基于有向异构图的发票明细税收分类方法

赵珮瑶[1,2]，郑庆华[1,2]，董博[3,4]，阮建飞[1,2]，罗敏楠[1,2]
[1] 西安交通大学计算机科学与技术学院
[2] 陕西省天地网技术重点实验室
[3] 大数据算法与分析技术国家工程实验室
[4] 西安交通大学继续教育学院

摘要

税收是国家赖以生存的物质基础。为加快税收现代化，方便纳税人便捷、规范开具增值税发票，国税总局规定纳税人在税控系统开票前选择发票明细对应的税收分类才可正常开具发票。提高税收分类的准确度，是构建税收风险指标和分析纳税人行为特征的重要基础。基于此，本文提出了一种基于有向异构图的税收分类方法 (Heterogeneous Directed Graph Attenton Network,HDGAT)，利用发票明细间的有向信息建模，引入外部信息，显著地提高了发票明细的税收分类准确度。

关键词： 税收分类 ； 图卷积网络 ； 有向异构图

Tax Classification of Invoice Details Based on Directed Heterogeneous Graph

Peiyao Zhao[1,2]，Qinghua Zheng[1,2]，Bo Dong[3,4]，Jianfei Ruan[1,2]，Minnan Luo[1,2]
[1] School of Computer Science and Technology, Xi'an Jiaotong University
[2]SPKLSTN Lab, Xi'an Jiaotong University
[3]National Engineering Lab for Big Data Analytics, Xi'an Jiaotong University
[4]School of Continuing Education, Xi'an Jiaotong University

Abstract

Taxation is the material basis for the survival of the country. In order to accelerate tax modernization and facilitate taxpayers to issue value-added tax invoices in a convenient and standardized manner, the State Administration of Taxation requires taxpayers to select the tax classification corresponding to the invoice details before issuing invoices in the tax control system Thus, improving the accuracy of tax classification has a crucial role in the construction of tax risk indicators and analysis of taxpayer behavior characteristics. This paper proposes a classification model based on the Heterogeneous Directed Graph Attention Network (HDGAT). The model significantly improves the accuracy of the tax classification of invoice details by using the directed information among the invoice details and external information.

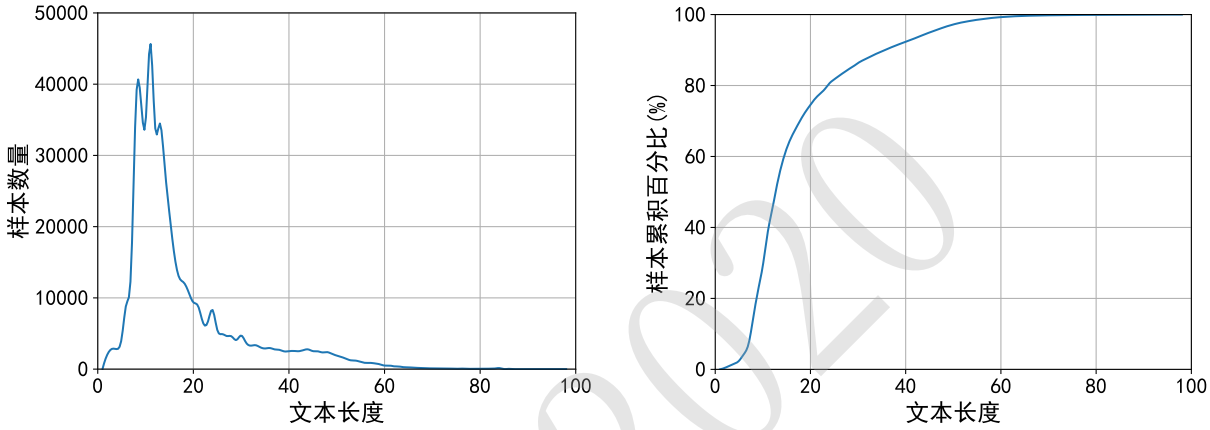
Keywords: Tax classification , Graph convolutional network , Directed heterogeneous graph

1 引言

第十九届中国计算语言学大会论文集，第771页-第782页，海南，中国，2020年10月30日至11月1日。

为加快税收现代化建设，方便纳税人便捷、规范开具增值税发票，国税总局自 2016 年 5 月 1 日起在全国范围内推行《商品和服务税收分类与编码》，纳税人在使用税控系统开票前，需要选择商品和服务（即发票明细）对应“税收分类编码”才能正常开具发票 (国家税务总局, 2016)。对于经济活动中不存在生产和加工环节的纳税人，购进和销售的商品原则上要保持一致，即进项货物和销项货物品类一致。通过对比商品与服务税收分类编码即可判断该类企业是否存在虚开或者开票不规范的行为。商品和服务的税收分类编码，作为各种税收风险指标的数据基础，其重要性不言而喻。提高商品和服务的税收分类精确度，是构建税收风险指标和分析纳税人行为特征的重要基础 (殷明霞, 2019)(孙懿, 2015)。

发票明细标注了对应的纳税人之间交易的商品和服务名称，选取中国某省纳税人数据进行分析，根据图（1）所示，发票明细平均文本长度为 17.62，在单词数量为 60 时，样本累计百分比达到了 98.76%，选取不同样本的统计结果有细微差别，但符合短文本的定义。因此，对针对发票明细的税收分类问题可转化为短文本的多分类问题。



