情报学报　2019 年 12 月　第 38 卷　第 12 期
Journal of the China Society for Scientific and Technical Information, Dec. 2019, 38(12): 1241-1249

DOI: 10.3772/j.issn.1000-0135.2019.12.001

基于三重维度的企业风险信息抽取方法研究

梁　娜[1]，姚长青[1]，王　峥[2]，高影繁[1]，李　岩[1]

（1. 中国科学技术信息研究所，北京　100038；2. 中国科学院文献情报中心，北京　100190）

摘　要　近年来，企业年报的篇幅越来越长，三大财务报表作为年报的主体，其内容几乎没有再增加，而财务报表之外的文字内容却愈加丰富，各种补充说明及解释成为了解公司生产经营现状的有益补充。其中，风险信息披露字段因其前瞻性和决策相关性逐渐成为学者们关注的焦点，如何从大量的风险信息中抽取出真正有价值的内容成为值得研究的问题。因此，本文以全部A股上市公司2016年半年报中披露的风险信息作为背景数据，提出三重维度的风险信息抽取方法，对风险描述文本中的风险信息进行抽取，使得抽取出的风险信息具有更丰富的信息含量，尽可能表征原始风险描述文本所要表达的信息。

关键词　风险信息；信息抽取；三重维度；年度报告；上市公司

Research on Enterprise Risk Information Extraction Method Based on Triple Dimensions

Liang Na[1], Yao Changqing[1], Wang Zheng[2], Gao Yingfan[1] and Li Yan[1]

(1. *Institute of Scientific and Technical Information of China, Beijing*　100038;
2. *National Science Library, Chinese Academy of Sciences, Beijing*　100190)

Abstract: In recent years, the length of the annual report of enterprises has been increasing. As the main body of the annual report, the content of the three financial statements has remained constant. However, the text content outside the financial statements has become increasingly descriptive. Various supplementary explanations have become a useful addition to understand the current situation of the company's production and operation. Among them, risk information disclosure has gradually become the focus of scholars due to its forward-looking and decision-making relevance. As such, the question of how to extract valuable content from the large amount of risk information has become a problem worth studying. Therefore, this paper considers the risk information disclosed in the 2016 semi-annual report of all A-share listed companies as the background data. It then proposes a three-dimensional risk information extraction method and extracts the risk information from the risk description text so that the extracted risk phrases are rich. The information content is extracted as much as possible to restore the information to be expressed in the original risk text.

Key words: risk information; information extraction; triple dimensions; annual report; listed company

1　引　言

上市公司披露的年度报告逐渐成为企业利益相关者了解企业生产经营现状的重要途径。近年来，年报的篇幅越来越长，但是作为年报主体的财务报表能披露的财务信息有限，而财务报表之外的文字

收稿日期：2019-05-22；**修回日期**：2019-09-25

基金项目：中国科学技术信息研究所重点工作项目“上市公司年报数据库建设及服务系统研发”（ZD2019-09）。

作者简介：梁娜，女，1995年生，硕士研究生，主要研究方向为科技大数据关键技术与应用服务研究，E-mail：liangna2017@istic.ac.cn；姚长青，男，1974年生，博士，研究员，硕士生导师，主要研究方向为情报理论与方法；王峥，女，1984年生，学士，馆员，主要研究方向为图书馆知识服务创新；高影繁，女，1974年生，博士，副研究员，硕士生导师，主要研究方向为文本挖掘、知识组织；李岩，男，1994年生，硕士，主要研究方向为数据挖掘算法研究。

内容却越来越丰富，各种补充说明及解释披露了越来越多的信息。年报中的文字信息具有一定的决策相关性，在帮助缓解信息不对称的同时又能够提高公司透明度、预测公司未来的发展情况。根据中国证监会发布的《公开发行证券的公司信息披露内容与格式准则第 2 号》相关规定，上市公司必须在年报中据实披露公司的风险信息。因此，相比于财务信息，年报中的风险信息被认为更具有前瞻性意义，是企业未来发展的先兆风险信息。如何从大量的风险信息中抽取出真正有价值的内容成为值得关注的问题。

上市公司由于执行相关规定、维护公司形象、维持与投资者之间的关系等原因，或被动或主动地定期向公众披露公司的资产、交易、年度报告等相关信息，为外界了解公司运转提供有效途径，保障了年报风险披露数据的开放和可获取。作为对以往基于财务数据的风险评估方法的重要补充，上市公司风险信息逐渐成为学者们关注和研究的焦点。

在此背景下，本文提出基于三重维度的企业风险信息抽取方法，以全部 A 股上市公司 2016 年半年报中的风险描述信息作为背景数据，采用多种数据挖掘手段和资源，对已有短语识别算法和分词工具进行改进，对风险描述中的风险短语进行抽取，使得抽取出的风险短语语义更加丰满，更具有价值。

