



(12) 发明专利

(10) 授权公告号 CN 116541594 B

(45) 授权公告日 2025. 06. 20

(21) 申请号 202310482963.9

G06F 18/24 (2023.01)

(22) 申请日 2023.04.28

G06F 18/25 (2023.01)

(65) 同一申请的已公布的文献号

G06N 3/0464 (2023.01)

申请公布号 CN 116541594 A

G06N 3/08 (2023.01)

(43) 申请公布日 2023.08.04

G06F 40/289 (2020.01)

G06F 40/284 (2020.01)

(73) 专利权人 吉林大学

(56) 对比文件

地址 130012 吉林省长春市前进大街2699号

CN 114020928 A, 2022.02.08

CN 114626369 A, 2022.06.14

(72) 发明人 李瑛 梁钰 张洁琳 杜伟 张晓
孙睿涵

审查员 袁冠群

(74) 专利代理机构 北京专赢专利代理有限公司
11797

专利代理师 李斌

(51) Int. Cl.

G06F 16/9535 (2019.01)

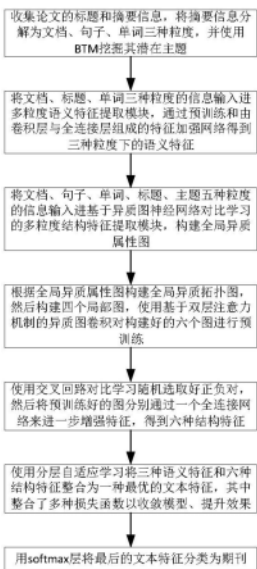
权利要求书3页 说明书9页 附图6页

(54) 发明名称

一种基于多粒度异质属性图对比学习的期刊推荐方法

(57) 摘要

本发明公开了一种基于多粒度异质属性图对比学习的期刊推荐方法,涉及数据分析技术领域,该期刊推荐方法包括以下步骤:数据处理、多粒度语义特征提取、基于异质图神经网络对比学习的多粒度结构特征提取和自适应学习。本发明基于异质图神经网络、对比学习、卷积神经网络和文本预处理方法,将文本处理为多粒度并从语义和结构方向下分别进行处理,使用分层自适应学习来将这些处理后的特征整合成一个最终的最优文本特征,使用softmax来得到最终的分类结果;并在考虑到论文和期刊的特殊性下整合了多个损失函数来训练模型。最终达到了只基于论文内容的针对特定研究领域的期刊推荐高质量效果。



1. 一种基于多粒度异质属性图对比学习的期刊推荐方法,其特征在于,包括以下步骤:

步骤1) 数据处理,收集论文的摘要和标题,并从中提取论文的文档粒度、句子粒度、单词粒度、主题粒度和标题粒度的信息;

步骤2) 多粒度语义特征提取,将文档粒度、句子粒度、单词粒度的信息输入进多粒度语义特征提取模块,通过预训练和由卷积层与全连接层组成的特征加强网络得到三种粒度下的语义特征 X_{word} 、 X_{abs} 和 X_{title} ;

步骤3) 基于异质图神经网络对比学习的多粒度结构特征提取,将文档粒度、句子粒度、单词粒度、主题粒度和标题粒度的信息输入进基于异质图神经网络对比学习的多粒度结构特征提取模块,构建全局异质属性图,根据全局异质属性图构建全局异质拓扑图,然后构建四个局部子图,使用基于双层注意力机制的异质图卷积对构建好的全局异质属性图、全局异质拓扑图和四个局部子图进行预训练,使用交叉回路图对比学习随机选取好正负对,然后将预训练好的图分别通过一个全连接网络来进一步增强特征,得到六种视角的结构特征 X_{graphs} ;

基于异质图神经网络对比学习的多粒度结构特征提取的具体步骤如下:

步骤S1: 结构角度下,输入文本的文档、句子、单词、标题和主题五个粒度的信息来构建全局异质属性图 G_{att} , G_{att} 具有D、S、T、W四种节点,D-W,D-S,D-T,T-W,S-W五种连接;

步骤S2: 通过将全局异质属性图 G_{att} 中的D和S节点的属性向量替换为无语义信息的TFIDF向量以及节点W的属性向量替换为无语义信息的one-hot向量,构成为全局异质拓扑图 G_{top} ;

步骤S3: 从全局异质拓扑图 G_{top} 中分离出三个局部子图,由除D以外的每种类型的节点和节点D组成:D-W,D-T,D-S,局部子图包含全局异质拓扑图 G_{top} 中两种类型的节点和对应的连接关系,将D-S局部图中S节点的属性向量替换为Sent2vec向量得到第四个局部子图D-Sv;

步骤S4: 使用基于双层注意力机制的异质图卷积来对以上构建的六个图进行预训练,得到预训练后的向量 X_{pre} ;提取全局异质拓扑图 G_{top} 、全局异质属性图 G_{att} 和四个局部子图中多维节点信息和结构,使用交叉回路图对比学习随机选取好正负对;

步骤S5: 将预训练过全局异质拓扑图 G_{top} 、全局异质属性图 G_{att} 和四个局部子图的D节点的属性向量分别输入到全连接层中进行进一步的特征增强,并将这些全连接层的输出,获得六种视角的结构特征 X_{graphs} ;

步骤4) 自适应学习,使用分层自适应学习来将三种语义特征和六种结构特征融合为一个最优的文本特征 X_{fin} ,其中,整合多种损失函数训练模型;

步骤5),使用softmax对 X_{fin} 进行分类,得到期刊推荐结果。

2. 根据权利要求1所述的基于多粒度异质属性图对比学习的期刊推荐方法,其特征在于,在所述步骤1)中,将论文分解为以下五种粒度:将整个摘要当成一段话的文档粒度、将摘要分解成若干句子的句子粒度、将摘要分解成若干单词的单词粒度,然后采用BTM方法挖掘出摘要的潜在主题形成主题粒度,再加上由论文的标题形成的标题粒度。

3. 根据权利要求1所述的基于多粒度异质属性图对比学习的期刊推荐方法,其特征在在于,在步骤2)中,对于单词粒度,使用Word2vec方法初步训练初始的单词嵌入,再通过一个由卷积层和全连接层组成的特征增强网络模块来对初始的单词嵌入进行特征增强以得到单词粒度的语义特征: $X_{word} = g(C(X_{ini}) * W_{fc} + b_{fc})$;

其中, X_{ini} 代表初始单词嵌入, $C(*)$ 代表卷积操作, W_{fc} 代表全连接层, b_{fc} 代表全连接层的偏置项, $g(*)$ 代表非线性激活函数。

4. 根据权利要求3所述的基于多粒度异质属性图对比学习的期刊推荐方法,其特征在在于,在步骤2)中,对于摘要和标题粒度的信息,先使用Doc2vec和BERT训练他们的初始嵌入,再通过特征增强网络模块分别得到相应粒度的语义特征 X_{abs} 和 X_{title} 。

