企业年报信息披露的战略内容分析
——基于语义分析系统

胡小鹏，袁琦，管东升
（中国电子信息产业发展研究院，北京，100048）

摘　要：介绍使用计算机系统分析企业年报信息披露的战略内容的国内外发展现状，提出利用USAS语义分析系统（UCREL Semantic Analysis System）提取中文年报战略内容核心术语和概念的研究方法。收集2007~2014年发表的10个有代表性的中国上市公司、且产业覆盖率较宽的57份年报，着重关注与企业战略描述密切相关的3个语义域——I（钱财和商业）、H（建筑和建筑物）和Y（科学技术），通过人工匿名评分，开展全面、客观的有效性评估。对实验结果的分析表明，利用开发的中文语义分析系统从企业年报信息披露文本中自动识别和提取中文关键术语用于企业战略信息的进一步分析是可行的。

关键词：企业年报信息披露；中文语义分析系统；语义域；企业财务信息环境（CFIE）项目
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**2095-8412 (2017) 01-001-07
工业技术创新 URL: http: //www.china-iti.com　　DOI: 10.14103/j.issn.2095-8412.2017.01.001

引言

近年来，采用自然语言处理（NLP，Natural Language Processing）方法自动分析大规模企业年报信息披露内容的研究受到日益关注，其研究目的是向关心企业和商业公司业绩的利益相关者（如投资者、供应商和监管机构）提供有关企业经营发展等战略内容的信息。所谓战略内容（Strategic Content），是指企业年报信息披露中涉及企业商业模式和财务陈述核心内容的部分，其中包括为达到预定的企业目标而采取的步骤和措施。

在经济领域，人们很早就认识到了研究和分析企业年报的重要性，并且进行了大量研究，但是上述研究中的大多数是基于小数据集或人工分析方法。目前我国企业年报信息披露质量评价仍然采用传统人工评价方法，几乎没有采用NLP方法的自动评价技术成果发表。

迄今为止，大多数关于企业年报文本内容自动分析的研究是针对英文年报进行的，而针对中文年报的相应研究成果还鲜为人知。本文的研究旨在消除这一差距，在国内外率先开展利用中文语义分析工具自动分析中文企业年报战略内容的研究与实验。具体而言，该项研究集中自动识别和提取企业年报核心中文术语和概念，在大量收集中文企业年报的基础上，这些核心术语和概念可用于搜索与商务战略相关的信息，并为进一步研究基于统计与语义相结合的模型挖掘年报战略内容奠定基础。本文首先介绍国内外基于计算机系统分析企业年报信息披露战略内容的研究现状；然后提出利用USAS语义分析系统提取中文年报战略内容核心术语和概念的研究方法；最后给出实验结果和结论，并提出展望。

1 研究进展及存在的问题

国际上利用NLP方法自动分析大规模企业年报文本内容的研究正在受到学术界的关注，包括自动剖析年报文档结构、识别和分析年报信息披露内容等。目前，使用计算机系统分析企业年报信息披露战略内容的研究大致包括四方面的内容：①自动识别和收集年报阐明公司战略内容的文本章节，这些章节的信息披露明确表达了公司将来努力的方向，以及实现其商业目标需要采取的步骤和措施，例如企业重组、提高产品和服务质量等；②挖掘年报特定语言表达，评估年报信息披露质量，利用语义分析工具实现抽象语义层面分析；③通过建立和使用可读性、前瞻性表述、模糊限制语、肯定/否定语气等年报特定语言表达的度量标准，评估年报的特征和质量，评价企业的业绩和置信度；④采用语义分析工具提取核心战略内容的语义模式，实现对企业年报的战略内容分析。

