(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107391577 A
(43)申请公布日 2017.11.24

(21)申请号 201710469315.4

(22)申请日 2017.06.20

(71)申请人 中国科学院计算技术研究所
地址 100080 北京市海淀区中关村科学院南路6号

(72)发明人 贾岩涛 蔡朋杉 王元卓 靳小龙 李曼玲 程学旗

(74)专利代理机构 北京律诚同业知识产权代理有限公司 11006
代理人 祁建国 梁挥

(51)Int.Cl.
G06F 17/30(2006.01)

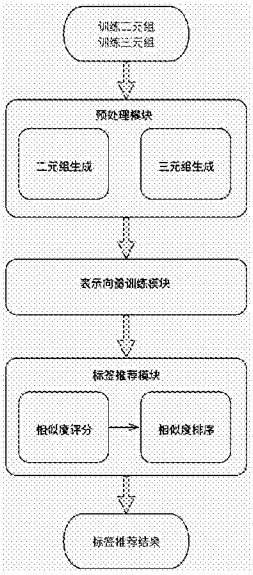
权利要求书2页 说明书12页 附图2页

(54)发明名称

一种基于表示向量的作品标签推荐方法和系统

(57)摘要

本发明涉及一种基于表示向量的标签推荐方法和系统，其特征在于，包括：获取多个作品，根据每个作品对应的标签、关系及对象，构建多个包含作品和标签的二元组信息和包含作品、关系及对象的三元组信息，根据二元组信息和三元组信息生成训练数据集；通过对训练数据集进行表示学习，分别得到各个作品的作品表示向量和各类标签的标签表示向量；通过计算各个作品表示向量和各类标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。本发明在学习表示向量的过程中，本发明同时考虑作品标签对二元组信息和作品的三元组信息。通过融入更多信息，使得学到的表示向量能够更准确地反映作品和标签的语义，从而更好地支持标签推荐这一任务。



CN 107391577 A

1.一种基于表示向量的标签推荐方法，其特征在于，包括：

步骤1，获取多个作品，每个该作品对应多个标签、多个关系及多个对象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及该对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集；

步骤2，通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量；

步骤3，通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。

2.根据权利要求1所述的基于表示向量的标签推荐方法，其特征在于，步骤1中该训练数据集生成的具体过程包括：

步骤S101，该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。

3.根据权利要求1所述的基于表示向量的标签推荐方法，其特征在于，该步骤2还包括：

步骤S102，该作品表示向量和该标签表示向量是通过将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习得到的。

4.根据权利要求3所述的基于表示向量的标签推荐方法，其特征在于，该损失函数为：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\|m-t\|+M_1-\|m'-t'\|\right]_+\right]+$$

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\|m+r-o\|+M_2-\|m'+r'-o'\|\right]_+\right]$$

其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′，或将t替换为其他标签t′得到的；负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′，o替换为其他对象o′，r替换为其他对象r′得到的，$[x]_+$表示返回x与0两者中的最大值；$\|\cdot\|$表示L_1或L_2范式。

5.根据权利要求1所述的基于表示向量的作品标签推荐方法，其特征在于，该步骤3还包括：用户输入距离阈值，根据该距离和该距离阈值，将各类标签中筛选出各个作品的推荐标签。

6.一种基于表示向量的标签推荐系统，其特征在于，包括：

信息获取模块，用于获取多个作品，每个该作品对应多个标签、多个关系及多个对象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及该对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集；

表示学习模块，用于通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量；

标签推荐模块，用于通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。

7.根据权利要求6所述的基于表示向量的标签推荐系统，其特征在于，该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。

8.根据权利要求6所述的基于表示向量的标签推荐系统，其特征在于，该作品表示向量和该标签表示向量是通过将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习得到的。

9.根据权利要求8所述的基于表示向量的标签推荐系统，其特征在于，该损失函数为：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\|m-t\|+M_1-\|m'-t'\|\right]_+\right]+$$

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\|m+r-o\|+M_2-\|m'+r'-o'\|\right]_+\right]$$

其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′，或将t替换为其他标签t′得到的；负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′，o替换为其他对象o′，r替换为其他对象r′得到的，$[x]_+$表示返回x与0两者中的最大值；$\|\cdot\|$表示L_1或L_2范式。

10.根据权利要求6所述的基于表示向量的作品标签推荐系统，其特征在于，该标签推荐模块还包括：用户输入距离阈值，根据该距离和该距离阈值，将各类标签中筛选出各个作品的推荐标签。

