(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 102880728 B
(45)授权公告日 2015.10.28

(21)申请号 201210427389.9

(22)申请日 2012.10.31

(73)专利权人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路95号

(72)发明人 徐常胜 邓拯宇

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 曹玲柱

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

(56)对比文件
CN 101477554 A,2009.07.08, 全文.
CN 101719145 A,2010.06.02, 全文.
CN 101901249 A,2010.12.01, 全文.
US 20110191339 A1,2011.08.04, 全文.

审查员 王秋平

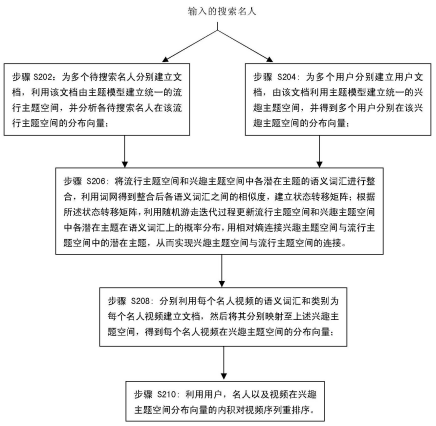
权利要求书2页 说明书7页 附图2页

(54)发明名称

名人视频搜索结果个性化排序的方法

(57)摘要

本发明提供了一种名人视频搜索结果个性化排序的方法。本方法同时考虑了用户和待搜索名人特点，在不同数据集上分析用户的兴趣分布和待搜索名人的流行分布，有效地表达了用户的兴趣主题和待搜索名人的流行主题，并对用户兴趣主题和名人流行主题进行有效关联，从而提高了个性化排序的准确性。



CN 102880728 B

1. 一种名人视频搜索结果个性化排序的方法，其特征在于，包括：

为多个预设待搜索名人分别建立文档，利用该文档由主题模型建立统一的流行主题空间，并分析各待搜索名人在该流行主题空间的分布向量；

利用用户与互联网的在线交互记录建立用户文档，由多个用户文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；

将流行主题空间和兴趣主题空间中各潜在主题的语义词汇进行整合，利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵；根据所述状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在整合后各语义词汇上的概率分布，用相对熵连接兴趣主题空间与流行主题空间中的潜在主题；

分别利用每个待搜索名人视频的语义词汇和类别为每个待搜索名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个待搜索名人视频在兴趣主题空间的分布向量；以及

利用用户，待搜索名人以及待搜索名人视频在兴趣主题空间分布向量的内积对视频序列重排序；

其中，所述为多个预设待搜索名人分别建立文档的步骤包括：收集整理多个待搜索名人分别的词条信息；利用词网过滤上述多个待搜索名人词条信息中的噪声，滤除所述多个待搜索名人词条信息除名词成分之外的其他成分；对于多个待搜索名人中的每一个，利用其对应的词条信息的名词成分建立待搜索名人文档；

其中，所述利用用户与互联网的在线交互记录建立用户文档的步骤包括：收集多个用户分别上传或收藏的互联网资源的语义词汇和类别；利用词网过滤上述语义词汇和类别中的噪声，滤除所述语义词汇和类别中除名词成分之外的其他成分；对于多个用户中的每一个，利用所述语义词汇和类别中的名词成分建立用户文档。

2. 根据权利要求 1 所述的方法，其特征在于，利用文档由潜在狄利克利分布模型建立统一的流行主题空间。

3. 根据权利要求 1 所述的方法，其特征在于，所述待搜索名人的词条信息取自于维基百科。

4. 根据权利要求 1 所述的方法，其特征在于，所述利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵的步骤中：

对于一个给定的包含 N 个语义词汇的语义词汇网络，每一个语义词汇被看成一个结点；状态转移矩阵用 P(N×N) 表示，该状态转移矩阵的元素 p_{ij} 表示从结点 i 到结点 j 的转移概率：

$$p_{ij} = s_{ij} / \Sigma_k s_{ik}$$

其中，s_{ij} 表示语义词汇 i 和 j 之间的语义相似度。

5. 根据权利要求 4 所述的方法，其特征在于，所述根据状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在语义词汇上的概率分布的步骤中，每一潜在主题随机游走的迭代公式为：

$$r_k = \lambda P t_{k-1} + (1-\lambda) y$$

其中，r_k 和 r_{k-1} 是两个列向量，分别表示某潜在主题各结点在随机游走过程中第 k 和 k-1 次迭代时的概率值，P 为状态转移矩阵，$\lambda \in (0,1)$ 是权重参数，y 是该潜在主题在语

