(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 111858915 A
(43)申请公布日 2020.10.30

(21)申请号 202010789845.9

(22)申请日 2020.08.07

(71)申请人 成都理工大学
地址 610059 四川省成都市成华区二仙桥东三路1号

(72)发明人 李冬芬 何菊兰 刘明哲 王惠明 唐小川 王林平 钟豪

(74)专利代理机构 成都金英专利代理事务所(普通合伙) 51218
代理人 袁英

(51)Int.Cl.
G06F 16/34(2019.01)
G06F 16/335(2019.01)
G06F 16/951(2019.01)
G06F 16/9535(2019.01)

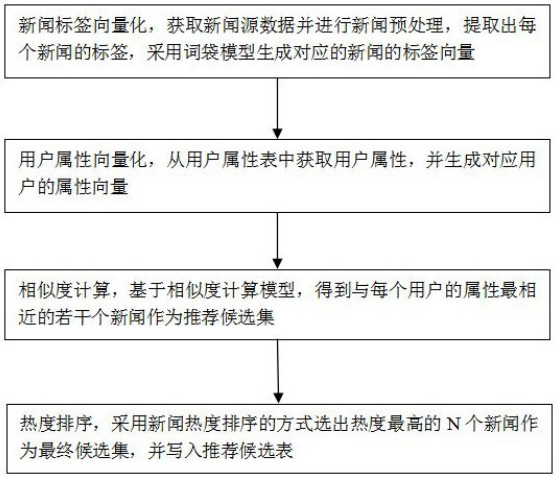
权利要求书1页 说明书4页 附图3页

(54)发明名称

基于标签相似度的信息推荐方法及系统

(57)摘要

本发明公开了基于标签相似度的信息推荐方法及系统，方法包括：新闻标签向量化，获取新闻源数据并进行新闻预处理，提取出每个新闻的标签，采用词袋模型生成对应的新闻的标签向量；用户属性向量化，从用户属性表中获取用户属性，并生成对应用户的属性向量；相似度计算，基于相似度计算模型，得到与每个用户的属性最相近的若干个新闻作为推荐候选集；热度排序，采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集，并写入推荐候选表。本发明还提供基于标签相似度的信息推荐系统，用于实现上述推荐方法。通过本方案能根据用户的属性和行为进行向量化分析处理，将热度最高且符合用户属性的新闻推荐给用户，解决推荐系统冷启动的问题。



CN 111858915 A

1.基于标签相似度的信息推荐方法,其特征在于,包括以下步骤:

S1,新闻标签向量化,获取新闻源数据并进行新闻预处理,提取出每个新闻的标签,采用词袋模型生成对应的新闻的标签向量;

S2,用户属性向量化,从用户属性表中获取用户属性,并生成对应用户的属性向量;

S3,相似度计算,基于相似度计算模型,得到与每个用户的属性最相近的若干个新闻作为推荐候选集;

S4,热度排序,采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集,并写入推荐候选表。

2.根据权利要求1所述的基于标签相似度的信息推荐方法,其特征在于,所述步骤S1中新闻预处理过程具体包括以下子步骤:

S101,利用爬虫机制并发从新闻数据库中爬取出半结构化或纯文本新闻源数据,并进行数据清洗和组织,生成结构化数据;

S102,采用TF-IDF算法对结构化数据进行关键字提取,将提取出的关键词进行重复检测,并作为每篇新闻的标签生成预处理后的新闻数据存入数据库中。

3.根据权利要求1所述的基于标签相似度的信息推荐方法,其特征在于,所述步骤S3中相似度计算过程具体包括:将用户自定义的标签作为该用户的兴趣表征,遍历新闻列表,采用余弦相似度的计算得到用户和新闻标签的相似距离,若相似距离超过设定阈值,则将新闻加入到推荐候选集中,直至所有新闻数据遍历完成。

4.根据权利要求3所述的基于标签相似度的信息推荐方法,其特征在于,所述采用余弦相似度的计算得到用户和新闻标签的相似距离过程具体包括:

