融合多源信息的专家学术专长语义匹配方法

谢小杰[1),2),3)] 梁 英[1),3)] 王梓森[1),2),3)] 刘政君[1),2),3)]

[1)](中国科学院计算技术研究所泛在计算系统研究中心 北京 100190)

[2)](中国科学院大学计算机科学与技术学院 北京 100049)

[3)](移动计算与新型终端北京市重点实验室 北京 100190)

摘 要 通过评议文档与专家库的专家学术专长匹配,可以输出领域相关的候选专家列表,是同行评议中专家遴选和专家推荐的重要参考依据.针对学术专长匹配存在语义鸿沟、无法反映专家和评议文档多源信息间语义关联的问题,首先对专家信息和评议文档的多源信息进行语义特征抽取,融合多类特征进行表示学习,利用卷积神经网络设计专家特征抽取器ExpFeat和评议特征抽取器RevFeat,采用词嵌入方法和注意力机制对专家专长标签、评议文档关键词、学科分类树语义特征进行抽取和融合,生成具有"小同行"特征的专家和评议文档语义特征向量表示,解决多类信息源间不同学术分类标准造成的语义差异,反映内在语义联系,利用低维稠密向量表达语义信息,降低匹配复杂度.然后,根据专家语义特征表示和评议文档语义特征表示进行学术专长语义匹配,将专家和评议文档特征向量映射到相同语义空间,计算向量间余弦相似度衡量语义相似性,引入负例专家进行模型训练,通过softmax函数计算最大化正例专家概率优化特征提取器参数,进一步提升语义差异的捕捉能力,解决专家信息和评议文档之间的语义鸿沟问题,提升专家匹配效果.最后,在开源的论文评审数据集和项目评审数据集上进行了实验对比和实例分析,结果表明,本文所提方法可以有效提升专家匹配精准度.

关键词 特征抽取;语义匹配;多源信息;专家推荐;同行评议

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2023.01045

Multi-Source Information Fused Academic Expertise Semantic Matching Method

XIE Xiao-Jie[1),2),3)] LIANG Ying[1),3)] WANG Zi-Sen[1),2),3)] LIU Zheng-Jun[1),2),3)]

[1)](*Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 100049)

[3)](*Beijing Key Laboratory of Mobile Computing and New Devices, Beijing* 100190)

Abstract By matching the review documents with the academic expertise of experts in the expert database, the list of candidate experts related to the field in peer review can be output, which is an important reference basis for expert selection and expert recommendation. Aiming at the problem of semantic gap and unable to reflect semantic association between multi-source information of experts and review documents in academic expertise matching, firstly, multi-source semantic features are extracted from expert information and evaluation documents, and multi-source features are extracted and fused for representation learning. The semantic expression ability of the model can be supplemented by fusing multi-source information, so that the model can fully extract the feature information of experts and review documents for reasonable correlation matching. Convolution neural network is used to design expert feature extractor ExpFeat and review feature extractor RevFeat, and word embedding method and attention mechanism is used to label expert expertise, review document keywords The semantic features of subject classification tree are extracted and

收稿日期:2021-11-22;在线发布日期:2022-09-21.本课题得到国家重点研发计划项目(2018YFB1004700)资助.**谢小杰**,硕士,主要研究方向为数据挖掘.E-mail:mailbox_of_xxj@126.com.**梁 英**(通信作者),高级工程师,主要研究方向为数据挖掘.E-mail:liangy@ict.ac.cn.**王梓森**,硕士研究生,主要研究方向为数据挖掘.**刘政君**,硕士,主要研究方向为数据挖掘.

fused to generate the semantic feature vector representation of expert and review documents with small peers, by separately modeling the characteristics of "small peers", the academic association information between projects and experts can be captured in a more granular way, solve the differences of academic classification standards among multiple information sources, and reflect the internal semantic connection. In addition, different source information is fused in the feature matching stage, and the original features of different source information are retained as much as possible through vector concatenating operation, so that the model can more fully capture the data distribution differences between information sources, and use low dimensional dense vectors to express semantic information to reduce the matching complexity. Then, the academic expertise matching degree is designed, the academic expertise semantic matching is carried out according to the expert semantic feature representation and the review document semantic feature representation, the expert and review document feature vectors are mapped into the same semantic space, and the cosine similarity between vectors is calculated to measure the semantic similarity. Through end-to-end learning of the neural network, the trainable parameters can converge to the global optimal direction, and the semantic correlation between expert information and review documents can be more accurately captured. And negative case experts are introduced for model training and the expert probability of maximizing positive cases is calculated by softmax function to optimize model parameters and improve the ability to capture semantic differences, so as to solve the semantic gap between expert information and review documents and improve expert matching effect. Finally, experimental comparison and analysis are carried out on the open-source paper review data set. The proposed method can effectively improve the accuracy of expert matching. And when the top$K$ value is 10, compared with the comparison method, the proposed method increases the hit rate by 9% and the gain rate by 10% in the paper review data set, and increases the hit rate by 9% and the gain rate by 6% in the project review data set, which verifies the effectiveness of the proposed method.

Keywords feature extraction; semantic matching; multi-source information; expert recommendation; peer review

1 引 言

学术数据由个人信息、学术成果、学术行为等信息构成,具有规模庞大、异质性强等特点,隐含了丰富的专家学术专长信息.利用大规模学术数据中的多源信息进行特征抽取,能够挖掘和识别出更加客观、精准、明确的专家学术特征,在补充专家库信息的同时,提升学术专长匹配效果.

专家学术专长匹配作为专家推荐的重要环节,通过综合考虑专家信息和评议文档内容,为专家推荐提供一批领域相关的候选专家,其匹配结果直接影响专家推荐和同行评议的质量,不恰当的匹配结果可能为同行评议带来严重的学术隐患.

"小同行"是保证科技评价科学性的基础,体现同行评议中学术同行的真正要义.由于学术数据是多源的,可以来自短文本、关键字标签、学科分类树等不同体系,例如:专家短文本为专家主要学术经历的文本简述,中图分类号包含论文的核心学科信息.因此,合理抽取专家信息和评议文档多源特征,有利于准确匹配领域相关的"小同行"专家,提升专家推荐和同行评议的效果.

现有学术专长匹配往往通过关键字搜索的方式检索专家库中的指定专家,缺少利用多源语义信息进行匹配的软匹配策略.例如专家短文本包含"机器学习"领域的专家和学科分类主题为"专家系统/知识工程"的评议文档之间虽然存在着潜在的语义关联,但在传统关键字检索方式下,由于两者存在语义鸿沟而无法有效匹配.

主题建模和语义匹配方法为跨越语义鸿沟提供

了有效途径，然而不同的信息源常属于不同的学术分类体系，在匹配过程中仍需要考虑分类标准不同所带来的语义信息差异，例如，专家近期发表过多篇中图分类号为“TP391([工业技术]-[自动化技术、计算机技术]-[计算技术、计算机技术]-[计算机的应用]-[信息处理])”且研究内容为“用户画像、推荐引擎”的论文，那么对于学科分类为“[图书馆、情报与文献学]-[情报学]-[情报检索学]”且涉及“知识检索系统”的项目而言，这位专家是密切相关的，但因为分类标准的不同，在传统的匹配模式下，专家和项目间会因上述分类词汇的差异存在一定的语义偏差. 并且随着近年来学术领域的飞速发展，跨学科领域研究和大量新兴专业术语的出现为多源信息的融合带来新的问题和难点. 因此，为了在同行评议中精准推荐领域相关的高质量专家，仍面临如下挑战：

(1) 如何深度挖掘专家和评议文档多源信息间语义关联，进一步提升匹配精度. 目前已有语义匹配方法大多基于专家和评议文档的短文本信息进行匹配，但学术信息是多源的，多类信息源间缺少统一的学术分类标准和规范，难以直接建立语义联系；不同类的信息源常包含着不同的特征，研究深层语义特征提取、深度挖掘语义信息关联，有助于提升匹配效果.

(2) 如何支持“小同行”评审专家推荐，确保评审专家的学术鉴别能力. 现有方法大多关注专家和评议文档间语义级别的关联匹配，较少考虑更细粒度的“小同行”专家特征，随着学科之间相互交叉渗透和新学科不断产生，在同行评议中需要进一步区分大学科与分支学科，而“小同行”专家是保证科技评价科学性的基础，可为同行评议提供更准确的决策支持.

针对上述挑战，本文提出了一种融合多源信息的学术专长语义匹配方法 ExpRec，通过特征表示抽取和语义特征匹配 2 个步骤建立专家信息和评议文档多源学术信息的内在语义联系，进一步提升学术专长匹配的精准度. 在开源论文评审数据集和真实项目评审数据集上进行实验对比和分析，验证了本文所提方法的有效性.

本文主要贡献包括：

(1) 提出一种多源语义特征抽取及融合表示方法. 针对多类信息源因分类标准不同产生的语义差异性，设计专家特征抽取器和评议文档特征抽取器，使用卷积神经网络和注意力机制抽取融合学术特征，将多类信息源映射到相同语义空间，并为“小同行”专家推荐提供支持.

(2) 提出一种学术专长语义精准匹配方法. 针对专家学术专长和评议文档匹配的泛化性，引入负例专家，基于深度语义匹配模型(Deep Structured Semantic Model，DSSM)框架训练特征抽取器，并通过 softmax 函数计算最大化正例专家概率，优化特征抽取器参数，增强对专家学术专长语义差异捕捉能力，提升匹配精准度.

2　相关工作

在专家和评议文档匹配任务中，为利用语义信息提升匹配效果，主题建模方法和语义匹配方法常用于对专家和评议文档间的语义特征进行捕捉.

