

● 彭继东^{1,2}, 谭宗颖²

(1. 中国科学院 研究生院, 北京 100049; 2. 中国科学院 文献情报中心, 北京 100190)

一种基于文本挖掘的专利相似度测量方法及其应用

摘 要: 介绍了专利相似度测量的现状及其局限性; 在此基础上提出了基于文本挖掘技术, 以专利标题、摘要、权利要求和说明书 4 个文本元素的加权相似度作为专利相似度的测量方法; 最后讨论了该方法的应用价值, 并以碳纳米管技术领域为例作了实证研究。

关键词: 专利; 文本挖掘; 测量方法; 应用

Abstract: The present situation and limitations of patent similarity measurement have been discussed. On this basis, the paper proposes a text mining-based measurement method which takes the weighted similarities of 4 text elements, that is, patent title, abstract, claims and description, as patent similarity. Finally, the paper discusses the application value of the method, and makes an empirical study of the technologies of carbon nanotube.

Keywords: patent; text mining; method of measurement; application

1 专利相似度的测量

专利数量巨大, 内容广泛。2007 年世界专利申请量和授权量分别为 185.4 万件和 76.5 万件^[1], 几乎涵盖人类生产活动的全部技术领域。专利信息具有很高的技术、情报、经济和法律价值, 在很多领域有重要的应用价值, 可用于获取竞争情报^[2]、制定技术路线图^[3]、评估专利侵权的风险^[4]、识别技术机会^[5], 等等。而这些应用都离不开对专利间相似度的测量。

专利相似度是指专利间技术内容上的相似性, 测量方法有多种, 大体可以分为两类, 一是专利引文分析, 二是专利内容分析。

1.1 利用专利引文分析测量专利相似度

在文献计量学中, 用引文分析法分析文献的相似性已被深入研究。1963 年, Kessler 就提出了文献耦合 (Bibliographic Coupling)^[6], 1973 年, Small 提出了共引分析 (Co-citation Analysis)^[7]。

专利引文具有文献引文的一般性特点, 可以通过引文分析来测量专利的相似度。研究人员利用专利引文分析法从不同角度对专利相似度进行了研究。Stuart 用专利的共引关系测量日本 10 家半导体企业的技术相似度^[8]。Lai 等^[9]构造了一个专利分类系统, 其核心就是用共引分析法来测量专利的相似度。McGill 和 Mowery 等人在分析专利联盟内企业间的关系时, 采用互引率 (Cross-citation Rate) 测量企业的专利相似度^[10-11]。

用引文分析法来测量专利的相似度存在一些缺陷。首

先, 引用仅能表明两件专利间的关系, 不能反映所有专利间的关系^[12]。其次, 引文分析忽略了专利文本信息中的技术、法律等信息。第三, 专利引用包括审查员引用和申请人引用, 二者在引用动机、引用对象和引用方式等方面都存在根本区别^[13]。最后, 引文分析很容易受数据源限制, 如中国国家知识产权局专利数据库就不具有引文著录项。

1.2 利用专利内容分析测度专利相似度

专利内容分析是从专利文本信息的角度来分析专利内容上的相似性。有专利语义分析法 (Semantic Patent Analysis)^[4,14]、发明功能树 (Invention Functional Tree)^[15]、文本挖掘^[12,17-18]等方法。这些方法都具有优势又具有局限性。

专利语义分析法是指利用自然语言处理技术解析出专利的主谓宾 (SAO) 结构, 然后分析专利间 SAO 结构的相似性来确定专利的相似度^[4,14]。语义抽取的准确性会受语义结构具有的复杂性与多变性的影响。

Cascini 等提出了一种改进的方法, 即构造专利的发明功能树, 通过比较该树中组件以及组件的功能和层次关系来确定专利的相似度^[15]。该方法计算出的相似度反映的是专利概念上的相似性, 而非内容上的相似性。