S301,定义标签类别,首先对可能出现的新闻标签进行所属类别定义;

S302,添加标签,通过用户自定义标签和文章关键字提取,分别为用户和新闻添加标签;

S303,标签向量化,采用oneHot编码的方式将用户和新闻标签编码为向量形式,将定义的所有标签设定为一个向量中的一位,对于用户或新闻中如果包含某个标签,那么对应向量中的那一位标签置为1,否则置为0;

S304,利用余弦相似度函数计算所有标签向量中每两个标签向量之间夹角的余弦值,值越大相似度越高。

5.基于标签相似度的信息推荐系统,其特征在于,包括

新闻预处理模块,用于从各个新闻源中爬取新闻信息,并对获取的新闻数据进行数据清洗和去重,生成结构化数据;

向量化模块,用于对结构化的新闻数据进行标签向量化,同时对用户属性进行向量化,获得对应的新闻的标签向量和用户的属性向量;

相似度计算模块,用于根据新闻的标签向量和用户的属性向量进行相似度计算,将与每个用户的属性最相近的新闻作为推荐候选集;

热度排序模块,用于根据热度排序算法从推荐候选集中选出热度最高的N个新闻作为最终候选集,并写入推荐候选表;

新闻推荐模块,用于根据推荐候选表中的新闻ID,在新闻数据库中进行匹配,得到对应的新闻内容数据,反馈给用户。

基于标签相似度的信息推荐方法及系统

技术领域

[0001] 本发明涉及文章推荐技术领域，尤其涉及基于标签相似度的信息推荐方法及系统。

背景技术

[0002] 随着信息技术和互联网的发展，人们逐渐从信息匮乏的时代走入了信息过载的时代。与此同时，无论是信息消费者还是信息生产者都遇到了很大的挑战：对于信息消费者，从海量信息中寻找到自己感兴趣的信息是一件非常困难的事；对于信息生产者，让自己生产的信息脱颖而出，受到用户的广泛关注，也是一件十分困难的事情。新闻是信息的重要载体之一，随着互联网的发展，浏览网络上即时发布的新闻成为人们获取信息的重要手段。而新闻推荐系统或新闻推荐装置就是解决这个矛盾的重要工具。它的任务是联系用户和信息，既帮助用户发现对自己有价值的信息，又让信息能够展现在对它感兴趣的用户面前，实现信息消费者和生产者的双赢。

[0003] 当前的信息推荐系统虽然能根据用户的兴趣为用户提供个性化的信息推荐服务，但是不能很好的解决推荐系统冷启动的问题。

发明内容

[0004] 本发明的目的在于克服现有技术的不足，提供基于标签相似度的信息推荐方法及系统，能根据用户的属性和行为进行向量化分析处理，将热度最高且符合用户属性的新闻推荐给用户，解决了推荐系统冷启动的问题。

[0005] 本发明的目的是通过以下技术方案来实现的：

基于标签相似度的信息推荐方法，包括以下步骤：

S1，新闻标签向量化，获取新闻源数据并进行新闻预处理，提取出每个新闻的标签，采用词袋模型生成对应的新闻的标签向量；

S2，用户属性向量化，从用户属性表中获取用户属性，并生成对应用户的属性向量；

S3，相似度计算，基于相似度计算模型，得到与每个用户的属性最相近的若干个新闻作为推荐候选集；

S4，热度排序，采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集，并写入推荐候选表。

[0006] 具体的，所述步骤S1中新闻预处理过程具体包括以下子步骤：

S101，利用爬虫机制并发从新闻数据库中爬取出半结构化或纯文本新闻源数据，并进行数据清洗和组织，生成结构化数据；

S102，采用TF-IDF算法对结构化数据进行关键字提取，将提取出的关键词进行重复检测，并作为每篇新闻的标签生成预处理后的新闻数据存入数据库中。

[0007] 具体的，所述步骤S3中相似度计算过程具体包括：将用户自定义的标签作为该用户的兴趣表征，遍历新闻列表，采用余弦相似度的计算得到用户和新闻标签的相似距离，若