1.1 国内外研究进展

近年来，研究主要集中在信息披露的语气、业绩和归因（EI-Haj等[1]）以及不确定性（Rayson等[2]）等语言特征研究。同时，近年来有关这一研究课题的学术会议和出版物不断增加，也体现了人们对这一研究领域的更多关注。最近两年的主要研究包括：Brennan等[3]从经济学、心理学、社会学、批判学这四个角度分析会计报表，研究了年报信息披露中的印象管理问题。Merkl-Davies等[4]研究了企业陈述性报告研究中的文本分析方法的类型，企业陈述性文档应用的方法论指导，以及评价此类研究质量的标准。Merkl-Davies等[5]也提出了一个针对企业陈述性报告研究文本分析方法的分类方案。Athanasakou和Hussainey[6]研究了前瞻性业绩信息披露的可信度。Athanasakou等[7]评测了度量企业信息披露战略内容的方法。Veronika Koller[8]采用USAS英文语义分析工具（Rayson等[9]）以及Wmatrix语料库分析网站（Rayson[10]）从语料库语言学和信息披露分析方面研究了英国企业年报中的情感和理性问题。近期的几次国际学术会议也展示了这一领域的最新研究成果，例如2014财务沟通话语方法国际学术会议(DAFC)[11]、第七届[12]和第八届[13]LSE/LUMS/MBS学术会议，其研究主题集中在“财务报告质量的构成要素是什么”和“探索陈述性报告在企业财务信息环境中的作用”。例如，Hoberg和Lewis[13]将语言学分析和统计指标（词汇向量的余弦相似度和潜在狄利克雷分布）用于检测企业虚假信息披露。Young[14]认为，NLP技术可以用于解决企业信息披露的各种问题，包括（1）识别虚假报告；（2）衡量情绪变化（市场和个股）；（3）投资战略和预期表现；（4）理解价格形成过程和目标调整；（5）挖掘观点；（6）衡量企业信息披露中的内容；（7）衡量企业特点；（8）评估股票分析师。Young和El-Haj[15]在文章中报告了采用NLP技术（例如可读性、搭配和重要词汇索引等）和Wmatrix系统（Rayson和El-Haj[16]）对英国企业年报样例的分析，并且提供了大量的文本分析观点，包括计算可读性指标、采用预设列表的字数统计（例如前瞻性表述、不确定性和语气等）、基于自定义词表的字数统计、与参考语料库的对比（单词、词性和语义）、重要词汇索引和组合等（Rayson和El-Haj[16]）。Lang和Stice-Lawrence[17]对全球企业年报财进行了研究。

虽然这一研究领域取得了重大发展，但瓶颈仍然存在，制约了研究成果在解决企业实际问题方面的应用。

1.2 存在的问题

目前仍缺乏能够自动分析大规模企业年报和信息披露的技术和软件工具。现有的大多数工具仍为人工或半自动工具，因此只能处理小规模的文件。最近，这一问题开始引起人们的关注，有代表性的是英国企业财务信息环境（CFIE）项目[18]。

企业财务信息环境（CFI）项目是涉及商学、金融学和计算语言学等高度跨学科的研发项目，在独立研发两年（2013~2014）后，目前作为社会科学语料库方法（CASS）项目的一部分继续实施。CFIE项目的目标是：

（1）通过开发一套基于语料库的自然语言处理（NLP）工具分析公司的陈述性沟通实践，推动对公司信息披露的词汇属性和陈述内容的研究；

（2）使用目标（1）中研发的方法和工具，度量企业强制性和自愿性信息披露的语言特征，找出这些语言特征随不同公司变化的决定因素，并将这些语言特征与信息披露的信息有效性联系起来；

（3）使用目标（1）中研发的方法和工具促进对企业自愿性信息披露和会计质量之间的相互关系研究，包括企业信息披露和盈余管理实践对于企业股利预期收益以及在公开的投资者关系质量排名中公司所处位置的联合效应；

（4）运用NLP评分法分析英国财经媒体的企业新闻报道，目的是研究和实现针对企业年报信息披露质量的更完整的度量方法；

（5）利用上述目标（1）~（4）的方法和见解，为盈余质量、信息披露质量和资本成本之间的联系提供新的证据[18]。

到目前为止，CFIE项目产生了四项主要成果，其中代表性的成果是创建了一个公开可用的网络工具，允许大批量评分英国企业年报的陈述性内容。图1显示了嵌入在CFIE网络工具中的处理流程。

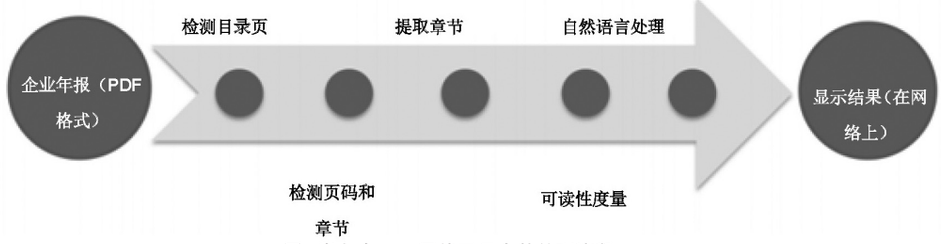

图1 嵌入在CFIE网络工具中的处理流程

处理流程的主要特征包括使用新颖的混合文本分析方法检测英国年报的结构，统计词频，计算可读性和其他信息披露质量度量标准，其目的是对年报的个别章节（如战略、绩效、治理等章节）进行自动评分。

