Accounting Research [J], 24: 489-522.

[33] KOTHARI S, LAGUERRE T E, LEONE A J 2002. Capitalization versus expensing: Evidence on the uncertainty of future earnings from capital expenditures versus R&D outlays. Review of Accounting Studies [J], 7: 355-382.

[34] KRISHNAN G V, PARSONS L M 2008. Getting to the bottom line: An exploration of gender and earnings quality. Journal of Business Ethics [J], 78: 65-76.

[35] LARCKER D F, ZAKOLYUKINA A A 2012. Detecting deceptive discussions in conference calls. Journal of Accounting Research [J], 50: 495-540.

[36] LECUN Y, BENGIO Y, HINTON G 2015. Deep learning. nature [J], 521: 436.

[37] LEHAVY R, LI F, MERKLEY K 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. The accounting review [J], 86: 1087-1115.

[38] LI F 2008. Annual report readability, current earnings, and earnings persistence. Journal of accounting and economics [J], 45: 221-247.

[39] LI F 2010(a). Survey of the Literature. Journal of accounting literature [J], 29: 143-165.

[40] Li F 2010(b). The information content of forward-looking statements in corporate filings—A naive Bayesian machine learning approach[J]. Journal of Accounting Research[J], 48(5): 1049-1102.

[41] LOUGHRAN T, MCDONALD B 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10‐Ks. The Journal of Finance [J], 66: 35-65.

[42] LOUGHRAN T, MCDONALD B 2016. Textual analysis in accounting and finance: A survey. Journal of Accounting Research [J], 54: 1187-1230.

[43] MANSO G 2011. Motivating innovation. The Journal of Finance [J], 66: 1823-1860.

[44] MERKLEY K J 2013. Narrative disclosure and earnings performance: Evidence from R&D disclosures. The accounting review [J], 89: 725-757.

[45] MIKOLOV T, SUTSKEVER I, CHEN K, et al. 2013. Distributed representations of words and phrases and their compositionality [C] //; City. 3111-3119.

[46] MILLER B P 2010. The effects of reporting complexity on small and large investor trading. The accounting review [J], 85: 2107-2143.

[47] MUKHERJEE A, SINGH M, ŽALDOKAS A 2017. Do corporate taxes hinder innovation? Journal of financial economics [J], 124: 195-221.

[48] MUSLU V, RADHAKRISHNAN S, SUBRAMANYAM K, et al. 2014. Forward-looking MD&A disclosures and the information environment. Management science [J], 61: 931-948.

[49] NAGAR V, NANDA D, WYSOCKI P 2003. Discretionary disclosure and stock-based incentives. Journal of