一种基于表示向量的作品标签推荐方法和系统

技术领域

[0001] 本发明涉及标签推荐领域，特别涉及一种基于表示向量的作品标签推荐方法和系统。

背景技术

[0002] 网络大数据时代的到来，使得网络上的数据呈爆炸式的增长。为了帮助用户更好，更快地了解网络上各种作品的信息，许多网站开放了分众分类体系(folksonomy)。分众分类体系允许用户给网站中的作品打标签。如在豆瓣电影网站上，电影《音乐之声》就被打上了标签“音乐剧”、“爱情”、“经典”等标签。这些标签可以帮助网站的用户更快速地找到信息，更便捷地发现信息，更容易地理解信息。

[0003] 然而，对于许多刚刚问世的作品，或者关注人数较少的作品而言，因为广大用户缺乏了解，所以难以给出准确的标签。这使得分众分类体系缺乏准确性和完整性，也进一步阻碍了更多用户了解到作品的相关信息。

[0004] 为了解决上述问题，我们需要一种方法，帮助我们利用已知的作品标签信息，发掘出潜在的作品标签，从而完成分众分类体系的自动扩充。

[0005] 现有技术包括：标签推荐技术和协同过滤推荐技术，前者是目前比较有效的方法是基于协同过滤算法的标签推荐方法；后者协同过滤推荐(Collaborative Filtering Recommendation)技术是在信息过滤和信息系统已经一项非常成熟的技术，与传统的基于内容过滤直接分析内容进行推荐不同，协同过滤分析用户兴趣，在用户群中找到指定用户的相似(兴趣)用户，综合这些相似用户对某一信息的评价，形成系统对该指定用户对此信息的喜好程度预测。

[0006] 具体地，在作品标签推荐这一任务中，协同过滤方法又可以细分为两种：

[0007] 1.基于作品的协同过滤：通过作品标签对来评测作品之间的相似性，基于作品之间的相似性，给作品推荐潜在的标签。

[0008] 2.基于标签的协同过滤：通过作品标签对来评测标签之间的相似性，基于标签之间的相似性，给标签推荐给潜在作品。

[0009] 但上述协同过滤技术在以下两点上存在缺陷：

[0010] (1)协同过滤技术仅仅适用于二元组式的数据结构，即其输入信息必须严格地服从“作品——标签”对形式。然而，众所周知的是，当前互联网上，作品的信息以多种形式存储。除了“作品标签”对这一形式之外，还存在丰富的三元组式的数据信息，如“音乐之声——导演——罗伯特·怀斯”这个三元组告诉我们，电影音乐之声的导演是罗伯特·怀斯。

[0011] 由于协同过滤技术无法处理这样的三元组信息，也使其在进行标签推荐的时候少了许多可用的资源。特别是对于某些标签较少的作品，由于缺乏足够多的标签表征其语义，标签推荐的结果往往不够准确。

[0012] (2)协同过滤技术仅仅依据显性的相似度进行推荐。如下表1所示，由于作品A和作

品C有着相似的标签集合(两作品均对应标签A、C)，故推测作品A也可能会对应标签D，而作品B和作品C的标签集合并不相似，故作品B不太可能会对应标签D。这一方法的缺陷在于，我们无从得知标签A,B,C之间的语义相似度。如果标签B的语义和标签A,C非常相似，那么我们同样有理由推测，作品B也对应标签D。然而，这种相似度是非显性的，应用协同过滤方法难以直接获得。

[0013] 表1：

[0014]

	标签A	标签B	标签C	标签D
作品A	√		√	推荐
作品B		√		不推荐
作品C	√		√	√

发明内容

[0015] 为了解决上述技术问题，本发明目的在于提供一种基于向量空间表示的标签推荐方法，该方法建模作品和标签之间关系的相关性，并利用此相关性计算损失函数，通过损失函数来体现系统性能，降低平均序值。

[0016] 具体地说，本发明公开了一种基于表示向量的标签推荐方法，其中包括：

[0017] 步骤1，获取多个作品，每个该作品对应多个标签、多个关系及多个对象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及该对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集；

[0018] 步骤2，通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量；

[0019] 步骤3，通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。

[0020] 该基于表示向量的标签推荐方法，其中步骤1中该训练数据集生成的具体过程包括：

[0021] 步骤S101，该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。

[0022] 该基于表示向量的标签推荐方法，其中该步骤2还包括：

[0023] 步骤S102，该作品表示向量和该标签表示向量是通过将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习得到的。

[0024] 该基于表示向量的标签推荐方法，其中该损失函数为：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\|m-t\|+M_1-\|m'-t'\|\right]_+\right]+$$

[0025]

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\|m+r-o\|+M_2-\|m'+r'-o'\|\right]_+\right]$$

[0026] 其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′，或将t替换为其他标签t′得到的；负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′，o替换为其他对象o′，r替换为其他对象r′得到的，$[x]_+$表示返回x与0两者中的最大值；$\|\cdot\|$表示L_1或L_2范式。