目前我国企业年报信息披露质量的评价仍然采用人工评价方法，自动分析企业年报信息披露的研究尚未受到应有的重视，截至目前很少有实用成果发表。XinyingQiu和Shengyi Jiang等[19]于2013年发表了研究中文年报信息披露质量的自动评估系统的成果，该评估系统的评估体系框架整合了不同的技术，其中包括中文文档建模，中文可读性指标构建，以及多级分类。Scott Piao（朴松林）和Xiaopeng Hu（胡小鹏）等[20]于2015年发表了使用中文语义分析系统从中文企业年报信息披露文档中自动识别和提取年报战略内容相关术语的研究成果。他们对实验结果的分析表明，利用语义分析器自动提取中文关键术语用于企业战略信息的进一步分析是可行的。

2 基于中文语义的课题研究

近年来，中国电子信息产业发展研究院与英国Lancaster大学合作，探讨和研究企业年报信息披露质量问题，即通过NLP技术检测语言特征，分析年报战略内容。所涉及的合作研究内容包括：（1）研究剖析中文企业年报文档结构，从中提取特定章节；（2）研究中文企业年报特定语言的度量标准，评估中文企业年报信息披露质量；（3）研究中文语义分析系统，对年报战略内容进行语义分析；（4）利用以上研究成果自动识别提取和分析企业年报的战略内容。本章重点介绍目前中文语义分析系统对年报战略内容进行语义分析所取得的阶段性研究成果。

2.1 中文语义分析系统机理

本文提出的对年报战略内容进行语义分析的中文语义分析系统采用Lancaster大学USAS语义分析系统（UCREL Semantic Analysis System）的框架和技术[21]。USAS语义分析系统的最初研发目的是提供英文文本的词汇语义标注。截至2016年，USAS语义分析系统已经扩展到包括中文在内的12种语言的大型多语种语义词典[22]。

本项研究所使用的中文语义分析系统是通过以下步骤转换生成的：

（1）通过自动映射和语义消歧生成中文语义词汇；

（2）扩充中文年报战略内容语义词汇；

（3）设计中文语义分析系统结构；

（4）分析不同语义域的词汇分布；

（5）分析和评估系统性能。

通过上述步骤生成的中文语义分析系统结构框图如图2所示。

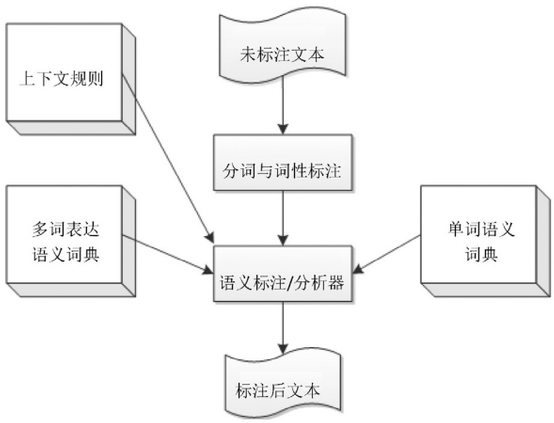

图2 中文语义分析系统结构框图

USAS系统与WordNet和HowNet截然不同。WordNet是基于同义词网络的词汇数据库，而HowNet是基于概念关系网络的知识系统。USAS系统不是纯粹的词汇数据库或知识库，而是一个无缝连接语义词汇和词义消歧工具的自然语言处理软件系统，其中包括分词工具、词性标注工具和其他辅助工具。除了标准语义词典词条，它还包含灵活的模板，可识别复杂结构和非连续结构的多词表达（MWE）。同时，由于USAS语义分类系统起源于稳定、通用的标准英文分类汇编（thesaurus）词典，因此使得USAS具有重大的实际应用价值，可执行大规模语义分析。由于USAS的语义词典和捆绑软件是紧密集成的，因此可以方便地通过在语义词典引入新的语义域和标记，调整USAS系统，使其适应分析新的语义领域。这对于企业年报内容分析具有重大的现实价值，因为除了通用语义域之外，还需要处理与企业年报和企业战略相关的特定领域的词义，同时还要处理模糊限制语、前瞻性表达等词汇语义。表1给出了USAS、WordNet和HowNet的重要特征对比。表2中列出了USAS顶层的21个标记。