文本挖掘技术出现之后, 文本相似度计算技术得到发展。专利的摘要、权利要求等都是典型的非结构化信息, 可以利用文本挖掘技术来测量专利的相似度。Magerman 等验证了文本挖掘技术测量专利相似度的可能性和准确性^[16]。刘玉琴等人提出基于文本挖掘技术计算得到的文

档相似度可用来定义专利相似度^[17]。陈苕熙等以专利名称和摘要的加权相似度作为专利分类的依据^[18]。Yoon 等利用文本挖掘技术对专利文献进行预处理,构建出专利的关键词向量,然后通过计算向量间的欧式距离得到专利相似度^[12]。

利用文本挖掘技术测量专利的相似度较之其他方法具有准确性高^[19]、处理数据量大、操作简单等优势。笔者认为,目前在利用文本挖掘技术测量专利相似度的研究中存在的普遍问题是忽略了专利文本信息具有的技术特性和法律特性。

为此,针对专利文本信息的这些特性,提出一种基于文本挖掘技术,以专利标题 (Title)、摘要 (Abstract)、权利要求 (Claims) 和说明书 (Description) 4 个文本元素 (简称专利四文本元素) 的加权相似度作为专利相似度的方法。

2 基于专利四文本元素加权的专利相似度测量

文本挖掘是从大量文本数据中识别出隐含的模式,抽取有价值的信息的方法与技术^[20],包括信息抽取、文本相似度计算、自然语言处理、文本分类、文本聚类、自动摘要等。

专利文本信息跟一般的文本信息有所不同,主要体现在两个方面:

- 1) 专利是技术创新的结果,具有技术特性。
- 2) 专利具有法律特性,书写受法律规范约束,导致专利四文本元素所包含的信息内容具有差异性。具体来说,专利标题和摘要高度概括了专利的技术内容,权利要求表达了请求保护的范围;说明书清楚、完整地描述了专利技术背景、技术特征、技术细节等内容。

在测量专利相似度时不仅要考虑专利在技术特征上的相似性,还要考虑专利四文本元素在信息内容上的差异性。为此,笔者提出把每件专利的文本分割成 4 个相对独立的部分,并用一个四元组来表示,即 $p = (t, a, c, d)$ 。元素 t 、 a 、 c 、 d 分别代表专利 p 的标题、摘要、权利要求和说明书。然后利用文本挖掘技术,分别计算出专利四文本元素的相似度,最后采用加权的方法,得到专利间的相似度。

具体过程如图 1 所示。

- 1) 专利数据获取。世界上主要国家 (如美国、日本和中国) 和专利组织 (如世界知识产权组织、欧洲专利组织) 都构建了专利网络数据库,可以公开获取。在专利相似度测量中应根据实际需求选择合适的专利数据库,采用一定的技术来获取专利数据集。

- 2) 数据预处理。文本抽取是指从专利文献中抽取

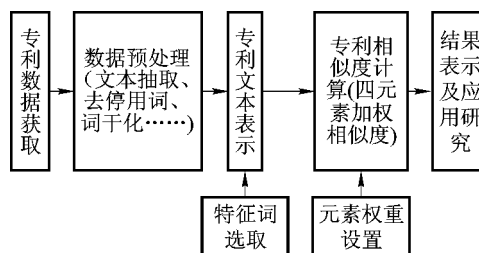


图1 基于文本挖掘技术以专利四文本元素加权的专利相似度测量方法流程

专利四文本元素,并把专利表示成四元组的过程。去停用词是去除出现频率太高、没有太大区别意义的词。本文选择康乃尔大学 Smart 系统的停用词表^[21],该表包括 429 个停用词。词干化是去除词缀只留词干的过程,其处理对象是英语中的名词复数、动词的多种时态变换以及形容词的比较级和最高级。本文采用 Porter^[22]的词干化算法。

3) 专利文本表示。专利文本是非结构化的数据,文本表示就是要将非结构数据结构化,以一定的特征词来表示文本信息。大量研究表明向量空间模型 (Vector Space Model, VSM)^[23]是一种适合于大规模语料的文本表示模型。

