doi:10.3772/j.issn.2095-915x.2018.02.007

利用 Doc2Vec 判断中文专利相似性

张海超[1] 赵良伟[2]

1. 中国科学技术信息研究所国家科技信息资源综合利用与公共服务中心 北京 100038
2. 邢台职业技术学院 邢台 054035

摘 要 目前专利侵权纠纷案件时有发生，企业一旦卷入专利侵权纠纷，通常会面临时间考验和经济损失。本文选取中文专利数据样本，抽取专利权利要求书形成训练语料，并利用 Doc2Vec 深度神经网络算法，计算权利要求书文本之间的相似度，得出与涉案专利相似性较高的专利。并且将上述方法应用到专利复审案件实验中，进行实证研究，取得了较好的效果。需要进一步提高训练数据的质量，对比其他算法的效果。利用该方法能够帮助专利审查人员和企业找到相似专利。

关键词： 专利相似度；专利侵权；Word2Vec；Doc2Vec
中图分类号： G305
开放科学（资源服务）标识码（OSID）

Judge Chinese Patents Similarity Based on Doc2Vec

ZHANG Haichao[1] ZHAO Liangwei[2]

1. National Engineering Research Center of Science and Technology Information, Institute of Scientific and Technical of Information of China, Beijing 100038, China;
2. Xingtai Polytechnic College, Xingtai 054035, China

Abstract Recently, patent infringement disputes occurred frequently. Once a company was involved in a patent infringement dispute, it usually faced the time test and economic loss. This paper chose the Chinese patents as data source, extracted the patent claims as the training corpus, and used the Doc2Vec deep neural network algorithm to calculate the similarity between the claims. Then, we obatined the patent with

基金项目： 基于文本内容相似性的中文专利侵权判定方法研究（YY2016-04）。

作者简介： 张海超（1989-），研究实习员，硕士，研究方向：专利信息挖掘，Email：zhanghc@istic.ac.cn；赵良伟（1989-），科员，硕士，研究方向：信息管理与数据分析，Email：zlwgongzuo@126.com。

higher similarity to the involved patent. Finally, the above methods was applied to the patent reexamination case to conduct the empirical research. The results indicated that this method can achieve good results. Moreover, the results also suggested that the method need to be further improved based on the high quality of training data and comparison with other algorithms. This method may help the patent reviewers and enterprises to find the similar patents.

Keywords：Similarity of Patents; patents Infringement; Word2Vec; Doc2Vec

1 背景

专利文献作为科技创新成果的重要载体和表现形式，内容新颖，蕴含了更前沿的科技信息，而企业一旦卷入专利侵权纠纷，通常会面临时间考验和经济损失。在专利侵权诉讼或者专利复审时，被告方通常采取无效请求的方式来应对。专利无效最重要的证据是在涉案专利申请之日前就已经有相同或相似的专利。本文从专利文本内容角度，利用Doc2Vec等深度神经网络算法，计算权利要求书文本之间的相似度，帮助专利审查人员和企业找到相似专利。

2 国内外研究现状

专利侵权判定的基本原则包括：全面覆盖原则、等同原则、禁止反悔原则、多余指定原则和实施自由公知技术不侵犯等原则[1]。除了从法律角度，从情报学、信息学角度来看，判定专利相似性、识别专利侵权的方法可以分成两类：基于引文的相似度计算方法和基于文本内容的相似度计算方法。

利用专利文献计量方法进行专利侵权研究，Sternitzke[2]主要进行专利引用分析，并且对两家公司的涉及诉讼案件的专利进行实证分析。Tijssen[3]基于专利与科技文献之间的互引，对二者之间的用专利与科学文献之间的相互引用，对科学和技术的关系以及其二者之间的知识流动进行了探讨。Lai等[4]提出通过同被引分析对高被引专利进行聚类的方法，并且构造了一个专利分类系统。McGill等[5]应用专利互引率测量企业的专利相似度。Cascini等[6]利用对比专利之间的发明功能树来进行专利相似度度量。Wu等[7]利用直接引用和间接引用来度量专利之间的相似度。Rodriguez等[8]利用两阶段相似度量方法对专利引文网络中的成对专利之间的相似度进行计算。彭爱东[9]利用专利文献同被引关系来计算专利相似度的方法，通过专利聚类等方式对公司及其竞争对手所拥有的专利进行分类管理。魏兵等[10]利用改进的模糊聚类方法实现对专利相关主题聚类。洪勇等[11]提出了专利互引、专利他引和专利自引等三种专利引用方式来比较企业间技术的相关性，并且分析出公路工程领域的技术溢出。