相似距离超过设定阈值，则将新闻加入到推荐候选集中，直至所有新闻数据遍历完成。

[0008] 具体的，所述采用余弦相似度的计算得到用户和新闻标签的相似距离过程具体包括：

S301，定义标签类别，首先对可能出现的新闻标签进行所属类别定义；

S302，添加标签，通过用户自定义标签和文章关键字提取，分别为用户和新闻添加标签；

S303，标签向量化，采用oneHot编码的方式将用户和新闻标签编码为向量形式，将定义的所有标签设定为一个向量中的一位，对于用户或新闻中如果包含某个标签，那么对应向量中的那一位标签置为1，否则置为0；

S304，利用余弦相似度函数计算所有标签向量中每两个标签向量之间夹角的余弦值，值越大相似度越高。

[0009] 基于标签相似度的信息推荐系统，包括

新闻预处理模块，用于从各个新闻源中爬取新闻信息，并对获取的新闻数据进行数据清洗和去重，生成结构化数据；

向量化模块，用于对结构化的新闻数据进行标签向量化，同时对用户属性进行向量化，获得对应的新闻的标签向量和用户的属性向量；

相似度计算模块，用于根据新闻的标签向量和用户的属性向量进行相似度计算，将与每个用户的属性最相近的新闻作为推荐候选集；

热度排序模块，用于根据热度排序算法从推荐候选集中选出热度最高的N个新闻作为最终候选集，并写入推荐候选表；

新闻推荐模块，用于根据推荐候选表中的新闻ID，在新闻数据库中进行匹配，得到对应的新闻内容数据，反馈给用户。

[0010] 本发明的有益效果：本方案能在用户初始访问时提供标签选择界面，根据用户的选择结果调用基于标签推荐的算法生成输出推荐结果，解决了现有新闻推荐系统的冷启动问题。

附图说明

[0011] 图1是本发明的方法流程图。

[0012] 图2是本发明的标签推荐功能数据图。

[0013] 图3是本发明的新闻预处理流程图。

[0014] 图4是本发明的标签推荐生成流程图。

[0015] 图5是本发明的系统功能模块图。

具体实施方式

[0016] 为了对本发明的技术特征、目的和效果有更加清楚的理解，现对照附图说明本发明的具体实施方式。

[0017] 本实施例中，如图1所示，基于标签相似度的信息推荐方法，主要包括以下步骤：

步骤1，新闻标签向量化，获取新闻源数据并进行新闻预处理，提取出每个新闻的标签，采用词袋模型生成对应的新闻的标签向量；

步骤2,用户属性向量化,从用户属性表中获取用户属性,并生成对应用户的属性向量;

步骤3,相似度计算,基于相似度计算模型,得到与每个用户的属性最相近的若干个新闻作为推荐候选集;

步骤4,热度排序,采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集,并写入推荐候选表。

[0018] 如图2所示,基于标签推荐功能主要包括新闻标签向量化、用户属性向量化、相似度计算和热度排序四个子功能。这里采用基于标签的推荐是为了解决冷启动问题,为初始用户提供基于初始化标签的推荐。从数据库中获取新闻数据,并提取出每个新闻的标签,采用词袋模型生成对应新闻的标签向量;同样从用户属性表(在用户体验功能部分生成)中获取用户属性,并生成对应用户的属性向量;然后经过相似度计算模型,得到与每个用户的属性最相近的若干个新闻作为推荐候选集;这里为了避免生成过多推荐候选集,采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集,并写入推荐候选表。

[0019] 其中,在进行新闻标签向量化的操作之前,还需对新闻数据进行预处理,新闻预处理过程主要利用爬虫获取实时新闻和文本数据预处理,使用的核心算法包含文本处理算法(数据清洗、关键字提取和重复检测),如图3所示,具体预处理过程如下:首先客户端输入想要获取的新闻源名称和url作为新闻源信息输入。系统需要判断数据库中是否有该新闻源的新闻输入,若不存在则为该新闻源创建一张表;若存在则调用爬虫机制进行新闻爬取。爬取后的新闻数据经过数据清洗、关键字提取和重复检测等机制,最终被存储到系统数据库中,作为系统的数据来源。

