(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 102929928 B
(45)授权公告日 2015.04.22

(21)申请号 201210355209.0

(22)申请日 2012.09.21

(73)专利权人 北京格致璞科技有限公司
地址 102399 北京市门头沟区妙峰山镇水丁路1号
专利权人 北京邮电大学

(72)发明人 叶小卫 曹一鸣 卢美莲 王明华 李佳珊 刘金亮

(74)专利代理机构 北京德琦知识产权代理有限公司 11018
代理人 夏宪富

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

(56)对比文件
CN 102376063 A,2012.03.14,
US 5867799 A,1999.02.02,
CN 101174273 A,2008.05.07,

审查员 张莹

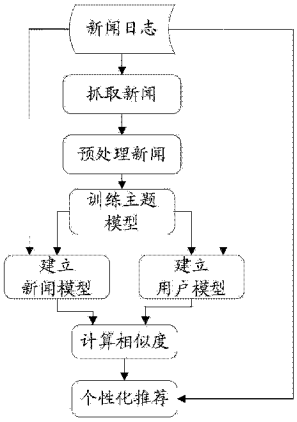
权利要求书3页 说明书9页 附图3页

(54)发明名称

基于多维相似度的个性化新闻推荐方法

(57)摘要

一种基于多维相似度的个性化新闻推荐方法：先从新闻日志抽取设定时间记录，根据新闻源地址抓取新闻并抽取标题和正文，对其进行分词和提取名词，并用主题模型分析该名词序列而得到该新闻的主题特征向量；再根据新闻的主题特征向量和用户行为数据，分别构建用户模型和新闻模型；然后根据用户模型、新闻模型和时间特征分别计算用户和新闻的内容相似度与行为相似度，并据此计算最终的用户相似度和最终的新闻相似度，并分别提取最相似的多个用户和多个新闻；最后，依据最近的新闻日志记录和与设定用户最相似的多个相似用户，生成基于用户的个性化推荐结果；或依据设定用户产生行为的新闻和与该新闻最相似的多个新闻，生成基于新闻的个性化推荐结果。



CN 102929928 B

1. 一种基于多维相似度的个性化新闻推荐方法，其特征在于：所述方法包括下列操作步骤：

(1) 抓取新闻：根据新闻日志中记录的新闻网页地址、即统一资源定位符 URL 抓取每篇新闻的标题和正文，并存储于新闻数据库中；

(2) 预处理新闻：从新闻数据库中取出新闻标题和正文，并使用分词系统对新闻正文进行分词、词性标注和提取其中名词，组成由新闻标识 id- 新闻名词序列构成的二维表，并存储于数据库中；

(3) 训练主题模型：采用潜在狄利克雷分布 LDA 和多个主题 k 对从数据库中读取的新闻 id- 新闻名词序列表进行主题模型训练，得到每篇新闻的主题模型、即主题特征向量 L = $(w_1, w_2, \ldots w_1 \ldots, w_k)$，且 $\sum_{l=1}^{k} w_l = 1$；式中，自然数下标 1 是主题序号，其最大值为主题总个数 k，w_1 是该新闻属于第 1 个主题的概率；

(4) 建立由两个特征组成的新闻模型：一个是行为特征 list($(u_1, t_1), (u_2, t_2), (u_3, t_3), \ldots$)，即从新闻日志中获取设定时间内对新闻产生浏览、评论、发布和推荐行为的用户 u 及其产生行为的时间 t 的序列；另一个是根据主题模型的训练结果得到每篇新闻的内容特征、即新闻主题特征向量 L = $(w_1, w_2, \ldots w_1 \ldots, w_k)$；

(5) 建立由两个特征组成的用户模型：一个是行为特征 list($(i_1, t_1), (i_2, t_2), (i_3, t_3), \ldots$)，即从新闻日志中获取设定时间内用户产生行为的各个新闻 i 及产生行为的时间 t 的序列；另一个是每篇新闻的内容特征，即用户具有历史行为的所有新闻的主题特征向量的平均值、即用户的主题特征向量 $u_{(w_1, w_2, \ldots, w_l, \ldots, w_k)} = \frac{\sum_{L \in n(u)} L}{|n(u)|}$，式中，n(u) 是用户 u 产生行为的新闻集合，自然数下标 i 是新闻序号，L 为新闻的主题特征向量；

(6) 利用用户模型、新闻模型和时间特征分别计算设定时间内所有用户之间的相似度和所有新闻之间的相似度：这两种相似度计算又各自分为行为相似度和内容相似度的计算，再对该两种相似度数值加权求和，作为用户之间和新闻之间的最终融合相似度，然后，分别提取最相似的多个用户和多个新闻存入数据库；

(7) 个性化推荐：分别依据最近的新闻日志记录，以及与设定用户最相似多个相似用户，生成基于用户的个性化推荐结果；或者依据设定用户当前产生行为新闻的最相似的多个新闻，生成基于新闻的个性化推荐结果；并实时更新推荐列表，如果当前尚未完成新闻的相似度的计算，则推荐结果维持不变。

2. 根据权利要求 1 所述的方法，其特征在于：所述步骤 (6) 中，计算用户相似度包括下列操作内容：

(61) 按照下述公式计算两个用户 u 和 v 的行为相似度 sim(u, v)：

$$sim(u,v) = \frac{\sum_{i \in n(u) \cap n(v)} \frac{1}{\log(1+|m(i)|)} e^{-\alpha |t_{ui} - t_{vi}|}}{n(u) \cup n(v)}$$；式中，n(u) 和 n(v) 分别为用户 u 和 v 产生过行为的新闻集合，m(i) 为对第 i 篇新闻产生行为的用户集合；t_{ui} 和 t_{vi} 分别为用户 u 和 v 对

第 i 篇新闻产生行为的时间，系数 α 是时间衰减因子，其数值取值范围为 [0,1]；

(62) 按照下述公式计算两个用户 u 和 v 的内容相似度、即余弦相似度 cos(u,v)：

$\cos(u,v)=\frac{\vec{u}\vec{v}}{|\vec{u}|\times|\vec{v}|}$，式中，$\vec{u}$ 和 $\vec{v}$ 分别为用户 u 和用户 v 的主题特征向量；

(63) 按照下述公式计算两个用户 u 和 v 的最终用户相似度 W(u,v)：

W(u,v) = βsim(u,v)+(1-β)cos(u,v)；式中，sim(u,v) 为该两个用户 u 和 v 的行为相似度，cos(u,v) 为该两个用户 u 和 v 的内容相似度，系数 β 是由实验确定的加权因子，其数值取值范围为 [0,1]。

