



(12) 发明专利

(10) 授权公告号 CN 113360776 B

(45) 授权公告日 2023.07.21

(21) 申请号 202110814971.X  
 (22) 申请日 2021.07.19  
 (65) 同一申请的已公布的文献号  
 申请公布号 CN 113360776 A  
 (43) 申请公布日 2021.09.07  
 (73) 专利权人 西南大学  
 地址 400715 重庆市北碚区天生路2号  
 (72) 发明人 肖国强 唐小琴 王晓蒙 吴松  
 程天宇  
 (74) 专利代理机构 北京海虹嘉诚知识产权代理  
 有限公司 11129  
 专利代理师 胡博文  
 (51) Int.Cl.  
 G06F 16/9535 (2019.01)  
 G06F 16/958 (2019.01)  
 G06F 16/22 (2019.01)  
 G06F 16/2458 (2019.01)  
 G06F 16/248 (2019.01)

G06F 16/26 (2019.01)  
 G06F 16/31 (2019.01)  
 G06F 16/335 (2019.01)  
 G06F 16/338 (2019.01)  
 G06F 16/34 (2019.01)  
 G06F 40/216 (2020.01)  
 G06F 40/284 (2020.01)  
 G06F 40/289 (2020.01)  
 G06F 40/117 (2020.01)

(56) 对比文件

CN 106815297 A, 2017.06.09  
 CN 112184334 A, 2021.01.05  
 CN 105512323 A, 2016.04.20  
 CN 111897999 A, 2020.11.06  
 CN 111931043 A, 2020.11.13  
 US 2014172627 A1, 2014.06.19  
 CA 3063243 A1, 2021.05.13

审查员 吕亦昕

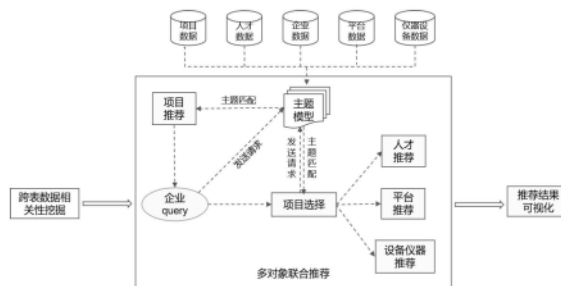
权利要求书2页 说明书4页 附图2页

(54) 发明名称

基于跨表数据挖掘的科技资源推荐方法

(57) 摘要

本发明公开了一种基于跨表数据挖掘的科技资源推荐方法,该方法通过前期数据语义分析,挖掘跨表数据或多对象属性之间的相关性,确定用于NLP主题模型的输入数据字段,在一定程度上优化了多对象之间的数据交流模式;通过采用的神经网络主题模型结构简洁,且无需先验假设,通过训练可获得质量更高的主题表示;通过对多对象推荐结果根据推荐指数和对象种类,进行不同大小和颜色的图模型展示及可视化,可实现推荐结果的直观、有效、合理显示,提升用户体验。



1. 一种基于跨表数据挖掘的科技资源推荐方法,其特征在于,包括步骤:

S1: 构建包括企业、人才、项目、平台和仪器设备属性数据的对象表,选取与对象表中各对象属性相关性最高的关联对象作为跨表数据交流的信息通道;

S2: 从所述对象表中提取与关联对象的属性数据对应的属性数据,并根据提取出的属性数据构建NLP主题模型形成文档数据;步骤S2具体包括:提取企业-业务范围数据、项目名称数据、人才-熟悉学科数据、仪器设备-主要功能数据和平台-研究方向数据输入NLP主题模型形成文档数据;其中,每个记录或样本对应的数据定义为一个文档数据 $d = \{d_1, d_2, \dots, d_N\}$ ,N表示文档总数;

S3: 对所述文档数据进行分词处理,然后将分词后的文档数据输入创建好的神经网络主题模型NTM进行训练,求得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ 及对应的权重矩阵 $W_\theta$ 和 $W_\phi$ ,并生成隐含层对应每个主题下的词汇集及其出现概率;

S4: 通过训练好的主题-词汇分布 $\phi$ ,求出与用户搜索的关键词匹配度最高的主题 $t^*$ ;再根据要求返回的对象,计算主题 $t^*$ 对应词汇集 $g^{t^*}$ 出现在每个对象文档数据d的概率 $p(g^{t^*} | d)$ ,然后对计算结果进行从大到小排序后将对应的对象ID作为推荐系数返回给企业用户。