设 P 是 n 件专利的集合, $P = \{p_1, \dots, p_i, \dots, p_n\}$, $i = 1, 2, \dots, n$ 。经过过程 2), 每件专利都被表示成一个四元组,即 $p_i = (p_i^t, p_i^a, p_i^c, p_i^d)$, p_i^t 、 p_i^a 、 p_i^c 、 p_i^d 分别表示专利 p_i 的标题、摘要、权利要求和说明书。这 4 个元素分别用 VSM 来表示,以标题元素 p_i^t 为例。

$p_i^t = (w_{i1}^t, \dots, w_{ij}^t, \dots, w_{im}^t)$, $i = 1, 2, \dots, n, j = 1, 2, \dots, m$ 。 w_{ij}^t 表示 p_i^t 的特征词 T_j 的权重, m 表示选取的特征词的个数。特征词的权重采用归一化的 TF-IDF 算法^[18],如公式 (1) 所示。

$$w_{ij}^t = \frac{\log(1 + ft_{ij}^t) \times \log(n/n_j^t + 0.01)}{\sqrt{\sum_{j=1}^m [\log(1 + tf_{ij}^t) \times \log(n/n_j^t + 0.01)]^2}} \quad (1)$$

其中 tf_{ij}^t 表示 p_i^t 的特征词 T_j 的词频, n 表示专利集合 P 中的专利件数, n_j^t 表示专利标题中出现特征词 T_j 的专利数量。

如何选取特征词是专利文本表示中的关键问题之一。依据需不需要类别信息,特征词选择可分为有监督和无监督两类^[24]。专利采用《国际专利分类表》(International Patent Classification, IPC) 对专利文献进行标识。因此可以用 IPC 分类号构建一个专利分类训练集,再选用某种有监督的特征词选择方法,如信息增益 (Information Gain),来确定特征词。

- 4) 基于专利四文本元素加权的专利相似度计算。本文以专利四文本元素的加权相似度作为专利间的相似度,

定义专利 p_i 和专利 p_j 的专利相似度为 $SIM(p_i, p_j)$, 计算方法如公式 (2) 所示:

$$SIM(p_i, p_j) = \sum w^s \sin^s(p_i, p_j) \quad (2)$$

其中 $s = \{t, a, c, d\}$, w^s , $\sin^s(p_i, p_j)$ 分别表示专利 p_i 和 p_j 的标题、摘要、权利要求和说明书的权重和相似度, $\sum w^s = 1$ 。 $\sin^s(p_i, p_j)$ 通过计算向量间的夹角余弦得到, 如公式 (3) 所示。 $SIM(p_i, p_j)$ 的取值范围是 $[0, 1]$, 值越大, 表示专利越相似。

$$\begin{aligned} \sin^s(p_i, p_j) &= \cos^s(p_i, p_j) \\ &= \frac{\sum_{k=1}^m w_{ik}^s \times w_{jk}^s}{\sqrt{\sum_{k=1}^m (w_{ik}^s)^2} \times \sqrt{\sum_{k=1}^m (w_{jk}^s)^2}} \quad (3) \end{aligned}$$

权重 w^s 的设置方法有两种, 即经验法和实验法。经验法就是借助技术领域专家、专利情报人员、专利审查员等专业人士的知识和经验来确定专利各元素的权重。如应用于评估专利侵权的风险时, 就应当加大权利要求元素的权重, 应用于技术相似性分析时, 就应当突出专利的技术特性, 赋予说明书元素较大的权重。

实验法要求事先建立一个标准的训练数据集, 然后在实验中不断调整各元素的权重, 直到实验结果与训练集中的结果拟合度达到一定阈值时, 就可以把该权重确定为最终的权重。

