基于表示学习的无监督跨语言专利推荐研究*

张金柱[1,2]　主立鹏[1]　刘菁婕[1]

[1]（南京理工大学经济管理学院 南京 210094）

[2]（江苏省社会公共安全科技协同创新中心 南京 210094）

摘要：【**目的**】减少双语词典和大规模双语语料库的构建，提高专利文本语义的揭示和利用，从文本语义表示角度设计无监督的跨语言专利推荐方法，提高跨语言专利推荐效果和领域适用能力。【**方法**】首先设计无监督跨语言词向量映射方法，通过线性变换将独立的中英专利词向量映射到统一语义向量空间，构建中英词语间的语义映射关系；然后利用平滑倒词频的词向量加权方法，形成基于跨语言专利词向量的专利文本语义表示方法，实现中英专利文本在同一向量空间中的语义表示；最后应用向量相似度计算指标，计算不同语言专利文本间的语义相似度，构建基于表示学习的无监督跨语言专利推荐方法，实现跨语言专利推荐。【**结果**】在无线通信领域的实验中，无监督跨语言专利推荐方法的 Top-1 和 Top-5 推荐准确率分别达到 55.63% 和 77.82%，较弱监督跨语言专利推荐方法分别提高了 0.66% 和 1.45%，较基于机器翻译的跨语言专利推荐方法分别提高了 4.29% 和 3.90%。【**局限**】仅对特定领域中英专利进行推荐，尚需扩展领域和语言范围。【**结论**】能够实现有效的中英跨语言专利推荐，并可扩展应用到其他领域和语种下的专利推荐中。

关键词：跨语言　专利推荐　表示学习　语义表示

分类号：G254

DOI：10.11925/infotech.2096-3467.2020.0272

引用本文：张金柱，主立鹏，刘菁婕. 基于表示学习的无监督跨语言专利推荐研究[J]. 数据分析与知识发现，2020，4（10）：93-103.(Zhang Jinzhu, Zhu Lipeng, Liu Jingjie. Unsupervised Cross-Language Model for Patent Recommendation Based on Representation[J]. Data Analysis and Knowledge Discovery, 2020, 4(10): 93-103.)

1　引　言

专利数据是世界上最大的技术信息集之一，几乎囊括了一切应用领域内的技术成果和发展动态，在技术、商业、法律等领域具有举足轻重的地位。随着经济全球化和科技的飞速发展以及知识产权保护国际化意识的逐渐增强，专利冲突与专利壁垒深深困扰着广大国内企业与研发机构，跟踪和研究国外专利技术发展的需求与日俱增。因此，及时准确地获取世界其他国家的专利信息变得十分必要，快速、有效地获取其他语言相关专利的跨语言专利推荐研究得到广泛关注。跨语言专利推荐可以帮助企业和个人遴选相关重要专利，发现相关技术发展趋势，跟踪最新技术进展，进而提供个性化信息推送和决策支持服务。

当前，专利推荐主要基于单语言专利数据，从专利文献题目、摘要或权利要求中提取技术关键词和主题等内容特征，研究单语言环境下的相似专利推荐，而专门针对双语或多语种专利文献数据的跨语言专利推荐还较少。跨语言专利推荐一般是将跨语

通讯作者：张金柱，ORCID：0000-0001-7581-1850，E-mail：zhangjinzhu@njust.edu.cn。

*本文系国家自然科学基金项目“基于表示学习的专利信息语义融合与深度挖掘研究”（项目编号：71974095）、江苏省社会科学基金项目“基于社团结构动态演化的主题突变监测与形成机制研究”（项目编号：17TQC003）和国家自然科学基金项目“基于被引科学知识突变的突破性创新动态识别及其形成机理研究”（项目编号：71503125）的研究成果之一。

言检索中涉及的词翻译和机器翻译等技术直接应用于跨语言专利推荐中，主要包括三种方法，分别是基于双语词典或多语词典的查询翻译方法[1]、基于双语或多语种平行语料库对齐技术的查询文本翻译方法[2]以及基于机器翻译的查询文本和检索结果翻译方法[3-4]。这些方法多从查询词和文本精确翻译的角度出发，往往需要大规模特定领域的双语词典、双语语料库以及高效准确的机器翻译方法来实现有监督的跨语言查询扩展，导致这些方法应用扩展到其他领域进行跨语言专利推荐的难度较大，成本较高，推荐准确性也需进一步提高；与此同时，由此推荐的专利大多是相似专利，推荐的多样性和相关性尚需进一步扩展，亟需从专利文本语义角度出发进行专利推荐，提供更好的决策支持服务。

因此，本文设计无监督的跨语言词向量映射方法，使用中英专利单语语料库独立训练得到单语专利词向量，然后通过线性变换将它们映射到统一向量空间，既不需要任何外部双语词典，也不需要大型双语语料库，就可以得到较好的中英专利相关词映射关系。接着利用平滑倒词频的词向量加权方法，形成基于跨语言专利词向量的专利文本语义表示方法，实现中英专利文本在同一向量空间中的语义表示，进而计算不同语言专利文本间的语义相似度，实现文本语义角度下的无监督跨语言专利推荐。