义词汇上的初始概率分布。

6. 根据权利要求 5 所述的方法，其特征在于，所述用相对熵连接兴趣主题空间与流行主题空间中的潜在主题的步骤中，相对熵表示为：

$$D_{KL}(z \,||\, x) = \frac{1}{2}\left(\sum_i z(i) ln \frac{z(i)}{x(i)} + \sum_i x(i) ln \frac{x(i)}{z(i)}\right)$$

其中，主题 z 和主题 x 分别来自兴趣主题空间和流行主题空间，z(i) 和 x(i) 表示主题 z 和主题 x 在语义词汇 i 上的概率值，主题 z 和主题 x 的相似度即为相对熵的倒数。

7. 根据权利要求 5 所述的方法，其特征在于，所述分别利用每个待搜索名人视频的语义词汇和类别为每个待搜索名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个待搜索名人视频在兴趣主题空间的分布向量的步骤中：

Φ 是一个 K×M 维的马尔可夫矩阵，其中，K 是兴趣主题空间潜在主题个数，M 是整合后语义词汇的个数；Φ 中的每一行表示某一主题在语义词汇上的概率分布，对于任一视频向量 $v_{M\times1}$，投影到兴趣主题空间后的分布向量为 $v'_{K\times1} = \Phi v_{M\times1}$。

8. 根据权利要求 6 所述的方法，其特征在于，所述利用用户，待搜索名人以及待搜索名人视频在兴趣主题空间分布向量的内积对视频序列重排序的步骤包括：

得到初始视频序列；

把与待搜索名人相关的视频分别映射到兴趣主题空间；

根据兴趣主题空间与流行主题空间的关联度对初始序列重排序。

9. 根据权利要求 8 所述的方法，其特征在于，所述兴趣主题空间与流行主题空间的关联度：

$$p(score\,|v,u,c)$$

$$=\sum_{i=1}^{K} P(z_i|v)p(z_i|u)p(z_i|c)$$

$$=\sum_{i=1}^{K} P(z_i|v)p(z_i|u)\sum_{j=1}^{L} P(x_j|c)p(z_i|x_j)$$

其中：K 是兴趣主题空间潜在主题个数，z_i是兴趣主题空间第 i 个潜在主题；

L 是流行主题空间潜在主题个数，x_j是流行主题空间第 j 个潜在主题；

$p(z_i|v)$ 和 $p(z_i|u)$ 分别表示视频 v 和用户 u 在主题 z_i上的概率；$p(z_i|x_j)$ 由相对熵计算得到；$\sum_{j=1}^{L} P(x_j|c)p(z_i|x_j)$间接表示名人 c 在主题 z_i上的概率。

10. 根据权利要求 1 至 9 中任一项所述的方法，其特征在于，所述的待搜索名人为在某一群体、某一领域内具有高知名度的人。

名人视频搜索结果个性化排序的方法

技术领域

[0001] 本发明涉及互联网搜索引擎技术领域，尤其涉及一种名人视频搜索结果个性化排序的方法。

背景技术

[0002] 随着WEB2.0的到来，在线视频的传播已经达到了前所未有的水平。虽然如此海量的视频数据能满足几乎所有用户的需求，但同时也使得搜寻和查找到用户真正感兴趣的视频成为了一件非常烦琐的事情。尽管搜索引擎已经成为了用户广泛使用的工具，但很少有搜索引擎能满足用户的个性化需求。往往对于同一个查询词，不同的用户表达的意思不尽相同。因此，个性化搜索对于信息爆炸的当今是非常必要的。

[0003] 在巨大的视频库中，有很大一部分是与名人相关的视频，由于“名人效应”，这些视频受到了广大用户的关注。传统搜索引擎根据视频与查询的相关性来排序。当用户搜索某一名人，搜索引擎通常返回一个包含各类视频的列表。其中，可能仅仅只有某一类视频是用户感兴趣的。在现有的搜索个性化排序方法中，一些研究者采用聚类算法辅助个性化搜索。比如，有人把社会语义词汇聚类成一些概念，然后通过这些概念连接用户和对象（视频、图像或文本等）。还有一些人采用概念或本体的层次集合，其中概念或本体的每一个结点都表示某一兴趣。进一步，有些研究者利用主题模型分析用户的兴趣主题分布。

