发明名称：一种基于用户评论的推荐方法

申请人：南京邮电大学

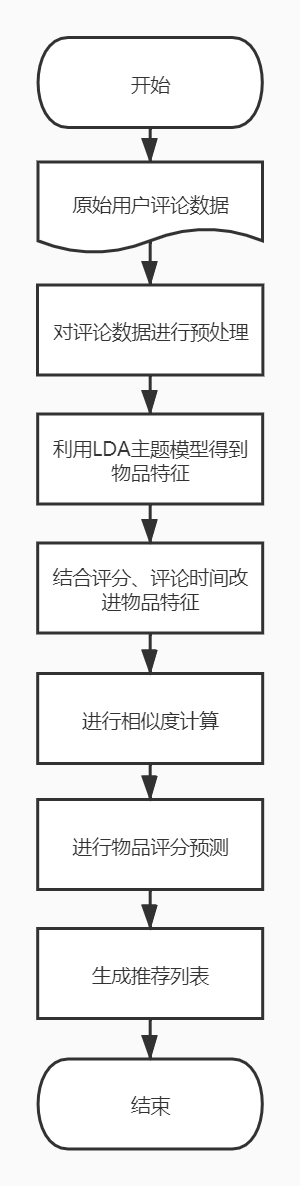
发明人：倪晓军、陈斌

第一发明人身份证号：320106196909012472

联系人：陈斌

电话：13851608260   
Email:1003931661@qq.com

本发明公开一种基于用户评论的推荐方法，针对传统基于物品的协同过滤算法矩阵稀疏的问题，该方法通过用户评论集，利用主题概率模型对物品的评论集进行主题概率分布的计算，再对每条评论加入用户评分以及评论时间权重，来提高物品特征计算的准确性，将得到的物品关于主题分布的概率作为物品的特征，利用物品特征计算物品相似度，通过用户的历史评分数据来对未评分物品进行评分预测，另外考虑到没有获得评论的物品将无法计算其特征向量，提出通过加入物品的描述信息作为一条物品的用户评论，以提高物品特征计算的准确性和全面性。



1.一种基于用户评论的推荐方法，其特征在于，包括以下步骤：

步骤1、对评论数据进行预处理，包括将所有单词转换为小写、去除标点符号等字符，对语句进行分词、断句以及词形还原等处理，去除停用词，过滤无用的空评论，词频统计与向量转换。对于没有获得用户评论的物品，加入物品的描述信息作为一条物品的用户评论。

步骤2、利用主题概率模型对物品的评论集进行主题概率分布的计算，将得到的物品关于主题分布的概率作为物品的特征。

步骤3、结合用户的评分、评论的时间对物品特征进行改进；

步骤4、利用余弦相似度对改进后的物品特征进行计算得到物品之间的相似度；

步骤5、根据计算得到的物品相似度，通过用户已评分的物品，对未评分物品进行评分预测；

步骤6、根据用户对物品的预测评分生成初始的推荐列表，然后过滤用户已评分物品，生成最终的推荐列表。

2.根据权力要求1所述的一种基于用户评论的推荐方法，其特征在于：在步骤2中，所述利用主题概率模型对物品的评论集进行主题概率分布的计算，将得到的物品关于主题分布的概率作为物品的特征。具体过程如下：

步骤2.1、首先通过文档集构建主题模型，然后根据物品的所有评论计算该物品在模型下的主题概率分布，得到一个关于主题分布的概率向量，把这个向量作为物品的特征向量。在对用户评论集建立主题模型前，作如下符号约定：

1. 评论集是个评论的集合，表示成
2. 每一条评论是个词的序列，用表示，是序列的第个词
3. 假设有个主题，主题模型为

步骤2.2、算法是常用的抽样算法，抽样算法可以看作是在已知数据集的情况下，通过参数聚集得到参数值的方法，根据算法可以得到评论中主题的概率。

 （1）

其中表示文档出现主题的次数，是主题的先验分布，是文档中主题的概率。

步骤2.3、通过计算得每个物品在主题上的概率分布，得到对应的矩阵：

 （2）

其中，表示成物品的个数，的每一项都是由公式（1）计算得到，由评论集得到的关于个主题的概率分布，作为物品关于评论集的特征向量。

3.根据权力要求1所述的一种基于用户评论的推荐方法，其特征在于：在步骤3中，所述结合用户的评分、评论的时间对物品特征进行改进，具体过程如下：

步骤3.1、用户对物品的评论为，物品的所有用户评论作为文档集，通过训练好的模型对每一条评论进行转换计算，生成维的主体分布，表示该条用户评论在个主题上的概率分布，从物品的角度来看，表示用户对物品的特征描述，代表物品的属性特征，当集合了所有用户对该物品的评论数据，则可以计算出物品的特征，其计算公式如下：