2.2 中文企业年报战略内容分析

实现自动分析业年报战略内容研究的第一步是进行提取年报战略内容核心术语的研究和实验，其目的是考察识别和提取企业年报中涉及描述企业商务策略的核心中文术语的可行性。采集到的这些术语可以帮助生成有关企业商务报告所反映出的公司战略的宏观总结，并为进一步的战略内容分析提供数据搜索点。显然，并非语义系统中所有的语义域都涉及到企业战

表1 USAS、WordNet和HowNet重要特征对比

特征	USAS	WordNet	HowNet
知识库	与标注软件结构融合的分类汇编（thesaurus）式词典知识库	基于同义词的词汇网络、有单独开发的搜索工具	基于知识系统的概念关联，有搜索工具
自然语言处理	包含全套自然语言处理工具：分词、POS 标注、词形还原、语义标注	不适用	不适用
灵活性	易于通过修改词库的和语义分类调整软件系统。	不易修改	不易修改
语义模板机制	有	无	无
适用语言	正在扩展到更多语言	正在扩展到更多语言	英文和中文
文本挖掘功能	与关键度（keyness）方法结合，检测关键语义域，用于社科领域文本挖掘。	无	无

表2 USAS 的大类语义分类标记

标记	语义域名称	标记	语义域名称	标记	语义域名称
A	一般术语和抽象术语	I	钱财和商业	Q	语言和通信
B	团体和个体	K	娱乐、运动和游戏	S	社交行动、状态和过程
C	艺术和工艺	L	生命和生活	T	时间
E	情感	M	移动、位置、旅行和运输	W	世界和环境
F	食物和耕种	N	数量和指标	X	心理活动、状态和过程
G	政府和公众	O	物质、材料、对象和设备	Y	科学技术
H	建筑和建筑物	P	教育	Z	名称和语法

表3 语义分析系统对10家公司信息披露的测试数据量和词汇覆盖率

	公司1	公司2	公司3	公司4	公司5	公司6	公司7	公司8	公司9	公司10	总计
文件数	5	4	7	5	6	6	6	7	6	5	57
字符数	336 133	251 686	245 275	264 416	342 622	276 025	570 537	593 088	422 743	282 431	3 584 956
词汇覆盖率	76.91%	75.41%	77.08%	75.57%	78.26%	75.95%	74.86%	75.57%	74.32%	78.16%	76.02%

略信息。因此在这项研究中，选择了我们认为是与公司年报战略内容密切相关的三个主要USAS语义域。这3个选定的语义域分别用英语字母表示如下：（1）I–钱财与商业；（2）H–建筑与建筑物；（3）Y–科学技术。

2.2.1 企业年报收集

作为实验测试数据，从中国证券公司网站（http://www.nbdqw.com/）下载了一批涉及企业信息披露的年报。这些公司年报为了达到商业促进的目的，公布了公司近年来的业绩和成就，并发布了它们将来的计划和战略。这种类型的数据为使用NLP方法分析和预测公司运营状况提供了有价值的资源。为了保证广泛的代表性和测试数据的高品质，人工选择了从2007~2014年发表的10个有代表性的中国上市公司的年报（在本文中用匿名代表这些公司），总共收集了57份年报包括3 584 956个中文字符。从行业领域来看，这些年报有较宽的产业覆盖率，涵盖高科技产业、建筑、汽车生产、化学工业、石油工业等。使用样本报告研究并实验了自动提取不同行业类型公司的商业战略信息的方法。在这个特定的实验中，使用样本数据测试自动识别有关商务战略的核心中文术语的方法。为了估计公司数据对于语义分析系统的词汇覆盖率，计算了语义分析系统识别的样本报告中的中文词的百分比。表3给出了10家公司的测试数据分布和中文语义分析系统的

词汇覆盖率。表中的各列表示各家公司，表中的三行分别表示给出的文件数、文件的中文字符量以及词汇覆盖率。

2.2.2 核心语义术语提取和人工评价

从PDF文档中提取的文本文件使用中文语义分析系统进行处理。分别处理每个公司年报是为了考察不同类型的企业业务如何影响这一方法和结果。然后，对于每一家公司，分别收集了用USAS的三个主要语义域I、H和Y标记标注的术语。接下来，收集了这些带有标记的术语的频率（例如，资产_I1.1），然后对每一家公司和3个主语义域中的每一个，选择频率最高的100个术语标记对。其结果是对于每个公司，获得3个术语标记频率列表，该列表含有表4所示格式的条目（方括号内是用于人工评价者输入评价分数）。

