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): 1268-1274

DOI: 10.3772/j.issn.1000-0135.2019.12.004

一种专利与企业相关性测度方法

高影繁[1]，王　峥[2]，胡小荣[1]，姚长青[1]，梁　娜[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　随着我国企业每年专利申请量的不断增多，一些专利申请乱象日渐显现：国家为专利申请制定了扶持与鼓励政策，有的企业为了争取这些政策奖励而胡乱申报或购买对于企业发展与利润提升无积极影响、且与企业主营产品无关的专利。在这种背景下，专利与企业相关性测度研究就成为一个有价值的研究课题。针对当前专利与企业相关性判断依赖领域专家的主观评价现状，提出一种专利与企业相关性的自动测度方法，为传统主观评价方法提供替代思路。采用信息论与深度学习方法进行专利的标引词识别，通过测度专利标引词在企业文本的映射程度来实现专利与企业相关性的自动量化。本研究对环保企业专利与企业相关性做整体评价判断，在任意选取的4325份环保领域专利中，有4065个一致性结果，260个不一致结果，其准确率达到94%，验证了本文方法的可行性。

关键词　专利与企业相关性；短语抽取；词向量过滤；标引词权重

A Method for Measuring the Relevance Between Patents and Enterprises

Gao Yingfan[1], Wang Zheng[2], Hu Xiaorong[1], Yao Changqing[1] and Liang Na[1]

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

Abstract: With the number of patent applications in China growing each year, some of these applications are becoming increasingly chaotic, as evident from the actions of certain enterprises in declaring or purchasing patents indiscriminately in order to win policy awards under the state's support and encouragement policies for patent applications. This can result in patents that have no positive impact on corporate development or profit improvement, and that are not related to the company's main products. In this context, patent and enterprise relevance measurement research has become a valuable research topic. In considering the subjective evaluation methods of experts in the field of patent and enterprise relevance judgment, this study proposes a more valuable, innovative application method and provides an alternative to the traditional subjective evaluation methods. Information theory and deep learning methods are used to identify the index of patents. Additionally, the degree to which patent index words are mapped in corporate texts is measured in order to achieve an automatic quantification of patent and enterprise relevance. This study conducts an overall evaluation of the relevance of patents and enterprises in the environmental protection industry. Among the 4325 patents in the environmental protection field, there are 4065 consistent results and 260 inconsistent results, or an accuracy rate of 94%. The feasibility of the method is thus verified.

Key words: patent and enterprise relevance; phrase extraction; filtering by word vector; weights for indexing words

收稿日期：2019-05-22；**修回日期**：2019-09-22

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：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，E-mail：gaoyingf@istic.ac.cn；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；胡小荣，女，1993年生，硕士研究生，主要研究方向为文本挖掘；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；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。

1　引　言

随着科学技术的日益发展，世界经济已经逐步迈向知识经济时代，企业的发展也逐渐向以知识和技术为核心的发展方向转变。专利作为企业创新活动的重要产物，已经成为企业核心技术与核心竞争力的体现。我国企业每年的专利申请量不断攀升，与此同时，专利的质量问题却逐渐显露。比如，国家出台相关政策鼓励专利申请，企业却为了获得政策奖励胡乱申报专利，专利质量不高，对企业主营产品（服务）无直接作用，且无益于企业自身发展或利润提升。“伪高新”现象就是一个典型案例，国家为高新技术企业提供了一系列税收优惠政策，有些企业为了得到免税优惠，存在购买专利或者申请与公司关联度不大的专利的情况，以“贝因美”为例，该公司连续3年研发支出比例过低，且所申报发明专利与主要产品不直接相关，最终于2011年9月29日被通告取消高新技术企业资格。

今年3月，科创板开始上线试行。科创板的特殊之处在于强调企业的科创实力，尤其以能够表征企业自主创新能力的专利数量作为重要衡量指标。在此背景下，如何判断企业的专利有效性，成为未来科创板健康、稳定发展的首要关注点。对于企业与审核机构而言，准确判断专利有效性能够提高专利申请效率、降低审核错误率，而判断专利有效性的一个重要因素为专利与企业的相关性。本文提出一种专利与企业相关性的测度方法，通过构建企业描述文本，从专利文本内容出发，根据专利文本领域相关性强、规范化等特点，使用自动标引、短语识别、相关性计算等机器学习方法自动量化专利与企业的相关性。

