论文编码：RUC-BK-071201

中国人民大学本科毕业论文

基于统计学前沿知识图谱的导师研究生科学双向选择模型

作　　者：许智彤
学　　院：统计学院
专　　业：统计学
年　　级：2017 级本科
学　　号：2017201661
指导教师：王星
论文成绩：
日　　期：2021 年 3 月 18 日

独创性声明

本人郑重声明：所呈交的论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写的研究成果，也不包含为获得中国人民大学或其他教育机构的学位或证书所使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者：_____ 日　　期：_____

关于论文使用授权的说明

本人完全了解中国人民大学有关保留、使用学位论文的规定，即：学校有保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

论文作者：_____ 日　　期：_____

指导老师：_____ 日　　期：_____

摘要

研究生教育关乎着国家教育和科学技术的发展，在这个人才是第一资源，创新是第一动力的时代，高等教育对国家的发展起到至关重要的作用。随着研究生招生规模的不断扩大，研究生与导师双向选择的过程中存在着信息不对称、培养理念不一致等问题并由此产生负面影响。本文以知识驱动为核心，分析如何科学高效地评估研究生与导师的合作预期，借助基于 DCMM 模型的统计学科前沿知识图谱，从 Web Of Science 导入导师的论文数据，结合从自建网站上收集到的学生数据，进行最优匹配。本工作可用于优化研究生导师双选模式下的人才培养。

关键词：研究生与导师　　双向选择　　知识图谱　　DCMM 模型　　匹配推荐

Abstract

Graduate education is related to the development of national science and technology. In this era when talents are the first resource and innovation is the first driving force, higher education plays a vital role in the development of the country. With the continuous expansion of the scale of graduate enrollment, there are some problems in the process of two-way selection between graduate students and tutors, such as asymmetric information, inconsistent training ideas and so on. This paper takes knowledge driven as the core, analyzes how to evaluate the cooperation expectation between graduate students and tutors scientifically and efficiently. With the help of the statistical frontier knowledge map based on DCMM model, this paper imports the tutor's paper data from web of science, and combines with the student data collected from the self built website to carry out the optimal matching. This work can be used to optimize the talent training under the mode of two-way selection between graduate students and tutors.

Key Words : Graduate and tutor two-way selection knowledge map DCMM model matching recommendation

目录

插图

表格

1 绪论

1.1 研究背景与问题提出

随着硕士研究生招收规模的持续扩大，出国留学潮流在国内兴起，越来越多的人在继续深造的时候需要进行对导师的选择。与此同时，网络论坛上充斥着对研究生生活的吐槽抱怨，新闻报道中也不乏恶性师生关系下造成的悲剧。研究生与导师的合作冲突问题越来越受到重视，目前对其的探讨和提出的解决方案并没有深入到问题的根本，缺少具体到执行的细度，暂时起不到决定性的改变。

其实，目前研究生在各个阶段都十分缺少方法。早期对学科不了解，在没有充分评估的情况下就做出选择，没有考虑万一后期进入不感兴趣怎么办。其了解信息的渠道，一般为百度、知乎等互联网知识问答平台，得到的信息比较分散。或是通过别人的评价，且主要是通过评价信息来做决策，非理性科学，容易受到他人主观片面观点的引导。初期进入研究生学习中，可能会遇到知识不够，跟不上课堂和科研进度的情况，并不知道如何得到帮助。中后期与导师合作论文时，找谁合作，在什么领域合作也是一大难题。

1.2 研究目的与研究意义

该研究的目的就是找到一种方法，在学生对学科知识有限时，帮助了解学科全貌，同时进行科学的自我评估，来判断是否适合在该专业领域上深入学习研究；在学生进入领域中时，帮助构建学生的知识体系，同时找到兴趣点，达到尊重学生个性的目的。现有的学生导师信息不对称容易产生良币驱逐劣币的现象，该研究致力于找到一个科学有效的评价系统和合作机制，一方面提升双方的信任度，一方面推动后期论文合作的高效开展。

这将不仅加速个人发展，促进合作交流，并且有利于和谐师生关系，一定程度上缓解研究生导师冲突的社会问题。习近平总书记在党的十九大报告中指出，人才是实现民族振兴、赢得国际竞争主动的战略资源。研究生作为国家创新拔尖人才的主要来源[1]，其教育与培养关系到我国“人才强国”与“教育强国”战略的实施[2]。该研究将对研究生的学业生涯产生十分重要的影响，进一步提高研究生的培养质量，从而为国家的建设输送更多优质人才。

1.3 研究思路与研究方法

本文将选择统计学为背景进行研究，首先考虑到统计学科拥有最新的关键词知识图谱，后续的所有研究将建立在这之上；其次不同于金融等热门学科，统计学科由于其难度需要学生有较强的数理背景，因此学生在学习研究上更需要规划指导；在大数据、人工智能的

浪潮下，近年来统计专业吸引了很多学生学习深造，逐渐变得热门，意味着本研究的受众将会很多，且时效性强。

目前，虽然大多数学校采用的研究生导师选择模式为“师生互选”这种双选模式，实际上研究生与导师之间存在上下位的关系，且最终选择的主导权在导师，造成入学后双方矛盾分歧的原因主要在于存在信息不对称[3]和培养理念不一致等问题。而目前已有的最优匹配方法都存在静态、主观的问题，匹配的都是过去时，如老师的论文数、科研项目数、地位，学生的初试、复试成绩等，而不是在专业知识上的交流碰撞，实际上这些都是外在的东西。研究生培养的目标应该是研究生和导师能在知识上形成合作，这也是一个好的师生关系的定义，只有知识作为基础，才能形成合作，而且知识掌握的情况，对知识的兴趣是会随时间而变化的，因此应该从知识的角度出发，动态预测研究生和导师的合作情况。本文尝试对知识进行画像，并基于知识进行推荐。

传统推荐可解释性较弱，关注不到内在的原因，而知识是有结构的，相互贯通的，知识图谱构建了实体与实体之间更深更远的联系，使得推荐算法的挖掘能力大大增强，从而提高推荐的精准性并达到多样化的结果，也有效地解决信息稀疏和缺失造成的冷启动问题，因此考虑将知识图谱应用于双选这一过程。