2 国内外研究现状

目前国内外有关上市公司年报中披露的风险信息的研究主要集中于两个方面：一是从理论和实践层面论证风险信息的价值性，二是面向不同的应用场景尝试从大量风险信息中提取出有用的内容。

2.1 风险信息价值研究

风险信息是企业根据自身所处的政治、经济、社会、市场等外部环境，同时结合各类财务及经营管理等内部环境，对与企业生存发展有关的现有或潜在因素做出的预判和警示，具有前瞻性和决策相关性意义。学者 Norden 等[1]认为由于存在盈余管理等因素，导致财务信息本质上是回顾性信息，而编写于会计年度结束之后的风险信息则是稳定的展望性信息，更具有预测能力。学者 Athanasakou 等[2-3]指出，风险信息能够帮助缓解信息不对称，并提高公司生产经营的透明度，尤其在风险预警方面，风险信息的价值含量高于一般的自愿性披露信息。学者 Gulin 等[4]指出目前上市公司非财务信息披露尚不规范，披露内容和程度仍取决于管理者的意愿。但是由于投资者的决策需要，迫使公司尽可能多地披露非财务信息，以增强投资者对公司的信心。Bochkay 等[5]评估了 MD&A（管理层讨论与分析）部分在预测所有者权益回报方面的预测价值，通过实验证明将 MD&A 信息纳入预测模型能够显著提高预测准确性。学者林钟高等[6]采用内容分析法，从董事会报告中提取风险、危机、危害、困难、困境、市场风险等具有风险提示意义的关键词，量化各企业的风险强度，通过回归分析得出年报中的风险信息与银行信贷决策呈显著负相关的结论，认为企业披露的风险信息越多，其能够获得的银行贷款数额会相应降低。学者孟庆斌等[7]采用文本向量化的方法度量上市公司年报中的 MD&A 字段的信息含量，并将 MD&A 字段分为回顾部分和展望部分。其中，回顾部分指公司对财务数据的解释部分，包括对主营业务、资产负债、利润等的分析；展望部分包括公司的竞争态势、战略规划以及集中揭示的风险事项和相应的应对措施等信息。在此基础上进行信息含量与股价崩盘风险的回归实验，结果表明 MD&A 中的信息含量能够显著降低股价崩盘风险，而且展望部分的信息含量发挥着主要作用。虽然学者们通过实证证实风险信息具有一定的价值，对于企业利益相关者进行预测和决策都有助益，但是学者申心吉[8]强调风险信息的价值仍会受到信息披露的及时性和可靠性影响。

2.2 风险信息抽取方法研究

回顾已有文献，针对上市公司年报风险信息字段的研究与实践并不多见。Li[9]指出，要更好地理解上市公司披露的文本信息，就需要对大量文本进行简化，将文本中包含的信息表示成数字变量。Hanley 等[10]使用文本内容分析法对首次公开招股说明书进行信息含量测度。该方法将 MD&A 文本表示为向量，向量中的元素表示文本中每个词语出现的频率，进而将 MD&A 分解为标准信息和真正具有价值的信息两个部分。在此基础上，DeAngelis 等[11]将潜在语义分析方法引入 MD&A 字段的分析中，使用向量空间模型将 MD&A 的文本表示为 n 维空间中的向量，其中 n 表示文档中不重复的单词数量，向量中的元素表示为单词出现的频次，然后通过 IDF 对其进行加权，进而计算不同 MD&A 之间的余弦相似度，揭示 MD&A 中的主题及其与同行业其他公司之间的差异。Yang[12]提出基于扩展的主题

模型识别上市公司风险信息中的风险情绪以及各种风险类型。Bochkay 等[5]采用词袋模型识别每个 MD&A 文本中的单词并计算词频，为了降低财务相关词汇的权重，学者们使用金融情感词典作为词汇计数的主要来源，将单词分为正面、负面、不确定性、诉讼性、强模态和弱模态六大类，从而将文本转化为变量并将其纳入统计模型。然而词袋模型忽略了文本的语义信息，造成特征提取高维度稀缺等问题[13]。Campbell 等[14]依据词汇统计方法对上市公司风险文本披露的信息量进行测量，通过风险信息总字数、风险信息关键词总数以及各风险类别下的关键词数量表征信息含量。国内学者周双文[15]设计了领域本体，并基于该领域本体对创业板公司年报中披露的风险类型、风险量化信息以及风险对策信息进行识别和抽取，实现自动、高效的风险信息抽取。胡小荣等[16]采用多因素拟合的风险短语识别技术，基于互信息、左右熵等统计量对环保行业不同风险主题下的主题短语进行抽取。