accounting and economics [J], 34: 283-309.
[50] NEKHILI M, BOUBAKER S, LAKHAL F 2012. Ownership structure, voluntary R&D disclosure and market value of firms: the French case. International Journal of Business [J], 17: 126.
[51] NEKHILI M, HUSSAINEY K, CHEFFI W, et al. 2016. R&D narrative disclosure, corporate governance and market value: Evidence from France. Journal of Applied Business Research [J], 32: 111-128.
[52] Price S M K, 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.
[53] RICHARDSON S A, SLOAN R G, SOLIMAN M T, et al. 2005. Accrual reliability, earnings persistence and stock prices. Journal of accounting and economics [J], 39: 437-485.
[54] RONG X 2014. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738 [J].
[55] SADKA G 2004. Financial reporting and product markets: Learning from competitors.
[56] SEGERSTROM P S 1991. Innovation, imitation, and economic growth. Journal of political economy [J], 99: 807-827.
[57] SLOAN R G 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Accounting Review [J]: 289-315.
[58] TIAN X, WANG T Y 2011. Tolerance for failure and corporate innovation. The Review of Financial Studies [J], 27: 211-255.
[59] UOTILA J, MAULA M, KEIL T, et al. 2009. Exploration, exploitation, and financial performance: analysis of S&P 500 corporations. Strategic Management Journal [J], 30: 221-231.
[60] WERNERFELT B 1984. A resource‐based view of the firm. Strategic Management Journal [J], 5: 171-180.
[61] XU C, YAN M 2014. Radical or incremental innovations: R&D investment around CEO retirement. Journal of Accounting, Auditing & Finance [J], 29: 547-576.
[62] 陈怡欣, 张俊瑞, 汪方军 2018. 卖空机制对上市公司创新的影响研究——基于我国融资融券制度的自然实验. 南开管理评论 [J], 21: 62-74.
[63] 窦欢, 陆正飞 2017. 大股东代理问题与上市公司的盈余持续性. 会计研究 [J]: 24-31+88.
[64] 方红星, 张志平 2013. 内部控制对盈余持续性的影响及其市场反应——来自 A 股非金融类上市公司的经验证据. 管理评论 [J], 25: 77-86.
[65] 冯根福, 温军 2008. 中国上市公司治理与企业技术创新关系的实证分析. 中国工业经济 [J]: 91-101.
[66] 韩鹏, 彭韶兵 2012. 研发信息披露质量测度及制度改进. 财经科学 [J]: 103-110.
[67] 韩鹏, 岳园园 2016. 企业创新行为信息披露的经济后果研究——来自创业板的经验证据. 会计研究 [J]: 49-55+95.
[68] 姜付秀, 石贝贝, 李行天 2015. “诚信”的企业诚信吗?——基于盈余管理的经验证据. 会计研究 [J]: 24-31+96.
[69] 蒋艳辉, 冯楚建 2014. MD&A 语言特征、管理层预期与未来财务业绩——来自中国创业板上市公司的经验证据. 中国软科学 [J]: 115-130.
[70] 蒋艳辉, 马超群, 熊希希 2014. 创业板上市公司文本惯性披露、信息相似度与资产定价——基于 Fama-French 改进模型的经验分析. 中国管理科学 [J], 22: 56-63.
[71] 李常青, 钟娟, 王毅辉 2008. 上市公司前瞻性信息披露动因研究. 统计与决策 [J]: 135-137.
[72] 林乐, 谢德仁 2017. 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. 管理世界 [J]: 125-145+188.
[73] 刘运国, 刘雯 2007. 我国上市公司的高管任期与 R&D 支出. 管理世界 [J]: 128-136.
[74] 孟庆斌, 杨俊华, 鲁冰 2017. 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. 中国工业经济 [J]: 132-150.
[75] 潘越, 潘健平, 戴亦一 2015. 公司诉讼风险、司法地方保护主义与企业创新. 经济研究 [J], 50: 131-145.

[76] 王红建, 曹瑜强, 杨庆, et al. 2017. 实体企业金融化促进还是抑制了企业创新——基于中国制造业上市公司的经验研究. 南开管理评论 [J], 20: 155-166.
[77] 王雄元, 高曦 2018. 年报风险披露与权益资本成本. 金融研究 [J]: 174-190.
[78] 王雄元, 李岩琼, 肖丞 2017. 年报风险信息披露有助于提高分析师预测准确度吗? 会计研究 [J]: 37-43+96.
[79] 魏江, 寒午 1998. 企业技术创新能力的界定及其与核心能力的关联. 科研管理 [J]: 13-18.
[80] 温军, 冯根福 2012. 异质机构、企业性质与自主创新. 经济研究 [J], 47: 53-64.
[81] 文芳, 胡玉明 2009. 中国上市公司高管个人特征与 R&D 投资. 管理评论 [J], 21: 84-91+128.
[82] 肖华, 张国清 2013. 内部控制质量、盈余持续性与公司价值. 会计研究 [J]: 73-80+96.
[83] 谢德仁, 林乐 2015. 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. 会计研究 [J]: 20-27+93.
[84] 薛爽, 肖泽忠, 潘妙丽 2010. 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. 管理世界 [J]: 130-140.
[85] 薛云奎, 王志台 2001. R&D 的重要性及其信息披露方式的改进. 会计研究 [J]: 20-26+65.
[86] 阎达五, 孙蔓莉 2002. 深市 B 股发行公司年度报告可读性特征研究. 会计研究 [J]: 10-17+64.
[87] 虞义华, 赵奇锋, 鞠晓生 2018. 发明家高管与企业创新. 中国工业经济 [J]: 136-154.
[88] 袁建国, 后青松, 程晨 2015. 企业政治资源的诅咒效应——基于政治关联与企业技术创新的考察. 管理世界 [J]: 139-155.
[89] 张兆国, 刘亚伟, 杨清香 2014. 管理者任期、晋升激励与研发投资研究. 会计研究 [J]: 81-88+97.
[90] 赵武阳, 陈超 2011. 研发披露、管理层动机与市场认同:来自信息技术业上市公司的证据. 南开管理评论 [J], 14: 100-107+137.