[0027] 该基于表示向量的作品标签推荐方法，其中该步骤3还包括：用户输入距离阈值，根据该距离和该距离阈值，将各类标签中筛选出各个作品的推荐标签。

[0028] 本发明还提供了一种基于表示向量的标签推荐系统，其中包括：

[0029] 信息获取模块，用于获取多个作品，每个该作品对应多个标签、多个关系及多个对象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及该对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集；

[0030] 表示学习模块，用于通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量；

[0031] 标签推荐模块，用于通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。

[0032] 该基于表示向量的标签推荐系统，其中该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。

[0033] 该基于表示向量的标签推荐系统，其中该作品表示向量和该标签表示向量是通过将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习得到的。

[0034] 该基于表示向量的标签推荐系统，其中该损失函数为：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\|m-t\|+M_1-\|m'-t'\|\right]_+\right]+$$

[0035]

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\|m+r-o\|+M_2-\|m'+r'-o'\|\right]_+\right]$$

[0036] 其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组

信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′,或将t替换为其他标签t′得到的;负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′,o替换为其他对象o′,r替换为其他对象r′得到的,$[x]_+$表示返回x与0两者中的最大值;$||\cdot||$表示L_1或L_2范式。

[0037] 该基于表示向量的作品标签推荐系统,其中该标签推荐模块还包括:用户输入距离阈值,根据该距离和该距离阈值,将各类标签中筛选出各个作品的推荐标签。

[0038] 综上,本发明通过考虑作品标签对二元组信息和作品的三元组信息,融入更多信息,解决了协同过滤方法无法融入异构信息的问题,并且本发明通过所学得的表示向量,不仅能够反映作品与作品之间,标签与标签之间显性的相似度,更能够反映出作品与作品之间,标签与标签之间非显性的相似度,使得学到的表示向量能够更准确地反映作品和标签的语义,从而更好地支持标签推荐这一任务。

附图说明

[0039] 图1为本发明系统模块示意图;

[0040] 图2为本发明方法流程示意图。

具体实施方式

[0041] 首先要声明的是,在本案说明书全文(包括权利要求书)中所使用的关键术语的具体定义如下:

[0042] 表示向量:分别用同一个空间中两组向量表示“作品”和“标签”。表示向量的两大特征分别是,低维——向量的维度不超过300维;稠密——对每个向量而言,99%维度上都是非零元素(几乎100%,但不能排除某些维度取值为0的可能性,主要是为了和高维稀疏向量表示产生对比,高维稀疏向量中,一般超过99%的维度上都是0);

[0043] 作品:指文学作品、影视作品等艺术作品,包含电影,音乐等具体作品,如十面埋伏、难忘今宵等,作品也指例如网购网站上的商品;

[0044] 标签:指用来描述某个作品特征的标签,如作品“音乐之声”具有标签“音乐剧”、“爱情”、“经典”等标签;

[0045] 本发明中每个三元组的组成包括作品的名称(主语)、关系(谓语),对象(宾语);

[0046] 本发明提出了一种基于表示学习的方法完成作品标签推荐的任务,输入为作品标签对的二元组信息和作品的三元组信息,对于每个作品,输出其可能对应的标签。

[0047] 为让本发明的上述特征和效果能阐述的更明确易懂,下文特举实施例,并配合说明书附图作详细说明如下。

[0048] 该方法具体包括以下几个步骤:第一步,根据输入的作品标签对二元组信息和作品三元组信息,生成正例集合和负例集合;第二步,利用第一步生成的正例集合和负例集合进行表示学习,得到作品和标签所对应的表示向量;第三步,对于每个作品,依次计算其对应的表示向量和所有标签的表示向量之间的距离,并按照距离从小到大的顺序将标签排序,排名越靠前的标签越可能是和作品相关的标签,具体实施时可通过预先设定一个距离阈值,若计算出的距离小于该距离阈值,则认为相应标签为该作品的相关标签,或者通过预先设定一排名阈值z,即取标签排序中前z名所对应的标签作为该作品的相关标签,其中z为

正整数。

[0049] 其中，针对目前协同过滤技术中的第一个缺陷，本发明第一步预处理产生了作品标签对二元组向量和作品三元组向量，第二步使用了新的目标函数，需要注意的是，旧的目标函数指的是仅利用作品-标签二元组信息进行表示学习的目标函数，新的目标函数是利用作品-标签二元组和作品-关系-对象三元组信息进行表示学习的目标函数，将第一步生成的多源异构信息(作品标签对二元组向量和作品三元组向量)融入表示向量的训练。本发明根据标签推荐任务的特点，在标签推荐过程中融入更加丰富的信息，使得表示向量能够更准确地反映作品和标签的语义。

[0050] 针对目前协同过滤技术中的第二个缺陷，本发明在第三步中采用表示向量进行标签推荐。通过表示学习所得到的表示向量能够反映出整张图的更能够发掘出作品之间非显性相似度，从而提升了推荐的准确率。

