情报学报 2017年7月 第36卷 第7期
Journal of the China Society for Scientific and Technical Information, July 2017, 36(7): 663-668

基于风险短语自动抽取的上市公司风险识别方法及可视化研究

胡小荣，姚长青，高影繁

（中国科学技术信息研究所，北京 100038）

摘 要 上市公司作为证券市场的基石，其财务状况与风险信息是众多投资者与研究人员的关注焦点，而上市公司年报中的风险信息披露字段因其权威性与公开性成为研究者评估上市公司风险的研究依据。目前针对风险信息披露字段内容的研究仅停留在基于分词与词频统计的风险分析层面，而单个的词并不能很好地揭示不同风险主题的具体表现和语义内容。本文采用基于多因素拟合的风险短语识别技术，对沪深两市环保行业76家上市公司年报中“风险因素”的文字描述字段进行处理，得到环保行业不同风险主题文本中的主题短语，最后使用jQCloud词云图对风险主题短语进行可视化展示。

关键词 上市公司风险评估；互信息；左右熵；多因素拟合；可视化

Risk Identification Method of Listed Companies Based on the Automatic Risk Phrase Extraction and Visualization

Hu Xiaorong, Yao Changqing and Gao Yingfan

(*Institute of Scientific and Technical Information of China, Beijing* 100038)

Abstract: The financial status and risk information of listed companies—the cornerstone of the securities market—is the focus of many investors and researchers, who usually conduct their researches based on the risk information invoked in annual reports of listed companies. The current methods are only based on word segmentation and frequency statistics, although a single word cannot capture the meaning of text and topics. This paper adapts the phrase extraction technology based on multi-factor fitting into the risk assessment of 76 listed companies in the environmental protection industry in Shenzhen and Shanghai stock markets. Finally, we use jQCloud to visualize the theme phrase.

Key words: risk assessment of listed companies; mutual information; information entropy; multi-factor fitting; visualization

1 引 言

企业的风险是“未来的不确定性对公司实现其目标的影响”，存在于企业生产经营活动的各个环节，在当前市场经济条件下，市场需求日趋多样，竞争程度愈加激烈，企业面临的风险日益加剧。而上市公司作为证券市场的基石，直面众多投资者，其经营状况会影响到投资者、证券市场乃至整个国

收稿日期：2017-05-02；**修回日期**：2017-06-20

基金项目：中央级公益性科研院所基本科研业务费专项资金项目“上市公司年报数据库建设及服务系统研发”（ZD2016-08），国家自然科学基金项目“科学基金项目产出专利对产业技术发展的影响研究”（L1624039），国家社会科学基金项目“面向科技型中小企业创新的技术竞争情报方法体系研究”（12CTQ030）。

作者简介：胡小荣，女，1993年生，硕士研究生，主要研究方向为文本挖掘，E-mail: huxr2016@istic.ac.cn；姚长青，男，1974年生，博士，副研究员，主要研究领域为情报理论与方法；高影繁，女，1974年生，博士，副研究员，主要研究方向为文本挖掘、知识组织。

民经济。因此，上市公司的风险信息成为众多投资者与研究人员的关注焦点。中国证监会发布的《公开发行证券的公司信息披露内容与格式准则 2 号》对上市公司年度报告中关于上市公司风险信息的披露做出相关规定，上市公司需针对自身的实际情况，充分、准确、具体地描述相关风险因素。

上市公司风险信息披露的数据公开可获取的特性使得众多研究者开始以此作为研究对象来评估上市公司风险。上市公司风险信息披露方式规定较为模糊[1]，大多数公司根据本公司面临的风险情况进行文字性描述。目前该类研究以基于分词与词频统计的文本分析方法为主，通过高频主题词分析文本的主旨[2]，而单个的词并不能很好地揭示不同种类风险的具体表现和语义内容。例如，“原材料价格”比“原材料”和“价格”两个单独的词表现出来的语义要更加丰富。因此，本文采用基于多因素拟合的风险短语抽取方法，以沪深两市环保行业 76 家上市公司年报为背景数据，通过对“风险因素”文字描述字段的分析，得到环保行业不同风险主题文本中的主题短语，以获取更丰富的语义表现,最后以词云图的方式进行可视化展示，使结果更加直观。