3. 根据权利要求 1 所述的方法，其特征在于：所述步骤 (6) 中，计算新闻相似度包括下列操作内容：

(6A) 按照下述公式计算两篇新闻 i 和 j 的行为相似度 sim(i,j)：

$sim(i,j)=\frac{\sum_{u\in m(i)\cap m(j)} e^{-\alpha|t_{ui}-t_{uj}|}}{m(i)\cup m(j)}$；式中，m(i) 和 m(j) 分别为对第 i 篇新闻和第 j 篇新闻产生过行为的用户集合，t_{ui} 和 t_{uj} 分别为用户 u 对第 i 篇新闻和用户 v 对第 j 篇新闻产生行为的时间，系数 α 是时间衰减因子，其数值取值范围为 [0,1]；

(6B) 按照下述公式计算两篇新闻的内容相似度、即余弦相似度 cos(i,j)：

$\cos(i,j)=\frac{\vec{i}\vec{j}}{|\vec{i}|\times|\vec{j}|}$；式中，$\vec{i}$ 和 $\vec{j}$ 分别为第 i 篇新闻和第 j 篇新闻的主题特征向量；

(6C) 按照下述公式计算两篇新闻 i 和 j 的最终新闻相似度 W(i,j)：

W(i,j) = βsim(i,j)+(1-β)cos(i,j)；式中，sim(i,j) 为两篇新闻 i 和 j 的行为相似度，cos(i,j) 为两篇新闻 i 和 j 的内容相似度，系数 β 是由实验确定的加权因子，其数值取值范围为 [0,1]。

4. 根据权利要求 1 所述的方法，其特征在于：所述步骤 (7) 中，基于用户的个性化推荐包括下列操作内容：

(71) 按照下述公式计算最近时间段内设定用户 u 对其未产生行为的每篇新闻的偏好程度：$p(u,i)=\sum_{v\in S(u,K)\cap m(i)} W(u,v)e^{-\gamma(t'-t_{vi})}$，式中，S(u,K) 为用户 u 的最相似的 K 个用户集合，m(i) 为对第 i 篇新闻产生行为的用户集合，W(u,v) 为两个用户 u 和 v 的最终用户相似度，t' 为当前时间，t_{vi} 为用户 v 对第 i 篇新闻产生行为的时间，系数 γ 为时间衰减因子，其取值范围 [0,1]；

(72) 根据最近时间段内设定用户 u 对其未产生行为的每篇新闻的偏好程度值的大小，对这些新闻进行降序排列，再选取其中偏好值高的多个新闻作为向该设定用户 u 个性化推荐的新闻列表。

5. 根据权利要求 1 所述的方法，其特征在于：所述步骤 (7) 中，基于新闻的个性化推荐包括下列操作内容：实时获取设定用户当前正在产生行为的新闻，再从数据库中选择和该新闻最相似的多篇新闻向该用户推荐；如果该新闻的相似新闻还未计算出来，即数据库中

不存在相似新闻时，则推荐列表维持不变；也就是，该用户对某篇新闻产生行为后，快速更新所推荐的新闻列表，以便实现对用户新闻兴趣偏好的即时追踪。

基于多维相似度的个性化新闻推荐方法

技术领域

[0001] 本发明涉及一种基于多维相似度的个性化新闻推荐方法，特别是涉及一种融合内容相似度、行为相似度和时间特征的个性化新闻推荐方法，属于基于协同过滤的个性化新闻推荐的技术领域。

背景技术

[0002] 随着互联网规模的迅速发展，人们获取信息的方式越来越多，信息呈爆炸式增长，用户逐渐由信息匮乏走向了信息过载时代——海量信息使得用户难以寻找到各自所需的信息。为了方便用户从海量信息中寻找其所需的内容，出现了很多解决方案：包括分类目录和搜索引擎。分类目录是把常用热门网站分门别类，便于用户查找信息。但是，随着互联网规模的扩大，分类目录只能覆盖少量热门网站。搜索引擎是用户只需把自己的需求转换成关键词的不同组合，再在网络中寻找其所需的信息。当用户有明确需求时，搜索引擎还是可行的，但是，用户在很多时候并不知道自己的需求。比如用户打开优酷视频网站，上面有大量视频内容信息，用户并没有明确的需求非要观看什么，这时如果推荐引擎能够自动给用户推荐一些视频，而这些视频恰好是用户所喜欢的，就能够很好地解决上述问题。著名电子商务网站 Amazon 销售额中的 35％是来自推荐系统。由此可见，推荐系统在提高了用户的满意度的同时，也提高了网站的黏性，增加了网站访问量，为网站带来巨大的商业利益。

[0003] 在用户需求模糊时，推荐引擎能够自动把用户感兴趣的内容推荐给用户，同时，过滤用户不感兴趣的大量内容，即为不同用户呈现不同的个性化内容。目前，推荐系统的实现方式很多，其中，协同过滤技术因其与内容无关，成为最为广泛使用的个性化推荐技术，被应用到电子商务、视频网站、个性化阅读、个性化广告等许多领域。

[0004] 目前，应用最广泛的协同过滤个性化推荐技术有两种方式（参见图 1）：基于用户的协同过滤和基于项目的协同过滤。前者主要包括三个步骤：用户行为数据表示；利用用户相似度计算方法，查找与目标用户最相似的多个用户；根据该多个相似用户对项目的行为来预测目标用户对项目的行为，并进行推荐。后者也包括三个步骤：项目行为数据表示；利用项目相似度计算方法，计算项目之间的相似度；把与用户产生行为的项目最相似的项目推荐给用户。

[0005] 下面详细介绍基于用户和基于项目的两种协同过滤的流程：

[0006] 基于用户的协同过滤技术中，用户行为数据表示为用户 - 项目二维矩阵，其中每行是用户对各列中各个项目的评分，通常的评分是 1 ～ 5。

[0007] 用户相似度的计算是协同过滤中最关键的操作，传统的相似度的计算有下述三种：余弦相似度、修正余弦相似度和皮尔逊相似度。

[0008] 余弦相似度（即 cosine 相似度）：将用户评分看作多维项目空间上的向量，如果用户对项目没有评分，则将该用户对该项目的评分设为 0；用户间相似度值是向量之间的余弦夹角值。余弦相似度的优点是：将用户没有评分的项目的评分值设为 0，有效提高了计算性能。但事实上，用户对未评分项目的喜好程度不可能全都相同、且都为 0。所以，在评

