(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 103744928 B
(45)授权公告日 2017.10.03

(21)申请号 201310743880.7

(22)申请日 2013.12.30

(65)同一申请的已公布的文献号
申请公布号 CN 103744928 A

(43)申请公布日 2014.04.23

(73)专利权人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号

(72)发明人 宿红毅 朱叶 王彩群 闫波 郑宏

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

(56)对比文件
CN 102421025 A,2012.04.18,
CN 102088626 A,2011.06.08,
US 2002078070 A1,2002.06.20,
US 2013332835 A1,2013.12.12,
张龙飞等.基于支持向量机元分类器的体育视频分类.《北京理工大学学报》.2006,第26卷(第1期),第41-44页.

审查员 田志方

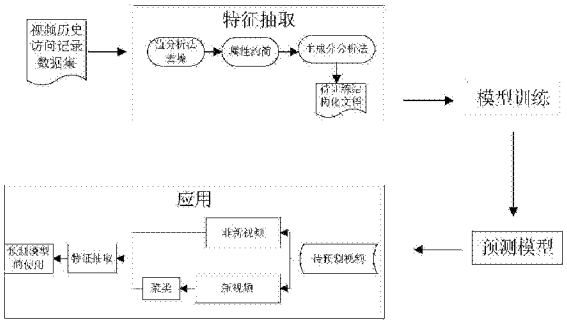
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种基于历史访问记录的网络视频分类方法

(57)摘要

本发明涉及一种基于历史访问记录的网络视频分类方法，属于计算机网络数据挖掘技术领域。首先通过对视频的历史访问记录数据集进行自动分析，抽取出有意义的特征后对其生成待用数据文件，通过所述数据文件将历史访问记录转化为可用于训练的结构化文档，然后用逻辑回归对所结构化文档进行机器学习得到预测模型。使用预测模型，根据待预测视频历史访问记录信息的完整程度，对其选用相应的方法进行分类预测。本发明对比现有技术，在减少人工代价的同时，使参与计算的参数更为精简，预测效果更为准确、花费的时间更少。同时，由于可以根据待预测视频历史访问记录信息的完整程度对其选择聚类与否的操作，使其模型的应用更为广泛。



CN 103744928 B

1.一种基于历史访问记录的网络视频分类方法，其特征在于，包括以下步骤：

步骤一、对视频历史访问记录数据集进行分析，抽取出属性特征并生成待用数据文件，通过所述待用数据文件将视频历史访问记录转化为待训练结构化文档；具体过程如下：

首先，对视频历史访问记录数据集，利用值分析方法去掉取值不正常的数据和属性，包括取值无变化的属性、缺失的或者噪音的数据以及去除播放次数小于某一阀值的视频记录，得到数据集U；

然后，利用基于互信息增益率的启发式属性约简算法，训练对数据集U的属性集进行约简；约简由核开始，逐步选择Z(c,R,D)达到最大的属性加入，直到所选择的属性子集分类能力与整个属性集的分类能力相同时结束，具体步骤如下：

第一步，将预测系统S定义为一个四元组：S＝(U,A,V,f)，其中U＝{u_1,u_2,…,u_n}是视频对象集，即论域；A是视频的属性集合；$V=\bigcup_{a\in A}V_a$为属性值的集合，V_a为属性a的值域；f是U×A→V_a的映射，它为U中各视频对象的属性指定唯一值；

