西北大学信息科学与技术学院

本科毕业设计开题报告/答辩登记表

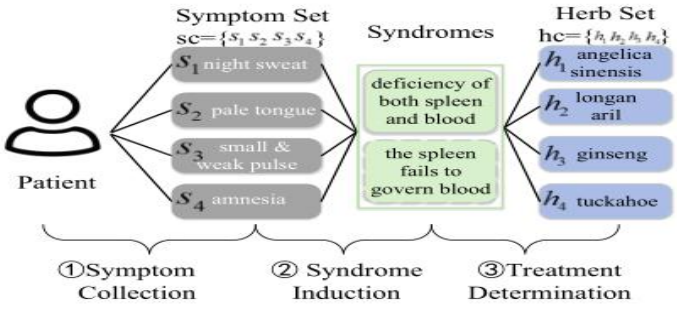
<table>
<tr><td>学生学号</td><td>2017111093</td><td>姓名</td><td>邹刘文</td><td>年级</td><td colspan="2">2017 级</td></tr>
<tr><td>专业</td><td colspan="6">软件工程</td></tr>
<tr><td>论文（设计）题　　目</td><td colspan="6">基于深度学习的中药推荐方法的研究与实现</td></tr>
<tr><td>指导教师姓　　名</td><td>张海波</td><td colspan="2">专业技术职务</td><td>副教授</td><td>开题报告日期</td><td>2021-1-13</td></tr>
<tr><td>企业导师姓　　名</td><td></td><td>文献综述成绩</td><td colspan="2"></td><td>开题报告成绩</td><td></td></tr>
<tr><td colspan="7">答辩小组成员（姓名，职称）：</td></tr>
<tr><td colspan="7">答辩小组组长签字：　　年　　月　　日</td></tr>
<tr><td colspan="7">开 题 报 告 内 容</td></tr>
<tr><td>选题来源</td><td colspan="6">1．教师指定（√）2．教师课题（ ）3．创新基金项目（ ）4．自选（ ）</td></tr>
<tr><td>设计选题的背景与意义、理论与实证准备、拟解决的问题、研究（设计）方法与技术路线</td><td colspan="6">1．设计选题的背景与意义
1.1 选题背景
中医，即由中国汉族劳动人民创建的传统医学，在现代社会中依然发挥着至关重要的作用。中药的治疗过程包含征集病人症状、从症状归纳出综合症、确认治疗方案三个核心步骤，其本质就是草药推荐，即根据一系列症状来推荐一组用于治疗的草药。现有的几种机器学习方法对草药和症状间的相互作用进行了建模，但是没有借鉴医生总结症状进而归纳出综合症这一核心做法。我们考虑到中医理论的整体性思想，将其纳入推荐模型的设计当中，并与其他经典算法对比，以期获得更科学合理的推荐效果。
论文选题有一定的应用价值
 

Fig. 1. An example of the therapeutic process in TCM.
1.2 选题意义
利用数据分析技术，在已有的经典症状-处方数据集上进行推荐模型的训练</td></tr>
</table>

和优化，我们能够提供智能化的中药推荐。其意义主要体现为：（1）具有广泛的应用：既能辅助专家医生开具中医处方，从而缓解专家医生稀缺的压力；又能提高普通人对中医药材的认识和作用，通过日常饮食来辅助疾病的预防和诊断，从而有效地减少医疗花费和提高治疗效率。（2）有助于让更多普通人了解和接受中医，在一定程度上推动了中医的传承、推广和创新、以及中医现代化。

2. 研究现状

2.1 推荐系统

2.1.1 目的

为了避免信息过载[1]，人们提出使用推荐系统的解决方案。

2.1.2 推荐任务

评级预测、排名预测（top-N 推荐）和分类三种。评级预测旨在填补用户项目评级矩阵中的缺失数据，top-N 推荐则是为用户生成包含有 N 个项目的推荐列表，分类任务旨在将候选项目分类为正确的推荐类别。

2.1.3 根据推荐策略，可以划分为：

（1）传统的推荐系统

包括基于内容的推荐[2]、协同过滤[3]以及混合推荐[4]三类。基于内容的推荐系统主要是基于用户偏好和项目的特征信息进行推荐，无需大量评分记录，但需要如文本，图像和视频等辅助信息；协同过滤推荐主要分为启发式和基于模型两类，前者通过用户和项目的显式或者隐式的历史交互信息计算相似度和效用，后者构建用户偏好模型展开预测；混合推荐指通过后融合、中融合、前融合等方式来组合推荐算法。

