情报科学
第39卷 第7期　2021年7月

在线技术供需文本匹配方法研究综述

余　辉,梁镇涛,谢　豪

(武汉大学 信息资源研究中心, 湖北 武汉 430072)

摘　要:【目的/意义】随着网络信息技术的发展及国家对技术转移的政策支持,大量的在线技术交易需求产生。在线技术转移服务平台作为技术供需交易的媒介,供需双方可以在平台上发布大量的技术供需文本信息,提高技术供需文本匹配效率,有助于提高技术交易成功率,促进技术转移。【方法/过程】在分析传统文本匹配方法的基础上,从基于关键词的匹配方法、基于句法分析和文本结构的匹配方法、基于深度学习的匹配方法和基于多维度视角的匹配方法四个方向对目前在线技术供需文本匹配方法现状进行梳理。【结果/结论】大多数研究都融合了多种匹配方法,从多维度视角进行匹配是研究发展趋势。在技术供需文本匹配未来研究中,除了继续将深度学习方法融合到已有的各种方法中,还应该从多维度、跨模态和可解释性方向来提高技术供需文本匹配效率。【创新/局限】本文将技术文本匹配方法进行归纳总结,能对技术文本的匹配方法提供参考,但现实中技术匹配还应该考虑其他影响技术转移的因素。

关键词: 技术转移;技术供需文本;文本表示;供需匹配;技术推荐

中图分类号: TP391.1;G254　　**DOI:** 10.13833/j.issn.1007-7634.2021.07.024

1 引　言

技术市场中存在的信息不对称影响技术交易[1],在线技术转移服务平台通过在线发布技术供需信息,减少了这种信息不对称的影响[2],发展和完善在线技术供需文本的匹配度研究有助于提高技术转移效率,并促进企业提升创新能力以及技术商业化[3],技术供需匹配是技术交易的重要环节,指引技术创新的发展[4],通过研究技术供需双方文本信息的匹配与推荐来提升技术转移服务效率成为技术创新和管理实践的重要研究之一。

匹配问题研究最早起源于婚姻匹配,匹配的目的是使匹配的双方尽可能找到满意的对方[5],在线技术供需信息匹配即对技术供需双方的技术需求与技术供给进行匹配,并协助完成技术交易,实现技术转移。目前在线技术供需信息匹配主要通过在线技术转移服务平台来实现,主要形式有在线技术交易合同登记以及专利转移两类业务。在线服务平台在技术供需双方间起到媒介的作用,提供服务平台给双方发布信息和专业的资源服务[6],发现技术转移机会[7],建立供需双方的信任[8],协助双方进行技术转移,监督双方在交易后的实施行为等。截止到2018年底,我国国家技术转移机构有453家,促成技术转移项目超过12万项,成交额超过2千亿,比上一年增长近20%[9]。2019年全国登记技术合同484 077项;成交额为22 398.4亿元,比上年增长26.6%,技术合同交易额本身也在逐年增长,可见技术转移机构还有很大的发展空间,研究技术转移中供需信息匹配,不仅是保障这些技术交易的需要,也是促进更大的技术市场的需要。

目前在线技术供需匹配服务平台的文本匹配工作可以分为文本的表示及文本的相似度计算两个步骤。技术供需文本表示方法是在文本表示的研究基础上,结合技术文本本身的特征来开展研究的,文本匹配目前研究简要分为三种类型。即传统基于关键词的词频、共现和网络结构图的研究;现在广泛应用的深度学习方法,利用词向量和文本向量对文本的语义特征表示,包括独热向量和词嵌入模型[10];以及学者们一直在尝试的基于多维度的匹配方法。根据不同的文本表示,语义相似度进行计算与匹配方法也不同,通常可以分为表象匹配和语义匹配,最终用统计方法或向量的余弦相似度来计算。表象匹配如字符串匹配,通常应用包括关键词共现和重复程度来衡量相似度[11],如在基于关键词对供需文本进行表示时可以采用此方法;语义匹配,文本语义向量化进行向量的相似度计算,如词袋模型,从分布上考虑上下文的相似程度来衡量文本的语义相似度[12],通常包括向量

收稿日期: 2020-09-22

基金项目: 国家重点研发计划项目"长江中游城市群综合科技服务平台研发与应用示范"(2018YFB1404300);国家自然科学基金重大课题"国家安全大数据综合信息集成与分析方法"(71790612),"大数据环境下基于特征本体学习的无监督文本分类方法研究"(71571064)

作者简介: 余辉(1993-),男,博士研究生,主要从事网络信息智能处理研究;梁镇涛(1996-),男,硕士研究生,主要从事信息分析与知识发现、信息计量研究;谢豪(1998-),男,硕士研究生,主要从事多模态数据融合研究.

空间模型(VSM)、潜在语义分析(LSA)、概率潜在语义分析(PLSA)和潜在狄利克雷分布(LDA)等[11];此外在张量神经网络中除了可以用采用余弦相似度,还可以用点积或双线性相似度来进行相似度的计算[13]。而技术供需文本匹配上,除了面临传统短文本处理中信息稀疏、表达不规范、异构和非结构化等挑战外,还存在着口语化与技术词汇表达不一致、部分重要匹配信息缺失以及语料库资源不足等难题。一些在线技术转移服务平台在分析技术需求文本的基础上,针对每个技术需求文本进一步分析出了相对应的需求功能,但并未能进行技术供给匹配。

本文在回顾文本处理的基础上,总结在线技术转移服务平台所需要的技术供需文本匹配研究,以文本的语义表示方法为重点,将研究分为基于关键词的匹配、基于句法分析和文本结构的匹配、基于深度学习的匹配以及基于多维度视角的匹配。

2 基于关键词的匹配

基于关键词的匹配方法是较早的方法,包括基于词频统计、关键词共现、词图网络,以及基于关键词语义的方法,这些方法的核心思想都是用关键词来进行文本表示,并以关键词的权重来衡量两篇文本的相似度。