主题建模方法利用主题模型提取专家文本和评议文档的主题信息，生成专家和评议文档的主题向量，并通过向量距离建模相关性和匹配度[1]. Maryam 等人[2]考虑了论文包含多个主题的情况，利用潜在语义分析(Latent Semantic Analysis，LSA)[3]主题模型和互信息聚类提取专家和论文的主题进行匹配，确保匹配到的评审专家能够覆盖所有主题. Stelmakh 等人[4]设计了一个基于增量最大流过程的评议文档分配算法，通过设定统计准确性目标实现同行评议中评议文档的公平性分配. Li 等人[5]将主题建模方法与学术网络相结合，通过利用协作距离和主题相似度对论文和评审专家间的匹配程度进行建模. Peng 等人[6]提出了一种感知时间和主题的评审专家分配模型，利用潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)[7]主题模型提取专家在特定时间段内的主题分布，结合 TFIDF 等统计特征提升专家与待评审论文之间的匹配度. Jin 等人[8]考虑专家研究兴趣的变化，利用作者主题模型(Author Topic Model，ATM)[9]对专家、研究主题和学术论文进行统一建模，并通过负相对熵估计专家和待评审论文之间的领域相关性. Anjum 等人[10]设计专家推荐系统 PaRe，通过专家数据收集和评审专家分配完成专家推荐任务，同时提出了一种共同研究领域的主题模型，以计算专家和待评审论文之间的研究领域重叠度. Charlin 等人[11]研发了一种论文匹配系统，利用 LDA 主题模型提取专家和评审文档的主题向量，并通过向量点积和线性回归计算专家分数，推荐得分高的专家作为匹配结果. Tan 等人[12]使用 LDA 主题模型和改进的语言模型提取并融合评审专家的词级主题特征，并基于归一化增益率度量专家和评审稿件的匹配程度. 由于主题模

型主要以词语为单位对专家和评议文档主题信息进行构建，当出现一词多义时，主题建模无法很好理解其中的共现信息[13]，从而导致对语义信息捕捉不够充分，让依赖于词的主题匹配方法难以收敛.

语义匹配方法对专家和评议文档的语义特征信息进行抽取，生成特征向量表示并计算语义匹配度，以获取专家和评议文档的领域相关性. Ogunleye 等人[14]使用 word2vec[15]词嵌入方法对文本语料进行预训练、生成词向量表示、度量专家和论文的匹配程度. 何柔萤等人[16]基于卷积神经网络提取文本特征表示，引入注意力机制提升表示精度，经向量语义度量完成相似度匹配，实现专家排序，以优化专家推荐质量. Duan 等人[17]基于预训练语言模型提取专家文本和评议文档的语义特征表示，利用文本中的语义信息捕捉专家与评议文档的深层语义关联. Huang 等人[18]提出了一种深度语义匹配模型 DSSM，通过单词哈希扩展语义模型容量，使用多层全连接层提取文本特征表示，并利用余弦相似度和 softmax 函数构建语义匹配层，提升了文本的匹配效果. Shen 等人[19]基于 DSSM 提出了卷积深度语义匹配模型(Convolutional Neural Network Deep Structured Semantic Model，CNN-DSSM)，通过 n-gram 和卷积神经网络(Convolutional Neural Network，CNN)提取文本特征表示，进一步捕捉了文本中的高维语义特征. Devlin 等人[20]提出了一种双向预训练语言模型(Bidirectional Encoder Representations from Transformers，BERT)，使用大规模文本语料训练掩码语言模型，预测下一语句概率，从而提升文本语义抽取能力和语义匹配准确度. Kou 等人[21]从专家发表的论文中提取领域知识生成专家表示向量，通过计算论文与专家表示向量间相似度实现专家匹配. Zhang 等人[22]使用 BERT 对文本进行全局编码，并使用 CNN 从本地角度捕捉关键语义信息，提取文本中深度语义关系. 与主题建模相比，语义匹配方法可以基于单词共现对语义信息建模，充分捕捉专家和评议文档间语义关联，因此本文基于语义匹配方法，通过融合专家和评议文档的深度语义特征信息，计算专家和评议文档间的语义相似性，以提升学术专长匹配的精准度. 除此之外，现有的语义匹配方法对多类信息源的语义匹配关注不足，且在特征提取时未考虑“小同行”评审专家特征，故仍需挖掘多源信息间的语义联系，捕捉更细粒度的“小同行”专家特征，从而确保评审专家的学术鉴别能力.

3 方法概述

3.1 问题描述

本文的专家信息特指用来描述专家学术专长的相关内容，共包含 2 类信息源，分别为专家短文本和学术专长标签集合. 其中专家短文本包括专家主要研究方向、科研教育经历、学术成果产出等学术信息，而学术专长标签常指代描述专家主要研究领域的学术短语标签.

评议文档由与被评议内容相关的若干文档构成，按照不同场景需求可以划分为：(1)项目评审场景的项目指南；(2)论文审稿场景的学术论文；(3)科技奖励场景中的申请书；(4)成果孵化场景的科研成果；(5)需求对接场景的需求说明书；(6)技术招标场景的招投标文件. 虽然不同场景需求下评议文档信息不同，但都可以通过短文本、学科分类路径和关键词集合三类信息源进行描述. 一般情况下，评议文档短文本概括了评议文档的学术主题信息，学科分类路径描述了评议文档所属的学科领域信息，关键词集合表达了评议文档的核心学术内容. 以中文期刊论文审稿场景为例，短文本指代学术论文的主题摘要，学科分类路径常指代中图分类法中根节点到论文中图分类号的学科路径，关键词集合由论文关键词构成或从论文正文中抽取得到.

专家信息和评议文档的每类信息源隶属于不同的分类体系，拥有独立的学术分类标准，不能直接进行学术专长匹配，但每类信息源间都存在语义关系，学术专长语义匹配的主要目的是捕捉专家信息和评议文档信息的内在语义关联，挖掘并融合不同类信息的关联语义特征，体现信息间的相关性. 对专家和评议文档的多源信息进行特征表示和语义匹配，有利于筛选领域相符的候选专家，并能够为细粒度的“小同行”专家推荐提供更加丰富的学术特征支持. 假设某专家的短文本描述了“机器学习”相关的科研经历，学术专长标签集合包含了“机器学习”、“信息检索”、“知识图谱”等学术专长标签，则对于摘要主题为“推荐系统”、学科分类路径为[工业技术]-[自动化技术、计算机技术]-[计算技术、计算机技术]-[计算机的应用]-[信息处理]、关键词集合包含“推荐系统”、“用户画像”的待评审论文而言，该专家是语义匹配的，且属于“小同行”专家. 但因为语义鸿沟问题，若直接使用关键字简单搜索的方式，则上述专家和评议文档无法得到有效匹配.

本文首先利用专家和评议文档的多源信息，通过专家特征抽取器和评议文档特征抽取器分别抽取专家特征和评议文档特征，并基于注意力机制融合“小同行”专家信息，生成语义特征表示，然后根据专家特征表示和评议文档特征表示计算余弦相似度，进行语义匹配及模型优化，建立专家信息和评议文档信息之间的语义关联，为专家遴选提供领域相关的候选专家列表，以提升同行评议推荐专家的学术匹配精准性.

3.2　概念定义

本文所提方法对专家信息和评议文档信息进行语义特征提取以实现专家和评议文档间的匹配，而专家信息和评议文档信息主要由短文本、标签、学科分类路径和关键词等构成，本节将对所涉及到的基本概念进行介绍.

专家集合 *EXP*. 专家集合由专家库中的所有专家构成，即 $EXP=\{exp_i \mid 1\leq i\leq |EXP|\}$，其中 exp_i 表示第 i 个专家，$|EXP|$ 表示专家数.

专家短文本 $exptxt_i$. 表达了专家知识技能、研究领域或科研经历等信息. 专家 exp_i 的短文本被定义为一段与专家相关的简短文本 $exptxt_i$.

学术专长标签 $expkey_{i,l}$. 为若干单词组成的短语标签，表达了专家的知识技能、研究领域等信息. $expkey_{i,l}$ 用来描述专家 exp_i 的第 l 个学术专长标签.

学术专长标签集合 $EXPKEY_i$. 专家 exp_i 的学术专长标签集合由一系列学术专长标签构成，即 $EXPKEY_i=\{expkey_{i,l} \mid 1\leq l\leq |EXPKEY_i|\}$，$|EXPKEY_i|$ 表示专家 exp_i 的学术专长标签数.

评议文档集合 *REV*. 评议文档集合由一系列评议文档构成，即 $REV=\{rev_j \mid 1\leq j\leq |REV|\}$，其中 rev_j 表示第 j 篇评议文档，$|REV|$ 表示评议文档数.

评议文档短文本 $revtxt_j$. 概括描述了评议文档的内容、主旨和主要细节等信息. 评议文档 rev_j 的短文本被定义为一段与评议文档相关的简短文本 $revtxt_j$.

学科分类路径 $SPATH_j$. 学科分类树由多个存在从属关系的学科节点构成，在学科分类树中，每一棵子树的根学科节点为其余学科节点的上级学科. 评议文档 rev_j 的学科分类路径为学科分类树中学科根节点到评议文档 rev_j 所属学科节点的学科路径 $SPATH_j=spath_{j,1}-\cdots-spath_{j,s}-\cdots-spath_{j,S}$，其中，$S$ 为学科分类路径 $SPATH_j$ 的学科数，$spath_{j,s}$ 表示学科分类路径 $SPATH_j$ 的第 s 个学科节点，$s\in[1,S]$.

关键词集合 $REVKEY_j$. 概括了评议文档的关键内容. 评议文档 rev_j 的关键词集合由一系列关键词构成，即 $REVKEY_j=\{revkey_{j,w} \mid 1\leq w\leq |REVKEY_j|\}$，其中，$revkey_{j,w}$ 表示评议文档 rev_j 的第 w 个关键词，$|REVKEY_j|$ 表示关键词数.

3.3　整体流程

本方法通过特征抽取与表示和语义特征匹配两个步骤捕捉专家信息和评议文档信息的内在语义联系，进一步提升学术专长匹配的精准度. 方法整体流程参见图 1.