2 国内外研究现状

国内外关于上市公司风险的分析主要基于上市公司年报或招股说明书中的风险披露字段及年报中的财务数据。最常见的分析方法是通过对上市公司年报中的财务数据进行量化分析来实现对上市公司财务风险的评估。着眼于对上市公司的风险披露及对策字段进行分析是近年来的一个研究热点，该类该部分的研究又分为两个维度：

（1）对风险披露方式存在的问题提出思考和改进。蒋巍等[3]对浙江航民股份有限公司年报的风险披露字段进行实证研究，认为我国上市公司风险信息披露的现状与规范要求仍有一定差距。张曾莲[4]从理论和实证的双重角度对我国上市公司风险披露的现状和存在的问题进行评价，从而得到我国上市公司的风险信息披露质量，并为其规范和准则提供理论。

（2）对风险字段进行文本内容分析，Meijer[5]采用内容分析法对荷兰上市公司 2005-2008 年的年报中揭示的企业风险信息的类型和性质进行了纵向研究，以度量风险披露的数量。Appiagyei 等[6]审查了加纳证券交易所（GSE）采用国际财务报告准则之前和之后的风险披露，并采用内容分析法对 2004-2011 年上市公司的年报进行了检验。国内吴运建等[7]分析了我国上证 A 股上市公司年报中的风险信息披露字段，应用内容分析法，对公司年报表示风险的关键词进行统计，用来表示风险信息披露的水平。种莉萍[8]运用内容分析法对上市公司招股说明书中的风险字段进行比较分析和趋势分析，得出不同市场拟上市企业信息披露的差异性及近四年内上市公司风险信息披露的趋势，并在此基础上运用了非参数分析及描述性统计的方法进行了实证比较。赵一鸣等[13]将基于文本主题可视化的方法应用在计算机应用服务行业上市公司的风险分析上，根据 J. Donohue 提出的高频低频词界分公式，得到代表主题的核心高频词汇，然后进行主题可视化分析，揭示该行业市场风险的具体表现和语义内容。徐静婷[9]在对招股说明书的文本内容进行分析时，引入了文本挖掘技术，进行了基于词频的首要风险。

现有的基于风险披露文本内容分析方法的主要问题在于：单个的词不能很好地揭示主题，并且会丢失掉一部分语义内容，本文采取基于多因素拟合的短语自动抽取技术提取不同种类风险的主题短语，本文方法基于“短语的信息表征能力比单个词要强很多，它们在确定集合主题时比单个的词更重要”的假设。例如，在市场竞争风险主题中，“技术优势”比“技术”与“优势”两个词表达的信息更丰富；在替代产品风险主题中，“新产品风险”比“新产品”与“风险”两个词更能揭示主题的具体表现。

3 基于多因素拟合的风险短语识别技术及可视化方法

3.1 方法流程

从统计学的视角来看，一个短语的内部词之间的结合紧密程度依赖于词语的共现频次[10]。因此，在进行短语提取时，最简单的方法就是统计候选词串的数量，即候选词串的内部词语之间的共现频次，但是这种方法会产生大量的噪音，结果中会包含很多不符合语法和语义的词串。为了消除词串统计方法的缺陷，基于多因素拟合的短语识别方法可以将词语的互信息、左熵、右熵这三个统计量相结合，首先对词串的互信息、左熵、右熵这三个统计量进行计算，用以判别词串的内部结合紧密程度和外部边界独立性，然后对互信息与左右熵进行综合计算，得到 score 值，根据 score 值得到候选短语序列。最后进行基于词频的短语过滤，筛选出符合语法与语义的风险主题短语。本文采用的基于多因素拟合的