总体而言，上市公司披露风险信息的前瞻性和决策相关性价值受到了诸多学者的肯定。但是在风险信息抽取方面，国内外学者主要依据风险词汇统计及文本分类统计算法对风险进行识别与度量[17]，存在以下问题：一是单个词汇造成大量语义的丢失，不能很好地揭示风险主题；二是胡小荣等[16]提出的主题短语抽取因算法和文本处理的缺陷，导致抽取结果覆盖度不足且容易引入噪声。故而本文在已有研究的基础上进行改进，提出三重维度的风险信息抽取方法，基于风险短语层面对风险因素进行抽取、揭示和预警。

3　方法研究

3.1　方法流程

对上市公司年报中的风险描述文本进行阅读可以发现，风险描述文本中的风险短语对于风险内容的揭示具有重要意义，如“人才短缺”、“毛利率下降”、“原材料价格上涨”、“经济下行压力”、“同比增长”等，这些风险短语明确揭示了风险的要素及成因，对于企业利益相关者了解、认识该风险的主要内容具有重要作用。因此，风险信息抽取的主要目标在于尽可能抽取出长度更长、信息量更大的风险短语，为后续风险信息文本挖掘及相关研究打好基础。

基于上述目标，本文提出三重维度的风险信息抽取方法，流程如图 1 所示。首先，由于年报风险信息披露的格式特点，每一段风险描述文本都以具体的风险类型作为小标题，总结本段风险信息所描述的具体内容。故而可以直接抽取该小标题作为风险信息的一维信息。其次，对传统的短语识别算法进行修正，通过修正后的风险短语识别方法对风险描述文进行风险短语识别，将该结果作为风险信息的二维信息。最后，在传统分词工具的基础上，优化分词词库，对原有工具的分词模型进行修正，在确保分词效率的前提下，优化分词效果，提高风险短语抽取的全面性，并将该结果作为风险信息的三维信息。经过三重维度的风险信息抽取结果，能够有效表征原始风险文本所要表达的信息。

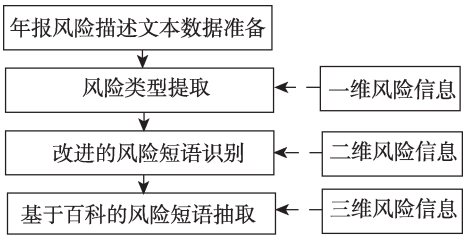

图 1　三重维度风险信息抽取方法流程

3.2　改进的风险短语识别方法

虽然学者们已经围绕风险信息展开诸多研究与探讨，但是现有的方法仍存在一定缺陷。一方面，短语的语义信息比单个词汇要更加丰富，使用单个词汇则意味着丢失更多的语义信息。孟庆斌等[7]采用《现代汉语词典》作为词语库构建全集向量，然而《现代汉语词典》中的词语长度短，无法揭示风险主题，例如，“毛利率”和“下降”两个分开的词汇，就不及“毛利率下降”所表达的含义那么明确；而林钟高等[6]提取出的风险关键词如“风险”、“危机”、“危险”、“危害”、“损害”等，信息量小且指代不明，不能明确表达风险内容。另一方面，胡小荣等[16]提出的风险主题短语抽取算法，在文本预处理时去除了标点符号，导致原本不相邻的词在去掉标点符号之后变得相邻，从而提取出“公司非晶”、“风险公司”、“风险原材料”等噪声词；且只提取名词短语的操作不符合风险信息抽取的实际需求，例如，“人才短缺”以名词+形容词来表示，“毛利率下降”是名词+动词形式，“原材料价格上涨”则以名词+名词+动词的形式存在，“PPP 模式”又以简称略语+名词的形式出现，其中，简称略语更是成为一种普遍存在的构词现象[18]。这些风险短

语的成词规则各异，难以通过穷举的方式列出，为避免风险信息的遗漏，不能限定词性及词性组合方式。故而本文提出改进的风险短语识别方法。

HanLP 提供的新词发现算法是目前广泛使用的短语识别算法之一。该算法主要基于互信息和左右熵进行短语识别。但是该算法在短语识别之前，过滤了包括标点符号在内的所有停用词，在词性标注后只保留名词过滤名词以外的其他词性，导致主题抽取结果噪声较大且覆盖面不全，无法直接应用于风险信息的抽取。故而本文在上述算法的基础上进行以下两点改进：在过滤停用词时，保留标点符号；在词性标注时，保留名词及名词以外的形容词、动词、副词等词性，保证重要词汇不被过滤。在上述预处理的基础上，采用互信息左右熵的计算规则识别风险短语。改进的风险短语识别方法具体流程如图 2 所示。