2　国内外研究现状

目前关于专利与企业相关性的研究多从企业专利申请量与企业利润的相关性角度出发进行分析。

首先，国外学者对专利与企业利润之间的相关性已经通过大量实证研究做过分析。Griliches[1]最早于1990年对专利申请数量与企业的市场价值之间的相关性做了相关表述，并选取美国340家企业在1970—1980年10年间的专利做了实证研究；紧接着，Ernst[2]于2001年对德国工具制造业50家企业的专利申请数量与企业绩效进行相关性分析，结果指出当年的专利申请量与企业滞后2~3年的绩效增长呈正相关。Reitzig[3]则于2004年对美国机械制造领域的企业专利进行研究，指出企业专利的活动对于企业绩效有正向作用，且效果较为显著；Anandarajan等[4]于2007年分别对中国台湾和美国的企业专利进行研究，结果指出专利授权对企业市场价值有正影响。

国外多数学者采用专利申请量作为指标衡量专利与企业的相关性，一方面是该数据可直接从专利数据库中获取；另一方面，专利申请量数据是可以作为研究的基础数据，而且结果均较为统一，认为专利申请量对于企业绩效有正向作用。而国内研究结果较为不一致。唐恒等[5]、赵远亮等[6]均以医药上市公司为样本进行研究，结果指出不同类型专利的数量对企业绩效会有不同的影响；苑泽明等[7]对高新技术上市企业的专利申请数量与企业未来的业绩的关系进行研究，结果发现二者并没有显著的相关关系。这与我国的实际情况相关，在国家税收优惠政策的大背景下，有一些公司追求专利数量而不看重专利质量，因此这些企业的专利技术含量较低，对企业的长期发展并不会起更大的作用。

而近年来，国外学者开始意识到专利与企业业绩之间的关系不能仅用专利申请量来衡量，原因在于专利本身的质量：有些企业专利质量高、影响力大，那么可以仅凭几项专利获取较大的利润，使其领先于领域内的其他企业；而有些企业专利质量差，这时即使该企业拥有大量专利，但专利影响力不大，技术含量低，并不能完全发挥作用[8]。因此许多学者将专利被引频次作为衡量专利质量的指标引入专利与企业利润的相关性分析中。这些学者认为，高被引专利具有对该领域较高的影响力与技术含量，因此其重要程度较高，对企业利润的影响也就越高。Chen等[9]对英国医药企业专利被引次数与企业的市场价值进行研究，结果发现为U形关系；Artz等[10]对专利质量与企业业绩增长进行研究，发现其呈正相关关系。而国内学者也开始注意到专利引用行为及施引频次对专利价值评估的重要性[11]，大多数研究表明高质量的专利能够帮助企业在行业中获得技术领先地位，生产出的新产品能够获得市场的认可，从而帮助企业提升业绩[12]。

综上所述，目前关于专利与企业相关性的研究集中于专利与企业业绩之间的关系，从专利申请数量与专利质量角度出发，宏观上研究专利对于企业利润的影响，而关于专利与企业主要产品（服务）的直接相关性判断通常由领域专家进行人工判断。本文提出了一种专利与企业相关性的自动量化方

法：构建企业描述文本，从专利文本内容出发，根据专利文本领域相关性强、规范化等特点，使用自动标引、短语识别、相关性计算等机器学习方法自动量化专利与企业的相关性，下面将对其中涉及的关键技术研究现状进行详细介绍。

3 专利与企业相关性界定与研究方法

3.1 专利与企业相关性界定

不同学者从多个角度对专利与企业相关性的定义进行描述与界定。

（1）关于专利与企业相关性的研究多从企业专利申请量与企业利润的相关性角度出发进行分析，多数学者认为企业技术创新的成果往往体现在企业专利申请量的增加和企业专利资产的增加[13]。