2.1 关键词词频和共现

词频统计是较早且较为直观的文本分析方法,直接根据词频来确定特定关键词的重要程度,也可以通过词频变化来对相关技术发展情况进行分析和预测。熊则见等人就采用关键词词频分析研究高技术产品研发的关键成功因素[14]。关键词共现与词图网络联合使用,即使用关键词共现来构建词图网络,通常可以用来识别重要节点[15]。谢玮等人基于TextRank的词影响力和逆文档频率IDF,构建了关键词图,并应用于论文与审稿专家的自动匹配系统[16]。Rahman和Roy应用TextRank识别并构建软件开发的任务变更需求和源代码位置之间的映射,为开发人员提供良好的检索条件[17]。

根据关键词的权重来衡量词在文本信息中的相关和重要程度,最经典的方法是TF-IDF[18],它的应用十分成熟,在易用性和实用性方面都有不错的应用效果,被应用到了多种文本处理场景,如Kuncoro等人用来进行关键词排序[19],Zheng Y等人用TF-IDF对短文本进行热点主题聚类与识别[20]。学者们经常结合TF-IDF与其他方法来提高单一方法的效果,如贺飞艳等人结合TF-IDF和方差统计对短文本进行多分类特征抽取[21],He结合词频统计方法和关键词位置,提升了对文本热点识别的效果[22]。

在技术供需文本匹配中,每一个技术需求文本或者供给文本为一个主体,语料库可以是对应的所有主体的需求文本或供给文本,或者是二者的并集。TF-IDF的具体表现为每一个技术关键词的权重与出现在各自主体中频次成正比,与出现在整体语料库的频次成反比。每个技术需求和供给文本都可以用关键词权重来向量化,通过计算文本向量之间的余弦相似度来计算技术供需文本的匹配度。杨德林等人用TF-IDF结合余弦相似度进行文本相似度的计算,对在线技术转移服务平台供需匹配效率进行分析研究,找出了供需文本信息的语言差异问题[23]。基于关键词的TF-IDF在技术供需匹配中运用流程如图1所示。

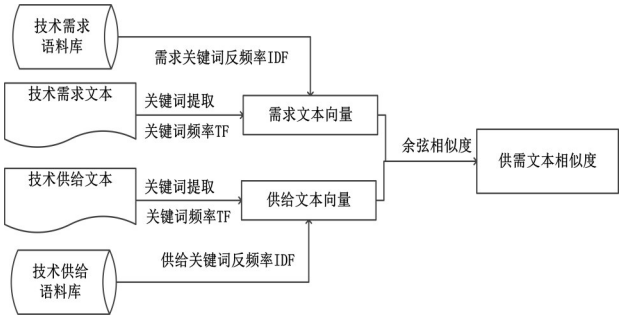

图1 基于关键词的TF-IDF方法

Figure 1 TF-IDF method based on keywords

基于关键词词频或者共现等方法,整体在分类上有一定的准确性,但是在不考虑语法和句法结构的基础上进行的相似度计算,忽略了文本语义和关键词之间的关系[24],并不一定能得到理想的匹配结果。例如,“企业需要大力发展新能源汽车驱动技术”和“企业不需要大力发展新能源汽车驱动技术”在词频和共现计算的相似度上是非常高的,但是却是在内燃机和电动机两个不同驱动技术上的需求。所以在实际研究工作中,关键词方法多与其他方法相结合以得到更好的匹配效果。

2.2 基于关键词语义

基于关键词词频和共现的统计方法忽略了关键词本身的含义,而关键词作为文本主题词的一种,本身就可以用来进行语义上的匹配,而不仅仅是表象上的匹配。基于关键词语义的匹配方法需要用到语义词典,也是基于关联知识和特征工程的匹配方法,不限于处理文本的语料库,借助其他语料库来获取更充分的信息,从整体上考虑确定语义,通常需要构建本体或者利用已有相关词典。

关键词的语义研究在文本匹配研究中可以被应用在关键词提取和文本相似度计算两个过程中,即在提取阶段就考虑关键词语义和匹配阶段考虑语义两种。在关键词提取阶段,王立霞和淮晓永把语义特征融入关键词提取过程中,解决了关键词词频和共现方法只考虑字面匹配忽略语义的问题[25];刘端阳和王良芳提出了基于《同义词词林》[26]语义词典的关键词语义相似度的提取算法,解决关键词提取过程中一词多义的问题[27]。在语义匹配阶段,任龙利用BM25算法,将其用法延伸至专利方面,将专利文本与背景知识库进行匹配计算[28],方俊等人基于英文词典WordNet计算关键词之间的内聚性[29],Gang等人基于知网HowNet计算中文语义相似度[30],Wei T结合WordNet和词汇链对文本进

行语义聚类[31]。除了知网的HowNet外，各个地区都有一些词典的代表，如欧洲国家的Eurowordnet、微软的Mindnet、韩国的Koreanwordnet等。而随着网络数据更新越来越快，一些学者认为传统词典在语义适应上与新的文本数据会有一些出入，于是以网络生成数据为训练样本的词典开始运用于语义分析上。如Wu Z等人认为单纯基于关键词的方法用关键词来表示文本信息在语义的表达上不足，所以他们以维基百科为基础语料库来进行语义匹配实现对文本信息的分类[32]。同年，Jiang Y 等人也运用维基百科中分类结构的概念，提出了解决语料库局限问题的信息相似度计算方法[33]。