 （3）

其中表示对物品作出评论的所有用户集。对所有评论计算得到的主题概率分布

进行累加，最后除以评论总数求出均值，得出物品从用户评论中得到的特征向量。

步骤3.2、评分是用户对物品评论的一个重要组成部分，评分通常由1至5的自然数来依次表示用户对物品的满意度，通过评分信息可以得知用户对物品的评价。对于评分较高的评论，认为其能更好地刻画用户更感兴趣的物品特征，赋予其较高的权重；对于评分较低的评论，认为描述的物品特征不能很好满足用户的兴趣，赋予其较低的权重。如此一来，突出了更符合用户喜好的物品特征，同时削弱了偏离用户喜好的物品特征，避免了物品特征趋于平均的问题。因此本方法提出将评分作为一个权值融合到物品的特征计算中。

用户对物品的评论为，同时给出的评分为，的取值范围为[1,5]的自然数，根据用户对物品的评分，对评论的主题概率分布加入如下的影响因子：

 （4）

其中表示物品的平均评分，影响因子的取值范围为(0,1)。

步骤3.3、新的评论数据更能反映当前物品的特征，用户的评论时间代表评论的新鲜度，对推荐的效果有一定的影响。基于这种分析，本方法提出结合时间权重，对于目标函数进行改进。

对于一个物品的所有用户评论，根据最晚发布的评论时间以及最早发布的时间，评估每条评论的时间权重。令表示物品所有评论中最早发表的时间，令表示物品所有评论中最晚发表的时间，表示用户对物品的评论时间，则评论的时间权重函数为：

 （5）

其中是计算得到的时间权重值，其大小随着时间而衰减，用户对于物品最早发布的评论，按照公式计算得到的权重为，这样的好处是即使是最早的评论也能对物品的特征有所影响，而不会认为该评论完全不起作用。对于最新发布的评论，其权重值为。

结合时间权重与评分权重，改进的物品特征最终表示为：

 （6）

4.根据权力要求1所述的一种基于用户评论的推荐方法，其特征在于：在步骤4中，所述利用余弦相似度对改进后的物品特征进行计算得到物品之间的相似度，具体如下：

步骤4.1、通过物品的评论集，融合评分的影响因子和时间权重，可以得到改进的物品特征向量，由物品特征向量可以计算物品间的相似度。物品和物品的相似度计算公式如下：

 （7）

其中，表示第个主题，表示物品在第个主题的特征分布，表示物品在第个主题的特征分布。相似度越大表示两者夹角越小，相似度越高。

5．根据权力要求1所述的一种基于用户评论的推荐方法，其特征在于：在步骤5中，所述根据计算得到的物品相似度，通过用户已评分的物品，对未评分物品进行评分预测。具体如下：

用户对物品的评分预测计算公式为：

 （8）

其中，表示物品属于用户已评论的物品集，用户对物品的评分预测需要基于已评分的物品集来进行计算，表示物品和物品之间的相似度。

6. 根据权力要求1所述的一种基于用户评分的推荐方法，其特征在于：在步骤6中，所述生成推荐列表是指首先按照相似度由高到低选取物品的邻近的个物品集合，定义为，通过公式（8）计算目标用户的评分预测，按照预测的评分值进行由高到低排序，过滤目标用户已经评价过的物品，选取评分最高的个物品生成列表进行推荐。

**一种基于用户评论的推荐方法**

**技术领域**

本发明涉及一种基于用户评论的推荐方法，属于推荐系统、自然语言处理、机器学习等交叉技术领域。

**背景技术**

随着信息技术和互联网的发展，信息过载问题日益严重，推荐系统的出现是解决这类问题的有效手段，通过用户的已有的历史记录自动从海量数据中发现满足用户需求的信息并进行推荐，提高了人们对信息获取的效率。

协同过滤算法是推荐算法中最为重要并且得到广泛应用的算法之一，但传统的算法在用户量以及物品量过大的情况下，会出现矩阵稀疏问题，这直接影响着算法计算的精确度以及运行效率。近年来随着文本处理技术的提升，越来越多的研究者们通过从用户评论入手，来获得用户和物品的特征，从而提高算法的准确度。本方法主要通过用户的评论信息，针对传统协同过滤算法的稀疏性和冷启动问题做出了改进。