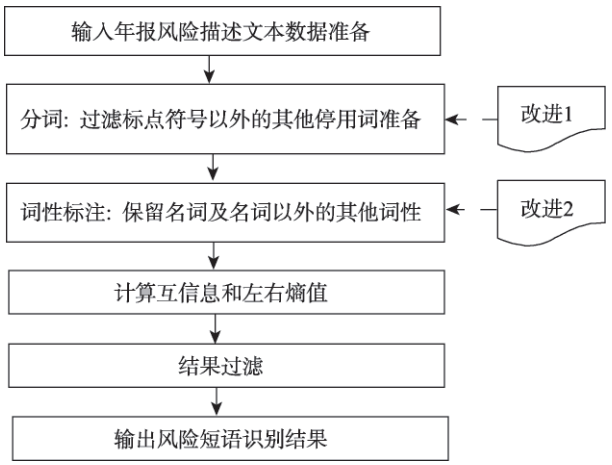

图 2 改进的风险短语识别方法流程

互信息的计算公式为

$$I(t)=\log\left(N\cdot\frac{n_t}{n_a n_b}\right) \tag{1}$$

其中，t 表示候选词串；N 表示集合中长度满足要求的所有候选词串的数量；n_t、n_a、n_b 分别表示词语 t、a、b 在文本中出现的频次。当互信息值越大，表明词语之间结合越紧密，成为短语的可能性越大；反之，互信息值越小，表明词语之间越不相关，越不可能构成短语。

左右熵计算公式为

$$E_{\mathrm{L}}(W)=-\sum_{\forall a\in A}P(aW|W)\cdot\log_2(P(aW|W)) \tag{2}$$

$$E_{\mathrm{R}}(W)=-\sum_{\forall b\in B}P(Wb|W)\cdot\log_2(P(Wb|W)) \tag{3}$$

其中，E_{L} 表示词串的左熵，E_{R} 表示右熵；W 表示候选词串集合；A 表示候选词串左边出现的所有词的集合，$a\in A$；B 表示候选词串右边出现的所有词的集合，$b\in B$。词串 E_{L} 和 E_{R} 的值越大，代表词串左右搭配的词语越丰富多样，该词串组成短语的概率更大。

3.3 基于百科的风险短语抽取方法

由于风险描述文本篇幅较短，经过上述风险短语识别之后，获得的风险短语数量较少，还不足以表征原始风险文本所要表达的信息。所以考虑基于其他资源的补充方案。复旦大学提供的中文通用百科知识图谱中涵盖百度百科、互动百科、中文维基百科等中文百科类网站的词条，包括具体事物、知名人物、抽象概念、文学著作、热点事件、专业术语、汉语字词或特定主题的组合等内容，从中获取共计 1200 多万个百科实体词，几乎覆盖全部领域，准确度高。因此，本文将在该百科实体词库的基础上进行风险短语抽取。

jieba 是目前最好用的 Python 中文分词组件，然而 jieba 自带词典所覆盖的风险短语数量有限，且在多数情况下将风险短语分割为多个词汇，故本文考虑在 jieba 的基础上，使用百科实体词库作为分词词典，替换 jieba 分词工具自带的原词典进行分词处理。考虑到 jieba 自带的分词词典规模约为 35 万，而百科实体词库规模在 1200 万左右，在规模上是原词典的 34 倍，直接使用百科实体词库作为分词词典并初始化 jieba，可能导致程序崩溃。由于 jieba 中采用了基于汉字成词能力的隐形马尔可夫模型（HMM），用于识别未登录词，本文将去除该模型以保证分词效率。具体流程如图 3 所示。

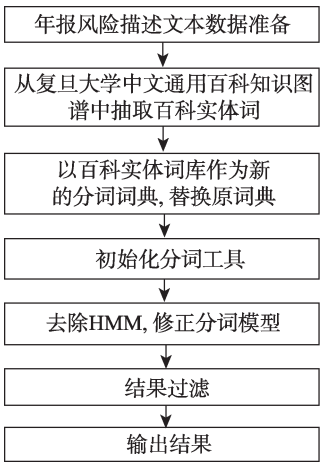

图 3 基于百科的风险短语抽取方法流程

首先，从复旦大学提供的中文通用百科知识图谱中抽取 1200 多万个百科实体词，组成实体词库，作为分词词典替换原词典，并初始化分词工具。其次，在分词时，添加 HMM=False 条件去除原分词工具中的隐马模型，修正分词模型，确保分词效率；最后进行结果过滤并输出风险短语抽取结果。

4　实验及结果分析

4.1　数据来源及预处理

由中国科学技术信息研究所自主建设的上市公司年报数据库，包含了沪深两市上市以来全部上市公司的年报数据。本文从中选取全部 A 股上市公司 2016 年半年报中的“风险因素”字段作为背景数据，共获得 14617 条风险描述文本，剔除几乎不包含风险信息的“避免同业竞争承诺”等文本，最终获取 6257 条风险描述文本作为最终的实验数据，对其进行风险信息抽取。