在技术供需文本匹配研究上，何喜军等人基于关键词语义构建了供需匹配视角下的技术需求识别模型，利用到了专利转让数据库进行语义匹配值的计算[34]，但总体上目前缺乏较为全面的技术领域的词典和数据库，除了利用已有词典和网络语料库外，还需要借助其他方法去获取技术词汇的语义关系，如刘翔等借助功能基和TRIZ原理进行规范表达，利用本体技术组织专利知识之间关系，提出了一种支持产品创新设计的专利知识库构建方法[35]，梁浩融入本体论、语义Web关键技术、TRIZ功能分析等，提出了一个能够适用于满足万维网的面向创新设计的语义知识表示模型，构建了基于林业机械领域本体知识的TRIZ融合架构[36]，张娴等基于语义TRIX理论与结构化本体构建方法，分别构建领域本体、专利技术供给本体和专利技术需求本体，设计供需匹配模型，并对专利技术供需信息匹配任务及其知识关联进行形式化描述[37]。此外，谷歌开发的语言建模工具Word2Vec[38]、Facebook开发的FastText[39]等用于深度学习的方法，能进一步提高技术供需文本匹配的准确率。

3　基于句法分析和文本结构的匹配

基于关键词匹配的方法不考虑文本结构可能会引起的文本匹配错误问题，而基于句法、语法以及知识结构的文本处理方法，对文本进行拆分，能有效辨识出是否存在否定词从而解决这个问题。此外依存句法和知识结构能在一定程度上体现关键知识点在文本中的重要性，如出现在技术供需文本的主题和正文中的技术关键词重要性不一样，出现在主语和宾语的关键词也可能有区别。

3.1 基于句法分析的匹配

基于句子分析的匹配，是借助依存句法关系、词性等，如句子的主谓宾结构和句子中的词汇之间的关系进行语义匹配研究。这些方法一般需要结合构建好的领域本体或行业字典，即通过多维度对语义匹配词进行确定和抽取。

在句子结构匹配研究中SAO理论用于技术供需文本匹配研究较多。SAO的概念来自于发明问题的解决理论(TRIZ)[40]，KIM H,等人探讨了如何利用主谓宾式三元组Subject－Action－Object(SAO)来描述Problems与Solutions[41]。詹文青和肖国华通过抽取技术需求文本和专利文献的动宾结构[42]，即认为SAO结构中主语并不重要，进行了技术信息表示的研究。在技术文本匹配研究中，本体理论常与SAO理论结合应用，本体可以用来表达技术这个特殊领域的形式语言[43]，有利于获取领域知识，在文本处理多种方法都利用本体作为一种语义分析工具[44]。何喜军等人在构建领域本体的基础上，通过对技术供需文本信息进行SAO结构的语义提取，提出了技术供需多维语义结构匹配模型[45]，在模型中把技术供需信息基于SAO结构分为三部分，并用技术语料库和产业政策文件构建的领域本体来进行这三个部分的语义相似度计算，最后用熵值法对三个部分的匹配度进行加权，得到技术供需文本的语义匹配值，简化流程如图2所示。

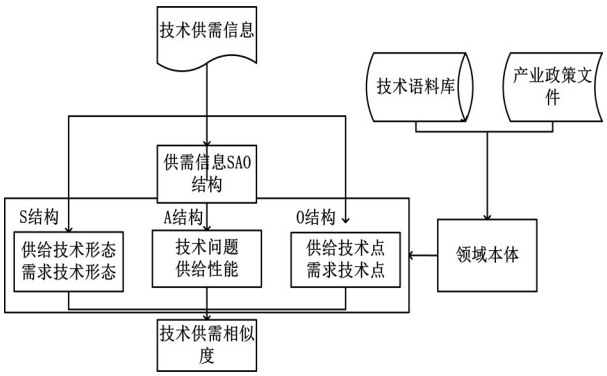

图2　基于本体和SAO结构的技术供需匹配流程

Figure 2　Technology supply and demand matching process based on ontology and SAO structure

除了基于SAO结构这种句法结构的研究外，还可以利用句子中词汇之间的依存关系将句法分析方法应用到技术供需文本匹配中。蒋振超等人借助依存句法关系和上下文关系对大规模无标记文本进行词向量的训练，并对比gram模型和CBOW模型，发现关系模型能更准确的表达词语的语义信息[46]。Ritam Dutt等人在依存句法分析的基础上结合词典来抽取需求文本和可用资源文本中的资源、数量、地理位置、来源、联系方式五类要素，识别资源需求信息和可用资源信息，并进行自动匹配[47]，其中供应资源、供应地理位置和资源数量识别抽取简要流程如图3所示。

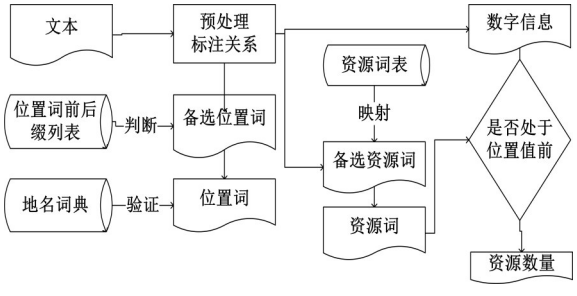

图3　依存句法供应资源识别

Figure 3　Identification of dependency syntax supply resources

其中，资源词由于其特殊性，只需要在抽取过程中利用好资源词典即可，而位置信息和资源数量信息需要用到词性

判断、地名词典验证以及判断是否位于资源词前等句子结构知识。此研究用于社交媒体上资源需求信息和可用供给信息的匹配，通过依存句法分析，对文本中相关关键词进行分析提取，并用词典进行验证，比单纯用领域本体或行业词典进行验证的准确率更高，如一个文本是数字文本，但不一定是资源数量。无论是SAO结构结合本体构建方法还是句子词汇关系模型结合词典研究都说明将依存句法分析结合其他文本方法应用于技术供需文本关键词识别与匹配有利于提高匹配准确率。