2. 根据权利要求1所述的基于跨表数据挖掘的科技资源推荐方法,其特征在于,所述步骤S3具体包括:

S31: 对文档集d进行n-gram分词得到词汇集g,构建神经网络主题模型NTM,并将每个文档集d及其n-gram词汇集g作为神经网络主题模型NTM的输入层;

S32: 添加n-gram词向量层,定义词向量维度为300,将每个词汇集g转换成数字向量 $le(g)$ 进行表示;

S33: 创建文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ 的两个隐含层 $ld(d)$ 和 $lt(g)$ ,主题数量为K;其中, $ld(d) = \text{softmax}(W_\theta(d))$ , $lt(g) = \text{sigmoid}(le(g) \times W_\phi)$ ,其中权重矩阵 $W_\theta$ 表示N个文档向量在K个主题上的分布,即 $W_\theta \in R^{N \times K}$ , $W_\theta(d)$ 为文档集d的权重矩阵; $W_\phi$ 表示主题-词汇层K个主题与词向量层300维词向量之间的权重矩阵,故 $W_\phi \in R^{300 \times K}$ 。因文档主题个数为K,则 $ld$ 和 $lt$ 均是一个K维向量;模型输出为文档集d关于词汇集g的分布概率 $p(g|d) = \phi \times \theta^T = lt(g) \times ld^T(d)$ ;

S34: 将步骤S31中每个样本数据(d,g),和通过统计标注获得的每个词汇集g在文档集d中出现的概率 $p(g|d)$ 分别作为神经网络主题模型NTM的输入和输出进行训练,获得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ ,以及对应的权重矩阵 $W_\theta$ 和 $W_\phi$ 。

3. 根据权利要求2所述的基于跨表数据挖掘的科技资源推荐方法,其特征在于,所述步骤S31中,文档集d采用unigram和bigrams模型生成词汇集 $g = \{g_1, g_2, \dots, g_V\}$ ,V表示文档的词汇数量。

4. 根据权利要求1所述的基于跨表数据挖掘的科技资源推荐方法,其特征在于,该方法还包括:

S5: 采用图结构对步骤S4得到的推荐结果进行可视化。

5. 根据权利要求4所述的基于跨表数据挖掘的科技资源推荐方法,其特征在于,所述步

骤S5具体包括：

S51:依据步骤S3的推荐指数 $p(g^{t^*} | d)$ 对图节点的大小进行定义,使推荐指数高的对象在图空间的节点面积最大,且距离图空间中该图节点最近。

6.根据权利要求5所述的基于跨表数据挖掘的科技资源推荐方法,其特征在于,所述步骤S5还包括：

S52:采用不同的颜色对不同对象进行区分和可视化。

## 基于跨表数据挖掘的科技资源推荐方法

### 技术领域

[0001] 本发明涉及一种基于跨表数据挖掘的科技资源推荐方法。

### 背景技术

[0002] 科技资源在国民经济发展中愈发重要,在科技活动中的共享和利用程度也得到相关部门和企业的高度重视,资源的多维大数据特征得以突显。

[0003] 科技资源具有领域性强、数据分散、地域性强的特点,当前的通用性推荐算法(如用户行为分析、协同过滤技术等)在实际的资源共享平台中的应用效果不佳,资源推荐准确率偏低。隐语义分析技术是一种基于机器学习的一系列方法,具有比较好的理论基础,目前部分算法在推荐系统中已经得到应用和肯定。但是,目前的推荐方法均不能进行跨表数据挖掘,从而实现基于多维度数据向企业准确推荐科技资源。

### 发明内容

[0004] 本发明的目的是提供一种基于跨表数据挖掘的科技资源推荐方法,能够为企业自动推荐科技资源。

[0005] 为解决上述技术问题,本发明提供一种基于跨表数据挖掘的科技资源推荐方法,包括步骤:

[0006] S1:构建包括企业、人才、项目、平台和仪器设备属性数据的对象表,选取与对象表中各对象属性相关性最高的关联对象作为跨表数据交流的信息通道;

[0007] S2:从所述对象表中提取与关联对象的属性数据对应的属性数据,并根据提取出的属性数据构建NLP主题模型形成文档数据;