4.2　一维风险信息抽取实验

对风险描述文本进行预处理后，编写适当的正则表达式对风险信息数据进行风险类型提取。经过去重处理，最终获得 1390 种不同表述的风险类型，随机选取 20 个文本所对应的风险类型抽取结果，如表 1 所示。对风险类型提取结果进行简单分析可以发现，尽管针对同一风险，不同公司因其所处行业、经营范围以及年报撰写者语言表述习惯等的不同，年报中提及的风险类型的描述详细程度大不相同。以“管理风险”为例，就有着“集团化管理风险”、“经销商管理风险”、“经营管理风险”、“内部管理风险”、“企业管理风险”、“存货管理风险”、“并购管理风险”等多达 60 余种不同表述。为了保证信息抽取结果能最大限度地还原年报中的风险信息，本文不对上述风险类型做合并处理。每个风险信息文本对应一种风险类型，该风险类型构成一维风险信息。

表 1　风险类型提取部分结果

风险描述文本	风险类型	风险描述文本	风险类型
Text1	环境保护风险	Text11	人才风险
Text2	核心技术、业务人员流失风险	Text12	募集资金使用的风险
Text3	PPP项目执行的风险	Text13	信息科技风险
Text4	规模扩张风险	Text14	管理及人力资源方面的风险
Text5	应收账款金额较大、账龄增长的风险	Text15	公司规模扩大后面临的管理风险
Text6	核心管理人员变动及人才短缺风险	Text16	市场竞争的风险
Text7	发展中的管理风险	Text17	钢铁行业持续不景气状况的风险
Text8	经济下行的风险	Text18	燃气业务风险
Text9	投资并购整合风险	Text19	股权收购所带来的管理风险
Text10	核心技术的流失及技术泄密风险	Text20	新的商业模式可能带来的市场风险

4.3　二维风险信息抽取实验

基于上述改进的风险短语识别算法对全部 6257 条风险描述文本进行风险短语抽取，设定阈值挑选共现概率高的候选词串（即互信息值高的词串），通过互信息值挑选出共现概率高的词，再选取左右熵值之和最高的前 20 个词，降序输出，过滤数字等无意义短语以及长度小于 4 的噪声词，最终获得 7551 个风险短语。对于同样数据，进行基于 HanLP 的短语识别，最终获得 58488 个短语。以第 4.2 节中的前 10 个文本为例，两种算法的抽取结果如表 2 所示。

对比短语识别结果可以发现，虽然多数情况下，改进后的风险短语识别方法识别出的结果数量明显少于基于 HanLP 的短语识别结果，但是从质量上来看，改进后的风险短语识别方法识别出的噪声短语更少，短语所表达的语义更加明确，更能够解释风险类型的具体内容。而以 Text2 的识别结果为例，改进后的风险短语识别方法能够识别出“人员流失”、“技术泄密”等名词+动词的组成形式；从 Text3 的识别结果来看，改进后的风险短语识别方法能够识别出英文简称缩略语+名词的短语组成形式，而在风险描述文本中出现的英文简称缩略语一般具有非常重要的作用，往往表示着明确的风险因素。因此，可以认为，改进后的风险短语识别算法抽取出的结果具有更大的价值。