(a) 样本数量统计　　(b) 样本累积百分比统计

图 1: 发票明细文本长度统计

因税务场景中发票上下游特有的方向性特征，现有的短文本分类方法并不适用于处理发票明细的税收分类问题。基于上述背景，本文提出了一种基于异构有向图的短文本分类模型 (Heterogeneous Directed Graph Attenton Network,HDGAT) 来进行发票明细的税收分类，有效地利用了发票明细间的有向信息。

2 相关工作

文本分类是自然语言处理中的经典应用，由于短文本的稀疏性和有限的标记数据，传统的文本分类方法不适用于短文本分类，目前主流方法主要包括基于主题模型的分类方法、基于深度学习的分类方法，根据数据集的不同，也可划分监督学习和半监督学习方法 (邓丁朋 et al., 2020)。

主题模型指通过推理策略获取短文本的主题特征，并将其与文档的原始特征进行融合，从而实现较好的分类效果。潜在语义分析模型（Latent Semantic Analysis，LSA）(Wiemer-Hastings et al., 2004) 通过奇异值分解将文档映射到低维语义空间里的进行向量表示，从概率生成的角度实现对文本的表示；概率潜在语义分析模型（probabilistic LSA，pLSA）(Hofmann, 1999) 从概率角度对 LSA 模型进行扩展，构建了一个以主题为隐变量的贝叶斯概率模型；隐含狄利克雷模型（Latent Dirichlet Allocation，LDA）(Blei et al., 2003) 是在 pLSA 的基础上扩展的一个文档、主题和词的三层贝叶斯概率模型，一定程度上克服了随着文档集增长 pLSA 参数过多而造成的过拟合问题。

随着神经网络的发展，RNN(Yin et al., 2017) 和 CNN(Lawrence et al., 1997) 等两个代表性的深度神经模型也开始在 NLP 任务中发挥了作用。TextCNN(Kim, 2014) 是将 CNN 用于句子分类的最初尝试，通过 word2vec 将单词转化为向量；循环神经网络 (Recurrent Neural Networks,RNN)(Yin et al., 2017) 是一类以序列数据为输入，在序列的演进方向进行递归且所有

节点按链式连接的递归神经网络；作为对 RNN 模型的改进，LSTM 网络模型 (Hochreiter and Schmidhuber, 1997) 被提出，通过长短期记忆单元来解决 RNN 梯度消失和指数爆炸问题。

针对短文本的特征词少，信息关联性不强以及存在大量样本的标注瓶颈问题，基于半监督学习的短文本分类问题越来越受到关注。预测文本嵌入 (Predictive Text Embedding,PTE) 模型 (Tang et al., 2015) 是一种半监督模型，该模型将有标签数据和无标签数据共同建模，然后将该网络降维到一个低维度的向量空间，得到文本的特征表示；分层注意力网络 (Hierarchical Attention Networks,HAN) 模型 (Yang et al., 2016) 是一种两层级的注意力模型。将文本分为单词-句子两层结构，分别建模后形成文本的向量表示。两层注意力模型分别应用于单词级别和句子级别，使得模型对于不同的单词和句子给予不同的权重，从而让文本分类更加精确；TextGCN 模型 (Yao et al., 2019) 将词和文本同时作为节点构建异构图，从而对文本进行表示；HGAT(Linmei et al., 2019) 模型通过在异构图上设立两层注意力机制，捕获不同节点的重要性。

目前金税三期的增值税发票管理系统中，会根据纳税人填写的发票明细，推荐其所属商品和服务税收分类编码，具体推荐规则为：前期通过各领域专家人工校定，得到大量有标签的样本，要推荐的发票明细通过与标签样本进行语义相似度计算，选取语义相似度最高的税收分类进行推荐。现有匹配规则忽略了发票间的关联关系以及文本间的共现关系，从而影响分类的质量。而上述短文本分类方法亦不能完全适用于税务领域，根据纳税人交易生成的发票信息可知，同一纳税人的进项发票和销项发票存在方向信息，而现有基于图的短文本分类方法是在无向图基础上建模，损失了发票的上下游信息。

本文提出了一种基于有向异构图的税收分类方法 HDGAT，通过融合标记数据和未标记数据间的关联关系，结合发票间的有向信息建模，引入标签概念和词语概念作为外部知识补充，设置双层注意力机制，通过多种粒度捕获关键信息，减少嘈杂信息的权重，显著地提高了发票明细的税收分类准确度。

3 基于有向异构图的税收分类

3.1 短文本异构图

为了充分利用发票间的上下游关系以及发票明细文本间的共现信息，本文将发票明细、标签、词语构建为有向异构图，如图（2）所示。

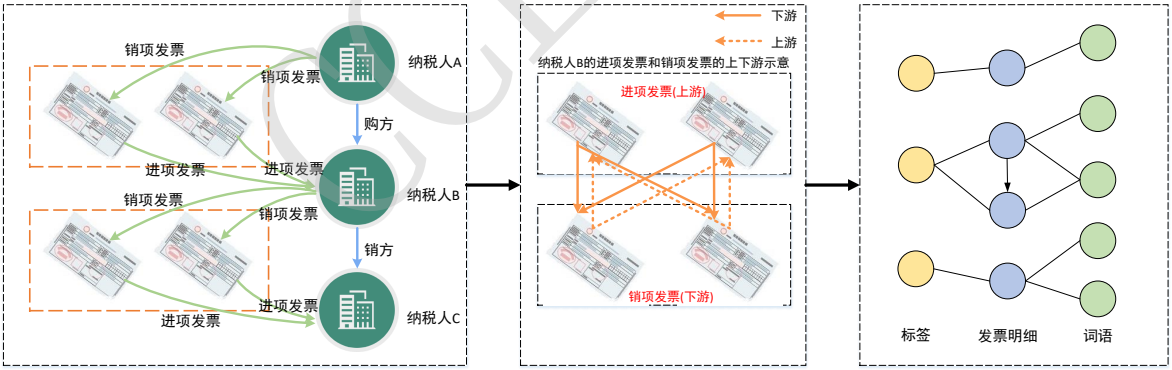


图 2: 税务异构图构建

发票间的上下游关系产生于纳税人间有向的交易关系，在图2中，纳税人 B 既可以作为与纳税人 A 交易中的购方纳税人，也可以作为与纳税人 C 交易中的销方纳税人，其交易产生的进项发票和销项发票间就存在上下游关系，可由此分析发票明细间的语义信息传递。