2 国内外研究现状

2.1 跨语言专利推荐

当前专利推荐主要集中于单语言相似专利推荐。通过从专利文献题目、摘要或权利要求中提取的技术关键词、主题等内容特征进行表示，形成了基于知识的[5]、基于协同过滤的[6]和基于内容的[7]相似专利推荐方法。而专门针对跨语言专利推荐的研究相对较少，其一般思路是将跨语言检索中涉及的词翻译和机器翻译等技术直接应用于专利推荐中，类似于跨语言专利信息检索。常用的方法主要有三种：第一种方法是基于词典的跨语言专利推荐[1]。该方法从特定领域的专利数据中提取/更新双语或者多语专利词典，并从中选择合适的词来替换查询词，进而实现跨语言查询扩展，运用已有单语专利推荐技术，实现相似专利推荐。第二种方法为基于语料库的跨语言专利推荐[2]。该方法抽取平行语料中双语句子或词语间的对齐关系，通过统计概率模型确定查询文本在目标语言中对应的词语或句子并过滤歧义项，最后结合单语推荐中的文本相似度计算方法，实现跨语言专利推荐。第三种方法为基于机器翻译的跨语言专利推荐[3-4]。该方法利用双语专利摘要等数据构建平行语料库，训练机器翻译模型，将源语言的查询词或句子翻译作为目标语言，将跨语言专利推荐问题转化为两个或多个对应的单语专利推荐。

跨语言专利推荐多从词语和文本篇章的精确翻译角度出发，往往需要大规模特定领域的双语词典、双语语料库以及高效准确的机器翻译方法来实现跨语言查询扩展，导致这些方法应用扩展到其他领域进行跨语言专利推荐的难度较大；与此同时，由此推荐的专利大多是相似专利，尚需从文本语义角度出发进行相关专利推荐，从而提供更好的决策支持服务。

2.2 表示学习

表示学习将研究对象的语义信息表示为稠密低维实值向量，能够充分利用对象间的关联关系，提高计算效率。表示学习对单词、短语、句子、文档、实体和社会网络等对象进行特征自动语义学习[8-9]，已经在中文分词、命名实体识别、实体消歧、情感分析、句法分析、信息检索、词向量映射等领域得到广泛应用[10-11]。跨语言专利推荐希望在同一语义空间中表示两种语言中词语和文本的语义向量，因此主要涉及词向量映射方法。

词向量映射作为表示学习的重要研究内容之一，可以有效地学习双语词向量表示[12]。现有跨语言词向量映射方法主要是利用双语平行语料库或双语词典学习词向量映射关系，主要包括三种方法：回归法使用最小二乘法将源语言词向量映射到目标语言空间中[12-13]；典型相关法使用典型关联分析法将两种语言向量映射到同一共享空间[14-15]；正交法是在正交变换的约束下，映射一种或两种语言词向量[16-17]。与此同时，基于词向量的分布信息学习跨语言词向量映射的完全无监督方法也逐渐兴起。例如，Barone[18]首先提出依赖于对抗性训练的自学习方法，使用编解码器将源语言词向量映射到目标语言中，并使用鉴别器区分映射后的源语言向量和目

标语言向量，尽管这种方法理论上可行，但模型的效果不如其他跨语言词向量映射的方法。随后，Zhang等[19]使用相似的架构，结合噪声注入等技术提高双语词典学习结果，在西英词翻译中准确率高达71.67%。Artetxe等[20]根据不同语言中的等价词应该具有相似分布的特点，提出一种新的无监督方法构建初始双语种子词典，并与鲁棒性强的自学习方法相结合，迭代改善映射效果，在英意、英德、英法词翻译上都取得了不错的效果。

目前，词向量映射方法多应用于机器翻译、跨语言自动链接、跨语言信息检索、命名实体识别、词性标注、依存关系分析等任务[11]。针对具体研究领域应用词向量映射进行深入解读和分析的研究还较少，特别是针对跨语言专利推荐这一特定研究问题，当前主要是从文本相似角度开展研究，尚需借鉴表示学习的相关理论和方法，设计针对性的无监督词向量映射方法，从文本内容语义相关的角度进行深入研究。

3　基于表示学习的无监督跨语言专利推荐方法

3.1　基于词向量映射的跨语言专利词语义表示

首先对中英专利文献进分词、去停用词等预处理，随后利用词表示学习方法 Word2Vec 中的 Skip-gram 模型对预处理后的中英专利文献分别进行词向量训练，得到中文专利词向量和英文专利词向量。在此基础上，设计无监督词向量映射方法，利用不同语言中互为翻译的两个词在各自向量空间中具有相似分布的特点，初始化中英文专利词间的相似关系，并不断迭代优化，实现中英文专利词向量在同一向量空间中的语义表示。主要包括如下三个步骤。

（1）中英专利词向量标准化。该步骤是后续步骤的基础和前提。标准化处理主要包括长度归一化和维度去均值中心化。首先，对独立训练的中英专利词向量的长度进行归一化，即将词向量的每个维度都除以该词向量的模；然后，对向量的每个维度的值进行去均值中心化，即把每个维度对应的值减去该维度所在列的均值形成新值；最后，再进行一次向量长度归一化处理。通过标准化预处理后，任何两个词向量的点积可以作为它们的相似性的度量。

（2）中英专利词向量间的语义相似关系初始化。虽然中英词向量是相互独立的，但它们在各自向量空间中的几何分布形态却是类似的。计算每个中文词向量 $\boldsymbol{x}$ 与其他中文词向量间的相似度，并对其相似度进行排序，形成这个单词的相似度向量 $\boldsymbol{x}_{sim}$。对英文词向量 $\boldsymbol{z}$ 做同样的计算，得到英文单词相似度向量 $\boldsymbol{z}_{sim}$。随后，计算中英文单词相似度向量间的相似度，形成中英文专利词间相似关系 $D(\boldsymbol{x}_{sim}, \boldsymbol{z}_{sim})$。由于不同语言中互为翻译的两个词在各自向量空间中具有相似分布，所以相似关系 D 值更高的两个单词互为翻译的概率越高，从而实现中英文单词间相似关系的初始化。