[0008] S3:对所述文档数据进行分词处理,然后将分词后的文档数据输入创建好的神经网络主题模型NTM进行训练,求得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ 及对应的权重矩阵 $W_\theta$ 和 $W_\phi$ ,并生成隐含层对应每个主题下的词汇集及其出现概率;

[0009] S4:通过训练好的主题-词汇分布 $\phi$ ,求出与用户搜索的关键词匹配度最高的主题 $t^*$ ;再根据要求返回的对象,计算主题 $t^*$ 对应词汇集 $g^{t^*}$ 出现在每个对象文档数据 $d$ 的概率 $p(g^{t^*}|d)$ ,然后对计算结果进行从大到小排序后将对应的对象ID作为推荐系数返回给企业用户。

[0010] 进一步地,所述步骤S2具体包括:提取企业-业务范围数据、项目-名称数据、人才-熟悉学科数据、仪器设备-主要功能数据和平台-研究方向数据输入NLP主题模型形成文档数据;其中,每个记录或样本对应的数据定义为一个文档数据 $d = \{d_1, d_2, \dots, d_N\}$ , $N$ 表示文档总数。

[0011] 进一步地,所述步骤S3具体包括:

[0012] S31:对文档集 $d$ 进行 $n$ -gram分词得到词汇集 $g$ ,构建神经网络主题模型NTM,并将每个文档集 $d$ 及其 $n$ -gram词汇集 $g$ 作为神经网络主题模型NTM的输入层;

[0013] S32:添加n-gram词向量层,定义词向量维度为300,将每个词汇集g转换成数字向量 $le(g)$ 进行表示;

[0014] S33:创建文档-主题分布 $\theta$ 和主题-词汇分布 $\Phi$ 的两个隐含层 $ld(d)$ 和 $lt(g)$ ,主题数量为K;其中, $ld(d) = \text{softmax}(W_0(d))$ , $lt(g) = \text{sigmoid}(le(g) \times W_\phi)$ ,其中权重矩阵 $W_0$ 表示N个文档向量在K个主题上的分布,即 $W_0 \in R^{N \times K}$ , $W_0(d)$ 为文档集d的权重矩阵; $W_\phi$ 表示主题-词汇层K个主题与词向量层300维词向量之间的权重矩阵,故 $W_\phi \in R^{300 \times K}$ 。因文档主题个数为K,则 $ld$ 和 $lt$ 均是一个K维向量;模型输出为文档集d关于词汇集g的分布概率

$$p(g|d) = \phi \times \theta^T = lt(g) \times ld^T(d);$$

[0015] S34:将步骤S31中每个样本数据 $(d, g)$ ,和通过统计标注获得的每个词汇集g在文档集d中出现的概率 $p(g|d)$ 分别作为神经网络主题模型NTM的输入和输出进行训练,获得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ ,以及对应的权重矩阵 $W_0$ 和 $W_\phi$

[0016] 进一步地,所述步骤S31中,文档集d采用unigram和bigrams模型生成词汇集 $g = \{g_1, g_2, \dots, g_v\}$ ,V表示文档的词汇数量。

[0017] 进一步地,该方法还包括:

[0018] S5:采用图结构对步骤S4得到的推荐结果进行可视化。

[0019] 进一步地,所述步骤S5具体包括:

[0020] S51:依据步骤S3的推荐指数 $p(g^*|d)$ 对图节点的大小进行定义,使推荐指数高的对象在图空间的节点面积最大,且距离图空间中该图节点最近。

[0021] 进一步地,所述步骤S5还包括:

[0022] S52:采用不同的颜色对不同对象进行区分和可视化。

[0023] 本发明的有益效果为:通过前期数据语义分析,挖掘跨表数据或多对象属性之间的相关性,确定用于NLP主题模型的输入数据字段,在一定程度上优化了多对象之间的数据交流模式;在采用多对象联合推荐技术推荐适合企业发展的科技项目,并为该项目准确推荐多种合适的科技资源如人才、仪器设备等。

## 附图说明

[0024] 此处所说明的附图用来提供对本申请的进一步理解,构成本申请的一部分,在这些附图中使用相同的参考标号来表示相同或相似的部分,本申请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。在附图中:

[0025] 图1为基于跨表数据挖掘的科技资源推荐方法实施方案;

[0026] 图2为面向企业的科技资源数据表达与挖掘方法示意图;

[0027] 图3为神经网络主题模型构建方法。

## 具体实施方式

[0028] 如图1所示的基于跨表数据挖掘的科技资源推荐方法,该方法包括步骤:

[0029] S1:构建包括企业、人才、项目、平台和仪器设备属性数据的对象表,选取与对象表中各对象属性相关性最高的关联对象作为跨表数据交流的信息通道;

[0030] S2:从所述对象表中提取与关联对象的属性数据对应的属性数据,并根据提取出

的属性数据构建NLP主题模型形成文档数据；

[0031] S3:对所述文档数据进行分词处理,然后将分词后的文档数据输入创建好的神经网络主题模型NTM进行训练,求得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ 及对应的权重矩阵 $W_\theta$ 和 $W_\phi$ ,并生成隐含层对应每个主题下的词汇集及其出现概率;

[0032] S4:通过训练好的主题-词汇分布 $\phi$ ,求出与用户搜索的关键词匹配度最高的主题 $t^*$ ;再根据要求返回的对象,计算主题 $t^*$ 对应词汇集 $g^{t^*}$ 出现在每个对象文档数据 $d$ 的概率 $p(g^{t^*}|d)$ ,然后对计算结果进行从大到小排序后将对应的对象ID作为推荐系数返回给企业用户。

[0033] 本发明通过前期数据语义分析,挖掘跨表数据或多对象属性之间的相关性,确定用于NLP主题模型的输入数据字段,在一定程度上优化了多对象之间的数据交流模式;在采用多对象联合推荐技术推荐适合企业发展的科技项目,并为该项目准确推荐多种合适的科技资源如人才、仪器设备等。

[0034] 根据本申请的一个实施例,所述步骤S2具体包括:提取企业-业务范围数据、项目名称数据、人才-熟悉学科数据、仪器设备-主要功能数据和平台-研究方向数据输入NLP主题模型形成文档数据;其中,每个记录或样本对应的数据定义为一个文档数据 $d = \{d_1, d_2, \dots, d_N\}$ , $N$ 表示文档总数。跨表数据包含企业、项目、人才、平台和仪器设备5种对象及其属性数据,通常这些属性对于不同的需求,其重要性和价值有所不同,附图2展示了各对象的部分属性。针对科技资源推荐应用场景,各资源数据的相关性主要体现在专业方向和实用价值方面,故定义一个用于关联多种数据对象的语义概念“研究方向”。随后,从每个数据对象表中选择与该语义概念最匹配的属性作为跨表数据交流的信息通道,并将该表中该属性对应的数据用于构建NLP主题模型。拟选择的对象属性包括:企业-“业务范围”、项目-“名称”、人才-“熟悉学科”、仪器设备-“主要功能”和平台-“研究方向”。

[0035] 根据本申请的一个实施例,所述步骤S3具体包括:

[0036] S31:对文档集 $d$ 进行 $n$ -gram分词得到词汇集 $g$ ,并将每个文档集 $d$ 及其 $n$ -gram词汇集 $g$ 作为神经网络主题模型NTM的输入层;

[0037] S32:添加 $n$ -gram词向量层,定义词向量维度为300,将每个词汇集 $g$ 转换成数字向量 $le(g)$ 进行表示;通过实现文本数据的量化表示,提高文本数据的可运算性和可操作性。

[0038] S33:创建文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ 的两个隐含层 $ld(d)$ 和 $lt(g)$ ,主题数量为 $K$ ;不同于传统概率主题模型,NTM无需指定先验分布,而是分别采用神经网络常用的softmax和sigmoid函数从权重矩阵中生隐含层 $ld$ 和 $lt$ ,即: $ld(d) = \text{softmax}(W_\theta(d))$ , $lt(g) = \text{sigmoid}(le(g) \times W_\phi)$ ,其中权重矩阵 $W_\theta$ 表示 $N$ 个文档向量在 $K$ 个主题上的分布,即 $W_\theta \in \mathbb{R}^{N \times K}$ , $W_\theta(d)$ 为文档集 $d$ 的权重矩阵。 $W_\phi$ 表示主题-词汇层 $K$ 个主题与词向量层300维词向量之间的权重矩阵,故 $W_\phi \in \mathbb{R}^{300 \times K}$ 。因文档主题个数为 $K$ ,则 $ld$ 和 $lt$ 均是一个 $K$ 维向量。模型输出为文档集 $d$ 关于词汇集 $g$ 的分布概率 $p(g|d) = \phi \times \theta^T = lt(g) \times ld^T(d)$ 。