5) 结果表示。经过上面的过程, 得到两两专利间的相似度。可以用 n 阶矩阵 M 表示专利间的相似关系, 并称 M 为专利相似度矩阵, 矩阵中的元素 a_{ij} 表示专利 p_i 和 p_j 的专利相似度, 显然满足 $a_{ij} \in [0, 1]$, $a_{ij} = a_{ji}$, $a_{ii} = 1$, $i = 1, 2, \dots, n, j = 1, 2, \dots, n$ 。

在专利相似度矩阵的基础上, 本文提出了相似专利和相似专利矩阵的概念。

定义 1 相似专利: 有专利 p_i 和 p_j , 当两件专利的相似度 a_{ij} 大于或等于某一阈值 α 时, 称专利 p_i 和 p_j 是相似专利, 并且称 α 为相似强度。

相似专利矩阵是一个 0-1 矩阵, 记为 M' 。 M' 中每个元素 b_{ij} 的取值如公式 (4) 所示。 b_{ij} 取 1 时, 表示专利 p_i 和 p_j 为相似专利, 取 0 时, 表示不是相似专利。

$$b_{ij} = \begin{cases} 1 & a_{ij} \geq \alpha \\ 0 & a_{ij} < \alpha \end{cases} \quad (4)$$

相似专利的概念可以扩展到专利与专利集, 假设有专利 p_i 和专利集 P , 对于 P 中的专利, 只要有一件专利与 p_i 为相似专利, 专利 p_i 就是专利集 P 的相似专利。

在实际应用中, 可能会关注专利集合之间的相似性。假设一个专利集合代表一种技术, 称专利集合间的相似度

为技术相似度, 定义如下:

定义 2 技术相似度: 有专利集合 P_A 和 P_B , 其专利数量分别为 m 和 n , 定义 P_A 和 P_B 的技术相似度为:

$$\sin(P_A, P_B) = \frac{c_{AB} + c_{BA}}{m + n} \quad (5)$$

$\sin(P_A, P_B) \in [0, 1]$, c_{AB} 表示 P_A 中是 P_B 的相似专利的专利数量, c_{BA} 表示 P_B 中是 P_A 的相似专利的专利数量。

3 应用研究

本节归纳总结了专利相似度的测量在多个方面的应用价值; 应用第 2 节提出的基于专利文本四元素加权的专利相似度测量方法进行实证研究。

3.1 专利相似度的测量的应用价值

1) 应用于评估专利侵权的风险。全球每年都会发生大量的专利侵权案件, 以我国为例, 2000—2007 年间, 我国受理的专利侵权纠纷案件共计 9 412 起^[25]。对一个企业而言, 一旦构成侵权, 就有可能面临着巨额的经济赔偿和信誉损失。所以企业在申请专利前, 必须认真地评估专利侵权的可能性。利用专利相似度的测量方法可以帮助专利申请人快速圈定那些相似度很高的授权专利, 进而确定是否专利侵权, 以降低专利侵权的风险。

2) 应用于专利信息检索与分析。可用于专利信息检索, 帮助检索人员寻找相似专利, 弥补关键词检索和分类检索的不足; 可用于专利自动分类, 减少了专利人工分类的工作量和分类错误; 可用于制作专利地图, 提高专利信息分析的可视性和可理解性。

3) 应用于企业技术竞争情报和企业创新管理。可用以衡量专利的价值和新颖度^[17]; 可以帮助企业分析企业间的技术相似性, 鉴定和跟踪其潜在的技术竞争对手, 寻找潜在的并购目标, 监测技术的变化趋势。

3.2 实证研究

本研究选取美国专利商标局 (USPTO) 专利网络数据库作为数据源, 以碳纳米管技术领域的授权发明专利 (以下简称碳纳米管专利) 为例, 利用专利相似度测量方法进行实证研究。

碳纳米管具有奇异的物理化学性能, 有着广阔的市场前景。美国 BCC 调查公司在报告中指出, 2009 年全球碳纳米管市场规模为 1.03 亿美元, 预计 2014 年将达到 10 亿美元^[26]。世界各国均投入大量的研发力量, 期望能占领该技术领域的制高点。