5. 根据权利要求1所述的基于多粒度异质属性图对比学习的期刊推荐方法,所述全局异质属性图 G_{att} 的构造过程如下:

(1) 将预处理过后的论文摘要作为D节点,将摘要分句和分词得到的句子和单词作为S节点和W节点,将每个D与其分句得到的S节点连接起来;

(2) 将使用BTM挖掘出摘要的潜在主题作为T节点,T节点的属性向量用单词的概率分布表示,对于每个D节点,为其分配概率最高的Top5个主题T并将他们连接起来;

(3) 使用Doc2vec训练文档得到D节点的属性向量,使用基于BERT的Sent2vec训练句子得到S节点的属性向量,使用Word2vec训练单词得到W节点的属性向量;

(4) 如果一个文档D包含一个单词W,在对应的D和W之间构建连接,同理,如果一个句子S包含一个单词W,则在对应的S和W之间构建连接,T-W的连接是基于每个主题的单词概率分布构建的;

(5) 针对每个论文的标题,考虑到标题的特殊性,使用BERT将其训练为嵌入后拼接在对应的D节点属性向量后,得到全局异质属性图 G_{att} 。

6. 根据权利要求1所述的基于多粒度异质属性图对比学习的期刊推荐方法,其特征在在于,所述步骤4) 自适应学习的具体步骤如下:步骤D1:将三种语义特征 X_{word} 、 X_{abs} 和 X_{title} 通过自适应学习融合成一个文本语义特征 X_{sem} ;

$$X_{sem} = \sum \alpha_i X_i;$$

其中, X_i 为 X_{word} 、 X_{abs} 和 X_{title} ;

步骤D2:将六种视角的结构特征 X_{graphs} 通过自适应学习融合成一个文本结构特征 X_{str} ;

$$X_{str} = \sum \beta_i X_i;$$

其中, X_i 为 X_{graphs} ;

步骤D3:分别将这两个特征通过全连接层再通过自适应学习融合成最终的最优文本特征 X_{fin} ;

$$X_{fin} = \sum \gamma_i X_i W_i;$$

其中, X_i 为 X_{str} , X_{sem} ; W_i 代表全连接层。

7. 根据权利要求6所述的基于多粒度异质属性图对比学习的期刊推荐方法, 其特征在于, 在步骤4) 中, 损失函数是在交叉熵的基础之上引入人脸识别领域的Asoftmax损失和对比学习领域的infoNCE损失, 其中, 模型的损失包括预训练损失、语义损失、结构损失和预测损失。

8. 根据权利要求7所述的基于多粒度异质属性图对比学习的期刊推荐方法, 其特征在于, 所述预训练损失为交叉熵, 其中, 图预训练损失 L_{HGCN} :

$$-\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c Y_{ij} * \log Z_{ij};$$

使用文本语义特征的Asoftmax作为语义损失, 其中, 语义损失 L_{sem} :

$$\frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right);$$

其中, $x \in X_{sem}$, $m \geq 1$ 代表控制边角距大小的整数, $m=1$ 时Asoftmax就会退化为普通的Softmax;

对比学习领域的infoNCE损失 L_{const} :

$$\sum_{H, H'} \sum_{d \in H} -\log \frac{\sum_{d+ \in H'} \exp\{(x_d)^T x_{d+}/t\}}{\sum_{d- \in H'} \exp\{(x_d)^T x_{d-}/t\}};$$

其中, H 和 H' 表示图对中的两个图, d 表示 H 中的一个D节点, $d+$ 表示 H' 中的一个同类D节点, $d-$ 表示 H' 中的一个不同类D节点, t 是一个超参数, 在softmax中称为温度系数;

将其与使用结构特征 X_{str} 计算的Asoftmax损失 L_{graph} 结合起来作为结构损失 L_{str} :

$$L_{str} = \delta L_{const} + L_{graph} + \eta \| \theta \|;$$

最终的文本特征 X_{fin} 通过Asoftmax计算预测损失 L_{pred} :

$$\frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right);$$

模型整体损失函数为:

$$L = \lambda_{HGCN} L_{HGCN} + \lambda_{sem} L_{sem} + \lambda_{graph} L_{graph} + \lambda_{const} L_{const} + \lambda_{pred} L_{pred}。$$

一种基于多粒度异质属性图对比学习的期刊推荐方法

技术领域

[0001] 本发明涉及数据分析技术领域,具体是一种基于多粒度异质属性图对比学习的期刊推荐方法。

背景技术

[0002] 近年来学术研究中出现了许多新兴的交叉研究领域,这些领域通常跨多个学科,其内容跨度较大,可发表的期刊选择也更广泛,这使得科研人员尤其是一些新关注、新领域的科研人员在选择合适的发表期刊时经常遇到困难。

[0003] 当前的期刊推荐方法大致可以分为两大类,不基于内容的与基于内容的方法。不基于内容的方法一般是通过发表历史和社交网络等信息进行期刊推荐,不适用于新学者。现存的基于内容的期刊推荐方法大多只针对特定的学术论文平台或期刊类型,目前基本没有针对特定研究领域尤其是新兴交叉研究领域的期刊推荐方法。基于内容的期刊推荐方法,本质上可以视为是一个文本分类问题,目前的文本分类方法可以分为以下四类:

[0004] 基于传统机器学习方法的文本分类:传统的机器学习方法,如SVM,需要一个特征工程步骤将文本转换为嵌入。方法的分类性能往往与嵌入的质量直接相关,由于高质量的特征工程依赖于专业的领域知识,这些方法难以获得良好的性能。

[0005] 基于深度神经网络的文本分类:基于深度学习模型将文本表示为嵌入。深度神经网络在文本分类领域得到了相当广泛的应用。比如TextCNN,LSTM,Transformer等,这些方法往往对数据集要求很高。

[0006] 基于对比学习的文本分类:对比学习最主要应用于自监督和无监督的研究任务中,在有监督的研究任务中的效果算不上很有竞争性,但与文本分类等需要额外训练成本的监督任务相比,对比学习成本较低,可以作为辅助方法。

[0007] 基于图神经网络(GNN)的文本分类:图神经网络与上述方法的最大区别在于,GNN不仅可以处理欧氏空间数据,还可以处理非欧氏空间数据,即图结构化数据。GNN已成功应用于文本分类并取得了很好的效果。然而,由于图结构的复杂性和多样性,基于GNN的文本分类方法在不同数据上的表现差异很大,但这也同时意味着基于GNN的方法往往有很高的提升空间。

发明内容

[0008] 本发明的目的在于提供一种基于多粒度异质属性图对比学习的期刊推荐方法,以解决背景技术中提出的问题。

[0009] 为实现上述目的,本发明提供如下技术方案:

[0010] 一种基于多粒度异质属性图对比学习的期刊推荐方法,包括以下步骤:

[0011] 步骤1) 数据处理,收集论文的摘要和标题,并从中提取论文的文档粒度、句子粒度、单词粒度、主题粒度和标题粒度的信息;