虽然这些推荐在实际应用中取得非常好的效果，但是在处理数据稀疏性、“冷启动”问题、推荐的有效性和精确度等方面，都存在着不同程度的局限性[5,6]。

（2）基于深度学习的推荐系统

其经典架构包含输入层、模型层和输出层。输入层包含用户的显示或隐式反馈数据、用户画像和项目内容以及评论等辅助信息；模型层使用各种深度学习模型；输出层将用户和项目的隐表示通过内积、softmax 等方法产生推荐列表。

由于深度学习的突出优势是以一种端到端的过程自动学习到数据的高层次特征表示形式[7]，其主要应用于基于内容的推荐系统中。具体而言，推荐中常用的深度模型有四种。

一是多层感知机 MLP，一方面其用于获取用户或项目的非线性的隐表示，比如 Cheng 等[8]在 Google Play 的 App 推荐中使用了深广学习（Wide&Deep Learning）模型以提高特征提取的记忆和泛化能力；另一方面其可以建模用户和项目的交互函数，例如 Deep Crossing[9]、Deep&Cross[10]、DEF[11]和 DIN[12]，以此来缓解输入特征的稀疏性和离散性问题。二是卷积神经网络，例如 Gong 等人[13]提出了一种基于注意力的卷积神经网络（CNN）来进行微博中的 HashTag 推荐，Lei 等人[14]在图像推荐的研究中利用 CNN 以及 MLP 产生用户和图像的隐表示并运用于比较深度学习方法（Comparative Deep Learning, CDL)中。三是循环神经网络，例如 Li 等人[15]提出了一种基于注意力的 LSTM 来进行微博中的 hashtag 推荐，RNN 与注意力的结合能够抓住文本序列特征并识别出最具信息量的词；Okura 等人[16]利用 RNN 与降噪自编码器（DAE）来进行新闻推荐。四是深度信念网络 DBN，Wang 等人[17]通过其与概率矩阵分解（Probabilistic Matrix Factorization,PMF)将特征提取和推荐两个过程统一起起来用于音乐推荐。

2.2 草药推荐

推荐系统的常用业务场景是针对某一用户推荐商品和服务，而现有的草药推荐通常是针对某一特定疾病领域，例如王等人[18]对根据舌苔图像来进行中药材推荐的研究，但此处研究的是针对一组症状推荐一组草药。

（1）传统的概率统计和数据挖掘方法

主要包括关联分析、简单的聚类和分类，但仅适用于单一疾病

（2）主题模型

主题模型用于处理自然语言形式的中药处方，其中处方是包含草药和症状等单词的文档。其主要思想是同一主题下的草药和症状相似。Ma 等人[19]提出了一个“症状-综合症”模型来挖掘症状和潜在综合症主题之间的相关性。 Ji 等人[20]认为“发病机制”是将症状和草药联系起来的潜在主题。Wang 等人[21]设计了一个不对称概率生成模型来同时对症状、草药和疾病进行建模。 Chen 等人[22]和 Wang 等人[23]将中医领域知识引入主题模型，以捕获草药相容性规律。

不幸的是，标准主题模型对短文本不是很友好。因此，处方的稀疏性[24]将在一定程度上限制主题模型在大规模处方上的性能。此外，他们无法全面分析各个实体之间的复杂相互关系。

（3）基于图神经网络

图是建模复杂关系数据的有效工具。基于图表示学习的草药推荐是当今研究的热点，其重点是获取中药实体的低维表示，然后基于嵌入来推荐草药。一些研究已经将深度学习技术引入了基于图的处方挖掘中。 Li 等人[25]采用 BRNN[26]对中医文献中的草药词进行文本表示学习，以达到补充治疗的目的。 [24]，[27]将自动编码器模型与元路径集成，以挖掘 TCM 异构信息网络。

