● 唐晓波，谭明亮，李诗轩，顾 娜（武汉大学信息管理学院，湖北 武汉 430072）

# 基于风险短语挖掘的知识聚合模型研究*

摘 要：［目的/意义］大数据时代，金融行业面临着海量多源异构信息源所带来的巨大挑战；针对大数据环境下存在的多源异构金融数据，通过对企业风险知识单元进行挖掘和聚合，从而使其有序化地收敛于高效的金融知识服务，这对于投资决策、风险管理和金融监管等金融决策支持过程具有十分重要的意义。［方法/过程］文章从知识挖掘、知识组织和知识服务的相关理论、方法和技术出发，构建了基于风险短语挖掘的知识聚合模型，该模型主要由知识采集模块、知识挖掘模块和知识服务模块等三大模块所组成。［结果/结论］文章利用N-gram算法来挖掘上市公司年报文本中的候选风险短语，并利用基于统计和基于规则的方法来实现候选短语的过滤，形成可复用的风险短语知识库；将短语作为知识聚合的粒度，利用聚类分析、共现分析和知识检索等技术进行了多种形式的知识聚合，从而为决策者提供智能化的金融知识服务。

关键词：短语挖掘；知识聚合；知识挖掘；知识服务

DOI：10.16353/j.cnki.1000-7490.2020.08.023

引用格式：唐晓波，谭明亮，李诗轩，顾娜．基于风险短语挖掘的知识聚合模型研究［J］．情报理论与实践，2020，43（8）：152-158，139.

## Research on Knowledge Aggregation Model Based on Risk Phrase Mining

**Abstract:** [Purpose/significance] In the era of big data, the financial industry faces enormous challenges from voluminous, dynamic and multi-source heterogeneous information sources. Aiming at the massive multi-source heterogeneous financial data under the big data environment, it is very important to mine and aggregate the enterprise risk knowledge units, so that it can converge to efficient financial knowledge services in an orderly manner, which is of great significance for financial decision-making support processes such as investment decision-making, risk management and financial supervision. [Method/process] From the theories, methods and technologies of knowledge mining, knowledge organization and knowledge service, this paper constructs a knowledge aggregation model based on risk phrase mining, which is mainly composed of three modules: knowledge collection module, knowledge mining module and knowledge service module. [Result/conclusion] In this paper, the N-Gram algorithm is utilized to mine candidate risk phrases from annual reports of listed enterprises, and the methods based on combination of statistics and rules are used to filter candidate phrases to form reusable risk phrase knowledge base. This paper takes phrase as the granularity of knowledge aggregation and makes use of cluster analysis, co-word analysis and knowledge retrieval to carry out various forms of knowledge aggregation, so as to provide intelligent financial knowledge services for decision-makers.

**Keywords:** phrase mining; knowledge aggregation; knowledge mining; knowledge service

金融业面临着大数据环境下的海量多源异构信息源所带来的巨大挑战。与此同时，作为典型的知识密集型行业，金融业对数据、信息和知识有着巨大的需求，投资决策、风险管理和金融监管等金融决策都需要大量的数据、信息和知识作为支撑[1]。大数据时代，从海量多源异构的金融数据资源中获取对金融决策支持有价值的知识，并对其进行有效的组织，从而为金融决策支持提供智能化的知识服务，这对于金融决策支持过程具有十分重要的意义。

上市公司年报文本中披露了大量有关企业的内外部风险、经营成果和未来发展等多方面的重要信息，从中挖掘出有价值的知识能够有效地辅助和支持决策者的金融决策，从而有效地提高金融决策的效率和质量。针对年报文本中描述企业内外部风险因素相关的文本信息，本文将基于统计和基于规则的短语挖掘方法相结合来抽取企业风险短语知识单元，并形成可复用的企业风险短语知识库；针对以往知识聚合研究中将词语、句子和文本作为主要知识聚合粒度所存在的不足，本文将短语作为知识聚合的粒

* 本文为国家自然科学基金项目“基于文本和Web语义分析的智能咨询服务研究”的成果之一，项目编号：71673209。

度，利用聚类分析、共现分析和知识检索等技术进行了多种形式的知识聚合，并开发了企业风险知识检索原型系统，从而更好地为企业风险识别、投资风险预警和金融监管等金融决策支持过程提供智能化的知识服务。

## 1 相关研究

1) 年报挖掘。作为会计和金融领域使用最为广泛的数据源之一，上市公司年报披露了有关企业的业务描述、经营情况、公司治理、内外部风险和未来发展等多方面的定量和定性信息，具有十分重要的价值。国内外的研究者们将年报中包含的财务指标等定量数据广泛地应用于年报欺诈检测、企业破产预测、企业财务危机预警、股票预测和企业信用风险评估等任务中。尽管年报中包含的财务指标等定量数据相对于定性文本数据更加易于处理和分析，但是这些定量数据有着其固有的内在缺陷，例如只能够反映企业过去的表现，不能够反映企业其他多个方面的重要信息等[2-3]。近年来越来越多的研究表明，定性文本信息能够有助于提高对定量会计信息的理解[4]；与此同时，上市公司年报等文本信息源中还包含了大量有关企业的潜在风险、财务状况、经营成果和发展战略等多个维度的补充信息[5-6]。国内外的研究者们开始运用分类[7]、聚类[8]、主题发现[9]、信息抽取[10]等技术来实现年报中定性文本的挖掘。