共现信息，指不同文本中的关联信息和特征项隐含的知识 (于游 et al., 2019)。例如，“七匹狼”这一发票明细表示的语义可以指香烟，也可以指衣物，在具体的分类任务中，如果没有其它信息的引入，很难得到正确的分类结果。若当前“七匹狼”的上下游发票明细中，含有衣物或是香烟的关联信息，则可对该文本的分类起到辅助作用。

除此外，本文考虑引入外部知识，通过处理《商品和服务税收分类编码》，提取具体类别的相关概念作为标签的外源信息；同时引用 Wikipedia 语料集，将词语对应的概念作为词语的外

源信息。

令 $\mathcal{G}=(\mathcal{V},\mathcal{E})$ 表示短文本异构图，节点 $\mathcal{V}=D\cup T\cup E$，包含短文本 $T=\{t_1,\ldots,t_m\}$，标签 $L=\{l_1,\ldots,l_k\}$ 和词语 $W=\{w_1,\ldots,w_n\}$ 。边集 $\mathcal{E}$ 代表三类节点间的关系，包括标签与文本间的关系 $Label-Text$、文本与文本间的关系 $Text-Text$、文本与词语间的关系 $Text-Word$。

短文本异构图细节描述如下。短文本间的关系通过纳税人的交易关系获得，并根据发票的方向信息得到短文本的上下游关系，因此，短文本间的边为有向边。不同的短文本可能带有税收分类标签，也可能没有税收分类标签。在 HDGAT 模型构建中，以基于谱域的方式构造有向图的拉普拉斯矩阵，利用发票明细间的上下游关系及三类实体间的关联关系对所有实体进行网络表示，包括有标签文本和无标签文本，最后对有标签的数据进行分类验证。

词语是从短文本数据预处理过程中得到的，通过分词工具，对短文本进行分词、去停用词，统计所有词语出现的次数，设定提取阈值 $\lambda=5$，提取出现次数大于 5 的词语加入异构图中。如果短文本包含某个词语，则将它们间进行连接，建立有向边。

标签来自短文本的标记信息，与短文本间的边为有向边，由标签指向短文本。

3.2 有向异构图卷积

图卷积网络 (Graph Convolutional Networks,GCN) 是一个多层神经网络，根据节点的邻域属性来推导其嵌入向量，最初应用于同构图 (Kipf and Welling, 2016)。根据更新方式可将 GCN 分为基于空间域的 GCN 和基于谱域的 GCN(Zhou et al., 2018), 本文构建的模型是以谱域的角度进行建模，模型整体架构如图（3）所示。

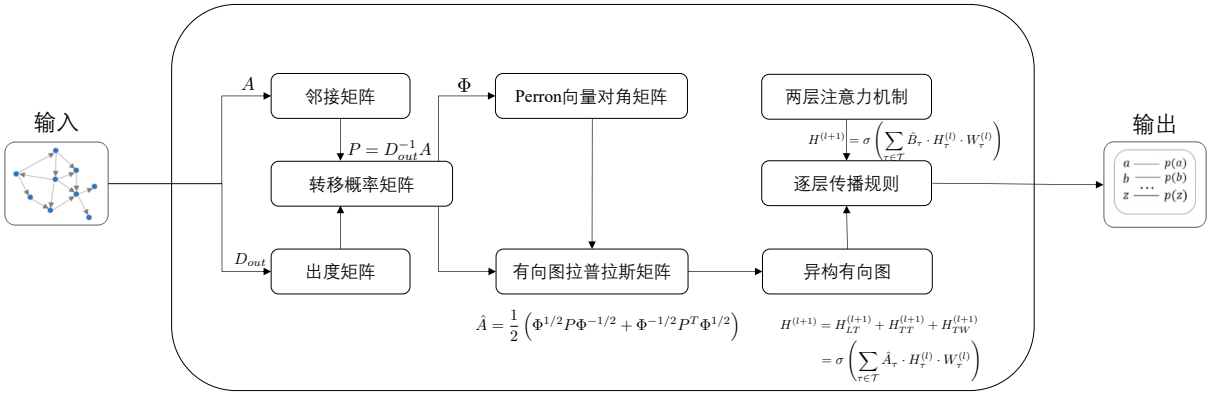


图 3: HDGAT 模型示意

定义有向图 $\mathcal{G}=(\mathcal{V},\mathcal{E})$，其中 $\mathcal{V}$ 和 $\mathcal{E}$ 分别代表节点和边的集合。如果图 G 的每对顶点之间在每个方向上都有一条路径，则此有向图 G 称为强连通。假设节点 u 指向节点 v，则这条边表示为 (u,v)，节点 u 和节点 v 互为一阶邻居。节点的出度指从该节点出发的边的条数，入度指进入该节点的边的条数。定义有向图的出度矩阵为 D_{out}，邻接矩阵为 A，则转移概率矩阵为表示为公式所示

$$P=D_{out}^{-1}A \tag{1}$$

根据 Perron-Frobenius 定理 (De Lathauwer et al., 2000)，具有非负项的不可约矩阵具有唯一的左本征向量，所有项均为正。将这一定理应用到有向图中，令 ρ 表示一个强连通有向图的转移概率矩阵 P 所有正特征向量的特征值，则 P 具有唯一的左本征向量 ϕ，其中 $\phi(v)>0$，满足等式 $\phi P=\rho\phi$，ϕ 表示行向量。根据 Perron-Frobenius 定理，令 $\rho=1$ ，而 P 的所有其它特征值的绝对值不大于 1。