（3）基于相似关系迭代优化的跨语言专利词向量语义统一表示。在初始化中英双语间的相似关系 D 后，设置目标函数，使映射后的统一语义空间里中英文词向量间的点积最大，迭代优化目标函数与相似关系 D 直至收敛，得到最终的跨语言专利词向量语义统一表示，目标函数如公式（1）所示。

$$\underset{W_x \cdot W_z}{\arg\max} \sum_i \sum_j D_{ij} \left((\boldsymbol{X}_{i*} W_X) \cdot (\boldsymbol{Z}_{j*} W_Z) \right) \tag{1}$$

其中，$\boldsymbol{X}_{i*}$ 为中文专利词向量；$\boldsymbol{Z}_{j*}$ 为英文专利词向量；$\boldsymbol{X}_{i*}W_X$ 和 $\boldsymbol{Z}_{j*}W_Z$ 为映射后的中英文专利词向量；D_{ij} 为步骤（2）中初始化的中英文单词间相似关系。

3.2　基于跨语言专利词向量的专利文本语义表示

利用映射后的中英专利词向量对中英专利文本进行表示，最简单的方法就是平均词向量，即将专利文本中所有词的词向量相加取平均，得到的向量即是最终的专利文本向量。但这种方法最明显的缺点是认为专利文本中的所有词对于表达文本含义同样重要。还有一种使用较多的方法是 TF-IDF 加权平均词向量，即对每个词向量按照 TF-IDF 加权，随后进行平均，得到最终的专利文本表示。Arora 等[21]对 TF-IDF 加权平均词向量的文本表示进行了改进，提出平滑倒词频（Smooth Inverse Frequency，SIF）方法，用于计算每个词的加权系数，并取得了更好的效果。因此，本文使用 SIF 方法实现跨语言专利文本语义表示，融合 SIF 加权信息的中英专利文本 s 的文本向量 $\boldsymbol{v}_s$ 通过公式（2）计算得到。

$$\boldsymbol{v}_s = \frac{1}{n} \sum_{w \in s} \frac{\alpha}{p(w) + \alpha} \boldsymbol{v}_w \tag{2}$$

其中，n为文本s中的单词数量；α为参数，常被设置为0.01；$p(w)$是单词在语料中预计出现的概率。对于词频更高的单词w，SIF权值$\alpha/(p(w)+\alpha)$更小。最终的专利文本向量需要在$\boldsymbol{v}_s$中每维都减去该列的第一个主成分，保留下来的文本向量更能够表示文本本身并与其他文本向量产生差距，随后进行奇异值分解得到最终的专利文本向量。

3.3 基于专利文本语义相似度的跨语言专利推荐

利用跨语言词向量映射方法和文本表示方法将中英专利文本表示成固定维度的语义向量之后，即可以运用多种向量相似度计算指标和方法计算专利间的语义相似度，进而实现跨语言专利推荐。本文通过常用的向量夹角余弦值计算相似度，以$\vec{\boldsymbol{x}}=(x_1,x_2,x_3,\cdots,x_i,\cdots,x_n)^{\mathrm{T}}$和$\vec{\boldsymbol{y}}=(y_1,y_2,y_3,\cdots,y_i,\cdots,y_n)^{\mathrm{T}}$分别表示两篇专利的向量，专利间的语义相似度可以通过公式(3)计算得到。

$$\mathrm{sim}(\vec{\boldsymbol{x}},\vec{\boldsymbol{y}})=\cos(\vec{\boldsymbol{x}},\vec{\boldsymbol{y}})=\frac{\vec{\boldsymbol{x}}\cdot\vec{\boldsymbol{y}}}{\|\vec{\boldsymbol{x}}\|\|\vec{\boldsymbol{y}}\|}=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{3}$$

3.4 跨语言专利推荐评价

在跨语言词向量映射实验中，常采用双语词汇提取(Bilingual Lexicon Extraction)的方法进行评价，该方法给定一个中文专利词向量$\boldsymbol{x}$，使用相似度检索的方式在共享语义空间中找到与其最相近的英文专利词y作为其跨语言映射结果，随后将其映射结果与对应英文匹配单词进行比较，从而衡量跨语言词向量映射的准确性。

衡量映射结果准确性的主要指标是准确率(Accuracy)。本文采用Top-1 Accuracy、Top-5 Accuracy对中英相关专利词映射结果进行评估。其中，Top-1 Accuracy是指源语言单词的词映射结果中排名第一的单词是其对应的目标语言单词的概率；Top-5 Accuracy是指源语言单词的映射结果中排名前五的单词包含其对应的目标语言单词的概率。

同样地，本文也将评测指标Top-1 Accuracy、Top-5 Accuracy应用于跨语言专利推荐结果的衡量中。Top-1 Accuracy是指跨语言专利推荐排名第一的专利是对应的专利翻译版本的概率；Top-5 Accuracy是指跨语言专利推荐排名前五的专利中包含对应的专利翻译版本的概率。

4 实验与分析

近年来，无线通信技术发展迅猛，已经成为全球通信业发展最受关注的产业领域之一，给人们的生活带来了巨大的便利和影响。截至2018年底，全球共有51亿人使用无线移动服务，占全球人口的67%。未来一段时间内，无线移动通信技术演进、智能终端和业务应用将形成广阔的市场空间，是全球通信业发展的重要推动力。因此，本文采用无线通信网络的中英文专利文献作为实验对象，进行无监督跨语言专利推荐的实证分析。