[0039] S34:将步骤S31中每个样本数据 $(d, g)$ ,和通过统计标注获得的每个词汇集 $g$ 在文档集 $d$ 中出现的概率 $p(g|d)$ 分别作为神经网络主题模型NTM的输入和输出进行训练,获得文档-主题分布 $\theta$ 和主题-词汇分布 $\phi$ ,以及对应的权重矩阵 $W_\theta$ 和 $W_\phi$ 。通过训练隐含层的主题模

型,生成同一主题下语义信息相似的词汇 $g^t$ ,且这些词汇隶属于该主题的概率 $\Phi$ 最大,如人才对象表中的以下词汇隶属于同一主题:{模式,识别,图像,处理,人工,智能,系统,计算机,机器,学习,深度}。

[0040] 根据本申请的一个实施例,所述步骤S31中,文档集 $d$ 采用unigram和bigrams模型生成词汇集 $g = \{g_1, g_2, \dots, g_v\}$ , $V$ 表示文档的词汇数量。

[0041] 根据本申请的一个实施例,该方法还包括:

[0042] S5:采用图结构对步骤S4得到的推荐结果进行可视化。

[0043] 根据本申请的一个实施例,所述步骤S5具体包括:

[0044] S51:依据步骤S3的推荐指数 $p(g^t | d)$ 对图节点的大小进行定义,使推荐指数高的对象在图空间的节点面积最大,且距离图空间中该图节点最近。

[0045] 根据本申请的一个实施例,所述步骤S5还包括:

[0046] S52:采用不同的颜色对不同对象进行区分和可视化。

[0047] 本申请通过根据推荐指数和对象种类,进行不同大小和颜色的图模型展示及可视化,可实现推荐结果的直观、有效、合理显示,提升用户体验。

[0048] 最后说明的是,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或者等同替换,而不脱离本发明技术方案的宗旨和范围,其均应涵盖在本发明的权利要求范围当中。