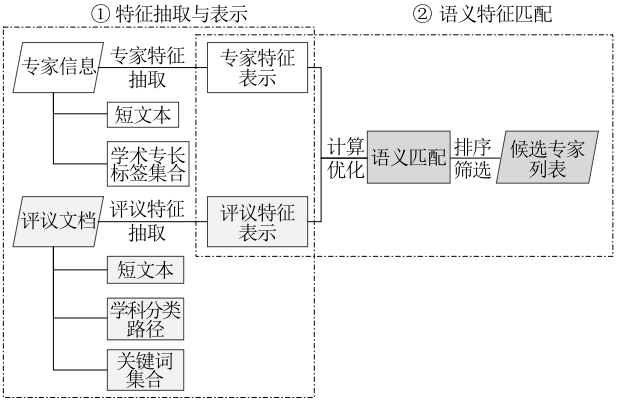


图 1　融合多源信息的学术专长语义匹配方法 ExpRec

特征抽取与表示通过设计专家特征抽取器对专家短文本和学术专长信息进行抽取和融合，获得专家特征表示；通过设计评议文档特征抽取器对评议文档短文本、学科分类路径、关键词集合进行抽取和融合，获得评议文档特征表示.

语义特征匹配使用专家特征表示和评议文档特征表示对专家和评议文档间的余弦相似度进行计算及模型优化，根据余弦相似度计算结果进行排序筛选，获得和评议文档相匹配的候选专家列表.

4　特征抽取与匹配

4.1　专家特征抽取

本文的专家特征从短文本和学术专长标签 2 类信息源抽取，不同源的专家特征常包含不同的信息量. 为充分挖掘专家多源特征带来的不同信息价值、提升本文匹配方法的质量和效果，本文使用专家特征抽取器 ExpFeat 对专家多源特征进行建模. 如图 2 所示，ExpFeat 利用 TextCNN[19] 模型对专家短文本进行特征编码，生成专家文本特征表示；通过向量加法对专家学术专长标签表示向量进行池化聚合，生成专家专长特征表示；最后，将 2 类特征拼接并输入到全连接层进行特征信息融合，生成专家特征表示.

计算机学报

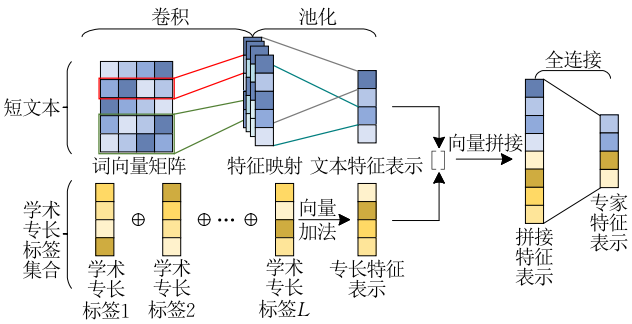


图 2 专家特征抽取器 ExpFeat

为融合专家信息源中短文本和学术专长标签等异构信息,需要对不同结构的信息源进行统一编码与映射.对于专家短文本,鉴于卷积操作对文本信息优秀的泛化性建模能力,以及较低的时空复杂度,使用 TextCNN 模型对专家短文本数据进行处理,抽取专家文本特征表示,如式(1)所示:

$$\boldsymbol{x}_{\text{txt}}(exp_i)=\text{TextCNN}(exptxt_i) \tag{1}$$

其中,$\boldsymbol{x}_{\text{txt}}(exp_i)\in\mathbb{R}^{d\_txt}$ 为专家 exp_i 的专家文本特征表示,d_txt 为文本特征表示的向量维度,$exptxt_i$ 为专家 exp_i 的短文本.

专家学术专长标签表达了专家的知识技能、研究领域等信息,既可以从网上获取专家公开的学术行为以及学术成果,也可以直接利用专家库的信息.本文收集了来自互联网的专家学术成果和学术行为的学术专长标签,在实际应用中还结合了专家库中专家研究领域信息,利用文献[23]的方法对学术专长标签进行再次筛选,筛选结果生成学术专长标签集合作为 ExpFeat 模型输入,确保了学术专长标签集合中的标签和专家研究方向的密切相关性.具体方法包括:首先构建学术网络进行节点表示学习生成学者异质语义表示,使用学者异质语义表示计算专家的专长标签相关度评分;然后依据相关度评分选择相贴切的词汇作为专家的学术专长标签,生成专家学术专长标签集合.

由于 embedding 方法[15]能够将稀疏离散的标签数据映射到低维稠密的连续向量空间,实现捕捉标签语义信息,同时方便模型计算和训练,因此本文利用 embedding 方法对学术专长标签集合进行处理,获得专家专长特征表示,进行池化融合.首先利用全连接网络拟合每个专长标签的低维语义表示向量矩阵,通过向量矩阵乘法对专家每个学术专长标签进行向量映射,以生成专长标签表示向量,如式(2)所示:

$$\boldsymbol{x}(expkey_{i,1})=\boldsymbol{c}(expkey_{i,1})\boldsymbol{W}_{\text{emb}} \tag{2}$$

其中,$\boldsymbol{x}(expkey_{i,l})$ 为专家 exp_i 的学术专长标签集合 $EXPKEY_i$ 中第 l 个标签的表示向量,$\boldsymbol{c}(expkey_{i,1})$ 为第 l 个标签的独热编码,$\boldsymbol{W}_{\text{emb}}\in\mathbb{R}^{|REVKEY_j|\times d\_vex}$ 为可学习的全连接矩阵,$|REVKEY_j|$ 表示关键词数,d_vex 为专家专长特征表示的向量维度.

与专家短文本不同,经过筛选后的专家学术专长标签集合的每个标签是对专家学术专长的精准刻画,为此通过向量加法池化对生成的专长标签表示向量进行融合,以保留全部标签的核心语义特征,得到专家专长特征表示 $\boldsymbol{x}_{\text{expkey}}(exp_i)\in\mathbb{R}^{d\_vex}$,如式(3)所示:

$$\boldsymbol{x}_{\text{expkey}}(exp_i)=\boldsymbol{x}(expkey_{i,1})\oplus\cdots\oplus\boldsymbol{x}(expkey_{i,l})\oplus\cdots\oplus\boldsymbol{x}(expkey_{i,L}) \tag{3}$$

其中,$\boldsymbol{x}_{\text{expkey}}(exp_i)$ 表示专家 exp_i 的专长特征表示,$\boldsymbol{x}(expkey_{i,l})$ 为专家 exp_i 的学术专长标签集合 $EXPKEY_i$ 中第 l 个标签的表示向量,$L=|EXPKEY_i|$ 为专家 exp_i 的学术专长标签数量,$\oplus$ 表示向量加法.

获得专家文本特征表示和专长特征表示后,利用全连接层对专家文本特征表示和专长特征表示进行融合,得到融合多源信息的专家特征表示,如式(4)所示:

$$\boldsymbol{x}_{\text{mat}}(exp_i)=\boldsymbol{W}_{\text{exp}}\cdot[\boldsymbol{x}_{\text{txt}}(exp_i),\boldsymbol{x}_{\text{expkey}}(exp_i)]\oplus\boldsymbol{b}_{\text{exp}} \tag{4}$$

其中,$\boldsymbol{x}_{\text{mat}}(exp_i)$ 为专家 exp_i 专家特征表示,$\boldsymbol{W}_{\text{exp}}\in\mathbb{R}^{d\_mat\times(d\_txt+d\_vex)}$ 和 $\boldsymbol{b}_{\text{exp}}\in\mathbb{R}^{d\_mat\times 1}$ 分别表示全连接层的权重和偏置,d_txt 为文本特征表示的向量维度,d_vex 为专家专长特征表示的向量维度,d_mat 为专家特征表示的向量维度,"·"表示矩阵乘法,[]表示向量拼接.

通过专家特征抽取,每位专家 exp_i 的相关信息均被融入了短文本和学术专长特征,并编码为低维稠密的向量表示 $\boldsymbol{x}_{\text{mat}}(exp_i)$,能有效描述专家知识技能、研究领域等信息,降低语义匹配复杂度.

4.2 评议特征抽取

评议文档信息包括短文本、学科分类路径、关键词等,具有多源、多类型的特点,不同学科的评议文档具有不同的语义信息分布,为语义特征匹配提供区分依据.为了充分挖掘评议文档蕴含的潜在价值,与专家特征抽取相似,使用评议文档特征抽取器 RevFeat 对评议文档多源特征进行建模.如图 3 所示,评议文档特征抽取器 RevFeat 使用 TextCNN 模型对评议文档短文本进行特征编码,生成评议文档文本特征表示;通过向量加法分别对学科节点和关键词的表示向量进行池化聚合,生成学科分类和

关键词特征表示；转化学科节点生成学科键值对集合，基于注意力机制融合“小同行”学科节点生成“小同行”特征表示；最后，拼接评议文档文本、学科分类、关键词特征表示、“小同行”特征表示，利用评议文档特征抽取器 RevFeat 的全连接层融合 4 类特征，生成评议文档特征表示.