4.1 数据获取与预处理

英文专利语料数据来自德温特专利数据库(Derwent Innovations Index，DII)，通过检索2016年到2018年间国际专利分类号(International Patent Classification，IPC)为H04W的无线通信网络英文专利数据，共获取71 202篇专利。对应的检索表达式为“IPC=(H04W*)时间跨度:2016-2018”。中文专利语料数据来自中国专利全文数据库(知网版)，其文献来源为国家知识产权局知识产权出版社。同样检索2016年到2018年间IPC分类号为H04W的无线通信网络中文专利数据，共获取46 333篇专利。对应的检索表达式为“分类号=H04W，公开日=2016年1月1日至2018年12月31日”。

为了后续跨语言专利推荐定量评价的开展，需要对英文专利和中文专利进行匹配。得益于德温特专利数据著录项中的专利号(Patent Number，PN)包含该专利在不同国家申请的专利号，中国专利作为世界知识产权的重要部分，其专利号同样包含在内。因此，通过完全匹配方式可以实现中文专利与英文专利的对应，最终得到43 600篇同时拥有中文与英文的专利数据，用于跨语言专利推荐定量评价。

数据预处理主要是抽取中文和英文数据中的专利标题和摘要，并进行文本预处理。由于专利数据含有大量的专业术语词组，英文数据如果简单采用空格分词，效果不是很好，所以使用NLTK(Natural Language Toolkit)对英文专利文本进行短语抽取，并

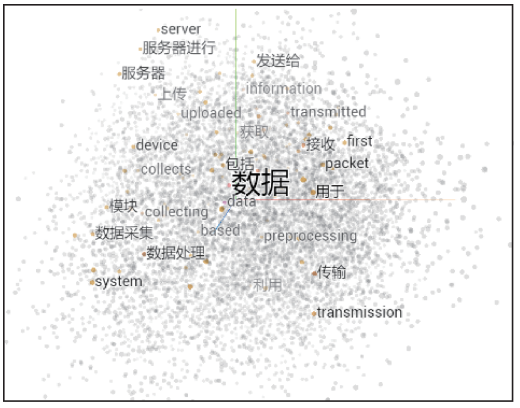
添加开放的《最新实用英汉电信词典》[①]为英文短语分词词典。中文专利数据使用jieba进行分词，同样添加《最新实用英汉电信词典》为自定义分词词典，从而提高预处理效果。接着，利用gensim提供的Word2Vec模块对预处理后的中英文专利文本数据分别进行单语专利词向量的构建，用于后续实验。其中重要的参数设置遵循推荐系统的一般通用设置："sg=1"表示使用Word2Vec中的Skip-gram模型，Skip-gram模型相较CBOW虽然训练时间长，但对于出现频率不高的单词，效果更好；"size=300"表示向量维度为300，即每条数据都表示为300维空间中的一个点；"window=5"表示取距离当前中心词位置为5的词作为上下文信息；"hs=1"表示使用Hierarchical Softmax优化策略。为了保证每次输出的词向量不变，设置"worker=1，seed=0"。应用TensorFlow[②]提供的向量可视化工具对单语词向量进行可视化，如图1所示。可以看出，通过降维到二维空间后，中英单语词向量明显分别处于两个不同的集群中，印证了在原始维度下，两者分属不同的语义空间。

图1映射前的中英单语专利词向量可视化

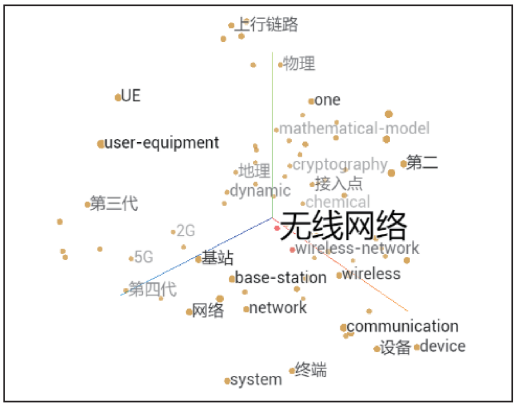
Fig.1　Visualization of Patent Word Vectors Before Mapping

4.2　基于词向量映射的中英专利词语义表示

利用无监督跨语言词向量映射方法对中英文专利词向量进行映射，实现中英文专利词向量在同一向量空间中的语义表示。应用TensorFlow提供的向量可视化工具，对映射后的中英专利词向量结果进行可视化展示，如图2所示。与图1所示映射前的中英专利单语词向量不同，映射后的中英专利词向量明显处于同一交叉语义空间。



(a) 映射后的中英专利词向量



(b) 语义相关词的位置关系示例

图2　基于无监督跨语言词映射的中英专利词向量表示可视化

Fig.2　Visualization of Patent Word Vectors After Mapping

本文在双语词汇提取评价中选择不同的相似度检索方式，分别为Cross-Domain Similarity Local Scaling检索（CSLS）[22]和传统的最近邻检索（K-Nearest Neighbor，KNN），并与弱监督[23]（使用25个

①http://idl.hbdlib.cn/book/00000000000000/pdfbook/018/016/166530.pdf.

②http://projector.tensorflow.org.