为了对实验结果进行评估，本文分别从基于 HanLP 的短语识别结果和改进的风险短语识别结果

表 2 两种算法的实验结果对比

风险描述文本	风险类型	基于HanLP的短语识别结果（top 20）	改进后的短语识别结果
Text1	环境保护风险	系统环保；公司垃圾焚烧；垃圾焚烧过程；垃圾焚烧项目；操作失误原因；故障操作失误；环境管理制度；风险垃圾焚烧；垃圾渗滤液；暂时性故障；废气污水；渗滤液系统；设备暂时性；污染物国家标准；污水污染物；烟气系统；厂区设施；过程废气；项目厂区；事故预案	环境保护；环保设施；污染物排放；生产过程；污染防治
Text2	核心技术、业务人员流失风险	业务人员风险；产品程度；人员公司；人员状况；人才制度；人才骨干；优势重大贡献；公司创新能力；公司技术；公司核心技术；公司能力；公司高端；创新能力技术；技术优势；新技术产品；有可能公司；核心技术业务人员；核心技术人员；程度专业人才；风险公司	人员流失；技术泄密；核心技术；流失风险
Text3	PPP项目执行的风险	业主手续；业务公共服务；公共服务领域；公司合同；公司市政；可控项目；市政污泥；我国模式；招投标项目；模式公共服务；污泥业务；经济效益风险；进度经济效益；部分项目；项目工期；项目进度；项目障碍；项目风险；风险公司；风险可控	PPP模式；PPP项目；项目实施；建设工期
Text4	规模扩张风险	体系体系；风险风险；战略管理水平；模式管理制度；管理制度制度；管理水平规模；环境海绵；海绵城市；财务风险；风险财务；资产规模；智慧环境；市场竞争力；风体系；规模风险；规模规模；规模业务；业务领域；公司智慧；制度风	管理水平；规模扩张；资产规模；规模扩大；管理风险；业务领域；公司资产；如果公司
Text5	应收账款金额较大、账龄增长的风险	业务单位；业务人员收账；业务量单位；余额问题；催收责任；公司业务人员；公司业绩；公司收账；公司清欠；单位收账；客户款项；情况业绩考核；收账余额；收账催收；收账情况；收账金额；时间业务量；欠款客户；款项清收；清欠小组	应收账款；相应增加；账龄增长；部分账龄；坏账准备
Text6	核心管理人员变动及人才短缺风险	业务人才；于本报告期；人才人才；人才公司；人才市场；人才技术；人才需求；人才风险；公司业务；公司于本；市场人才；技术人才；报告期董事；换届选举管理人员；核心管理人员；监事换届选举；管理人员人才；管理层人员；董事监事；风险公司	管理人才；管理人员；人员变动；人才短缺；公司未来
Text7	发展中的管理风险	业务人力资源；业务规模；业务范围遍布全国；人力资源风险；人员业务；公司业务；公司产业；公司损失；公司管理层；公司管理水平；公司跨度；公司过程；公司风险；内部管理机制；员工潜能；复杂多变外部环境；方面要求；最大限度缺陷；组织结构；组织协调能力	快速发展；管理失衡；管理能力；管理难度；提高管理；管理水平；业务规模；公司管理层；运营管理；管理风险
Text8	经济下行的风险	工业民间；经济压力	下行压力；进出口贸易；民间投资；经济下行
Text9	投资并购整合风险	上海业务范围；上海凌霞固；业务领域；产业规模；充分利用平台；公司上海；公司内生；公司协议；内生外延；凌霞固事宜；协议合作方；合作方项目；外延战略；平台公司；战略模式；控股子公司管理制度；方面风险；管理制度凌霞固；规模业务；部分股权	投资项目；整合风险；业务领域；凌霞固废；并购整合；控股子公司
Text10	核心技术的流失及技术泄密风险	膜组件；核心技术人员；产品性能；专利非专利；严格执行技术；产品膜法；公司严格执行；国家大力发展；大力发展重点；快速增长重大贡献；技术保密制度；收入快速增长；核心技术技术；程度核心技术；膜法水资源；非专利技术；忠诚度凝聚力；膜膜；团队忠诚度；组件产品	核心技术；技术泄密；研发人员；一定程度；产品性能；人员签订；保持核心

中随机抽取1000个风险短语，分别由3位情报学领域的研究生判断是否构成短语以及具体的构词规则。针对同一个短语，若有两位或两位以上的研究生判断结果相同，则取该结果作为最终评判结果。最终得到的结果如表3所示。可以看出，基于HanLP的短语识别算法准确率约为70.5%，而本文提出的改进后的风险短语识别方法准确率约为80.6%，且后者识别出的风险短语构词规则更多样化。相比之下，改进后的风险短语识别方法能够识别出更多准确短语，且构成短语的词汇词性更多样化。

4.4 三维风险信息抽取实验

虽然改进后的风险短语识别算法能够更好地将原始风险文本中的风险因素抽取出来，但是抽取结果仍不足以表征原始风险文本所要表达的全部信息，尤其是原始风险文本中提及的风险应对措施相关内容。故而本文在上述实验的基础上，进行基于百度百科词汇库的资源补充，以抽取出更多风险信

表3　人工评判结果比较

	基于HanLP的短语识别结果	改进后的短语识别结果
随机抽取短语数	1000	1000
准确短语数	705	806
不准确短语数	295	194
准确率	70.5%	80.6%
短语的构词规则	名词(词组)+名词(词组)	名词+名词,名词+动词,动词+名词,简称缩略语+名词,动词+动词,副词+动词等

息对上述风险短语识别结果进行补充。

在第3.3节的基础上，以包含1200多万百科实体的词库作为分词词典，对6257条风险描述文本进行分词处理，对分词结果过滤停用词、数字等不规范短语以及长度小于4的常用词语，最终获得46370个百科实体词。同样以上述10个Text为例，风险短语抽取对应结果如表4所示。对Text1进行分析，结合其原始风险描述文本，如图4所示，在Text1的风险信息抽取结果中，短语“环境保护风险”、“环境管理”、“环境保护”是对风险类型的补充，短语“垃圾焚烧发电”、“垃圾焚烧”、“操作失误”、“生产过程”等描述了公司在项目运营过程中可能导致风险的因素，短语“应急预案”、“污染防治技术”、“环保设施”、“国家标准”等表达公司对上述风险因素的应对措施，这些短语都在一定程度上补充了公司的风险信息。