2) 短语挖掘。短语是文本中连续出现的词汇序列，是在文本中特定的上下文语境下所形成的语义单元[11]。短语是比词汇更高阶的语义单元，是词与上下文结合的表达，具有一定的语法句法规则；与词汇相比，短语在语法和语义结构上比词汇要更为完整，包含了更加丰富和清晰的语义语境信息，且具有更强的可读性[12-14]。例如，描述医药制造企业政策风险的短语“仿制药一致性评价”比“仿制药”“一致性”和“评价”三个单独的词汇所揭示的语义要更加丰富和具体。短语挖掘已成为自然语言处理和数据挖掘等领域所广泛关注的重要课题，在机器翻译、语义检索、本体构建、文本分类和自动摘要等多个领域有着非常重要的应用[15-16]。当前短语挖掘的方法主要包括三类：基于规则的方法、基于统计的方法和将前两者方法相结合的混合方法[16]。基于规则的方法准确率较高，但是需要人工确定规则模板，且可移植性较差[17]。基于统计的方法中的统计机器学习模型的训练需要大规模高质量的语料[16]。混合方法则将两种方法进行有效整合来实现短语的抽取，是当前短语挖掘的主流方法[18]。

3) 知识聚合。针对高度分散且无序分布的知识资源为知识组织和知识服务带来的机遇与挑战，学者们提出了知识聚合的理念和方法。作为近年来图书情报领域新兴的研究热点之一，知识聚合旨在对知识碎片进行动态关联和筛选组织，以实现知识单元的有机连接和知识资源的多维组合，从而为用户提供个性化和智能化的知识服务[19-21]。研究者们基于本体、关联数据、计量分析等方法针对以馆藏资源为代表的学术资源的知识聚合开展了大量的研究[21-24]。部分研究者也尝试将知识聚合的理论、方法和技术应用于医疗[25]、旅游[26]、问答社区[27]等多个领域中。当前研究主要从词语、句子与文本三种粒度进行知识聚合[28]，其中词语和文本是常见的粒度。但是，文本粒度的知识单元粒度粗且信息冗余度高；词语粒度的知识单元表征资源内涵能力有限，对主题内容描述完整性较低，容易导致语义的缺失；句子相较词语语义表征能力更强，相较文本信息冗余更小，但受技术限制，目前句子粒度的资源聚合研究较少[28]。尽管研究者们针对短语挖掘进行了大量的研究，但是以短语作为知识聚合粒度的研究还较为缺乏；本文利用短语挖掘技术来从年报文本中抽取风险短语，并基于短语粒度开展金融领域的知识聚合。

## 2 基于风险短语挖掘的知识聚合模型

为了更好地支持决策者的企业风险识别、投资风险预警和金融监管等金融决策，从而有效地提高金融决策的效率和质量；本文从知识挖掘、知识组织和知识服务的相关理论、方法和技术出发，构建了基于风险短语挖掘的知识聚合模型，如图 1 所示。该模型主要由知识采集模块、知识挖掘模块和知识服务模块三大模块所组成。其中，知识采集模块是知识挖掘模块和知识服务模块的基础，主要完成知识资源的采集；知识挖掘模块主要利用基于统计和基于规则的方法相结合来挖掘风险短语，形成可复用的风险短语知识库；知识服务模块则是基于风险短语知识库利用聚类分析、共现分析和知识检索等方法和技术完成多种形

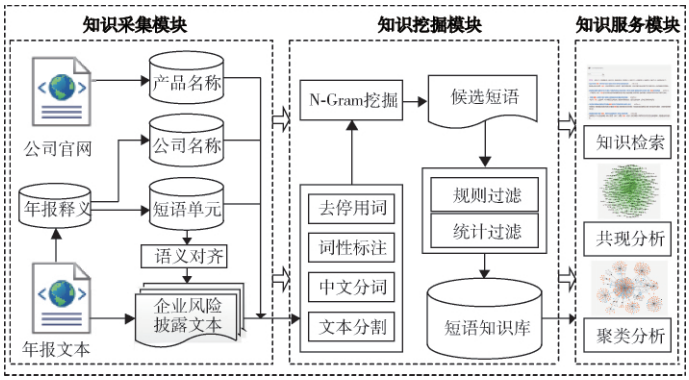


图 1 基于风险短语挖掘的知识聚合模型

式的知识聚合，从而为决策者提供智能化的金融知识服务。

## 3 知识资源采集

本文从巨潮资讯网下载医药制造类上市公司于2018年披露的年报，医药制造类上市公司的名单从中国证券监督管理委员会发布的行业分类结果中获取。我们去除没有披露企业所面临的风险信息的年报，最终共获得205家医药制造类上市公司的年报文本；将年报中披露企业风险信息的文本抽取出来，形成企业风险披露文本集合；年报中典型的风险信息披露形式如图2所示，我们将图2中的风险披露文本保存为三个独立的文本片，这三个文本片分别描述了该公司的原料血浆供应不足风险、产品研发风险和行业垄断风险。