[0051] 下举实施例对应上述步骤，以进一步阐述本发明实施细节。本发明适用于作品标签推荐的场景，即给定作品标签的二元组信息和作品的三元组信息，推测出作品可能相关的其他标签。例如，已知电影A对应标签a，标签b等一系列标签，自动推测出电影A可能相关的标签x等一个或多个标签。

[0052] 如图1所示，本发明一共包括三个步骤：

[0053] 步骤1、获取多个作品，每个该作品对应多个标签、多个关系及对象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集，其中该作品包括音像作品、画作和商品。具体包括，根据输入的二元组和三元组信息，生成训练数据集，包括两个子模块：(a)二元组生成模块，正例二元组集合即训练数据中的二元组集合θ_1，通过调换任意正例二元组中的标签构建负例二元组集合θ_2，使得$\theta_1 \cap \theta_2 = \Phi$，其中$\Phi$的代表含义为空集；(b)三元组生成模块，正例三元组集合即训练数据中的二元组集合θ_3，通过随机调换任意正例三元组中的头作品或尾作品构建负例三元组集合θ_4，使得$\theta_3 \cap \theta_4 = \Phi$，其中$\Phi$的代表含义为空集。

[0054] 步骤2、通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量。具体包括，基于从步骤1获得的训练数据，学习并生成作品、标签的向量表示。通过随机梯度下降函数，对目标函数进行优化，通过多次迭代学习得到作品和关系的向量表示，即根据训练数据优化目标函数，根据优化后的目标函数得到的作品和标签的向量表示。

[0055] 步骤3、通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。具体包括，根据学得的表示向量，给作品推荐其可能相关的标签。该步骤3包括两个子步骤：(a)相似度评分模块，对于任意给定作品，依次评判其与每类标签的相似度，即计算该作品与各类标签间的距离；(b)相似度排序模块，根据相似度评分将向量排序，向作品推荐相似度评分最高的几个标签。

[0056] 本实施例以电影标签推荐为例，介绍相关的作品标签推荐技术方案。例如，已知电影A，电影A相关的标签$a_1, a_2, \ldots, a_n$，电影A相关的三元组信息$t_1, t_2, \ldots, t_n$需要发掘出和电影A相关的其他标签$b_1, b_2, \ldots, b_n$。其中任意三元组$t_k \in \{t_1, t_2, \ldots, t_n\}$以<电影A，关系，对象>的形式存储，关系可以是“导演”，“编剧”，“类型”等任意关系，“对象”与关系相对应，如“张艺谋”，“王朔”，“情感”等。

[0057] 训练数据包括两个部分，其一是电影及其已知相关标签的二元组信息(train.txt)，其二是电影的三元组信息(triple.txt)。文件均由行组成，每个数据文件中行的格式说明如下：

[0058] 1)train.txt行数据包含两列，第一列数据为电影名称；第二列为电影相关的标签。其中，列和列之间用Tab隔开。如：

[0059]

```
伊万王子和灰狼	新问题
顺藤而上的你	好男人
顺藤而上的你	世故人情
顺藤而上的你	大结局
暴力罗曼史	清澈的眼神
暴力罗曼史	满脑子
暴力罗曼史	大结局
七个隆咚锵咚锵	小电影
七个隆咚锵咚锵	好多情节
七个隆咚锵咚锵	小城市
七个隆咚锵咚锵	大银幕
无憾	真实的故事
宇宙兄弟	一般的人
宇宙兄弟	新观众
魔幻手机Ⅱ：傻妞归来	高科技
魔幻手机Ⅱ：傻妞归来	真的大丈夫
疯狂的蠢贼	老男孩
疯狂的蠢贼	好的剧本
疯狂的蠢贼	复杂的人物关系
```

[0060] 2)triple.txt：行数据包含三列，第一列电影名称；第二列为关系(dir：导演，sw：编剧，type：电影类型)，第三列为对象名称。其中，列和列之间用Tab隔开。如：

[0061]

喜盈门	dir	赵焕章
喜盈门	sw	辛显令
喜盈门	type	剧情
喜盈门	type	家庭
多谢款待	dir	木村隆文
多谢款待	dir	小林大児

[0062]

多谢款待	dir	福岡利武
多谢款待	dir	盆子原誠
多谢款待	sw	森下佳子
多谢款待	type	剧情
大蓝湖	dir	曾翠珊
大蓝湖	sw	曾翠珊
大蓝湖	type	剧情
醉生梦死之湾仔之虎	dir	罗杰承
醉生梦死之湾仔之虎	sw	罗诚
醉生梦死之湾仔之虎	sw	伍立光
醉生梦死之湾仔之虎	sw	梁恩东
醉生梦死之湾仔之虎	type	动作