[0004] 图1为现有技术进行名人视频搜索结果个性化排序的流程图。如图1所示，现有技术名人搜索结果个性化排序的流程包括：

[0005] 步骤S102，为多个用户分别建立用户文档，由该文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；

[0006] 步骤S104，分别利用每个名人视频的语义词汇和类别为每个名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个名人视频在兴趣主题空间的分布向量；

[0007] 步骤S106，利用用户和视频在兴趣主题空间分布向量的匹配程度对视频序列重排序。

[0008] 发明人发现上述名人视频搜索结果个性化排序的方法存在如下技术缺陷：

[0009] 1）只考虑了用户的兴趣分布，而没有考虑搜索对象（名人）的流行分布，个性化排序准确性差；

[0010] 2）建立兴趣主题空间时，没到考虑兴趣主题空间中语义词汇之间的相关性，影响了兴趣主题空间的准确表达。

发明内容

[0011] （一）要解决的技术问题

[0012] 为解决上述的一个或多个问题，本发明提供了一种名人视频搜索结果个性化排序的方法，以提高个性化排序的准确性。

[0013] （二）技术方案

[0014] 根据本发明的一个方面，提供了一种名人视频搜索结果个性化排序的方法。该方法包括：为多个预设待搜索名人分别建立文档，利用该文档由主题模型建立统一的流行主题空间，并分析各待搜索名人在该流行主题空间的分布向量；利用用户与互联网的在线交互记录建立用户文档，由多个用户文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；将流行主题空间和兴趣主题空间中各潜在主题的语义词汇进行整合，利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵；根据所述状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在整合后各语义词汇上的概率分布，用相对熵连接兴趣主题空间与流行主题空间中的潜在主题；分别利用每个待搜索名人视频的语义词汇和类别为每个待搜索名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个待搜索名人视频在兴趣主题空间的分布向量；以及利用用户，待搜索名人以及视频在兴趣主题空间分布向量的内积对视频序列重排序。

[0015] （三）有益效果

[0016] 从上述技术方案可以看出，本发明名人视频搜索结果个性化排序的方法具有以下有益效果：

[0017] (1) 同时考虑了用户和待搜索名人特点，在不同数据集上分析用户的兴趣分布和待搜索名人的流行分布，有效地表达了用户兴趣主题和待搜索名人的流行主题，从而提高了个性化排序的准确性；

[0018] (2) 利用随机游走迭代过程加强流行主题空间和兴趣主题空间中各潜在主题的语义词汇之间的关联，提高了流行主题空间和兴趣主题空间中各潜在主题的准确性；同时，随机游走过程使得流行主题空间和兴趣主题空间中各潜在主题的概率分布遍布整个词汇集，从而可以有效关联兴趣主题空间和流行主题空间。

附图说明

[0019] 图 1 为现有技术利用传统方法进行互联网搜索结果个性化排序的流程图；

[0020] 图 2 为本发明实施例名人视频搜索结果个性化排序方法的流程图。

具体实施方式

[0021] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明进一步详细说明。

[0022] 需要说明的是，在附图或说明书描述中，相似或相同的部分都使用相同的图号。且在附图中，以简化或是方便标示。再者，附图中未绘示或描述的实现方式，为所属技术领域中普通技术人员所知的形式。另外，虽然本文可提供包含特定值的参数的示范，但应了解，参数无需确切等于相应的值，而是可在可接受的误差容限或设计约束内近似于相应的值。

[0023] 本发明的目的是实现名人个性化搜索。该问题存在如下挑战。首先，我们通常不易知道名人活跃在那个领域；另外，由于隐私问题，用户也很少明确表达自己的兴趣分布；最后，用户兴趣主题空间与名人流行主题空间基于不同的数据集，两个空间不存在显示相关性，如何对这两个空间进行有效关联也是一个难点。

[0024] 在本发明的一个示例性实施例中，提出了一种名人视频搜索结果个性化排序的方

法。图 2 为本发明实施例名人视频搜索结果个性化排序方法的流程图。如图 2 所示，本实施例包括：

[0025] 步骤 S202，为多个待搜索名人分别建立文档，利用该文档由主题模型建立统一的流行主题空间，并分析各待搜索名人在该流行主题空间的分布向量；

[0026] 通常情况下，在互联网上进行名人搜索的对象为通常所说的“名人”，此处的名人，为在某一群体、某一领域内具有较高知名度的人，如克林顿、成吉思汗、耶稣、贝克汉姆、张靓颖等。