[0012] 步骤2) 多粒度语义特征提取,将文档粒度、句子粒度、单词粒度的信息输入进多粒

度语义特征提取模块,通过预训练和由卷积层与全连接层组成的特征加强网络得到三种粒度下的语义特征 X_{word} 、 X_{abs} 和 X_{title} ;

[0013] 步骤3) 基于异质图神经网络对比学习的多粒度结构特征提取,将文档粒度、句子粒度、单词粒度、主题粒度和标题粒度的信息输入进基于异质图神经网络对比学习的多粒度结构特征提取模块,构建全局异质属性图,根据全局异质属性图构建全局异质拓扑图,然后构建四个局部子图,使用基于双层注意力机制的异质图卷积对构建好的全局异质属性图、全局异质拓扑图和四个局部子图进行预训练,使用交叉回路图对比学习随机选取好正负对,然后将预训练好的图分别通过一个全连接网络来进一步增强特征,得到六种视角的结构特征 X_{graphs} ;

[0014] 步骤4) 自适应学习,使用分层自适应学习来将三种语义特征和六种结构特征融合为一个最优的文本特征 X_{fin} ,其中,整合多种损失函数训练模型;

[0015] 步骤5),使用softmax对 X_{fin} 进行分类,得到期刊推荐结果。

[0016] 在上述技术方案的基础上,本发明还提供以下可选技术方案:

[0017] 在一种可选方案中:在所述步骤1)中,摘要分解为以下三个粒度:将整个摘要当成一段话的文档粒度、将摘要分解成若干句子的句子粒度和将摘要分解成若干单词的单词粒度,采用BTM方法挖掘出摘要的潜在主题形成主题粒度,论文的标题形成标题粒度。

[0018] 在一种可选方案中:在步骤2)中,对于单词粒度,使用Word2vec方法初步训练初始的单词嵌入,再通过一个由卷积层和全连接层组成的特征增强网络模块来对初始的单词嵌入进行特征增强以得到单词粒度的语义特征:

[0019] $X_{word} = g(C(X_{ini}) * W_{fc} + b_{fc})$

[0020] 其中, X_{ini} 代表初始单词嵌入, $C(*)$ 代表卷积操作, W_{fc} 代表全连接层, b_{fc} 代表全连接层的偏置项, $g(*)$ 代表非线性激活函数。

[0021] 在一种可选方案中:在步骤2)中,对于摘要和标题粒度的信息,先使用Doc2vec和BERT训练他们的初始嵌入,再通过特征增强网络模块分别得到相应粒度的语义特征 X_{abs} 和 X_{title} 。

[0022] 在一种可选方案中:所述步骤3) 基于异质图神经网络对比学习的多粒度结构特征提取的具体步骤如下:

[0023] 步骤S1:结构角度下,输入文本的文档、句子、单词、标题和主题五个粒度的信息来构建全局异质属性图 G_{att} , G_{att} 具有D、S、T、W四种节点,D-W,D-S,D-T,T-W,S-W五种连接;

[0024] 步骤S2:通过将全局异质属性图 G_{att} 中的D和S节点的属性向量替换为无语义信息的TFIDF向量以及节点W的属性向量替换为无语义信息的one-hot向量,构成为全局异质拓扑图 G_{top} ;

[0025] 步骤S3:从全局异质拓扑图 G_{top} 中分离出三个局部子图,由除D以外的每种类型的节点和节点D组成:D-W,D-T,D-S,局部子图包含全局异质拓扑图 G_{top} 中两种类型的节点和对应的连接关系,将D-S局部图中S节点的属性向量替换为Sent2vec向量得到第四个局部子图D-Sv;

[0026] 步骤S4:基于双层注意力机制的异质图卷积来对以上构建的六个图进行预训练,得到预训练后的向量 X_{pre} ;提取全局异质拓扑图 G_{top} 、全局异质属性图 G_{att} 和四个局部子图中多维节点信息和结构,通过交叉回路图对比学习随机选取好正负对;

[0027] 步骤S5:将预训练过全局异质拓扑图 G_{top} 、全局异质属性图 G_{att} 和四个局部子图的D节点的属性向量分别输入到全连接层中进行进一步的特征增强,并将这些全连接层的输出,获得六种视角的结构特征 X_{graphs} 。

[0028] 在一种可选方案中:所述全局异质属性图 G_{att} 的构造过程如下:

[0029] (6)将预处理过后的论文摘要作为D节点,将摘要分句和分词得到的句子和单词作为S节点和W节点,将每个D与其分句得到的S节点连接起来;

[0030] (7)将使用BTM挖掘出的摘要的潜在主题作为T节点,T节点的属性向量用单词的概率分布表示,对于每个D节点,为其分配概率最高的Top5个主题T并将他们连接起来;

[0031] (8)使用Doc2vec训练文档得到D节点的属性向量,使用基于BERT的Sent2vec训练句子得到S节点的属性向量,使用Word2vec训练单词得到W节点的属性向量;

[0032] (9)如果一个文档D包含一个单词W,在对应的D和W之间构建连接,同理,如果一个句子S包含一个单词W,则在对应的S和W之间构建连接,T-W的连接是基于每个主题的单词概率分布构建的;

[0033] 针对每个论文的标题,考虑到标题的特殊性,使用BERT将其训练为嵌入后拼接在对应的D节点属性向量后,得到全局异质属性图 G_{att} 。

[0034] 在一种可选方案中:所述步骤4)自适应学习的具体步骤如下:步骤D1:将三种语义特征 X_{word} 、 X_{abs} 和 X_{title} 通过自适应学习融合成一个文本语义特征 X_{sem} :

$$[0035] \quad X_{sem} = \sum \alpha_i X_i$$

[0036] 其中, X_i 为 X_{word} 、 X_{abs} 和 X_{title} ;

[0037] 步骤D2:将六种视角的结构特征 X_{graphs} 通过自适应学习融合成一个文本结构特征 X_{str} :

$$[0038] \quad X_{str} = \sum \beta_i X_i$$

[0039] 其中, X_i 为 X_{graphs} ;

[0040] 步骤D3:分别将这两个特征通过全连接层再通过自适应学习融合成最终的最优文本特征 X_{fin} :

$$[0041] \quad X_{fin} = \sum \gamma_i X_i W_i$$

[0042] 其中, X_i 为 X_{str} 、 X_{sem} ; W_i 代表全连接层。

[0043] 在一种可选方案中:在步骤4)中,损失函数是在交叉熵的基础之上引入人脸识别领域的Asoftmax损失和对比学习领域的infoNCE损失,其中,模型的损失包括预训练损失、语义损失、结构损失和预测损失。

[0044] 在一种可选方案中:所述预训练损失为交叉熵,其中,图预训练损失 L_{HGCN} :

$$[0045] \quad -\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c Y_{ij} * \log Z_{ij}$$