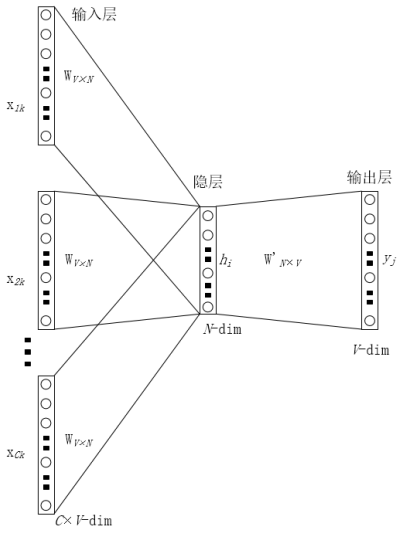
附 录

附录 1. 年报文本处理流程

WinGo 财经文本数据平台是一家以研究财务报告等上市公司披露文本为主的人工智能财经数据平台，本文从该平台获取了中国沪深 A 股上市公司年报的相关文本数据，对年报文本的处理主要包括以下步骤：（1）从上海证券交易所、深圳证券交易所、巨潮资讯等网站下载原始 PDF 文件，并结合 PDF 文件解密技术、OCR（光学字符识别）技术将其解析成 TXT 格式文本；（2）对文本进行去除表格、去除页眉页脚等清洗操作；（3）使用中文财经专用分词系统对文本进行分词，将连续的年报文本序列切分成若干单独的词汇，用于计算词汇的词频等信息。不同于天然以空格为间隔符的英文文本，中文分词是进行中文文本分析时最为重要的预处理环节。相比于直接在未经分词的文本中查询关键词计数的词频统计方法，使用经过中文分词处理后的关键词词频方法的优势在于：（1）当进行关键词查询如“创新”时，“再创新高”这类和“创新”无关的词不会被错误地计入关键词词频；（2）当进行关键词查询如“资产”时，“资产负债表”这类和“资产”本身所指代对象差别较大的词不会被错误地计入关键词频。因此，使用基于财经专用系统的分词结果保证了本研究中描述性创新指标的构建结果更为准确有效。

附录 2. Word2Vec 神经网络模型简介

Word2Vec 神经网络模型由 Mikolov et al. (2013) 提出，是近年来深度学习领域的里程碑式成果 (LeCun et al., 2015)。Word2Vec 模型根据上下文内容将词语表征为实数值向量，并通过向量的相似度计算得到词语之间的语义相似性 (Bengio et al., 2003)。在实际应用中，Woed2Vec 主要分为 CBOW (Continuous Bag of Word) 和 Skip-gram 两种模型，我们采用 CBOW 模型进行训练。CBOW 的基本思想是根据上下文来预测当前词语的概率，该模型的基本框架 (Rong, 2014) 如附图 1 所示。



附图 1 Word2Vec (CBOW)的基本框架

在图 1 所示 CBOW 模型中，输入词序列 S 中的第 k 个词 w(k)，用它的上下文来预测它本

身，即根据附近的 C 个词，预测第 k 个词。

模型中隐层的输出为：

$$\mathrm{h} = \frac{1}{C} W^T (x_1 + x_2 + \cdots + x_C)$$

$$= \frac{1}{C} (v_{w_1} + v_{w_2} + \cdots + v_{w_C})^T$$

其中，W 是所有词对应的输入向量矩阵，$w_1, \ldots, w_C$ 是 w(k)的上下文，C 是上下文窗口大小，x 是上下文的 one-hot 向量，v 是上下文对应的输入向量。

则词汇表中词w_j的得分u_{w_j}为：

$$u_{w_j} = {v'_{w_j}}^T h$$

其中v'_{w_j}是词汇表中词w_j对应的输出向量。当上下文为 $w_I = (w_1, w_2, \ldots, w_C)$时，w(k)出现的概率为：

$$\mathrm{p}(w_k | w_I) = y_k = \frac{\exp(u_{w_k})}{\sum_{j=1}^{V} \exp(u_{w_j})}$$