[0027] 上述为搜索的特定名人建立文档，可以是搜索引擎提供商编辑的文档，也可以是利用互联网上的与该特定名人相关的已有文档，例如维基百科、百度百科或搜狗百科等网站提供的该特定名人的词条。考虑到公正性和权威性，优选采用维基百科的相关词条。

[0028] 一般来说，名人都有维基文档，这些文档全面介绍他们的职业、成就和生活等方面，从而反映了名人的流行分布。但是这类文本中含有大量噪声和没有实际意义的词汇，不利于名人流行主题分布的表达。为了解决这个问题，我们利用词网来过滤这些信息，并且只保留了名词性成分，因为名词具有最强的语义性。此处，词网即 WorldNet，是一种包含了语义信息的字典。WorldNet 根据词条的意义将它们分组，并为每一个同义词集合提供了简短，概要的定义，并记录不同集合之间的语义关系。

[0029] 基于上述讨论，本步骤又可以分为以下子步骤：

[0030] 步骤 S202a，收集整理多个名人分别的维基百科词条信息；

[0031] 步骤 S202b，利用词网过滤上述多个名人词条信息中的噪声，滤除所述多个名人词条信息除名词成分之外的其他成分；

[0032] 步骤 S202c，对于多个名人中的每一个，利用其对应的名人词条信息的名词成分建立名人文档；

[0033] 步骤 S202d，利用多个名人文档由主题模型建立流行主题空间，并得到每个名人在该流行主题空间的分布向量。

[0034] 上述主题模型可以选择潜在狄利克利分布模型 (LDA)，当然也可以选择本领域公知的其他模型，例如：概率潜在语义分析模型 (PLSA) 或关联主题模型 (CTM)。

[0035] 步骤 S204，利用用户与互联网的在线交互记录建立用户文档，由多个用户文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；

[0036] 用户对某一视频的主动行为（如上传或收藏）反映了用户的兴趣爱好。因此我们利用用户上传或收藏视频的语义词汇和类别来建立用户文档。但是这类语义词汇通常由网络用户提供，含有大量噪声，如无意义的词汇及误输入。这里我们同样采用词网进行过滤，具体步骤与 S202 类似，可以分为以下子步骤：

[0037] 步骤 S204a，收集多个用户分别上传或收藏的互联网资源的语义词汇和类别；

[0038] 步骤 S204b，利用词网过滤上述语义词汇和类别中的噪声，滤除所述语义词汇和类别中除名词成分之外的其他成分；

[0039] 步骤 S204c，对于多个用户中的每一个，利用所述语义词汇和类别中的名词成分建立用户文档。

[0040] 步骤 S204d，由多个用户文档利用潜在狄利克利分布主题模型建立兴趣主题空间，

并得到多个用户分别在该兴趣主题空间的分布向量。

[0041] 经过步骤 S202 和 S204，分别获得了用户兴趣主题空间与名人流行主题空间，接下来通过连接这两个主题空间的潜在主题来关联兴趣主题空间与流行主题空间。

[0042] 步骤 S206，将流行主题空间和兴趣主题空间中各潜在主题的语义词汇进行整合，利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵；根据所述状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在语义词汇上的概率分布，用相对熵 (Relative Entropy) 连接兴趣主题空间与流行主题空间中的潜在主题，从而实现兴趣主题空间与流行主题空间的连接。

[0043] 由于流行主题空间和兴趣主题空间分别来自不同的数据集，因此他们的词汇表（空间中所有词汇的集合）是不一致的，换句话说，具有相似意义的主题在不同空间所包含的词汇是不一致的。

[0044] 因此，通过把流行主题空间和兴趣主题空间中各潜在主题的语义词汇进行整合，并利用词网得到词汇之间的语义相关性，建立状态转移矩阵，然后采用随机游走迭代过程更新每个主题在所有词汇上的概率分布，使每个主题的概率分布拓展到整个融合后的词汇集，此时便可计算各个主题之间的相对熵，从而连接兴趣主题空间与流行主题空间。