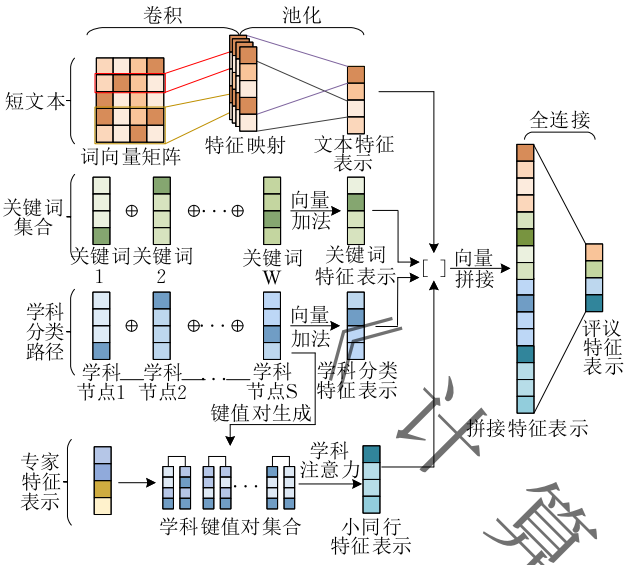


图 3　评议特征抽取器 RevFeat

与专家特征抽取类似，为了融合评议文档的多源异构信息，对不同结构的信息源进行统一编码与映射，首先对评议文档短文本特征进行抽取，利用 TextCNN 模型抽取评议文档文本特征表示，如式(5)所示：

$$\boldsymbol{x}_{\mathrm{txt}}(rev_j)=\mathrm{TextCNN}(revtxt_j) \tag{5}$$

其中，$\boldsymbol{x}_{\mathrm{txt}}(rev_j)\in\mathbb{R}^{d\_txt}$ 为评议文档 rev_j 的评议文档文本特征表示，d_txt 为文本特征表示的向量维度，$revtxt_j$ 为评议文档 rev_j 的短文本.

然后生成评议文档的学科分类树和关键字的向量表示，为保证与专家学术专长标签的向量空间一致性，同样采用 embedding 和加法池化方法进行向量表示和融合，实现对评议文档的学科分类树和关键字中语义信息的提取.

利用 embedding 方法，学习学科分类路径中每个学科节点 $spath_{j,s}$ 的向量表示 $\boldsymbol{x}(spath_{j,s})$，通过向量加法⊕进行池化聚合，得到学科分类特征表示 $\boldsymbol{x}_{\mathrm{spath}}(rev_j)$，如式(6)所示：

$$\boldsymbol{x}_{\mathrm{spath}}(rev_j)=\boldsymbol{x}(spath_{j,1})\oplus\cdots\oplus\boldsymbol{x}(spath_{j,s})\oplus\cdots\oplus\boldsymbol{x}(spath_{j,S}) \tag{6}$$

其中，$\boldsymbol{x}_{\mathrm{spath}}(rev_j)$ 表示评议文档 rev_j 的学科分类特征，$\boldsymbol{x}(spath_{j,s})$ 为学科分类路径 $SPATH_j$ 中第 s 个学科节点的表示向量，S 为学科分类路径 $SPATH_j$ 中的学科数.

利用 embedding 方法学习关键词集合中每个关键词 $revkey_{j,w}$ 的向量表示 $\boldsymbol{x}(revkey_{j,w})$，通过向量加法⊕进行池化聚合，得到关键词特征表示 $\boldsymbol{x}_{\mathrm{revkey}}(rev_j)$，如式(7)所示：

$$\boldsymbol{x}_{\mathrm{revkey}}(rev_j)=\boldsymbol{x}(revkey_{j,1})\oplus\cdots\oplus\boldsymbol{x}(revkey_{j,w})\oplus\cdots\oplus\boldsymbol{x}(revkey_{j,W}) \tag{7}$$

其中，$\boldsymbol{x}_{\mathrm{revkey}}(rev_j)\in\mathbb{R}^{d\_key}$ 表示评议文档关键词特征表示，d_key 为关键词特征表示的向量维度，$\boldsymbol{x}(revkey_{j,w})$ 为关键词集合 $REVKEY_j$ 中关键词 $revkey_{j,w}$ 的表示向量.

在学科分类路径中，根学科节点反映了评议文档隶属的“大学科领域”信息，而靠近叶子的学科节点反映了评议文档隶属的“细分学科领域”信息，若直接通过加法池化对学科节点进行融合，虽然能够捕捉评议文档学科分类的核心学术语义特征，但容易造成与“小同行”专家相匹配的“细分学科领域”信息被“大学科领域”信息稀释. 鉴于评议文档评审环节中“小同行”专家的重要学术意义，为此，引入注意力机制动态调节学科节点的融合权重，提取与评议专家密切相关的“细分学科领域”分类信息，生成“小同行”特征表示 $\boldsymbol{x}_{\mathrm{small}}(rev_j)$，如式(8)所示：

$$\begin{aligned}&\boldsymbol{key}_{j,s}=\boldsymbol{x}(spath_{j,s})\boldsymbol{W}_{\mathrm{key}},\\&\boldsymbol{value}_{j,s}=\boldsymbol{x}(spath_{j,s})\boldsymbol{W}_{\mathrm{value}},\\&\boldsymbol{x}_{\mathrm{small}}(rev_j)=\sum_s(\boldsymbol{x}_{\mathrm{mat}}^{\mathrm{T}}(exp_i)\boldsymbol{key}_{j,s})\boldsymbol{value}_{j,s}\end{aligned} \tag{8}$$

其中，$\boldsymbol{x}(spath_{j,s})$ 为学科分类路径 $SPATH_j$ 中第 s 个学科节点的表示向量，$\boldsymbol{W}_{\mathrm{key}},\boldsymbol{W}_{\mathrm{value}}\in\mathbb{R}^{d\_mat\times d\_spath}$ 为学科节点键值对转化矩阵，d_spath 为学科分类特征表示的向量维度，d_mat 为评议文档特征表示的向量维度，$\boldsymbol{x}_{\mathrm{mat}}(exp_i)$ 为专家 exp_i 专家特征表示.

利用全连接层融合评议文档文本特征表示、学科分类特征表示、关键词特征表示、“小同行”特征表示，生成评议文档特征表示，如式(9)所示：

$$\boldsymbol{x}_{\mathrm{mat}}(rev_j)=\boldsymbol{W}_{\mathrm{rev}}\cdot[\boldsymbol{x}_{\mathrm{txt}}(rev_j),\boldsymbol{x}_{\mathrm{spath}}(rev_j),\boldsymbol{x}_{\mathrm{revkey}}(rev_j),\boldsymbol{x}_{\mathrm{small}}(rev_j)]\oplus\boldsymbol{b}_{\mathrm{rev}} \tag{9}$$

其中，$\boldsymbol{x}_{\mathrm{mat}}(rev_j)$ 为评议文档 rev_j 的评议文档特征表示，$\boldsymbol{W}_{\mathrm{rev}}\in\mathbb{R}^{d\_mat\times(d\_txt+d\_spath+d\_key+d\_mat)}$ 和 $\boldsymbol{b}_{\mathrm{rev}}\in\mathbb{R}^{d\_mat\times 1}$ 分别表示全连接层的权重和偏置，d_txt 为文本特征表示的向量维度，d_spath 为学科分类特征表示的向量维度，d_key 为关键词特征表示的向量维度，d_mat 为评议文档特征表示的向量维度，其中评议文档特征表示和专家特征表示的向量维度相同.

通过评议特征抽取，每篇评议文档不同类型的短文本、学科分类路径以及关键词集合描述的评议

计算机学报

文档信息被提取出多种特征，并最终转换为包含深层语义信息的评议文档特征表示，同时描述了评议文档的概要信息.

4.3 语义特征匹配

为实现专家和评议文档间的匹配，本节详细说明语义特征匹配的具体内容和计算步骤. 首先，在抽取出专家特征表示和评议文档特征表示后，语义特征匹配根据专家和评议文档特征表示之间的向量余弦相似度，衡量专家和评议文档的专长匹配程度，以捕捉专家和评议文档的语义关联性. 然后，为增强模型对学术语义差异的捕捉能力，引入负例专家参与模型训练，通过最大化正例专家概率优化特征抽取器参数，重新匹配并筛选专家库中合适的专家，以确定与评议文档专业背景相关的候选专家列表，完成专家和评议文档的匹配.

考虑到向量空间的可解释性，使用余弦相似度描述专家和评议文档的专长匹配程度. 具体地，给定专家 exp_i 的专家特征表示 $\boldsymbol{x}_{\text{mat}}(exp_i)$ 和评议文档 rev_j 的评议文档特征表示 $\boldsymbol{x}_{\text{mat}}(rev_j)$，余弦相似度计算如式(10)所示：

$$score_{i,j}=\boldsymbol{x}_{\text{mat}}(exp_i)\odot\boldsymbol{x}_{\text{mat}}(rev_j) \tag{10}$$

其中，$score_{i,j}\in[-1,1]$ 为余弦相似度，描述专家和评议文档的专长匹配程度，$\odot$ 表示余弦相似度函数. 通过选取余弦相似度最大的前 K 个专家作为专家候选列表，完成专家和评议文档的匹配.

式(10)中，语义特征匹配的效果 $score_{i,j}$ 依赖于专家特征抽取器 ExpFeat 和评议文档特征抽取器 RevFeat. 在语义匹配过程中，若直接使用标注数据训练特征抽取器，可能缺失非匹配专家的语义特征，导致与评议文档不相关的部分专家获得较高的余弦相似度，进而影响匹配效果.

为了优化特征抽取器参数，本文参考 DSSM[18] 训练框架对特征抽取器进行训练，通过在训练过程中随机引入负例来提升神经网络对语义差异的捕捉能力. 根据实际评议记录选取正例专家，随机负采样构建负例专家. 利用 DSSM 特征抽取器的训练框架(如图 4)，通过为一条训练数据随机引入若干负例专家，使模型在训练过程中增强对特征的泛化性建模，进一步提高模型对语义差异的捕捉能力，降低不相关专家的语义关联，提升匹配的精准度.