对于预测系统S，将属性集合A分为条件属性集C和决策属性集D，A＝C∪D，且C∩D＝φ，其中属性集C中包含的元素有视频ID c_1、标题c_2、类型c_3、时长等级c_4、URL c_5、URL信誉度c_6、播放次数c_{10}、评论次数c_{12}、分享次数c_{15}，收藏次数c_{16}，下载次数c_{17}，分享率c_{18}，收藏率c_{19}，下载率c_{20}，点赞率c_{21}，播放次数增长率c_{22}，好评率c_{23}，时间戳c_{24}，被观看时长c_{25}，被观看时长占的比率c_{26}；决策属性集合D包括受欢迎与否d；将该做了上述变化的预测系统S命名为决策系统L；由于在S中，对于属性集$G\subseteq A$，构造对应的二元等价关系，当I_G＝{(x,y)∈U×U；$\forall a\in G$，有a(x)＝a(y)}，称I_G为由G构造的不可分辨关系，则对决策系统L＝(U,C∪D,V,f)，设$R\subset C$，I_R和I_D导出的划分分别为X＝{X_1,X_2,…X_n}和Y＝{Y_1,Y_2,…Y_n}，则R的熵定义为$H(R)=-\sum_{i=1}^{n}p(X_i)\lg p(X_i)$，其中p($X_i$)＝card($X_i$)/card(U)；R相对D的条件熵定义为$H(D/R)=-\sum_{i=1}^{n}p(X_i)\sum_{j=1}^{m}p(Y_j/X_i)\lg p(Y_j/X_i)$，其中p($Y_j$/$X_i$)＝card($Y_j$∩$X_i$)/card($X_i$)；决策属性集D和条件属性子集R的互信息定义为：W(R;D)＝H(D)-H(D/R)，属性重要性的度量方法定义为：Z(c,R,D)＝(W(R∪{c};D)-W(R;D))/H(c)，其中$H(c)=-\sum_{i=1}^{m}p_i\lg p_i$，$p_i$是属性取值为$x_i$的对象的个数占总对象数N的比例，设属性c有m种取值x_1,x_2.…,x_m，N为总对象数；

第二步，计算条件属性集C和决策属性集D的互信息W(C;D)＝H(D)-H(D/C)；

第三步，计算核R＝$CORE_D$(C)，并计算W(R;D)，其中核的计算过程为：

设$CORE_D$(C)＝φ，对于条件属性集C中的所有属性r，如果H({d}/C)<H({d}/C-{r})，则$CORE_D$(C)＝$CORE_D$(C)∪{r}；

第四步，令$C_{candidate}$＝C-R，按Z(c,R,D)＝(W(R∪{c};D)-W(R;D))/H(c)计算$C_{candidate}$中各属性的重要性，并选择Z(c,R,D)达到最大的属性c_i；

第五步，令R＝R∪{c_i}，若W(C;D)＝W(R;D)，则终止，并将约简后的属性集所对应的数据集用U′表示；否则转第四步继续执行；

之后，对数据集U′进行主成分分析，得到彼此不相关的若干个主成分，具体步骤如下：

第一步，对数据集U′进行Z标准化，得到数据集U″；

第二步，对数据集U″进行主成分分析，得出各主成分的特征值、方差贡献率及累计方差贡献率，其中，对各个主成分的特征值按由大到小的方式进行排序；根据主成分累计方差贡献率大于85％的个数来确定主成分的个数k，根据主成分分析时得到的因子荷载表，写出k个主成分与数据集U″中的各个属性之间的关系式，如下所示，其中Z_k代表着第k个主成分，β_{km}代表着Z_k的第m个因子载荷，c_m为数据集U″中的第m个属性的值，$c_m \in$｛视频ID c_1、标题c_2、类型c_3、时长等级c_4、URLc_5、URL信誉度c_6、播放次数c_{10}、评论次数c_{12}、点赞率c_{21}、分享率c_{18}，收藏率c_{19}，播放次数增长率c_{22}，被观看时长占的比率c_{26}｝：

$$\begin{cases} Z_1 = \beta_{11}c_1 + \beta_{12}c_2 + \cdots + \beta_{1m}c_m \\ Z_2 = \beta_{21}c_1 + \beta_{22}c_2 + \cdots + \beta_{2m}c_m \\ \vdots \quad\quad\quad \vdots \\ \vdots \quad\quad\quad \vdots \\ Z_k = \beta_{k1}c_1 + \beta_{k2}c_2 + \cdots + \beta_{km}c_m \end{cases}$$

步骤二、用逻辑回归方法，对所述结构化文档进行机器学习，得到预测模型，具体过程如下：

对步骤二得到的各主成分值进行二元逻辑回归分析，得出逻辑回归模型：

$$p = \frac{e^{\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \cdots + \alpha_k z_k}}{1 + e^{\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \cdots + \alpha_k z_k}}$$

其中$\alpha_1, \alpha_2, \cdots, \alpha_k$为预测模型经过训练后得到的参数，P的值越接近于1，说明待分类视频越受欢迎，P越接近于0，说明待分类视频为越不受欢迎，若$p \geq 0.5$，则待分类视频为受欢迎视频；若$p < 0.5$，则待分类视频为不受欢迎视频；