[0045] 本步骤 S206 中“利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵”具体包括：

[0046] 使用 s_{ij} 表示语义词汇 i 和 j 之间的语义相似度。对于一个给定的包含 N 个语义词汇的语义词汇网络。每一个语义词汇被看成一个结点。状态转移矩阵用 P(N×N) 表示。该状态转移矩阵的元素 p_{ij} 表示从结点 i 到结点 j 的转移概率，即语义词汇 i 和 j 的相似度。

[0047] $p_{ij} = s_{ij} / \sum_k s_{ik}$ (2)

[0048] 本步骤 S206 中“根据所述状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在语义词汇上的概率分布”具体包括：

[0049] 用 $r_k(i)$ 表示结点 i 在随机游走迭代过程中第 k 次迭代时的概率值，那么，所有结点的概率值形成一个列向量 $r_k = [r_k(i)]N\times 1$。因此，随机游走迭代过程的表达式为

[0050] $r_k = \lambda P r_{k-1} + (1-\lambda) y$ (3)

[0051] 其中，r_k 和 r_{k-1} 是两个列向量，分别表示某潜在主题各结点在随机游走过程中第 k 和 k-1 次迭代时的概率值，y 是潜在主题在语义词汇上的初始概率分布，$\lambda \in (0,1)$ 是权重参数。λ 越大则随机游走迭代过程的作用越强。随机游走迭代过程使得相似的语义词汇有相近的概率分布，同时使得近义词越多的词汇得到更多的强化。随机游走迭代过程使得每个潜在主题的概率分布拓展到整个融合后的词汇集。

[0052] 本步骤 S206 中“用相对熵 (Relative Entropy) 连接兴趣主题空间与流行主题空间中的潜在主题”具体包括：

[0053] 采用计算兴趣主题和流行主题之间的相对熵。因为相对熵是与方向有关的，所以，采用两个方向的平均相对熵。假定主题 z 和主题 x 分别来自兴趣主题空间和流行主题空间。相对熵表示为

[0054] $$D_{KL}(z \,||\, x) = \frac{1}{2}\left(\sum_i z(i) ln \frac{z(i)}{x(i)} + \sum_i x(i) ln \frac{x(i)}{z(i)}\right) \quad (4)$$

[0055] 其中 z(i) 和 x(i) 表示主题 z 和主题 x 在语义词汇 i 上的概率值。主题 z 和主题

x 的相似度即为相对熵的倒数。

[0056] 经过步骤 S206，我们实现了潜在语义主题层面关联用户与名人，从而提高了个性化排序的准确性。

[0057] 步骤 S208：分别利用每个名人视频的语义词汇和类别为每个名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个名人视频在兴趣主题空间的分布向量；

[0058] 具体地说，假定 Φ 是一个 K×M 维的马尔可夫矩阵。其中，K 是兴趣主题空间潜在主题个数，M 是语义词典的维数，Φ 中每一行表示某一主题在语义词汇上的概率分布。对于任一视频向量 $v_{M\times1}$，投影到兴趣主题空间后的分布向量为 $v'_{k\times1}=\Phi v_{M\times1}$。

[0059] 步骤 S210：利用用户，名人以及视频在兴趣主题空间分布向量的内积对视频序列重排序。

[0060] 给定任一用户（用 u 表示），当该用户搜索某位名人（用 c 表示），我们首先从传统搜索引擎得到初始视频序列。然后把与名人相关的视频（初始视频序列的前 N 个视频）分别映射到兴趣主题空间。然后我们根据兴趣主题空间与流行主题空间的关联度对初始序列重排序，具体步骤如下：

[0061] 对于任一名人视频 v，他与某一用户的相关性得分由该名人和该用户及该视频在兴趣主题空间的分布向量共同决定，具体表达式如下：

$$p(score\,|v,u,c) \tag{5}$$

[0062]

$$=\sum_{i=1}^{K}P(z_i|v)p(z_i|u)p(z_i|c)$$

$$=\sum_{i=1}^{K}P(z_i|v)p(z_i|u)\sum_{j=1}^{L}P(x_j|c)p(z_i|x_j)$$

[0063] 其中：K 是兴趣主题空间潜在主题个数，z_i 是兴趣主题空间第 i 个潜在主题；

[0064] L 是流行主题空间潜在主题个数，x_j 是流行主题空间第 j 个潜在主题；

[0065] $p(z_i|v)$ 和 $p(z_i|u)$ 分别表示视频 v 和用户 u 在主题 z_i 上的概率；$p(z_i|x_j)$ 由相对熵近似（如公式 4）计算得到。$\sum_{j=1}^{L}P(x_j|c)p(z_i|x_j)$ 间接表示名人 c 在主题 z_i 上的概率。该公式表明，我们计算视频得分时，不仅考虑视频与搜索词的相似性，还考虑用户本身的兴趣分布。为每个视频重新计算他们与用户的相关性得分后，我们再根据这一得分调整视频序列，返回给该用户。