利用 USPTO 专利网络数据库, 在专利标题和摘要字段中检索 “(carbon AND nano AND tube \$) OR (carbon AND nanotube \$)”, 该检索式并没囊括到所有碳纳米管

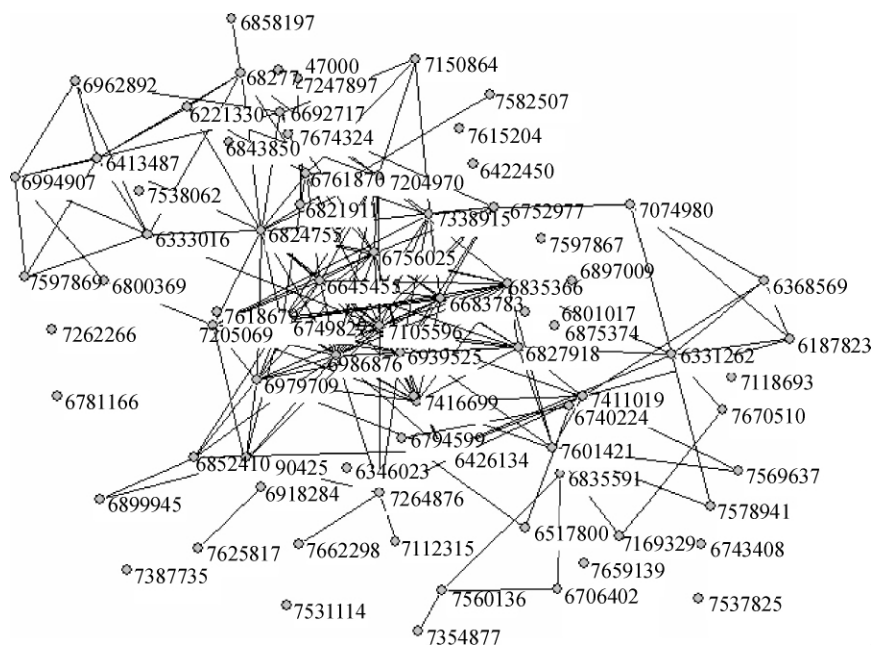


图2 相似强度 $\alpha = 0.21$ 时碳纳米管技术领域
的部分授权专利的专利相似性网络图

专利,比如还有 CNT, SWNT, BUCKYTUBE 等。笔者在平衡考虑检全率和检准率后,选择了构建这个简单的检索式。数据截至 2010 年 3 月 1 日,共检索到 1 286 件碳纳米管专利。然后利用自行开发的专利分析平台下载这些专利并解析出每件专利的专利号、专利权人、标题和摘要等。按照第 2 节的流程,生成专利相似度矩阵。标题、摘要、权利要求、说明书四元素的权重分别设置为 0.20、0.22、0.28 和 0.30。在第 2 节中给出了权重设置的两种思路,受时间限制,本实证采用经验法,突出专利的技术特性,赋予说明书最大权重。在实证中反复试验了取不同权重时的分析结果,发现权重的设置对结果影响较大。所以,最好建立一个标准的训练集,通过实验法来设定权重,这也是下一步的研究重点。

1) 相似专利分析。在相似强度 $\alpha = 0.21$ 时,相似强度 α 取值越大,每件专利的相似专利就越少。通过调节 α 的大小,得到不同相似强度下的专利相似矩阵,进而得到不同的分析结果。本实证中 α 取 0.21 主要是考虑专利相似性网络图(见图 2)在该值下分布比较均匀。另外,实验还发现 α 取值对技术相似性网络图(见图 3)的影响不大。生成碳纳米管专利的相似专利矩阵。利用社会化网络软件 Pajek 可视化展示专利间的相似关系。图 2 展示了碳纳米管技术领域的部分专利(含有 977/750 分类号)的专利相似性关系,从图 2 中可以看到与专利 7105596 相似的专利最多,该专利是关于单壁碳纳米管复合材料的制造方法。7262266、6781166 等专利没有相似专利。