（三）公司可能面临的风险

1、原料血浆供应不足风险

血液制品的原材料为健康人血浆，由于来源的特殊性及生活水平的提高，目前整个行业原料血浆供应较为紧张，原料血浆供应量直接决定血液制品生产企业的生产规模。

2、产品研发风险

生物制药行业是创新型行业，产品创新度越高失败的风险越大，且新产品研发成功并产业化后仍存在能否符合市场需求的风险，具有周期长、投入大、风险高的特点。

3、行业垄断风险

由于不断加强的监管政策提高了行业准入门槛和企业的经营成本，促使行业优胜劣汰和强强联合，促进了行业集中度的提高。虽然与国外比较，国内血液制品市场较为分散，但是集中化的趋势已经非常明显，一些同行业企业已启动并购步伐，走在行业前列，公司将加快发展步伐，促进企业做强做大，防范行业垄断风险。

图2 年报中典型的风险信息披露形式

上市公司年报的释义中不仅对公司名称等缩写进行了解释，同时还包含了部分短语知识单元；例如，年报释义中披露国家食药监总局指国家食品药品监督管理总局，因此我们可以从年报释义中获取部分风险短语。接下来我们需要利用从释义中获取的风险短语对企业风险披露文本进行语义对齐，具体规则是将语义对齐到使用频次最多的短语上。除此之外，我们从年报文本的释义部分获取公司名称，以及从公司的官网上获取企业的产品名称；并将公司名称、产品名称和已获取的风险短语存为用户词典，以便后续更好地对文本进行分词。

## 4 风险短语挖掘

在对知识资源采集的基础上，我们需要对企业风险披露文本集合进行挖掘以获取风险短语，从而形成风险短语知识库。首先，我们需要对企业风险披露文本集合进行文本预处理；然后，使用N-Gram算法来抽取候选风险短语；最后，利用基于统计和基于规则的方法相结合来实现候选短语的过滤，并通过领域专家进行人工辅助判断与筛选，形成可复用的医药制造类企业风险短语知识库。

### 4.1 文本预处理

在进行风险短语挖掘之前，我们首先需要对企业风险披露文本集合进行文本预处理。由于风险短语中不可能含有逗号、分号、句号和感叹号等标点符号，因此我们将这些标点符号定义为文本分割符并完成企业风险披露文本集合的分割，从而有助于过滤掉大量含有文本分割符的无意义的候选短语。随后，我们将公司名称、产品名称和已获取的风险短语作为用户词典，将哈尔滨工业大学停用词表作为停用词典，使用Jieba分词工具对企业风险披露文本集合进行分词、词性标注和去停用词，从而形成经过文本预处理后的企业风险披露文本语料。

### 4.2 候选风险短语抽取

我们利用N-gram算法来抽取候选风险短语，N-gram算法是一种基于统计语言模型的算法，其基本思想是通过一个大小为$N$的滑动窗口对文本内容进行切分，形成长度为$N$的字节片段序列，每一个字节片段称为gram，对所有的gram的出现频度进行统计，并且按照事先设定好的阈值进行过滤，形成关键gram列表。由于N-gram算法具有与语种无关、不需要语言学处理、不需要词典和规则等优点[17]，因此本文采用N-gram算法来抽取候选风险短语。我们将最小阈值设定为4，将$N$设定为2至4，即识别长度分别为2-gram、3-gram、4-gram的风险短语，从而形成候选风险短语集合。

### 4.3 基于规则的候选短语过滤

我们从以往短语挖掘的研究文献中获取部分短语过滤规则，同时针对企业风险披露文本的特点补充了部分规则，最终形成如下八条候选风险短语的过滤规则：

规则一：含有文本分割符的候选短语；

规则二：若候选短语是百科词条则直接确定为短语，例如“药品集中招标采购”等短语；

规则三：短语的首词不得为介词、连词、助词、区别词、后接成分；

规则四：短语的尾词不得为介词、连词、助词、前接成分、方位词、时间词、区别词；

规则五：短语中不得含代词、叹词、语气词、助词、拟声词、成语、标点符号、状态词、公司名；

规则六：短语中至少含有名词、动词、形容词、习用语、简称略语、后接成分；

规则七：若父串的频次和子串的频次相同，则只保留父串，例如“行业整合趋势”与“行业整合趋势加剧”的频次都为4，此时则保留“行业整合趋势加剧”；

规则八：含有“万元”等词的候选短语。

### 4.4 基于统计的候选短语过滤

参考刘彤等[29]的研究，本文使用扩展的多元互信息指标[30]来度量候选风险短语的紧密度。对于一个N-gram的候选风险短语 $A = A_1, A_2, \cdots, A_n$，$p(A)$ 和 $p(A_i)$ 分别为候选风险短语 $A$ 和词单元 $A_i$ 在语料中出现的频率，则候选风险短语 $A$ 的互信息为：

$$\mathrm{EMI}(A) = \log \frac{p(A)}{\sqrt{\prod_{i=1}^{n} p(A_i)}} \tag{1}$$

针对不同长度的N-gram候选风险短语的互信息不具有可比性的问题，我们将具有相同长度的N-gram候选风险短语进行归一化处理[29]，从而形成最终度量N-gram紧密度的指标：

$$\overline{\mathrm{EMI}(A)} = \frac{\mathrm{EMI}(A)}{\frac{1}{S_{|A|}} \sum_{A' \in S_{|A|}} \mathrm{EMI}(A')} \tag{2}$$