[0020] 爬虫采用scrapy框架进行爬取。在爬取新闻的过程中,可能会遇到IP检测、代理重复访问、cookies封锁和ajax异步传输等问题,使得对于某些新闻网站无法正常获取数据。具体的解决方案包括:

针对IP检测问题,首先爬取免费代理网站的IP,利用状态码返回值来检测IP的可用性,最终生成一个可用代理IP列表。

[0021] 针对代理重复访问的问题,采用随机User-Agent的方式,利用fake_useragent库,伪装请求头。

[0022] 针对cookies封锁,采取禁用cookies,COOKIES_ENABLED = False的方式。

[0023] 针对ajax异步传输,在浏览器开发者界面选择network标签,再次刷新,我们会发现异步请求的文件展现在name文件列表,找到它之后获取headers中的url并向其发送post请求即可爬取信息。

[0024] 新闻预处理结束后,进行新闻标签向量化操作,如图4所示,将用户自定义的标签作为该用户的兴趣表征,遍历新闻列表,采用余弦相似度的计算得到用户和新闻标签的相似距离,若相似距离超过设定阈值,则将新闻加入到推荐候选集中,直至所有新闻数据遍历完成。由于这样计算完成后,可能会造成某些用户的推荐结果过多,因此这里采用热度排序的方式,选出匹配度最高的10个新闻作为最终推荐。

[0025] 其中,余弦相似度算法通过计算在同一个向量空间中两个向量的夹角余弦值,来表征两个个体(用向量表示)之间差异距离的大小。具体来说就是,如果得出的余弦值越接近1,那么夹角就越接近0,这样的结果就能够说明两个向量越相似;反之亦然。

[0026] 本方法将利用余弦相似度计算的方法来基于标签来计算用户和新闻之间的相似

程度，具体计算过程包括四个步骤：1、定义标签类别，首先对可能出现的新闻标签进行所属类别定义。

[0027] 2、添加标签，通过用户自定义标签和文章关键字提取，分别为用户和新闻添加标签。3、标签向量化，采用oneHot编码的方式将用户和新闻标签编码为向量形式，将定义的所有标签设定为一个向量中的一位，对于用户或新闻中如果包含某个标签，那么对应向量中的那一位标签置为1，否则置为0。4、利用余弦相似度函数计算所有标签向量中每两个标签向量之间夹角的余弦值，值越大相似度越高。

[0028] 本实施例中，如图5所示，基于标签相似度的信息推荐系统，主要包括新闻预处理模块、向量化模块、相似度计算模块、热度排序模块和新闻推荐模块。其中，新闻预处理模块，用于从各个新闻源中爬取新闻信息，并对获取的新闻数据进行数据清洗和去重，生成结构化数据。向量化模块，用于对结构化的新闻数据进行标签向量化，同时对用户属性进行向量化，获得对应的新闻的标签向量和用户的属性向量。相似度计算模块，用于根据新闻的标签向量和用户的属性向量进行相似度计算，将与每个用户的属性最相近的新闻作为推荐候选集。热度排序模块，用于根据热度排序算法从推荐候选集中选出热度最高的N个新闻作为最终候选集，并写入推荐候选表。新闻推荐模块，用于根据推荐候选表中的新闻ID，在新闻数据库中进行匹配，得到对应的新闻内容数据，反馈给用户。

[0029] 本实施例中，系统中的新闻预处理模块的主要作用是从第三方库爬取新闻源数据，并进行新闻内容的清洗和深度分析，最终输出为经过预处理的结构化新闻数据。接口设计包括向外提供爬虫接口和新闻输出接口，由系统内核调用并连接数据库。