基于文本内容的专利相似度研究是从专利文本信息的角度来分析专利内容上的相似性,主要是基于统计学的相似度计算方法和基于语义的相似度计算方法。

doi:10.3772/j.issn.2095-915x.2018.02.007

目前最常用的方法是利用向量空间模型（VSM）来计算专利相似度[12]。其基本思想是假设词语词之间是不相关的，一般将专利的名称、摘要、权利要求书和说明书等文本信息转化为词向量模型。Bergmann[13]提出了利用句子语义结构计算专利相似度，识别专利侵权风险，并且以DNA芯片技术专利侵权案件为例，证实此方法的可行性。Lee等[14]运用基于词频的空间向量计算专利相似度，从而对专利进行分类，分析了PDA领域的技术空白点。Taghaboni-Dutta等[15]采用TF-IDF与本体相结合的方法构建基于关键词的向量，利用余弦相似度算法计算出专利之间的相似度，最后进行聚类分析出该RFID领域企业间的相关技术竞争形势。

国外对专利文本的表示已经不再限于关键词。Moehrle等[16]利用专利中的“组合概念”即方法、过程以及导致的结果组合起来进行文本内容的分析专利之间的相似度。Tseng等[17]指出在专利分析中，单个概念会很多很复杂，因此不能准确表达专利的内容，从而引入了SAO结构（“主语-谓语-宾语”结构）。Cascini等[18]指出SAO结构能够清晰表征专利成分之间的关系。其中主语和宾语表示专利内容的组成成分，通过谓语的连接能够表达作用效果。Magerman等[19]基于潜在语义分析对专利文本进行文本挖掘，以检测专利文献和科学出版物之间的相似性。Yoon等[20]在SAO语义相似度的基础上进行专利相似度计算，并构建专利地图。Indukuri等[21]利用自然语言处理的语法分析和语义分析对专利相似度计算并且将其工具化。McNamee等[22]利用专利之间的相似度和距离对专利进行分级分类，并且以美国专利分类体系作为文章的实证研究。

本文研究着眼于中文专利侵权，为侵权判定从文本内容挖掘方面提出了一种新的尝试，提出一种基于文本内容的中文专利相似性判定方法。本文不同于传统的向量空间模型，通过选取中文专利数据样本，抽取专利权利要求书形成训练语料，并利用Doc2Vec深度神经网络算法，计算权利要求书文本之间的相似度，得出与涉案专利相似性较高的专利。力求从文本挖掘的角度进行专利侵权研究，为陷入其中的企业提供参考，更为专利复审委员会提供相似专利参考。

3 基于 Doc2Vec 计算中文专利相似度

本文的整体思路如下图所示：

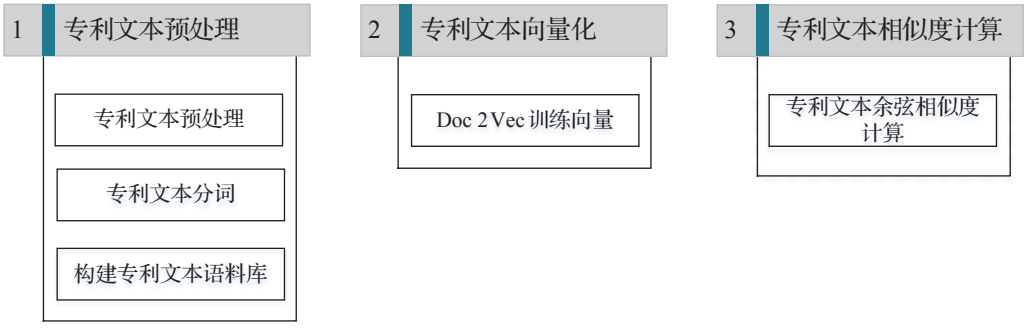

图 1 论文整体思路

3.1 专利文本预处理和构建语料库

在做专利文本向量化工作之前，首先要对中文专利文本进行预处理。主要包括分词、去除停用词和去除特殊符号等预处理工作。目前中文分词的算法大致包括三种：基本词典的分词方法、基于统计的分词方法和基于组合的分词方法。针对中文专利文本，很多学者也提出一些改进的方法，本文选用jieba中文分词组件[23]对中文专利文本进行分词处理。