表4 术语标记频率列表格式示例

频率值	词语-标记	人工评价
2866	资产_I1.1	[ ]

最后，邀请熟悉企业年报的评估人员使用表5中的5级评分标准对每个术语打分。

表5 评估人员评分等级标准

评分等级	评分判据	举例
5	与企业战略信息披露紧密相关	资产、工厂设备
4	与企业战略信息披露比较相关	方法、计划
3	一般性名词、动词等	目录、计算
2	有意义的单字词	层、台
1	无意义的单字	璃，志

根据表5的评分等级标准，评级分值3指示中性用语。因此，如果一个术语的评分是5或4的分数，它被认为是与企业战略信息披露相关。此外，如果一组术语的平均评分等级大于3，则表明该组术语全部带有企业战略相关的信息，并且分数值越高，它们携带的战略信息越多。按照这个判据，在这一实验中，通过观察由该方法提取出的术语所获得的平均分数测量该方法的性能，平均得分大于3表示基本成功，平均得分为5表示最大成功。

2.2.3 有效性评估

根据人工评价者给出的分数，结合依据每个USAS语义域对与企业战略相关的术语提取的有效性，评估方法的性能和效果。具体而言，对于每家公司，对USAS的I、H和Y三个语义域各自计算评估人员的平均评分，以考察划入这些语义域的术语携带企业战略信息披露相关的信息达到何种程度。

表6所示为评分结果。语义域I产生了最佳平均评分结果，为3.60。鉴于评价得分为3.0表示中性术语，平均得分超过3.0意味着很多提取的术语包含与企业战略内容相关的某些信息。此外，可观察到由PDF文档转换到文本的错误以及中文分词过程错误引起的一些断裂的词也包括在术语列表中，可能影响结果。因此，过滤单字词和无意义的单字，并重新计算人工评分的统计信息，结果如表7所示。

初步比较表6和表7，表明破碎的词语确实影响了结果。例如，通过去除中文单字词和单字，语义域I的平均打分提高了0.13。图3、图4和图5分别直观地表示出语义域I、H和Y通过过滤所实现的改进，其中实线和虚线分别表示三个语义域不带有和带有单字词和单字的平均评分线，充分证明过滤之后的评分明显得到了改进。

表6 10家公司年报的语义域人工评分初步统计

语义域	公司1	公司2	公司3	公司4	公司5	公司6	公司7	公司8	公司9	公司10	平均
I	4.33	3.52	3.66	3.8	4.18	3.44	3.27	3.32	3.28	3.25	3.60
H	3.43	3.07	3.02	3.15	3.06	3.06	3.28	3.07	3.4	3.28	3.182
Y	3.13	2.9	3.04	3.53	3.07	2.83	2.83	2.93	2.97	2.92	3.01

表7 过滤后10家公司年报的语义域人工评分统计

语义域	公司1	公司2	公司3	公司4	公司5	公司6	公司7	公司8	公司9	公司10	平均
I	4.5	3.65	3.76	3.88	4.30	3.55	3.40	3.45	3.41	3.38	3.73
H	3.89	3.48	3.44	3.51	3.38	3.37	3.31	3.29	3.77	3.61	3.50
Y	3.45	3.30	3.33	3.83	3.45	3.17	3.15	3.21	3.15	3.18	3.32