式中，$S_{|A|}$ 表示与短语 $A$ 具有相同长度的N-gram集合。同时，我们通过计算候选风险短语的左信息熵和右信息熵来衡量其左右边界的不确定性。信息论中的信息熵表示随机变量的不确定性，随机变量越不确定，其熵值越大。若候选风险短语的边界越不确定，则其信息熵越高，越可能是一个完整的短语[31]。左信息熵、右信息熵的公式[32]表述如下：

$$\mathrm{LE}(s) = -\sum_{l \in L} p(ls \mid s) \log_2 p(ls \mid s) \tag{3}$$

$$\mathrm{RE}(s) = -\sum_{r \in R} p(sr \mid s) \log_2 p(sr \mid s) \tag{4}$$

式中，$s$ 是候选风险短语；$l$ 和 $r$ 分别是候选风险短语的左邻接词和右邻接词；$L$ 表示 $s$ 的所有左邻接词的集合，$R$ 表示 $s$ 的所有右邻接词的集合。$p(ls \mid s)$ 表示在 $s$ 出现的情况下，$l$ 和 $s$ 共现的条件概率；$p(sr \mid s)$ 表示在 $s$ 出现的情况下，$s$ 和 $r$ 共现的条件概率。$\mathrm{LE}(s)$ 和 $\mathrm{RE}(s)$ 越大，表明候选短语的左右邻接词越不固定，则 $s$ 适合筛选为一个完整的短语[33]。

同时，我们还通过计算候选风险短语的C-value值来实现统计过滤。C-value值是由Frantzi等[34]提出，其计算主要基于如下的考虑：一个候选短语的C-value值与其在领域语料中的词频成正比，与其长度成正比；候选短语如果被嵌套，其权重会相应降低。C-value值的计算公式如下所示：

$$\text{C-value}(A) = \begin{cases} \log_2 |A| \times f(A) & A \text{ 未被嵌套} \\ \log_2 |A| \times f(A) - \dfrac{1}{c(A)} \sum_{i=1}^{c(A)} f(b_i) & \text{其他} \end{cases} \tag{5}$$

式中，$A$ 是某个候选短语；$|A|$ 是候选短语 $A$ 的长度；$f(A)$ 表示候选短语 $A$ 在领域语料中的词频；$b_i$ 表示嵌套 $A$ 的候选短语；$c(A)$ 表示嵌套 $A$ 的候选短语的数量。C-value值计算简单、适用性强，且具有语言和领域无关性。此外，C-value值考虑了候选短语的长度和嵌套性，在长短语与嵌套短语抽取方面具有一定的优势[35]。

为了获得高质量的风险短语，从而构建可复用的风险短语知识库，我们除了将基于规则和基于统计方法来实现短语的过滤外，还通过领域专家进行人工辅助判断与筛选[16,36]。我们将挖掘到的风险短语与从年报释义中获取的风险短语进行合并、去重，形成可复用的医药制造类企业风险短语知识库，知识库中共包含风险短语1334个。

## 5 基于风险短语的知识聚合

我们需要在形成的可复用的医药制造类企业风险短语知识库的基础上利用聚类分析、共现分析和知识检索等技术进行多种形式的知识聚合，从而为决策者提供智能化的金融知识服务。

### 5.1 基于风险短语的文本聚类分析

我们将经过文本预处理后的语料中的词汇序列替换为企业风险短语知识库中相应的风险短语，例如将“行业整合趋势加剧”替换为短语“行业整合趋势加剧”；然后去除语料中的非风险短语知识单元，形成只含风险短语的文本集合；最后使用向量空间模型VSM来对该文本集合进行表示，并将TF-IDF方法作为文本表示单元转换为向量的权重计算方法，最后使用K-means算法对该文本集合进行聚类。我们针对医药制造类上市公司的风险信息披露情况和专家的领域知识，将K-means算法聚类的簇设定为10；完成聚类后，我们根据短语描述为每个类簇确定一个风险标签，各个类簇中的排序前10的核心风险短语如表1所示。

我们可以直观地根据基于风险短语的文本聚类结果看出医药制造类上市公司面临的风险主要包括企业管理风险、行业政策风险、商誉减值风险、产品研发风险、产品质量风险、投资项目风险、公司经营风险、市场竞争风险、药品降价风险和企业盈利风险。与此同时，各个类簇的核心风险短语也描述了某一风险的具体情况，这些风险短语比单个的词包含了更加清晰和丰富的语义信息，且具有更强的可读性。

### 5.2 基于风险短语的共现分析

我们将经过文本预处理后的语料中的词汇序列替换为企业风险短语知识库中相应的风险短语，并去除语料中的非风险短语知识单元，形成只含风险短语的文本集合；然后使用中国医科大学开发的书目共现分析系统Bicomb来