定义有向图的拉普拉斯矩阵为 L_{dir}，具体表示为

$$L_{\text{dir}}=I-D^{-1/2}AD^{-1/2}=I-D^{1/2}PD^{-1/2}=I-\Phi^{1/2}P\Phi^{-1/2} \tag{2}$$

其中，$\Phi=\text{diag}\left(\phi_{\text{norm}}(v)\right)$ 为 P 的 Perron 向量对角矩阵，因上述公式（2）中, 有向图转移概率矩阵 P 为非对称矩阵，进一步修改为以下公式保证拉普拉斯算子的对称性 (Horn and

Johnson, 2012)。

$$L_{dir} = I - \frac{1}{2}\left(\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^T \Phi^{1/2}\right) \tag{3}$$

定义有向图的分层传播规则如下

$$H^{(l+1)} = \sigma\left(\hat{A} \cdot H^{(l)} \cdot W^{(l)}\right) \tag{4}$$

其中，$\hat{A} = \frac{1}{2}\left(\tilde{\Phi}^{1/2}\tilde{P}\tilde{\Phi}^{-1/2} + \tilde{\Phi}^{-1/2}\tilde{P}^T\tilde{\Phi}^{1/2}\right)$ 表示有向图的拉普拉斯矩阵，$H^{(l)} \in \mathbb{R}^{|\mathcal{V}|\times q}$ 表示 l^{th} 层中节点的隐藏表示，第一层设置 $H^{(0)} = X$，$W^{(l)}$ 表示使用梯度下降训练神经网络权重矩阵。$\sigma(\cdot)$ 表示激活功能，例如 ReLU。

在短文本异构图中，有三类节点：短文本、词语、标签。将边类型集合表示为 $\mathcal{T} = \{\tau_1, \tau_2, \tau_3\}$，具体来说，$\tau_1 = Label - Text$，$\tau_2 = Text - Text$，$\tau_3 = Text - Word$，则对于短文本异构图的传播函数定义为各部分的加和，将它们各自的权重矩阵投影到一个隐式的公共空间中，如公式 (5) 所示:

$$H_{LT}^{(l+1)} = \sigma\left(\hat{A}_{LT} \cdot H_{LT}^{(l)} \cdot W_{LT}^{(l)}\right) \tag{5a}$$

$$H_{TT}^{(l+1)} = \sigma\left(\hat{A}_{TT} \cdot H_{TT}^{(l)} \cdot W_{TT}^{(l)}\right) \tag{5b}$$

$$H_{TW}^{(l+1)} = \sigma\left(\hat{A}_{TW} \cdot H_{TW}^{(l)} \cdot W_{TW}^{(l)}\right) \tag{5c}$$

$$H^{(l+1)} = H_{LT}^{(l+1)} + H_{TT}^{(l+1)} + H_{TW}^{(l+1)} = \sigma\left(\sum_{\tau\in\mathcal{T}} \hat{A}_\tau \cdot H_\tau^{(l)} \cdot W_\tau^{(l)}\right) \tag{5d}$$

其中 $\hat{A}_\tau \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}_\tau|}$ 是 $\hat{A}$ 的子矩阵，$\hat{A}$ 的行代表所有节点，列代表类型为 τ 的相邻节点，$H^{(l+1)}$ 的表示第 $l+1$ 层的传播函数，是通过聚集三部分 $H_\tau^{(l)}$ 的特征的信息而获得的，第一层设置 $H_\tau^{(0)} = X_\tau$。权重矩阵 $W_\tau^{(l)} \in \mathbb{R}^{q(l)} \times q^{(l+1)}.$，需考虑不同特征空间的差异。

3.3 注意力机制

通常，对于指定节点，不同类型的邻居节点可能会对它产生不同的影响。相同类型的邻居节点可能会携带更多有用的信息，并且，相同类型中的不同邻居节点也可能具有不同的重要性。例如，在短文本异构图中，短文本节点、标签节点和词语节点可能会对指定短文本节点具有不同的影响，而与该节点相连的其它短文本节点也可能具有不同的重要性。为了同时捕获节点级别和类型级别的不同重要性，本文引入双层注意力机制 (Velikovi et al., 2017)。

1）类型级别的注意力机制

给定特定节点 v，类型级别的注意将学习不同类型的相邻节点对 v 的影响权重。具体来说，首先将类型 τ 的嵌入表示为 $h_\tau = \sum_{v'} \hat{A}_{vv'} h_{v'}$，$h_\tau$ 是相邻节点特征 $h_{v'}$ 的总和，其中节点 $v' \in \mathcal{N}_v$ 并且类型为 τ。然后，基于当前节点 h_v 和嵌入表示 h_τ 来计算类型级别的注意力得分，如公式（6）所示。

$$\hat{a}_\tau = \sigma\left(\mu_\tau^T \cdot [h_v \| h_\tau]\right) \tag{6}$$

其中 μ_τ 是类型 τ 的注意力向量，$\|$ 表示“连接”，$\sigma(\cdot)$ 表示激活函数，例如 ReLU。

使用 softmax 函数对所有类型的注意力得分进行归一化，从而获得类型级别的注意力权重，如公式（7）所示。

$$\hat{\alpha}_\tau = \frac{\exp\left(\hat{a}_\tau\right)}{\sum_{\tau'\in\mathcal{T}} \exp\left(\hat{a}_{\tau'}\right)} \tag{7}$$