图 2 中每个节点代表一件专利,节点的标签是该专利

的专利号。如果两节点间有连线,表示这两件专利为相似专利。

2) 技术相似性分析。在技术竞争情报分析中,需重点监测某技术领域的研发单位之间的技术相似性,进一步分析它们间的竞争与合作关系。技术相似度分析基于一个假设,即两研发单位的专利越相似,技术就越相似。依照公式(5),可得到碳纳米管技术领域的主要研发单位间的技术相似度,如图 3 所示。从图 3 可以看出,三星、莱斯大学、鸿海精密等研发单位的专利数量居前,是碳纳米管技术领域的研发主体。清华大学与鸿海精密的技术相似度最高,究其原因是在大陆与富士康(鸿海精密在大陆的分公司)合建了纳米科技研究中心,具有合作关系。Intel 和 IBM 两家企业的技术相似度也很高,说明这两家企业在碳纳米管技术领域的技术布局比较相似,形成竞争关系。

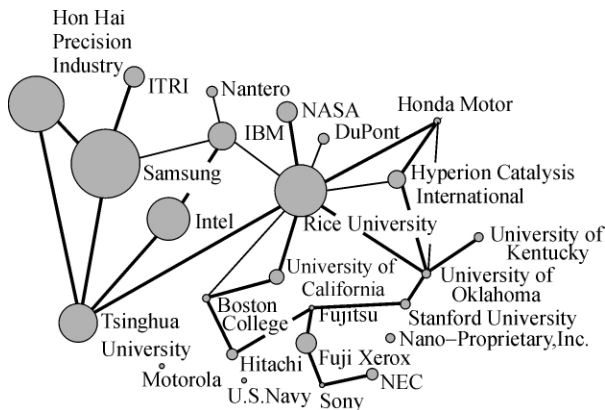


图3 相似强度 $\alpha = 0.21$ 时碳纳米管技术领域
授权专利量居于前 20 的研发单位的技术相似性网络图

图 3 中每个节点代表一个研发单位,节点的大小代表该研发单位的碳纳米管专利的数量,节点越大,专利越多;边的粗细代表技术相似度的大小,边越粗,表示技术越相似。图 3 中只显示了相似度大于 0.45 的边。

4 结束语

本文介绍了几种专利相似性的测量方法,分析了这些方法的优劣,认为文本挖掘技术是一种比较适合的方法。同时指出,专利文本与普通文本相比,具有法律特性和技术特性,测量专利的相似度也应该考虑这一点。为此提出

了以专利的4个文本元素的加权相似度作为专利相似度测量的方法,并设计了该方法的实施操作流程,最后利用本文提出的方法在碳纳米管技术领域做了实证研究,表明了该方法在实际应用中是比较有效的。

研究中的一个关键点是四元素权重的设定,在实证中采用的是经验法。下一步的研究重点将针对某技术领域建立一个标准的训练集,然后采用实验法来确定权重。另外,专利包括结构化和非结构化信息,本文提出的专利相似度测量方法完全是基于非结构化信息,如何融入结构化的信息及其有效性也是下一步值得研究的问题。总之,该方法还需要在实践中不断地验证、改进和完善。□