本研究将借助统计学科关键词知识图谱，通过 WOS 核心期刊网获取 Bin Yu 老师、Candes 老师两位国际知名统计学者的论文数据，提取出关键词信息，检索知识图谱得到其分布情况，并进行描述性统计分析。并在自建的网站上发布问卷，收集到来自国内统计学相关方向研究生和本科生对统计学科知识图谱中关键词的了解情况，得到分布。假定意向导师为 Bin Yu 老师和 Candes 老师，开发基于知识图谱的算法计算关键词分布之间的距离，分析样本学生知识结构的差异，并给样本学生在两位导师之间进行推荐。

2 文献综述

2.1 对学生和导师双选的研究

导师选择模式的研究有：李洁茗 (2010); 苗玥明 等 (2019) 研究结果表明，对于学生来说，由于缺乏科学指导，他们的选择是盲目的。“师生互选”经常导致扎堆现象。这是因为学生不清楚在选择导师时应考虑哪些因素，而只能根据外部的，自私的和每个人都接受的标准来选择导师；就导师而言，导师在教师和学生的相互选择上相对被动，并且与学生的互动受到限制，在老师和学生的相互选择中，导师通常只能通过学生提交的单一形式的“意向书”来了解学生，很少有学生有机会与他们的导师进行面对面的交流，由于时间太短，理解也可能是“肤浅的”；就教育公平而言，可能存在复杂的人际关系，资源失配和缺乏公平[4]。

关于导师和研究生的研究，对于导师与研究生之间的关系研究居多。如王志栋 (2006) 研究了影响研究生和导师双向选择的因素，得出的结论是影响研究生选择导师的因素有：学历、指导水平、研究方向、导师和毕业生的风格、学生自身的因素，影响导师选择研究生的因素是研究生的知识、兴趣倾向、人格和情商，影响研究生与导师双向选择的客观环境因素是信息获取环境和社会环境，并从导师招收研究生的角度提出了相应的对策[5]。任旭 等 (2020) 研究结果表明，导师和研究生的角色期望是一致的，“朋友型关系”与“合作伙伴型关系”是最为师生认可的 2 种关系；导师与研究生之间的角色互动受导师合作质量和研究生合作质量的影响；导师与研究生的角色互动中存在认知差异，主要是合作质量上的认知差异，合作情感上的认知差异和学科上的认知差异[2]。

对于研究生与导师的双向选择的最优匹配问题的研究有：徐豪华 等 (2005) 通过建立指标体系，获得了研究生对每个导师的满意度和导师对每个研究生的满意度，满意度矩阵乘以该值可用于评估研究生与导师之间的匹配度[6]。该方法存在的主要问题是评价指标的设计过于简单，存在无法量化的指标，且缺少专业性方面的评估。王红霞 等 (2007); 周琴 (2019) 使用层次分析法分析研究生和导师的匹配情况[8]。该方法存在的主要问题是更多是从宏观层面上给出建议，不涉及合作理论的应用。向冰 等 (2016) 在单个导师，系统具有足够的信息，并且确定了每个导师带的研究生人数的情况下，使用 Gale-Shapley 算法，以研究生先做出选择的方式最佳地匹配研究生和导师的双向选择[9]。该方法与上两个方法同时存在的问题是静态和主观，没有考虑到导师对研究生的培养是长期的过程，双方的关系是动态变化的，因此结论可靠性不高。黄宏涛 等 (2016) 基于协同推荐模型，构建面向导师和研究生的双向选择系统。根据用户访问项目的历史记录以及项目之间的相关信息，建立用户兴趣模型，对复杂信息进行过滤，向用户推荐感兴趣的对象[10]。该方法存在的主要问题是研究生导师双选的场景下，协同推荐并不适用。协同推荐的前提以商品推荐举例为：存在大量商品，目的是让用户能看到更多可能喜欢的商品，从而把更多商品卖出去，而导师

却是稀缺的，因此并不符合该限定。并且还存在无法处理冷启动、很难包含查询用户的侧面特征等问题。

2.2 知识图谱简介

2012 年 5 月 17 日，Google 正式提出了知识图谱（Knowledge Graph）的概念，其初衷是为了优化搜索引擎返回的结果，增强用户搜索质量及体验[11]。目前，随着智能信息服务应用的不断发展，知识图谱已广泛应用于智能搜索，智能问答，个性化推荐等领域[12]。知识图谱，本质上，是一种揭示实体之间关系的语义网络。信息是指外部的客观事实，知识是对外部客观规律的归纳和总结，在信息的基础上，建立实体之间的联系，就能行成知识，换句话说，知识图谱是由一条条知识组成[13]。知识图谱的数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储，如果用（实体 1，关系，实体 2）、（实体、属性、属性值）这样的三元组来表达事实，可选择图数据库作为存储介质[12]。

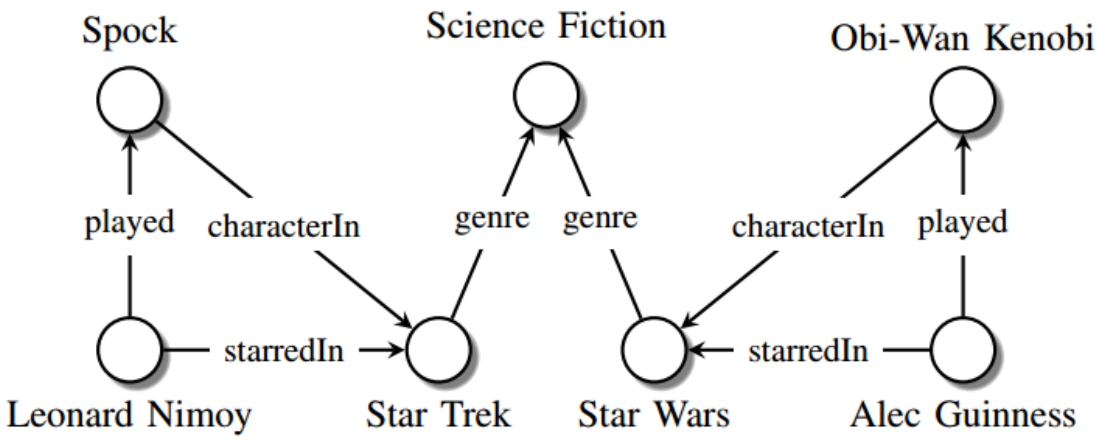

图 **2.1** 知识图谱示例

与语义网络的区别在于，知识图谱具有广泛的数据源，并且对诸如知识表示和知识融合之类的技术给予了更多关注。同时，知识图谱和知识库在理论和方法上也有许多相似之处。不同之处在于，知识库在某个机构领域中包含更多的知识，而知识图谱是一个更大的知识库的知识集合。知识图具有逻辑推理，可解释性，自然关联，有效资源发现，透明共享和可视化的优点，具有广阔的应用前景，目前典型的知识图谱主要有开放领域知识图谱(Freebase、Dbpedia、Wikidata、YAGO、Babel Net、Web Data Commons)、垂直领域知识图谱 (Linked life data、Linked movie data Set、Concept Net、Microsoft Concept Graph)、中文知识图谱 (Open KG、CN Dbpedia、Xlore、PKU PIE、Belief Engine)[14]。