单词的双语词典作为初始双语相似关系）方法进行对比，结果如表1所示。可以看出，在中英专利词映射准确率上，本文使用的无监督跨语言词向量映射的方法明显要高于弱监督的方法。而选择CSLS检索方式的准确率不管是在无监督还是弱监督中都比最近邻检索（KNN）有小幅度的提高。

表1　中英跨语言词映射准确率（%）

Table 1　Accuracy of Cross-Language Word Mapping

检索方法	CSLS	KNN
弱监督	46.72	43.68
无监督	49.08	46.87

考虑到出现频率较低的单词可能会对无监督跨语言专利词映射效果产生影响，图3展示了基于频率排序的中英专利单词数量对无监督词向量映射准确率的影响。可以看出，随着单词的数量增多，低频率的单词越来越多，中英跨语言专利词映射的准确率有所下降。但对于前30 000的单词来说，Top-5 Accuracy仍然比较高，约为55%。同时可以看出最常见的5 000个单词映射效果最好。

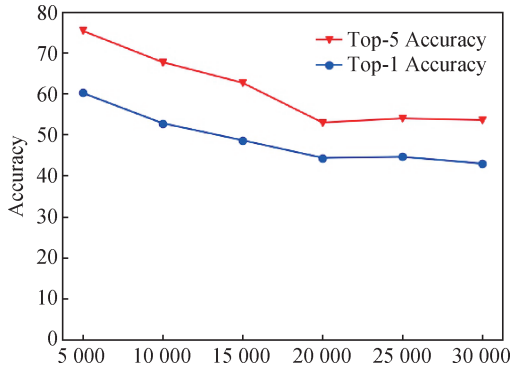


图3专利单词数量对跨语言词映射准确率的影响

Fig.3　Influence of the Number of Patent Words on Mapping Accuracy

中英跨语言专利相关词映射的部分实例如表2所示。可以看出，名词映射效果较动词和形容词更好，“接入点”相似度最高的三个映射词除了包含常见匹配单词外，还提供了英文缩写“AP”。由于英文动词拥有不同的时态，导致相似度最高的英文映射单词大多都不是动词的原形。而对于形容词，英文形容词拥有许多不同的表达，如“快速的”在英文中有多个翻译“fast”“quick”“rapid”等，而且形容词和副词在很多时候都是相通的，所以“quickly”“rapidly”也是相关词。

表2　中英跨语言专利相关词映射实例

Table 2　Examples of Cross-Language Patent Word Mapping

中文单词	英文映射单词	常见匹配单词
移动终端	mobile-terminal；terminal；mobile-phone	mobile-terminal
接入点	access-point；AP；access-points	access-point
选择	selecting；selection；selected	select
检测	reducing；reduced；reduce	reduce
快速的	quickly；rapid；rapidly	fast
准确的	accurately；accuracy；accurate	accurate

4.3　基于中英专利文本语义表示的跨语言专利推荐评测

利用无监督跨语言专利词向量对中英专利名称和摘要进行文本向量表示，进而实现跨语言专利推荐。同样应用TensorFlow提供的向量可视化工具，对中英专利文本向量进行可视化展示，如图4所示。其中，cn代表中文专利，en代表英文专利，同一篇专利使用相同的编号。例如，cn1849和en1849表示公开号为CN106376069A专利的中文版本与英文版本。从该示例可以看出，中英专利文本不仅处于同一语义空间，而且在距离上是最近的，两者的相似性最高。

为了验证方法有效性，本文选取跨语言推荐中效果较好的机器翻译方法进行对比分析[3]。参照跨语言推荐的一般思路，首先选取翻译效果较好的谷歌翻译将英文专利文本批量翻译为中文文本；然后使用文本向量表示方法[24]实现中英专利文本向量表示，为了保证比较的公平性，使用与其他方法一致的TF-IDF加权词向量表示专利文本向量；最后使用相同的余弦相似度计算向量相似度，实现中英跨语言专利推荐。

与此同时，将弱监督词映射方法和无监督词映射方法与不同专利文本表示方式进行结合，比较基于表示学习的中英跨语言专利推荐效果，结果如表3所示。平均词向量文本表示与无监督词向量映射相结合的跨语言专利推荐方法由于没有添加任何单词重要性的信息，准确率相对较低，Top-1 Accuracy只

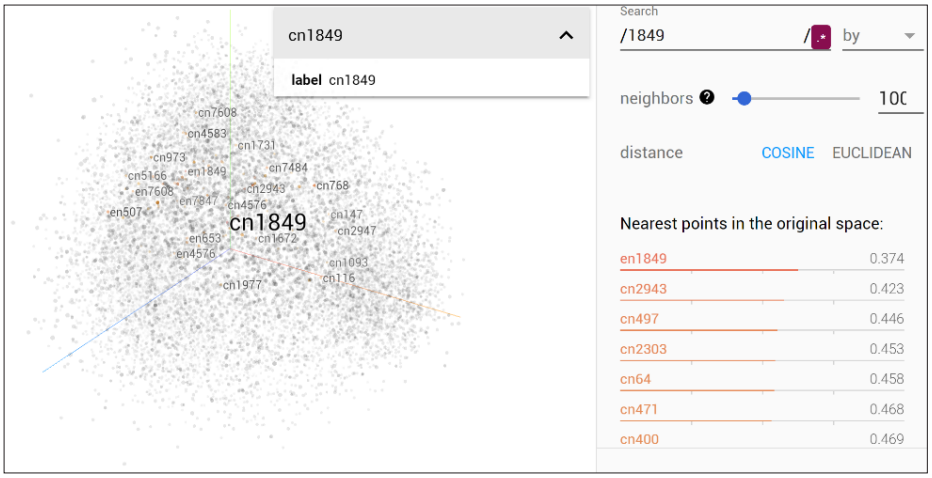


图 4　基于跨语言专利词向量表示的中英专利文本表示

Fig.4　Visualization of Patent Text Representation

有 33.75%。SIF 加权与 TF-IDF 加权相比，基于 SIF 加权的无监督跨语言专利推荐效果更好，Top-1 和 Top-5 推荐准确率分别达到 55.63% 和 77.82%，较弱监督和机器翻译的跨语言专利推荐方法均有小幅提高。