分数据稀疏情况下，余弦相似度方法就无法准确计算用户之间的相似度和项目之间的相似度；同时，余弦相似度并未考虑用户评分尺度的问题。

[0009] 修正的余弦相似度度量：将用户对项目的评分减去用户对项目的平均评分，以改善余弦相似度度量方法的缺陷，这种度量方法考虑了不同用户的评价尺度问题。与余弦相似度性类似，它也是将用户未评分项目的评分值设为 0，在稀疏矩阵情况下，也不能准确地计算出用户 / 项目之间的相似度。

[0010] 皮尔森相似度：只在用户间共同评分的项目上计算相似度，比修正的余弦相似度计算方法中直接用 0 来填充，具有更好的推荐质量。

[0011] 推荐方法也有两种：评分预测和 Top-N 推荐。其中，评分预测是先计算用户对项目的预测评分，然后选择评分高的项目推荐给用户。评分预测有两种方法：第一种是简单加权平均：$r_{u,i}=\frac{\sum_{v\in s(u,K)} sim(u,v)R_{v,i}}{\sum_{v\in s(u,K)} sim(u,v)}$；其中，s(u,K) 为用户 u 的 K 个相似用户，$r_{u,i}$ 和 $R_{v,i}$ 分别为用户 u 及其邻居用户 v 对第 i 个项目的预测评分和实际评分。sim(u,v) 为两个用户 u 和 v 的行为相似度。第二种是考虑各个用户间不同的评分尺度的用户评分偏移加权平均：$r_{u,i}=\overline{R_u}+\frac{\sum_{v\in s(u,K)} sim(u,v)(R_{v,i}-\overline{R_v})}{\sum_{v\in s(u,K)} sim(u,v)}$。

[0012] Top-N 推荐是计算用户的兴趣偏好程度，选择其中最高的 N 个项目推荐。Top-N 推荐公式是：$r_{u,i}=\sum_{v\in s(u,K)} sim(u,v)R_{v,i}$。

[0013] 目前的很多研究表明，Top-N 推荐优于评分预测推荐，因为向用户最终推荐的项目准确率的评判标准是用户是否查看，而不是看完以后的评价是多少分。

[0014] 基于项目的协同过滤和基于用户的协同过滤的最大不同是：前者是计算出相似项目后，通过查找和目标用户产生行为的项目最相似的若干项目作为推荐。

[0015] 项目行为数据是由用户的行为数据来表示，项目相似度的计算方法是把用户相似度计算公式中的用户替换成项目、项目替换成用户。

[0016] 传统相似度的计算方法得到了广泛应用，但依然存在很多问题。例如：两个用户之间的相似度无法计算或计算的结果错误；随着推荐系统的规模扩大，数据稀疏性使得上述问题更加严重；而且，传统相似度计算不适用于集中评分数据，例如对于 1-5 的评分项目，用户的大部分评分可能集中在 3-4 之间。

[0017] 另外，某些情况的相似度无法计算：如果共同评分项目是一个，则皮尔逊相似度就无法计算，因为其分母为 0。如果用户的评分非常平稳时，比如 <1,1,1,>,<3,3,3> 或 <4,4,4> 时，皮尔逊相似度的分母也为 0，同样无法计算相似度。

[0018] 某些情况计算出的相似度不准确：如果两个评分向量位于同一条直线上，例如评分 <1,1>、<4,4>，很显然，用户的喜好还是不同的，但是其余弦相似度为 1，即很相似。如果两个用户评分是线性相关，例如 v1<1,2,1,2,1>、v2<4,5,4,5,4>、v2 = v1+3，那么皮尔逊相似度是 1。实际上，这两个用户是显然不相同的。如果两个用户评分向量为 <4,5,4,5,4> 和

<5,4,5,4,5>，虽然这两个评分向量很相似，但皮尔逊相似度为 -1，即负相关。

[0019] 目前，在数据稀疏时，共同评分的项目很少，计算结果往往不准确或无法计算。随着推荐系统的规模越来越大，用户和项目的数量都急剧扩大，因为每个用户只会选择少数项目，这样，用户间选择相同项目的可能性越来越小，使得数据稀疏性的问题越来越严重，进而导致相似度无法计算或计算结果不准确。以实验常用的数据集为例，MovieLens 的稀疏度是 95.5%，Netflix 的稀疏度是 98.8%，Delicious 的稀疏度是 99.954%。

[0020] 为了解决数据稀疏性导致相似度无法计算或计算不准确问题，目前，已经提出了矩阵填充、矩阵降维以及其他相似度计算方法，下面简要说明之。

[0021] （一）矩阵填充 - 缺省填充：解决数据稀疏性问题的最简单方法是把矩阵的空位置添上数值，称为矩阵填充。

[0022] 大多数情况下，缺省填充值设置为中值或稍小的数值，也可设置为用户的评分均值或项目的评分均值。但是，该填充方法的问题是：用户对项目的评分不可能完全相同，以这种方法填充的评分矩阵的可信度不高。

[0023] 众数法：采用一组数据中出现频率最高的数对未评分项目进行赋值，即采用目标用户所有评分的众数作为未评分项目的预测值。但是，众数法存在“多众数”（即有两个或两个以上的评分值出现次数是最多时）和“无众数”（所有评分值的出现次数都相同）的问题，导致这种方法应用的局限性很大。

[0024] （二）矩阵填充 - 预测填充：通过预测评分来填充，有代表性的是基于项目评分预测的协同过滤推荐方案。该方案是：先计算经过两个用户 u 和 v 评分的项目集合的并集 P_{uv}。两个用户 u 和 v 在项目集合 P_{uv} 中未评分的项目则由用户对相似项目的评分预测出来，然后，在项目集合 P_{uv} 上采用修正余弦相似度或皮尔逊相似度计算这两个用户 u 和 v 之间的相似度。最后找到最相似 K 个用户产生推荐。

[0025] 这种方法不仅有效解决相关相似度度量方法中用户共同评分数据比较少的情况，而且，有效解决余弦相似度度量方法和修正的余弦相似度度量方法中对所有未评分项目的评分均相同的问题（均为 0），使得计算得到的目标用户的最近邻居比较准确。但在实际应用中，用户 - 项目二维表已经很庞大，对稀疏的地方进行填充，不仅增加计算量，而且消耗大量内存空间，从实际效果来看，矩阵填充技术对评分预测的准确度提高有限。

[0026] （三）矩阵降维 - 云模型：为解决数据稀疏性问题，提出了云模型方案：将某个用户对多个项目的评价情况进行统计，称为用户评分频度向量。根据用户评分频度向量，再利用逆向云算法可以计算用户的评分特征向量，记为 $q = (E_x, E_n, H_e)$，其中，期望 E_x 为用户对所有项目的平均满意度，属于偏好水平；熵 E_n 为用户打分的集中程度，反映投票偏好的离散度；H_e 为熵的稳定度。对于两个用户云模型 q1 和 q2，这两个云之间的相似度：$sim(q_1, q_2) = \frac{q_1 \cdot q_2}{|q_1| \times |q_2|}$ 就是其特征向量的余弦相似度。这样把评分矩阵转化成 3 个指标，解决了数据稀疏性问题，但是，因为采用了降维技术，丢失了大量相关信息，推荐效果不理想。