2）节点级别的注意力机制

对于特定节点 v, 节点级别的注意力机制将捕获不同相邻节点的重要性并减少噪声节点的权重。形式上，给定类型为 τ 的特定节点 v 及为不同类型 τ' 的相邻节点 $v' \in \mathcal{N}_v$，基于类型级别

的嵌入 τ'，计算节点 v 和节点 v' 基于节点级别的注意力得分，注意力权重表示为 $\alpha_{\tau'}$ ，如公式（8）所示。

$$\hat{b}_{vv'} = \sigma\left(\nu^T \cdot \hat{\alpha}_{\tau'}\left[h_v \| h_{v'}\right]\right) \tag{8}$$

其中 ν 是注意力向量。使用 softmax 函数对节点级别的注意力得分进行归一化，如公式（9）所示。

$$\hat{\beta}_{vv'} = \frac{\exp\left(b_{vv'}\right)}{\sum_{i\in\mathcal{N}_v}\exp\left(b_{vi}\right)} \tag{9}$$

最后，通过替换等式，我们将包括类型级别和节点级别注意的双重注意机制集成到有向异构图卷积中。整体的分层传播规则如公式所示：

$$H^{(l+1)} = \sigma\left(\sum_{\tau\in\mathcal{T}}\hat{B}_\tau \cdot H_\tau^{(l)} \cdot W_\tau^{(l)}\right) \tag{10}$$

其中，$\hat{B}_\tau$ 表示注意力矩阵，$\beta_{vv'}$ 表示在 v^{th} 行 v'^{th} 列的元素。为了实现后续的分类任务，将短文本嵌入 $H^{(L)}$ 馈送到 softmax 层进行分类，在模型训练过程中，通过 L2 范数交叉熵损失训练数据，表示为：

$$\mathcal{L} = -\sum_{i\in D_{\text{train}}}\sum_{j=1}^{C} Y_{ij}\cdot\log Z_{ij} + \eta\|\Theta\|_2 \tag{11}$$

其中，C 表示分类数目，D_{train} 表示的短文本索引集合，用于训练，Y 表示标签矩阵，Θ 表示模型参数，η 表示正则化因子。训练过程采用梯度下降算法优化模型。

4 实验设计

针对本文提出的 HDGAT 模型，本节进行相应的对比实验，说明文本间的共现信息和发票间的有向信息对于发票明细分类的有效性，同时，引入的两层注意力机制能够进一步提高模型的分类效果。

4.1 数据集描述及预处理

本文实验评测的数据集采用从我国某省国税局获取的纳税人交易数据、国税总局推出的《商品和服务税收分类编码》以及中文维基百科语料库。国税总局 2018 年发布的《商品和服务税收分类编码》，包含 4206 种商品和服务分类，每种分类都有具体的概念说明，例如，货物“甲类卷烟”，属于“烟草制品”，对应的概念描述为“每标准条 (200 支) 调拨价格在 70 元 (不含增值税) 以上 (含 70 元) 的卷烟”。概念描述为判断货物的商品和服务类别提供了依据。

中文维基百科语料库是由原始的维基百科网页处理得到，提取网页中的纯文本信息进行词模型训练，得到大量词条的向量表示，用于对短文本异构图中词语节点进行外源信息的补充。

根据我国某省国税总局中的纳税人虚开名单，根据其交易关系，选取一阶邻居节点，包含 284 种商品和服务类别的共计 50000 条发票明细文本，涉及到的纳税人数量为 4987 个，经过数据预处理后得到 12733 个词语，作为构建短文本异构图的原始数据。数据集描述如表1所示。

表 1: 数据集基本描述

数据集	文本	标签比例	词语	标签	边
TAX50K	50000	14%	12733	284	941488

短文本异构图中有向边的度分布呈幂律分布，如图4所示，大部分的节点度极小，小部分的节点度极大，且入度和出度的幂律分布有细微差别。

通过计算可知，异构图的边密度为 1.6865×10^{-5}。根据 (Goswami et al., 2018) 的研究，边密度表示图实际具有的边与其潜在边的比率，并不能很好的展现加权图的稀疏性特征，因此

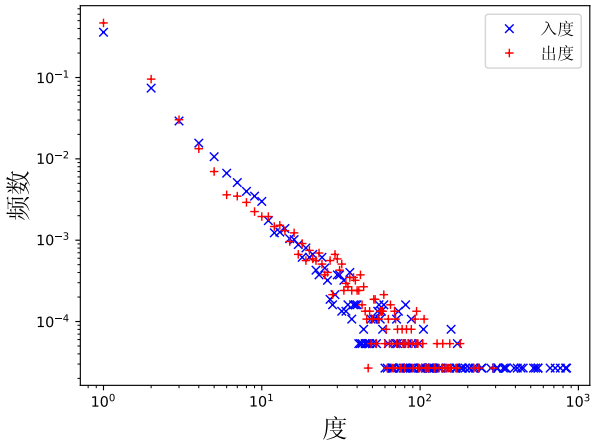


图 4: 税务异构图的度分布

引入 *Gini* 系数分析图的稀疏性。Gini 衡量一定数量的节点之间连接度的不平等程度，公式为 $GI = 1 - 2 \cdot \left[\sum_{i=1}^{n} \frac{b_i}{T}\left(\frac{n-i+\frac{1}{2}}{n}\right)\right]$，其中 T 表示所有节点度的和，b_i 表示节点的度按升序排列后第 i 个值。由表2可知，构造的税务异构图 Gini 系数为 0.6530，稀疏性较高，因此 HDGAT 通过学习局部结构的文本表示来构建发票明细的税收分类模型。