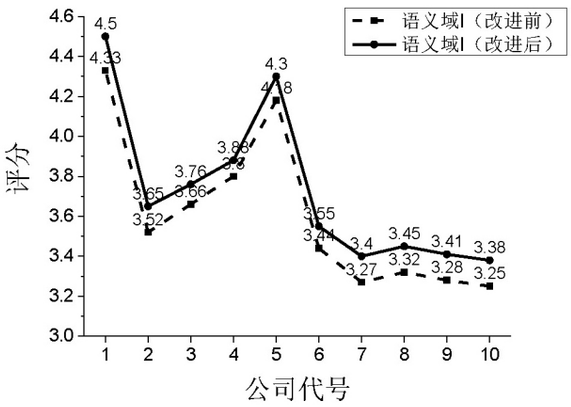

图3 10家公司年报语义域I的评分对比

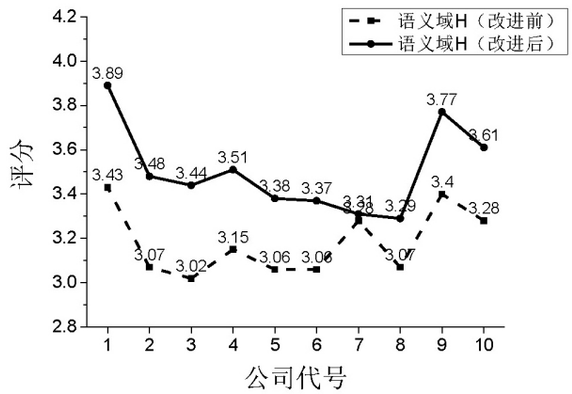

图4 10家公司年报语义域H的评分对比

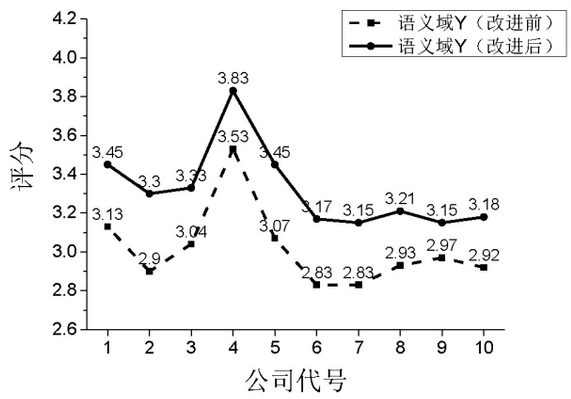

图5 10家公司年报语义域Y的评分对比

研究和实验分析结果表明，通过语义域的精心挑选和文本噪音的适当净化处理，中文语义分析系统有实现快速自动提取企业年报战略相关的术语和概念的潜力。这是在基于大规模数据的大数据背景下，实现及时向利益相关者和客户提供企业关键商业信息至关重要的第一步。

3 结论与展望

本文提出了使用中文语义分析系统，从企业年报信息披露文档中自动识别和提取中文企业年报战略内容相关术语这一技术。对实验结果的分析表明，利用语义分析器自动提取中文关键术语用于企业战略内容的进一步分析是完全可行的。

随着今后研究的深入和年报语料库规模的扩大，我们将运用词语搭配和其他统计方法，提取规模更大的年报战略内容相关术语，扩充现有语义分析系统中的单词和MWE词库，更大限度地提高对企业年报的分析能力。

致谢

感谢英国Lancaster大学Paul Rayson博士，Scott Piao博士和新加坡E-C Consulting公司Anthony Wong 高级顾问对本文研究给予的支持。

基金项目：

国家自然科学基金应急管理项目——基于统计与语义模型的企业年报战略内容挖掘方法研究（No. 61540029）。

参考文献

[1] El-Haj, Mahmoud El-Haj, Paul Rayson, Steven Young, Andrew Moore, Martin Walker, Thomas Schleicher, Vasiliki Athanasakou. Learning Tone and Attribution for Financial Text Mining [C]. Accepted by LREC 2016 Conference, Portorož, Slovenia.

[2] Rayson, et al. http://ucrel.lancs.ac.uk/cfie/annual-report-scores.php. [OL].

[3] Brennan, Niamh M. and Merkl-Davies, Doris M. Accounting Narratives and Impression Management [J]. In: Russell J. Craig, Jane Davison and Lisa Jack (eds.). The Routledge Companion to Accounting Communication. London: Routledge, 2013: 109-132.

[4] Merkl-Davies, Doris M., Niamh M. Brennan and Petros Vourvachis. Content Analysis and Discourse Analysis in Corporate Narrative Reporting Research: A Methodological Guide [C]. Centre for Impression Management in Accounting Communication (CIMAC) Conference, 2014, Bangor, UK.

[5] Merkl-Davies, Doris M., Brennan, Niamh M. and Vourvachis, Petros. A Taxonomy of Text and Discourse Analysis Approaches in Corporate Narrative Reporting Research [C]. International Conference on Discourse approaches to financial communication, 2014, Monte Verità Ascona, Switzerland.