[0063] 该基于表示向量作品标签推荐方案包括以下步骤：

[0064] 步骤1、根据已知和电影相关的标签信息及电影三元组信息，构建训练数据集。训练数据集分为两个部分，1)二元组训练数据，又可进一步分为正例二元组集合θ_1，负例二元组集合θ_2。(2)三元组训练数据，又可进一步分为正例三元组集合θ_3，负例三元组集合θ_4，主要注意的是本实施例以电影为作品，但不以此为限，作品可为艺术作品、影视作品、戏曲作品等。

[0065] 步骤2、作品和关系的向量空间表示学习。初始化电影，标签，关系，对象及其向量表示，需要注意的是本发明利用作品-关系-对象的三元组信息去学习得出更准确的作品、标签的表示向量，从而更好地完成标签推荐的任务，比如，如果知道了“夏洛特烦恼-主演-沈腾”这个三元组信息，会给“夏洛特烦恼”这个作品推送“喜剧”这个标签，因为在大量的训练数据中，其他沈腾主演的电影也被打上了“喜剧”这个标签，因此，在整个表示学习过程中，电影，标签，关系，对象的表示向量都是需要学习的，只是最后能用上的只有电影和标签的表示向量而已。具体为利用随机梯度下降法自适应地学习并得到作品和标签的向量表示，标签推荐的任务是：发现和作品潜在相关的标签。比如，知道电影“夏洛特烦恼”，也知道“喜剧”这个标签，但不知道二者是否相关，于是就通过对“夏洛特烦恼”和“喜剧”这两个标

签进行计算，去判断二者的相关程度，所以，此处的表示向量，是所有作品和标签的表示向量。而反观现有技术中的做法仅利用电影和其已知对应标签的信息为电影推荐新标签，这使得训练数据中已知相关标签较少的电影往往不能通过训练获得较为准确的表示向量，从而影响标签推荐的准确性。

[0066] 步骤3、为某个未知标签(待推荐标签)的电影作品推荐可能相关的标签。对某个给定电影作品，根据向量空间(表示学习将每个作品，标签，关系和对象表示为一组低维空间中的向量，这里的向量空间，就是这个低维空间)中作品和各个标签之间的距离，构建打分函数，计算给定作品和所有标签打分函数分值，取分值排名靠前的标签作为发掘出和作品相关的标签。已有的作品标签推荐方法没有基于表示向量的。

[0067] 其中，参考图2为所述作品和标签的向量空间表示学习流程。具体地，该方法包括：制造二元组和三元组训练数据集(步骤S101)；计算并优化目标函数，最终得到电影和标签在向量空间的表示向量(步骤S102)；根据电影和标签在向量空间的表示向量完成电影标签推荐的任务(步骤S103)。

[0068] 下面对某些步骤进行详细描述。

[0069] 步骤S101：该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。具体包括构造训练数据。训练数据集分为两个部分，1)在二元组训练数据中，正例二元组集合即训练数据中的二元组集合θ_1，通过调换任意正例二元组中的标签(随机调换，比如将“夏洛特烦恼-搞笑”这个二元组中的标签调换为“悲剧”，即构建了“夏洛特烦恼-悲剧”这个二元组)，构建负例二元组集合θ_2，使得$\theta_1 \cap \theta_2 = \Phi$；(2)在三元组训练数据中，正例三元组集合即训练数据中的正例三元组集合θ_3，通过随机调换任意正例三元组中的对象构建负例三元组集合θ_4，使得$\theta_3 \cap \theta_4 = \Phi$。

[0070] 步骤S102：将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习，以得到该作品表示向量和该标签表示向量。具体包括将作品，标签，关系和对象初始化成给定维度的向量，可以采用平均分布初始化、伯努利分布初始化等。其中该给定维度是人为设定的，一般初始化为100-300维。理论上来说，维度决定了模型的表达能力，即维度越高，模型能承载的信息量越大。然而，受到训练数据量的影响，过于复杂的模型，即维度过高的模型，可能会在训练过程中欠拟合，即训练数据量过小，模型没有收敛，效果不升反降。直观的理解，就像人有时候把一个简单的问题想得太复杂，反而做不好，因此，具体设为多少维，需要结合具体的任务进行多次尝试，并保留最优结果的参数。

[0071] 利用随机梯度下降法计算目标函数(损失函数)。根据已初始化的作品、标签，关系和对象建模表示知识图谱。知识图谱：由作品及作品间关系构成的网络，其中，网络中的点是作品，网络中的边是作品间的关联关系。并基于假设，作品和其相关的标签在向量空间中彼此临近，即作品所对应的向量m和相关标签所对应的向量t在向量空间中满足$m \approx t$；此外，作品所对应的向量m，关系所对应的向量r和对象所对应的向量o在向量空间中满足$o+r \approx t$，损失函数使得正例作品关系对尽量满足此假设，负例作品关系对尽量不满足此假设，损失函数如下：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\left\|m-t\right\|+M_1-\left\|m'-t'\right\|\right]_+\right]+$$