短语提取技术流程如图 1 所示。

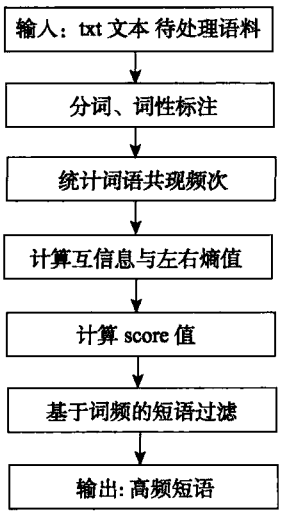

图 1　基于多因素拟合的短语提取流程

3.2　互信息的计算

为了简化文本中的统计计算，在该方法中对互信息与左右熵的计算方法进行了重新定义。二元互信息的计算公式（1）如下：

$$\mathrm{MI}(t)=\log\left(\frac{p(t)}{p(x)p(y)}\right)=\log\left(N\cdot\frac{n_t}{n_a\cdot n_b}\right) \quad (1)$$

其中，t 为候选词串，x,y 为候选词串 t 的内部词语。$p(t),p(x),p(y)$ 分别表示 t,x,y 的概率。采用简单的归一化频率形式来估计概率：$p(t)=n_t/N$，$p(x)=n_x/N$，$p(y)=n_y/N$．n_t,n_x,n_y 分别表示 t，x，y 在语料中出现的频次，N 是集合中所有长度满足阈值的候选词串的总数。互信息体现了两个词之间的相互依赖程度，互信息值越高，表明 x 和 y 相关性越高，则 x 和 y 组成短语的可能性越大；反之，互信息值越低，x 和 y 之间相关性越低，则 x 和 y 之间存在短语边界的可能性越大。

3.3　左右熵的计算

左熵的计算公式（2）如下：

$$E_L(W)=-\sum_{\forall a\in A}P(aW\mid W)\cdot\log_2P(aW\mid W) \quad (2)$$

右熵的计算公式（3）如下：

$$E_R(W)=-\sum_{\forall b\in B}P(Wb\mid W)\cdot\log_2P(Wb\mid W) \quad (3)$$

其中，E_L 与 E_R 分别表示词串的左熵和右熵，W 表示候选词串，$W=\{w_1,w_2,\cdots,w_n\}$；A 表示候选词串左边出现的所有词的集合，a 表示集合 A 中的某一个词；B 表示候选词串右边出现的所有词的集合，b 表示集合 B 中的某一个词；如果某个词串的 E_L 与 E_R 值越大，则该词串左右出现的词越多，搭配越丰富，那么该词串更有可能是短语。

3.4　基于多因素拟合的 score 值计算

score 值是对互信息与左右熵的综合计算。通过对互信息、左熵、右熵归一化之后求和得到，具体计算公式（4）如下：

$$\text{score=MI/total\_MI+LE/total\_LE+RE/total\_RE} \quad (4)$$

其中，MI、LE、RE 分别为某一候选短语的互信息值、左熵与右熵，total_MI、total_LE、total_RE 分别为候选短语序列的互信息值之和、左熵之和与右熵之和。本文采取由 score 值排序得到的抽取结果作为候选短语序列，并将其按照词频排序，通过词频降低噪声词的权重，以进行短语过滤。

3.5　可视化技术与词云图

在当前的大数据时代背景下，数据与信息量呈爆炸式增长，这就使得数据的处理变得更加复杂化，从大量数据中提取有效信息也变得更加困难。自然语言分析技术可以较好地从文本大数据中挖掘出重要信息[11]，但是挖掘出来的这些信息则需要一种更为直观、具象的组织表达形式，才能更加便于人们进行理解、浏览、传播与应用。可视化技术为此提供了一种可能的解决方式，它又被称为科学知识图谱[12]。因此，目前基于大数据可视化的研究成为热点，越来越多的可视化形式也随之被研究者所开发。词云是由词汇组成类似云的彩色图形，是近年来最受欢迎的信息可视化形式之一。它将文本中的词语按照一定顺序和规律进行排列，比如词频或者字典顺序，并用文字的大小来表示词语在文本中的重要程度。本文采用词云图的方式呈现抽取到的风险短语。