在专利文本预处理之后，需要构建文本语料库。本文选择专利数据，针对具体领域，检索专利数据，构建专利文本语料库。

3.2 专利文本向量化

词的向量化，简言之就是将自然语言中的词语进行数学化，把一个词语表示成一个向量，常见的词向量化方式主要有三种：

（1）one-hot representation方式

one-hot representation方式是最简单的方式，用一个很长的向量（长度通常为词典的长度）来表示一个词。向量的分量只有一个1，其余全部是0。例如："番茄"表示为[0 0 0 0 0 1 0 0 0 …]，"西红柿"则表示为 [0 0 1 0 0 0 0 0 0 …]。这种方式虽然简单明了，但确不能有效的表征语义信息。

（2）Distributed representation（词向量）

这种方式最早由Hinton提出，是将词映射到一个低维、稠密的实数向量空间，这样词义越近在空间的距离也就越近。这种方式比第一种更合适比较词向量的相似度。

（3）word2vec模型

word2vec[24]是Mikolov等在Bengio的NNLM（Neural Network Language Model）[25]模型和Hinton的Log_Linear模型[26]的基础上提出的语言模型。word2vec模型能够根据给定的语料库，通过优化后的训练模型快速有效的将一个词语表达成向量形式，其核心架构包括CBOW模型和Skip-gram模型，如图2所示。NNLM模型是神经网络概率语言模型的基础模型，与其相比，CBOW模型和Skip-gram模型去掉了隐含层。有实践证明，word2vec训练产生的词向量的精确度可能不如NNLM模型，但通过增加训练语料方法来提高。

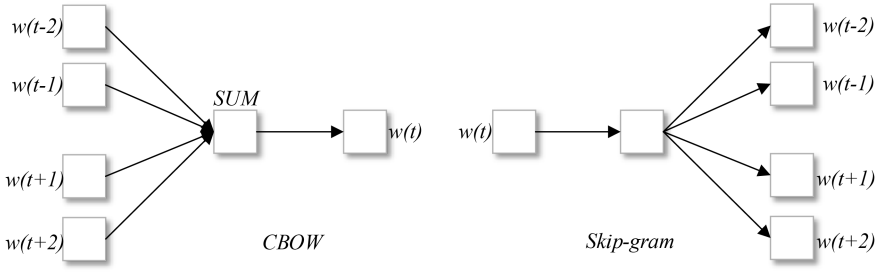

图 2 CBOW 模型（左图）和 Skip–gram 模型（右图）

其中CBOW模型能够计算出上下文决定当前词$w(t)$出现的概率，并且上下文所有的词对当前词$w(t)$出现概率的影响权重是一样的。而Skip-gram模型正好相反，它是利用词$w(t)$去预

doi:10.3772/j.issn.2095-915x.2018.02.007

测$w(t)$前后各个词。

（4）Doc2Vec模型

Doc2Vec模型[24]与word2vec模型类似，增加了一个段落向量，可作为处理段落可变长度文本的模型。和word2vec一样，该模型也有两种方法：Distributed Memory（DM）和Distributed Bag of Words（DBOW）。DM 试图在给定上下文和段落向量的情况下预测单词的概率。在一个句子或者文档的训练过程中，段落 ID 保持不变，共享着同一个段落向量。DBOW 则在仅给定段落向量的情况下预测段落中一组随机单词的概率。