具体地，给定评议文档 rev_j 信息、正例专家 exp_i 信息和 $Z(Z>0)$ 个负例专家 $exp_z(z\neq i)$ 信息，首先利用评议文档特征抽取器和专家特征抽取器抽取特征表示，然后，结合余弦相似度计算专家特征表示和

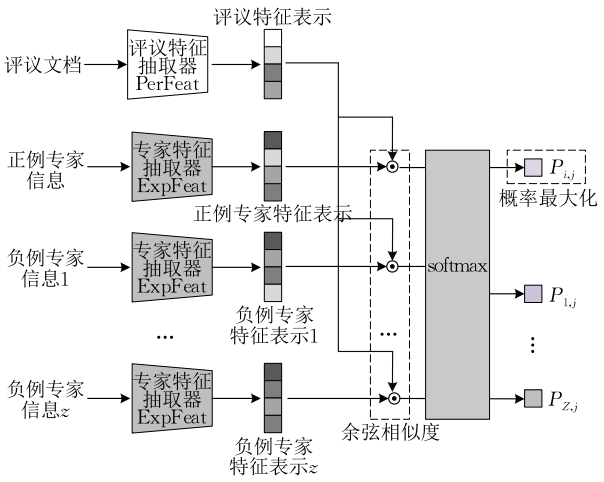


图 4 DSSM 特征抽取器训练框架

评议文档特征表示间向量距离，将专家和评议文档特征映射入相同语义向量空间，最终，使用 softmax 函数计算正例专家概率，通过最大化正例专家概率来优化特征抽取器参数，提升衡量专长匹配效果的余弦相似度的准确性. 正例专家概率 $P_{i,j}$ 计算如式(11)所示：

$$\begin{aligned}P_{i,j}&=\text{softmax}(\boldsymbol{x}_{\text{mat}}(exp_i)\odot\boldsymbol{x}_{\text{mat}}(rev_j))\\&=\frac{e^{x_{\text{mat}}(exp_i)\odot x_{\text{mat}}(rev_j)}}{e^{x_{\text{mat}}(exp_i)\odot x_{\text{mat}}(rev_j)}+\sum_{z=1}^{Z}e^{x_{\text{mat}}(exp_z)\odot x_{\text{mat}}(rev_j)}}\end{aligned} \tag{11}$$

其中，$P_{i,j}$ 为正例专家概率，表示需要最大化的正例专家 exp_i 和评议文档 rev_j 余弦相似度的优化概率，exp_z 表示随机采样的第 $z(z<Z)$ 个负例专家，$\boldsymbol{x}_{\text{mat}}(exp_i)$ 为正例专家 exp_i 的专家特征表示，$\boldsymbol{x}_{\text{mat}}(exp_z)$ 为负例专家 exp_z 的专家特征表示，$\boldsymbol{x}_{\text{mat}}(rev_j)$ 为评议文档 rev_j 的评议文档特征表示.

训练过程中通过最大化正例专家概率 $P_{i,j}$ 训练特征抽取器，进一步提升余弦相似度准确性. 经过正负例训练，获得参数优化后的专家特征提取器 ExpFeat 和评议文档特征提取器 RevFeat，然后基于优化后的特征提取器为每个评议文档生成候选专家列表，完成专家和评议文档间的语义匹配.

5 实验及效果评估

5.1 实验数据

(1)论文评审数据集

本文使用开源的论文评审数据集 Review Data 进行学术专长语义匹配实验，以评估 ExpRec 方法的有效性. Reviewer Data 是由文献[1]作者贡献的开源论文评审数据集，基于 SIGIR2007 发表和录用的论文构建，包括 73 篇论文、189 位专家和 25 个研

究主题，每篇论文和每位专家都具有一个或多个研究主题.

在神经网络模型训练中，负例数据的比例会影响方法性能，随着负例比例的增加，模型性能首先略有提升而后基本保持不变. 因此，常使用 1∶1～1∶5 的比例进行负例数据采样[24-25]. 本文参照文献[24]采样结论，使用 1∶4 的正负比例，通过论文和专家研究主题向量的余弦相似度构建了 3650 条评审标注记录，并利用论文和专家的文本分别提取了 30 个专家学术专长标签和论文关键词. 同时，结合 ACM Computing Classification System 2012 学科分类标准和研究主题①构建了论文的学科分类树.

为了训练学术专长匹配模型并进行效果评估，本文将 3650 条论文评审标注记录按 8∶2 的比例划分为训练集和测试集. 其中，训练集包含 58 篇论文，测试集包含 15 篇论文，每篇论文均对应 10 位正例评审专家和 40 位负例评审专家.

(2) 项目评审数据集

为了对本文所提学术专长匹配方法进行应用效果评估和验证，本文根据近年网络公开的项目评审信息，收集并构建了专家库数据和项目评审数据. 其中，专家库数据共包含 2902 位专家，主要涉及个人主页抽取的专家姓名、工作单位等基本信息以及科研经历、学术专长等描述信息；项目评审数据共包含 1666 个评审项目，主要包括项目标题、项目指南、学科分类路径以及模拟项目专家评审产生的 22 448 条评审记录.

在进行学术专长匹配实验时，专家短文本为专家库中的科研经历，专家学术专长标签从专家学术成果和学术行为以及专家库中提取后，经专家画像得到. 项目短文本为项目标题和项目指南，项目学科分类路径由[专项]-[一级学科]-[二级学科]-[三级学科]-[四级学科]构成，评审项目的关键词集合通过 TextRank4ZH 工具②从项目短文本中提取.

此外，本文对数据进行了预处理，构建标注数据. 将评审记录中的评审专家作为正例专家，根据学科分类路径中的专项学科随机对专家进行负采样，构建负例专家. 预处理后的训练集和测试集均包含约 8 万条标注记录，其中，训练集共有 18384 条正例标注数据、64 916 条负例标注数据，用于训练学术专长语义匹配方法 ExpRec；测试集共有 4060 条正例标注数据，79 236 条负例标注数据，每个项目包含 50 位专家，正例专家数为 1～5 位，用于评价 ExpRec 方法是否能从 50 位专家中匹配出相关正例专家.

5.2　对比方法

为了评估本研究所提方法的有效性，实验与具有代表性的多个主题建模方法和语义匹配方法进行对比. 主题建模方法的实验内容包括构建语义矩阵，训练主题模型，计算专家短文本和评议短文本的主题向量，并根据主题向量计算专家和评议文档的专长匹配程度，实现专家和评议文档的匹配. 语义匹配方法的实验内容包括利用神经网络模型编码专家短文本和评议短文本，学习语义特征表示，并使用语义特征表示进行计算，实现专家和评议文档匹配. 具体对比方法如表 1 所示.

表 1　对比方法概览表

方法名称	方法类型	方法实现向量生成	学术专长匹配
LDA[7]	主题建模	构造文档-关键词矩阵，训练 LDA 模型，计算主题向量	余弦相似度
ATM[9]	主题建模	构造专家-评议文档矩阵以及文档-关键词矩阵，训练 ATM 模型，计算主题向量	余弦相似度
LSA[3]	主题建模	构造文档-关键词矩阵，训练 LSA 模型，计算主题向量	余弦相似度
BTM[26]	主题建模	抽取短文本二元词组，训练 BTM 模型，计算主题向量	余弦相似度
WSIM[12]	主题建模	使用语言模型统计单词重要性，基于 LDA 模型生成主题向量	使用增益率和词频计算主题关联度
word2-vec[15]	语义匹配	利用 CBOW 模型编码专家短文本和评议文档短文本，通过均值池化计算句向量的方式抽取文本特征，生成特征表示	余弦相似度
CNN-DSSM[19]	语义匹配	利用 TextCNN 模型编码专家短文本和评议文档短文本，生成特征表示	使用 DSSM 框架训练，计算匹配度
BERT[20]	语义匹配	利用 BERT 编码专家短文本、学术专长标签集合以及评议文档短文本、学科分类路径、关键词集合等信息，生成特征表示	对特征表示向量进行均值池化，利用正负例训练模型，计算匹配度
R2R[22]	语义匹配	基于 BERT 模型生成专家和评议文档的全局语义编码，基于 CNN 模型构建局部语义编码，通过向量拼接生成特征表示	利用多层感知机对特征表示向量进行处理，计算匹配度

ExpRec 是本文所提的融合多源信息的学术专长语义匹配方法，通过特征表示抽取编码专家的短文本、学术专长标签集合以及评议文档的短文本、学科分类路径、关键词集合等信息，并利用正负例专家进一步训练并优化专家特征抽取器 ExpFeat 和评议文档特征抽取器 RevFeat，计算余弦相似度，进行专家信息和评议文档的语义匹配，筛选合适的专家.

① https://dl.acm.org/ccs#10002951
② https://github.com/letiantian/TextRank4ZH

计算机学报

5.3 评价指标

为了评估主题建模方法和语义匹配方法在测试集上的学术专长匹配效果，本实验使用增益率（$NDCG@K$）[26]和命中率（$HR@K$）[26-27]作为评价指标.

增益率 $NDCG@K$ 是位置敏感的评价指标，用于评价前 K 个匹配结果中正例专家的排名情况，正例专家排名越靠前，增益率越大，如式(12)：

$$NDCG@K = Z_K \sum_{k=1}^{K} \frac{2^{r_k}-1}{\log_2(1+k)} \tag{12}$$

其中，r_k 表示相关性因子，取值为 0 或 1；$r_k=0$ 时表示匹配专家为负例专家，与评议文档内容不匹配；$r_k=1$ 时表示匹配专家为正例专家，与评议文档内容匹配. Z_K 表示归一化因子.

命中率 $HR@K$ 是位置不敏感评价指标，用于评价前 K 个匹配结果中正例专家的比例，正例专家比例越大，命中率越高，如式(13)：

$$HR@K = cvr/K \tag{13}$$

其中，cvr 表示前 K 个结果中正例专家的个数.

此外，式(12)和式(13)同样适用于“小同行”匹配的指标.