[0066] 为了便于理解，以下以一具体的搜索结果为例进行说明，例如，特定用户 A 对特定名人“贝克汉姆”进行搜索，具体步骤如下：

[0067] 我们首先 1) 利用维基百科对多个名人分别建立文档；利用多个用户分别与互联网资源的在线交互记录建立用户文档。其中名人“贝克汉姆”($d_{贝克汉姆}$)，“用户 A”($d_{用户A}$) 及部分其他名人文档 ($d_{嘎嘎女士}$，$d_{罗伯茨}$) 与用户文档 ($d_{用户B}$，$d_{用户C}$) 示意如下。

[0068] $d_{贝克汉姆}$ = {season united league club match real cup final team player premier young goals madrid goal champions scoring players youth england president title scored number shirt played football injury competition games london reached transfer barcelona matches family company produced night featured number work late school took age received california father appearance

appeared working interview…}

[0069] $d_{嘎嘎女士}$ = {album music released song songs performed country tour records billboard concert band chart artist musical awards sold albums record live hit solo debut award recorded release grammy rock copies october pop performing singles studio dance fame addition nominations fusari monster born critically worldwide countries art sgband judas creative tried positive radio starlight…}

[0070] $d_{罗伯茨}$ = {film role starred appeared played movie character television award star comedy cast series films performance drama acting office reviews success supporting production opposite box actress episode festival adaptation roles september years york announced february american november april world series…}

[0071] ……

[0072] $d_{用户A}$ = {robin gary norris baba comedy bob pack soccer football salem engineer training free real pitch goal retard film driver limo battle swerve mike kick curve festival technique madrid jimmy perfect rock tutorial drunk corner casino martin stockbroker hotel league crazy blue porn crone gymnastics riley shot iris dice news manchester nike penny…}

[0073] $d_{用户B}$ = {gaming music play wedding quality dream nancy drew definition song academy screen viva description hq princess album filmanimation disney knowledge white real studio game firefly story official vision coliseum capsule mac beauty voyage soundtrack vega monster version secret edition slot…}

[0074] $d_{用户C}$ = {filmanimation comedy bang theory merchant raj book leonard night talk super future penny diary list animal bucket idiot host italia animation funny interview work ice question world television opening twins stupid humor roads episode head headache guinness plumbing coming sky spot office guest strike warwick…}

[0075] ……

[0076] 2) 然后利用潜在狄利克利分布主题模型建立流行主题空间和兴趣主题空间，并得到所有名人分别在流行主题空间的分布向量及所有用户分别在兴趣主题空间的分布向量。

[0077] 3) 利用词网得到各语义词汇之间的相似度，建立状态转移矩阵 $p_{N\times N}$。然后利用随机游走迭代过程更新各潜在主题在语义词汇上的概率分布，最后用相对熵连接兴趣主题空间与流行主题空间中的潜在主题，从而实现兴趣主题空间与流行主题空间的连接。

[0078] $$P_{N\times N}=\begin{pmatrix}0.00123,0.00015,0.00000,0.00025,0.00016,0.00006\ldots\\0.00031,0.00249,0.00000,0.00000,0.00000,0.00011\ldots\\0.00000,0.00000,1.00000,0.00000,0.00000,0.00000\ldots\\0.00046,0.00000,0.00000,0.00226,0.00028,0.00000\ldots\\0.00027,0.00000,0.00000,0.00025,0.00207,0.00000\ldots\\0.00012,0.00010,0.00000,0.00000,0.00000,0.00231\ldots\\0.00024,0.00000,0.00000,0.00023,0.00025,0.00000\ldots\\0.00045,0.00000,0.00000,0.00043,0.00028,0.00000\ldots\\\ldots\ldots\end{pmatrix}$$