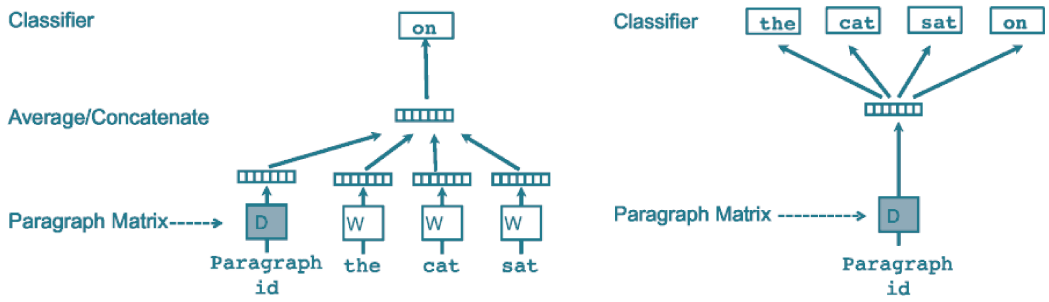

图 3 DM 模型（左图）和 DBOW 模型（右图）

3.3 专利文本相似度计算

利用Doc2Vec模型训练专利文本的向量之后，需要进行专利文本相似度的计算，本文采用余弦相似度计算公式，如下所示：

$$\mathrm{sim}(A,B)=\cos(\theta)=\frac{A\cdot B}{\|A\|\|B\|} \quad (1)$$

通过专利文本相似度计算，我们可以找到若干相似度较高的专利。然后将相似度较高的专利进行技术特征分解，找到可以作为审查或者无效的证据专利。

4 实验

4.1 案例背景

本文选取了中国国家知识产权局专利复审委员会2014年度十大重大案件中“‘具有位置可变的平衡重的移动式提升起重机’发明专利权无效宣告请求案”，案件详情如表1所示：

表 1 案件详情

案件编号	专利号	发明名称	专利权人	无效宣告请求人	审查结论
4W102283	ZL200810092407.6	具有位置可变的平衡重的移动式提升起重机	Manitowoc（马尼托瓦克起重机有限公司）	三一重工股份有限公司	全部无效

该案件的双方分别是美国和中国在重型机械领域的龙头企业，Manitowoc（马尼托瓦克公司）成立于1902 年，是一家全球性起重机企业。提供全方位的吊运解决方案，包括履带式、吊臂、伸缩臂或塔式起重机，同时还确立了起重机客户服务综合服务计划，提供专业化诊断和修理服务。三一重工股份有限公司由三一集团投资创建于1994年，目前是全球装备

制造业领先企业之一。产品包括混凝土机械、挖掘机械、起重机械、桩工机械、筑路机械、建筑装配式预制结构构件，其中泵车、拖泵、挖掘机、履带起重机、旋挖钻机、路面成套设备等主导产品已成为中国第一品牌。

马尼托瓦克公司2011年获得该专利的中国授权，并享有两项美国优先权。三一重工股份有限公司向专利复审委员提起无效宣告请求，最终经过专利复审委员会审理后作出无效决定，宣告其专利全部无效。

希望通过该实验，利用Doc2Vec计算中文专利权利要求书文本的相似度，力求为陷入其中的企业提供参考，更为专利复审委员会审查专利时提供相似专利参考。

4.2 数据来源

本实验以三一重工和马尼托瓦克之间的专利无效案件为背景，专利号为ZL200810092407.6（称为：涉案专利）。以该专利的IPC分类号B66C23/76和B66C23/36为检索依据，在中国专利信息中心的专利之星检索系统（patentstar）检索并筛选到中国专利687件，并抽取出专利的权利要求项。

本实验过程利用Python实现，并且调用了jieba和gensim等主要程序包。

4.3 专利权利要求书文本向量化

（1）预处理

首先对专利权利要求项进行了去掉一些特殊字符，例如“【 】、[]、/、•、”等。然后进行分词处理，本实验选用jieba中文分词组件进行分词，并且调用了哈工大停用词表去除相关停用词，停用词表如下所示：

图 4 停用词表

（2）权利要求书文本向量化

在专利文本预处理之后，针对检索到的687件专利，去掉涉案专利，将剩余的686件专利作为文本训练语料，构建专利文本语料库，最后利用gensim提供的Doc2Vec向量化模块进行文本向量化。

4.4 专利文本相似度计算

在对专利文本进行向量化之后，利用余弦相似度计算公式（公式一）进行相似度计算，本实验选择与涉案专利相似度在0.5以上的专利，如表2所示：

通过表2可以看出，相似度在0.5以上的专利中，有3件专利和涉案专利同属于马尼托瓦克起重机有限公司，并且申请号CN201010624732.X和涉案专利共同享有两项美国优先权。此外，该公司在2007年已经有另外两件相似专利在申请。