4　实验及结果分析

4.1　数据来源

中国科学技术信息研究所自建上市公司年报数据库，包含沪深两市上市以来全部上市公司的年报数据。本文从上市公司年报数据库中选取沪深两市环保行业 76 家上市公司年报“风险因素”的文字描述字段，对此进行分析处理。中国科学技术信息研究所上市公司年报数据库在进行年报加工时，已经

由人工抽取的方式完成了上市公司年报中风险因素的归类，风险类别主要包括：供应商风险、顾客风险、技术风险、经济风险、潜在竞争对手风险、市场竞争风险、替代产品风险、政治法律风险、其他风险等。本文实验中将年报数据库中每一类风险因素字段内容的构成初始的文本集合，分别以不同风险类别命名。

4.2 数据预处理

风险文本的预处理过程包括分词、去停用词、名词过滤、统计词频等，考虑到名词的文本表征能力和风险文本的特点，我们在对文本进行分词之后进行名词提取，在后续处理中将进行以名词为中心的风险短语抽取处理。本文在对风险文本进行分词时，利用 HanLP 汉语言处理包进行。HanLP 是由一系列模型与算法组成的 Java 工具包，不仅仅是分词，而是提供词法分析、句法分析、语义理解等完备的功能。HanLP 完全开源，包括词典，不依赖其他 jar，其官方模型训练自 2014 年人民日报语料库。通过工具类 HanLP 可以一句话调用所有功能，也可以进行二次开发，因此本文选取 HanLP 进行文本处理。本文首先在 Myeclipse 中搭建了 HanLP 环境，然后采用了能够灵活支持过滤器的 NotionalTokenizer 分词器进行分词。

4.3 短语抽取实验结果及分析

4.3.1 基于多因素拟合的风险短语抽取结果及分析

对风险数据进行预处理之后，按照上述算法进行短语抽取。以“供应商风险”短语抽取为例，按照互信息、左熵、右熵、score 值排序得到的短语抽取结果如表 1 所示。

在表 1 中可以看到，基于多因素拟合的短语自动提取技术会产生噪声词，如“公司非晶”、“风险化工产品”、“风险原材料”、“风险公司”等。不成词的成因有两方面：第一，在文本预处理中分词、去符号之后，没有对文本进行进一步处理，导致文本中不相邻的词在去掉符号之后变得相邻。以“风险公司”为例，该抽取结果频率较高的原因在于原文本中，“风险”与“公司”以“风险：公司”的形式出现，去掉冒号之后，两个词相邻，且出现频率高，因此将其判断为一个短语。第二，该方法自身不能很好地识别噪声词的缺陷，由于此原因生成的抽取结果虽然不成词，但是对研究仍有一定意义。以“风险能力”为例，该词并不是符合人们语言习惯的短语，但是在原文本中常以“提高公司的抵御风险能力”及“抗风险能力”等的形式出现，因此，虽然不成词，但是可以得出原风险文本中不仅披露了风险信息，也对提高企业抗风险能力做出了描述。

表 1 按照互信息、左熵、右熵、score 值排序得到的短语抽取结果 top10

MI（互信息）	LE（左熵）	RE（右熵）	Score 值
非晶合金	原材料价格	原材料价格	原材料价格
合金铁芯	苯乙烯	风险公司	非晶合金
产品价格风险	原材料电解铜	半导体多晶硅	苯乙烯苯胺
公司非晶	业务原材料	苯乙烯苯胺	苯乙烯
苯乙烯苯胺	非晶合金	公司原材料	价格风险
原材料产品价格	风险原材料	公司产品	原材料电解铜
期货市场手段	价格风险	非晶合金	中央空调业务
中央空调业务	公司非晶	苯乙烯	钢材铝材
风险化工产品	中央空调业务	业务原材料	风险公司
铁芯质量	原材料供货商	价格风险	电解铜价格