知识图谱的分类根据研究内容进行划分可以分为文本知识图谱、视觉知识图谱和多模式知识图谱，文本知识图谱主要以文本为研究内容，并用文本样本构造而成，对文本知识进行知识表示和知识推理，它主要用于语义检索、深度搜索和情报分析；视觉知识图主要以图像为研究内容，以图像样本为基础，对图像进行知识表示、知识处理、推理更新等操

作，难以获得实体之间的复杂关系、建模困难，它主要用于语义图像检索，判断文本关系的正确与否等；多模态知识图的构建需要知识表示、知识推理和更新等操作，其构建过程的每个步骤都需要所有多模态样本，它在生活中具有更广泛的应用，例如实现视觉和文本相结合的知识问答等[14]。

知识图谱为互联网上海量、异构、动态的大数据表达、组织、管理以及利用提供了一种更为有效的方式，使得网络的智能化水平更高，更加接近于人类的认知思维[12]。目前迫切需要用于特定领域中的整合和相关性的资源，知识图谱可以提供更准确和标准化的行业数据以及丰富的表达，从而帮助用户更方便地获取行业知识。

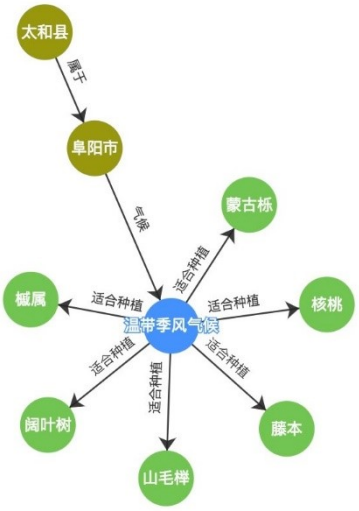

图 **2.2** 农业知识图谱示例

3 模型构建与模拟

3.1 数据说明

统计学科关键词知识图谱的数据来源为 2006-2018 年 Web Of Science 的 22 个统计学顶级期刊数据，共 31681 篇论文的基本信息，数据项包括论文标题、作者、简介、关键词等。经过同义词处理，后生成关键词的邻接矩阵。两个关键词之间存在一条边，如果这两个关键词至少在同一篇论文的关键词中出现。

表 **3.1** WOS 统计学代表性期刊

ID	Journal Directory	ID	Journal Directory
1	The Annals of Statistics	12	Machine Learning
2	Journal of the American Statistical Association	13	Bioinformatics
3	Journal of the Royal Statistical Society (Series B)	14	Journal of Machine Learning Research
4	Biometrika	15	Econometrica
5	Statistica Sinica	16	Journal of Econometrics
6	Scandinavian Journal of Statistics	17	Statistical Methods in Medical Research
7	The Annals of Applied Statistics	18	Technometrics
8	Bernoulli	19	The IEEE Transactions on Pattern Analysis and Machine Intelligence
9	Journal of Business & Economic Statistics	20	Psychometrika
10	Biostatistics	21	Test
11	Biometrics	22	Artificial Intelligence

3.1.1 度校正混合会员概率归属估计模型 DCMM

度校正混合会员网络提取模型（Degree-Corrected Mixmembership Model, DCMM）是 Jiashun Jin(2016,2018) 社会网络统计分析模型基础上发展起来的一种以估计混合社群归属概率为估计目标的研究方法，旨在通过稀疏随机矩阵方法估计社群结构以形成节点社群归属概率的方法。该模型不仅考虑了网络社群结构影响，同时也分析节点本身的会员属性，能够较为全面地揭示大型网络中节点是主要受到每一个社群内结构影响还是用于连接不同片区的结构，提供可能性的作用机理，DCMM 常被用于解释社会网络的形成机制。例如，针对合作者引用网络，根据合作者不同研究兴趣网络形成过程中的合作效应、传承机制和主题扩展机制[15]。Jiashun Jin 和他的团队逐渐将 DCMM 用于知识网络的研究中，从知识网络结构特性、社会因素、语义因素等方面探讨合作者网络和引文网络等的形成机制。

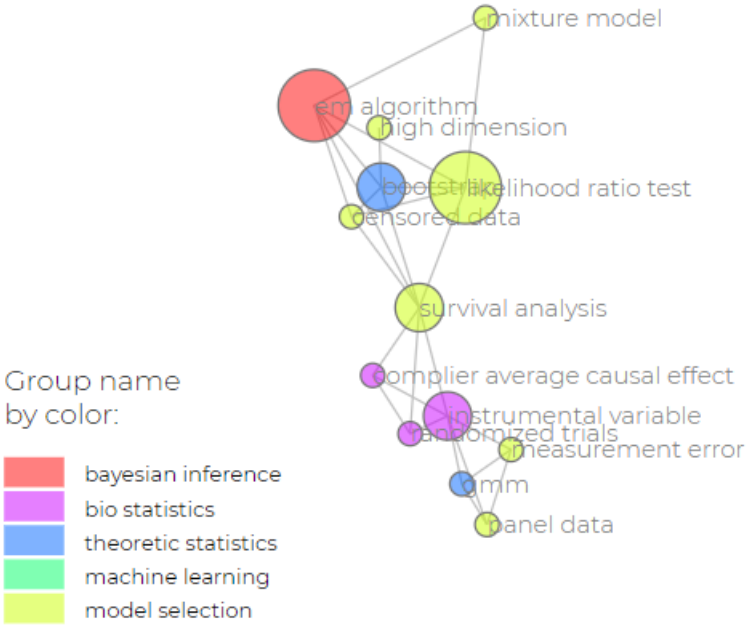

图 **3.1** 统计学科关键词知识图谱局部可视化

然而，目前尚未有利用该模型和算法研究关键词共词网络的形成过程的研究，共词网络是计量学领域分析学科生态和研究结构的基本数据类型，有研究认为它比合作者网络和引用网络更加稀疏，估计难度更大。鉴于 DCMM 的优势和当前研究现状，统计学科知识图谱作者借助 DCMM 网络模型揭示统计学内部不同领域的共词网络生成机制，期望较为全面地剖析网络结构、学科属性、时间等多个因素对学科内不同领域知识系统形成的影响。