表 1 基于风险短语的文本聚类结果

| 类簇序号 | 类簇标签 | 类簇中的排序前 10 的核心风险短语 |
|---|---|---|
| 1 | 企业管理风险 | 管理风险，更高要求，公司管理，业务规模，管理团队，市场开拓，生产管理，公司发展，组织结构，公司业务 |
| 2 | 行业政策风险 | 两票制，医保控费，行业政策风险，一致性评价，公立医院改革，二次议价，分级诊疗，行业政策变化风险，重大影响，医药行业未来发展 |
| 3 | 商誉减值风险 | 经营业绩，不利影响，商誉减值风险，实现收益，重大不利影响，减值测试，企业合并，公司主导产品，企业会计准则，汇率波动 |
| 4 | 产品研发风险 | 新药研发，研发风险，药品研发，临床试验，国家监管法规，注册法规严格，投入大，产品研发，投入大量资金，临床前研究 |
| 5 | 产品质量风险 | 产品质量，生产经营，药品质量，产品质量风险，不利影响，产品质量问题，药品生产，工艺复杂，特殊商品，质量问题 |
| 6 | 投资项目风险 | 募集资金投资项目，市场环境，发展战略，投资项目，销售渠道，gmp，产品销售，不利影响，重大变化，公司产品销售 |
| 7 | 公司经营风险 | 公司产品，公司经营，不利影响，公司业绩，经营业绩，生产销售，产品销售，公司主要产品，医药产品，原材料价格波动风险 |
| 8 | 市场竞争风险 | 环保风险，市场竞争，不利影响，市场竞争风险，应收账款，公司业绩，市场风险，更高要求，环保标准，产品销售 |
| 9 | 药品降价风险 | 药品价格，药品降价风险，招标采购，产品销售价格，公司产品，国家发改委，医保支付，市场竞争，医保目录，政府定价 |
| 10 | 企业盈利风险 | 新产品，盈利能力，不利影响，竞争优势，新产品开发，临床试验，医药市场，不可预测因素，体外诊断试剂，体外诊断行业 |

创建风险短语（频次高于 10 次）的共现矩阵，并使用社会网络分析工具 gephi 对风险短语共现矩阵进行挖掘和可视化，gephi 的风险短语可视化结果如图 3 所示。

根据风险短语的共现分析可视化图中风险短语节点的大小，可以直观地看出医药制造类上市公司的核心风险短语主要有“不利影响”“经营业绩”“盈利能力”“行业政策”“新药研发”“医保控费”“药品质量”“两票制”“一致性评价”和“市场竞争”等，这些风险短语较为集中地反映和体现了医药制造类上市公司所面临的行业政策风险、产品质量风险、产品研发风险和市场竞争风险等重要核心风险。

除此之外，我们还能针对具体的某一风险短语来呈现与之共现的风险短语；例如与“研发风险”短语共现的风险短语如图 4 中红色节点所示（图中为计算机显示颜色），主要包括“药品研发”“临床试验”“新产品研发”“临床研究”“高投入”“投入大”“难度大”“周期较长”“较大不确定性”“研发能力”“新药研发”“技术研发”“研发过程”和“药品注册”等风险短语。

与此同时，我们还可以针对具体的某一公司的风险短语进行可视化，从而为决策者提供更加细粒度和更加精准的金融知识服务。例如披露图 2 风险信息的 A 公司的风险短语可视化结果如图 5 所示，A 公司主要面临三种风险；第一种风险主要体现在原材料供应方面，第二种风险主要体现在产品研发方面，第三种风险主要体现在行业政策方面。

### 5.3 基于风险短语的知识检索

为了帮助决策者便捷有效地获取金融知识从而辅助其做出更加合理的决策，本文利用信息检索技术和软件工程技术实现了一个基于风险短语的企业风险知识检索原型系统；基于 MVC 设计模式和 B/S 系统架构，使用 Java 程序设计语言与全文搜索引擎库 Lucene 对原型系统进行了设计与实现。当用户以一定的检索词进行检索时，该原型系统会呈现出与用户检索词相关的风险短语以及风险短语所对应的文本；若用户输入的检索词为风险短语知识库存在的风险短语时，原型系统会为用户推荐与该风险短语相似的风险短语。我们使用 Ochiia 系数来度量风险短语之间的相似度，其计算公式如下：