表 3　中英跨语言专利推荐准确率(%)

Table 3　Accuracy of Cross-language Patent Recommendation

跨语言专利推荐方法	Top-1 Accuracy	Top-5 Accuracy
机器翻译	51.34	73.92
无监督+平均词向量	33.75	56.50
无监督+TF-IDF	42.01	65.45
弱监督+SIF	54.97	76.37
无监督+SIF	55.63	77.82

4.4　无监督跨语言专利推荐示例

本文选择近期热门的无线通信领域中有关 5G 网络的一篇中文专利作为测试专利，展示中英跨语言专利推荐结果。该专利的专利号为 CN108476513A，专利名称为“用于提供 5G 上行链路请求的设备和方法”，推荐结果如表 4 所示。这篇专利主要是用于 5G 系统中上行链路数据调度方法的研究，选取推荐结果排名前三的英文专利进行结果分析。排名第 1 的英文专利是测试专利的英文版，可以看出，虽然两篇专利是同一专利的不同语言版本，但由于英文专利有固定的格式，中英专利的名称和摘要不是完全按句子互为翻译。即便如此，由于它们之间的语义相关性，仍然在结果中排在第一位。

推荐排名第 2、3 的英文专利，虽然都与 5G 网络无关，但都与下行链路传输的物理通道相关。其中，对上下行链路的表达则选取了英文“uplink”和“downlink”的缩写“UL”和“DL”。这表明无监督跨语言专利推荐方法不仅可以有效减少查询构建的工作量，而且能够较好地处理同义词、缩写等不同的词汇表达，进而推荐语义相关的跨语言专利。

同样地，本文对英中跨语言专利推荐也进行示例说明。选取一篇实际应用较强的英文专利作为测试专利，专利号为“CN109005522A”，专利名称为“Wireless sensor network based intelligent forest fire monitoring system，has single sink node for converging environment monitoring data and transmitting received environment monitoring data to monitoring terminal”，主要内容为设计了一个基于无线传感器网络的森林防火智能监控系统。选取推荐结果排名前五的中文专利进行结果分析，发现在前三篇专利中并没有测试专利的中文版本，进一步查看发现第 5 篇中文推荐专利为测试专利的中文版本。主要原因在于该专利对应的中文版摘要非常简单，导致其推荐排名靠后，进而也会导致无监督跨语言专利推荐准确率降低。虽然如此，推荐排名第 1 的中文专利与测试专利高度相关，同样将无线传感器网络用于森林环境监测。而推荐排名第 2 的中文专利是无线传感器网

表4　中英跨语言专利推荐结果示例

Table 4　Examples of Cross-language Patent Recommendation

专利元数据名称	元数据内容
中文测试专利号	CN108476513A
中文测试专利名称	用于提供5G上行链路请求的设备和方法
中文测试专利摘要	一般地描述了在5G系统中调度上行链路数据请求的设备和方法。UE在预留用于5G调度请求的或未预留的5G或LTE链路资源上向eNB发送调度请 求(SR)或5G物理随机接入信道(xPRACH)。消息取决于使用哪个链路进行发送。取决于是否使用预留的资源和预留的逻辑信道ID,在发送SR之后并且响应于接收到针对该SR的上行链路许可,UE向eNB发送BSR和可能的波束测量报告。然后,响应于在最优波束上接收到包括针对数据的5G上行链路许可的5G物理下行链路控制信道,UE发送5G物理上行链路共享信道。在发送xPRACH时,使用减少的随机接入响应。
英文推荐排名第1的专利号	CN108476513A
英文推荐排名第1的专利名称	User equipment apparatus
英文推荐排名第1的专利摘要	NOVELTY - The apparatus has a processing circuitry for decoding fifth generation (5G) physical downlink control channel (PDCCH) containing 5G uplink grant received from a 5G evolved NodeB (eNB) (506) on selected beam, where the 5G uplink grant comprises resources allocated for transmission of uplink data to the 5G eNB. The processing circuitry generates 5G physical uplink shared channel (xPUSCH) comprising the data for transmission to the 5G eNB using the resources. USE - User equipment (UE) apparatus i.e. eNB apparatus. ADVANTAGE - The apparatus reduces number of messages between the UE and the 5G eNB so as to reduce uplink access latency. DETAILED DESCRIPTION - An INDEPENDENT CLAIM is also included for a computer-readable storage medium comprising a set of instructions for using UE apparatus.
英文推荐排名第2的专利号	CN108605364A
英文推荐排名第2的专利名称	Wireless communication method for uplink (UL) data transmission
英文推荐排名第2的专利摘要	NOVELTY - The wireless communication method involves sending contention-based UL message based on the selected resource block by the user equipment (UE) to the base station, with the UL message indicating to reserve the corresponding UL resources associated with the selected resource block. A response signal is received from the base station in response to the UL message, with the positive response signal indicating success reservation of the corresponding UL resources. UL data is transmitted on the corresponding UL resources upon receiving the positive response signal. USE - Wireless communication method for UL data transmission e. g. contention-based UL resource reservation. ADVANTAGE - Allows UE to select resource block from resource pool, sends UL message based on selected resource block requesting resource reservation with or without reservation resource request (RRR) and receives response signal indicating success of resource reservation. Reduces signaling overhead by the implementation of contention-based UL resource reservation. Defines one-to-one mapping to avoid blind detection in UE to save power consumption. The utilization of UL resource is improved since the resource reserved for RRR is much smaller than the one for contention-based UL message with large payload.
英文推荐排名第3的专利号	CN107113911A
英文推荐排名第3的专利名称	Scheduling method used in a Base Station (BS) for scheduling a User Equipment (UE) involves determining uplink (UL) control channel position for UE, based on allocated downlink (DL) resource blocks (RB)
英文推荐排名第3的专利摘要	NOVELTY - The scheduling method (900) involves allocating (910) one or more DL RB for transmitting DL data to the UE, then determining (920) an UL control channel position for the UE transmitting a Hybrid Automatic Repeat Request (HARQ) feedback of the DL data, based on the allocated DL RB. The DL data is transmitted (930) to the UE using the allocated DL RB. USE - Scheduling method used in BS for scheduling UE. ADVANTAGE - Eliminates HARQ index confliction between dynamic scheduling and Semi-Persistent Scheduling (SPS), while improving physical uplink control channel (PUCCH) resource usage efficiency, by indicating UL control channel position for the UE transmitting a HARQ feedback of the DL data using the allocated DL RB.