[0030] 向量化模块和相似度计算模块的主要作用是分别从数据库和用户属性表中提取新闻标签和用户属性，并将其向量化，经过余弦相似度的计算得出与该用户属性最相近的新闻推荐。其接口设计包括调用数据库和用户属性表两个数据接口；向外提供基于标签的推荐候选表接口，由用户调用。

[0031] 以上显示和描述了本发明的基本原理和主要特征和本发明的优点。本行业的技术人员应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化和改进都落入要求保护的本发明范围内。本发明要求保护的范围由所附的权利要求书及其等效物界定。

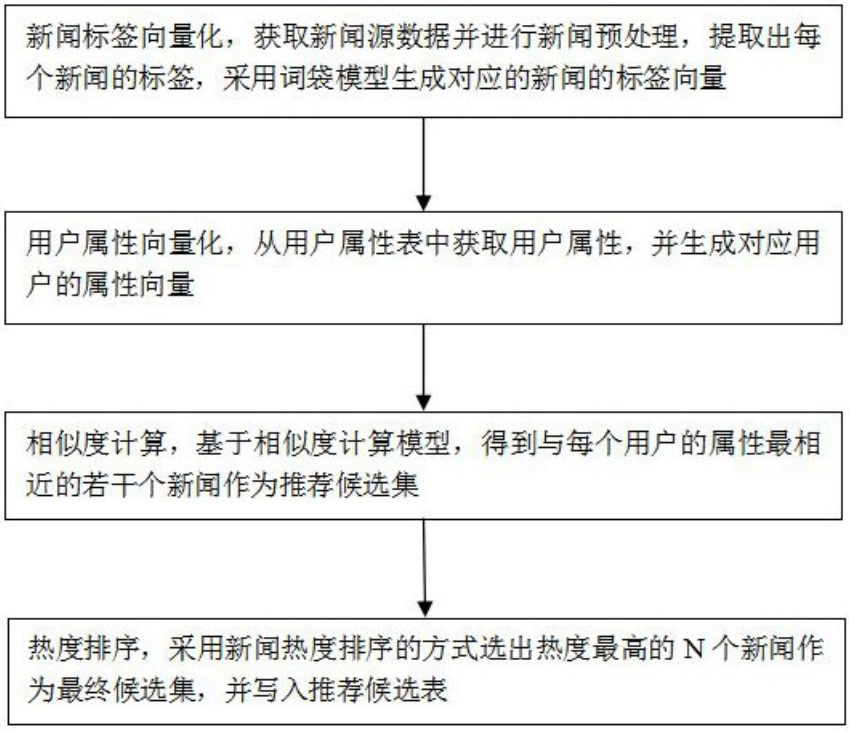

新闻标签向量化，获取新闻源数据并进行新闻预处理，提取出每个新闻的标签，采用词袋模型生成对应的新闻的标签向量
用户属性向量化，从用户属性表中获取用户属性，并生成对应用户的属性向量
相似度计算，基于相似度计算模型，得到与每个用户的属性最相近的若干个新闻作为推荐候选集
热度排序，采用新闻热度排序的方式选出热度最高的N个新闻作为最终候选集，并写入推荐候选表