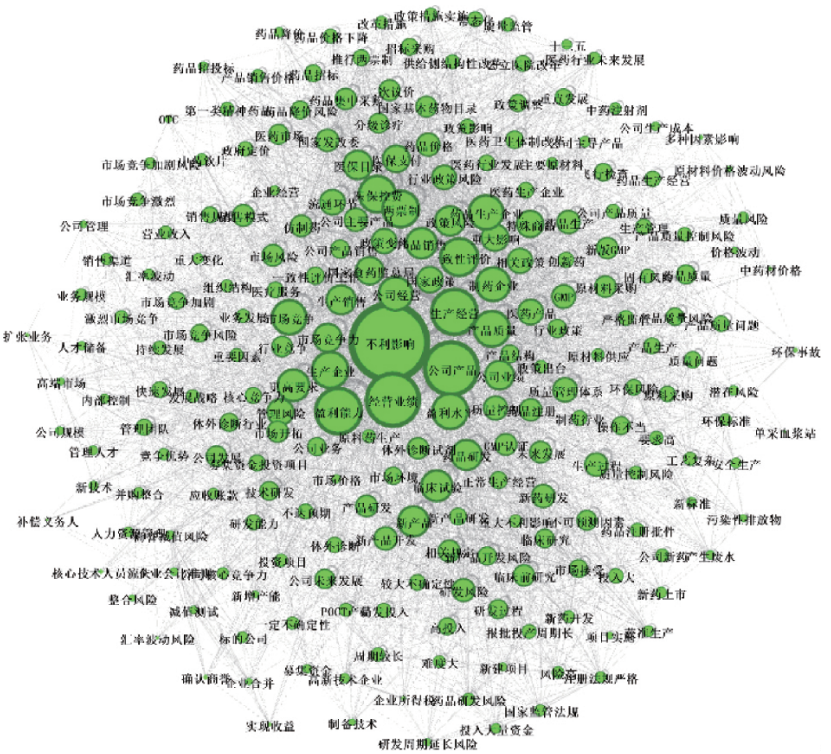


图 3 基于风险短语的共现分析可视化

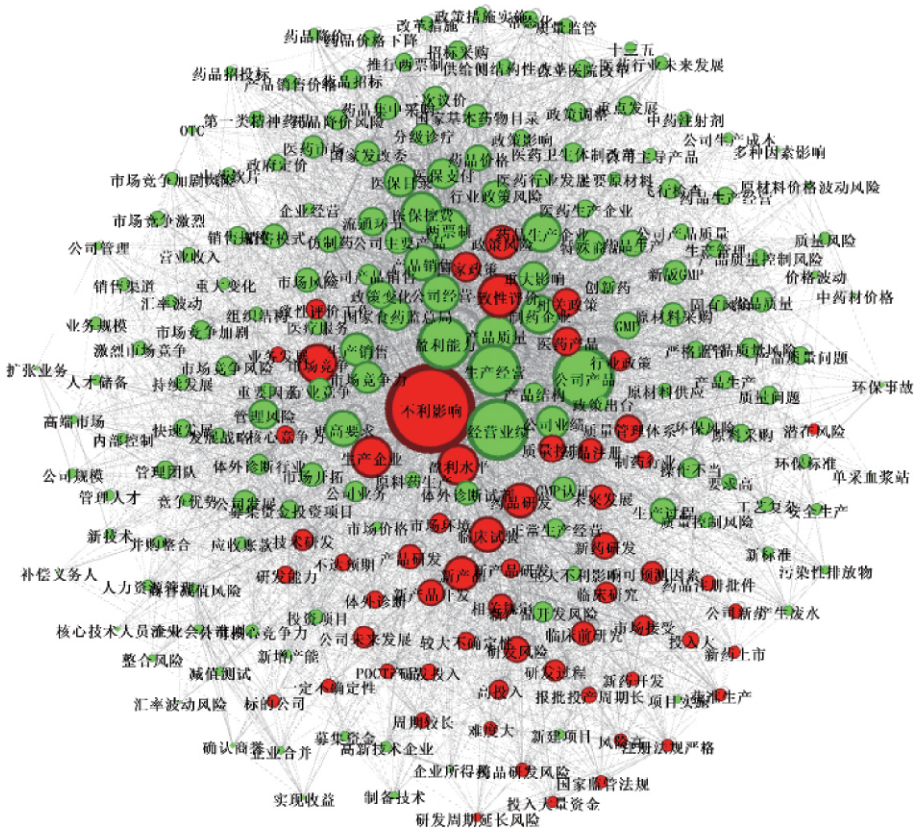


图 4 “研发风险”短语的共现风险短语

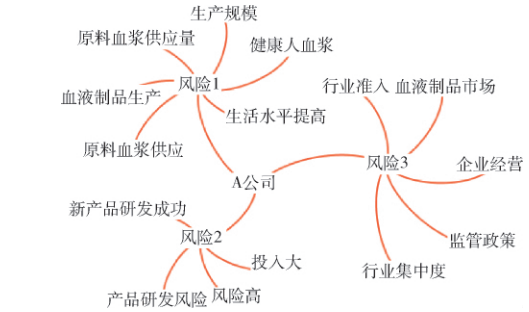


图 5 A 公司的风险短语可视化

$$\mathrm{Ochiia}_{ij} = \frac{N_{ij}}{\sqrt{N_i \times N_j}} \quad (6)$$

式中，$\mathrm{Ochiia}_{ij}$ 表示风险短语 $i$ 与风险短语 $j$ 的 Ochiia 系数；$N_{ij}$ 表示风险短语 $i$ 与风险短语 $j$ 的共现频次；$N_i$、$N_j$ 分别表示风险短语 $i$ 与风险短语 $j$ 出现的频次。

例如当用户输入检索词“两票制”时，原型系统的检索结果如图 6 所示。该原型系统会呈现出与检索词“两票制”相关的风险短语以及风险短语所对应的文本，由于短语“两票制”存在于风险短语知识库中，因此该原型系统同时还会呈现出与“两票制”这一风险短语最相似的 10 个风险短语及其相似度，如“一致性评价”“分级诊疗”“药品集中采购”“医保控费”和“行业政策变化风险”等风险短语。


图 6 基于风险短语的知识检索原型系统

## 6 总结与展望

大数据时代，金融行业积累的丰富数据资源为金融知识服务带来了新的契机，但这也对知识获取、知识组织和知识服务提出了更高的要求。本文从知识挖掘、知识组织和知识服务的相关理论、方法和技术出发，构建了基于风险短语挖掘的知识聚合模型；我们利用基于统计和基于规则的短语挖掘方法相结合来抽取企业风险短语知识单元，形成可复用的企业风险短语知识库；文章将短语作为知识聚合的粒度，并利用聚类分析、共现分析和知识检索等技术进行了多种形式的知识聚合，从而为决策者提供智能化的知识服务。

今后我们的研究主要集中于以下三个方面：首先，需要研究如何利用挖掘到的风险短语来实现本体知识库的半自动构建，这会涉及短语之间的语义对齐和语义融合，以

及短语之间语义关系的识别和挖掘等关键问题；然后，如今深度学习技术被广泛地应用于图像处理、语音识别和机器翻译等领域中，部分研究者也尝试将深度学习技术应用于经济金融领域的数据分析中[37-38]，然而深度学习模型的训练往往需要大量经过标注的样本，因此如何从百科词条、论文关键词等知识资源中获取大规模的短语样本是我们接下来要研究的关键问题之一；最后，需要利用语义本体、关联数据等技术将风险短语和其他与金融、经济、商务相关的多源开放数据进行链接、集成和整合，从而为构建一个具备互链接和互操作的全局金融商务信息生态系统奠定基础。□

## 参考文献


[1] 唐晓波，刘广超. 基于两层知识融合的金融知识服务模型研究 [J]. 图书馆学研究，2018（16）：79-85.