上式即为目标函数，最大化该目标函数，等价于最小化损失函数 E：

$$\mathrm{E} = -\log P(w_k | w_I)$$

$$= -\log \frac{\exp(u_{w_k})}{\sum_{j=1}^{V} \exp\left(u_{w_j}\right)}$$

$$= -u_{w_k} + \log \sum_{j=1}^{V} \exp\left(u_{w_j}\right)$$

$$= -{v'_{w_k}}^T h + \log \sum_{j=1}^{V} \exp\left({v'_{w_j}}^T h\right)$$

通过最小化上述损失函数，可优化词表的输入矩阵和输出矩阵，最终得到词对应的 Word2Vec 词向量。

附录 3. 描述性创新代表性企业示例

附表 1 列示了农、林、牧、渔业，制造业，电力、热力、燃气及水生产和供应业以及科学研究和技术服务业按描述性创新指标排序的前五名和后五名的企业。农、林、牧、渔业描述性创新指标最大的五个企业中，排名第一、第四的登海种业和隆平高科分别是以“南袁北李”享誉中国种业界的李登海和袁隆平发起设立的企业；制造业的前五名企业中，排名第三的朗科科技获得过中国知识产权的最高奖项—第十五届中国专利金奖；排名第五的中科曙光连续 8 年蝉联中国高性能计算机 TOP100 排行榜市场份额第一，是中国高性能计算、服务器、云计算、大数据领域的领军企业。

附表 1 部分行业描述性创新代表性企业示例

行业	排名前五位的企业	排名后五位的企业
农、林、牧、渔业	登海种业	海南橡胶
	神农基因	福成股份
	东方海洋	香梨股份

	隆平高科	开创国际
	荃银高科	中水渔业
制造业	神思电子	华资实业
	机器人	富奥股份
	朗科科技	深中华A
	海利生物	山东金泰
	中科曙光	中再资环
电力、热力、燃气及水生产和供应业	科林环保	国新能源
	迪森股份	深圳燃气
	富春环保	新天然气
	节能风电	新疆浩源
	创业环保	百川能源
科学研究和技术服务业	博济医药	中矿资源
	华电重工	贝瑞基因
	柏堡龙	华建集团
	电科院	中国海诚
	国检集团	百花村

注：以上企业基于 2007-2016 年全样本数据（不包括 ST 企业）的上市公司描述性创新指标均值排序生成，数据未经任何剔除和缩尾操作。

下文列示了各行业描述性创新指标排名前五企业中部分企业的年报内容摘录。

（1）农、林、牧、渔业：登海种业（2009）

作为全国首批农业首个国家创新型企业，科研创新是公司历年工作的重中之重。2009 年公司围绕“自主创新，建设创新型登海种业”这一中心任务，整合优化现有科研资源，持续增加科研投入，推动科研扩规模、增总量、提层次，壮实力，取得了丰硕的科研成果。

报告期内，公司科研项目立项新增 9 项，其中“紧凑耐密型高产玉米新品种配套技术试验示范”、“高产优质多抗玉米新品种培育”等 5 项主承担项目，参与承担项目 4 项（含国家转基因重大专项课题 3 项）；公司共有 4 个玉米新品种通过审定，其中登海 661 通过山东省审定，登海 662 通过国家、山东省及河南省审定，登海 701 通过山东省及河南省审定，登海 3769 通过国家审定；承担国家级、省级及国内外公司安排的试验 33 项，申请品种权 15 项，获得品种权 24 项；2009 年 5 月，“一种绿色糯玉米的选育方法”获得发明专利授权。

2009 年 9 月，公司通过“山东省专利明星企业”复审；10 月，公司被评为“山东省知识产权试点单位”。经山东省知识产权局批准，公司被确定为“山东省专利能力培育单位”。

（2）制造业：朗科科技（2009）

本公司基于闪存应用及移动存储领域内持续自主创新的全球领先技术及专利，专业从事闪存应用及移动存储产品的研发、生产、销售及相关技术的专利运营业务。本公司在全球范围内拥有闪存盘相关领域的系列原创性基础发明专利、闪存应用及移动存储领域其他核心技术及其专利，凭借专业的技术创新与研发平台、成熟的专利运营体系、知名的品牌和多渠道营销网络，与多家全球知名企业建立了战略合作关系，面向全球进行产品销售和专利授权许可，以实现公司长期可持续发展。

（3）电力、热力、燃气及水生产和供应业：科林环保（2016）

公司持续优化生产经营管理模式，进一步巩固国内外业务市场，在垃圾、固废焚烧烟气治理等新兴市场得到有效拓展。依托企业博士后工作站以及科林国家级企业技术中心平台，不断完善和提升了各项超低排放技术。紧紧围绕市场需求，加大工程技术和各类袋式除尘装备的技术集成和研发，并延伸脱硫、脱硝、脱重金属、脱二噁英和 VOCs 等技术的研发和示范应用。加快固废领域及其他行业袋式除尘器专有技术的发展及工程成套技术与装备的设计优化；加快对高中温烟气 SCR 脱硝技术、超高温除尘器技术、塔式湿法脱硫脱硝技术、高效除雾等技术的合作与研发。