[0027] （四）PIP 相似度计算模型：用于解决冷启动问题的 PIP 方法，把两个用户的每对评分都划分为三个因素来计算每对评分的相似度，最后综合所有评分的相似度，得出用户之间的相似度。PIP 相似度计算方法主要由下述三部分组成：临近度 (Proximity) 是两个用户评分差距，影响度 (Impact) 是两个用户对项目喜好程度，普及度 (Popularity) 是目前

评分与项目平均分的差距。对于任意两项的评分公式为：$PIP(r_1, r_2) = Proximity(r_1, r_2) \times Impact(r_1, r_2) \times Popularity(r_1, r_2)$。该方案在一定程度上解决了协同过滤数据稀疏性问题，但对于评分较多的两个用户，这种计算方式要计算所有可能组合的评分对，不仅计算结果不准确，而且计算工作量很大。

[0028] 传统的相似度计算方法适合数据平均分散的情况，人们通过对数据集研究发现，大部分评分数据是集中的，也就是在 1 ～ 5 的评分中，用户更倾向于 3、4 的评分，很少评为 1 或 5；比如，MovieLens 数据集的评分方差在 1.2，更加表明评分数据的集中性。该方案提出用户评分基本表明用户的喜好：小于 3 分为不喜欢，大于 3 分为喜欢。该方式把用户评分矩阵转换成两个列表：喜欢项目和不喜欢项目。再计算两个用户喜欢列表的 Jaccard 相似度和项目评分的均方位移乘积作为两个用户的相似度，则两个用户 u1 和 u2 的 Jaccard 相似度为：$Jaccard(u_1,u_2)=\frac{like(u_1)\cap like(u_2)}{like(u_1)\cup like(u_2)}$；其中 like(u) 为用户所喜欢的项目集合。最终推荐系统准确率和召回率都有明显提高。

[0029] 上述方案是先把评分矩阵转换成喜欢内容的列表，再用 Jaccard 相似度计算用户相似度。它较好地解决了传统相似度计算方式不适合集中性数据问题，但是，当数据稀疏时，用户选择相同项目的可能性越来越小，Jaccard 相似度也很难计算出用户间相似度。

[0030] 总之，协同过滤推荐作为目前推荐系统采用的主要技术，得到了广泛应用。相似度的计算作为过滤推荐系统的核心，直接决定了相似邻居计算和推荐结果的质量。然而，直至今天，仍然存在很多问题（如：数据稀疏性、冷启动等），其本质是在无评分数据或评分数据稀少时，用户之间相似度如何计算的问题。

[0031] 例如：(1) 传统相似度计算问题：余弦相似度、修正余弦相似度和皮尔逊相似度适用于离散分布的评分数据，然而，实际评分数据往往是集中的；同时当两个评分向量选取特定值时，会导致相似度无法计算或计算结果出现严重偏差。

[0032] (2) 数据稀疏性问题：为了解决数据稀疏性问题，现有的数据填充方案主要缺陷是：数据填充不仅增加了计算量，还要占用大量内存，但对评分预测的准确率并未明显提高。数据降维方案的缺陷是：在降维过程中丢失了大量有用信息，最终推荐结果并不理想；相似度计算方式 PIP 也并未有效改善推荐结果。

[0033] 总之，相似度的计算是基于协同过滤的推荐系统中最关键的操作步骤，目前的相似度计算方法存在的无法计算、计算准确度不高和占用大量内存等多个问题，都有待于业内科技人员尽快解决之。

发明内容

[0034] 有鉴于此，本发明的目的是提供一种基于多维相似度的个性化新闻推荐方法，本发明根据新闻推荐领域的特殊性，结合新闻的内容特征、用户的行为特征和时间特征来计算用户或新闻的相似度。这样即使两个用户行为数据或新闻内容的重合度很低，但是，因新闻的内容相似，也可以计算出用户或新闻之间的相似度。同时，在相似度计算中加入时间特征，使得相似度的计算更加准确。最后，基于该相似度计算方法提供一种具有较高准确率的个性化新闻推荐方法。

[0035] 为了达到上述目的，本发明提供了一种基于多维相似度的个性化新闻推荐方法，

其特征在于：所述方法包括下列操作步骤：

[0036] (1) 抓取新闻：根据新闻日志中记录的新闻网页地址、即统一资源定位符 URL 抓取每篇新闻的标题和正文，并存储于新闻数据库中；

[0037] (2) 预处理新闻：从新闻数据库中取出新闻标题和正文，并使用分词系统对新闻正文进行分词、词性标注和提取其中名词，组成由新闻标识 id- 新闻名词序列构成的二维表，并存储于数据库中；

[0038] (3) 训练主题模型：采用潜在狄利克雷分布 LDA 和多个主题 k 对从数据库中读取的新闻 id- 新闻名词序列表进行主题模型训练，得到每篇新闻的主题模型、即主题特征向量 $L = (w_1, w_2, \ldots w_1 \ldots, w_k)$，且 $\sum_{l=1}^{k} w_l = 1$；式中，自然数下标 l 是主题序号，其最大值为主题总个数 k，w_1 是该新闻属于第 l 个主题的概率；

[0039] (4) 建立由两个特征组成的新闻模型：一个是行为特征 $list((u_1, t_1), (u_2, t_2), (u_3, t_3), \ldots)$，即从新闻日志中获取设定时间内对新闻产生浏览、评论、发布和推荐行为的用户 u 及其产生行为的时间 t 的序列；另一个是根据主题模型的训练结果得到每篇新闻的内容特征、即新闻主题特征向量 $L = (w_1, w_2, \ldots w_1 \ldots, w_k)$；

[0040] (5) 建立由两个特征组成的用户模型：一个是行为特征 $list((i_1, t_1), (i_2, t_2), (i_3, t_3), \ldots)$，即从新闻日志中获取设定时间内用户产生行为的各个新闻 i 及产生行为的时间 t 的序列；另一个是每篇新闻的内容特征，即用户具有历史行为的所有新闻的主题特征向量的平均值、即用户的主题特征向量 $u_{(w_1, w_2, \ldots, w_l, \ldots, w_k)} = \frac{\sum_{L \in n(u)} L}{|n(u)|}$，式中，n(u) 是用户 u 产生行为的新闻集合，自然数下标 i 是新闻序号，L 为新闻的主题特征向量；