[6] Athanasakou, Vasiliki and Khaled Hussainey. The perceived

credibility of forward-looking performance disclosures [M]. Accounting and Business Research, 2014, 3: 227-259. DOI: 10.1080/00014788.2013.867403.

[7] Athanasakou, Vasiliki, Mahmoud El-Haj, Paul Rayson, Martin Walker, Steven Young. Computer-based Analysis of the Strategic Content of UK Annual Report Narratives [C]. American Accounting Association Annual Meeting, 2014, Atlanta, USA.

[8] Koller, Veronika. Emotion and Rationality in Annual Reports: A semantic Domain Analysis (Slides) [C/OL]. The 8th LSE/LUMS/MBS Conference, ICAEW, 2014, London. http://ucrel.lancs.ac.uk/cfie/Conf2014/VeronikaKoller.pdf.

[9] Rayson, Paul, Dawn Archer, Scott Piao, Tony McEnery. The UCREL semantic analysis system [C]. Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 2004: 7-12.

[10] Rayson, Paul. Wmatrix: A web-based corpus processing environment, Computing Department, Lancaster University, 2009. http://ucrel.lancs.ac.uk/wmatrix/. [OL].

[11] International Conference on Discourse Approaches to Financial Communication 2014 (DAFC), Monte Verità Ascona, Switzerland. http://www.dafc.usi.ch/. [OL].

[12] The 7th LSE/LUMS/MBS Conference "What constitutes Financial Reporting Quality?" [C/OL]. London School of Economics, 2013, London, UK. http://ucrel.lancs.ac.uk/cfie/Conf2013/LSE-LUMS-MBSProgramme-2013.pdf.

[13] Hoberg, Gerard and Craig M. Lewis. Do Fraudulent Firms Strategically Manage Disclosure? [C/OL]. The 8th LSE/LUMS/MBS Conference, ICAEW, 2014, London. http://ucrel.lancs.ac.uk/cfie/Conf2014/CraigLewis.pdf.

[14] Young, Steven. Textual Analysis and Investment Decisions: An Overview (Slides) [C/OL]. Sell-Side Meeting: Citi Investment Research & Analytics, 2014 Citi Quantitative Finance Conference. Valencia, Spain. http://ucrel.lancs.ac.uk/cfie/CitiQuantsYoungJune2014.pdf.

[15] Young, Steven and Mahmoud El-Haj. Computer-based Analysis of UK Annual Report Narratives (Slides) [C/OL]. Financial Reporting & Business Communication Conference, 2014, Bristol, UK. http://www.lancaster.ac.uk/staff/elhaj/docs/ElhajFRBCslides.pdf.

[16] Rayson, Paul and Mahmoud El-Ha. Natural Language Processing of UK Annual Report Narratives (Slides) [C/OL]. The 8th LSE/LUMS/MBS Conference, ICAEW, 2014, London. http://ucrel.lancs.ac.uk/cfie/Conf2014/RaysonElHaj.pdf.

[17] Lang, Mark H. and Stice-Lawrence, Lorien. Textual Analysis and International Financial Reporting: Large Sample Evidence [C/OL]. Feb 14, 2015. SSRN: http://ssrn.com/abstract=2407572. DOI: 10.2139/ssrn.2407572.

[18] CFIE project [Z/OL]. http://ucrel.lancs.ac.uk/cfie/objectives.php.

[19] Xinying Qiu, Shengyi Jiang, and Kebin Deng. Automatic Assessment of Information Disclosure Quality in Chinese Annual Reports [C]. Natural Language Processing and Chinese Computing 2013, CCIS 400.

[20] Scott Piao, Xiaopeng Hu, Paul Rayson. Towards A Semantic Tagger for Analysing Contents of Chinese Corporate Reports [C]. ISCC 2015, Dec 18-19, 2015, Guangzhou, China.

[21] Piao, Scott and Bianchi, Francesca and Dayrell, Carmen and D'egidio, Angela and Rayson, Paul. Development of the multilingual semantic annotation system [C]. The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 1268-1274.

[22] Scott Piao, Paul Rayson, et al. Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages [C]. Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC. May 23-28, 2016, Portorož, Slovenia.

作者简介：

胡小鹏（1973-），男，博士，高级工程师。研究方向：自然语言处理、机器翻译。

袁琦（1939-），男，研究员。研究方向：自然语言处理、机器翻译。

管东升（1976-），男，硕士，高级工程师。研究方向：自然语言处理、文本挖掘。

英文摘要转至p12