[0079] 4) 分别利用每个名人视频的语义词汇和类别为每个名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个名人视频在兴趣主题空间的分布向量。视频文档示例如下：

[0080] d_v= {trailer, teaser, prelude, new, video, marry, night, born, way, preview, Lady, Marry, Night, Mother, Monster, Little, Monsters}

[0081] 5) 利用用户，名人以及视频在兴趣主题空间分布向量的内积对视频序列重排序。

[0082] 为了评估本发明，我们从福布斯 (Forbes) 得到最受欢迎和最具影响力且活跃在多领域的 106 位名人。同时，我们从视频分享网站 YouTube 采集了 143 位用户。每个用户都上传或收藏过一定量的视频，并且这些视频中的某些视频与上述 106 位名人中的某一位相关。我们假设用户 u 上传或收藏的视频中含有与名人 c 相关的视频。实验中，我们假定用户 u 对名人 c 进行搜索，然后统计用户 u 上传或收藏的视频中与名人 c 相关的视频在返回视频序列中的数目。为了评价我们的发明的性能，我们比较了 1) 非个性化搜索方法，2) 传统方法。性能评价方法是 F 值（一种搜索的测量方式，同时考虑了准确度与召回率，其中准确度是指返回结果中正确结果所占的比例，召回率是指返回结果中正确结果占所有正确结果的比例）。

[0083] 分析实验结果我们发现，本发明的方法要明显好其他两种方法。如返回序列前 20 视频的平均 F 值，我们的方法是 0.4262，传统方法为 0.2696，而非个性化方法只有 0.0456。

[0084] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

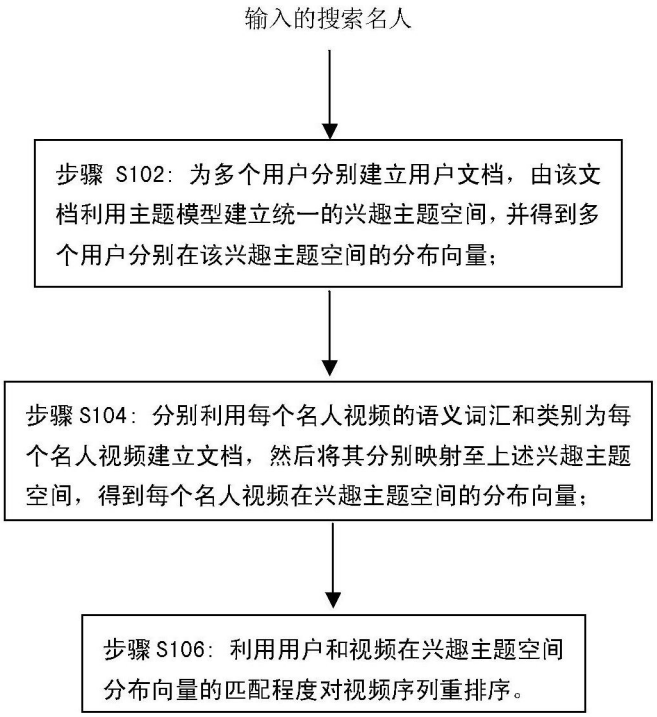

输入的搜索名人
步骤 S102：为多个用户分别建立用户文档，由该文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；
步骤 S104：分别利用每个名人视频的语义词汇和类别为每个名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个名人视频在兴趣主题空间的分布向量；
步骤 S106：利用用户和视频在兴趣主题空间分布向量的匹配程度对视频序列重排序。