（2）在专利价值评估指标体系中，有学者将专利与企业相关性作为市场因素下的一个二级指标。企业因为需要保护核心技术、增加核心竞争力而申请专利，因此企业在申请专利前可先对该专利与企业的关联程度进行分析，判断是否有必要申请该专利，从而提高专利申请效率，减少专利维护费用。

（3）在国家高新技术企业认定条件中规定，企业需要对其主要产品的核心技术拥有自主知识产权的所有权，而在核心自主知识产权的评价指标中包含该知识产权对主要产品（服务）在技术上发挥核心支持作用，即认为专利应主要与企业主要产品（服务）直接相关。

基于以上观点，本文认为，专利与企业相关性指的是专利内容与企业主要产品、主营业务、核心技术等文本描述的相关性，并可根据相关性计算结果的大小来判断专利与企业的相关性强弱。

3.2 研究方法

本文提出的专利与企业相关性测度流程如图 1 所示。

3.2.1 专利特征标引词抽取及权重计算方法

1）融合多统计量的短语抽取方法

本文利用互信息的值来计算候选短语内部组成词的结合紧密度，利用左邻接熵熵与右邻接熵来判断短语的组成边界，并结合词频进行短语提取，将互信息、左熵、右熵与词频进行拟合，得到拟合值 Score，并设置阈值进行短语识别，Score 值的计算方式为

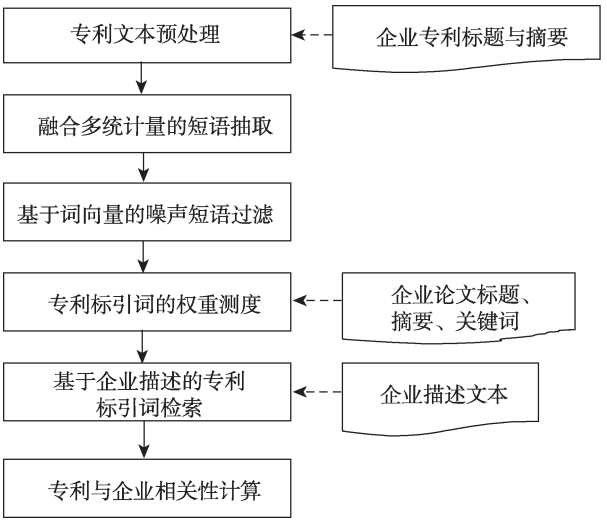

图 1 专利与企业相关性测度流程

$$\text{Score} = \left(\text{NorFreq} + \text{NorMI} + \text{NorLE} + \text{NorRE}\right)/4 \quad (1)$$

其中，NorFreq、NorMI、NorLE、NorRE 分别为词频、互信息、左右邻接熵归一化后的值，采用最大最小归一化处理方法。Score 值越高，代表该候选短语成为一个短语的可能性更高；反之，则说明该候选短语成为一个短语的可能性更低。

2）基于词向量的噪声短语过滤方法

用词向量表达文本中的词语是将深度学习算法的一个典型应用。词向量通过训练神经网络语言模型对文本特征进行描述，用 WordEmbedding 形式进行词表示，把词映射到 K 维向量空间，能够表达词语更丰富的语义信息，在向量空间上距离相近的词语在语义上也有近似的关系。本文认为，在语义上相近的两个词组成短语的可能性更高，以此作为噪声短语过滤的依据。本文采用 word2vec 进行词向量的训练，通过词向量之间的距离（如 cosine 相似度、欧氏距离等）来判断词之间的语义相似度，若两个词的语义相关性越大，其组成短语的可能性更大。本方法在融合多统计量进行短语抽取的基础上进行短语过滤，用以提高短语识别的准确性。

3）专利标引词权重计算方法

经过上述两个步骤之后，需将抽取的短语与专利预处理后的词进行合并去重后作为候选标引词。标引词权重的计算需要企业论文数据的支持，计算方法为

$$w_i = \text{idf}_i * \left(p\_\text{tf}_i + c\_\text{tf}_i\right) \quad (2)$$

其中，p_tf_i 表示在专利标题和摘要中该词的词频强度，用(出现次数+1)/(总词数+1)计算，对于没有出现的词，+1 可以起到平滑的作用；c_tf_i 表示公司发