为了更好的解释哪些专利可以作为无效证据专利，我们根据优先权日按照时间序列进行排序（若无优先权，则按申请日进行排序），如图5所示：

doi:10.3772/j.issn.2095-915x.2018.02.007

表 2 与涉案专利相似度在 0.5 以上的专利列表

相似度	序号	申请号	申请日	优先权号	优先权日	申请人
涉案专利	P1	CN200810092407.6	2008.04.09	US11/733104; US12/023902	2007.04.09;2008.01.31	马尼托瓦克起重机有限公司
0.79	P2	CN200710192985.2	2007.10.26	US60/863265; US11/733104	2006.10.27;2007.04.09	马尼托瓦克起重机有限公司
0.71	P3	CN201210253579.3	2007.10.26	US60/863265; US11/733104	2006.10.27;2007.04.09	马尼托瓦克起重机有限责任公司
0.56	P4	CN201010624732.X	2008.04.09	US11/733104; US12/023902	2007.04.09;2008.01.31	马尼托瓦克起重机有限公司
0.55	P5	CN201220729203.0	2012.12.25	--	--	青岛滨海学院
0.53	P6	CN201310300559.1	2013.07.17	--	--	中联重科股份有限公司
0.52	P7	CN201210229336.6	2012.07.03	--	--	徐工集团工程机械股份有限公司
0.52	P8	CN201310653459.7	2013.12.05	--	--	长沙中联消防机械有限公司;中联重科股份有限公司
0.51	P9	CN200680041384.5	2006.10.12	DE102005055694.9; DE102006015307.3	2005.11.17; 2006.03.29	特雷克斯－德马格有限及两合公司

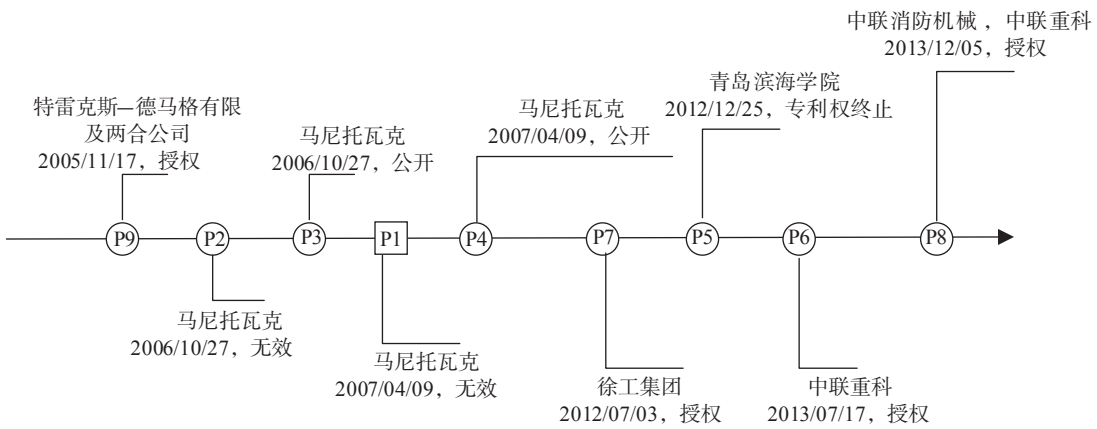

图 5 相似专利时间序列

从表2和图5综合可以看出，在涉案专利（P1）申请日之前共有3件专利（申请号分别为：CN200680041384.5（P9）、CN200710192985.2（P2）、CN201210253579.3（P3））。再结合法律状态，可以判定CN200680041384.5（P9）和CN201210253579.3（P3）可以作为涉案专利复审和无效请求的证据专利。

5 总结

专利文献作为科技创新成果的重要载体和表现形式，内容新颖，蕴含了更前沿的科技信息，而企业一旦卷入专利侵权纠纷，通常会面临时间考验和经济损失。从情报学、信息学角度来看，判定专利相似性、识别专利侵权的方法可以分成两类：基于引文的相似度计算方法

和基于文本内容的相似度计算方法。

本文提出一种基于文本内容的中文专利相似性判定方法，选取中文专利数据样本，抽取专利权利要求书形成训练语料，并利用Doc2Vec深度神经网络算法，计算权利要求书文本之间的相似度，得出与涉案专利相似性较高的专利。最后，我们将相似专利识别方法应用到具体案例，选取了中国国家知识产权局专利复审委员会2014年度十大重大案件中“‘具有位置可变的平衡重的移动式提升起重机’发明专利权无效宣告请求案”。经过一系列的处理和分析，我们发现实验结果很好的验证了本文提出的相似专利识别算法的有效性，而且指出了双方潜在专利侵权纠纷的技术点。

因此，本文的中文专利相似性判定方法为专利侵权判定从文本内容挖掘方面提出了一种新的尝试。并且不同于传统的向量空间模型，本文利用Doc2Vec深度神经网络算法，通过实验结果验证该方法的有效性。由于专利文本具有领域性强、专业术语多等特别，本文在文本处理，特别是分词的准确性、新词识别方面还有待于进一步的提高。

参考文献

[1] 程永顺. 专利侵权判定实务[M]. 北京: 法律出版社, 2009.

[2] Sternitzke C, Bartkowski A, Schramm R. Visualizing Patent Statistics by Means of Social Network Analysis Tools[J]. World Patent Information, 2008, 30(2):115-131.