图1

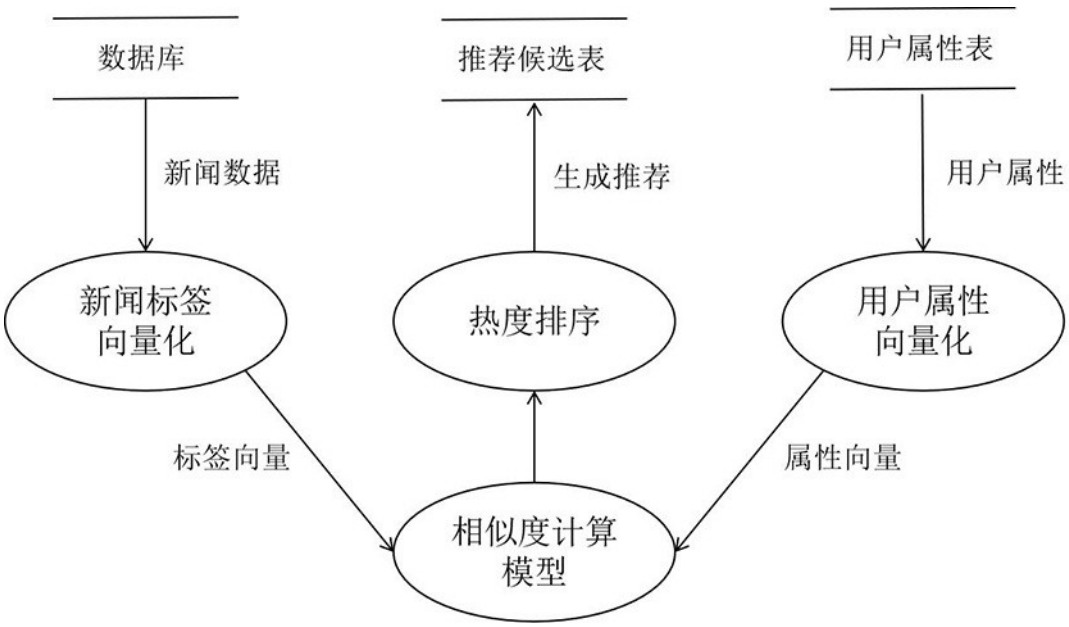

数据库
推荐候选表
用户属性表
新闻数据
生成推荐
用户属性
新闻标签向量化
热度排序
用户属性向量化
标签向量
属性向量
相似度计算模型

图2

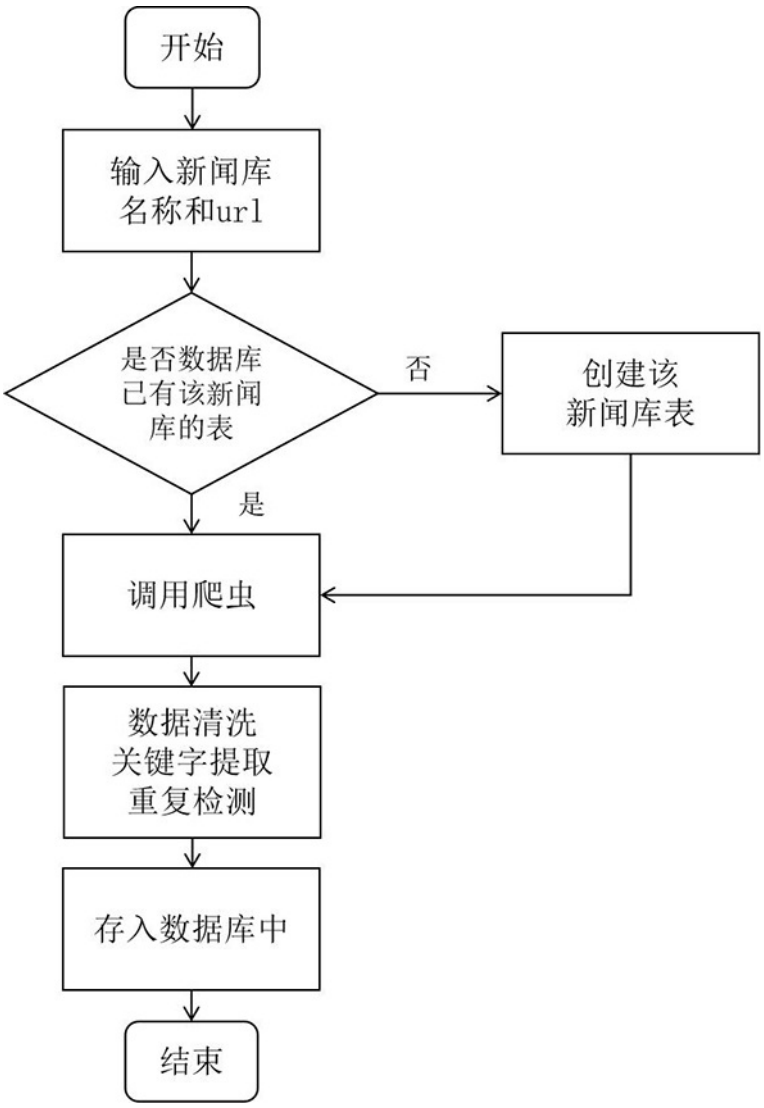

开始
输入新闻库
名称和url
是否数据库
已有该新闻
库的表
否
创建该
新闻库表
是
调用爬虫
数据清洗
关键字提取