3.2 基于文本结构的匹配

从基于依存句法的文本表示与匹配研究中可以看出，部分研究还用到了供需文本的上下文关系，句子层面是对主谓宾细粒度的区分，文本层面是对文本整体信息的把握，对文本结构和上下文环境的分析能更有效率的进行文本匹配。如Fabio Benedetti 等人基于知识的上下文语义分析，提出CSA方法可以在考虑上下文环境下有效计算文本之间的相似度[48]。Zhu G 等人在语义相似性实体消歧研究时，采用了基于上下文和实体信息词语义模型，同时利用知识和语料库[49]。陈颖等人基于专利结构、语法、线索词特征的技术词、功效词的识别方法，提高了技术词和功效词的识别效果[50]。

在技术供需匹配研究上，有学者提出先对文本内容进行划分处理，依据一定的规则对技术供需文本结构进行分析，从而对信息进行提纯来进行供需匹配。杨德林等人在对在线技术转移平台供需匹配效率分析时，用TF-IDF和余弦相似度对中国技术交易信息服务平台上的供需信息进行匹配，发现服务平台上整体信息匹配度不高，并发现了是因为供需文本语言差异造成的。接着他们对文本结构进行解析，发现技术供需文本都可以分为技术信息和非技术信息两个部分，其中非技术信息的差异过大，影响了相似度的计算。随后采取了人工标注和机器学习的方法对技术供需文本中含技术信息的文本进行提纯再进行供需文本的相似度计算，研究简化流程如图4所示。

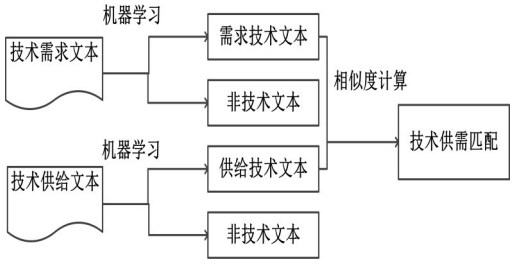

图4 技术文本划分匹配流程

Figure 4 Matching flow of technical text Division

研究结果表明提纯后的文本匹配度得到了提升，该研究思路为技术供需文本匹配提供了新的方向，即从文本结构上对选取关键信息进行匹配，排除非关键因素的噪声干扰，既避免了过高的计算量和过拟合又提高了匹配效率，在技术转移中具有实际应用价值。

4 基于深度学习的匹配

随着大数据和机器学习的发展得到了广泛的应用，深度学习有强大的特征表示学习能力，适合用于复杂的文本匹配，学者们利用深度学习算法进行图像、文本等数据处理的研究越来越多[51]，基于深度学习的技术文本匹配方法成为研究和应用的重要方法。深度学习方法与其他语义的方法，如基于关键词语义的方法区别在于，深度学习可以将整个文本都向量化，相比传统的只基于一个文档特征（如词频），能生成一个更高维度的向量，然后计算这个文本向量的相似度进行分类和匹配。

深度学习方法与传统文本匹配方法聚焦于选择合适的算法得到最优的匹配模型不同，深度学习文本匹配可以总结为利用深度学习方法进行文本向量化表达并进行分类的过程，并且聚焦于文本的表示研究。通常基于深度学习的技术文本匹配的语义表达就是词和文本的向量化，输入技术供需文本，经文本预处理为较小文本单元，输入到中间表示层，然后再连接输出文本表示向量用于计算相似度，目前深度学习生成文本表示用于文本相似度计算的中间表示层方法有基于全连接神经网络、基于卷积神经网络和基于循环神经网络等，基于深度学习的文本匹配框架如图5所示。

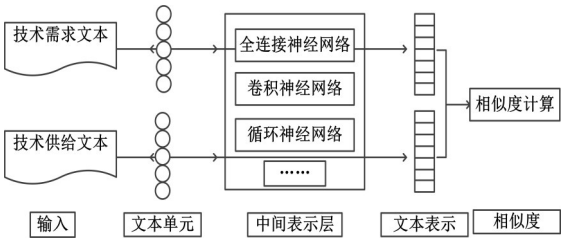

图5 基于深度学习的文本匹配框架

Figure 5 Text matching framework based on deep learning

4.1 全连接神经网络

全连接神经网络是深度学习的基础模型，在文本语义匹配应用中典型代表是深度结构语义模型[52]。深度结构语义模型提出时应用于查询与解答之间的匹配，把每个文本用五层网络进行向量化，目的是将一个没有处理过的高维的文本术语向量，经过模型输出为低维的语义特征向量，深度结构语义模型如图6所示。

在深度结构语义模型作为深度学习的中间层时，把各文档软化为术语向量后，由基于n-gram的词哈希方法对术语向量进行降维处理，然后再采用全连接方式对文本进行处理得到主题层面的特征向量（此模型定为128个维度），然后依次计算查询文本与各资源文本的相似度进行匹配。HUANG等人通过真实网站数据进行实验，证明了深度结构语义模型优于其他潜在语义模型[52]。但此模型在运用词哈希时，可能会出现产生错误，n-gram方法可能把两个不同的词表示

为相同的向量，而且全连接神经网络由于自身参数较多，缺乏对词语语序的考虑，在文本数量规模不大且复杂度不高时，可以考虑全连接神经网络生成文本向量。

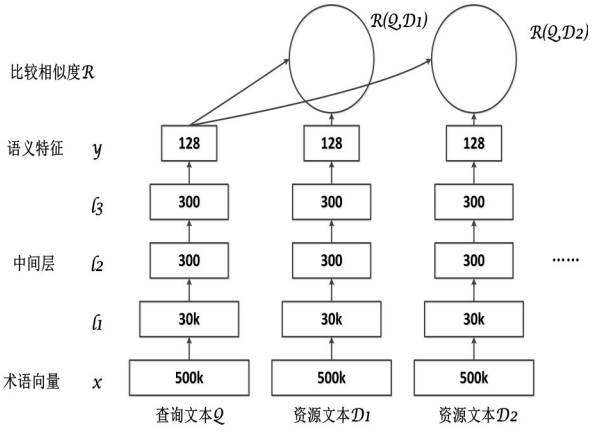

图6 深度结构语义匹配模型

Figure 6 Deep structure semantic matching model

4.2 卷积神经网络

卷积神经网络相比全连接神经网络可以减少参数，并且按序移动的卷积核（滑动窗口）使它考虑了词序特征。卷积神经网络一开始用于图像识别领域，后来Kim在句子分类中提出用卷积神经网络用于句子语义的表达[53]，学者开始提出了一些应用卷积神经网络进行文本匹配的研究，如Hu提出的ARC-I模型，用两个定长的向量表示两个句子；Qiu等人提出用张量神经网络（CNTN）来计算卷积神经网络生成的文本表示相似度[54]。其中运用的基础模型为基于单词序列的卷积深度语义结构模型（CDSSSM），相比于全连接的深度学习模型，把全连接层换成了卷积层。