上述基于图的模型的弱点在于，所应用的深度学习技术最初是为欧几里德空间数据设计的，并且缺乏非欧几里德空间图数据的可解释性和推理能力。

3.理论与实证准备

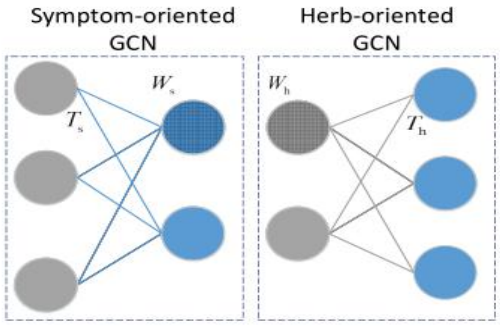
3.1 理论准备

3.1.1SMGCN 模型

（1）Bipar-GCN 二部卷积图

构建此组件原因： 建模草药和症状两种实体间的关系

组件结构：



说明：　A.图卷积通常有两类，一类是谱域图卷积，即利用图上的傅里叶变化将空间域的卷积转化到谱域上，如 SCNN、ChebNet、GCN；另一类是空域图卷积，即利用结点在空间中的邻居结点聚集信息，与矩阵的卷积很类似，如 GraphSAGE、GAT、PGC 方法；这里使用的是后一类； B.具体的卷积过程：选择邻居结点、构造消息、聚合消息、高阶传播； C.草药嵌入和症状嵌入的方法略有不同，区别在于第 0 层即初始嵌入层的对象不同

（2）SGE（Synery Graph Encoding）协同图编码

构建此组件原因：需要建模同种实体间的关系；引入更多信息有助于缓解中药处方的数据稀疏性问题

组件结构：　草药-草药图，症状-症状图

说明：　草药-草药图、症状-症状图实际上是同质图，不过在逻辑上可以看成是同类实体构成的特殊二部图；将处方中草药间和症状间同时出现的次数作为衡量协同效果的强度，设置相应的阈值，超过阈值才被视作具有协同效果

（3）信息融合

将从 Bipar-GCN 和 SGE 获取的嵌入进行相加，得到合并后的草药和症状嵌入

（4）MLP 综合症感知层

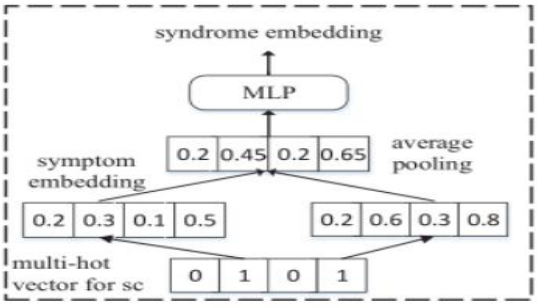


Fig. 4. The MLP-based method for syndrome induction.