表 2: 数据集稀疏性描述

数据集	边密度	平均集聚系数	度平均值	Gini 系数
TAX50K	1.6865×10^{-5}	3.1×10^{-3}	1.7902	0.6530

表 3: 样本集数量表

总样本集 (10)	训练集 (6)	验证集 (2)	测试集 (2)
51000	31000	10000	10000

数据集按照 3:1:1 的比例划分为训练集、验证集、测试集，各样本数量如表3所示。数据集中的边共有 941488 条，其中，边类型为 $Text - Label$ 的有 50000 条，边类型为 $Text - Text$ 的有 723451 条，边类型为 $Text - Word$ 的有 168037 条。对词语补充概念文本、税收分类补充概念文本进行长度统计，统计结果如图5所示。由图可知，概念类文本长度的数值分布与发票明细的有明显区别，概念类文本长度分布不连贯，且比发票明细文本更长。

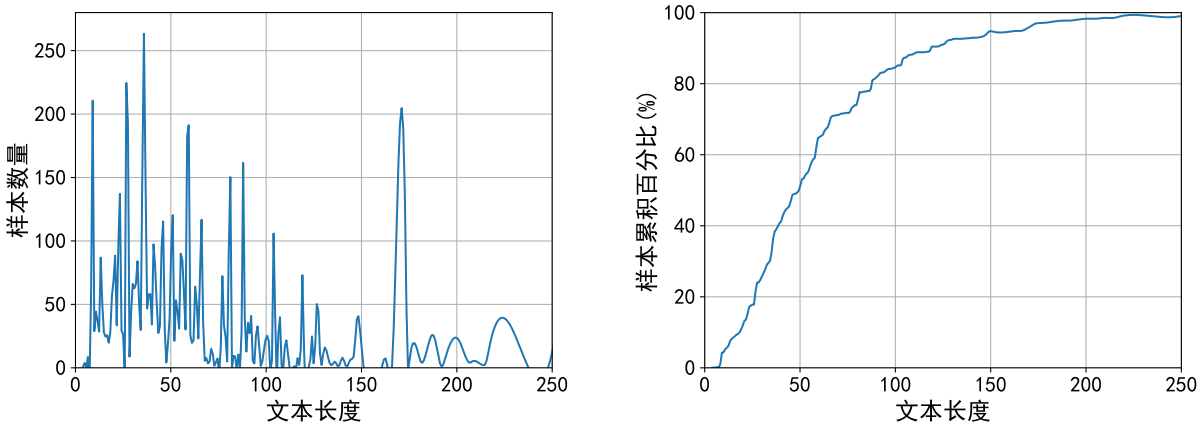
对于文本的数据预处理过程，主要包括发票明细短文本的预处理以及标签、词语的外源信息预处理两部分，描述如下。

预处理过程尝试采用 Word2Vec+TF-IDF 和 Bert 两种方法分别获取文本的向量表示，并分别进行实验对比。

1）Word2Vec + TF-IDF

Step1: 分词。根据《商品和服务税收分类编码》中对税收分类的说明构建交易明细专有词典，并根据货物名称人工添加词条作为补充，构建的交易明细专业词典包括“五金”、“浇铸”、“锡锭”、“绝地求生”、“晨光”等各行业交易明细专业性词汇共计 4467 个单词；基于 Jieba 分词工具，将交易明细专业词典设为自定义词典，作为对原有词典的补充，对发票明细短文本进行分词处理；

Step2: 去停用词。根据实际货物描述特点，将表示规格、体积的描述词添加到停用词词典中，例如：“52° 五粮液 1618 瓷瓶 500ml”中，添加“52°”、“500ml”至停用词词典，利用 Jieba



(a) 样本数量统计　　(b) 样本累积百分比统计

图 5: 概念类文本长度统计

工具对分词后的发票明细进行去停用词处理；

Step3: 获取词向量。基于 Python 的 gensim 库中的 Word2Vec 工具对纳税人的文本特征进行 Word Embedding 处理，采用 Skip-gram 模型，以词向量上下文最大距离为 5，词向量维度为 100 进行文本特征的向量化，得到词语的 Embedding 结果；

Step4: 计算 TF-IDF。TF 表示词频，词频 = 某个词出现次数/总词数；IDF 表示逆文档频率，逆文档频率 =log(语料库的文档总数/(包含该词的文档数 +1))，由 TF-IDF=TF×IDF 计算得到不同词的权重；

Step5: 获取句向量。根据词向量和词权重进行加权平均，作为该条发票明细整体的向量表示。

2)Bert

Bert 是谷歌于 2018 年提出的基于 Transformer 的双向编码器的端到端表示模型 (Devlin et al., 2018)。旨在通过联合调节所有层中的上下文来预先训练深度双向表示，并通过后期微调的训练策略提高对不同文本的向量表示能力。本文选用 Bert 旨在对比其表示效果，对细节原理不做详述。

Step1: 获取中文预训练模型，根据模型所在路径设置加载地址；

Step2: 使用 Python 的第三方库 bert-as-service，分别完成客户端 bert-serving-client 和服务端 bert-serving-server 的配置；