5.4 实验结果

表 2 和表 3 展示了对比方法和本文 ExpRec 方法在论文数据集上的实验结果，表 4 和表 5 展示了在项目数据集上的实验结果，可以得出如下结论：

表 2 论文评审数据集学术专长匹配增益率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	4.01	4.77	21.58	23.05	44.82
ATM	27.81	40.63	49.96	58.54	63.59
LSA	34.53	44.57	55.14	61.67	67.51
BTM	18.58	33.04	43.27	50.67	57.80
WSIM	22.37	33.05	44.03	52.82	58.12
word2vec	45.81	60.31	66.51	69.94	74.18
CNN-DSSM	74.54	83.06	86.14	88.96	91.08
BERT	75.95	81.90	84.67	87.23	90.17
R2R	77.77	81.10	84.84	88.28	91.21
ExpRec	**86.69**	**93.00**	**95.54**	**96.10**	**96.64**

表 3 论文评审数据集学术专长匹配命中率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	3.33	4.67	41.33	44.67	100
ATM	24.00	47.33	67.33	87.33	100
LSA	28.33	47.33	70.00	85.33	100
BTM	16.67	42.67	64.67	82.00	100
WSIM	23.33	42.67	66.00	86.67	100
word2vec	42.00	68.00	81.33	89.33	100
CNN-DSSM	66.00	81.33	88.00	94.67	100
BERT	70.00	80.67	86.67	92.67	100
R2R	70.67	76.67	84.67	92.67	100
ExpRec	**80.67**	**92.00**	**97.33**	**98.67**	**100**

表 4 项目评审数据集学术专长匹配增益率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	4.80	4.80	7.51	8.33	26.28
ATM	16.14	22.26	25.90	28.62	36.02
LSA	34.53	44.57	55.14	61.67	65.71
BTM	36.16	42.81	45.80	47.89	49.65
WSIM	36.60	42.14	45.97	48.69	51.05
word2vec	40.00	47.98	50.83	51.72	52.11
CNN-DSSM	81.35	85.39	86.49	86.96	87.20
BERT	82.57	85.14	86.14	86.72	87.03
R2R	77.92	80.54	81.53	82.08	82.51
ExpRec	**89.15**	**91.41**	**94.49**	**94.50**	**94.51**

表 5 项目评审数据集学术专长匹配命中率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	6.03	6.03	15.32	17.05	100
ATM	24.93	43.68	56.97	67.84	100
LSA	29.33	47.33	70.00	85.33	100
BTM	53.57	73.90	84.31	92.52	100
WSIM	50.44	67.15	80.99	91.80	100
word2vec	58.31	83.06	94.20	98.26	100
CNN-DSSM	80.00	92.07	96.70	98.79	100
BERT	81.80	91.52	95.73	98.45	100
R2R	83.65	91.88	95.66	98.51	100
ExpRec	**91.68**	**94.60**	**97.90**	**98.94**	**100**

首先，ExpRec 方法能准确捕捉专家和评议文档之间的语义相关性. 由表 2 到表 5 的实验结果可知，本文所提的 ExpRec 方法在命中率和增益率上达到了最佳效果，并且 top$K$ 取值为 10 时，本文方法与对比方法相比，在论文评审数据集中命中率提升 9%、增益率提升 10%，在项目评审数据集中命中率提升 9%、增益率提升 6%，验证了 ExpRec 方法的有效性，同时也说明了通过融合多源信息可以补充模型的语义表达能力，从而让 ExpRec 方法能够充分提取专家和评议文档的特征信息，进行合理的相关性匹配.

其次，语义匹配方法的匹配效果优于主题建模方法. 实验结果显示，语义匹配方法（word2vec、CNN-DSSM、BERT、R2R、ExpRec）在命中率和增益率上的性能要优于主题建模方法（LDA、ATM、LSA、BTM、WSIM），说明与仅建模浅层语义信息的主题建模方法相比，由于语义匹配方法通过对神经网络进行端到端的学习，能够使可训练的参数向全局最优方向收敛，更能准确捕捉专家信息和评议文档之间的语义相关性，提升专长匹配效果. 其中，BTM 和 WSIM 这 2 个主题建模方法除挖掘主题关联外，还会对模型参数进行迭代优化，因此在主题建模方法中有较好的表现，然而由于 BTM 和 WSIM

仅基于主题关联性优化参数，使模型参数并不能向最优方向进行收敛，从而导致匹配结果低于语义匹配方法.

最后，除本文方法外，可以发现 CNN-DSSM、BERT 和 R2R 也获得较好的匹配效果. 其中，因为专家匹配任务文本特征长度较短，且 CNN-DSSM 模型结构较精简，同时泛化性更强，导致 CNN-DSSM 实验结果略好于 BERT 方法. R2R 通过对 BERT 模型提取的全局特征进行文本卷积实现专家匹配，虽然卷积操作能够提升模型的泛化能力，但过早的引入 BERT 导致模型放大了数据中原始噪音对结果的影响，造成 R2R 实验结果虽高于 BERT 方法但略低于 CNN-DSSM 模型. 此外，CNN-DSSM、BERT 和 R2R 均在特征提取前对多源信息进行融合，使得特征匹配环节丧失对不同信息源的感知，进而影响了方法对不同信息源的敏感度，而本文方法是在特征匹配阶段才对不同源信息进行融合，并且通过向量拼接操作尽可能保留了不同源信息的原始特征，让模型可以更加充分地捕捉到信息源间的数据分布差异，从而实现对专家和评议文档多源数据的有效建模，最终使本文所提 ExpRec 方法匹配效果优于 CNN-DSSM、BERT 和 R2R.

为验证本文方法对“小同行”专家的匹配效果，本文基于评议文档学科分类路径的叶子节点构建并标注“小同行”专家，并以“小同行”专家为正例，展开匹配实验. 表 6 和表 7 展示了对比方法和本文 ExpRec 方法在公开论文评审数据集上的“小同行”匹配结果，表 8 和表 9 展示了在项目评审数据集上的“小同行”匹配结果. 从表中可知，本文方法在“小同行”命中率和增益率上均取得了最佳效果，说明本文通过额外对“小同行”特征进行单独建模，可以更加细粒度地捕捉项目和专家间的学术关联信息，验证了本文方法在“小同行”匹配上的有效性.

表 6　论文评审数据集“小同行”匹配增益率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	2.10	2.10	6.90	21.00	42.28
ATM	26.50	39.08	46.88	57.36	61.84
LSA	32.89	44.77	53.47	61.36	65.86
BTM	23.31	36.95	44.42	53.79	60.42
WSIM	27.53	36.67	45.94	55.92	61.80
word2vec	37.78	50.71	58.76	66.42	69.20
CNN-DSSM	72.94	81.28	86.82	88.54	90.14
BERT	76.17	82.21	85.33	88.74	91.42
R2R	78.67	82.01	85.67	88.54	92.17
ExpRec	**86.00**	**92.27**	**94.45**	**95.31**	**95.83**

表 7　论文评审数据集“小同行”匹配命中率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	2.70	2.70	13.30	48.00	100
ATM	24.67	47.33	64.00	88.67	100
LSA	30.00	51.33	70.00	88.67	100
BTM	20.67	45.33	61.33	83.33	100
WSIM	26.14	47.07	65.67	85.67	100
word2vec	34.00	57.33	74.67	88.00	100
CNN-DSSM	65.33	80.00	92.00	96.00	100
BERT	68.00	78.67	85.33	93.33	100
R2R	69.77	79.33	83.14	94.21	100
ExpRec	**80.67**	**92.00**	**96.67**	**98.67**	**100**

表 8　项目评审数据集“小同行”匹配增益率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	4.80	4.80	6.51	9.33	26.28
ATM	16.14	22.26	25.53	27.62	37.51
LSA	34.53	44.57	51.03	56.13	59.37
BTM	36.16	41.03	45.80	47.89	49.65
WSIM	36.53	40.54	44.51	48.03	51.05
word2vec	40.00	46.53	50.83	52.31	55.11
CNN-DSSM	79.01	82.39	85.49	85.96	86.20
BERT	81.31	82.14	86.79	85.99	86.03
R2R	80.67	82.16	82.94	84.11	85.65
ExpRec	**88.15**	**90.41**	**92.49**	**94.50**	**94.51**

表 9　项目评审数据集“小同行”匹配命中率对比

方法名称	top$K$ 候选专家				
	10	20	30	40	50
LDA	5.03	5.03	12.57	16.05	100
ATM	24.53	42.61	53.21	68.04	100
LSA	34.33	47.33	68.00	86.33	100
BTM	51.57	72.10	84.07	91.33	100
WSIM	51.33	69.61	80.52	90.94	100
word2vec	52.31	78.11	91.20	97.26	100
CNN-DSSM	77.11	89.07	93.10	97.21	100
BERT	78.66	87.52	92.91	97.45	100
R2R	82.33	89.12	93.07	97.22	100
ExpRec	**90.68**	**94.60**	**96.98**	**98.12**	**100**

为验证信息源和正负采样比对 ExpRec 方法的影响，本文在论文评审数据集上开展消融实验，对不同信息源和正负采样比下 ExpRec 的匹配效果进行分析和验证. 此外，在项目评审数据集上可以得到类似实验结果，不再赘述.

表 10 展示了不同信息源对本文所提方法 ExpRec 增益率和命中率实验结果的影响，由表 10 可观察到，在 2 个评价指标上增加模型输入信息源能够提升本文方法语义匹配效果. 同时使用专家短文本和学术专长标签与仅使用 1 种专家信息源相比，语义匹配效果得到提升. 增加评议文档学科分类路径和关键词与仅使用评议文档短文本的方法相比，实验效果有较

大提升.在使用专家的短文本、学术专长标签集合以及评议文档的短文本、学科分类路径、关键词集合全部信息的情况下,专家和评议文档间的语义关联更强,匹配效果得到进一步完善,模型效果最佳,进一步印证了通过融合多种信息源,可以为模型引入更加丰富的语义信息量,有助于提升专家匹配精度.

此外,从表 10 结果可知,与未筛选专家学术专长标签相关程度的方法相比,本文方法通过筛选操作能够减弱专长标签中样本噪音的影响,从而提升专家与评议文档的学术关联程度,使得最终匹配结果在增益率和命中率 2 个指标上均有提高,语义匹配效果更优.