先对所有症状的合并嵌入使用一层平均池化层得到平均嵌入向量，再用一层 MLP 获取综合症嵌入

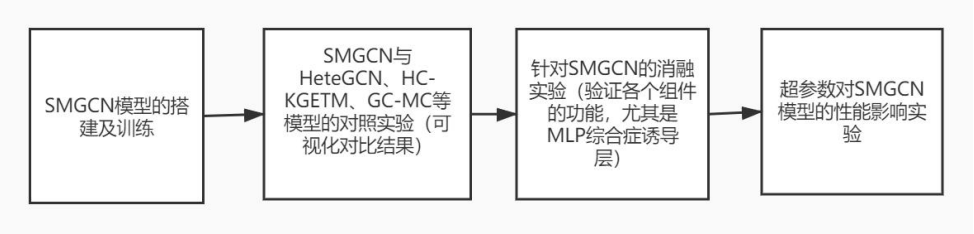
（5）模型训练

将一组草药推荐的问题视为多标签分类，使用多标签损失函数进行训练

3.1.2 其他对照模型（基线）

（1）主题模型 HC-KGETM

	将从中医知识图谱中获取的 TransE 嵌入集成到主题模型中，不仅考虑中医处方中的共现信息，还考虑知识图谱中症状和草药的全面语义相关性 （2）图神经网络模型 GC-MC 模型利用 GCN 获得症状和草药的表示，并对其设置了一个图卷积层； PinSage 是 GraphSAGE 在项目-项目图上的工业应用，这里我们将其应用于症状-草药相互作用图，并在其之后设置两个卷积层； NGCF 是目前基于图的最好的协同过滤方法。显式构造了一个二部症状-草药图，以对高阶连通性进行建模，以此为症状和草药获得更具表达力的嵌入形式； HeteGCN 用于基于内容的异质图表示学习问题，主要过程包括重定位的随机游走采样邻居节点、聚合异质的节点内容嵌入、聚合同一类型的邻居节点以及聚合不同类型的邻居节点。这里我们将症状-症状、草药-草药、症状-草药三种图合并为一个异质图。 3.1.3 推荐性能指标 我们在实验中使用准确率（Precision）、回召率（Recall）和归一化折损累计增益（Normalized Discounted Cumulative Gain, NDCG）来作为推荐性能指标。 3.2 实证准备 （1）数据准备 数据来自于基准中医数据集，其包含 98,334 份原始医疗处方和 33,765 份经过处理的医疗处方（仅由症状和草药组成）。为了获取更多数据集，可能需要进行爬虫 （2）代码准备 通过对现有的开源算法进行分析学习，融合本选题背景设计和实现核心算法。 4.拟解决的问题 （1）质量可靠的中药症状-处方数据集获取 （2）SMGCN 算法模型中单个节点公式向矩阵形式的推导 （3）SMGCN 模型的算法实现及其消融实验的展开 （4）对照模型 HeteGCN、HC-KGETM、GC-MC 等的算法实现 （5）模型间的性能比较及其可视化 5. 研究（设计）方法与技术路线 5.1 研究方法 （1）阅读文献法：通过阅读大量深度学习推荐算法、草药推荐任务等方面的文献，发现了图表示学习算法在获取高质量的实体嵌入进而提高推荐效果方面的优异性能，我们认为其适用于草药推荐问题并对此展开相关研究。 （2）实验验证法：



5.2 技术路线

（1）完善好相关数学公式，设计并编写算法的伪代码

（2）使用 tensorflow 或者 pytorch 框架实现算法

（3）设定初始参数后在服务器上运行，并根据实验结果调整实验初始参数，最后按照一定的规则保存训练过程的模型

（4）使用 matplotlib 等工具可视化实验结果以及各个模型的性能比较

参考文献

[1] 高占林.浅谈信息过载的影响及消除[J].天水行政学院学报,2010,(06):63-65

[2] Mooney R J, Roy L. Content-based book recommending using learning for text categorization//Proceedings of the 5th ACM Conference on Digital libraries. San Antonio, USA,2000:195-204

[3] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering/Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Madison, USA,1998:43-52

[4] Balabanovic M, Shoham Y. Fab: content-based, collaborative recommendation. Communications of the ACM,1997,40(3):66-72

[5] Mcnee S M, Riedl J, Konstan J A. Being accurate is not enough: how accuracy metrics have hurt recommender systems[C]. CHI 06 Extended Abstracts on Human Factors in Computing Systems. ACM,2006:1097-1101

[6] Castells P. Rank and relevance in novelty and diversity metrics for recommender systems[C]. ACM Conference on Recommender Systems. ACM,2011:109-116

[7] He X, Liao L, Zhang H, et al. Neural Collaborative Filtering[J].2017

[8] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning forrecommender systems//Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. Boston, USA,2016:7-10

[9] Shan Y. Hoens T R, Jiao J, et al. Deep Crossing: Web-scale modeling without manually crafted combinatorial features/Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA,2016:255-262

[10] Wang R, Fu B, Fu G et al. Deep & cross network for ad click predictions. arXiv preprint arXiv:1708.05123,2017

[11] Zhu J, Shan Y, Mao J C, et al. Deep embedding forest: forest-based serving with deep embedding features. arXiv preprint arXiv:1703.05291,2017

[12] Zhou G. Song C, Zhu X, et al. Deep interest network for click-through rate prediction. arXiv preprint arXiv:1706.06978,2017