表论文中该词的词频强度，c_tf的计算需要获取企业的论文数据，即以该企业名（包括现用名、曾用名）来检索学术论文数据库，获取以该企业为作者单位的学位论文，提取论文标题、摘要、关键词等字段的信息，计算方法为

$$c\_tf = \sum_{i \in I} c\_tf(i) * c\_weight(i) \tag{3}$$

其中，I为企业论文中的字段集合，包括论文标题、摘要、关键词3个字段，i为第i个字段；$c_tf(i)$为标引词在论文第i个字段中出现的次数；$c_weight(i)$为论文中第i个字段的权重，实验中依据经验为论文标题、摘要和关键词设定了固定权重。

3.2.2　基于企业描述文本的专利标引词检索方法

对获取的专利标引词，可以通过在企业描述文本中进行检索的方式，来获知专利标引词与企业描述的重合程度。本文认为，专利标引词在企业描述文本中出现的位置和频率可以反映该标引词的重要性。本研究选取了上市公司年报中的研发重点项目、行业关键技术、核心竞争力、董事会讨论、主要产品、经营范围、风险字段共7个字段构建企业描述文本，对这7个字段设置不同的权重来计算专利标引词在企业描述文本中的词频强度r_tf，

$$r\_tf = \sum_{i \in I} r\_tf(i) * r\_weight(i) \tag{4}$$

其中，I为企业描述文本中的字段集合，包括上面提到的7个字段，i为第i个字段；$r_tf(i)$为专利标引短语在描述文本第i个字段中出现的次数；$r_weight(i)$为企业描述文本中第i个字段的权重。

3.2.3　专利与企业相关性计算方法

本文认为，专利文本的标引词如果在企业描述文本中出现，那么该词就将专利文本与企业描述文本进行了关联。专利标引词的权重及其在企业文本中出现的词频强度可以反映专利与企业的相关性强弱程度。在第3.2.1节和第3.2.2节的计算基础上，本文以专利标引短语出现在企业描述文本中的词频强度与标引词权重的乘积作为评价专利与企业相关性强弱程度的评价指标。具体相关性强度计算方法为

$$r = \sum_{k_i \in K} w(k_i) * r\_tf(k_i) \tag{5}$$

其中，K为专利的标引词集合，k_i为第i个标引词；$w(k_i)$为标引词k_i在专利中的权重（计算方法见公式(3)）；$r_tf(k_i)$为标引词在企业描述文本中的词频强度（计算方法见公式(4)）；r为专利与企业相关度值。

计算得到专利与企业相关性强弱数值后，设置不同阈值，将专利与企业相关性强度分为“强相关”、“弱相关”、“不相关”3个等级，然后与人工标注结果进行对比分析，用准确率对本文方法效果进行评价。

4　实验及结果分析

4.1　数据来源及预处理

实验数据主要包括企业专利数据、企业论文数据以及词向量库。

（1）企业年报数据来自中国科学技术信息研究所自建的上市公司年报数据库，本文选取年报中的研发项目重点、行业技术_关键技术、核心竞争力、董事会讨论、主要产品、经营范围、风险字段共7个字段构建企业描述文本。其中，①研发项目重点、行业技术_关键技术、核心竞争力、主要产品4个字段为企业自身主营产品（服务）、关键技术等的直接描述，用词专业、领域性强，权重最高；②经营范围字段是本公司经营业务的总体介绍，范围较大、用词较为笼统，风险字段从市场、技术、政策等方面描述企业所面临的风险，这两类字段的权重设定为第二个等级；③董事会讨论字段是管理层对于本企业过去经营状况的评价分析以及对企业未来发展趋势的前瞻性判断，描述范围更加宽泛，权重最低。本文采用简单排序编码法，按照自然数顺序将3类权重分别设为3、2、1，归一化处理后，分别对应0.5、0.3和0.2。

（2）企业专利数据来自中国科学技术信息研究所中文专利数据库，共获取环保领域专利数据4801条，提取发明专利号、专利名称、摘要、专利权人、法律状态、专利类型等字段。