表 10 不同信息源的实验结果对比

信息源						top$K$ 候选专家									
						NDCG					*HR*				
专家短文本	专家学术专长标签(未筛选)	专家学术专长标签	评议文档短文本	评议文档关键字	评议文档学科分类路径	10	20	30	40	50	10	20	30	40	50
√	×	×	√	×	×	74.38	80.31	85.61	87.60	90.00	67.33	78.00	89.33	94.00	100
×	×	√	√	×	×	71.64	79.08	83.09	85.63	87.49	67.33	80.67	89.33	95.33	100
√	×	√	√	×	×	76.92	84.26	87.34	89.60	91.45	70.00	83.33	90.00	95.33	100
√	×	×	√	√	√	86.69	92.11	94.93	95.71	96.52	80.67	90.00	96.00	95.33	100
×	×	√	√	√	√	86.26	93.04	95.14	96.07	96.52	80.00	90.00	96.67	98.00	100
√	√	×	√	√	√	85.42	91.61	93.83	94.68	95.75	79.33	90.67	95.33	97.33	100
√	×	√	√	√	√	**87.11**	**93.24**	**95.57**	**96.10**	**96.64**	**81.33**	**92.00**	**97.33**	**98.67**	**100**

表 11 和表 12 分别展示了不同正负采样比对本文所提方法 ExpRec 增益率和命中率实验结果的影响,从结果中可以看到,随着负采样比例的增加,本文方法的推荐效果由上升逐渐趋于稳定,说明通过在模型训练过程中引入一定比例的负例,能够提高模型匹配效果,同时让模型学习正负样本专家间的差异,可以提高模型对不相关专家的辨别能力.

表 11 不同正负采样比在增益率上的实验结果对比

正负采样比	top$K$ 候选专家				
	10	20	30	40	50
1∶0	41.16	51.96	59.99	65.63	71.45
1∶1	81.81	88.39	91.50	92.64	93.44
1∶2	85.15	91.20	94.65	94.94	95.75
1∶3	86.10	91.90	94.46	95.61	96.14
1∶4	87.11	93.24	95.57	96.10	96.64
1∶5	84.51	91.48	93.35	94.48	95.54
1∶6	85.30	92.11	93.15	95.61	95.72

表 12 不同正负采样比在命中率上的实验结果对比

正负采样比	top$K$ 候选专家				
	10	20	30	40	50
1∶0	35.33	54.67	72.00	85.33	100
1∶1	76.67	88.67	95.33	98.00	100
1∶2	79.33	90.00	97.33	98.00	100
1∶3	80.00	90.67	96.00	98.67	100
1∶4	81.33	92.00	97.33	98.67	100
1∶5	78.00	90.67	94.67	97.33	100
1∶6	81.10	91.67	93.12	97.32	100

为进一步验证 ExpRec 方法的语义匹配效果,本文在论文评审数据集和项目评审数据集上分别对 ExpRec 方法的实验匹配结果展开了实例分析.鉴于 BERT 方法在语义匹配任务中至今仍是首选对比方法,有着公认的普适性和推演能力,因此选取 BERT 作为基线展开实例分析.

图 5 和图 6 是对论文数据集最差匹配结果的分析.其中,图 5 展示了ExpRec最差匹配结果中 ExpRec 和 BERT 的匹配结果,图 6 展示了 BERT 最差匹配结果中 ExpRec 和 BERT 的匹配结果,标注为 1 表示正例候选专家,标注为 0 表示负例候选专家.从图中可以看出,无论在哪个较差示例下,ExpRec 方法都能相对将更多标注为 1 的正例候选专家召回并排在匹配结果前面,将更多标注为 0 的负例候选专家排在靠后的位置,从侧面说明了 ExpRec 方法在学术专长语义匹配上的有效性.

评审论文	候选专家	标注	专长匹配度
P067	R032	1	0.6916
P067	R043	1	0.6802
P067	R060	1	0.6798
P067	R153	1	0.6793
P067	R067	1	0.6509
P067	R003	1	0.6320
P067	R119	0	0.5520
P067	R087	0	0.5497
P067	R100	0	0.5017
P067	R188	0	0.4773
P067	R160	1	0.4672
P067	R134	0	0.4285
P067	R110	1	0.4196
P067	R096	0	0.4020
P067	R133	0	0.4019
P067	R103	0	0.3955
P067	R121	0	0.3934
P067	R102	0	0.3927
P067	R085	0	0.3796
P067	R130	0	0.3723

(a) ExpRec的匹配结果

评审论文	候选专家	标注	专长匹配度
P067	R043	1	0.7066
P067	R003	1	0.6845
P067	R153	1	0.6751
P067	R032	1	0.6553
P067	R060	1	0.6470
P067	R067	1	0.6285
P067	R119	0	0.5928
P067	R087	0	0.5577
P067	R160	1	0.4498
P067	R099	0	0.4454
P067	R093	0	0.4440
P067	R133	0	0.4417
P067	R116	0	0.4301
P067	R121	0	0.4248
P067	R134	0	0.3998
P067	R100	0	0.3834
P067	R128	0	0.3705
P067	R122	0	0.3679
P067	R096	0	0.3599
P067	R130	0	0.3591

(b) BERT的匹配结果

图 5 ExpRec 较差匹配结果与 BERT 匹配结果对比示例

《计算机学报》

评审论文	候选专家	标注	专长匹配度
P071	R025	1	0.7178
P071	R145	1	0.7151
P071	R174	1	0.7020
P071	R138	1	0.6693
P071	R098	0	0.6631
P071	R057	1	0.6530
P071	R061	1	0.6212
P071	R006	1	0.6020
P071	R027	1	0.5929
P071	R100	0	0.5811
P071	R078	0	0.5411
P071	R168	0	0.5328
P071	R087	0	0.5213
P071	R164	0	0.5204
P071	R118	0	0.4529
P071	R103	0	0.4349
P071	R105	1	0.4135
P071	R080	0	0.4018
P071	R110	0	0.3895
P071	R188	0	0.3811

(a) ExpRec的匹配结果

评审论文	候选专家	标注	专长匹配度
P071	R078	0	0.7178
P071	R025	1	0.7151
P071	R145	1	0.7020
P071	R057	1	0.6693
P071	R168	0	0.6631
P071	R027	1	0.6530
P071	R087	0	0.6212
P071	R098	0	0.6020
P071	R138	1	0.5929
P071	R174	1	0.5811
P071	R006	1	0.5411
P071	R164	0	0.5328
P071	R100	0	0.5213
P071	R061	1	0.5204
P071	R116	0	0.4529
P071	R096	0	0.4349
P071	R099	0	0.4135
P071	R163	0	0.4018
P071	R103	0	0.3895
P071	R084	0	0.3811

(b) BERT的匹配结果

图 6　BERT 较差匹配结果与 ExpRec 匹配结果对比示例

图 7 展示了用 ExpRec 方法判断为专业背景相符的编号为 R025 的候选专家和对应的编号为 P071 待评审论文的具体匹配信息. 由图 7 可以观察到，本文 ExpRec 方法能够准确捕捉专家信息和待评审论文信息之间的语义相关性，有效解决学术专长匹配时的语义鸿沟问题，提升学术专长匹配度. 具体地，专家信息中的“user query”、“search engine”、“information retrival”、“search intention”、“retrieved web page”、“web search engine system”等短语与待评审论文信息中的“information retrieval system”、“conceptual retrieval model”、“retrieval model”、“IR”等短语是语义相关的，进一步验证了本文 ExpRec 方法在学术专长语义匹配上的有效性.

评审论文P071的短文本:

... knowledge-intensive conceptual retrieval and passage extraction of ... incorporating domain-specific knowledge (i.e., information about concepts and relationships between concepts in a certain domain) in an information retrieval (IR) system ... conceptual retrieval model ...

评审论文P071的学科分类路径:

information systems-information retrieval-IR models and ranking-NLP for IR

评审论文P071的关键词集合:

information retrieval, conceptual retrieval model, domain specific knowledge, effectiveness, performance contribution, ...

候选专家R025的短文本:

... a query is submitted to a search engine, the search engine returns a dynamically generated ... a retrieved web page ... extract search result records ... pages returned by search engines ... map a user query to a set of categories, which represent user's search intention ...

候选专家R025的学术专长标签集合:

user query, information retrieval, word sense, word disambiguation, web search engine system, ...

图 7　ExpRec 的语义特征匹配示例

为验证 ExpRec 方法对“小同行”专家的匹配效果，本文选择匹配效果较好的 BERT 方法作为基线，进行“小同行”匹配实例分析. 图 8 展示了项目评审数据集上 ExpRec 最差匹配结果中 ExpRec 和 BERT 的“小同行”匹配结果，其中标注为 1 表示“小同行”专家，可以观察到与 BERT 方法相比，本文方法能够将“小同行”专家排在匹配结果前面，验证了本文方法对专家和评议文档“小同行”学术特征的捕捉能力，侧面说明了 ExpRec 在“小同行”匹配上的有效性.