络在其他一些场景的应用,涉及使用无线传感器进行数据采集以及后续数据传输和数据接收等,也具有很强的语义相关性,验证了该方法在语义表达上的优势。

除了可以推荐跨语言专利外，该方法同样也可以推荐同语言的相关专利，从而实现多语言专利推荐。同样以中文专利号为CN108476513A，专利标题为“用于提供5G上行链路请求的设备和方法”的专利进行推荐。结果显示，中文推荐结果中排名第2的专利与英文推荐结果中排名第3的专利是同一专利号“CN107113911A”的不同语言版本，从侧面表明本文提出的跨语言推荐方法的有效性。推荐的其他中英文专利也均与该专利具有间接关联，其内容为不同通信链路或设备间的传输技术，具有语义上的高关联性。

综上，无监督跨语言专利推荐的方法可以有效地实现相关专利推荐，并减少了用户构建跨语言查询的工作量。在不需要任何外部双语词典和大型双语语料库的情况下，只需要输入整篇专利就可以得到目标语言的相关专利推荐。推荐结果具有的高语义相关性和多样性，为更全面有效地发现相关领域的前沿技术、新兴技术和发展态势提供了数据基础，也可以便利地向专利审查员、专利申请人等提供详细准确的参考引用、方法创新和产品革新信息。

5　结　语

针对多语言专利文献数量巨大，难以有效获取跨语言相关专利的迫切现实需求，本文借鉴深度学习的理论和方法，设计无监督跨语言词向量映射方法，形成基于跨语言专利词向量的专利文本语义向量表示，应用向量相似度计算指标计算不同语言专利文本间的语义相似度，构建基于表示学习的无监督跨语言专利推荐方法，实现跨语言专利推荐。该方法一方面减少了双语词典和大规模双语语料库的需要，另一方面从文本语义角度出发进行相关而非相似专利推荐，可以得到更丰富的专利推荐结果，提高查询效果，提供更好的决策支持服务。

该方法在无线通信领域中的实验结果证实，基于无监督跨语言词映射的中英专利词向量表示与专利文本表示相结合的方法可以更有效全面地实现跨语言专利推荐。在未来的研究中，可以将无监督词向量映射方法扩展到其他领域和多种语言中，进一步验证方法的有效性。同时，可以将基于无监督跨语言词映射的专利词向量表示与专利文本表示结合的方法应用到其他的专利挖掘任务中，丰富不同语言专利间的深度挖掘研究，扩展专利情报分析方法。

参考文献：

[1] Jochim C, Lioma C, Schütze H, et al. Preliminary Study into Query Translation for Patent Retrieval[C]//Proceedings of the 3rd Workshop on Patent Information Retrieval.2010: 57-66.

[2] Magdy W, Jones G J F. Studying Machine Translation Technologies for Large-Data CLIR Tasks: A Patent Prior-Art Search Case Study [J]. Information Retrieval, 2014, 17(5): 492-519.

[3] Magdy W, Jones G J F. An Efficient Method for Using Machine Translation Technologies in Cross-Language Patent Search[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. 2011: 1925-1928.

[4] Shen X, Huang H Y, Li L Z, et al. A Parallel Cross-Language Retrieval System for Patent Documents[C]//Proceedings of the 6th IEEE International Conference on Software Engineering and Service Science. 2015: 672-676.

[5] Lee C S, Wang M H, Hsiao Y C, et al. Ontology-Based GFML Agent for Patent Technology Requirement Evaluation and Recommendation[J]. Soft Computing, 2019, 23(2): 537-556.

[6] Ji X, Gu X J, Dai F, et al. Patent Collaborative Filtering Recommendation Approach Based on Patent Similarity[C]// Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery. 2011: 1699-1703.

[7] Rui X H, Min D. HIM-PRS: A Patent Recommendation System Based on Hierarchical Index-Based MapReduce Framework[C]// Proceedings of UCAWSN 2016, CUTE 2016, CSA 2016: Advances in Computer Science and Ubiquitous Computing. 2016: 843-848.

[8] 李枫林, 柯佳. 词向量语义表示研究进展[J]. 情报科学, 2019, 37(5): 155-165. (Li Fenglin, Ke Jia. Research Progress of Word Vector Semantic Representation[J]. Information Science, 2019, 37 (5): 155-165.)