3.1.2 混合成员谱聚类算法 (Mixed-SCORE)

Mixed-SCORE 算法[16] 是专门针对网络结构稀疏，需要通过纯节点来勾勒出社区边界进而将混合节点和纯节点进行区分的方法，它适用于节点度具有较大异质性的全连通网络，本质上是通过节点的混合归属概率估计来进行混合和单纯网络结构功能的区分。稀疏网络中，由于很多节点的度都很小，运用传统的谱聚类算法很难有效的社区划分，估计其异质性参数误差很大。与经典谱聚类方法相比，Mixed-SCORE 算法的优点在于它是以主特征向量和其他特征向量的元素比例为聚类依据，很大程度上抑制了异质节点的强干扰性，还实现了将具有混合社区属性的节点嵌入到单纯社区属性所围成的轮廓之中，从而可用于理解网络结构的微观连接机制。Jin J, Ke Z (2018) 文章里中对度极不平衡的节点的会员估计算法的一致性进行完整的理论分析。

3.1.3 统计学科知识图谱

构建该知识图谱时首先要求它是全连通的，其次还要覆盖大部分的关键词，而且该知识图谱是要可用于中国人的，也就是说这些关键词是外国人和中国人所发论文中都有的。

使用 topic-coherence 得到最佳主题数为 5，按照节点数由多到少依次为主题命名为：模型选择（Model Selection），机器学习（Machine Learning），统计理论（Theoretic Statistics），贝叶斯推断（Bayesian Inference）和生物医学统计（Bio Statistics）。

因此，在最大的连通子图上考虑一个 DCMM 模型，假设（a）有 5 个社群，分别称为 bayesian inference, bio statistics, theoretic statistics, machine learning, model selection 以及（b）一些节点在 5 个社区中具有混合成员身份。使用 mixed-score 法，得到七百多个关键词和它们在 5 个社群上的 mixed-score，即归属度。之后为这些词人工贴上标签，是属于 level1（基础知识）, level2（专业知识）还是 level3（研究热点），得到经过专家加工的数据，部分数据展示见表3.2。

老师近期研究方向的数据输入为 Web Of Science 核心期刊上近六年来为论文作者的文章关键词词频统计。

表 **3.2** 部分数据展示

Keyword	Community Membership	Model Selection	Machine Learning	Theoretic Statistics	Bayesian Inference	Bio Statistics	Type
multiple imputation	Bio Statistics	0.249	0	0	0.105	0.646	1
meta-analysis	Bayesian Inference	0.244	0.012	0	0.428	0.316	3
imputation	Bio Statistics	0.291	0	0	0.16	0.549	2
sparse representation	Machine Learning	0.235	0.765	0	0	0	3
causal discovery	Bio Statistics	0	0.296	0.015	0.262	0.427	2
compressed sensing	Machine Learning	0.329	0.671	0	0	0	2
marginal likelihood	Bayesian Inference	0.372	0	0	0.556	0.072	3

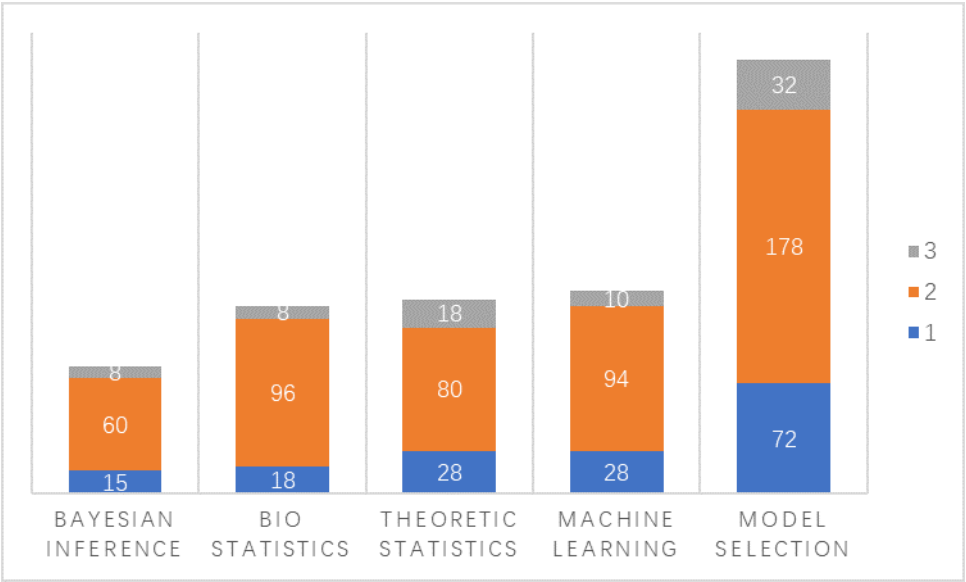

图 **3.2** 统计学科知识图谱中 5 个专业方向上关键词的分布情况

样本数据来源于自建网站上的问卷调查，调查分为两个部分，基本信息采集和学科知识了解情况。其中，学科知识了解情况通过受试者在网页上点选自己掌握的关键词知识得到。输入模型的数据为所有关键词及 0-1 标签，1 表示受试者了解掌握该关键词，0 表示不熟悉。

图 **3.3** 关键词点选页面示例

3.2 模型构建

3.2.1 符号说明

n 表示统计学科知识图谱的节点数
S_i 表示学生点选第 i 个节点
CM_i 表示第 i 个节点的专业特征 ($CM_i = 1, 2, 3, 4, 5$)
L_i 表示第 i 个节点的层级 ($L_i = 1, 2, 3$)
A 表示学生点选关键词的集合
O 表示老师论文关键词的集合
M 表示混杂度大的关键词的集合
T_i 表示老师论文涉及第 i 个节点
H_i 表示老师论文涉及第 i 个节点的频数

3.2.2 假设

这个推荐模型是针对研究生期间的目标为拿到学位的基础上，在导师指导下完成一篇高质量的硕士论文，并获得老师的认可，以及在这两个目标的基础上，为获得国内外一流大学博士学位做好准备，打好基础的申请者。所以样本学生进入的规则为至少掌握 10 个 level2 关键词。