Step3: 服务端开启服务后，客服端根据服务端 ip 进行连接；

Step4: 依次遍历发票明细、词语补充概念、税收分类概念，根据 BertClint.encode() 函数获取对应的 768 维句向量。

4.2 评价指标和参数设置

实验任务为实体分类，选取精准率 Accuracy、F1 值作为评价指标。Accuracy 的计算方法如公式12所示，F1 的计算方法如公式13所示。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{13}$$

其中，TP 表示正例被判定为正例，FP 表示负例被判定为正例，FN 表示正例被判定为负例，TN 表示负例被判定为负例。

针对 HDGAT 模型，我们探索了不同参数对实验结果的影响。实验基于随机梯度下降进行模型训练，在训练中，使用固定学习率 λ，在 $\lambda = 0.001, 0.002, 0.005$ 的范围内进行尝试，训练时的 dropout 在 $0.5, 0.6, 0.7, 0.8, 0.9$ 中进行尝试。最终选择如下最优参数：$\lambda = 0.005$，$dropout = 0.8$，

网络的隐藏层维度设置为 512，根据 Word2Vec+TF-IDF 和 Bert 训练得到向量维度的不同，分别将输入维度设置为 100 或 768。正则化因子设置为 $\eta = 5e-6$，避免模型过拟合。

4.3 实验结果与分析

1) 有效性分析

本节将 HDGAT 与 LSTM、PTE、TextGCN、HAN 四类基准算法在税务数据集上进行实验对比，用来评估 HDGAT 用于半监督短文本分类的效果。其中，LSTM 是长短期记忆网络(Hochreiter and Schmidhuber, 1997)，将发票明细的向量表示直接输入到模型中，进行文本分类任务；PTE(Tang et al., 2015) 是最早的异构网络的表示学习方法，可用于文本数据的半监督分类任务；TextGCN(Yao et al., 2019) 是将最早将异构 GCN 模型应用于文本分类任务的，通过将文本和词语作为节点，学习二者之间的关联，进而输出向量表示进行半监督的文本分类任务；HAN(Yang et al., 2016) 通过元路径，将文本异构图转化为几个同构子网络的加和，然后在应用图注意力机制进行文本分类任务。HGAT(Linmei et al., 2019) 将注意力机制引入异构图中进行文本分类的任务。我们用 W 表示模型输入的向量是通过 Word2Vec+TF-IDF 生成的，用 B 表示模型输入的向量是通过 Bert 生成的。“-directed”表示在 HDGAT 模型中去除有向信息，观察实验结果；“-attention”表示在 HDGAT 模型中去除双层注意力机制，观察实验结果。实验结果如表4所示。

表 4: 实验结果对比

评价指标	Accuracy(%)	F1(%)
LSTM	43.67	42.19
PTE	45.32	42.53
TextGCN	72.61	60.98
HAN	65.64	59.77
HGAT	79.49	74.12
HDGAT(W)	84.45	77.67
HDGAT(B)	**89.32**	**81.55**
HDGAT(W-directed)	79.33	72.34
HDGAT(B-directed)	80.98	73.76
HDGAT(W-attention)	73.67	61.97
HDGAT(B-attention)	75.56	62.69

由表4可知，HDGAT 模型的分类效果比基准方法有明显提升，其中 HDGAT 基于 Bert 获得向量表示比基于 Word2Vec+TF-IDF 获得向量表示对后续模型训练分类的效果更好，说明优质的词向量模型对于后续模型训练的重要性，词向量模型的训练语料越丰富，得到词语的向量表示越准确。对于基准方法的实验效果，PTE 模型的分类结果较差，原因可能是，PTE 仅依靠词语间的共现信息来学习文本嵌入的，而短文本的共现信息相对于长文本来说较少，使得 PTE 分类效果不理想。基于图神经网络的模型 HGAT、TextGCN 和 HAN 模型较 CNN、LSTM、PTE 分类效果理想。

针对 HDGAT 的消去实验的分类结果，去除有向信息的分类结果与 HGAT 分类效果基本一致，优于其它基准方法，说明引入发票明细的上下游信息捕获了短文本间的语义信息传递，同时也说明注意力机制捕获了不同相邻节点和不同类型节点的重要性（减少了嘈杂信息的权重）；去除注意力机制的分类结果仍优于基准方法，说明将发票明细间的上下游的方向信息输入到模型的有效性。

HDGAT 模型是基于谱域的图卷积模型，模型训练过程中需计算全部图的邻接矩阵，不适合超大规模图的计算，相对于基于空间域的图卷积模型，不会随机选取邻居节点进行建模，可解释性更强。

2）可解释性分析

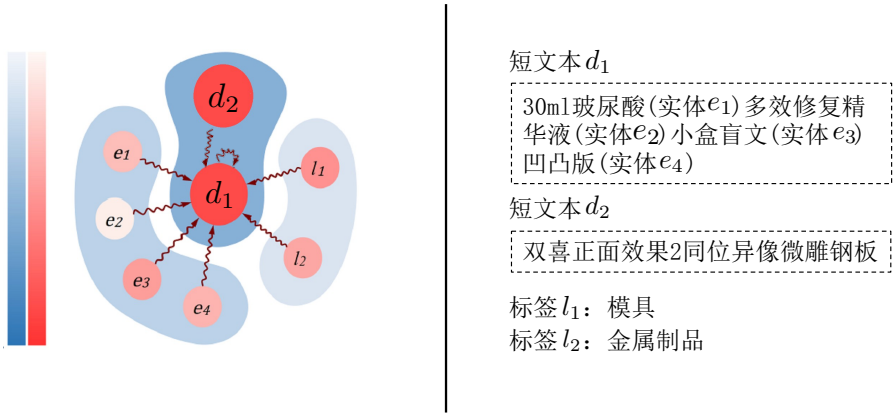


图 6: HDGAT 可视化分析

以发票明细“30ml 玻尿酸多效修护精华液小盒盲文凹凸版”为例，其正确分类为“印刷专用设备”。该发票明细通过分词、去停用词、超过阈值条件被选定关联的词语依次为“玻尿酸”、“精华液”、“盲文”、“凹凸版”。通过下游发票明细“双喜正面效果 2 同位异像微雕钢板”关联的标签为“模具”，“陶瓷玻化砖钻头”关联的标签为“金属制品”等等，构成的局部短文本异构图如图（6）所示。