[9] 涂存超, 杨成, 刘知远, 等. 网络表示学习综述[J]. 中国科学:信息科学, 2017, 47(8): 980-996. (Tu Cunchao, Yang Cheng, Liu Zhiyuan, et al. Network Representation Learning: An Overview [J]. SCIENTIA SINICA Informationis, 2017, 47(8): 980-996.)

[10] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261. (Liu Zhiyuan, Sun Maosong, Lin Yankai, et al. Knowledge Representation Learning: A Review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.)

[11] 彭晓娅, 周栋. 跨语言词向量研究综述[J]. 中文信息学报, 2020, 34(2): 1-15. (Peng Xiaoya, Zhou Dong. Survey of Cross-Lingual Word Embedding[J]. Journal of Chinese Information Processing, 2020, 34(2): 1-15.)

[12] Mikolov T, Le Q V, Sutskever I. Exploiting Similarities among Languages for Machine Translation[OL]. arXiv Preprint, arXiv: 1309.4168.

[13] Dinu G, Baroni M. Improving Zero-Shot Learning by Mitigating the Hubness Problem[C]// Proceedings of the 3rd International Conference on Learning Representations. 2014. DOI: 10.1007/978-3-319-23528-8_9.

[14] Faruqui M, Dyer C. Improving Vector Space Word Representations Using Multilingual Correlation[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.2014: 462-471.

[15] Lu A, Wang W, Bansal M, et al. Deep Multilingual Correlation for Improved Word Embeddings[C]//Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 250-256.

[16] Smith S L, Turban D H P, Hamblin S, et al. Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax[C]//Proceedings of the 5th International Conference on Learning Representations.2017.

[17] Xing C, Wang D, Liu C, et al. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation[C]// Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015: 1006-1011.

[18] Barone A V M. Towards Cross-Lingual Distributed Representations Without Parallel Text Trained with Adversarial Autoencoders[C]//Proceedings of the 1st Workshop on Representation Learning for NLP.2016: 121-126.

[19] Zhang M, Liu Y, Luan H B, et al. Adversarial Training for Unsupervised Bilingual Lexicon Induction[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.2017. DOI: 10.18653/v1/P17-1179.

[20] Artetxe M, Labaka G, Agirre E. A Robust Self-Learning Method for Fully Unsupervised Cross-Lingual Mappings of Word Embeddings[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.2018. DOI: 10.18653/v1/P18-1073.

[21] Arora S, Liang Y, Ma T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings[C]//Proceedings of the 5th International Conference on Learning Representations. 2017.

[22] Conneau A, Lample G, Ranzato M A, et al. Word Translation Without Parallel Data[C]//Proceedings of the 6th International Conference on Learning Representations. 2017.

[23] Artetxe M, Labaka G, Agirre E. Learning Bilingual Word Embeddings with (Almost) no Bilingual Data[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 451-462.

[24] Oh S, Lei Z, Lee W C, et al. CV-PCR: A Context-Guided Value-Driven Framework for Patent Citation Recommendation[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013: 2291-2296.

作者贡献声明：

张金柱：提出研究思路，设计研究方案，修改论文；
主立鹏：实验设计与实现，论文起草；
刘菁婕：实验设计与实现，论文修改。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

支撑数据由作者自存储，E-mail：liujingjie@njust.edu.cn。
[1] 主立鹏, 刘菁婕. H04W16-18PP. xlsx. 原始实验数据.
[2] 主立鹏, 刘菁婕. 16-18cnPPP. txt/16-18enppp. txt. 预处理后中/英文数据.
[3] 主立鹏, 刘菁婕. 16-18cnw. txt/16-18enw. txt. 训练后的中/英文词向量.
[4] 主立鹏, 刘菁婕. 16-18cnwmap. txt/16-18enwmap. txt. 无监督词向量映射后的中/英文词向量.
[5] 主立鹏, 刘菁婕. doc_sifcn. txt/doc_sifen. txt. 文本表示后中/英专利文本向量.
[6] 主立鹏, 刘菁婕. doc_sif. xlsx. 跨语言专利推荐计算结果.

收稿日期：2020-03-31
收修改稿日期：2020-07-31

Unsupervised Cross-Language Model for Patent Recommendation Based on Representation

Zhang Jinzhu[1, 2]　Zhu Lipeng[1]　Liu Jingjie[1]

[1](School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)

[2](Jiangsu Provincial Social Public Safety Science and Technology Collaborative Innovation Center, Nanjing 210094, China)

Abstract: [**Objective**] This paper designs a cross-language recommendation model for patents based on text semantic representation, aiming to reduce the number of bilingual dictionaries and large-scale corpus, as well as improve the ability of domain adaptation. [**Methods**] First, we designed a word vector mapping method with unsupervised cross-language algorithm. Then, we mapped Chinese and English word vectors to the unified semantic vector space with linear transformation, which constructed the semantic mapping relationship between Chinese and English words. Third, we created semantic representation of patent texts based on cross-language word vector with smooth inverse frequency (SIF) reweighting method. It realized the semantic representation of Chinese-English patent texts in the same vector space. Finally, we calculated the semantic similarity between patent texts and recommend the cross-language patents. [**Results**] We examined the proposed method with patents on "wireless communication" and the recommendation accuracy rate of the top 1 and the top 5 reached 55.63% and 77.82%, which were 0.66% and 1.45% higher than those of the weak supervised based cross-language recommendation. They were also 4.29% and 3.90% better than the machine translation based ones. [**Limitations**] We only examined the proposed method with Chinese and English patents from one specific field. [**Conclusions**] This proposed method could recommend Chinese and English patents effectively, which help future research in cross-language patent recommendations.

Keywords: Cross-Language　Patent Recommendation　Representation Learning　Semantic Representation