[2] WANG G，CHEN G，CHU Y. A new random subspace method incorporating sentiment and textual information for financial distress prediction [J]. Electronic Commerce Research and Applications，2018，29：30-49.

[3] 宋彪，朱建明，李煦. 基于大数据的企业财务预警研究 [J]. 中央财经大学学报，2015（6）：55-64.

[4] CHUNG W. BizPro：extracting and categorizing business intelligence factors from textual news articles [J]. International Journal of Information Management，2014，34（2）：272-284.

[5] 陈艺云. 大数据时代基于文本信息的信用风险管理研究 [J]. 金融理论与实践，2017（4）：14-20.

[6] HAJEK P，OLEJ V，MYSKOVA R. Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making [J]. Technological and Economic Development of Economy，2014，20（4）：721-738.

[7] HUANG K W，LI Z. A multilabel text classification algorithm for labeling risk factors in SEC form 10-K [J]. ACM Transactions on Management Information Systems，2011，2（3）：1-19.

[8] GLANCY F H，YADAV S B. A computational model for financial reporting fraud detection [J]. Decision Support Systems，2011，50（3）：595-601.

[9] 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.

[10] 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.

[11] SHANG J，LIU J，JIANG M，et al. Automated phrase mining from massive text corpora [J]. IEEE Transactions on Knowledge and Data Engineering，2018，30（10）：1825-1837.

[12] 荣垂田，李银银，王琰. 中文关键短语自动提取方法研究 [J]. 计算机科学与探索，2018，12（11）：1-14.

[13] 俞琰，赵乃瑄. 融入术语知识的专利主题发现方法 [J]. 图书情报工作，2018，62（21）：118-126.

[14] 马佩勋，高琰. 基于TF* PDF的热点关键短语提取 [J]. 计算机应用研究，2013，30（12）：3610-3613.

[15] HASAN K S，NG V. Automatic keyphrase extraction：a survey of the state of the art [C] //Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics，2014：1262-1273.

[16] 颜端武，李兰彬，曲美娟. 基于N-gram复合分词的领域概念自动获取方法研究 [J]. 情报理论与实践，2014，37（2）：122-126.

[17] 段宇锋，鞠菲. 基于N-Gram的专业领域中文新词识别研究 [J]. 现代图书情报技术，2012，28（2）：41-47.

[18] 汤青，吕学强，李卓，等. 领域本体术语抽取研究 [J]. 现代图书情报技术，2014，30（1）：43-50.

[19] 赵雪芹. 知识聚合与服务研究现状及未来研究建议 [J]. 情报理论与实践，2015，38（2）：132-135.

[20] 李亚婷. 知识聚合研究述评 [J]. 图书情报工作，2016，60（21）：128-136.

[21] 陈果，吴微，肖璐. 知识共聚：领域分析视角下的知识聚合模式 [J]. 图书情报工作，2018，62（8）：115-122.

[22] 赵蓉英，王嵩，董克. 国内馆藏资源聚合模式研究综述 [J]. 图书情报工作，2014，58（18）：138-143.

[23] 邱均平，王菲菲. 基于共现与耦合的馆藏文献资源深度聚合研究探析 [J]. 中国图书馆学报，2013，39（3）：25-33.

[24] 李洁，毕强. 数字图书馆资源知识聚合可视化模型构建研究 [J]. 情报学报，2016，35（12）：1273-1284.

[25] 陈果，肖璐，孙建军. 面向网络社区的分面式导航体系构建——以丁香园心血管论坛为例 [J]. 情报理论与实践，2017，40（10）：112-116.

[26] 吕琳露，李亚婷. 游记文本中的知识发现与聚合——以蚂蜂窝旅行网杭州游记为例 [J]. 情报杂志，2017，36（7）：176-181.

[27] 陶兴，张向先，郭顺利. 基于DPCA的社会化问答社区用户生成答案知识聚合与主题发现服务研究 [J]. 情报理论与实践，2019，42（6）：94-98.

[28] 肖璐. 基于知识超网络的网络社区学术资源多粒度聚合研究 [J]. 情报杂志，2018，37（12）：182-187.

[29] 刘彤，倪维健，柳梅. 面向搜索引擎查询日志的领域术语自动识别方法 [J]. 现代图书情报技术，2016，32（2）：25-33.

[30] VAN DE CRUYS T. Two multivariate generalizations of pointwise mutual information [C] //Proceedings of the Workshop on Distributional Semantics and Compositionality. Association for Computational Linguistics，2011：16-20.


（下转第139页）

information in online consumer reviews diagnostic over time? The role of review relevancy, factuality, currency, source credibility and ranking score [J]. Computers in Human Behavior, 2018, 80: 122-131.

[3] 孟美任，丁晟春. 在线中文商品评论可信度研究 [J]. 现代图书情报技术，2013 (9): 60-66.

[4] 黄婷婷，曾国荪，熊焕亮. 基于商品特征关联度的购物客户评论可信排序方法 [J]. 计算机应用，2014，34 (8): 2322-2327.