[0041] (6) 利用用户模型、新闻模型和时间特征分别计算设定时间内所有用户之间的相似度和所有新闻之间的相似度：这两种相似度计算又各自分为行为相似度和内容相似度的计算，再对该两种相似度数值加权求和，作为用户之间和新闻之间的最终融合相似度，然后，分别提取最相似的多个用户和多个新闻存入数据库；

[0042] (7) 个性化推荐：分别依据最近的新闻日志记录，以及与设定用户最相似多个相似用户，生成基于用户的个性化推荐结果；或者依据设定用户当前产生行为新闻的最相似的多个新闻，生成基于新闻的个性化推荐结果；并实时更新推荐列表，如果当前尚未完成新闻的相似度的计算，则推荐结果维持不变。

[0043] 本发明推荐方法的优点是：针对新闻领域的特殊性，在计算用户相似度和新闻相似度时，不仅考虑传统的用户行为相似度、即从用户行为数据的相似度出发，还融合新闻内容、即从新闻内容角度挖掘用户或新闻之间的相似性，以提高相似度计算的准确性。尤其在用户行为数据稀疏时，本发明方法比传统相似度算法更能挖掘用户相似性，使得推荐结果的准确率和召回率都得到明显提升。同时，本发明把时间特征引入到推荐过程中的各个环节：计算用户相似度和新闻相似度，以及 Top-N 推荐过程中，都考虑了时间特征，使得相似度的计算结果更加准确，推荐的新闻更具有时效性，以及最终的推荐结果准确率和召回率都得到显著提高，从而，提高了本发明基于多维相似度的个性化新闻推荐方法的推荐质量。

附图说明

[0044] 图 1(A)、(B) 分别是基于用户和基于项目的两个协同过滤操作流程图。

[0045] 图 2 是本发明基于多维相似度的个性化新闻推荐方法的操作步骤流程图。

[0046] 图 3 是建立用户模型和新闻模型的操作步骤示意图。

[0047] 图 4 是相似度计算和个性化推荐的操作步骤示意图。

具体实施方式

[0048] 为使本发明的目的、技术方案和优点更加清楚，下面结合附图对本发明作进一步的详细描述。

[0049] 本发明是针对新闻领域的特殊性所提出的一种融合了用户行为相似度和新闻内容相似度，并结合时间特征的多维相似度的个性化新闻推荐方法，用于提高个性化新闻推荐方法的推荐质量。

[0050] 众所周知，个性化推荐新闻时，由于新闻日志系统存储的是大量隐式行为数据（包括浏览、评论、发布等），而不是显式评分数据，如何有效利用这些数据来计算用户或新闻的相似度是首先要解决的问题。目前的相似度计算方法仅仅利用用户行为数据来计算相似度，忽略了新闻的内容信息，更没有考虑时间特征。本发明方法是：先从新闻日志抽取设定时间的日志记录，根据日志记录的新闻源地址抓取新闻内容；并从该新闻内容中抽取标题和正文，对其进行分词处理和提取名词，以及采用主题模型对所得到的名词序列进行分析，得到该新闻的主题特征向量；接着，根据新闻的主题特征向量和用户行为数据，分别构建用户模型和新闻模型；根据用户模型、新闻模型和时间特征分别计算用户的内容相似度、行为相似度，以及新闻的内容相似度、行为相似度；再基于内容相似度和行为相似度计算最终的用户相似度和最终的新闻相似度，并分别提取最相似的多个用户和多个新闻；然后，分别依据最近的新闻日志记录和与设定用户最相似的多个相似用户，生成基于用户的个性化推荐结果；或者依据设定用户产生行为的新闻和与该新闻最相似的多个新闻，生成基于新闻的个性化推荐结果。

[0051] 参见图 2，介绍本发明方法的具体操作步骤：

[0052] 步骤 1，抓取新闻：根据新闻日志中记录的新闻网页地址、即统一资源定位符 URL(Uniform Resource Locator) 抓取每篇新闻的标题和正文，并存储于新闻数据库中。

[0053] 步骤 2，预处理新闻：从新闻数据库中取出新闻标题和正文，并使用中科院 ICTCLAS 分词系统对新闻正文进行分词、词性标注和提取其中名词，组成由新闻标识 id- 新闻名词序列构成的二维表，并存储于数据库中。

[0054] 步骤 3，训练主题模型：采用潜在狄利克雷分布 LDA(Latent Dirichlet Allocation) 和多个主题 k 对从数据库中读取的新闻 id- 新闻名词序列表进行主题模型训练，得到每篇新闻的主题模型、即主题特征向量 $L = (w_1, w_2, \ldots w_1 \ldots, w_k)$，且 $\sum_{l=1}^{k} w_l = 1$；式中，自然数下标 1 是主题序号，其最大值为主题总个数 k，w_1 是该新闻属于第 1 个主题的概率。

[0055] 步骤 4，建立由两个特征组成的新闻模型（参见图 3）：一个是行为特征 list($(u_1, t_1), (u_2, t_2), (u_3, t_3), \ldots$)，即从新闻日志中获取设定时间内对新闻产生浏览、评论、发布和

推荐行为的用户 u 及其产生行为的时间 t 的序列；另一个是根据主题模型的训练结果得到的每篇新闻的内容特征、即新闻主题特征向量 L = (w_1, w_2, ... w_1..., w_k)。

[0056] 步骤 5，建立由两个特征组成的用户模型（参见图 3）：一个是行为特征 list((i_1, t_1), (i_2, t_2), (i_3, t_3), ...)，即从新闻日志中获取设定时间内用户产生行为的各个新闻 i 及产生行为的时间 t 的序列；另一个是每篇新闻的内容特征，即用户具有历史行为的所有新闻的主题特征向量的平均值、即用户的主题特征向量 $u_{(w_1,w_2,...,w_k,...,w_k)}=\frac{\sum_{L\in n(u)} L}{|n(u)|}$，式中，n(u) 是用户 u 产生行为的新闻集合，自然数下标 i 是新闻序号，L 为新闻的主题特征向量。

[0057] 步骤 6，利用用户模型、新闻模型和时间特征分别计算设定时间内所有用户之间的相似度和所有新闻之间的相似度（参见图 4）：这两种相似度计算又各自分为行为相似度和内容相似度的计算，再对该两种相似度数值加权求和，作为用户之间和新闻之间的最终融合相似度，然后，根据最终融合相似度分别提取最相似的多个用户和多个新闻存入数据库。