（3）论文数据来源于中国科学技术信息研究所中文论文数据库，共获取环保领域上市企业论文906篇。统计论文所发表的期刊及其出现频次，并按照频次降序排列，选择出现频次最高的前10个期刊作为环保领域的核心期刊。在中文论文数据库中，查找10个期刊的全部论文并下载其包括关键词在内的著录信息，获取论文4万多篇。最终获得并统计84家环保企业论文关键词，共3564个。

（4）本实验利用中文专利全库采用word2vec模型来训练中文专利词向量库，结果如表1所示；中文专利的预处理过程包括分词、去停用词、词性标

表 1 词向量库

词向量库	语料来源	文本大小	向量维度	词向量库大小
专利词向量库	中文专利数据库	约 10G,千万条文本,不重复词 1082189 个	100 维	990M

注等，根据词性标注结果提取实词作为候选词。

4.2 短语抽取实验及评价结果

本实验采用第 2.2.1 节描述的融合多统计量的短语抽取方法和基于词向量的噪声短语过滤方法来抽取专利标引短语，即：首先融合词频、互信息、左右邻接熵等统计量进行初步短语识别，然后利用专利词向量进行过滤，以提高短语识别准确率，通过实验最终识别出 4894 个短语，前 20 个短语如表 2 所示。

表 2 短语抽取结果中的前 20 个短语

Top 1~5	Top 5~10	Top 10~15	Top 15~20
吸收式热泵	煤粉燃烧器	电压电流	定子铁芯
功率单元	信号电路	热力膨胀阀	烟气余热
溴化锂吸收式	湿式电除尘器	余热锅炉	风冷冷凝器
袋式除尘器	高压变频器	烟气换热器	脱硫脱硝
非晶合金	板式换热器	空调机组	电流电压

为了对实验结果的正确性进行评价，采用人工标注的方式将结果分为正确短语与噪声短语。邀请 3 位情报学领域研究生通过原文本回溯与网络检索的方式对初步识别的 4894 个短语进行标注，当 3 人的判断结果中有 2 人判断为正确时即可标注为正确。分别选取前 500、1000 和 2000 个短语，根据标注结果判断短语识别的正确率，结果如表 3 所示。

表 3 短语识别结果准确率评价

短语数	前 500 个	前 1000 个	前 2000 个
准确率	96.2%	94.5%	92.0%

4.3 短语识别阈值调整方案

在依据第 2.2.1 节方法确定词频、互信息、左右邻接熵、Score 等的阈值时，需要通过反复实验进行确定。根据本文所选环保领域专利语料，经过反复实验与结果比较，确定了词频、互信息、左右邻接熵的阈值。首先将词频阈值设置为 2，将词频大于等于 2 的词串加入候选词串。互信息的值越高，候选短语内部组成词的结合程度较为紧密；将互信息值大于 0 作为候选短语组成词内部紧密结合程度的一个判定标准，将左右邻接熵的阈值均设置为 0.1。在利用词向量进行短语过滤时，经过反复实验，最终采用 Score×0.8+词向量相似性值×0.2 的方式进行拟合，噪声短语过滤结果最好。Score 值越高的短语在实验结果中排序越靠前的可能性越大。短语识别实验最终获取到短语共 2681 个短语，与数据预处理时获得的候选词集合进行合并去重，最终得到 52230 个候选词集合用于专利标引。

4.4 专利短语标引实验结果

使用第 3.2 节得到的词集合作为专利标引候选词，并以第 2.2 节专利中出现的短语及其频率为依据。以北京碧水源科技股份有限公司部分专利为例，短语识别结果如表 4 所示（由于论文篇幅限制，只展示前 5 个短语）。

4.5 专利与企业相关性计算实验结果

依据第 2.2.3 节专利与企业相关性计算方法，以北京碧水源科技股份有限公司部分专利为例，计算结果如表 5 所示。为了确定“强相关”、“弱相关”、“不相关”3 类专利的阈值，需通过设定不同阈值进行尝试。多次实验结果表明，当专利在企业描述文本中未命中时，确认该专利与企业“不相关”；当相关性归一化值低于 0.2 时，该专利与企业“弱相关”；当相关性归一化值大于 0.2 时，该专利与企业“强相关”。