[13] Gong Y, Zhang Q. Hashtag recommendation using attention-based

convolutional neural network//Proceedings of the 25th Intemational Joint Conference on Artificial Intelligence. New York, USA, 2016:2782-2788
[14] Lei C, Liu D, Li W, et al. Comparative deep learning of hybrid representations for image recommendations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA,2016:2545-2553
[15] Li Y, Liu T, Jiang J, et al. Hashtag recommendation with topical attention-based LSTM/Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan,2016:943-952
[16]Okura S, Tagami Y, Ono S, et al. Embedding-based News Recommendation for Millions of Users//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, Canada,2017:1933-1942
[17] Wang X. Wang Y. Improving content-based and hybrid music recommendation using deep learning/Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, USA,2014:627-636
[18] 王科文. 基于深度神经网络的中医药材推荐[D].华南理工大学,2018:1-52
[19] J. Ma and Z. Wang, “Discovering syndrome regularities in traditional chinese medicine clinical by topic model,” in 3PGCIC, 2016, pp. 157 - 162
[20]W. Ji, Y . Zhang, X. Wang, and Y . Zhou, “Latent semantic diagnosis in traditional chinese medicine,” WWW, vol. 20, no. 5, pp. 1071 - 1087, 2017
[21] S. Wang, E. W. Huang, R. Zhang, X. Zhang, B. Liu, X. Zhou, and C. Zhai, “A conditional probabilistic model for joint analysis of symptoms, diseases, and herbs in traditional chinese medicine patient records,” in BIBM, 2016, pp. 411 - 418
[22]X. Chen, C. Ruan, Y . Zhang, and H. Chen, “Heterogeneous informa-tion network based clustering for categorizations of traditional chinese medicine formula,” in BIBM, 2018, pp. 839 - 846
[23] X. Wang, Y . Zhang, X. Wang, and J. Chen, “A knowledge graph en-hanced topic modeling approach for herb recommendation,” in DASF AA, 2019, pp. 709 - 724
[24] C. Ruan, J. Ma, Y . Wang, Y . Zhang, and Y . Yang, “Discovering regularities from traditional chinese medicine prescriptions via bipartite embedding model,” in IJCAI, 2019, pp. 3346 - 3352.
[25]W. Li and Z. Yang, “Distributed representation for traditional chi-nese medicine herb via deep learning models,” arXiv preprint arXiv:1711.01701, 2017
[26]M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural net-works,” IEEE Trans. Signal Processing, vol. 45, no. 11, pp. 2673 - 2681, 1997
[27]C. Ruan, Y . Wang, Y . Zhang, and Y . Yang, “Exploring regularity in traditional chinese medicine clinical data using heterogeneous weighted networks embedding,” in DASF AA, 2019, pp. 310 - 313

论文写作提纲	（除题目外，具体到三级标题）： 1.绪论 1.1 选题背景与意义 1.2 研究现状 1.2.1 现有研究方法 1.2.2 存在的问题 1.3 主要研究内容 1.4 论文组织结构 2.SMGCN 模型 2.1 模型的核心思想 2.2 模型的原理 2.3 模型的优缺点分析 3.对照模型 3.1 主题模型 3.1.1 HC-KGETM 3.2 图神经网络模型 3.2.1 GC-MC 3.2.2 PinSage 3.2.3 NGCF 3.2.4 HeteGCN 4.实验准备 4.1 数据集 4.1.1 数据集的获取 4.1.2 数据集的整理分析 4.2 实验指标 4.3 实验平台 5.实验过程 5.1 消融实验 5.1.1 实验步骤 5.1.2 实验结果 5.1.3 结果分析 5.2 对比实验 5.2.1 实验步骤 5.2.2 实验结果 5.2.3 结果分析 5.3 超参数影响实验 5.3.1 实验步骤 5.3.2 实验结果 5.3.3 结果分析 6.总结与展望 6.1 本文的工作总结

	6.2 展望 7.参考文献
工作步骤与时间安排	2020.12.01-2021.01.08 查阅文献，明确研究内容，拟定试验方案，完成开题报告初稿 2021.01.07-2021.01.13 整理开题内容，进行开题答辩 2021.01.14-2021.03.13 针对老师提的问题，完善毕设内容和方案，修改开题报告，提交开题报告终稿 2021.02.01-2021.02.29 学习需要的理论知识和查阅文献，并进行环境搭建 2021.03.01-2021.03.31 毕设完成阶段，算法设计和实现过程、论文写作开始 2021.04.01-2021.04.30 优化算法，并完成论文初稿 2021.05.01-2021.05.10 检验修正算法实现，提交毕业论文初稿并查重 2021.5.11-2021.5.19 进一步修改完善论文，准备答辩 2021.5.20-2021.5.30 准备答辩 PPT，进行毕业论文答辩
开题答辩评语	（从选题、理论与实证准备、研究（设计）方法、工作安排等方面给出评价，并提出指导意见） 指导教师签名： 年　月　日

注：此表由学生填写后交指导教师签署意见，并交院系教务办保存，否则不得开题；此表将作为毕业设计最终评分的依据。