图 1

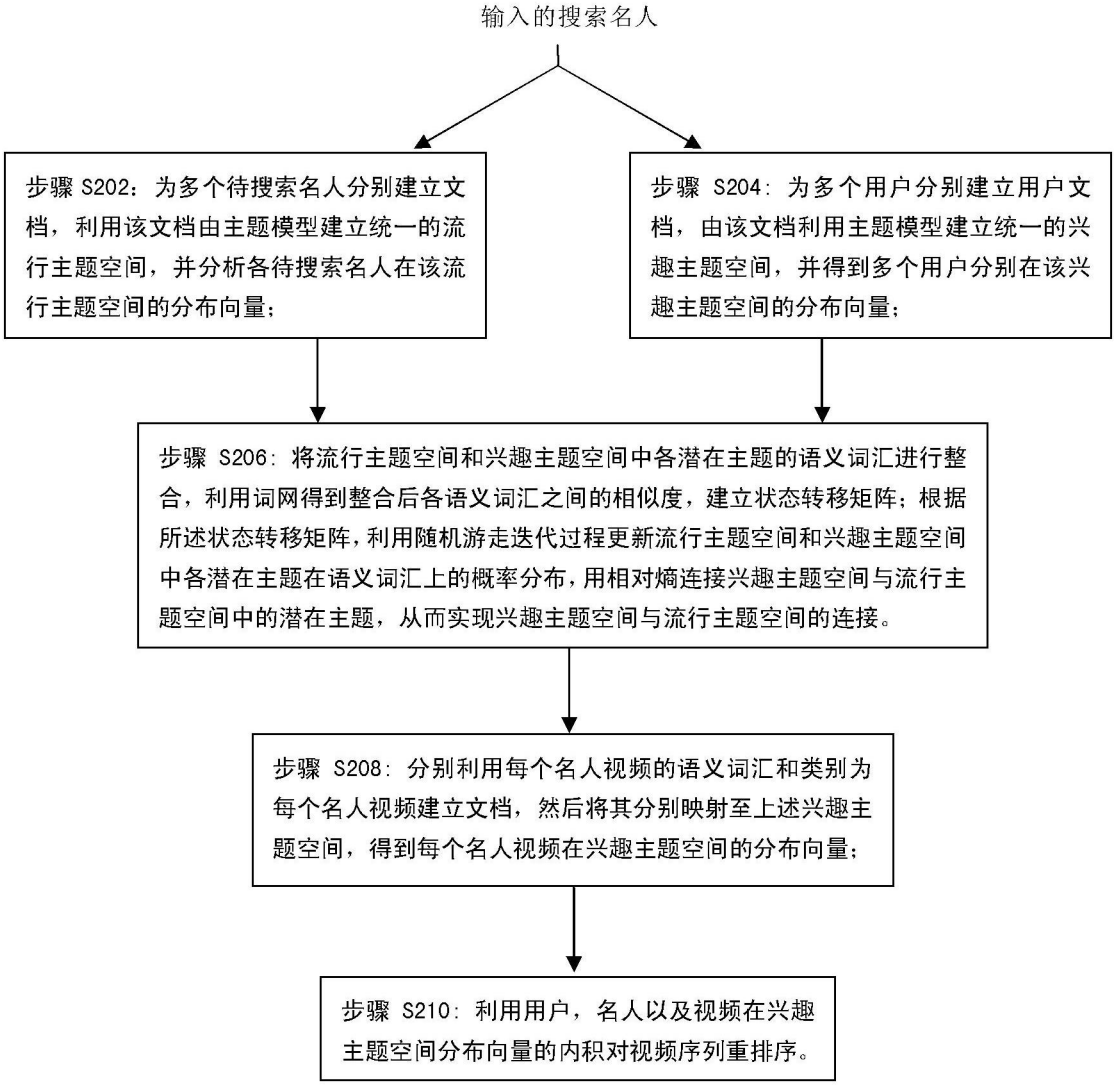

输入的搜索名人
步骤 S202：为多个待搜索名人分别建立文档，利用该文档由主题模型建立统一的流行主题空间，并分析各待搜索名人在该流行主题空间的分布向量；
步骤 S204：为多个用户分别建立用户文档，由该文档利用主题模型建立统一的兴趣主题空间，并得到多个用户分别在该兴趣主题空间的分布向量；
步骤 S206：将流行主题空间和兴趣主题空间中各潜在主题的语义词汇进行整合，利用词网得到整合后各语义词汇之间的相似度，建立状态转移矩阵；根据所述状态转移矩阵，利用随机游走迭代过程更新流行主题空间和兴趣主题空间中各潜在主题在语义词汇上的概率分布，用相对熵连接兴趣主题空间与流行主题空间中的潜在主题，从而实现兴趣主题空间与流行主题空间的连接。
步骤 S208：分别利用每个名人视频的语义词汇和类别为每个名人视频建立文档，然后将其分别映射至上述兴趣主题空间，得到每个名人视频在兴趣主题空间的分布向量；
步骤 S210：利用用户，名人以及视频在兴趣主题空间分布向量的内积对视频序列重排序。

图 2