[0046] 使用文本语义特征的Asoftmax作为语义损失,其中,语义损失 L_{sem} :

$$[0047] \quad \frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right)$$

[0048] 其中, $x \in X_{sem}$, $m \geq 1$ 代表控制边角距大小的整数, $m=1$ 时, Asoftmax就会退化为普

通的Softmax;

[0049] 对比学习领域的infoNCE损失 L_{const} :

$$[0050] \quad \sum_{H, H'} \sum_{d \in H} -\log \frac{\sum_{d+ \in H'} \exp\{(x_d)^T x_{d+}/t\}}{\sum_{d- \in H'} \exp\{(x_d)^T x_{d-}/t\}}$$

[0051] 其中, H 和 H' 表示图对中两个图, d 表示 H 中一个D节点, $d+$ 表示 H' 中的一个同类D节点, $d-$ 表示 H' 中的一个不同类D节点, t 是一个超参数, 在softmax中称为温度系数;

[0052] 将其与使用结构特征 X_{str} 计算的Asoftmax损失 L_{graph} 结合起来作为结构损失 L_{str} :

$$[0053] \quad L_{\text{str}} = \delta L_{\text{const}} + L_{\text{graph}} + \eta || \Theta ||$$

[0054] 最终的文本特征 X_{fin} 通过Asoftmax计算预测损失 L_{pred} :

$$[0055] \quad \frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right)$$

[0056] 模型整体损失函数为:

$$[0057] \quad L = \lambda_{\text{HGCN}} L_{\text{HGCN}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{graph}} L_{\text{graph}} + \lambda_{\text{const}} L_{\text{const}} + \lambda_{\text{pred}} L_{\text{pred}}.$$

[0058] 相较于现有技术, 本发明的有益效果如下:

[0059] 本发明基于论文的内容进行期刊推荐, 从多粒度、语义结构双方向出发, 充分提取了文本中的各种信息; 通过多粒度和特征加强网络充分提取了文本在各个粒度下的语义信息。通过构建全局异质图和局部异质图, 配合提出的交叉回路图对比学习, 充分提取并挖掘了文本在各种视角下的结构信息; 并考虑到了以期刊作为标签的特殊性整合了多种损失函数以达到同时挖掘文本信息间的差异性和共性、全局性和局部性, 类内差异过大, 标签含有噪声等性质和问题。完成了针对特定领域的论文期刊推荐工作; 并已经通过实验证明了模型的效果优于现有的文本分类方法。

附图说明

[0060] 图1为本发明一种基于多粒度异质属性图对比学习的期刊推荐方法的整体流程图。

[0061] 图2为本发明中的数据预处理示意图。

[0062] 图3为本发明中的多粒度语义特征提取示意图。

[0063] 图4为本发明中的基于异质图神经网络对比学习的多粒度结构特征提取示意图。

[0064] 图5为本发明中的自适应学习示意图。

[0065] 图6为本发明中的LncRNA数据集上的推荐实例示意图。

[0066] 图7为本发明中的MicroRNA数据集上的推荐实例示意图。

具体实施方式

[0067] 为了使本发明的目的、技术方案及优点更加清楚明白, 以下结合附图及实施例, 对本发明进行进一步详细说明。本发明所列举的各实施例仅用以说明本发明, 并非用以限制本发明的范围。对本发明所作的任何显而易知的修饰或变更都不脱离本发明的精神与范围。

[0068] 在一个实施例中,如图1-图5所示,提供了一种基于多粒度异质属性图对比学习的期刊推荐方法,该方法包括以下步骤:

[0069] 一、数据预处理

[0070] 提取出论文的摘要和标题,将摘要分解为以下三个粒度:1)将整个摘要当成一段话的文档粒度。2)将摘要分解成若干句子的句子粒度。3)将摘要分解成若干单词的单词粒度。

[0071] 然后使用BTM方法挖掘出摘要的潜在主题,再加上论文的标题,得到了论文文档、句子、单词、主题、标题五种粒度的信息。将预处理后数据分别从语义和结构两个角度进行处理。

[0072] 二、多粒度语义特征提取

[0073] 在语义角度下,处理文本的文档、单词、标题粒度信息以充分挖掘文本的语义信息。具体来说,对于单词粒度,首先使用Word2vec方法初步训练初始的单词嵌入,然后通过一个由卷积层和全连接层组成的特征增强网络模块来对初始的单词嵌入进行特征增强以得到单词粒度的语义特征 X_{word} :

$$[0074] \quad X_{word} = g(C(X_{ini}) * W_{fc} + b_{fc})$$

[0075] 其中, X_{ini} 代表初始单词嵌入, $C(*)$ 代表卷积操作, W_{fc} 代表全连接层, b_{fc} 代表全连接层的偏置项, $g(*)$ 代表非线性激活函数。

[0076] 对于摘要和标题粒度的信息,分别先使用Doc2vec和BERT训练他们的初始嵌入,然后通过类似的特征增强网络模块分别得到相应粒度的语义特征 X_{abs} 和 X_{title} 。

[0077] 三、基于异质图神经网络对比学习的多粒度结构特征提取

[0078] 结构角度下,首先,输入文本的文档、句子、单词、标题和主题五个粒度的信息来构建全局异质属性图 G_{att} ,它有D、S、T、W四种节点,D-W,D-S,D-T,T-W,S-W五种连接。构造过程详细描述如下:

[0079] 1)将预处理过后的论文摘要作为D(Document)节点,将摘要分句和分词得到的句子和单词作为S(Sentence)节点和W(Word)节点,将每个D与其分句得到的S节点连接起来。

[0080] 2)将使用BTM挖掘出的摘要的潜在主题作为T(Topic)节点,T节点的属性向量用单词的概率分布表示,对于每个D节点,为其分配概率最高的 T_{op5} 个主题T并将他们连接起来,在实验中挖掘了200个主题T。

[0081] 3)使用Doc2vec训练文档得到D节点的属性向量,使用基于BERT的Sent2vec训练句子得到S节点的属性向量,使用Word2vec训练单词得到W节点的属性向量。

[0082] 4)如果一个文档D包含一个单词W,在对应的D和W之间构建连接,同理,如果一个句子S包含一个单词W,则在对应的S和W之间构建连接,T-W的连接是基于每个主题的单词概率分布构建的。

[0083] 5)除了以上信息,还有每个论文的标题,考虑到标题的特殊性,使用BERT将其训练为嵌入后拼接在对应的D节点属性向量后,将这样构造出的全局异质属性图定义为 G_{att} 。

[0084] 为了更好地捕捉图的拓扑结构,构造了全局异质拓扑图 G_{top} ,他是没有语义属性的图 G_{att} 的拓扑版本,具体来说,将图 G_{att} 的D和S节点的属性向量替换为无语义信息的TFIDF向量,将节点W的属性向量替换为无语义信息的one-hot向量,其他结构不变,将这样构造出的图定义为全局异质拓扑图 G_{top} 。