表4　基于百科的风险短语抽取结果

风险描述文本	风险类型	改进后的短语识别结果	基于百科的风险短语抽取结果
Text1	环境保护风险	环境保护；环保设施；污染物排放；生产过程；污染防治	环境管理；应急预案；污染防治技术；管理制度；垃圾渗滤液；环境保护；垃圾焚烧发电；环境保护风险；环保设施；垃圾渗滤液处理；污染防治；垃圾焚烧；操作失误；生产过程；国家标准
Text2	核心技术、业务人员流失风险	人员流失；技术泄密；核心技术；流失风险	核心技术；技术优势；人员流失风险；持续创新；创新能力
Text3	PPP项目执行的风险	PPP模式；PPP项目；项目实施；建设工期	项目实施；项目执行；公共服务；经济效益；项目建设工期；基础设施；项目建设；市政污水处理；污水处理；建设工期；污泥处置
Text4	规模扩张风险	管理水平；规模扩张；资产规模；规模扩大；管理风险；业务领域；公司资产；如果公司	资产规模；管理制度；智慧环境；管理风险；激励体系；发展战略；公司发展战略；公司发展战略和管理；市场竞争；综合能力；财务风险；内控制度
Text5	应收账款金额较大、账龄增长的风险	应收账款；相应增加；账龄增长；部分账龄；坏账准备	坏账准备；应收账款；应收账款催收；法律手段；风险评估；账款回收；业绩考核；经营业绩
Text6	核心管理人员变动及人才短缺风险	管理人才；管理人员；人员变动；人才短缺；公司未来	管理人员；技术研发；经营管理；管理团队；市场开拓；换届选举；高级管理人员
Text7	发展中的管理风险	快速发展；管理失衡；管理能力；管理难度；提高管理；管理水平；业务规模；公司管理层；运营管理；管理风险	资产规模；管理能力；人力资源；股权投资；管理组织；综合管理；管理风险；投资项目；组织协调能力；资金运营；管理机制；组织结构；业务范围；协调能力；经营决策；管理跨度；公司管理；激励员工；公司管理层
Text8	经济下行的风险	下行压力；进出口贸易；民间投资；经济下行	下行压力；出口贸易；经济发展；经济下行；民间投资；经济下行压力；工业投资
Text9	投资并购整合风险	投资项目；整合风险；业务领域；凌霞固废；并购整合；控股子公司	管理制度；并购整合风险；投资项目；拓展公司；业务范围；控股子公司；产业规模；外延式发展；公司管理；公司管理制度；并购整合；发展模式
Text10	核心技术的流失及技术泄密风险	核心技术；技术泄密；研发人员；一定程度；产品性能；人员签订；保持核心	高新技术领域；核心技术；保密制度；竞业限制；技术进步；整体解决方案；技术垄断；产品性能；专利技术；非专利技术；解决方案；高新技术；保密协议；市场竞争；技术保密；发达国家

环境保护风险：垃圾焚烧发电项目在运营过程中会产生废气、污水和固废等污染物，在具体执行过程中，可能由于设备暂时性故障或人为操作失误等原因导致环境保护风险，从而对公司的项目运营造成不利影响，对此公司将采用一系列污染防治技术和措施，加强了环境管理制度的建立和落实；在各项目厂区配备了应急设施和设备，并制定了环境事故应急预案；同时建立了一套由烟气处理系统、垃圾渗滤液处理系统等多个环保设施系统构成的环保执行体系，确保公司垃圾焚烧发电生产过程符合环保要求，废气、污水和固废等污染物排放达到国家标准。

图4 Text1的原始风险描述文本

经过上述实验，针对每个风险描述文本都能够获得三重维度的风险信息抽取结果。以Text1为例，其风险信息抽取结果如表5所示。

表5 三重维度风险信息抽取结果——以Text1为例

维度	风险信息抽取结果
一维风险信息	环境保护风险
二维风险信息	环境保护；环保设施；污染物排放；生产过程；污染防治
三维风险信息	环境管理；应急预案；污染防治技术；管理制度；垃圾渗滤液；环境保护；垃圾焚烧发电；环境保护风险；环保设施；垃圾渗滤液处理；污染防治；垃圾焚烧；操作失误；生产过程；国家标准

4.5 实验结果分析

一维风险信息主要利用年报中原始风险描述文本的特点直接抽取而获得，不存在噪声问题，准确性很高。只是需要注意的是，有极少部分原始风险描述文本在叙述时没有指出具体的风险类型，造成极少数风险描述文本一维风险信息的缺失。可以考虑通过文本相似性计算等方式，使用相似文本的风险类型作为该风险描述文本的一维信息。