4.3.2 基于短语与基于分词的抽取结果对比

本节根据上节的短语抽取结果，按照词频进行过滤并选取 top10 的短语。以“供应商风险”与“政治法律风险”为例，将其结果与单纯基于分词与词频的方式得到的主题词进行对比，并统计了供应商风险中各主题短语的文档频率，以揭示其具体分布情况。结果如表 2 和表 3 所示。

在表 2 中，供应商风险出现频率最高的短语为“原材料价格”、“价格风险”、“非晶合金”、“原材料风险”等，结合原始风险分类文本，供应商风险主要包括原材料供应不足的风险、原材料涨价风险、

表 2 供应商风险主题短语与主题词对比

基于短语的抽取结果			基于分词的统计结果	
主题短语	词频	文档频率/总文档数	主题词	词频
原材料价格	16	9/16	公司	63
价格风险	13	8/16	原材料	55
非晶合金	10	2/16	价格	42
原材料风险	6	6/16	风险	33
苯乙烯	5	2/16	产品	13
苯乙烯苯胺	5	2/16	乙烯	12
公司原材料	5	7/16	多晶硅	11
风险公司	4	8/16	合金	10
电解铜价格	4	2/16	成本	10

表 3　政治法律风险主题短语与主题词对比

基于短语的抽取结果			基于分词的统计结果	
主题短语	词频	文档频率/总文档数	主题词	词频
政策风险	127	102/218	公司	643
宏观经济政策	60	61/218	政策	510
宏观经济	57	67/218	风险	423
公司业务	56	17/218	国家	410
国家政策	55	37/218	行业	264
风险公司	52	191/218	环保	152
行业政策	52	24/218	业务	151
产业政策	49	67/218	产业政策	122
政策性风险	43	39/218	宏观经济	99
风险能力	35	24/218	标准	94

原材料供应及产品价格变动风险等，该风险中主要涉及原材料的问题，而在环保行业中常用到的原材料包括非晶合金、苯乙烯、电解铜等。出现最高的词为“公司”、“原材料”、“价格”、“风险”等，其中，“公司”与“风险”为各类风险文本中的通用词，并不能反映风险主题，而“原材料价格”比“原材料”、“价格”两个单独的词所表达的语义更为丰富。在短语的文档频率统计中，“原材料价格”、“价格风险”、“原材料风险”、“公司原材料”等短语的文档频率较高，其文档分布范围较广，这揭示了供应商风险主要与原材料相关。

在表 3 中，政策法律风险出现频率最高的主题词为“政策风险”、“宏观经济”、“国家政策”、“宏观经济政策”、“公司业务”等，结合原始文本，“政策风险”、“宏观政策风险”、“产业政策风险”等出现频率较高的短语出现在公司年报风险文本的小标题处，而“国家政策”则主要出现在风险字段最后“避免和减少因国家政策变化对公司产生的不利影响”中，可以看出，政治法律风险的主题是该行业所面对的国家政策、宏观经济政策、产业政策发布或改变时可能带来的风险，而企业应当根据政策调整公司战略、进行技术升级、提高公司的应变能力。对比“政策”、“国家”、“风险”等词，其表达效果更好。在短语的文档频率统计中，“政策风险”、“宏观经济政策”、“产业政策”等短语的文档频率较高，其文档分布范围较广，这揭示了政治法律风险主要与国家政策相关。