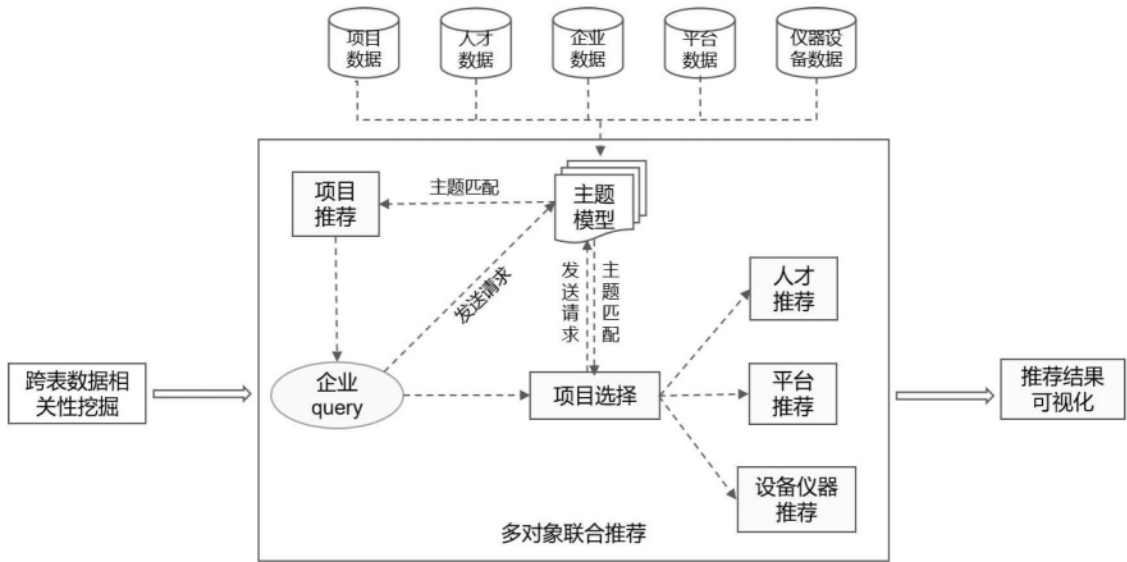


图1

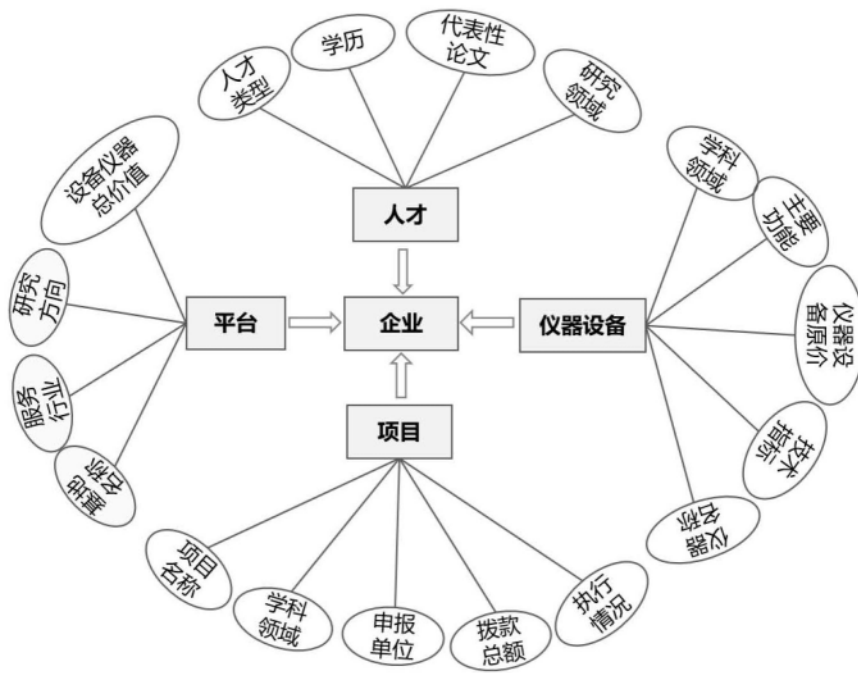


图2



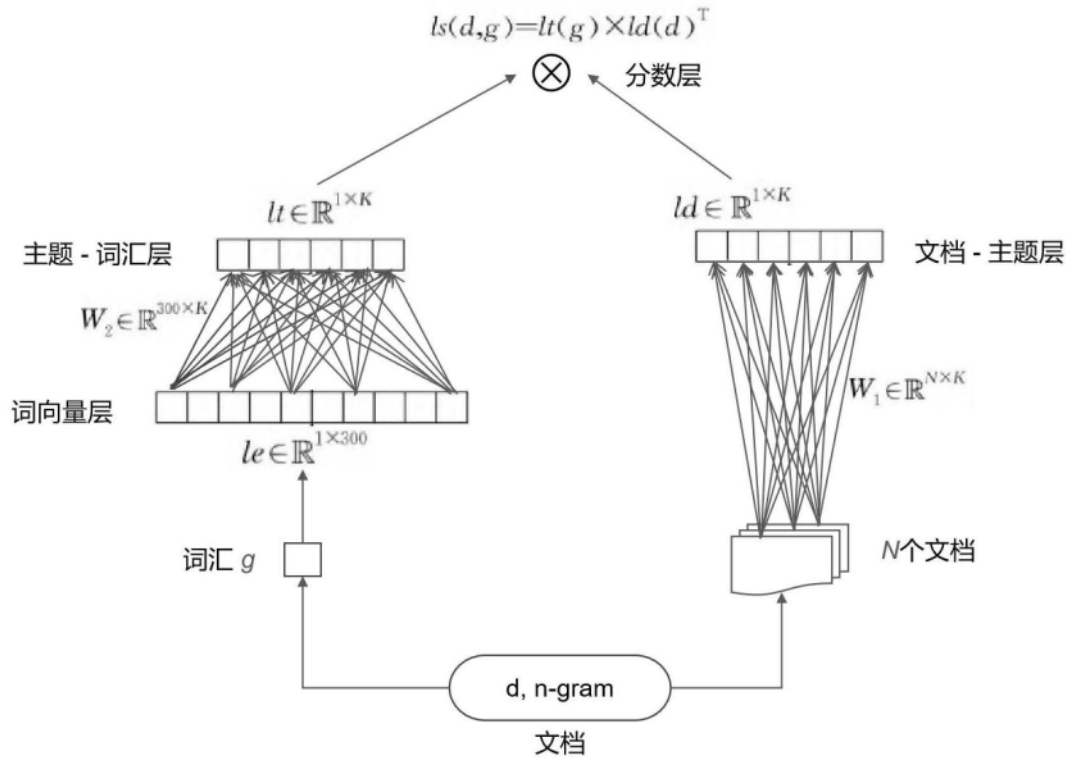


图3