同时，考虑申请者有主观的初筛等前期准备，最后保留下两位意向导师，模型在其中进行推荐。

要想达到学生和导师高效合作的目标，需要从师生关系的角度考虑。作为申请者，二者中较为主动的一方，要从导师的角度出发，这样申请的成功率才会更高。调查研究发现，在师生互动中，导师对双方的合作素质更为看重，其认为自身对于研究生的指导与管理，以及研究生所具备的科研素质等更能影响二者的关系构建[2]。这意味着，学生的科研素质起着极为重要的作用，而科研素质又建立在基础性知识掌握的前提下。并且，对导师而言，其对自身的角色认知仍以“教育者”为主，并期望研究生不仅作为其学生，还能够作为其科研伙伴，从而构建以科研为基础的师生关系[2]。这就要求学生在老师研究的前沿也要有所了解。因为老师的研究也会有所变化，主要是近六七年发表的文章有参考价值。

因此在选择的时候遵循以下原则和流程，首先是需要有扎实的基础知识，然后是较丰富的专业知识，另外还要对导师近期研究的方向有所了解。

考虑到申请者的背景可能很丰富，存在跨专业申请等情况。如果申请者在导师近期论文中高频出现、混杂度大的关键词其对应的大的专业领域上有所知识储备，将更有利于日后的研究[17]。因为合作的前提是存在交集，混杂度大的关键词意味着同时涉及多个领域方向，即使只对其中某一个比较了解，也可以就此与导师进行合作，在过程中再对其他方向的知识进行补充，所以混杂度大的关键词上更容易形成合作。

基于上述假设，在计算时给予 Level1、Level2、Level3 的关键词不同的权重。

3.2.3 KL 距离

KL 距离，是 Kullback-Leibler 差异 (Kullback-Leibler Divergence) 的简称，也叫做相对熵 (Relative Entropy)。它衡量的是相同事件空间里的两个概率分布的差异情况。其物理意义是：在相同事件空间里，概率分布 $P(x)$ 对应的每个事件，若用概率分布 $Q(x)$ 编码时，平均每个基本事件（符号）编码长度增加了多少比特。我们用 $D(p||q)$ 表示 KL 距离，计算公式如下：

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)} \tag{3.1}$$

其中，$p(x)$ 与 $q(x)$ 是两个概率分布，KL 距离的大小反映了分布之间的远近。

3.2.4 推荐算法——基于研究领域的匹配

用 $p(x)$ 刻画知识结构的分布，则 $p(x)$ 的计算公式如下：

$$p_j = \frac{\sum_{i=1}^{n} L_i I(S_i = 1, CM_i = j)}{\sum_{i=1}^{n} L_i I(S_i = 1)} (j = 1, 2, 3, 4, 5) \tag{3.2}$$

每个人的知识结构就对应了这样的一个离散分布，用 $q(x)$ 刻画老师研究领域的分布，则 $q(x)$ 的计算公式如下：

$$q_j = \frac{\sum_{i=1}^{n} L_i I(T_i = 1, CM_i = j) H_i}{\sum_{i=1}^{n} L_i I(T_i = 1) H_i} (j = 1, 2, 3, 4, 5) \tag{3.3}$$

通过两位导师的论文关键词数据可以得到他们研究领域的分布，同时样本在网页上通过点选也得到了知识结构的分布，比较样本分布与两位导师的分布的距离远近。这里使用 KL 距离来定义两个分布的距离，KL 距离的大小将用来判断样本更适合与哪位导师合作论文。

及，如果

$$D(Student||Teacher1) < D(Student||Teacher2) \tag{3.4}$$

则样本的知识结构与 Teacher1 的研究领域更近，更适合选择 Teacher1 作为导师。

3.2.5 推荐算法——基于研究关键词的匹配

V 为 S 与 O 的交，即学生掌握的知识与老师所发论文涉及的知识的交，意味着在该关键词上，双方可以进行合作。交的关键词越多，关键词的 Level 更高，则合作越顺利；根据理论，混杂度大的关键词上更容易形成合作，且是跨学科合作的切入点，意味着混杂度大的关键词相较纯的关键词更容易产出论文。因此交的关键词情况、混杂度大的关键词占所有交的关键词比例衡量了学生与老师合作的效率预期。

定义合作效率预期的计算公式如下：

$$V = A \cap O \tag{3.5}$$

$$\text{Score} = \frac{\sum_{i=1}^{n} L_i I\left(S_i = 1, T_i = 1\right) H_i}{\sum_{i=1}^{n} L_i I(T_i = 1) H_i} + \frac{\sum_{v \in O} I(v \in M) I(v \in V)}{\sum_{v \in O} I(v \in V)} \tag{3.6}$$

如果样本与 Teacher1 计算得出的 Score 大于与 Teacher2 计算得出的 Score，则更适合选择 Teacher1 作为导师。

3.3 模拟实验

选择声誉差不多、年龄相近、性别不同、国籍不同、不同学校的两个老师 Bin Yu 和 Candes 作为意向导师。

Bin Yu 老师是加州大学伯克利分校统计系和电子工程与计算机科学系的校长教授。她目前的研究兴趣集中在统计学和机器学习理论、方法学以及解决高维数据问题的算法上。她的团队与来自基因组学、神经科学和遥感的科学家一起从事跨学科研究。

Candes 老师是数学和统计学教授，也是斯坦福大学的电气工程教授。主要工作是在数学、统计学、信息论、信号处理和科学计算的界面上，寻找新的信息表示方法，从复杂数据中提取信息。他在过去十年左右的工作主要是统计性质的。令人感兴趣的是机器学习的最新进展，它为我们提供了许多潜在的有效工具，可以从不断增加的数据集中学习并做出有用的预测。其中一些工具已被证明是强大的，同时又极其复杂。最近的工作是开发了广泛的方法，可以包装在任何黑匣子周围，以产生可信任的结果。

3.3.1 描述性统计

从两位老师历年来所发论文的关键词分布来看，Bin Yu 老师相比 Candes 在 Bio Statistics 和 Theoretic Statistics 上的研究更多，而 Candes 老师在 Model Selection 和 Machine Learning 上的研究更多。不过总体来看，两位老师主要的研究方向都在 Machine Learning 和 Model Selection 上。