步骤三、使用上述预测模型对视频进行欢迎与否的测试，具体过程如下：

首先，判断视频历史访问记录的信息完整性，如果待预测视频是新视频，即该视频的历史访问记录不存在，根据视频的特征信息计算tf-idf值，用tf-idf矩阵作为聚类模型的输入，得到新视频的最相似视频，并将其历史访问记录信息设为新视频的历史访问记录信息；如果待预测视频不是新视频，直接进行下一步；

然后，对待预测视频的历史访问记录数据进行相应的转化，即进行特征抽取；

最后，使用预测模型对其进行欢迎与否的分类。

一种基于历史访问记录的网络视频分类方法

技术领域

[0001] 本发明涉及一种网络视频分类方法，属于计算机网络数据挖掘技术领域。

背景技术

[0002] 随着数据库技术的迅速发展、数据库管理系统的广泛应用和Internet的迅速普及，互联网上的视频(以下简称视频)历史访问记录数据量急剧增长。激增的数据后面蕴涵着大量的“宝藏”，即事先未知而潜在有用的信息。面对大规模的海量数据，数据挖掘技术应运而生，从大量的、不完全的、有噪声的、模糊的、随机的、实际应用的数据中提取隐含在其中的、人们不知道的但又有用的信息和知识的过程。

[0003] 数据挖掘的任务主要有分类、预测、关联分析、时序模式、聚类、偏差检测等。每种问题都有许多具体的数据挖掘或统计模型来加以解决。

[0004] 其中，分类是根据数据集的特点构造一个分类器，利用分类器对未知类别的样本赋予类别的一种技术。构造分类器的过程一般分为模型训练和使用模型分类两个步骤。在模型训练阶段，分析训练数据集的特点，为每个类别产生一个对相应数据集的准确描述或模型。在模型使用阶段，根据待分类对象的数据描述信息，利用模型对其进行分类。

[0005] 分类算法主要包括神经网络方法、决策树分类法、统计方法等。其中，统计方法主要有回归和朴素贝叶斯分类算法。回归分类包括一般的线形回归和Logist回归(或称为逻辑回归)，都是将数据分为两类。普通的Logist回归是用事件发生的概率与不发生该事件的概率之比来进行分类的，对于多分类问题则会采取Logist回归的一种自然扩展Logit回归。目前，应用最为广泛的是基于逻辑回归的预测方法：通过对数据集进行分析、建模，对待分类的对象进行二分类预测。然而，数据集中的知识(属性)并不是同等重要的，还存在冗余，这不利于做出正确而简洁的决策。而较优的数据集拥有如下指标：个数较少；属性的规则数目较少；最终范化规则数目较少等。但是，现存的基于逻辑回归的预测方法在数据集的精简方面都存在一些局限性，如只对属性的重要度进行排序而忽略了取值的离散分布、没有考虑属性之间的相关性等等。

发明内容

[0006] 本发明的目的是为了克服当前基于逻辑回归的预测方法在数据集精简方面所存在的不足，提出一种基于历史访问记录的网络视频分类方法。

[0007] 本发明所述方法在保持知识库的分类和决策能力不变的条件下，通过对数据集特征的抽取过程进行优化，删除不相关或不重要的属性，避免了变量之间所反映信息的重叠，从而使数据集达到了最为精简，并减少了人工代价。由于参与计算的参数更为精简，使预测效果更为准确、时间效果更为提升。本方法简单、易行，适合目前广泛流行的分布式计算应用。

[0008] 本发明所述方法包括以下步骤：

[0009] 步骤一、建立预测模型

[0010] 首先，进行特征抽取。通过对视频的历史访问记录数据集进行自动分析，抽取出最精简的属性特征后生成待用数据文件，通过所述数据文件将历史访问记录数据集转化为可用于训练的结构化文档。

[0011] 然后，进行模型训练。采用逻辑回归方法，对所述结构化文档进行机器学习，得到预测模型。

[0012] 步骤二、采用预测模型，对视频进行欢迎程度预测

[0013] 首先，判断视频历史访问记录的信息完整性。如果视频是新视频，即历史访问记录信息不完整，则使用聚类方法找到与其相似性最高的视频，将其历史访问记录信息设为新视频的历史访问记录信息。如果视频不是新视频，即历史访问记录信息完整，直接进行下面的操作。