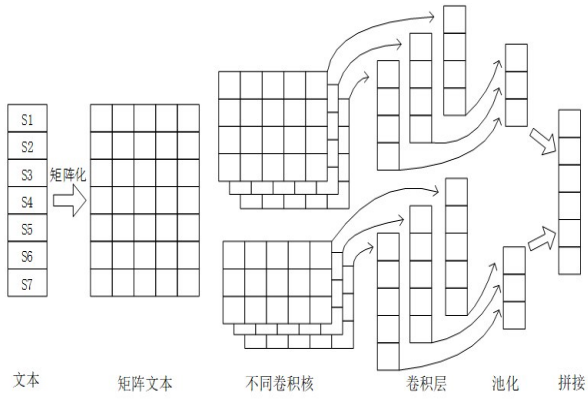

图7 基于卷积循环网络的文本表示

Figure 7 Text representation based on convolutional cyclic network

基于卷积的神经网络的文本匹配，先要对文本单元进行矩阵化，使文本的输入和图像处理的输入形式一致，然后用卷积核进行滑动（与文本矩阵进行哈达马积运算），得到多个特征图。多个卷积核能得到更多卷积特征图，多个卷积核包括不同形状和相同形状不同数值两种，矩阵文本和不同的卷积核运算能得到不同的卷积层的特征图，所以通常卷积层有很多个特征图。池化处理对不同位置的特征进行聚合统计，实现降维并防止过拟合，通常取向量中最大值或者均值来进行。经过多个卷积和池化过程后，对最终池化向量进行拼接即生成文本的向量表示。采用卷积循环网络为深度学习的中间层表示过程简图如图7所示。

卷积神经网络考虑了语序，在相关度方面相对全连接神经网络有一定的提升。但是对于句子中距离较远的关系无法表示，不能进行过于复杂的语义表示。适合较短的文本进行相似度计算，如果技术供需文本过长，可以对标题和简介进行卷积神经网络进行匹配，对于正文较长内容，则需要用到循环神经网络的文本表示方法。

4.3 循环神经网络

循环神经网络利用句子间的依存关系和知识网络等对长距离句子进行表示，典型的模型如基于长短时记忆（LSTM）文本模型。长短时记忆的循环神经网络在循环网络中加入了输入门、遗忘门和输出门三个门，来控制每个时刻的“记忆”，既可以保存当前时刻的信息，又可以保存比如之前某个时刻保存下来的“记忆”，从而保证模型能保留长距离的信息。有学者指出长短时记忆循环神经网络更偏向保存离当前位置更近的信息[55]，于是又有人提出双向的循环神经网络[56]，即从两个相反的方向对文本进行扫描，每个句子会得到两个不同的表示，最后计算两文本的交互张量，得到一个相似度矩阵。基于深层双向循环网络（Deep Bidirectional RNNs）的模型即把基础深度学习框架中的中间层方法选用深层双向循环网络结构，中间层模型如图8所示[57]。

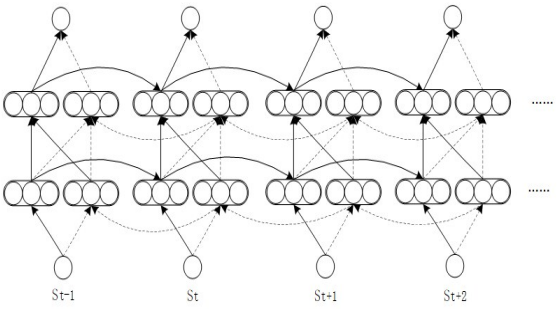

图8 深层双向循环神经网络结构

Figure 8 Deep bidirectional circulation neural network structure

在深层双向循环网络作为中间层时，最下面为划分的句子或词单元，中间是循环网络控制门结构，实线和虚线分别代表两个方向上得到的信息。在技术供需文本中发布者对部分信息描述不清时，用双向循环神经网络能从两个方面补足缺失值。在深层双神经网络的基础上，Google mind团队以人类注意力机制为启发使用了注意力模型，注意力模型在双向循环神经网络的基础上为每个词赋予了权重，即S_t和S_{t+1}的最后输出的文本表示权重不一定相同，以便能直观地反应每个词对于分类类别的重要性。起初模型是应用于图像分类，

后来Bahdanau等人将注意力机制应用到了文本处理中[58]，从而使技术文本处理也能应用此方法。在技术供需文本匹配中，注意力模型能很好的兼顾个别技术关键词的重要作用和整个文本的语义表示，由于不同权重的加入，促进了研究向多维度和多粒度匹配方向发展。多粒度神经网络，即从文本的不同粒度去对文本进行向量表示并计算相似度，如同时进行局部性文本表示和全局性文本表示的递归神经网络及其变体，这些方法能在一定程度上弥补深度学习过程中的信息损失问题。多维度匹配方法则是同时考虑多个表现因素，最后通过赋值不同的权重得到一个综合的文本匹配值。赵洪在运用深度学习的自动文摘研究中表明生成式的文本表示更接近人工摘要过程[59]，具有重要的研究意义，而从总体上看，在深度学习研究方向上，技术供需文本匹配用得较多的文本表示方法仍然属于抽取式，所以生成式方法的应用是未来基于深度学习匹配的另一个研究方向。