[0085] 局部结构比全局图更能代表单一类型节点对文本的影响;为了进一步捕获图的局部结构,构造了四个局部图。

[0086] 具体来说,从全局异质拓扑图 G_{top} 中分离出三个局部子图,由除D以外的每种类型的节点和节点D组成:D-W,D-T,D-S,这些局部图只包含全局异质拓扑图 G_{top} 中两种类型的节点和对应的连接关系。将D-S局部图中S节点的属性向量替换为Sent2vec向量得到第四个局部图D-Sv。

[0087] 众所周知,GCN是图神经网络最常用的方法。对于任意图 $G=(V,E)$,设 X 为图 G 的节点及其特征向量矩阵,设 $A'=A+I$ 为图 G 添加了自连接的邻接矩阵,设 M 为度矩阵,其中 $M_{ii}=\sum_j A'_{ij}$,则GCN的传播规则如下:

$$[0088] \quad H^{(l+1)} = \sigma(\tilde{A} \cdot H^{(l)} \cdot W^{(l)})$$

[0089] 其中 $\tilde{A} = M^{-\frac{1}{2}} A' M^{-\frac{1}{2}}$, $W^{(l)}$ 是一个可训练的变换矩阵, $\sigma(\cdot)$ 是一个非线性激活函数 $H^{(1)}$ 表示第1层节点的隐藏表示, $H^{(0)}=X$ 。

[0090] 然而,由于异质性,GCN不能简单的应用于异质图神经网络。我们使用异质图卷积(HGCN),为异质图中的每一类节点都设置一个变换矩阵,通过各类变化矩阵将图中不同类型的节点投射到同一隐式公共空间中,HGCN的传播规则如下:

$$[0091] \quad H^{(l+1)} = \sigma\left(\sum_{r \in \gamma} \tilde{A}_r \cdot H_r^{(l)} \cdot W_r^{(l)}\right)$$

[0092] 其中 $\tilde{A}_r$ 是 $\tilde{A}$ 的子矩阵,表示相邻的类型为 r 的节点, $W_r^{(l)}$ 为将 r 类型的节点投射到隐式公共空间的可训练矩阵。 $H_0^{(r)} = X_r$ 。

[0093] 双层注意力机制的异质图卷积使用类型级和节点级注意力计算不同类型节点之间的重要性权重,传播规则如下:

$$[0094] \quad H^{(l+1)} = \sigma\left(\sum_{r \in \gamma} \beta_r \cdot H_r^{(l)} \cdot W_r^{(l)}\right)$$

[0095] 其中 β_r 代表双层注意力矩阵。

[0096] 我们使用基于双层注意力机制的异质图卷积来对以上构建的六个图进行预训练来得到预训练后的向量 X_{pre} :

$$[0097] \quad X_{pre} = H^{(L)}$$

[0098] 双层注意力机制只捕获了不同类型节点信息之间的差异性,而没有考虑他们之间的共性,这些共性反映了文本对不同类型信息和结构的共同依赖,也是需要捕获的重要信息。通过构建上述图提取了多维节点信息和结构,接下来通过引入对比学习来捕获多维信息和结构的共性。

[0099] 提出了交叉回路图对比学习,这是一种同时包含属性--拓扑两种结构,全局--全局,全局--局部,局部--局部三种连接的对比学习方法。

[0100] 具体地说,将上述六个图中的每两个连接为一对,将图对中同类的D节点视为正样本对,不同类的D节点视为负样本对。在这一步基于交叉回路图对比学习随机选择好正样本

对和负样本对。然后将这六个预训练过的图的D节点的属性向量分别输入到全连接层中进行进一步的特征增强,并将这些全连接层的输出记录为 X_{graphs} 。

[0101] 四、自适应学习

[0102] 至此,获得了三种粒度的语义特征 X_{word} 、 X_{abs} 和 X_{title} 和六种视角的结构特征 X_{graphs} 。使用分层自适应学习来将其融合为一个最优的文本特征。

[0103] 具体地说,首先将三种粒度的语义特征通过自适应学习融合成一个文本语义特征 X_{sem} :

$$[0104] \quad X_{\text{sem}} = \sum \alpha_i X_i$$

[0105] 其中, X_i 为 X_{word} 、 X_{abs} 和 X_{title} 。

[0106] 同时将六种视角的结构特征 X_{graphs} 通过自适应学习融合成一个文本结构特征 X_{str} :

$$[0107] \quad X_{\text{str}} = \sum \beta_i X_i$$

[0108] 其中, X_i 为 X_{graphs} ;

[0109] 然后分别将这两个特征通过全连接层再通过自适应学习融合成最终的最优文本特征 X_{str} :

$$[0110] \quad X_{\text{str}} = \sum \beta_i X_i$$

[0111] 其中, X_i 为 X_{graphs} ;

[0112] 然后分别将这两个特征通过全连接层再通过自适应学习融合成最终的最优文本特征 X_{fin} :

$$[0113] \quad X_{\text{fin}} = \sum \gamma_i X_i W_i$$

[0114] 其中, X_i 为 X_{str} 、 X_{sem} ; W_i 代表全连接层;

[0115] 使用softmax对 X_{fin} 进行分类,得到期刊推荐结果。

[0116] 五、损失函数

[0117] 没有依赖于以往文本分类方法使用的交叉熵损失函数。而是在交叉熵的基础之上引入了人脸识别领域的Asoftmax损失和对比学习领域的infoNCE损失。将模型的损失分为四部分:

[0118] 图预训练损失,语义损失,结构损失,预测损失。

[0119] (1) 图预训练损失:

[0120] 发现虽然Asoftmax更适配于的研究问题,但随着模型的参数变多、层数变深,Asoftmax损失函数的效果下降的要比传统交叉熵更快,所以在图预训练这一步仍然沿用交叉熵作为图预训练损失 L_{HGCN} :

$$[0121] \quad -\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^c Y_{ij} * \log Z_{ij}$$

[0122] (2) 语义损失:

[0123] 相比传统的交叉熵损失,Asoftmax可以在扩大类间差距的同时显著减小类内差距,很适合于这种标签类内差距大于类间差距的研究问题。所以使用语义特征 X_{sem} 的Asoftmax作为语义损失 L_{sem} :

$$[0124] \quad \frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right)$$

[0125] 其中, $x \in X_{\text{sem}}$, $m \geq 1$ 代表控制边角距大小的整数, $m=1$ 时, Asoftmax 就会退化为普通的 Softmax。

[0126] (3) 结构损失:

[0127] 在结构角度下, 使用前文交叉回路对比学习选定好的正负对, 引入对比学习领域的 infoNCE 损失计算对比学习损失 L_{const} :

$$[0128] \quad \sum_{H, H'} \sum_{d \in H} -\log \frac{\sum_{d+ \in H'} \exp\{(x_d)^T x_{d+}/t\}}{\sum_{d- \in H'} \exp\{(x_d)^T x_{d-}/t\}}$$