**发明内容**

**技术问题：**本发明的目的在于缓解传统的协同过滤推荐算法面临数据稀疏性和冷启动问题，通过用户评论集，利用主题概率模型对物品的评论集进行主题概率分布的计算，再对每条评论加入用户评分以及评论时间权重，来提高物品特征计算的准确性，将得到的物品关于主题分布的概率作为物品的特征，通过计算物品之间关于特征向量的余弦相似度来获得物品间的相似程度，从而根据物品相似度，通过用户的历史评分数据来对未评分物品进行评分预测。

**技术方案：**为了实现上述目的，本发明采用以下技术方案：

一种基于用户评论的推荐方法，包括以下步骤：

步骤1、对评论数据进行预处理，包括将所有单词转换为小写、去除标点符号等字符，对语句进行分词、断句以及词形还原等处理，去除停用词，过滤无用的空评论，词频统计与向量转换。对于没有获得用户评论的物品，加入物品的描述信息作为一条物品的用户评论。

步骤2、利用主题概率模型对物品的评论集进行主题概率分布的计算，将得到的物品关于主题分布的概率作为物品的特征。

步骤3、结合用户的评分、评论的时间对物品特征进行改进；

步骤4、利用余弦相似度对改进后的物品特征进行计算得到物品之间的相似度；

步骤5、根据计算得到的物品相似度，通过用户已评分的物品，对未评分物品进行评分预测；

步骤6、根据用户对物品的预测评分生成初始的推荐列表，然后过滤用户已评分物品，生成最终的推荐列表。

在步骤2中，所述利用主题概率模型对物品的评论集进行主题概率分布的计算，将得到的物品关于主题分布的概率作为物品的特征。具体过程如下：

步骤2.1、首先通过文档集构建主题模型，然后根据物品的所有评论计算该物品在模型下的主题概率分布，得到一个关于主题分布的概率向量，把这个向量作为物品的特征向量。在对用户评论集建立主题模型前，作如下符号约定：

(1)评论集是个评论的集合，表示成

(2)每一条评论是个词的序列，用表示，是序列的第个词

(3)假设有个主题，主题模型为

步骤2.2、算法是常用的抽样算法，抽样算法可以看作是在已知数据集的情况下，通过参数聚集得到参数值的方法，根据算法可以得到评论中主题的概率。

 （1）

其中表示文档出现主题的次数，是主题的先验分布，是文档中主题的概率。

步骤2.3、通过计算得每个物品在主题上的概率分布，得到对应的矩阵：

 （2）

其中，表示成物品的个数，的每一项都是由公式（1）计算得到，由评论集得到的关于个主题的概率分布，作为物品关于评论集的特征向量。

在步骤3中，所述结合用户的评分、评论的时间对物品特征进行改进，具体过程如下：

步骤3.1、用户对物品的评论为，物品的所有用户评论作为文档集，通过训练好的模型对每一条评论进行转换计算，生成维的主体分布，表示该条用户评论在个主题上的概率分布，从物品的角度来看，表示用户对物品的特征描述，代表物品的属性特征，当集合了所有用户对该物品的评论数据，则可以计算出物品的特征，其计算公式如下：

 （3）

其中表示对物品作出评论的所有用户集。对所有评论计算得到的主题概率分布

进行累加，最后除以评论总数求出均值，得出物品从用户评论中得到的特征向量。

步骤3.2、评分是用户对物品评论的一个重要组成部分，评分通常由1至5的自然数来依次表示用户对物品的满意度，通过评分信息可以得知用户对物品的评价。对于评分较高的评论，认为其能更好地刻画用户更感兴趣的物品特征，赋予其较高的权重；对于评分较低的评论，认为描述的物品特征不能很好满足用户的兴趣，赋予其较低的权重。如此一来，突出了更符合用户喜好的物品特征，同时削弱了偏离用户喜好的物品特征，避免了物品特征趋于平均的问题。因此本方法提出将评分作为一个权值融合到物品的特征计算中。

用户对物品的评论为，同时给出的评分为，的取值范围为[1,5]的自然数，根据用户对物品的评分，对评论的主题概率分布加入如下的影响因子：

 （4）

其中表示物品的平均评分，影响因子的取值范围为(0,1)。

步骤3.3、新的评论数据更能反映当前物品的特征，用户的评论时间代表评论的新鲜度，对推荐的效果有一定的影响。基于这种分析，本方法提出结合时间权重，对于目标函数进行改进。