4.6 结果分析

由表 4 可以看到，专利文本标引词中短语的比例很高，但是仍然存在“膜”等单字词和“过滤”等二字词，这是因为考虑到环保领域的特性、词在领域中的重要性以及主题表达能力，与基于词的专利标引词相比，短语的表达具有更为丰富的语义信息。

为了对本文方法专利与企业相关性的计算结果进行评价，我们选择了 3 位具有化学、环保领域本科专业背景的、情报学领域的研究生对专利与企业的相关性进行人工判断，为降低评价难度，将评价等级分为了“相关”和“不相关”两种，如果有两个或两个以上判断结果相同，则取相同结果作为最终人工判断结果。本实验对环保企业专利与企业相关性做整体评价判断，在任意选取的 4325 份环保领域专利（约全部环保企业专利的 90%）中，有 4065 个一致性结果，260 个不一致结果，其准确率达到 94%。

表 4　北京碧水源科技股份有限公司专利标引结果示例

专利标题	专利标引短语
一种带有吊装部件的大型膜生物反应器组器	处理；膜；污水；生物反应器；膜生物反应器
脉冲错流式膜生物反应器	系统；膜；曝气；生物反应器；反应器
一种多孔膜表面的永久亲水改性方法与采用所述方法得到的多孔膜	膜；多孔膜；膜表面；聚合物；亲水改性
一种有机废水的双膜处理方法	膜；处理；废水；污水；膜生物反应器
一种生物难降解污水的深度处理方法	降解；污水；处理方法；难降解；深度处理
一种脉冲曝气式膜生物反应器装置	曝气；膜；生物反应；曝气装置；生物反应器
一种控制由膜生物反应器混合液造成的严重膜污染的方法	膜；调控剂；混合液；膜污染；污染
一种中空纤维膜,所述膜的生产方法与用途	清洗；膜；回收；中空纤维；中空纤维膜
一种异质结构的中空纤维膜的制备方法	膜；中空纤维膜；纤维膜；异质结构；制备方法
生物填料摇动床	生物填料；填料；处理；废水；水中有机物
一种微污染水的处理方法	处理方法；污染水；微污染水；生物反应器；膜
一种有机废水处理方法	废水；生物反应器；膜；膜生物反应器；处理
一种干态中空纤维膜的通量恢复方法	膜；中空纤维；中空纤维膜；膜的通量；恢复
一种带衬型中空纤维复合膜的制备方法及其产品	膜；制膜液；中空纤维；纤维复合膜；制备方法
一种两级 A/O-MBR 脱氮除磷装置	膜；脱氮；除磷；脱氮除磷 ;MBR
中空纤维膜清水通量的测定装置	膜；中空纤维；水通量；纤维膜；清水通量

表 5　北京碧水源科技股份有限公司结果示例

申请号	专利标题	相关性值	相关性判断
CN2007201548841	一种带有吊装部件的大型膜生物反应器组器	1	强相关
CN2009102434752	一种控制由膜生物反应器混合液造成的严重膜污染的方法	0.934507	强相关
CN2006101408460	膜生物反应器-臭氧联合工艺生产再生水的方法	0.891963	强相关
CN2008101131315	一种有机废水的双膜处理方法	0.856615	强相关
CN2011104097712	一种异质结构的中空纤维膜的制备方法	0.835186	强相关
CN2007101784603	一种中空纤维膜,所述膜的生产方法与用途	0.814684	强相关
CN2008100974272	强化内源反硝化的膜-生物反应器脱氮除磷工艺及装置	0.810281	强相关
CN2011104150451	一种多孔膜表面的永久亲水改性方法与采用所述方法得到的多孔膜	0.793269	强相关
CN2010106095505	双环沟 MBR 废水处理系统	0.787767	强相关
CN201020685005X	双环沟 MBR 废水处理系统	0.787767	强相关
CN2011103991110	一种干态中空纤维膜的通量恢复方法	0.777899	强相关
CN2007201414866	一种脉冲曝气式膜生物反应器装置	0.774292	强相关
CN2009202991499	脉冲错流式膜生物反应器	0.77019	强相关
CN2009201092382	一种处理微污染地表水的装置	0.768097	强相关
CN2011103478806	一种带衬型中空纤维复合膜的制备方法及其产品	0.758822	强相关
CN2012102300887	一种城市污水深度处理方法	0.749018	强相关
CN2008100557170	一种共混膜,所述膜的生产方法与用途	0.747348	强相关