5 基于多维度视角的匹配

融合多维特征的技术文本供需匹配模型能有效解决供需文本中的特征稀疏和表述差异问题，从而提高供需匹配的准确性[2]。蒋振超等人用的词向量模型是基于词语关系上，结合了关键词、依存句法和深度学习三种方法来提高了词语的语义表达准确性[46]。谷重阳等人运用TF-IDF值建立文本的向量空间，用以解决传统基于关键词没有考虑语义信息问题，同时解决了词向量文本忽略了关键词在语料库中的分布问题[60]。Ferreira Rafael 等人基于语汇和句法，用语义相似度构建了一个识别系统[61]。荆琪等人以维基百科为语料库，通过词形、语序等句法结构和关键词权重来对句子相似度进行计算和特征扩展[62]。Liang Y 等人基于梯度文档生成，对混合主题建模和句子排序，提取文本中用户感兴趣的信息[63]。Xia X等人在错误识别研究中，提出了基于多维特征建立了主题模型(MTM)，扩展了LDA模型[64]。

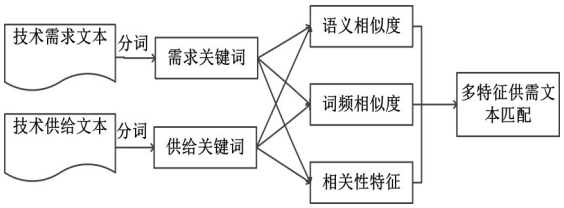

图9 技术供需文本多维度匹配模型

Figure 9 Multi dimension matching model of technology supply and demand text

何喜军等人在融合多特征进行技术供需文本匹配研究中，将技术供需文本的语义特征、词频特征和相关性特征三个维度融合建模，提高了在线技术转移平台的供需匹配效率[2]，其中语义特征是以专利数据库和维基百科作为基础语料库，用深度学习方法进行文档的向量化处理得到的，最后把三个维度计算得到的相似度用熵值法进行赋值得到多维度下的技术供需文本匹配值，多维度视角下技术供需文本匹配模型流程如图9所示。

在该研究中，虽说有从多个维度对技术供需文本进行匹配，但主体仍以关键词为基础，缺少对文本结构的考虑，深度学习的应用也只是基础的文本向量化，准确来说研究是从多维视角对技术供需文本关键词进行了匹配。同样他们在对专利技术交易主体的研究中，对交易主体关系、个体属性和交易网络等多维度建立了预测模型[65]，在这个模型的多维度中考虑是有关主体的多维度。此外该多维度文本研究除了从多维度视角对供需文本关键词进行匹配外，还基于文本匹配的结果对在线技术转移服务平台的匹配效率进行分析，在技术供需领域和技术供需主体区域邻近性进行了研究，得出了技术供需匹配中领域交叉特征不明显与地理邻近性特征明显的结论，即从文本的多维特征进行了供需匹配的结果中探索了除供需文本外其他影响因素对技术转移供需匹配的影响，这也为未来的多维度匹配提出了方向。李志义等人在表示学习的检索模型和特征抽取研究中表明，现有研究仍以单模态研究为主，跨模态研究仅限于图文对齐[66]，可以看出多维度的研究受限于来源数据的不同模态影响，未来在研究多维度匹配方向上，多来源跨模态信息的利用也是重点研究对象之一。

除了直接用各种方法融合外，考虑了时序的不同方法的先后使用也是一种多维度的匹配方法，如在庞亮等人提出的直接建模匹配方法中，先考虑关键词是否匹配，再考虑关键词相对位置是否匹配，最后分析整个文本语义是否匹配，综合打分得出句子的匹配值[13]。学者们在各种方法研究中都或多或少的用到了其他方法，如基于关键词的语义匹配中，要用到语料库以及词向量等，词向量的生成涉及到了深度学习的方法，而在词形语序以及循环神经网络中，也都离不开对依存句法的研究。综合各种方法能克服或减弱单一方法缺点带来的影响。

6 结 语

从传统基于关键词的研究到现在主流的基于深度学习的匹配研究，首先可以发现技术文本匹配并不是一个单纯研究方向上的转换，而是不断的在已有的方法上加入其他方法来提供匹配效果；其次后面深度学习的研究也并不是不注重前面的研究，而是在研究过程中利用好关键词、依存句法、上下文和特征工程等的基础上，对传统方法中的一些步骤进行改进，以达到更好的匹配效果。

目前无论是文本匹配还是技术文本匹配的研究都已经在向着多维度的方向发展，从更高的视角观察，随着计算能力的提升，人类决策的依据大多是从多个维度进行权衡，评价指标的设计是尽可能的考虑得更全面，多维度视角融合是未来的趋势。在技术供需文本匹配研究中，技术供需信息的匹配，并不一定是严格计算文本相似度，因为还有各个主体间的供需关系、对应关系，单纯计算文本相似度并不一定能很好的进行匹配，后续研究应当考虑其他因素，并且在跨模

态和可解释性做出研究。现有研究虽然有从多维度视角去考虑匹配问题，但是大多还是限于文本信息，如从文本的关键词的词频和语义，此外，有学者指出现有研究仅使用技术供需单一数据源，并不能全面对技术进行供需匹配研究[67]。而已有学者证实了地理邻近性特征明显[2]，领域交叉性不明显等其他因素对技术供需匹配的影响，但总体上在技术文本供需匹配时考虑供需主体的合作关系、地理邻近性和政策影响等其他因素的研究还是较少。技术转移的实现是一个受多方因素影响的过程，而这一过程应当与技术交易双方主体的决策过程一致，尽可能的考虑更多的因素，然后做合理的取舍，会比直接不考虑一部分因素更有说服力，从而提高技术供需匹配推荐的效率。

参考文献

1 喻昕. 技术市场信息不对称问题研究[J]. 情报科学,2011,29(4):515-519.