[5] 郝玫，杨晓媛. 中文网络客户评论可信度研究 [J]. 现代图书情报技术，2015 (2): 55-63.

[6] LU T C, YU T, CHEN S H. Information manipulation and web credibility [C] //International Symposium on Distributed Computing and Artificial Intelligence. Springer, Cham, 2017: 86-95.

[7] 王忠群，吴东胜，蒋胜，等. 一种基于主流特征观点对的评论可信性排序研究 [J]. 数据分析与知识发现，2017，1 (10): 32-42.

[8] 王腾，朱青，王珊. 基于语义相似度的 Web 信息可信分析 [J]. 计算机学报，2013，36 (8): 1668-1681.

[9] 邓莎莎，张朋柱，张晓燕，等. 基于欺骗语言线索的虚假评论识别 [J]. 系统管理学报，2014，23 (2): 263-270.

[10] BHATTACHERJEE A, SANFORD C. Influence processes for information technology acceptance: an elaboration likelihood model [J]. MIS Quarterly, 2006, 30 (4): 805-882.

[11] SEN S, LERMAN D. Why are you telling me this? An examination into negative consumer reviews on the web [J]. Journal of Interactive Marketing, 2007, 21 (4): 76-94.

[12] CHEUNG M Y, SIA C L, KUAN K K Y. Is this review believable? A study of factors affecting the credibility of online consumer reviews from an ELM perspective [J]. Journal of the Association for Information Systems, 2012, 13 (8): 618-635.

[13] JENSEN M L, AVERBECK J M, et al. Credibility of anonymous online product reviews: a language expectancy perspective [J]. Journal of Management Information Systems, 2013, 30 (1): 293-324.

[14] FILIERI R. Why do travelers trust TripAdvisor? Antecedents of trust towards consumer-generated media and its influence on recommendation adoption and word of mouth [J]. Tourism Management, 2015, 51: 174-185.

[15] 陈德华，殷苏娜，乐嘉锦，等. 一种面向临床领域时序知识图谱的链接预测模型 [J]. 计算机研究与发展，2017，54 (12): 2687-2697.

[16] 王忠群，皇苏斌，修宇，等. 基于领域专家和商品特征概念树的在线商品评论深刻性度量 [J] 现代图书情报技术，2015 (9): 17-25.

[17] 马飞翔，廖祥文，於志勇，等. 基于知识图谱的文本观点检索方法 [J]. 山东大学学报 (理学版)，2016，51 (11): 33-40.

[18] HanLP [EB/OL]. [2019-04-01]. http://hanlp.linrunsoft.com/.

[19] HIT-CIR Tongyici Cilin (Extended) [EB/OL]. [2014-12-28]. http://ir.hit.edu.cn/demo/ltp/Sharing_Plan.htm/.

[20] 张艳丰，李贺，彭丽徽，等. 基于情感语义特征抽取的在线评论有用性分类算法与应用 [J]. 数据分析与知识发现，2017，1 (12): 74-83.

**作者简介：** 王忠群 (ORCID: 0000-0002-5307-5706)，男，1965 年生，教授。研究方向：信息管理与信息系统。叶安杰，男，1996 年生，硕士生。皇苏斌，男，1986 年生，讲师。研究方向：信息系统。陈云霞，女，1967 年生，副教授。研究方向：医学信息管理。

**作者贡献声明：** 王忠群，设计研究方案，论文最终版本修订。叶安杰，论文初稿、实验验证。皇苏斌，提出部分研究方案，部分实验工作。陈云霞，提出部分研究方案，初稿修改。

**录用日期：** 2020 - 03 - 09

(上接第 158 页)

[31] 李丽双，王意文，黄德根. 基于信息熵和词频分布变化的术语抽取研究 [J]. 中文信息学报，2015，29 (1): 82-87.

[32] 任禾，曾隽芳. 一种基于信息熵的中文高频词抽取算法 [J]. 中文信息学报，2006，20 (5): 40-43.

[33] 唐晓波，胡华. 中文社会化媒体的本体概念抽取研究 [J]. 情报科学，2014，32 (4): 9-15.

[34] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms: the c-value/nc-value method [J]. International Journal on Digital Libraries, 2000, 3 (2): 115-130.

[35] 熊李艳，谭龙，钟茂生. 基于有效词频的改进 C-value 自动术语抽取方法 [J]. 现代图书情报技术，2013，29 (9): 54-59.

[36] 韩红旗，朱东华，汪雪锋. 专利技术术语的抽取方法 [J]. 情报学报，2011，30 (12): 1280-1285.

[37] TSAI M F, WANG C J, CHIEN P C. Discovering finance keywords via continuous-space language models [J]. ACM Transactions on Management Information Systems, 2016, 7 (3): 1-17.

[38] KRAUS M, FEUERRIEGEL S. Decision support from financial disclosures with deep neural networks and transfer learning [J]. Decision Support Systems, 2017, 104: 38-48.

**作者简介：** 唐晓波，男，1962 年生，教授，博士生导师。研究方向：知识组织与情报研究。谭明亮 (通讯作者)，男，1990 年生，博士生。研究方向：知识组织与商务智能。李诗轩，女，1993 年生，博士生。研究方向：语义分析与商务智能。顾娜，女，1995 年生，硕士生。研究方向：知识组织与情报研究。

**录用日期：** 2020 - 01 - 21