[0058] 下面分别介绍该步骤中的用户相似度的计算和新闻相似度的计算，其中，计算用户相似度的操作包括下列内容：

[0059] (61) 按照下述公式计算两个用户 u 和 v 的行为相似度 sim(u, v)：

[0060] $sim(u,v)=\frac{\sum_{i\in n(u)\cap n(v)}\frac{1}{\log(1+|m(i)|)}e^{-\alpha|t_{ui}-t_{vi}|}}{n(u)\cup n(v)}$；式中，n(u) 和 n(v) 分别为两个用户 u 和 v 产生过行为的新闻集合，m(i) 为对第 i 篇新闻产生行为的用户集合；t_{ui} 和 t_{vi} 分别为用户 u 和 v 对第 i 篇新闻产生行为的时间，系数 α 是时间衰减因子，其数值取值范围为 [0, 1]。

[0061] 这里先介绍两个用户 n(u) 和 n(v) 的 Jaccard 相似度计算公式为：$Jaccard(n(u),n(v))=\frac{n(u)\cap n(v)}{n(u)\cup n(v)}$，本发明在计算 n(u) 和 n(v) 的 Jaccard 相似度的基础上，增加了两个维度：热度和时间。其中热度采用了 John S. Breese 提出的思想，时间特征是本发明提出的。

[0062] 众所周知，John S. Breese 提出两个用户对冷门物品采取过相同行为，更能说明两者兴趣的相似度。同样地，热门新闻对于用户相似度的贡献就没有冷门新闻对用户相似度的贡献大。例如，两个用户都看过伦敦奥运会开幕式的新闻，很难说明其兴趣相似；因为很多用户都会关注重大新闻事件。相反，如果两个用户都看过推荐领域最新进展的新闻，说明两个用户比较相似。用户对冷门物品有过相同行为，更能说明用户之间的相似性。因此根据 John S. Breese 的思想，上述用户行为相似度 sim(u, v) 公式对每篇新闻 i 赋予了惩罚处理因子：$\frac{1}{\log(1+|m(i)|)}$，简单地说，就是新闻越热门，对用户间的相似度贡献越少。

[0063] 两个用户兴趣相似，说明这两个用户喜欢的新闻内容很多是相同的，或者对很多相同新闻产生过行为。进一步说，如果两个用户在设定时间内对相同新闻产生行为，更能说明这两个用户间的相似性。因此，用户行为相似度的计算公式还引入了时间特征：$e^{-\alpha|t_{ui}-t_{vi}|}$，这个因子是两个用户 u 和 v 阅读相同新闻的时间间隔的反函数，其表明：这两个用户的阅读

时间越相近，$e^{-\alpha|t_{ui}-t_{vi}|}$数值越大。

[0064] (62) 按照下述公式计算两个用户 u 和 v 的内容相似度、即余弦相似度 cos(u,v)：$\cos(u,v)=\frac{\vec{u}\vec{v}}{|\vec{u}|\times|\vec{v}|}$，式中，$\vec{u}$和$\vec{v}$分别为用户 u 和用户 v 的主题特征向量；

[0065] (63) 融合行为相似度 sim(u,v) 和内容相似度 cos(u,v)，按照下述公式计算两个用户 u 和 v 的最终用户相似度 W(u,v)：W(u,v) = β sim(u,v)+(1-β)cos(u,v)；式中，系数 β 是由实验确定的加权因子，其数值取值范围为 [0,1]。

[0066] 该步骤中，计算新闻相似度操作包括下列内容：

[0067] (6A) 按照下述公式计算两篇新闻 i 和 j 的行为相似度 sim(i,j)：

[0068] $sim(i,j)=\frac{\sum_{u\in m(i)\cap m(j)} e^{-\alpha|t_{ui}-t_{uj}|}}{m(i)\cup m(j)}$；式中，m(i) 和 m(j) 分别为对第 i 篇新闻和第 j 篇新闻产生行为的用户集合，t_{ui}和 t_{uj}分别为用户 u 对第 i 篇新闻和用户 v 对第 j 篇新闻产生行为的时间，系数 α 是时间衰减因子，其数值取值范围为 [0,1]；该公式考虑时间特征，因用户在相近时间看过两个新闻，更说明了这两个新闻间的相似性。

[0069] (6B) 按照下述公式计算两篇新闻的内容相似度、即余弦相似度 cos(i,j)：$\cos(i,j)=\frac{\vec{i}\vec{j}}{|\vec{i}|\times|\vec{j}|}$；式中，$\vec{i}$和$\vec{j}$分别为第 i 篇新闻和第 j 篇新闻的主题特征向量。

[0070] (6C) 融合两篇新闻 i 和 j 的行为相似度 sim(i,j) 和内容相似度 cos(i,j)，按照下述公式计算这两篇新闻的最终新闻相似度 W(i,j)：W(i,j) = β sim(i,j)+(1-β)cos(i,j)；式中，系数 β 是由实验确定的加权因子，其数值取值范围为 [0,1]。

[0071] 步骤 7，个性化推荐（参见图 4）有两种，一种是基于用户推荐：依据最近的新闻日志记录，以及与设定用户最相似多个相似用户，生成基于用户的个性化推荐结果；另一种是基于新闻推荐：依据设定用户当前产生行为新闻的最相似的多个新闻，生成基于新闻的个性化推荐结果；并且，都要实时更新推荐列表，如果当前尚未完成新闻的相似度的计算，则推荐结果维持不变。

[0072] 其中基于用户的个性化推荐包括下列操作内容：

[0073] (71) 实际系统中比较常用 Top-N 推荐，而不是评分预测。因为推荐给用户的新闻是否正确的判断标准是：用户是否喜欢，而不是其看完新闻后对该新闻的评分。本发明在传统新闻推荐基础上加入时间特征，使得推荐结果具有时效性。按照下述公式计算最近时间段内设定用户 u 对其未产生行为的每篇新闻的偏好程度：$p(u,i)=\sum_{v\in S(u,K)\cap m(i)} W(u,v)e^{-\gamma(t'-t_{vi})}$，式中，S(u,K) 为用户 u 的最相似的 K 个用户集合，m(i) 为对第 i 篇新闻产生行为的用户集合，W(u,v) 为两个用户 u 和 v 的最终用户相似度，t' 为当前时间，t_{vi}为用户 v 对第 i 篇新闻产生行为的时间，系数 γ 为时间衰减因子，其取值范围 [0,1]。若邻居用户 v 对对第 i 篇新闻产生行为的时间越远，则该对第 i 篇新闻出现在推荐列表中的可能性越小。也就是邻居用户 v 最近产生行为的新闻对用户推荐结果影响较大。