2 何喜军,马珊,武玉英,等. 多特征融合下在线技术转移平台供需匹配研究——以京津冀区域数据为例[J]. 情报杂志,2019,38(6):174-181.

3 LIU Y,LI K W. A two-sided matching decision method for supply and demand of technological knowledge [J]. JOURNAL OF KNOWLEDGE MANAGEMENT,2017,21(3):592-606.

4 薛伟贤,田鹏,孙姝羽. 战略性新兴产业技术供需协同研究:以陕西为例[J]. 科研管理,2016,37(S1):515-524.

5 郭韧,陈福集. 知识供需匹配的研究综述[J]. 情报理论与实践,2013,36(12):114-118.

6 LICHTENTHALER U,ERNST H. External Technology Commercialization in Large Firms: Results of a Quantitative Benchmarking Study[J]. R & D Management,2007,37(5):383-397.

7 LI C,LAN T,LIU S J. Patent attorney as technology intermediary: A patent attorney-facilitated model of technology transfer in developing countries[J]. World Patent Information,2015,(43):62-73.

8 ALBERS A,BURSAC N,MAUL L,et al. The Role of In-house Intermediaries in Innovation Management-Optimization of Technology Transfer Processes from Cross-industry [J].Procedia Cirp,2014,(21):485-490.

9 科技部火炬中心技术市场管理处. 2019年技术市场统计年度报告[R].2019.

10 BENGIO Y,DUCHARME R,VINCENT P,et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research,2003,3(2):1137-1155.

11 陈二静,姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现,2017,(6):1-11.

12 HARRIS Z S. Papers in Structural and Transformational Linguistics[M]. Netherlands:Springer Netherlands,1970.

13 庞亮,兰艳艳,徐君,等. 深度文本匹配综述[J]. 计算机学报,2017,40(4):985-1003.

14 熊则见,杨敏,赵雯. 高技术产品研发关键成功因素的文献计量分析[J]. 科研管理,2011,(10):38-47.

15 邓小龙,李欲晓. 面向应急管理的大图重要节点中介度高效近似计算方法[J]. 系统工程理论与实践,2015,35(10):89-101.

16 谢玮,沈一,马永征. 基于图计算的论文审稿自动推荐系统[J]. 计算机应用研究,2016,33(3):798-801.

17 Rahman M M ,Roy C K . TextRank based search term identification for software change tasks[C]//IEEE International Conference on Software Analysis. IEEE,2015:540-544.

18 SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing & Management,1988,24(5):513-523.

19 Kuncoro B A ,Iswanto B H . TF-IDF method in ranking keywords of Instagram users´ image captions[C]// 2015 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE,2016:1-5.

20 ZHENG Y, MENG Z, XU C. A Short-Text Oriented Clustering Method for Hot Topics Extraction[J]. INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 2015, 25(3):453-471.

21 HE F,HE Y,LIU N,et al. A microblog short text oriented multi-class feature extraction method of fine-grained sentiment analysis[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2014,(50):48-54.

22 GUOWAN H,JIE W,YAFENG Z,et al. Keyword extraction of web pages based on domain thesaurus[C]//2014 IEEE 3rd International Conference on Cloud Computing and Intelligence Systems,2014:48-54.

23 杨德林,夏青青,马晨光. 在线技术转移平台的供需匹配效率分析[J]. 管理科学,2017,30(6):104-112.

24 KIM H G,LEE S,KYEONG S. Discovering hot topics using Twitter streaming data: social topic detection and geographic clustering[C]//Proceedings of the 2013 IEEE / ACM International Conference on Advances in Social Networks Analysis and Mining,2013:1215-1220.

25 王立霞,淮晓永. 基于语义的中文文本关键词提取算法[J]. 计算机工程,2012,38(1):1-4.

26 梅家驹,竺一鸣,高蕴琦,等. 同义词词林[M]. 上海:上海辞书出版社,1993.

27 刘端阳,王良芳. 结合语义扩展度和词汇链的关键词提取算法[J]. 计算机科学,2013,40(12):264-269.

28 任龙,姜学民,傅晓晓. 基于专利权转移的中国区域技术流动网络研究[J]. 科学学研究,2016,34(7):993-1004.