5　结　论

本文提出了一种融合深度学习、信息论与检索策略的专利与企业相关性计算方法，从文本内容的角度出发对专利与企业相关性进行自动量化，并对其中涉及的自动标引、短语识别、相关性计算等关键技术进行研究。本文首先将专利与企业相关性概念定义为专利对企业主要产品、主营业务、主要服务在技术上的相关性，然后基于专利文本领域性强、规范性的特点，构建标引词集合对专利文本进行自动标引。在专利文本与企业描述文本的相关性计算上，本文并未采用传统文本相似性计算方法，而是根据专利文本特点，构建企业描述文本集合，在专利标引词抽取的基础上，面向企业描述文本进行专利标引词检索，根据标引词的检索命中情况及权重计算策略判断专利与企业相关性。与人工标注

结果对比的实验结果验证了本文方法的可行性。

本文的研究工作同时也存在一些不足之处。一是本文采用的基于词向量的短语过滤策略的准确性有待改进，需要加入更多过滤方法提高短语识别结果的准确性；二是由于本文采用的专利语料仅使用了标题和摘要字段，长度不足会造成专利标引词覆盖范围有限，需采用合理的方式扩充专利的标引词范围，从而增大专利标引词与企业文本的命中程度，提高专利与企业相关性计算的准确率；三是在判断相关性计算结果的正确性时，人工标注可能会因为标注者的主观理解不同而存在判断偏差，继而给实验结果评价带来误差。

参 考 文 献

[1] Griliches Z. Patent statistics as economic indicators: A survey[J]. Journal of Economic Literature, 1990, 28(4): 1661-1707.

[2] Ernst H. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level[J]. Research Policy, 2001, 30(1): 143-157.

[3] Reitzig M. Strategic management of intellectual property[J]. MIT Sloan Management Review, 2004, 45(3): 35-40.

[4] Anandarajan A, Chin C L, Chi H Y, et al. The effect of innovative activity on firm performance: The experience of Taiwan[J]. Advances in Accounting Incorporating Advances in International Accounting, 2007, 23: 1-30.

[5] 唐恒, 金志成. 我国医药制造上市公司专利水平与公司业绩协整分析[J]. 科技管理研究, 2012, 32(8): 180-183, 188.

[6] 赵远亮, 周寄中, 侯亮, 等. 医药企业知识产权与经营绩效的关联性研究[J]. 科研管理, 2009, 30(4): 175-183.

[7] 苑泽明, 严鸿雁, 吕素敏. 中国高新技术企业专利权对未来经营绩效影响的实证研究[J]. 科学学与科学技术管理, 2010, 31(6): 166-170.

[8] Chang K C, Chen D Z, Huang M H. The relationships between the patent performance and corporation performance[J]. Journal of Informetrics, 2012, 6(1): 131-139.

[9] Chen Y S, Chang K C. The relationship between a firm's patent quality and its market value—The case of US pharmaceutical industry[J]. Technological Forecasting and Social Change, 2010, 77(1): 20-33.

[10] Artz K W, Norman P M, Hatfield D E, et al. A longitudinal study of the impact of R&D,patents,and product innovation on firm performance[J]. Journal of Product Innovation Management, 2010, 27(5): 725-740.

[11] 李睿, 张谙宁. 论专利法视域下专利引文的情报学意义[J]. 情报工程, 2016, 2(5): 8-17.

[12] 李忆, 马莉, 苑贤德. 企业专利数量、知识离散度与绩效的关系——基于高科技上市公司的实证研究[J]. 情报杂志, 2014, 33(2): 194-200.

[13] 康婧, 谢怡, 宋佳颖, 等. 专利信息系统分析与研究[J]. 情报工程, 2017, 3(5): 112-123.

（责任编辑 王克平）