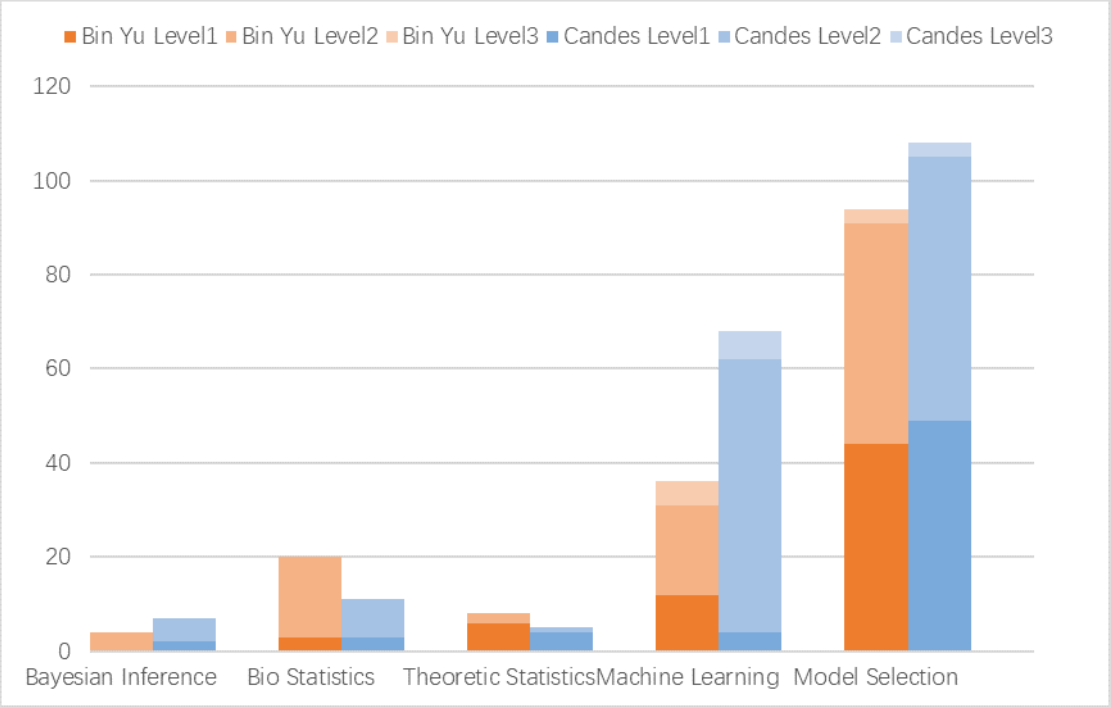

图 **3.4** 历年来两位老师的研究方向

从近六年来两位老师的研究方向上看，Candes 老师主要的发文方向为 Model Selection，且明显高于在其他方向上的频率。Bin Yu 老师在四个方向上则比较均衡，且可以看到研究方向从过去发 Model Selection 上的文章更多变成发 Machine Learning 更多。且在 Bio Statistics 上发的也比较多，从个人简介上来看，她的团队确实与来自基因组学、神经科学

的科学家一起从事跨学科研究。这两位老师近年来都没有在 Bayesian Inference 方向上发文。

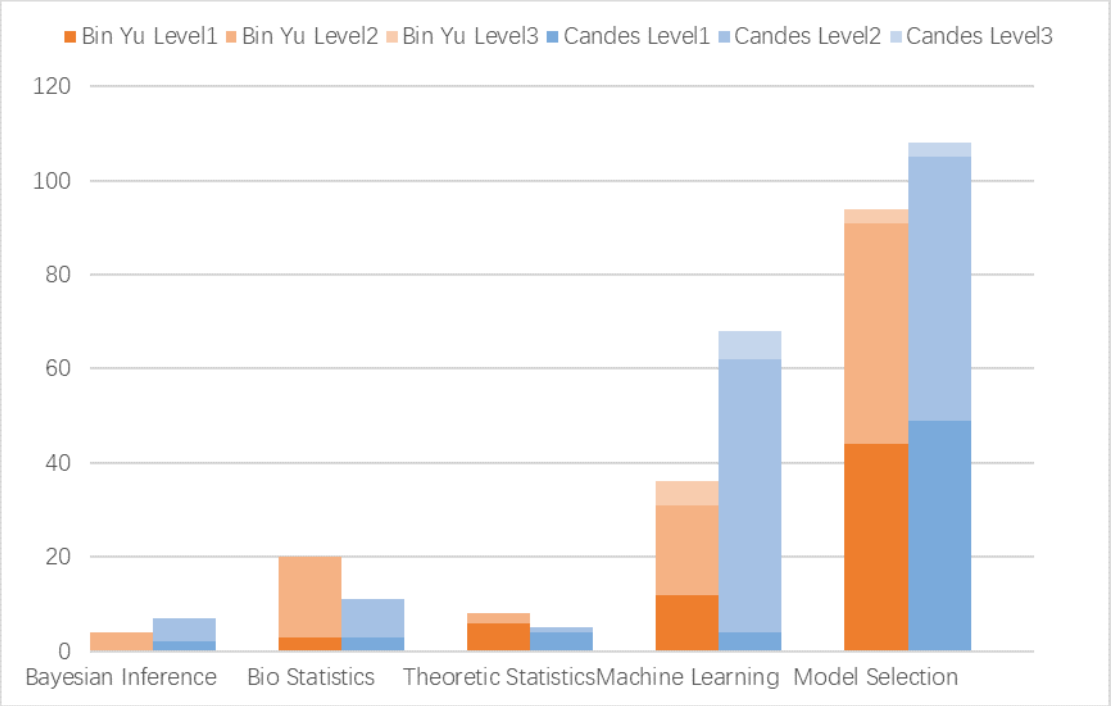

图 **3.5** 近六年来两位老师的研究方向

从网页问卷上收集的数据中选取的样本学生的知识结构见图3.6。

可以看到除了在 model selection 这个本身比较大的方向上掌握的知识最多，另外在 Bayesian Inference 和 Bio Statistics 上的知识丰富，相比在 Machine Learning 上的知识则较少。接下来将在两位老师中对其进行推荐，首先根据上文提到的计算方法得到两位老师研究方向的分布和样本知识结构的分布，对比历年来和近六年来的情况，结果如图3.7、图3.8。

3.3.2 基于研究领域的匹配

已知分布，根据 KL 距离计算公式，得到两位老师近六年来研究方向和样本知识结构的距离比较如下:

$D(BinYu||Sample) = 0.522, D(Candes||Sample) = 0.365$

两位老师历年来研究方向和样本知识结构的距离比较如下：

$D(BinYu||Sample) = 0.254, D(Candes||Sample) = 0.491$

考虑合作的时效性的话，即与老师近期发表论文方向更匹配的角度来说，推荐该样本与 Candes 老师合作；如果考虑合作时的沟通效率的话，因为知识结构与 Bin Yu 老师历年研究方向更接近，意味着研究背景、经历、兴趣更相近，合作时交流可能更轻松方便，心理负担也较小。