29 方俊,郭雷,王晓东. 基于语义的关键词提取算法[J]. 计算

机科学,2008,35(6):148-151.
30 Gang L ,Qiangbin D ,Quan W .A new approach to compute semantic relevance of Chinese words[C]//IEEE 2010 International Conference on Artificial Intelligence and Education,2010:610-613.
31 WEI T,LU Y,CHANG H,et al. A semantic approach for text clustering using WordNet and lexical chains[J]. Expert Systems with Applications An International Journal,2015,42(4):2264-2275.
32 ZONGDA,WU,HUI,et al. An efficient Wikipedia semantic matching approach to text document classification[J]. Information Sciences,2017,(393):15-28.
33 JIANG Y,BAI W,ZHANG X,et al. Wikipedia-based information content and semantic similarity computation[J]. Information Processing & Management,2017,53(1):248-265.
34 何喜军,张婷婷,武玉英,等.供需匹配视角下基于语义相似聚类的技术需求识别模型[J].系统工程理论与实践,2019,39(2):476-485.
35 刘翔,李彦,李文强,等.构建支持产品创新设计的专利知识库[J].机械设计与研究,2010,26(6):7-11.
36 梁浩.面向创新设计的语义知识表示方法的研究与应用[D].哈尔滨:东北林业大学,2014.
37 张娴,胡正银,茹丽洁,等.专利技术供需信息关联知识组织模式研究[J].图书情报工作,2016,60(8):118-125.
38 MIKOLOV T,CHEN K,CORRADO G S,et al. Efficient Estimation of Word Representations in Vector Space[C]. Proceedings of Workshop at ICLR,2013.
39 JOULIN A,GRAVE E,BOJANOWSKI P,et al. Bag of Tricks for Efficient Text Classification[J]. https://arxiv.org/pdf/1607.01759v2.pdf,2016.
40 杜玉锋,季铎,姜利雪,等.基于SAO的专利结构化相似度计算方法[J].中文信息学报,2016,30(1):30-35.
41 KIM H,HYEOK Y,KIM K. Semantic SAO network of patents for reusability of inventive knowledge[C]. 2012 IEEE 6th International Conference on Management of Innovation and Technology,ICMIT,2012:510-515.
42 詹文青,肖国华.面向技术需求的潜在技术转移专利识别[J].情报理论与实践,2019,42(5):117-121.
43 张沪寅,温春艳,刘道波,等.改进的基于本体的语义相似度计算[J].计算机工程与设计,2015,36(8):2206-2210.
44 WEI G,YUN G,ZHU L. Ranking based ontology learning algorithm for similarity measuring and ontology mapping using representation theory[J]. Journal of Information & Optimization Sciences,2016,37(2):303-320.
45 何喜军,马珊,武玉英.基于本体和SAO结构的线上技术供需信息语义匹配研究[J].情报科学,2018,36(11):95-100.
46 蒋振超,李丽双,黄德根.基于词语关系的词向量模型[J].中文信息学报,2017,31(3):25-31.
47 DUTT R,BASU M,GHOSH K,et al. Utilizing microblogs for assisting post-disaster relief operations via matching resource needs and availabilities[J]. Information Processing & Management,2019,56(5):1680-1697.
48 BENEDETTI F,BENEVENTANO D,BERGAMASCHI S,et al. Computing inter-document similarity with Context Semantic Analysis[J]. Information Systems, 2019, 80(FEB.):136-147.
49 ZHU G,IGLESIAS C A. Exploiting semantic similarity for named entity disambiguation in knowledge graphs[J]. Expert Systems with Applications,2018,101(JUL.):8-24.
50 陈颖,张晓林.专利中技术词和功效词识别方法研究[J].现代图书情报技术,2011,(12):24-30.
51 吕正东,李航.深度匹配学习在语言匹配中的应用[J].中国计算机学会通讯,2015,11(8):30-37.
52 HUANG P,HE X,GAO J,et al. Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data: CIKM '13[C]//New York,NY,USA,2013.
53 KIM Y. Convolutional Neural Networks for Sentence Classification[J]. Computer Science,2014,3(9):1746-1751.
54 QIU X,HUANG X. Convolutional Neural Tensor Network Architecture for Community-Based Question Answering[C]//IJCAI'15,2015:1315-1311.
55 BAHDANAU D,CHO K,BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. ArXiv,2014,1409.
56 GRAVES A,SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural networks :the official journal of the International Neural Network Society, 2005, (18):602-610.
57 GRAVES A,MOHAMED A,HINTON G. Speech Recognition with Deep Recurrent Neural Networks[C]. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings,2013.
58 BAHDANAU D,CHO K,BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate[C]//ICLR,2015.
59 赵洪.生成式自动文摘的深度学习方法综述[J].情报学报,2020,39(3):330-344.
60 谷重阳,徐浩煜,周晗,等.基于词汇语义信息的文本相似度计算[J].计算机应用研究,2018,35(2):391-395.
61 FERREIRA R,CAVALCANTI G D C,FREITAS F,et al. Combining sentence similarities measures to identify paraphrases[J]. Computer Speech & Language, 2018, 47(JAN.):59-73.
62 荆琪,段利国,李爱萍,等.基于维基百科的短文本相关度计算[J].计算机工程,2018,44(2):197-202.

63 LIANG Y,LIU Y,CHEN C,et al. Extracting topic-sensitive content from textual documents—A hybrid topic model approach[J]. Engineering Applications of Artificial Intelligence, 2018,(70):81-91.
64 XIA X,LO D,DING Y,et al. Improving Automated Bug Triaging with Specialized Topic Model[J]. IEEE Transactions on Software Engineering,2016,(43):1.
65 何喜军,董艳波,武玉英,等. 基于ERGM的科技主体间专利技术交易机会实证研究[J]. 中国软科学,2018,(3):184-192.
66 李志义,黄子风,许晓绵. 基于表示学习的跨模态检索模型与特征抽取研究综述[J]. 情报学报,2018,37(4):422-435.
67 张婷婷. 基于多源数据的企业潜在技术需求点识别[D]. 北京:北京工业大学,2018.

(责任编辑:徐 波)

A Review of Supply and Demand Matching of Online Technological Text

YU Hui, LIAGN Zhen-tao, XIE Hao

(Center for Studies of Information Resources, Wuhan University, Wuhan 430072*,China)*

Abstract:【Purpose/significance】With the development of network information technology and national policy support for technology transfer, a large number of online technology transaction needs arise. As the intermediary of technology supply and demand transaction, a large number of text information of technology supply and demand can be released on the platform. Improving the efficiency of text matching of technology supply and demand is helpful to improve the success rate of technology transaction and promote technology transfer.【Method/process】Based on the analysis of traditional text matching methods, this paper combs the current situation of online technology supply and demand text matching methods from four directions: keyword based matching method, syntactic analysis and text structure based matching method, deep learning based matching method and multi-dimensional perspective based matching method.【Result/conclusion】The results show that most of the studies have integrated a variety of matching methods, and matching from a multi-dimensional perspective is the research trend. In the future research of technology supply and demand text matching, in addition to continuing to integrate deep learning methods into existing methods, researchers should also improve the efficiency of technology supply and demand text matching from the directions of multi-dimensional, cross modal and interpretable.【Innovation/limitation】This paper summarizes the technology text matching method, which can provide reference for the technology text matching method, but technology matching should also consider other factors affecting technology transfer in reality.

Keywords: technology transfer; technological text for supply and demand; textrepresentation; supply and demand matching; bag of words; technical recommendation