4.4　风险短语云图

本文采用 jQCloud 对抽取结果进行可视化展示。jQCloud 是一款基于 jQuery 的标签云插件，它使用较少的 HTML 和 CSS 来构建多彩的标签云效果，综合应用字体大小、粗细、颜色三个视觉特征来表示关键词的重要程度的差异，其中，可以自定义字体的颜色和大小，也可以由插件随机生成，常见于 WordPress 的侧边栏标签云。jQCloud 能够兼容大多数主流的浏览器。本文使用 jQCloud 对抽取结果进行可视化展示时，采用由插件随机生成的字体大小与颜色。在此之前，需要首先对抽取出的短语-词频文件进行预处理，通过程序将其文本处理为如下格式：

{"text":"原材料价格","weight":16},{"text":"价格风险","weight":13},{"text":"非晶合金","weight":10}

最后进行可视化处理。以“供应商风险词云图”与“政治法律风险词云图”为例，结果如图 2 和图 3 所示。

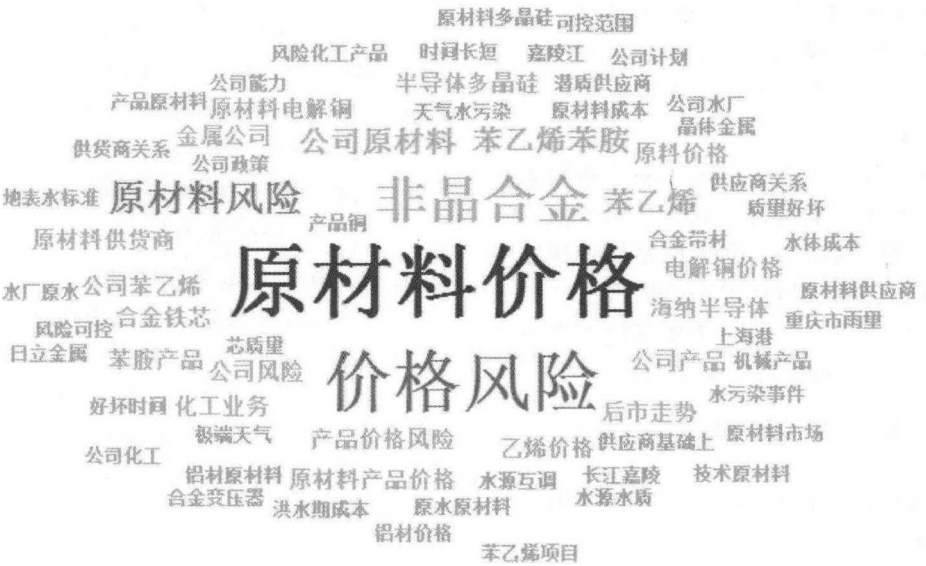

图 2　供应商风险主题可视化

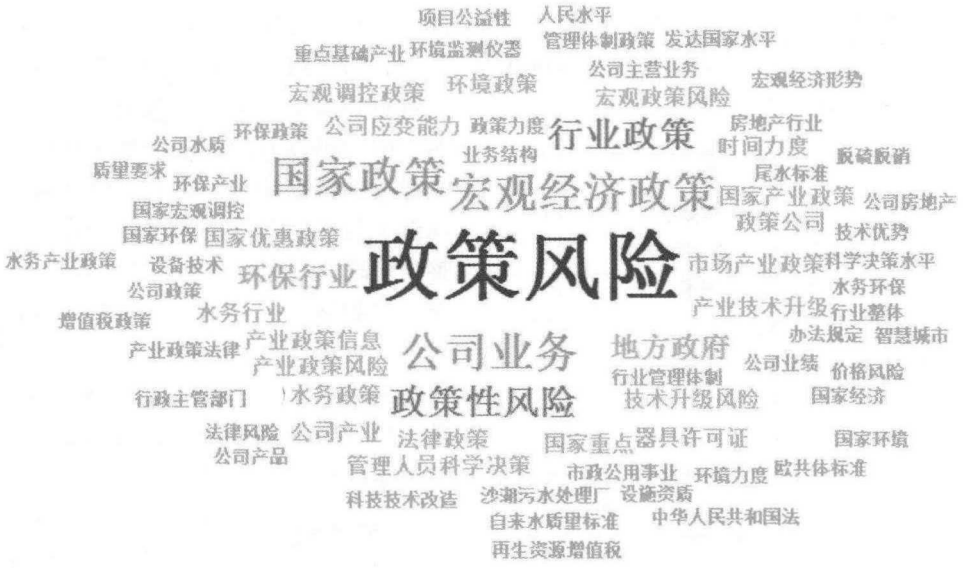

图 3 政治法律风险主题可视化

5 结 论

本文采用基于多因素拟合的短语提取技术对沪深两市环保行业76家上市公司年报中的风险字段进行分析，并进行主题可视化展示。文中所采用的上市公司风险评估方法的优势在于：首先，该方法完全通过编写代码，由计算机自动实现，因此可以快速、批量地处理数据，这在大数据时代是极为重要的优势。其次，中国科学技术信息研究所上市公司年报数据库数据加工工作已经完成，2016年之前（不含2016年）沪深两市主板全部上市企业的年报数据已完成结构化、碎片化处理，因此在数据资源方面极具优势。第三，相对于单个词而言，作为整体的短语携带的信息比其组成词语的总和更多，能够更为丰富地展示不同种类风险的具体表现和语义内容。第四，抽取结果可视化表示使得界面对于用户而言更加直观、友好，更方便研究者发现数据内部隐藏的信息及规律。

本文也存在一些待改进的地方：基于多因素拟合的短语提取技术还尚未完善，只能根据熵值、互信息值等提取出可能的短语，而不能很好地识别噪声词，因此，该方法需要进行进一步的改进。词云图的可视化方式较为单一，需要添加其他因素或者可视化方式如热力图等对处理结果进行展示。

参 考 文 献