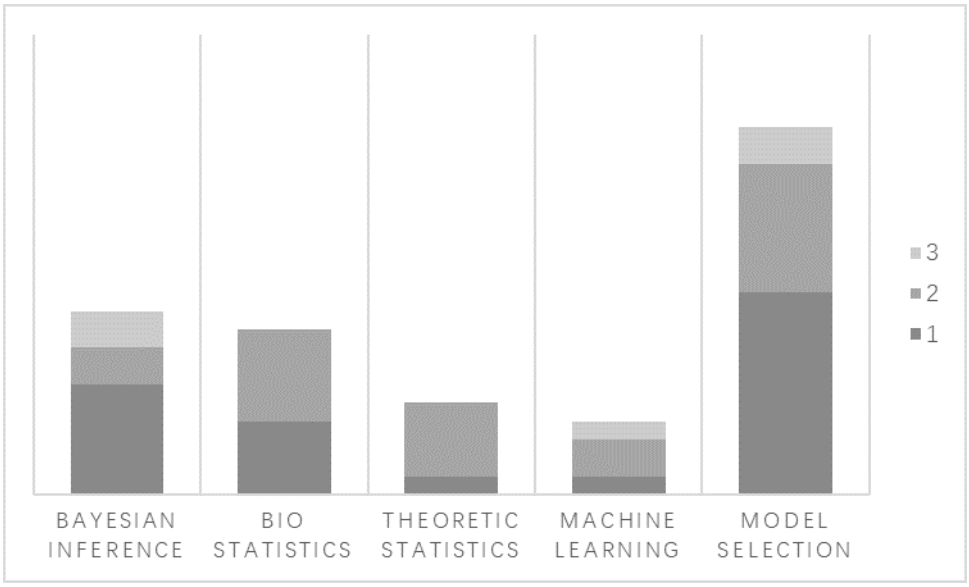

图 **3.6** 样本学生的知识结构

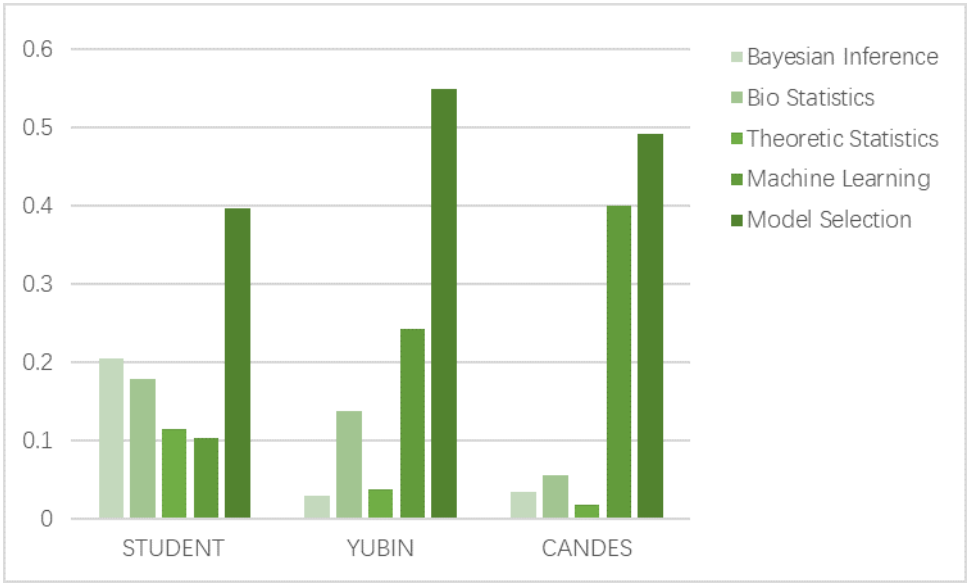

图 **3.7** 历年来研究方向的分布与样本知识结构

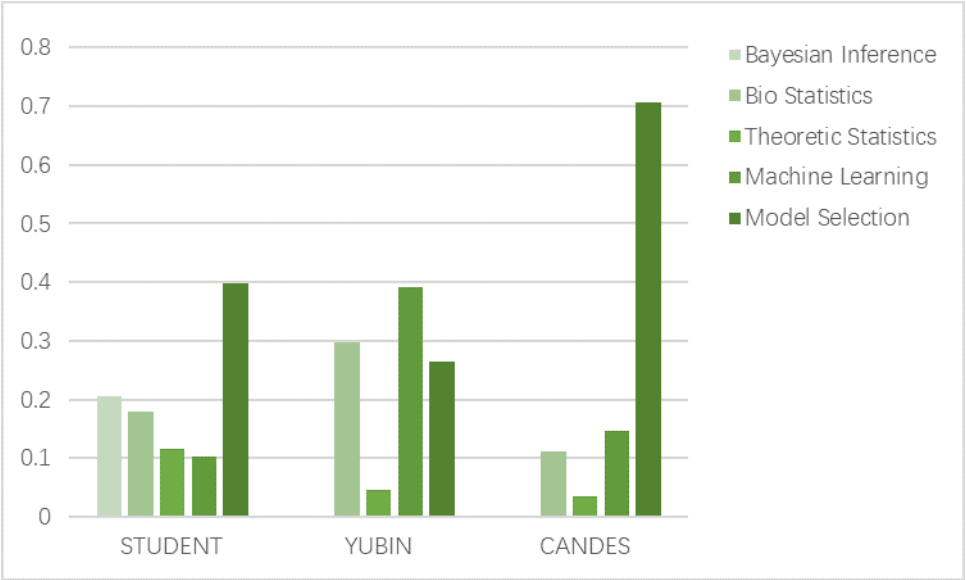

图 **3.8** 近六年来研究方向的分布与样本知识结构

3.3.3 基于研究关键词的匹配

两位老师近年来所发论文关键词的词云图如下：

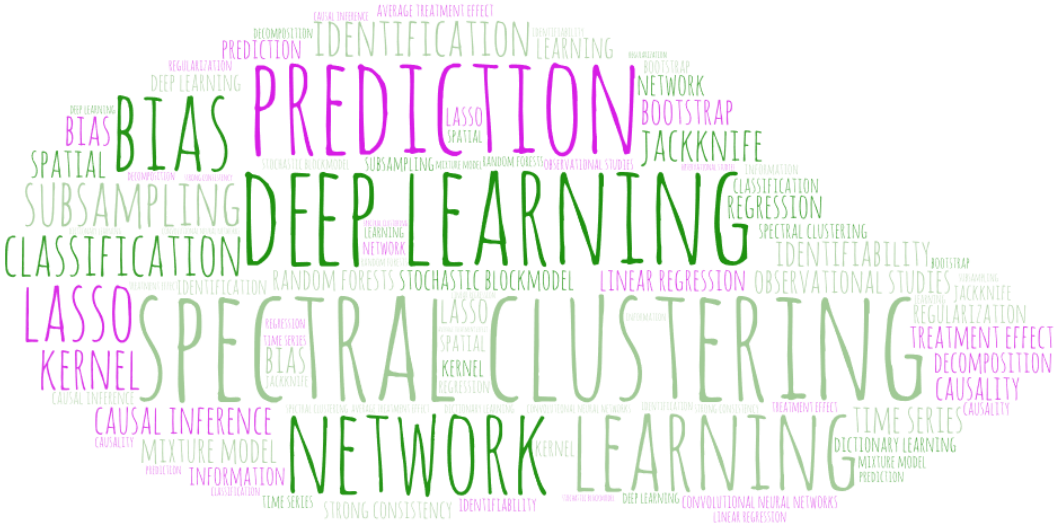

图 **3.9** Bin Yu 老师的词云图

合作预期效率的计算结果为：

$Score(withBinYu) = 0.234, Score(withCandes) = 0.457$

样本在 level1 与 level2 词语中与两位导师的交的个数差别都不大，整体上与 Bin Yu 老师的交集稍多。可以看到样本在两位导师论文高频词中的知识掌握情况差别也不大，但 Bin Yu 老师近些年来并没有明显高频的词语，也即发文方向比较广。而 Candes 老师有明显的主要研究方向，如 False Discovery Rate、Variable Selection、Lasso 等，而在这些方面

图 **3.10** Candes 老师的词云图

上该样本申请者并没有充足的知识基础。混杂度大的关键词 membership 主要是在 Model Selection 和 Machine Learning 两个方向上，样本与 Candes 老师所交关键词中这样混杂度大的占到四分之一，而与 Bin Yu 老师所交关键词则比较纯。最终得到样本与 Candes 老师的合作效率预期要高于与 Bin Yu 老师。

4 总结与建议

研究生的知识掌握情况和研究兴趣是会随时间动态变化的，导师的研究方向亦是如此。目前已有的研究生导师最优匹配研究都是从静态出发，匹配的是过去的状态，而不是去预测未来。把研究生和导师合作出高质量的论文作为目标的话，应该预测未来研究生和导师是否是一个好的匹配。本文考虑从知识的角度进行画像，然后形成推荐，达到动态预测的目的。

传统的基于协同过滤算法的推荐系统需要大量的数据支持且可解释性较差，而知识图谱可以从特征、结构、可解释性、语义等方面优化。对于研究生申请导师这样一个信息不是很充分公开的过程，因为导师是稀缺资源，也不会有大量如浏览物品产生的偏好标签等数据支撑，即使有申请成功的案例参考，但申请者的详细知识结构和研究经历并未披露，参考价值并不大。为了打破这种信息不对称，考虑通过第三方的知识图谱工具来进行匹配，搭建起学生和老师之间的桥梁。

本文以统计学科为例，借助统计学科知识图谱进行基于研究领域及研究关键词的推荐。以样本为例，在国际知名统计学者 Bin Yu 老师和 Candes 老师中给其推荐合适的论文合作者，并从历年来研究方向和近六年研究方向，近六年研究关键词中相交关键词层级、混杂度等角度出发给出推荐理由。研究定义了知识结构和研究方向匹配公式及合作效率预期得分的计算方法，并应用到实际数据中，进行推荐。