根据二维风险信息抽取实验结果可以得出，本文提出的改进后的风险短语识别方法识别出的准确短语数量占比在原短语识别方法的基础上提升了10%，识别出的短语噪声更少，且短语语义更为丰富，所表达的信息更明确。从短语的构词规则来看，改进后的风险短语识别方法除了能够识别名词+名词（例如：资金实力、资产负债率、原材料价格、新产品市场等）所组成的短语以外，还能够识别出名词+动词（例如：业务减少、业绩下降、人才流失、季节性亏损等）、动词+名词（例如：审批风险、收购资产、投资回报率、销售规模、增长速度等）、简称缩略语+名词/动词（例如：PPP模式、LED行业、ERP系统、同比下降、同比衰退等）、动词+动词（例如：并购重组、整改监督、质押担保、追加投资等）以及副词+动词（例如：高速增长、大幅下降、快速增长等）等多种词性构成的短语，抽取出的风险信息更全面。

对于三维风险信息抽取实验，在同一数据样本的基础上，使用原jieba分词工具耗时21秒，而优化词典后的jieba分词工具耗时357秒，考虑到优化后的分词词典规模是原词典的34倍，可以认为优化后分词效率并没有降低。该实验能够在确保分词效率的基础上，抽取出更多的风险信息以弥补风险短语识别实验的不足。

总体而言，经过上述三重维度的风险信息抽取，能够获得质量高、信息量大而且更为全面的风险信息。

5 结 论

本文提出三重维度的企业风险信息抽取方法，并对全部A股上市公司2016年半年报中的风险描述信息进行抽取，验证了本文方法的可行性和有效性。文中所采用的企业风险信息抽取方法，其优势在于：采用多种数据挖掘手段和资源，对已有的算法和分词工具进行了改进，对风险描述文本进行不同维度的信息抽取，提高了风险短语抽取的准确性和全面性，使得抽取的风险短语语义更丰富、更具有应用价值。在已有的风险信息相关研究中，大多数还是基于词汇层面，本研究在短语层面进行的尝试，有望为其他学者提供新的思路和方法。

此外，本文尚有不足之处。风险描述文本篇幅短小，加大了风险短语提取的难度，本文提出的改进后的风险短语识别算法依然会产生少量噪声词。同时，由于缺乏特定的领域词典，本文使用百科实体词库作为资源补充，虽然能帮助抽取更多的信息，但是效果仍有望提升，而且在结果过滤时，没有对短语的同频子集进行过滤，可能导致风险信息的重复抽取。

参 考 文 献

[1] Norden L, Weber M. Credit line usage, checking account activity, and default risk of bank borrowers[J]. Review of Financial Studies, 2010, 23(10): 3665-3699.

[2] Athanasakou V, Hussainey K. Forward-looking performance disclosure and earnings quality[R]. London School of Economics

and University of Stirling Working Paper, 2010.

[3] Athanasakou V, Hussainey K. The perceived credibility of forward-looking performance disclosures[J]. Accounting and Business Research, 2014, 44(3): 227-259.

[4] Gulin D, Hladika M, Mićin M. Disclosure of non-financial information: The case of croatian listed companies[C]// Proceedings of the Conference on Consumer Behavior, Organizational Strategy and Financial Economics. Cham: Springer, 2018, 9: 159-175.

[5] Bochkay K, Levine C B. Using MD&A to improve earnings forecasts[J]. Journal of Accounting, Auditing & Finance, 2019, 34(3): 458-482.

[6] 林钟高, 杨雨馨. 风险提示信息与银行信贷决策——基于A股上市公司年报文本信息的研究[J]. 安徽师范大学学报(人文社会科学版), 2017, 45(2): 245-255.

[7] 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.

[8] 申心吉. 中国上市公司信息披露质量状况研究——基于深交所信息披露考评的经验证据[J]. 时代金融, 2017(12): 152-154.

[9] Li F. Textual analysis of corporate disclosures: A survey of the literature[J]. Journal of Accounting Literature, 2011, 29: 43-65.

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

[11] DeAngelis M D. Uncommon information in firm disclosures[D]. East Lansing: Michigan State University, 2014.

[12] Yang B. Extending topic models for text analysis of corporate risk disclosures[D]. Singapore: National University of Singapore, 2013.

[13] 翟文洁, 闫琰, 张博文, 等. 基于混合深度信念网络的多类文本表示与分类方法[J]. 情报工程, 2016, 2(5): 30-40.

[14] Campbell J L, Chen H, Dhaliwal D S, et al. The information content of mandatory risk factor disclosures in corporate filings[J]. Review of Accounting Studies, 2014, 19(1): 396-455.

[15] 周双文. 基于领域本体的创业板公司年报风险信息抽取方法研究[D]. 长沙: 湖南大学, 2013.

[16] 胡小荣, 姚长青, 高影繁. 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[J]. 情报学报, 2017, 36(7): 663-668.

[17] 肖浩, 詹雷, 王征. 国外会计文本信息实证研究述评与展望[J]. 外国经济与管理, 2016, 38(9): 93-112.

[18] 张秋子, 陆伟, 程齐凯, 等. 基于最大熵模型的学术缩写自动识别[J]. 情报工程, 2015, 1(2): 64-72.

（责任编辑　王克平）