[0014] 然后，对待预测视频的历史访问记录信息进行特征抽取，使用预测模型对其进行欢迎与否的分类。

[0015] 有益效果

[0016] 本发明采用基于历史访问记录的网络视频分类方法，对视频的欢迎与否进行预测。通过对视频的历史访问记录数据集进行特征抽取等属性约简，进而建立相应的预测模型。完整的历史访问记录分析方法，在减少人工代价的同时，使参与计算的参数更为精简，预测效果更为准确、花费的时间更少。同时，由于可以根据待预测视频历史访问记录信息的完整程度对其选择聚类与否的操作，使其模型的应用更为广泛。

附图说明

[0017] 图1为本发明方法的流程图。

具体实施方式

[0018] 下面结合附图及实施例对本发明的具体实施方式做进一步详细说明。

[0019] 如图1所述，一种基于历史访问记录的网络视频分类方法，包括以下步骤：

[0020] 步骤一、对视频历史访问记录数据集进行分析，抽取出最精简的属性特征并生成待用数据文件。通过所述待用数据文件将视频历史访问记录转化为待训练结构化文档。具体过程如下：

[0021] 首先，对视频历史访问记录数据集，利用值分析方法去掉取值不正常的数据和属性，包括取值无变化的属性、缺失的或者噪音的数据以及去除播放次数小于某一阀值的视频记录，得到数据集U。

[0022] 然后，利用基于互信息增益率的启发式属性约简算法，训练对数据集U的属性集进行约简。约简由核开始，逐步选择Z(c,R,D)达到最大的属性加入，直到所选择的属性子集分类能力与整个属性集的分类能力相同时结束。具体步骤如下：

[0023] 第一步，将预测系统S定义为一个四元组：S＝(U,A,V,f)，其中U＝{$u_1,u_2,\cdots,u_n$}是视频对象集，即论域；A是视频的属性集合；$V=\bigcup_{a\in A}V_a$为属性值的集合，V_a为属性a的值域；f是U×A→V_a的映射，它为U中各视频对象的属性指定唯一值。

[0024] 对于预测系统S，将属性集合A分为条件属性集C和决策属性集D，A＝C∪D，且C∩D

＝φ，其中属性集C中包含的元素有视频IDc1、标题c2、类型c3、时长等级c4、URLc5、URL信誉度c6、播放次数c10、评论次数c12、分享次数c15，收藏次数c16，下载次数c17，分享率c18，收藏率c19，下载率c20，点赞率c21，播放次数增长率c22，好评率c23，时间戳c24，被观看时长c25，被观看时长占的比率c26；决策属性集合D包括受欢迎与否d。将该做了上述变化的预测系统S命名为决策系统L。由于在S中，对于属性集$G \subseteq A$，可构造对应的二元等价关系，当$I_G = \{(x,y) \in U \times U; \forall a \in G, 有 a(x) = a(y)\}$，称$I_G$为由G构造的不可分辨关系。则对决策系统L＝(U,C∪D,V,f)，设$R \subset C$，I_R和I_D导出的划分分别为X＝{X_1,X_2,…X_n}和Y＝{Y_1,Y_2,…Y_n}，则R的熵定义为$H(R) = -\sum_{i=1}^{n} p(X_i) \lg p(X_i)$，其中p($X_i$)＝card($X_i$)/card(U)。R相对D的条件熵定义为$H(D/R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j/X_i) \lg p(Y_j/X_i)$，其中p($Y_j$/$X_i$)＝card($Y_j$∩$X_i$)/card($X_i$)。决策属性集D和条件属性子集R的互信息定义为：W(R;D)＝H(D)-H(D/R)，属性重要性的度量方法定义为：Z(c,R,D)＝(W(R∪{c};D)-W(R;D))/H(c)，其中$H(c) = -\sum_{i=1}^{m} p_i \lg p_i$，$p_i$是属性取值为$x_i$的对象的个数占总对象数N的比例，设属性c有m种取值x_1,x_2.…,x_m,N为总对象数。

[0025] 第二步，计算条件属性集C和决策属性集D的互信息W(C;D)＝H(D)-H(D/C)；

[0026] 第三步，计算核R＝$CORE_D$(C)，并计算W(R;D)。其中核的计算过程为：

[0027] 1.设$CORE_D$(C)＝φ；

[0028] 2.对于条件属性集C中的所有属性r，如果H({d}/C)<H({d}/C-{r})，则

[0029] $CORE_D$(C)＝$CORE_D$(C)∪{r}.