[0072]

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\left\|m+r-o\right\|+M_2-\left\|m'+r'-o'\right\|\right]_+\right]$$

[0073] 其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′，或将t替换为其他标签t′得到的；负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′，o替换为其他对象o′，r替换为其他对象r′得到的，$[x]_+$表示返回x与0两者中的最大值；$||\cdot||$表示L1或L_2范式。需要注意的是，该损失函数在本发明中即为目标函数。

[0074] 在梯度下降更新时，向量更新方式如下：

[0075] $\forall i\in\{0,1,2,\ldots\dim\}$；

[0076] $m_i=m_i-\mu_1*2*|t_i-m_i|-\mu_2*2*|o_i-m_i-r_i|$

[0077] $t_i=t_i+\mu_1*2*|t_i-m_i|$

[0078] $r_i=r_i-\mu_2*2*|o_i-m_i-r_i|$

[0079] $o_i=o_i+\mu_2*2*|o_i-m_i-r_i|$

[0080] $m'_i=m'-\mu_1*2*|t'_i-m'_i|-\mu_2*2*|o'_i-m'_i-r'_i|$

[0081] $t'_i=t'_i+\mu_1*2*|t'_i-m'_i|$

[0082] $r'_i=r'_i-\mu_2*2*|o'_i-m'_i-r'_i|$

[0083] $o'_i=o'_i+\mu_2*2*|o'_i-m'_i-r'_i|$

[0084] 其中，dim是向量空间的维度，m_i代表m的第i维向量。μ_1和μ_2为学习率。

[0085] 步骤S103：用户输入距离阈值，根据该距离和该距离阈值，将各类标签中筛选出各个作品的推荐标签。具体包括根据步骤3中该距离生成一打分函数，根据打分函数对所有作品关系对(m',t)打分，打分函数基于向量表示相似度和视频属性相似度定义：

[0086] $f_{predict}(m,t)=\left\|m-t\right\|_{l_1}$，

[0087] 其中$\left\|m-t\right\|_{l_1}$衡量向量之间的距离。

[0088] 具体地，$\left\|m-t\right\|_{l_1}=\sum_{i=1}^{\dim}\left|m_i-t_i\right|$

[0089] 为验证本发明提供的基于向量空间表示的作品对齐方法，以标签推荐任务为例。发明人获取了部分电影作品标签对和电影作品三元组，并将电影作品标签对分为训练数据和测试数据。发明人采用本发明提供的方法，在真实训练数据集上进行了表示向量的学习。在测试过程中，对于任意一个给定的电影作品，计算在向量空间中和它距离最近的十个标签，并按照距离从小到大将作品标签对依次排序。评价指标为：所有的测试电影作品标签对中，有多少比例被排在前三，前五及前十名(前三命中率，前五命中率，前十命中率)，实验参数如下：

[0090] 数据集采集自豆瓣(www.douban.com)网站上获取了2015-2016年部分电影作品信息。其中电影作品标签对共38398对，其中训练数据共30398对，测试数据共8000对。电影三元组信息共24026条，包括3种关系(导演、编剧、类型等)。

[0091] 学习过程使用的学习率μ＝0.001，向量的维度d＝100，批处理大小B＝100，选用L_2范式衡量相似度。

[0092] 得到标签推荐的前三命中率为18.85％，前五命中率为27.45％，前十命中率为38.625％。

[0093] 表格2数据集信息

[0094]

数据源	训练数据集大小	测试数据集大小
豆瓣（www.douban.com）	30398	8000

[0095] 学习得到的向量表示为100维向量，例如，电影“山河故人”的向量表示为：

[0096]

0.098736 0.148169 -0.006748 -0.033977 -0.147386 0.008067
0.154854 -0.129785 -0.148183 0.113105 0.052244
0.148388 0.040527 0.058668 0.029696 -0.101465
-0.000045 0.051474 0.051823 0.138922 0.044317
0.003595 0.068311 -0.130376 -0.101556 -0.152162
0.129072 0.022441 0.055113 -0.069564 -0.081825
-0.091650 0.083833 -0.073370 0.105618 -0.067246
-0.076510 -0.026746 -0.139388 0.117846 -0.009062
-0.155884 -0.053244 0.159078 -0.170914 -0.060208
0.142106 0.142205 0.139162 0.137175 -0.017455
-0.056536 0.120213 0.145928 -0.146345 -0.054259
0.125724 -0.043677 -0.128898 0.129607 0.156227
-0.132797 -0.088120 0.087432 -0.138939 0.109464
-0.054793 -0.106385 -0.077519 0.087118 0.058175
0.007658 -0.025107 -0.001242 -0.159152 -0.056204
-0.021440 0.059051 -0.059941 -0.115516 -0.062110
-0.088237 0.032899 0.056564 -0.123168 0.040342
0.144112 -0.111303 -0.024808 0.162416 -0.122204
0.150433 0.159387 -0.037686 0.066868 -0.100892
-0.093795 0.037118 0.061321 -0.064903