对于一个物品的所有用户评论，根据最晚发布的评论时间以及最早发布的时间，评估每条评论的时间权重。令表示物品所有评论中最早发表的时间，令表示物品所有评论中最晚发表的时间，表示用户对物品的评论时间，则评论的时间权重函数为：

 （5）

其中是计算得到的时间权重值，其大小随着时间而衰减，用户对于物品最早发布的评论，按照公式计算得到的权重为，这样的好处是即使是最早的评论也能对物品的特征有所影响，而不会认为该评论完全不起作用。对于最新发布的评论，其权重值为。

结合时间权重与评分权重，改进的物品特征最终表示为：

 （6）

在步骤4中，所述利用余弦相似度对改进后的物品特征进行计算得到物品之间的相似度，具体如下：

步骤4.1、通过物品的评论集，融合评分的影响因子和时间权重，可以得到改进的物品特征向量，由物品特征向量可以计算物品间的相似度。物品和物品的相似度计算公式如下：

 （7）

其中，表示第个主题，表示物品在第个主题的特征分布，表示物品在第个主题的特征分布。相似度越大表示两者夹角越小，相似度越高。

在步骤5中，所述根据计算得到的物品相似度，通过用户已评分的物品，对未评分物品进行评分预测。具体如下：

用户对物品的评分预测计算公式为：

 （8）

其中，表示物品属于用户已评论的物品集，用户对物品的评分预测需要基于已评分的物品集来进行计算，表示物品和物品之间的相似度。

在步骤6中，所述生成推荐列表是指首先按照相似度由高到低选取物品的邻近的个物品集合，定义为， 通过公式（8）计算目标用户的评分预测，按照预测的评分值进行由高到低排序，过滤目标用户已经评价过的物品，选取评分最高的个物品生成列表进行推荐。

**有益效果：**本发明采用以上技术方案与现有技术相比，具有以下技术效果：

（1）结合主题模型与传统协同过滤算法的理论，提出融合用户评论的物品协同过滤算法的实现方法，通过物品的评论数据集提取物品特征，相比评分用户评论可以更准确地反映用户偏好的物品特征，评论的信息量丰富，可用性高。

（2）当评分数据稀疏时，会对协同过滤算法造成极大影响，而通过使用少量评论就可以建立物品特征。

（3）通过物品的元数据信息作为评论集数据解决新增物品的冷启动问题，同时结合用户评分、评论的时间作为评论的权重来提高特征提取的准确度，能够有效地解决传统协同过滤算法的矩阵稀疏问题。

**附图说明**

图1是本发明的方法流程图。

### 具体实施方式

下面结合附图对发明的技术方案做进一步详细的说明：

如图1所示，一种基于用户评论的推荐方法，包括以下步骤：

步骤1、对评论数据进行预处理，包括将所有单词转换为小写、去除标点符号等字符，对语句进行分词、断句以及词形还原等处理，去除停用词，过滤无用的空评论，词频统计与向量转换。对于没有获得用户评论的物品，加入物品的描述信息作为一条物品的用户评论。

步骤2、利用主题概率模型对物品的评论集进行主题概率分布的计算，将得到的物品关于主题分布的概率作为物品的特征。

具体地讲，所述步骤2包括以下步骤：

步骤2.1、首先通过文档集构建主题模型，然后根据物品的所有评论计算该物品在模型下的主题概率分布，得到一个关于主题分布的概率向量，把这个向量作为物品的特征向量。在对用户评论集建立主题模型前，作如下符号约定：

(1)评论集是个评论的集合，表示成

(2)每一条评论是个词的序列，用表示，是序列的第个词

(3)假设有个主题，主题模型为

步骤2.2、算法是常用的抽样算法，抽样算法可以看作是在已知数据集的情况下，通过参数聚集得到参数值的方法，根据算法可以得到评论中主题的概率。

 （1）

其中表示文档出现主题的次数，是主题的先验分布，是文档中主题的概率。

步骤2.3、通过计算得每个物品在主题上的概率分布，得到对应的矩阵：

 （2）

其中，表示成物品的个数，的每一项都是由公式（1）计算得到，由评论集得到的关于个主题的概率分布，作为物品关于评论集的特征向量。

步骤3、结合用户的评分、评论的时间对物品特征进行改进；

具体地讲，所述步骤3包括以下步骤：

步骤3.1、用户对物品的评论为，物品的所有用户评论作为文档集，通过训练好的模型对每一条评论进行转换计算，生成维的主体分布，表示该条用户评论在个主题上的概率分布，从物品的角度来看，表示用户对物品的特征描述，代表物品的属性特征，当集合了所有用户对该物品的评论数据，则可以计算出物品的特征，其计算公式如下：