[3] Tijssen R J W. Global and Domestic Utilization of Industrial Relevant Science: Patent Ctation Analysis of Science–technology Interactions and Knowledge Flows[J]. Research Policy, 2001, 30(1): 35-54.

[4] Lai K K, Wu S J. Using the Patent Co-citation Approach to Establish a New Patent Classification System[J]. Information Processing & Management, 2005, 41(2): 313-330.

[5] McGill J P. Technological Knowledge and Governance in Alliances among Competitors [J]. International Journal of Technology Management, 2007, 38(1-2):69-89.

[6] Cascini G, Zini M. Measuring Patent Similarity by Comparing Inventions Functional Trees[M]. Computer-aided Innovation (CAI). Springer US, 2008: 31-42.

[7] Wu H C, Chen H Y, Lee K Y, et al. A Method for Assessing Patent Similarity using Direct and Indirect Citation Links[C]. Industrial Engineering and Engineering Management (IEEM), 2010 IEEE International Conference on. IEEE, 2010: 149-152.

[8] Rodriguez A, Kim B, Turkoz M, et al. New Multi-stage Similarity Measure for Calculation of Pairwise Patent Similarity in a Patent Citation Network[J]. Scientometrics, 2015, 103(2): 565-581.

[9] 彭爱东. 基于同被引分析的专利分类方法及相关问题探讨[J].情报科学, 2008, 26(11):1676-1679.

[10] 魏兵, 李亚非. 基于同被引矩阵的专利引文分析方法[J]. 计算机工程与设计, 2012, 31(8):1779-1781.

[11] 洪勇, 康宇航. 基于专利引文的企业间技术溢出可视化研究[J]. 科研管理, 2012, 33(7):81-87.

[12] 曾文, 徐红姣, 李颖, 等. 基于VSM的科技期刊文献与专利文献的相似度计算方法研究[J]. 情报工程, 2016, 2(3):37-42.

[13] Bergmann I, Butzke D, Walrter L, et al. Erdmann. Evaluating the Risk of Patent Infringement by Means of Semantic Patent Analysis: The Case of DNA Chips [J]. R&D Management, 2008(38):550-562.

[14] Lee S, Yoon B, Park Y. Anapproach to Discovering New Technology Opportunities:

doi:10.3772/j.issn.2095-915x.2018.02.007

Keyword based Patent Map Approach [J]. Technovation, 2009, 29(6-7):481-497.

[15] Taghaboni - Dutta F, Trappey A J C, Trappey C V, et al. An Exploratory RFID Patent Analysis[J]. Management Research News, 2009, 32(32):1163-1176.

[16] Moehrle M G, Gerken J M. Measuring Textual Patent Similarity on the Basis of Combined Concepts: Design Decisions and their Consequences[J]. Scientometrics, 2012, 91(3): 805-826.

[17] Tseng Y H, Lin C J, Lin Y I. Text Mining Techniques for Patent Analysis[J]. Information Processing & Management, 2007, 43(5):1216-1247.

[18] Cascini G, Zini M. Measuring Patent Similarity by Comparing Inventions Functional Trees [J]. Computer-Aided Innovation (CAI), 2008(277):31-42.

[19] Magerman T, Looy B V, 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.

[20] Yoon J, Kim K. Generation of Patent Maps using SAO-based Semantic Patent Similarity[J]. Entrue Journal of Information Technology, 2011, 10(1): 19-27.

[21] Indukuri K V, Ambekar A A, Sureka A. Similarity Analysis of Patent Claims using Natural Language Processing Techniques[C]. Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on. IEEE, 2007(4): 169-175.

[22] McNamee R C. Can't see the Forest for the Leaves: Similarity and Distance Measures for Hierarchical Taxonomies with a Patent Classification Example[J]. Research Policy, 2013, 42(4): 855-873.

[23] fxsjy. Jieba中文分词组件[EB/OL]. (2016-08-05) [2016-11-20]. https://github.com/fxsjy/jieba,.

[24] Le Q V, Mikolov T. Distributed Representations of Sentences and Documents [J]. Eprint Arxiv, 2014(4):1188—1196.

[25] Bengio Y, Schwenk H, Senécal J S, et al. Neural Probabilistic Language Models[M]. Innovations in Machine Learning. Springer Berlin Heidelberg, 2006.

[26] Mnih A, Hinton G. Three New Graphical Models for Statistical Language Modelling[C]. Proceedings of the 24th International Conference on Machine Learning. ACM, 2007:641—648.