[0030] 3.结束。

[0031] 第四步，令$C_{candidate}$＝C-R，按Z(c,R,D)＝(W(R∪{c};D)-W(R;D))/H(c)计算$C_{candidate}$中各属性的重要性，并选择Z(c,R,D)达到最大的属性c_i；

[0032] 第五步，令R＝R∪{c_i}，若W(C;D)＝W(R;D)，则终止，并将约简后的属性集所对应的数据集用U′表示；否则转第四步继续执行。

[0033] 之后，对数据集U′进行主成分分析，得到彼此不相关的若干个主成分。具体步骤如下：

[0034] 第一步，对数据集U′进行Z标准化，得到数据集U″；

[0035] 第二步，对数据集U″进行主成分分析，得出各主成分的特征值、方差贡献率及累计方差贡献率，其中，对各个主成分的特征值按由大到小的方式进行排序。根据主成分累计方差贡献率大于85％的个数来确定主成分的个数k，根据主成分分析时得到的因子荷载表，写出k个主成分与数据集U″中的各个属性之间的关系式，如下所示，其中Z_k代表着第k个主成分，β_{km}代表着Z_k的第m个因子载荷，c_m为数据集U″中的第m个属性的值，c_m∈{视频IDc1、标题c2、类型c3、时长等级c4、URLc5、URL信誉度c6、播放次数c10、评论次数c12、点赞率c21、分享率c18，收藏率c19，播放次数增长率c22，被观看时长占的比率c26}：

[0036]
$$\begin{cases} Z_1 = \beta_{11}c_1 + \beta_{12}c_2 + \cdots + \beta_{1m}c_m \\ Z_2 = \beta_{21}c_1 + \beta_{22}c_2 + \cdots + \beta_{2m}c_m \\ \vdots \quad\quad\quad \vdots \\ \vdots \quad\quad\quad \vdots \\ Z_k = \beta_{k1}c_1 + \beta_{k2}c_2 + \cdots + \beta_{km}c_m \end{cases}$$

[0037] 步骤二、用逻辑回归方法，对所述结构化文档进行机器学习，得到预测模型。具体过程如下：

[0038] 对步骤二得到的各主成分值进行二元逻辑回归分析，得出逻辑回归模型：

[0039]
$$p = \frac{e^{\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \cdots + \alpha_k z_k}}{1 + e^{\alpha_1 z_1 + \alpha_2 z_2 + \alpha_3 z_3 + \cdots + \alpha_k z_k}}$$

[0040] 其中$\alpha_1,\alpha_2,\cdots,\alpha_k$为预测模型经过训练后得到的参数，P的值越接近于1，说明待分类视频越受欢迎，P越接近于0，说明待分类视频为越不受欢迎，若$p \geq 0.5$，则待分类视频为受欢迎视频；若$p < 0.5$，则待分类视频为不受欢迎视频；

[0041] 步骤三、使用预测模型对视频进行欢迎与否的测试，具体过程如下：

[0042] 首先判断视频历史访问记录的信息完整性。如果待预测视频是新视频，即该视频的历史访问记录不存在，但是其自身的一些特征信息是有的，比如视频ID、查询ID、视频的标题、描述、关键词等等，根据视频的特征信息计算tf-idf值，用tf-idf矩阵作为聚类模型的输入。运用tf-idf便能从文本的内容上进行聚类，得到新视频的最相似视频，并将其历史访问记录信息设为新视频的历史访问记录信息；如果待预测视频不是新视频，直接进行下一步。