[0074] (72) 根据最近时间段内设定用户 u 对其未产生行为的每篇新闻的偏好程度值的大小，对这些新闻进行降序排列，再选取其中偏好值高的多个新闻作为向该设定用户 u 个性化推荐的新闻列表。

[0075] 基于新闻的个性化推荐包括下列操作内容：实时获取设定用户当前正在产生行为的新闻，再从数据库中选择和该新闻最相似的多篇新闻向该用户推荐（比如：新闻，最相似新闻 1，最相似新闻 2，最相似新闻 3。这里只选取前 3 个相似的新闻）；如果该新闻的相似新闻还未计算出来，即数据库中不存在相似新闻时，则推荐列表维持不变；这样，该用户对某篇新闻产生行为后，可快速更新所推荐的新闻列表，以便实现对用户新闻兴趣偏好的即时追踪。

[0076] 本发明已经进行了大量的实施试验，下面简要说明实施试验的情况：实施例中的数据集采集于抽屉网站 7 月份的访问记录。

[0077] 1、主题模型提取：通过中科院的 ICTCLAS 汉语分词系统，对新闻集进行分词，过滤选取其中的名词，再删除其中词语数少于 5 个的新闻。选取的主题数为 150，然后通过 LDA 方法进行主题模型训练，得到新闻的主题特征向量 $I = (w_1, w_2, w_3, \ldots, w_n)$，其中 $\sum_{l=1}^{k} w_l = 1$，w_1 是新闻 I 属于主题 l 的概率。读取用户浏览历史表，计算用户的主题特征向量 $u_{(w_1,w_2,w_3,\ldots,w_n)} = \frac{\sum_{I \in n(u)} I}{|n(u)|}$。其中 n(u) 是用户 u 浏览历史。

[0078] 2、基于用户协同过滤推荐和基于项目协同过滤推荐：根据实验确定 $\alpha = 4\times10^{-10}$、$\beta = 0.2$、$\gamma = 4\times10^{-9}$。再分别根据下述三个公式计算用户相似度：$sim(u,v) = \frac{\sum_{i \in n(u) \cap n(v)} \frac{1}{\log(1+|m(i)|)} e^{-\alpha|t_{ui}-t_{vi}|}}{n(u) \cup n(v)}$，$\cos(u,v) = \frac{\vec{u}\vec{v}}{|\vec{u}| \times |\vec{v}|}$ 和 W(u,v) = β sim(u,v)+(1-β) cos(u,v)。

[0079] 然后，再根据公式 $p(u,i) = \sum_{v \in S(u,K) \cap m(i)} W(u,v) r_{vi} e^{-\gamma(t-t_{vi})}$ 对每个用户分别计算其对所有新闻的兴趣偏好程度。最后，把兴趣偏好程度数值最高的前 10 篇新闻推荐到用户推荐列表中。

[0080] 然后，根据下述三个公式：$w_{ij} = \frac{\sum_{u \in m(i) \cap m(j)} e^{-\alpha|t_{ui}-t_{uj}|}}{m(i) \cup m(j)}$，$\cos(i,j) = \frac{\vec{i}\vec{j}}{|\vec{i}| \times |\vec{j}|}$ 和 W(i,j) = β sim(i,j)+(1-β)cos(i,j) 计算新闻相似度，并把其中前三个相似的新闻存储起来。当用户访问完一个新闻后，立即把该 3 个最相似的新闻呈现给用户。