[0097] 标签“文艺”的向量表示为：

[0098]

0.036184	-0.009332	-0.132870	-0.096886	-0.018262	-0.054635
-0.022058	-0.110106	0.138795	-0.122196	-0.133257	
-0.138437	-0.104106	-0.140142	0.049812	-0.057475	
0.040610	-0.073355	0.127355	0.131401	-0.109707	
0.161415	0.047945	-0.047072	-0.145430	-0.065574	
0.015665	0.046578	0.040839	-0.139704	-0.120163	
-0.132189	-0.134460	-0.006240	0.147563	-0.096106	
-0.040107	0.123002	0.074933	0.124272	-0.007448	
-0.135036	-0.056038	0.074840	-0.067716	-0.032693	
-0.140884	-0.138045	0.116983	-0.088303	0.108735	
0.135465	-0.052583	0.083590	0.032579	0.157741	
0.109106	0.093032	0.092659	-0.119091	-0.059389	
-0.048071	0.026506	-0.093365	-0.095610	-0.088956	
0.088453	-0.047203	0.078419	0.039758	0.128609	
-0.122362	-0.037266	0.024148	0.117004	-0.158162	
0.106510	-0.004874	-0.160402	-0.129818	0.141589	
0.057890	0.096468	0.144083	0.082553	0.151821	
-0.052995	0.134880	-0.076478	-0.130280	0.119114	
-0.157158	-0.056668	-0.024751	0.049683	-0.128142	
0.012738	-0.136701	0.016516	0.104242		

[0099] 给定电影“山河故人”，根据打分函数计算出标签“文艺”的排位为：第二位。

[0100] 在第二步表示向量的训练过程中，该发明使用了表示学习的方法，无须经过人工进行复杂的特征选取，就可以将作品和标签的相关度信息用两组表示向量表示出来。

[0101] 此外，该发明在表示向量的训练过程中，同时使用了作品标签对二元组信息和作品三元组信息。将两种信息同时融入目标函数，使得训练处的表示向量能够更准确地反映作品和标签的语义信息，从而更好地支持标签推荐任务。

[0102] 本发明拟出了基于表示学习的标签推荐方法，使得大规模分众分类体系(Folksonomy)的构建不需要投入大量人力。同时标签推荐的效果能达到实际应用的需求。

[0103] 以下为与上述方法实施例对应的系统实施例，本实施系统可与上述实施方式互相配合实施。上述施方式中提到的相关技术细节在本实施系统中依然有效，为了减少重复，这里不再赘述。相应地，本实施系统中提到的相关技术细节也可应用在上述实施方式中。

[0104] 本发明还提供了一种基于表示向量的标签推荐系统，其中包括：

[0105] 信息获取模块，用于获取多个作品，每个该作品对应多个标签、多个关系及多个对

象，构建多个包含该作品和该标签的二元组信息，同时构建包含该作品、该关系及该对象的三元组信息，根据该二元组信息和该三元组信息生成训练数据集；

[0106] 表示学习模块，用于通过对该训练数据集进行表示学习，得到各个作品和各类标签的表示向量，分别称为作品表示向量和标签表示向量；

[0107] 标签推荐模块，用于通过计算各个该作品表示向量和各类该标签表示向量之间的距离，从各类标签中筛选出各个作品的推荐标签。

[0108] 该基于表示向量的标签推荐系统，其中该训练数据集包括：该二元组信息、该三元组信息、负例二元组信息和负例三元组信息；其中通过将该二元组信息中的标签调换为除该标签外的任意该多个标签中之一，构建该负例二元组信息，通过将该三元组信息中的关系调换为除该关系外的任意该多个关系中之一，构建负例三元组信息。

[0109] 该基于表示向量的标签推荐系统，其中该作品表示向量和该标签表示向量是通过将该训练数据集中作品、标签、关系及对象分别初始化为给定维度的向量，并将该向量输入损失函数，通过随机梯度下降法对该损失函数进行迭代表示学习得到的。

[0110] 该基于表示向量的标签推荐系统，其中该损失函数为：

$$L=\sum_{(m,t)\in\theta_1}\left[\sum_{(m',t')\in\theta_2}\left[\left\|m-t\right\|+M_1-\left\|m'-t'\right\|\right]_+\right]+$$

[0111]

$$\sum_{(m,r,o)\in\theta_3}\left[\sum_{(m',r',o')\in\theta_4}\left[\left\|m+r-o\right\|+M_2-\left\|m'+r'-o'\right\|\right]_+\right]$$