评审论文	候选专家	标注	专长匹配度
P956	R5634	1	0.7019
P956	R3358	0	0.6991
P956	R9820	1	0.6953
P956	R1686	1	0.6911
P956	R6346	1	0.6331
P956	R1679	1	0.6275
P956	R3795	1	0.6212
P956	R9787	1	0.6101
P956	R6415	1	0.5813
P956	R5445	0	0.5305
P956	R3644	0	0.5211
P956	R1647	0	0.5103
P956	R5672	0	0.4913
P956	R1327	0	0.4712
P956	R2592	0	0.4239
P956	R0754	0	0.4111
P956	R8995	0	0.3995
P956	R6598	0	0.3818
P956	R2535	0	0.3801
P956	R2966	0	0.3801

(a) ExpRec的匹配结果

评审论文	候选专家	标注	专长匹配度
P956	R3358	0	0.6121
P956	R5445	0	0.6051
P956	R3795	1	0.5983
P956	R9787	1	0.5871
P956	R6415	1	0.5766
P956	R5634	1	0.5751
P956	R9820	1	0.5603
P956	R1679	1	0.5552
P956	R1647	0	0.5344
P956	R3644	0	0.5305
P956	R1686	1	0.5296
P956	R6346	1	0.5253
P956	R5672	0	0.5103
P956	R1647	0	0.5001
P956	R1327	0	0.4923
P956	R8995	0	0.4894
P956	R0754	0	0.4531
P956	R6598	0	0.4322
P956	R2535	0	0.4121
P956	R2966	0	0.4091

(b) BERT的匹配结果

图 8　“小同行”匹配示例分析

6　总结与展望

为提升学术专长匹配精度并实现细粒度“小同行”专家匹配，本文提出融合多源信息的学术专长语义匹配方法 ExpRec，包括特征抽取与表示和语义特征匹配 2 个步骤. 对于给定的专家信息（短文本、学术专长标签集合）和评议文档信息（短文本、学科分类路径、关键词集合）中的多源信息，首先进行特征抽取，基于注意力机制融合“小同行”专家学术特征，设计专家特征抽取器、评议文档特征抽取器分别获得专家特征表示和评议文档特征表示，有效融合多源学术语义信息并为“小同行”专家匹配提供特征支撑. 然后进行语义特征匹配，基于专家特征表示和评议特征表示计算余弦相似度，引入负例专家进行模型训练，优化模型参数，建立专家信息和评议文档的深层语义关联，提升模型语义差异捕捉能力. 最后，本文在开源论文评审数据以及根据网络公开数据构建的项目评审数据上分别进行了实验验证和实例分析，结果表明本文所提方法合理、有效，能够提高专家匹配精准度. 在未来工作中，将考虑分析专家信息的时序性，期望通过筛选过滤专家具有时间维度的学术行为和学术成果信息，进一步提升匹配效果，并应用于专家精准推荐领域.

参　考　文　献

[1]　Han Ya-Nan, Liu Jian-Wei, Luo Xiong-Lin. A survey on

probabilistic topic model. Chinese Journal of Computers, 2021, 44(6): 1095-1139(in Chinese)
(韩亚楠, 刘建伟, 罗雄麟. 概率主题模型综述. 计算机学报, 2021, 44(6): 1095-1139)

[2] Maryam K, Zhai C X, Belford G. Multi-aspect expertise matching for review assignment//Proceedings of the ACM Conference on Information and Knowledge Management. New York, USA, 2008: 1113-1122

[3] Scott C D, Susan T D, Thomas K L, et al. Indexing by latent semantic analysis. Journal of the Association for Information Science & Technology, 2010, 41(6): 391-407

[4] Stelmakh I, Shah N B, Singh A. PeerReview4All: Fair and accurate reviewer assignment in peer review//Proceedings of the 30th International Conference on lgorithmic Learning Theory(PMLR). Chicago, USA, 2019, 98: 828-856

[5] Li K, Cao Z, Qu D. Fair reviewer assignment considering academic social network//Asia-Pacific Web (APWeb) and Web-Age Information Management(WAIM) Joint Conference on Web and Big Data. Cham, Swiss: Springer, 2017: 362-376

[6] Peng H W, Hu H J, Wang K Q, et al. Time-aware and topic-based reviewer assignment//Proceedings of the International Conference on Database Systems for Advanced Applications. Cham, Swiss: Springer, 2017: 145-157

[7] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[8] Jin J, Qian G, Zhao Q, et al. Integrating the trend of research interest for reviewer assignment//Proceedings of the World Wide Web Conference. New York, USA, 2017: 1233-1241

[9] Michal R Z, Thomas L G, Mark S, et al. The author-topic model for authors and documents//Proceedings of the Conference in Uncertainty in Artificial Intelligence. Arlington, USA, 2004: 487-494

[10] Anjum O, Gong H, Bhat S, et al. PaRe: A paper-reviewer matching approach using a common topic space//Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing. Stroudsburg, USA, 2019: 518-528

[11] Charlin L, Zemel R S. The Toronto paper matching system: An automated paper-reviewer assignment system//Proceedings of the International Conference on Machine Learning. New York, USA, 2013: 1-11

[12] Tan S, Duan Z, Zhao S, et al. Improved reviewer assignment based on both word and semantic features. Information Retrieval Journal, 2021, 24: 175-204

[13] Huang Jia-Jia, Li Peng-Wei, Peng Min, et al. Review of deep learning-based topic model. Chinese Journal of Computers, 2020, 43(5): 827-855(in Chinese)
(黄佳佳, 李鹏伟, 彭敏等. 基于深度学习的主题模型研究. 计算机学报, 2020, 43(5): 827-855)

[14] Ogunleye O, Ifebanjo T, Abiodun T, et al. Proposed framework for a paper-reviewer assignment system using word2-vec//Proceedings of the 4th Covenant University Conference on E-Governance. Ogun State, Nigeria, 2017: 211-218

[15] Mikolov T, Sutskever I, Chen K, et al. Distributed representation of words and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013: 3111-3119

[16] He Rou-Ying, Xu Jian. Expert recommendation for trouble tickets using attention-based CNN model. Journal of Nanjing University of Science and Technology, 2019, 43(1): 13-21 (in Chinese)
(何柔萤, 徐建. 基于注意力卷积神经网络的工作票专家推荐方法. 南京理工大学学报(自然科学版), 2019, 43(1): 13-21)

[17] Duan Z, Tan S C, Zhao S, et al. Reviewer assignment based on sentence pair modeling. Neurocomputing, 2019, 366: 97-108

[18] Huang P S, He X D, Gao J F, et al. Learning deep structured semantic models for web search using clickthrough data//Proceedings of the ACM International Conference on Information and Knowledge Management. New York, USA, 2013: 2333-2338

[19] Shen Y L, He X D, Gao J F, et al. A latent semantic model with convolutional-pooling structure for information retrieval//Proceedings of the ACM International Conference on Information and Knowledge Management. New York, USA, 2014: 101-110

[20] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, USA, 2019: 4171-4186

[21] Kou N M, U L H, Mamoulis N, et al. Weighted coverage based reviewer assignment//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. Melbourne, Australia, 2015: 2031-2046

[22] Zhang K, Wu L, Lv G, et al. Making the relation matters: Relation of relation learning network for sentence semantic matching//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021, 35(16): 14411-14419

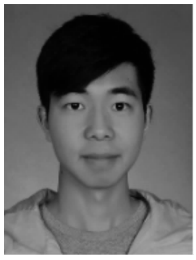
[23] Xie Xiao-Jie, Liang Ying, Wang Zi-Sen, Liu Zheng-Jun. Heterogeneous network node classification method based on graph convolution. Journal of Computer Research and Development, 2022, 59(7): 1470-1485(in Chinese)
(谢小杰, 梁英, 王梓森, 刘政君. 基于图卷积的异质网络节点分类方法. 计算机研究与发展, 2022, 59(7): 1470-1485)

[24] He X N, Liao L Z, Zhang H W, et al. Neural collaborative filtering//Proceedings of the World Wide Web Conference. New York, USA, 2017: 173-182

[25] Song B, Yang X, Cao Y, et al. Neural collaborative ranking //Proceedings of the ACM Conference on Information and Knowledge Management. New York, USA, 2018: 1353-1362

[26] Yan X, Guo J, Lan Y, Cheng X. A biterm topic model for short texts//Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 112-127

[27] He X N, Chen T, Kan M Y, et al. TriRank: Reviewaware explainable recommendation by modeling aspects//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. Melbourne, Australia, 2015: 1661-1670



XIE Xiao-Jie, M. S. His research interest is data mining.

LIANG Ying, senior engineer. Her main research interest is data mining.

WANG Zi-Sen, M. S. candidate. His main research interest is data mining.

LIU Zheng-Jun, M. S. His main research interest is data mining.

Background

As an important part of expert recommendation, expert academic expertise matching provides related candidate experts for expert recommendation by calculating the academic similarity between experts and review documents. The matching results directly affect the quality of expert recommendation and peer review. The existing approaches use keyword search to retrieve the designated experts in the database. However, under the keyword retrieval method, due to the existence of semantic gap and the lack of unified academic standards, it's difficult to establish semantic links between relevant experts and review documents. To improve the matching effect by using semantic information, topic modeling methods and semantic matching methods are often used. The topic modeling methods extract the subject information and quantify the matching degree through subject vector distance. However, because the topic model constructs in terms of words, when a word is polysemy, it is unable to understand the semantic information in words. Compared with topic modeling, semantic matching methods can model semantic information based on co-occurrence feature vectors. However, it doesn't pay enough attention to semantic matching of multi-source information, and don't consider the characteristics of "small peer" when feature extraction, so it's still necessary to deeply mine the association between information sources to capture more fine-grained feature, and ensure the academic identification ability of the experts in the matching results.

In order to solve the above problems, the Multi-source Information Fused Academic Expertise Semantic Matching method (ExpRec) is proposed. Through unified coding of different information sources, heterogeneous information is transformed into the same semantic space to solve the differences caused by classification standards. And considering the importance of "small peer", if subject nodes are integrated through addition, it's easy to cause the information of "sub discipline fields" with "small peer" information to be diluted by the "university discipline fields". Therefore, attention mechanism is used to dynamically adjust the fusion weight, extract the "sub discipline fields" related to the review experts, and generate the representation of the "small peers". Then, dense vectors are used to reduce the complexity, and measure similarity by calculating cosine similarity between vectors. In order to further improve the model's ability to capture semantic differences, negative experts are introduced for model training, and the feature extractor is optimized by maximizing positive experts' probability through softmax function. Finally, experiments are conducted on the open source paper review data set and project review data set. The results show that the method proposed in this paper can effectively improve the accuracy, and when the top$K$ value is 10, compared with the comparison method, the hit rate in the paper review data set is increased by 9%, the gain rate is increased by 10%, the hit rate in the project review data set is increased by 9%, and the gain rate is increased by 6%, which verify the effectiveness of the method.

The main achievement of this paper is to solve a part of the academic classification and recommendation system based on stereo accurate portrait in the National Key Research and Development Program of China (No. 2018YFB1004700).