通过分析纳税人间的交易信息可知，发票明细间的上下游关系可以分为通用类货物销向关系和加工类货物销向关系，通用类货物指电脑、中性笔等不局限于纳税人的经营范围而购入的货物，加工类货物指根据纳税人的经营范围购入经过加工或散装销出的货物。根据短文本 d_1 关联到的上游节点 d_2 可知，二者不属于通用类货物。d_2 作为上游节点，可能与 d_1 存在加工制作的关联性。

类型级别的注意权将高权重（0.75）分配给短文本本身，而将低权重（0.2 和 0.05）分配给实体和主题，这意味着短文本本身的语义比实体和主题对分类的贡献更大，即上游发票明细“双喜正面效果 2 同位异像微雕钢板”对于分类的影响大于其它类型节点，d_2 自身的标签为“模具”，关联为 d_1 的二阶邻居。节点级别的注意力为短文本关联的节点分配了不同的权重，属于同一类型的节点的节点级权重之和为 1，如图所示，实体 e_3（盲文）、e_4（凹凸版）的权重比 e_1（玻尿酸）、e_2（精华液）的权重更高。经过 HDGAT 模型中 softmax 层选出可能性最大的两个类别依次为“印刷专用设备”、“美容护肤品”，选定可能性最大的“印刷专用设备”作为分类结果。该发票明细的二阶邻居节点，标签 l_1（模具）和 l_2（金属制品）对于将短文本分类为“印刷专用设备”具有近似相等的影响。该案例表明，模型中引入发票明细间的有向信息及双层注意力机制可以以多种粒度来捕获关键信息，并减少嘈杂信息的权重，从而影响模型的分类结果。

5 总结

本文提出了一种基于有向异构图的税收分类方法 HDGAT。该方法通过信息传播有效利用了标记数据和未标记数据间的关联关系，构建短文本异构图，并集成标签概念和词语概念作为外源信息补充，利用了发票明细间的方向信息及双层注意力机制，有效提高了模型的分类效果。本文的实验部分利用了中国某省税务数据，证明了以下结论：

1）利用发票明细间的上下游信息，并引入双层注意力机制，通过多种粒度捕获关键信息，减少嘈杂信息的权重，有效提升了分类效果；2）通过对比基于 Word2Vec+TF-IDF 与基于 Bert 获得向量表示对后续模型训练的分类效果，说明优质的词向量模型对于后续模型训练的重要性，词向量模型的训练语料越丰富，得到词语的向量表示越准确。3）HDGAT 是一种基于谱域的图卷积模型，训练过程中需用到全部图的邻接矩阵进行计算，不适用于超大规模图计算。

该模型是根据税务场景进行构建，若其它场景中也存在短文本间具有有向信息传递的现象，可尝试进行迁移，不适用于一般类型的短文本分类任务。未来可结合空间域的图卷积模型进行改进，并从采样方式中寻求可解释性，以适应于超大规模图的计算。

致谢

本研究得到如下项目资助：国家重点研发计划“云计算与大数据”重点专项课题（2016YFB1000903），教育部创新团队 (IRT_17R8)，国家自然科学基金（61721002、61532015），西安交通大学-税友集团人工智能联合实验室项目。

参考文献

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Swati Goswami, CA Murthy, and Asit K Das. 2018. Sparsity measure of a network graph: Gini index. *Information Sciences*, 462:16–39.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. *international acm sigir conference on research and development in information retrieval*, 51(2):50–57.

Roger A Horn and Charles R Johnson. 2012. *Matrix analysis*. Cambridge university press.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Steve Lawrence, C Lee Giles, Ah Chung Tsoi, and Andrew D Back. 1997. Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, 8(1):98–113.

Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4823–4832.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174.

Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks.

Peter Wiemer-Hastings, K Wiemer-Hastings, and A Graesser. 2004. Latent semantic analysis. In *Proceedings of the 16th international joint conference on Artificial intelligence*, pages 1–14. Citeseer.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Chuanlong Yin, Yuefei Zhu, Jinlong Fei, and Xinzheng He. 2017. A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access*, 5:21954–21961.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

于游, 付钰, and 吴晓平. 2019. 中文文本分类方法综述. 网络与信息安全学报, (5).

国家税务总局. 2016. 国家税务总局关于开展商品和服务税收分类与编码试点工作的通知. [EB/OL]. http://www.chinatax.gov.cn/n810341/n810755/c2417702/content.html Accessed Feb 25, 2016.

孙懿. 2015. 大数据时代对税务工作的挑战与对策. 学术交流, (2015 年 06):133–139.

殷明霞. 2019. 基于金税三期下的企业税务风险管理研究. 纳税, (33).

邓丁朋, 周亚建, 池俊辉, and 李佳乐. 2020. 短文本分类技术研究综述. 软件.