[0129] 其中, H 和 H' 表示图对中两个图, d 表示 H 中一个 D 节点, $d+$ 表示 H' 中的一个同类 D 节点, $d-$ 表示 H' 中的一个不同类 D 节点, t 是一个超参数, 在 softmax 中称为温度系数。

[0130] 将其与使用结构特征 X_{str} 计算得的 Asoftmax 损失 L_{graph} 结合起来作为结构损失 L_{str} :

$$[0131] \quad L_{\text{str}} = \delta L_{\text{const}} + L_{\text{graph}} + \eta || \Theta ||$$

[0132] (4) 预测损失:

[0133] 得到了最终的文本特征 X_{fin} 之后, 同样使用 Asoftmax 计算得预测损失 L_{pred} :

$$[0134] \quad \frac{1}{N} \sum_i -\log \left(\frac{e^{\|x_i\| \varphi(\theta_{y_i}, i)}}{e^{\|x_i\| \varphi(\theta_{y_i}, i)} + \sum_{j \neq y_i} e^{\|x_i\| \varphi(\theta_{y_i}, i)}} \right)$$

[0135] 模型整体损失函数为:

$$[0136] \quad L = \lambda_{\text{HGCN}} L_{\text{HGCN}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{graph}} L_{\text{graph}} + \lambda_{\text{const}} L_{\text{const}} + \lambda_{\text{pred}} L_{\text{pred}}.$$

[0137] 1. LncRNA 相关论文数据集实验结果

[0138] 从 Pubmed 上获取了近五年来的 LncRNA 相关文章, 这是一个典型的新兴交叉研究热点领域。经过分析与筛选, 选择了其中发表文章最多的前 50 个期刊及其论文作为数据集, 共含有 LncRNA 相关论文 15114 篇。

[0139] 模型的效果如表 1 所示, 模型的期刊推荐实例如图 6 所示。

[0140] 表 1. LncRNA 数据集上的 Top1-Top10 分类正确率结果

	TOP5	TOP1	TOP2	TOP3	TOP4	TOP6	TOP7	TOP8	TOP9	TOP10
	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac
[0141] Our method	78.03	43.68	58.84	67.50	73.53	80.41	72.79	85.24	86.96	88.42

[0142] 2. MicroRNA 相关论文数据集实验结果

[0143] MicroRNA 也是近年来新兴的交叉研究领域, 关于其研究也是科研人员的热点, 从 Pubmed 上获取了近五年来的 MicroRNA 相关文章。经过分析与筛选, 选择其中发表文章最多的前 50 个期刊及其论文作为的数据集, 共含有 MicroRNA 相关论文 29094 篇。

[0144] 模型的效果如表2所示,模型的期刊推荐实例如图7所示。

[0145] 表2.MicroRNA数据集上的Top1-Top10分类正确率结果

[0146]

	TOP5	TOP1	TOP2	TOP3	TOP4	TOP6	TOP7	TOP8	TOP9	TOP10
	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac	Ac
Our method	76.01	43.55	56.41	65.31	71.16	78.62	81.47	83.67	85.08	86.94

[0147] 以上所述,仅为本公开的具体实施方式,但本公开的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本公开揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本公开的保护范围之内。因此,本公开的保护范围应以权利要求的保护范围为准。