参考文献

- [1] World Intellectual Property Organization. World intellectual property indicators 2009 [EB/OL]. [2010-04-25]. <http://www.wipo.int/ipstats/en/statistics/patents/index.html>.
- [2] SHIH M J, LIU D R, HSU M L. Discovering competitive intelligence by mining changes in patent trends [J]. Expert Systems with Applications, 2010, 37 (4): 2882-2890.
- [3] LEE S, YOON B, LEE C, et al. Business planning based on technological capabilities: patent analysis for technology-driven roadmapping [J]. Technological Forecasting & Social Change, 2009, 76 (6): 769-786.
- [4] BERGMANN I, BUTZKE D, WALTER L, et al. Evaluating the risk of patent infringement by means of semantic patent analysis: the case of DNA chips [J]. R&D Management, 2008, 38 (5): 550-562.
- [5] YOON B, PARK Y. A systematic approach for identifying technology opportunities: keyword-based morphology analysis [J]. Technological Forecasting and Social Change, 2005, 72 (2): 145-160.
- [6] KESSLER M M. An experimental study of bibliographic coupling between technical papers [J]. IEEE Transaction on Information Theory, 1963, 9 (1): 49.
- [7] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents [J]. Journal of American Society for Information Science, 1973, 24 (4): 265-269.
- [8] STUART T B, PODOLY J M. Local search and the evolution of technological capabilities [J]. Strategic Management Journal, 1996, 17 (2): 21-28.
- [9] LAI K K, WU S J. Using the patent co-citation approach to establish a new patent classification system [J]. Information Processing and Management, 2005, 41 (2): 313-330.
- [10] McGill J P. Technological knowledge and governance in alliances among competitors [J]. International Journal of Technology Management, 2007, 38 (1-2): 69-89.
- [11] MOWERY D C, OXLEY J E, SILVERMAN, B S. Technological overlap and interfirm cooperation: implications for the resource-based view of the firm [J]. Research Policy, 1998, 27 (5): 507-523.
- [12] YOON B U, PARK Y T. A text-mining-based patent network: analytical tool for high-technology trend [J]. Journal of High Technology Management Research, 2004, (15): 37-50.
- [13] 李睿, 孟连生. 论专利间引用关系分析中存在的问题 [J]. 情报理论与实践, 2009, 32 (7).
- [14] MOEHRLE M G, WALTER L, GERITZ A, et al. Patent-based inventor profiles as a basis for human resource decisions in research and development [J]. R&D Management, 2005, 35 (5): 513-524.
- [15] CASCINI G, ZINI M. Measuring patent similarity by comparing inventions functional trees [C] // IFIP International Federation for Information Processing: Computer-aided Innovation (CAI). Boston: Springer, 2008: 31-42.
- [16] MAGERMAN T, VAN LOOY B, SONG X. Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications [J]. Scientometrics, 2010, 82 (2): 289-306.
- [17] 刘玉琴, 朱东华, 吕琳. 基于文本挖掘技术的产品技术成熟度预测 [J]. 计算机集成制造系统, 2008, 14 (3): 506-510.
- [18] 陈苡熙, 顾新建, 陈国海, 等. 基于向量空间模型和专利文献特征的相似专利确定方法 [J]. 浙江大学学报: 工学版, 2009, 43 (10): 1848-1852.
- [19] AHLGREN P, COLLIANDER C. Document-document similarity approaches and science mapping: experimental comparison of five approaches [J]. Journal of Informetrics, 2009, 3 (1): 49-63.
- [20] FELDMAN R, SANGER J. The text mining handbook: advanced approaches in analyzing unstructured data [M]. London: Cambridge University Press, 2007: 1-48.
- [21] Stop word list 1 [EB/OL]. [2010-05-19]. <http://www.lextek.com/manuals/onix/stopwords1.html>.
- [22] PORTER M F. An algorithm for suffix stripping [J]. Program, 1980, 14 (3): 130-137.
- [23] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18 (11): 613-620.
- [24] 郭庆琳, 李艳梅, 唐琦. 基于VSM的文本相似度计算的研究 [J]. 计算机应用研究, 2008, 25 (11): 3256-3258.
- [25] 国家知识产权局. 国家知识产权局统计年报 [EB/OL]. [2010-05-10]. <http://www.sipo.gov.cn/sipo2008/tjxx>.
- [26] BCC Research. Carbon nanotubes: technologies and global markets [EB/OL]. [2010-05-10]. <http://www.bccresearch.com/report/NAN024D.html>.

作者简介: 彭继东, 男, 1984年生, 硕士生。

谭宗颖, 女, 1957年生, 研究员, 博士生导师。

收稿日期: 2010-07-26