[0081] 总之，本发明的实施试验是成功的，实现了发明目的。

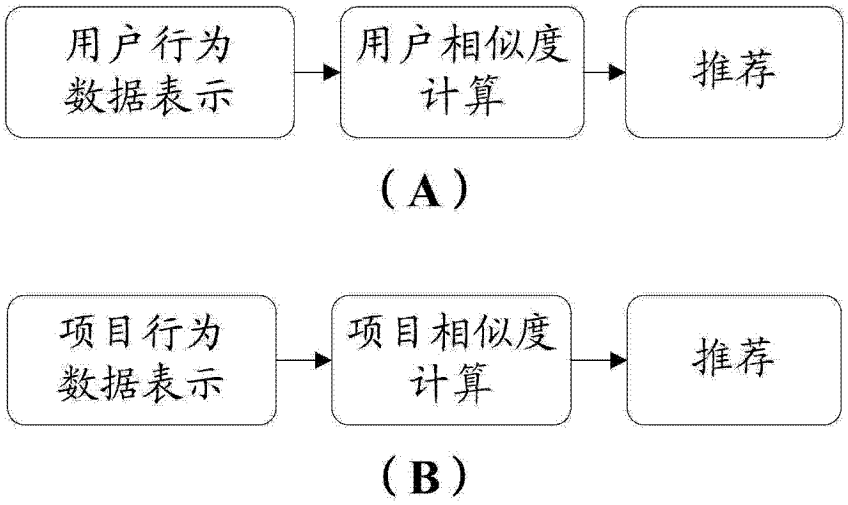

用户行为数据表示
用户相似度计算
推荐
（A）
项目行为数据表示
项目相似度计算
推荐
（B）

图 1

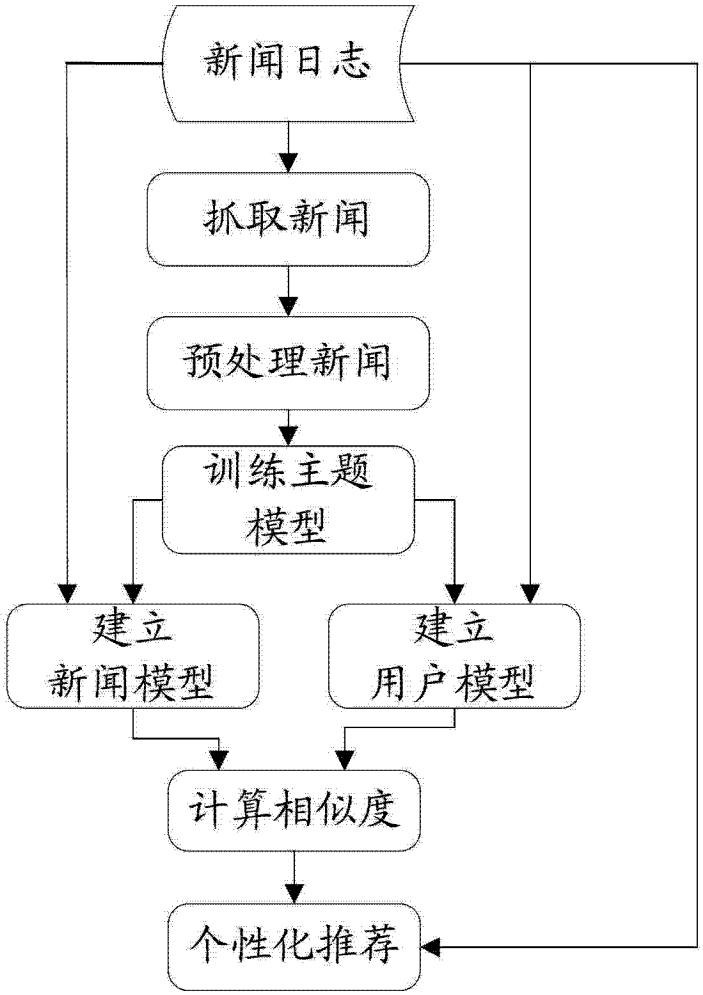

新闻日志
抓取新闻
预处理新闻
训练主题模型
建立新闻模型
建立用户模型
计算相似度
个性化推荐

图 2

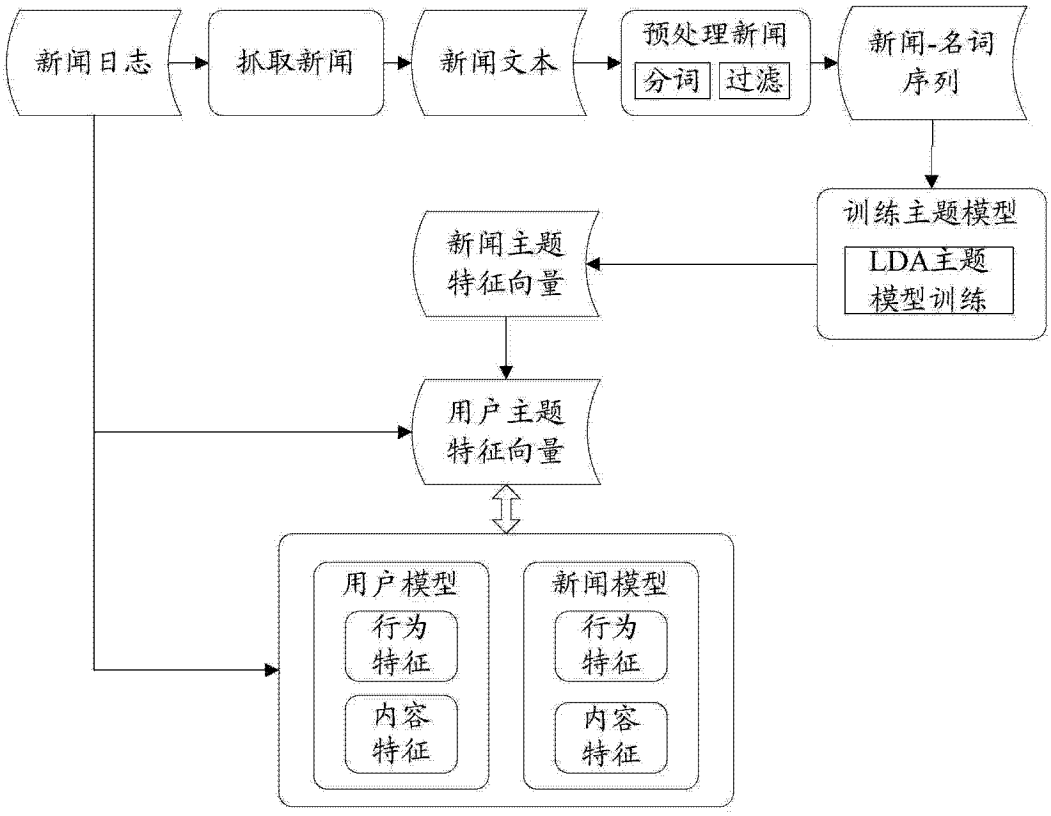

新闻日志
抓取新闻
新闻文本
预处理新闻
分词
过滤
新闻-名词序列
训练主题模型
LDA主题模型训练
新闻主题特征向量
用户主题特征向量
用户模型
行为特征
内容特征
新闻模型
行为特征
内容特征

图 3

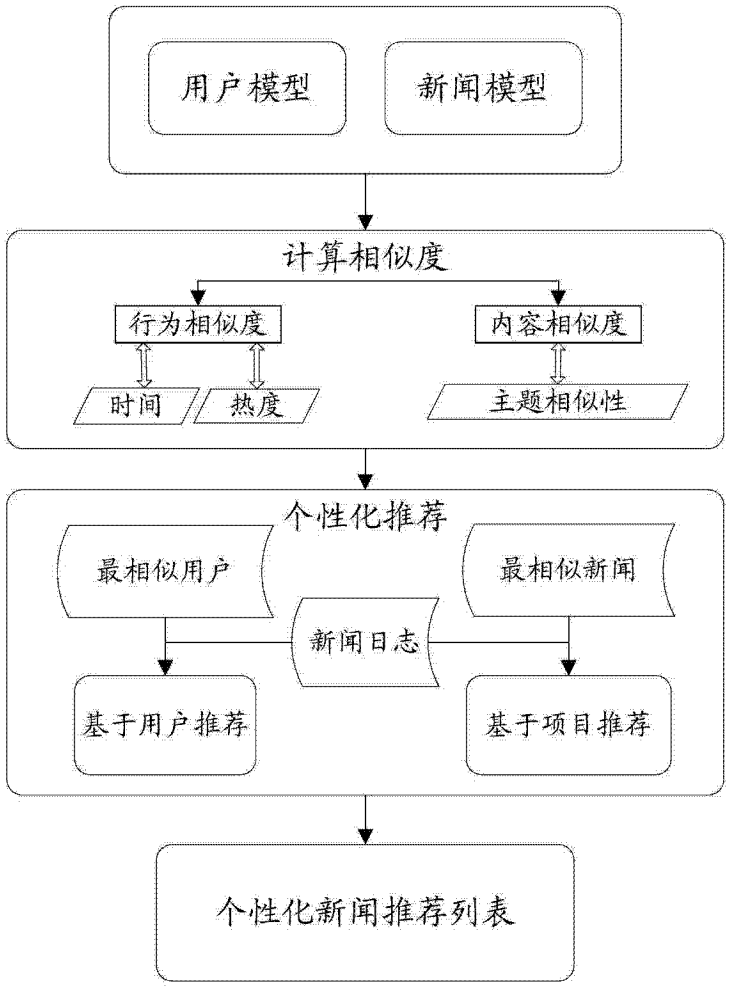

用户模型
新闻模型
计算相似度
行为相似度
内容相似度
时间
热度
主题相似性
个性化推荐
最相似用户
最相似新闻
新闻日志
基于用户推荐
基于项目推荐
个性化新闻推荐列表

图 4