[0043] 然后对待预测视频的历史访问记录数据进行相应的转化，即进行特征抽取。

[0044] 最后使用预测模型对其进行欢迎与否的分类。

[0045] 实施例

[0046] 本发明方法包括三阶段，第一阶段为对视频的历史访问记录进行特征抽取阶段，第二阶段为预测模型的训练阶段，第三阶段为待分类视频欢迎与否的预测阶段。

[0047] 参阅图1，下面详细叙述本实施例第一阶段的具体过程：

[0048] 步骤1：根据视频的历史访问记录数据量大小，去除播放次数小于某一阀值的视频访问记录。具体地，根据对一些数据集的分析，这些历史访问记录在一定程度上都服从长尾效应，即包含许多点击次数不够多的视频记录，所以处理的第一步，应该设定Q为阀值，移除点击次数低于此阀值的视频记录。然后去掉一些取值无变化的属性列，从而得到初步输入数据集U；

[0049] 步骤2：对数据集U的属性集进行约简，约简由核开始，逐步选择重要的属性加入，直到所选择的属性子集分类能力与整个属性集U的分类能力相同时结束。具体地，经过步骤1的初步筛选后，初步得到的输入数据集中条件属性集合＝{视频IDc_1、标题c_2、类型c_3、时长等级c_4、URLc_5、URL信誉度c_6、视频上传者IDc_7、上传者粉丝级别c_8、上传时间c_9、播放次数c_{10}、不同IP地址观看人数c_{11}、评论次数c_{12}、好评数c_{13}、视频画质c_{14}、分享次数c_{15}，收藏次数c_{16}，下载次数c_{17}，分享率c_{18}，收藏率c_{19}，下载率c_{20}，点赞率c_{21}，播放次数增长率c_{22}，好评率c_{23}，时间戳c_{24}，被观看时长c_{25}，被观看时长占的比率c_{26}}。先计算条件属性C与决策属性D的互信息

W(C,D)＝0.283，以及相对核属性$K_D(C)$＝{视频ID、URL信誉度、上传者粉丝级别、播放次数、评论次数}，然后分别计算剩余属性的重要性分别为

[0050] $Z(c_{21},R,D)=(W(R\cup\{c_{21}\};D)-W(R;D))/H(c_{21})=0.2182$，

[0051] $Z(c_9,R,D)=(W(R\cup\{c_9\};D)-W(R;D))/H(c_9)=0.2180$，

[0052] $Z(c_4,R,D)=(W(R\cup\{c_4\};D)-W(R;D))/H(c_4)=0.2160$，

[0053] …，

[0054] $Z(c_{14},R,D)=(W(R\cup\{c_{14}\};D)-W(R;D))/H(c_{14})=0.0110$，由重要度的高低次序，将属性加入条件属性集合-得到C′＝{视频IDc_1、标题c_2、类型c_3、时长等级c_4、URLc_5、URL信誉度c_6、视频上传者IDc_7、上传者粉丝级别c_8、上传时间c_9、播放次数c_{10}、评论次数c_{12}、分享次数c_{15}，收藏次数c_{16}，分享率c_{18}，收藏率c_{19}，点赞率c_{21}，播放次数增长率c_{22}，被观看时长占的比率c_{26}}；

[0055] 步骤3：对条件属性集合C′进行主成分分析，得到彼此不相关的若干个主成分。具体步骤如下：

[0056] i)对条件属性集合C′对应的数据集U′进行Z标准化得到数据集U″；

[0057] ii)对数据集U″进行主成分分析，求出各主成分的特征值(由大到小的方式进行排序)、方差贡献率及累计方差贡献率，根据主成分累计方差贡献率大于85％的个数来确定主成分的个数k，根据主成分分析时得到的因子荷载表，写出k个主成分与条件属性集合C′中的各个属性之间的关系式，如：

[0058]
$$\begin{cases} Z_1 = \beta_{11}c_1 + \beta_{12}c_2 + \cdots + \beta_{1m}c_m \\ Z_2 = \beta_{21}c_1 + \beta_{22}c_2 + \cdots + \beta_{2m}c_m \\ \vdots \quad\quad\quad \vdots \\ \vdots \quad\quad\quad \vdots \\ Z_k = \beta_{k1}c_1 + \beta_{k2}c_2 + \cdots + \beta_{km}c_m \end{cases}$$