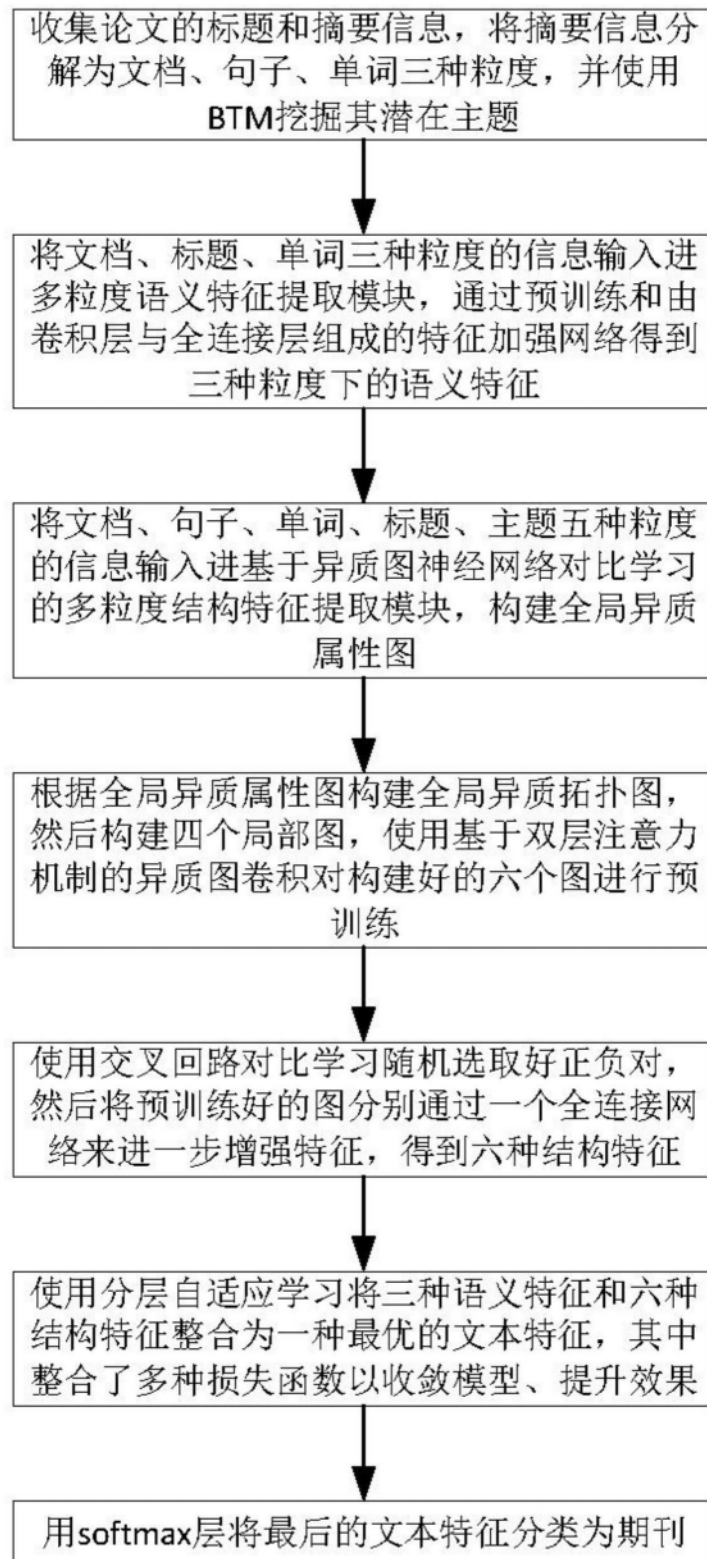


图1

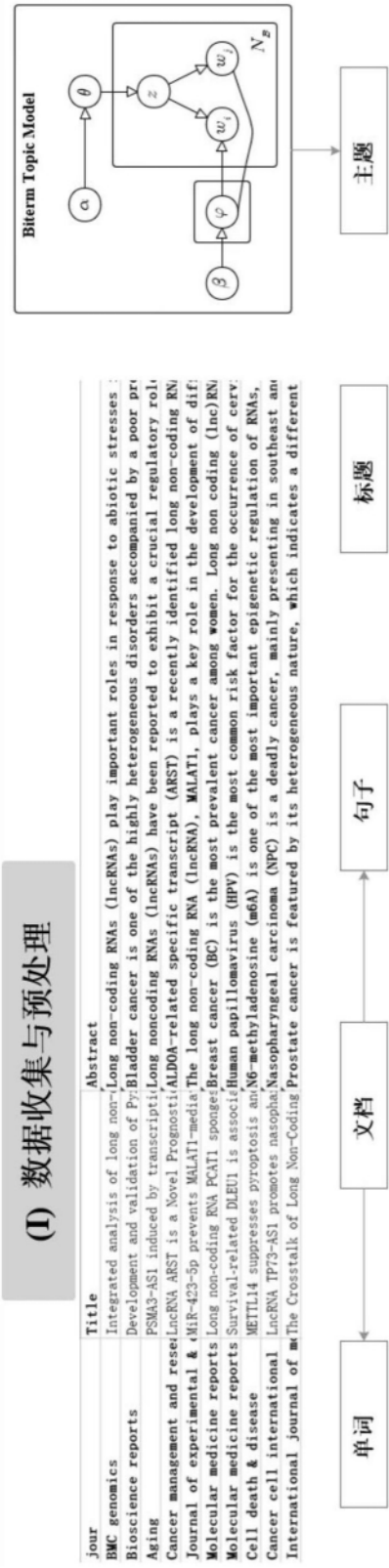


图2

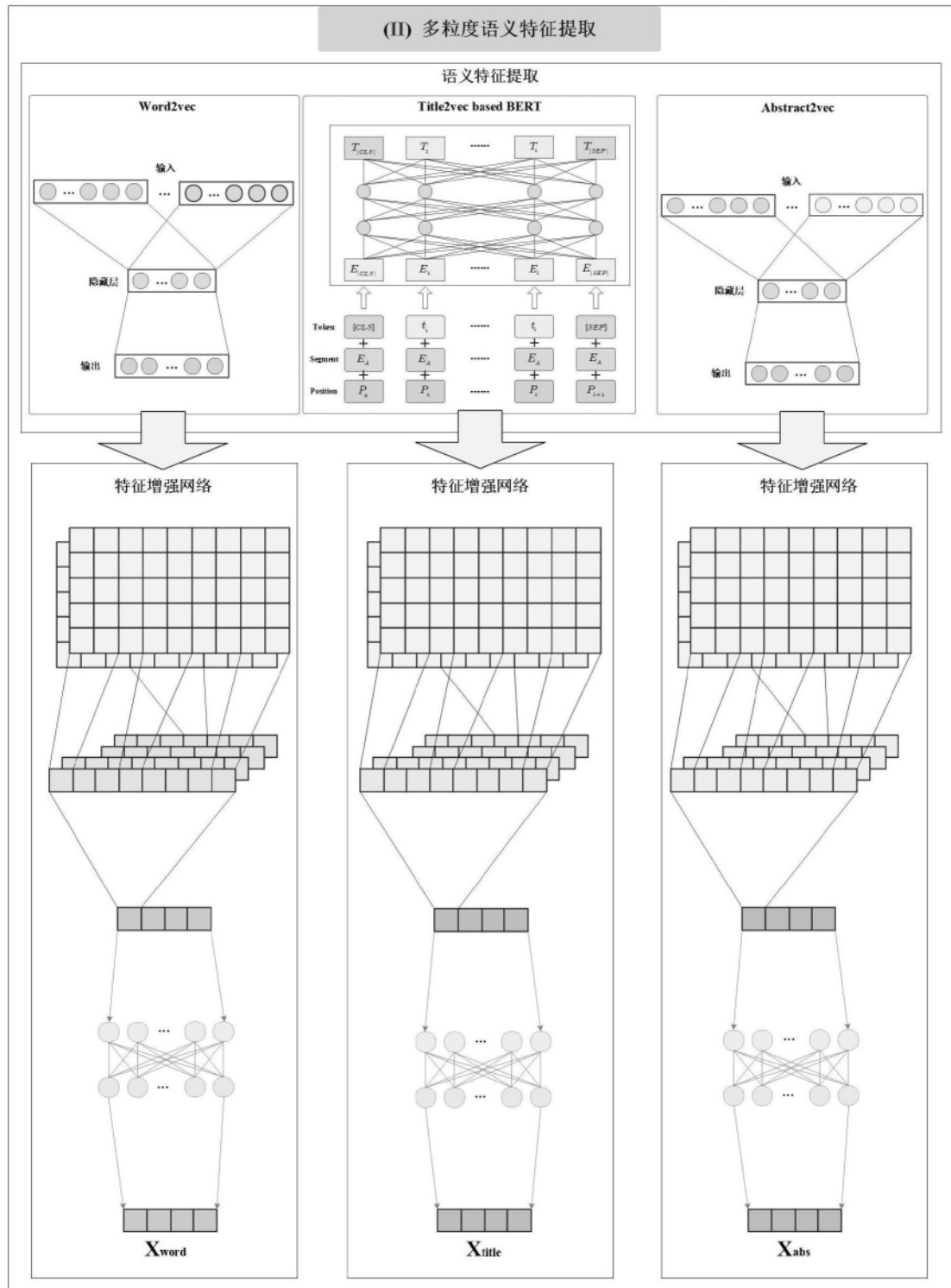


图3

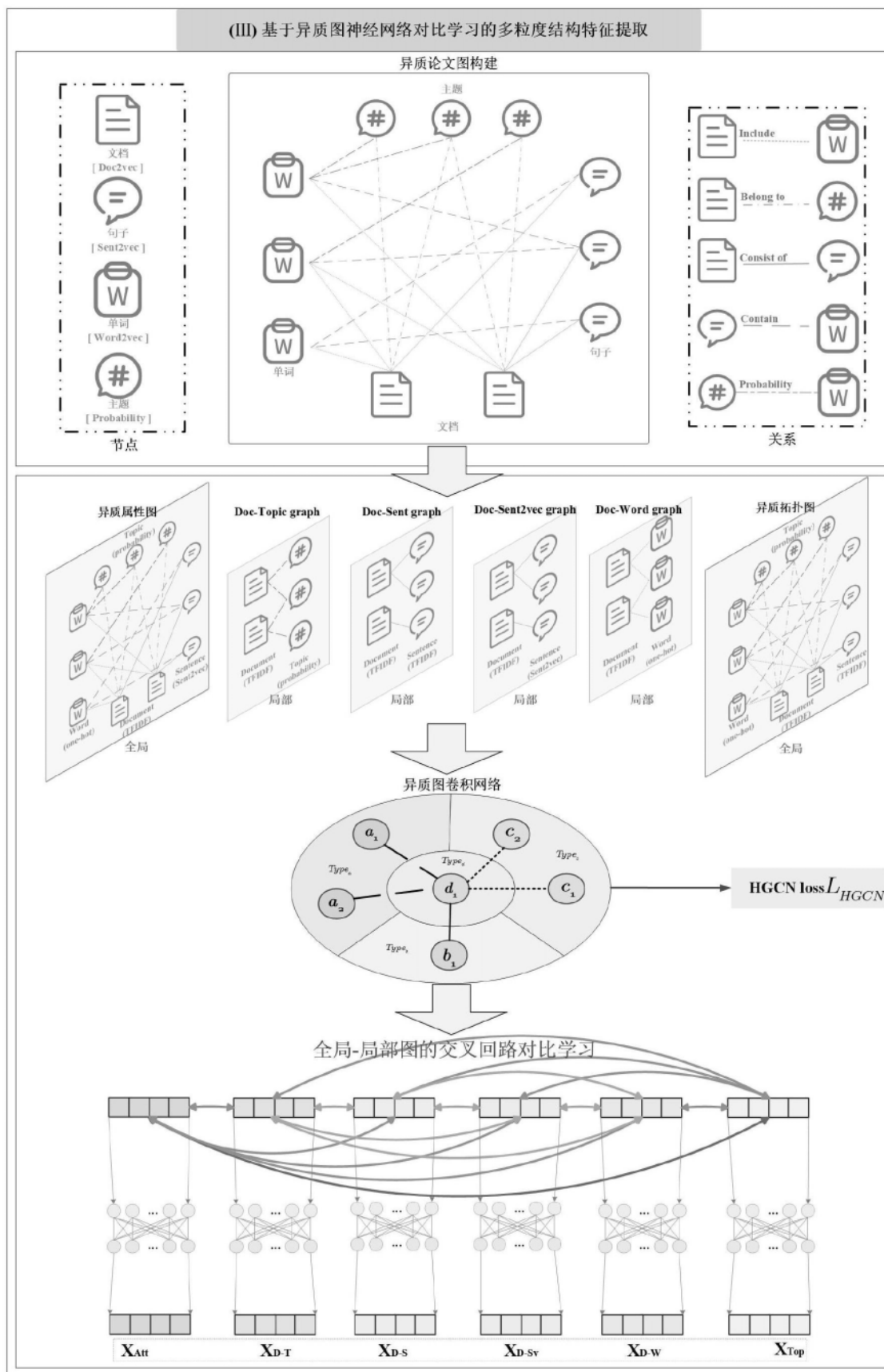


图4

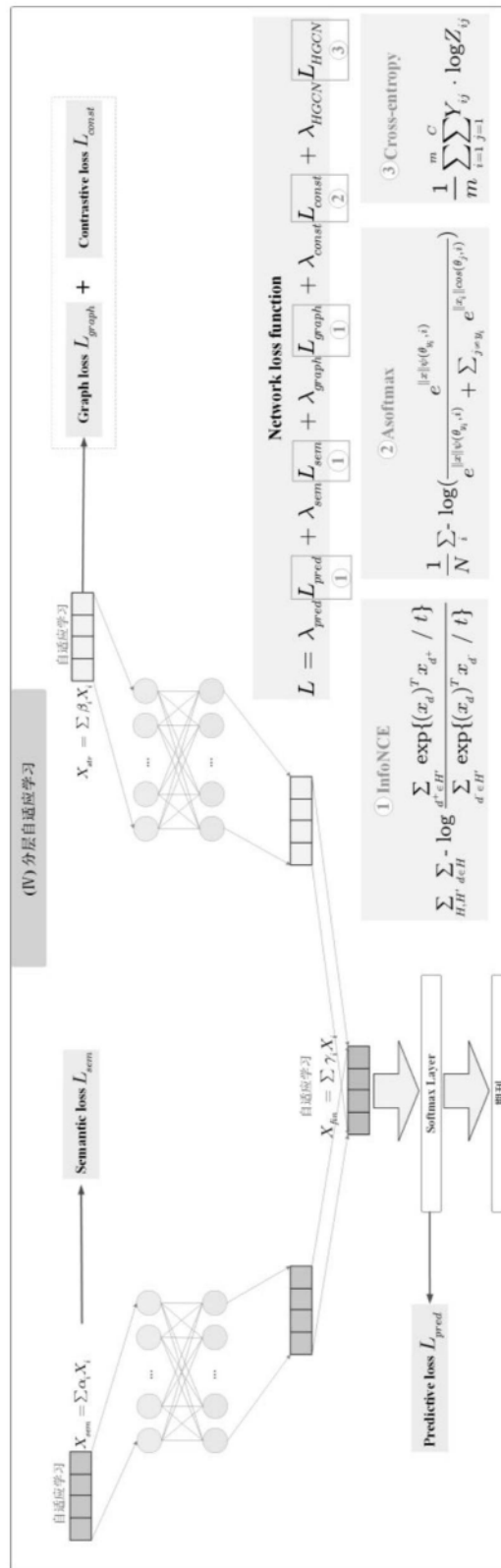


图5

Title and abstract (example)

Trends in the development of miRNA bioinformatics tools

MicroRNAs (miRNAs) are small noncoding RNAs that regulate gene expression via recognition of cognate sequences and interference of transcriptional, translational or epigenetic processes. Bioinformatics tools developed for miRNA study include those for miRNA prediction and discovery, structure, analysis and target prediction. We manually curated 95 review papers and ~1000 miRNA

Briefings in bioinformatics https://academic.oup.com/bib
Methods in molecular biology (Clifton, N.J.) https://link.springer.com/bookseries/7651
Medical science monitor : international medical journal of experimental and clinical research https://medscimonit.com/
International journal of molecular sciences http://www.mdpi.com/journal/ijms
Frontiers in genetics http://www.frontiersin.org/genetics

图6

Title and abstract (example)

bioinformatics tools published since 2003. We classified and ranked them based on citation number or PageRank score, and then performed network analysis and text mining (TM) to study the miRNA tools development trends. Five key trends were observed: (1) miRNA identification and target prediction have been hot spots in the past decade; (2) manual curation and TM are the main methods for collecting miRNA knowledge from literature; (3) most early tools are well maintained and widely used;

Briefings in bioinformatics https://academic.oup.com/bib
Methods in molecular biology (Clifton, N.J.) https://link.springer.com/bookseries/7651
Molecular biology reports https://link.springer.com/journal/11033
Genes http://www.mdpi.com/journal/genes
Frontiers in genetics http://www.frontiersin.org/genetics

图7