[1] 王谊. 上市公司风险信息披露规范中美比较及启示[J]. 财会通讯, 2016(1): 121-123.

[2] 刘琦岩, 张泽玉, 张运良, 等. 从主题词词频变化看我国科技政策演变[J]. 情报工程, 2016, 2(4): 14-19.

[3] 蒋巍, 杨颖颖. 我国上市公司年报风险信息披露现状——以浙江航民为例[J]. 生产力研究, 2015(11): 1-3.

[4] 张曾莲. 上市公司年报风险信息披露研究[M]. 北京: 清华大学出版社, 2016: 33-60.

[5] Meijer M G H. Risk disclosures in annual reports of Dutch listed companies during the years 2005-2008[D]. Delden: University of Twente, 2011.

[6] Appiagyei K, Boateng C A, Onumah J M. Risk disclosures in the annual reports of firms in Ghana[J]. International Journal of Management Practice, 2016, 9(2): 142.

[7] 吴运建, 商行. 上证A股上市公司年报中风险信息的价值研究[J]. 财经问题研究, 2013(6): 66-73.

[8] 种莉萍. 基于内容分析法的 IPO 招股说明书中的风险信息披露研究[D]. 济南: 山东财经大学, 2012: 33-46.

[9] 徐静婷. 基于文本挖掘技术的创业板股票招股说明书风险分析[D]. 上海: 上海师范大学, 2015: 19-31.

[10] 刘荣, 王奕凯. 利用统计量和语言学规则提取多字词表达[J]. 太原理工大学学报, 2011, 42(2): 133-137.

[11] 唐家渝, 孙茂松. 新媒体中的词云: 内容简明表达的一种可视化形式[J]. 中国传媒科技, 2013(11): 18-19.

[12] 赵阳, 文庭孝, 姜坷炘. 我国智库研究文献可视化分析——以CNKI为例[J]. 情报工程, 2016, 2(4): 40-48.

[13] 赵一鸣, 张进. 文本主题可视化及其在上市公司风险分析中的应用[J]. 图书情报工作, 2014, 58(2): 102-108.

（责任编辑 宋 扬）