未来会继续完善自我评估的功能，系统使用者可以从点选的关键词看到自己的知识结构在整个图谱中的分布情况，将知识图谱与推荐系统结合，提供个性化的提升路线定制服务。

合作中的学生可以借助该系统提供的匹配度与心理主观感受是否相符，并对以后的学习研究进行规划，考虑选择合作对象的学生也应该利用该工具进行科学的选择与评估。

参考文献

[1] 施鹏, 张宇. 论研究生教育中和谐师生关系及其构建路径[J]. 学位与研究生教育, 2015 (05): 37-41.

[2] 任旭, 于倩, 王康冉, 等. 基于角色理论的医药类研究生交互合作型师生关系研究[J]. 药学教育, 2020, 36(03): 71-78+83.

[3] 李洁茗. 从信息不对称理论看研究生导师双向选择[J]. 新西部, 2010(12): 178+171.

[4] 苗玥明, 肖磊. 硕士生“导师选择模式”探析[J]. 天津市教科院学报, 2019(04): 42-47.

[5] 王志栋. 硕士研究生与导师双向选择影响因素分析[J]. 山西医科大学学报 (基础医学教育版), 2006, 8(03): 322-324.

[6] 徐豪华, 董志明, 唐志武, 等. 研究生录取的优化模型[J]. 数学的实践与认识, 2005(07): 115-119.

[7] 王红霞, 朱喜林, 赵丽平. 研究生录取问题的数学建模[J]. 太原理工大学学报, 2007(05): 467-470.

[8] 周琴. 层次分析法在考研专业选择中的应用[J]. 高师理科学刊, 2019, 39(09): 36-39.

[9] 向冰, 刘文君. 硕士研究生与导师的双向选择的最优匹配[J]. 未来与发展, 2016, 40(04): 91-94.

[10] 黄宏涛, 骆玉璞, 彭利园, 等. 基于协同推荐模型的导师研究生双向选择系统[J]. 软件导刊, 2016, 15(05): 74-75.

[11] 罗丹. 一种基于表示学习的知识图谱融合算法与系统实现[D]. 中国: 浙江大学, 2018.

[12] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报(4 期): 589-606.

[13] 丑晓慧. 面向中文知识图谱构建中的知识推理方法研究[D]. 中国: 国防科学技术大学.

[14] 欧艳鹏. 知识图谱技术研究综述[J]. 电子世界, 2018, No.547(13): 56+58.

[15] JIN J, KE Z, LUO S. Score+ for network community detection[Z]. 2018.

[16] JIN J, KE Z T, LUO S. Estimating network memberships by simplex vertex hunting[Z]. 2019.

[17] 赵秀敏. 寻找跨学科研究旨趣的交集——建筑学研究生跨学科国际合作培养模式探索[J]. 高等工程教育研究, 2004(05): 77-81.