 （3）

其中表示对物品作出评论的所有用户集。对所有评论计算得到的主题概率分布

进行累加，最后除以评论总数求出均值，得出物品从用户评论中得到的特征向量。

步骤3.2、评分是用户对物品评论的一个重要组成部分，评分通常由1至5的自然数来依次表示用户对物品的满意度，通过评分信息可以得知用户对物品的评价。对于评分较高的评论，认为其能更好地刻画用户更感兴趣的物品特征，赋予其较高的权重；对于评分较低的评论，认为描述的物品特征不能很好满足用户的兴趣，赋予其较低的权重。如此一来，突出了更符合用户喜好的物品特征，同时削弱了偏离用户喜好的物品特征，避免了物品特征趋于平均的问题。因此本方法提出将评分作为一个权值融合到物品的特征计算中。

用户对物品的评论为，同时给出的评分为，的取值范围为[1,5]的自然数，根据用户对物品的评分，对评论的主题概率分布加入如下的影响因子：

 （4）

其中表示物品的平均评分，影响因子的取值范围为(0,1)。

步骤3.3、新的评论数据更能反映当前物品的特征，用户的评论时间代表评论的新鲜度，对推荐的效果有一定的影响。基于这种分析，本方法提出结合时间权重，对于目标函数进行改进。

对于一个物品的所有用户评论，根据最晚发布的评论时间以及最早发布的时间，评估每条评论的时间权重。令表示物品所有评论中最早发表的时间，令表示物品所有评论中最晚发表的时间，表示用户对物品的评论时间，则评论的时间权重函数为：

 （5）

其中是计算得到的时间权重值，其大小随着时间而衰减，用户对于物品最早发布的评论，按照公式计算得到的权重为，这样的好处是即使是最早的评论也能对物品的特征有所影响，而不会认为该评论完全不起作用。对于最新发布的评论，其权重值为。

结合时间权重与评分权重，改进的物品特征最终表示为：

 （6）

步骤4、通过物品的评论集，融合评分的影响因子和时间权重，可以得到改进的物品特征向量，由物品特征向量可以计算物品间的相似度。物品和物品的相似度计算公式如下：

 （7）

其中，表示第个主题，表示物品在第个主题的特征分布，表示物品在第个主题的特征分布。相似度越大表示两者夹角越小，相似度越高。

步骤5、根据计算得到的物品相似度，通过用户已评分的物品，对未评分物品进行评分预测，用户对物品的评分预测计算公式为：

 （8）

其中，表示物品属于用户已评论的物品集，用户对物品的评分预测需要基于已评分的物品集来进行计算，表示物品和物品之间的相似度。

步骤6、首先按照相似度由高到低选取物品的邻近的个物品集合，定义为，通过公式（8）计算目标用户的评分预测，按照预测的评分值进行由高到低排序，过滤目标用户已经评价过的物品，选取评分最高的个物品生成列表进行推荐。

以上所述仅是本发明的优选实施方式，应当指出：对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

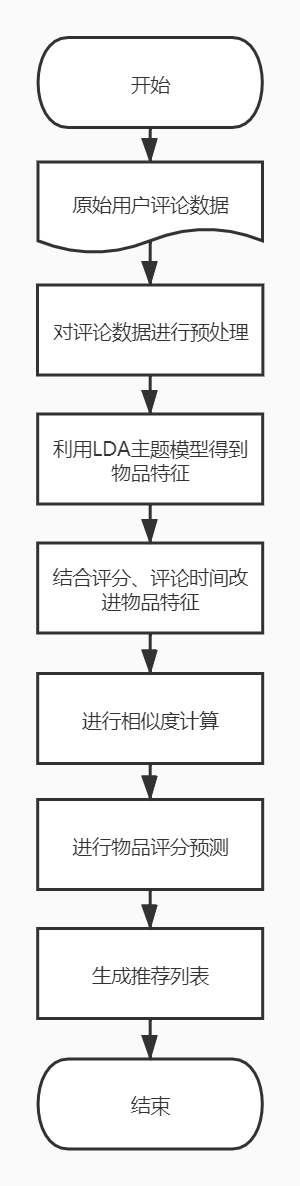


图1