重复检测
存入数据库中
结束

图3

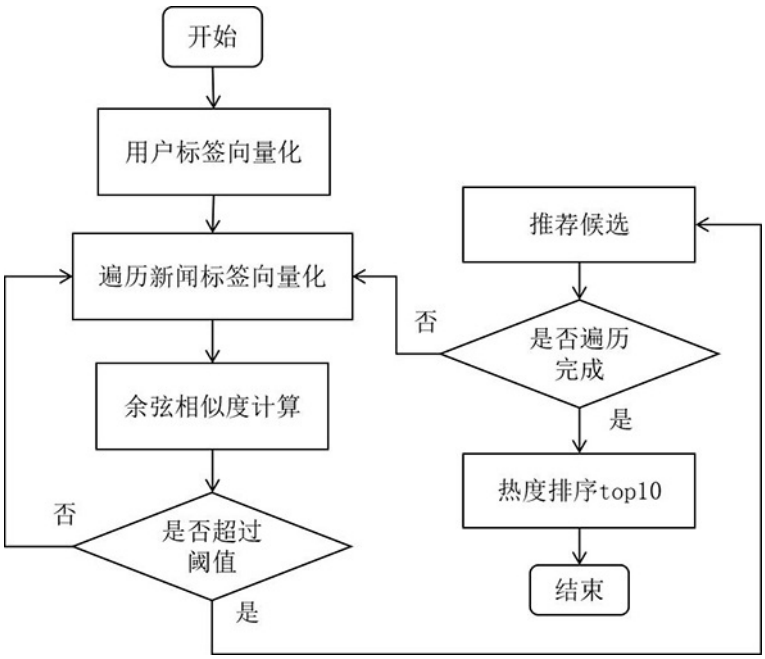

开始
用户标签向量化
遍历新闻标签向量化
余弦相似度计算
是否超过阈值
否
是
是否遍历完成
否
是
推荐候选
热度排序top10
结束

图4

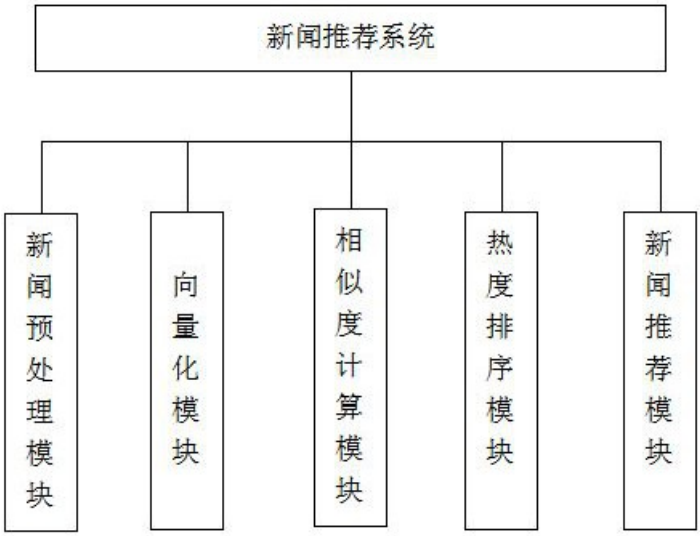

新闻推荐系统
新闻预处理模块
向量化模块
相似度计算模块
热度排序模块
新闻推荐模块

图5