[0059] 以上步骤1-3为本实施例第一阶段的特征抽取阶段的具体过程，得到了结构化的文档用作后续模型训练的输入。

[0060] 在第一阶段后，进入第二阶段，即模型训练阶段，此阶段用逻辑回归对第一阶段到的结构化文档进行机器学习，得到预测模型。

[0061] 在众多机器学习算法中，逻辑回归是一种高效又表现理想的算法。逻辑回归会充分用到所有的特征练预测模型，如得出的逻辑回归模型：

[0062]
$$p=\frac{e^{\alpha_1 z_1+\alpha_2 z_2+\alpha_3 z_3+\cdots+\alpha_k z_k}}{1+e^{\alpha_1 z_1+\alpha_2 z_2+\alpha_3 z_3+\cdots+\alpha_k z_k}}$$

[0063] 第三阶段为视频欢迎与否的预测阶段，具体包括以下阶段：

[0064] 步骤1：判断待预测视频的描述信息的完整性；

[0065] 步骤2：若待预测视频不是新视频，即有一定的历史访问记录数据，则对其数据进行特征值的抽取，转化成结构化文档形式，而后代入预测模型进行欢迎与否预测；

[0066] 步骤3：若待预测视频是新视频，使用聚类找到与其相似性最高的视频，并将新的描述信息设为待预测视频的描述信息，然后进其进行相应的预测操作；

[0067] 具体的，将如何预测出新视频欢迎与否的问题转换成找到与此视频最相似的集

合，即转换成了聚类问题。

[0068] 本发明针对待预测视频的条件属性计算tf-idf值，用tf-idf矩阵作为聚类模型的输入，运用tf-idf能从数据集的内容上进行聚类，以此方法计算出的相似性更为准确。

[0069] 由此，本实施例通过三个步骤的处理，得到了对新视频欢迎与否的预测，使得视频能够得到更准确的预测，和更精准的投放。

[0070] 以上所述的具体实例是对本发明的进一步解释说明，并不用于限定本发明的保护范围，凡在本发明原则和精神之内，所做的更改和等同替换都应是本发明的保护范围之内。

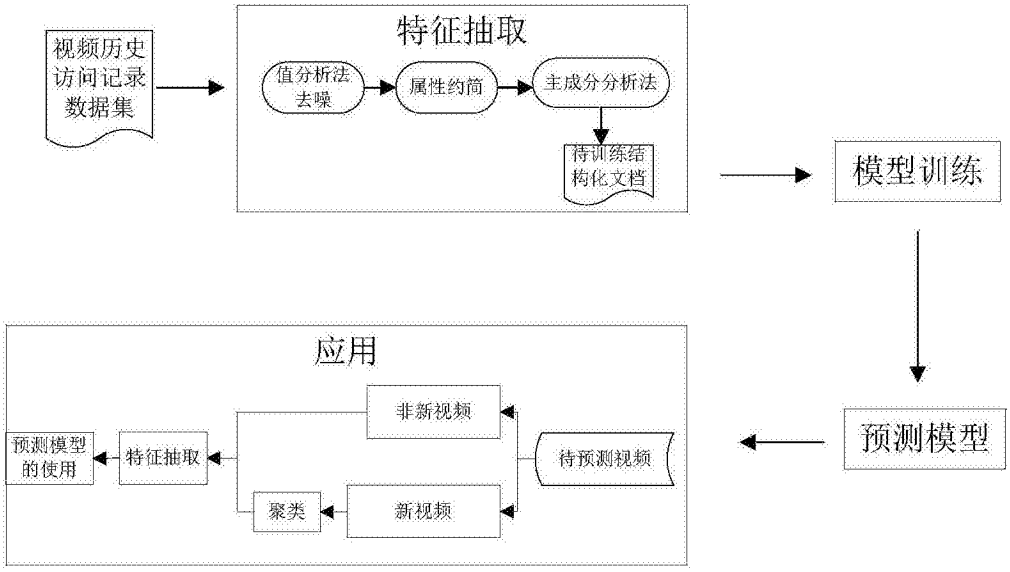

视频历史访问记录数据集
特征抽取
值分析法去噪
属性约简
主成分分析法
待训练结构化文档
模型训练
预测模型
应用
非新视频
预测模型的使用
特征抽取
待预测视频
聚类
新视频

图1