[0112] 其中，L为该损失函数，m、t、r、o依次对应作品、标签、关系及对象的向量，θ_1为该二元组信息，θ_2为该负例二元组信息，θ_3为该三元组信息，θ_4为该负例三元组信息，负例二元组信息$(m',t')\in\theta_2$是将$(m,t)\in\theta_1$中的m替换为其他作品m′，或将t替换为其他标签t′得到的；负例三元组信息$(m',r',o')\in\theta_4$是将$(m,r,o)\in\theta_3$中的m替换为其他作品m′，o替换为其他对象o′，r替换为其他对象r′得到的，$[x]_+$表示返回x与0两者中的最大值；$||\cdot||$表示L1或L_2范式。

[0113] 该基于表示向量的作品标签推荐系统，其中该标签推荐模块还包括：用户输入距离阈值，根据该距离和该距离阈值，将各类标签中筛选出各个作品的推荐标签。

[0114] 虽然本发明以上述实施例公开，但具体实施例仅用以解释本发明，并不用于限定本发明，任何本技术领域技术人员，在不脱离本发明的构思和范围内，可作一些的变更和完善，故本发明的权利保护范围以权利要求书为准。

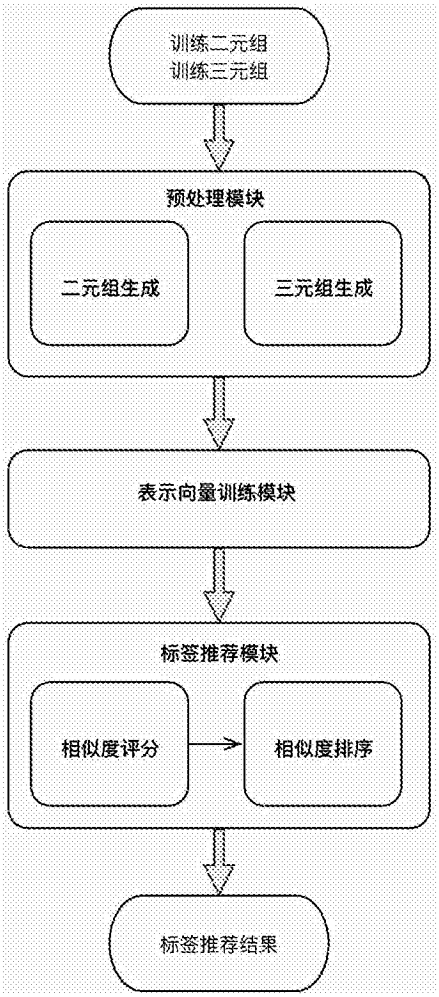

训练二元组
训练三元组
预处理模块
二元组生成
三元组生成
表示向量训练模块
标签推荐模块
相似度评分
相似度排序
标签推荐结果

图1

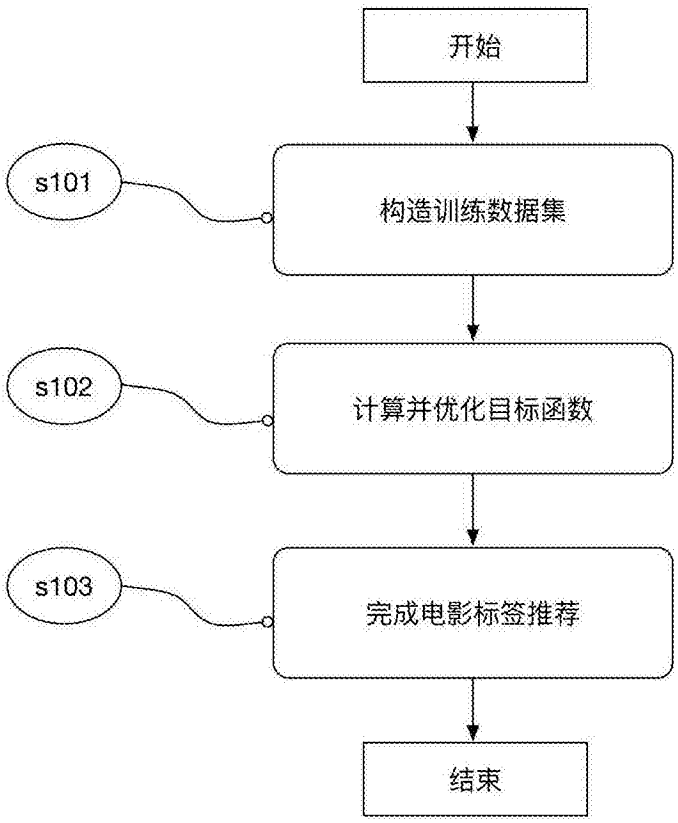


图2