[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/60

[12] 发明专利申请公开说明书

[21] 申请号 02822478.7

[43] 公开日 2005 年 2 月 23 日

[11] 公开号 CN 1585954A

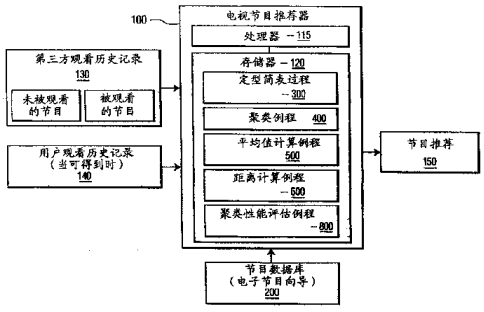
[22] 申请日 2002.10.22 [21] 申请号 02822478.7
[30] 优先权
[32] 2001.11.13 [33] US [31] 10/014,180
[86] 国际申请 PCT/IB2002/004413 2002.10.22
[87] 国际公布 WO2003/042879 英 2003.5.22
[85] 进入国家阶段日期 2004.5.12
[71] 申请人 皇家飞利浦电子股份有限公司
地址 荷兰艾恩德霍芬
[72] 发明人 S·V·R·古特塔 K·库拉帕蒂
[74] 专利代理机构 中国专利代理(香港)有限公司
代理人 程天正 王 勇

权利要求书 2 页 说明书 14 页 附图 8 页

[54] 发明名称 在项目推荐器中评估这些项的接近度的方法及装置

[57] 摘要

公开了一种在可得到用户的观看历史记录或购买历史记录之前，向用户推荐感兴趣的项目，诸如电视节目推荐的方法及装置。 处理第三方观看或购买历史记录以生成反映由有代表性的观众所选定的典型项目样式的定型简表。 用户能够从所生成的定型简表中选择最相关的定型，从而用最接近他或她自己的兴趣的项目来初始化他或她的简表。聚类例程把第三方观看或购买历史记录(数据集)划分成群，以使得在一个群内的点(例如，电视节目)比其它任何群更接近该群的平均值。 距离计算例程根据给定的电视节目与给定的群的平均值之间的距离来评估电视节目与各个群的接近度。



ISSN1008-4274

1. 一种用在推荐器（100）内用于评估两个项（205，210，220）的接近度的方法，所述项（205，210，220）中的每一项都用至少一个符号特征来表征，所述方法包括步骤：

根据对于所述符号特征值的每个可能值的所有例子的总的分类相似性来计算所述两个项（205，210，220）的相应的符号特征值之间的距离；以及

合计各个所述符号特征值之间的距离以确定所述两个项（205，210，220）的接近度。

2. 权利要求 1 的方法，被安排用于把一项（205，210，220）分配给一组项或多组项，所述项（205，210，220）中的每一项都用至少一个符号特征来表征，所述方法包括步骤：

计算所述项的相应的符号特征值与各个所述组内的至少一项之间的距离，所述距离是基于对于所述符号特征值的每个可能值的所有例子的总的分类相似性；

合计各个所述特征值之间的距离以确定所述项与各个所述组内的至少一项之间的接近度；以及

把所述项分配给与最小距离值有关的所述组。

3. 权利要求 1 或 2 的方法，其中所述计算步骤采用值差量度（VDM）技术来计算所述符号特征之间的距离。

4. 权利要求 1 或 2 的方法，其中所述计算步骤采用修改的值差量度（VDM）技术来计算符号特征之间的所述距离。

5. 权利要求 1 或 2 的方法，其中用下式给出对于一特定符号特征的两个值，V1 与 V2 之间的所述距离 δ：

$$\delta(V1, V2) = \sum | C1i/C1 - C2i/C2 |^r$$

其中 C1i 是 V1 被分到类别 i 内的次数，而 C1 是 V1 出现于数据集内的总次数。

6. 权利要求 1 或 2 的方法，其中所述项（205，210，220）为节目，感兴趣的类别为“被观看”和“未被观看”，并用下式给出对于一特定符号特征的两个值，V1 与 V2 之间的所述距离 δ：

$$\delta(V1, V2)=\left|\frac{C1\_watched}{C1\_total}-\frac{C2\_watched}{C2\_total}\right|+$$

$$\left|\frac{C1\_not\_watched}{C1\_total}-\frac{C2\_not\_watched}{C2\_total}\right|$$

其中C1i是V1被分到类别i内的次数，而C1_total是V1出现于数据集内的总次数。

7. 权利要求1或2的方法，其中所述项（205，210，220）中的一项为群平均值。

8. 权利要求1或2的方法，其中所述项（205，210，220）为节目。

9. 权利要求1或2的方法，其中所述项（205，210，220）为内容。

10. 权利要求1或2的方法，其中所述项（205，210，220）为产品。

11. 一种用在推荐器（100）内的用于评估两个项（205，210，220）的接近度的系统（100），所述项（205，210，220）中的每一项都用至少一个符号特征来表征，所述系统包括：

用于根据对于所述符号特征值的每个可能值的所有例子的总的分类相似性来计算所述两个项（205，210，220）的相应的符号特征值之间的距离的装置；以及

用于合计各个所述符号特征值之间的距离以确定所述两个项（205，210，220）的接近度的装置。

12. 权利要求11的系统，进一步包括：

存储器（120），用于存储计算机可读代码；以及

处理器（115），可操作地耦合于所述存储器（120），所述处理器（115）被配置成：

根据对于所述符号特征值的每个可能值的所有例子的总的分类相似性来计算所述两个项（205，210，220）的相应的符号特征值之间的距离；以及

合计各个所述符号特征值之间的距离以确定所述两个项（205，210，220）的接近度。

13. 一种计算机程序产品，使可编程的设备在当执行所述计算机程序产品时起到如权利要求11所限定的系统的作用。

在项目推荐器中评估这些项的接近度的方法及装置

本发明与名为“Method and Apparatus for Partitioning a Plurality of Items into Groups of Similar Items in a Recommender of Such Items”（代理人案卷号为US010568）的美国专利申请，名为“Method and Apparatus for Generating A Stereotypical Profile for Recommending Items of Interest Using Item-Based Clustering”（代理人案卷号为US010569）的美国专利申请，名为“Method and Apparatus for Recommending Items of Interest Based on Preferences of a Selected Third Party”（代理人案卷号为US010572）的美国专利申请，名为“Method and Apparatus for Recommending Items of Interest Based on Stereotype Preferences of Third Parties”（代理人案卷号为US010575）的美国专利申请，以及名为“Method and Apparatus for Generating A Stereotypical Profile for Recommending Items of Interest Using Feature-Based Clustering”（代理人案卷号为US010576）的美国专利申请相关，每一件均同此发明同时申请，均被转让给本发明的受让人，并在此并入作为参考。

本发明涉及用于推荐感兴趣的项，诸如电视节目的方法及装置，并且更具体地，涉及用于在可得到用户的购买或观看历史记录之前推荐感兴趣的节目或其它项的技术。

随着电视观众可用的频道数目增多，以及存在于这些频道上的节目的多样性，对于电视观众来说，识别感兴趣的电视节目已经日益变得复杂。电子节目向导（EPG）通过例如，名称、时间、日期以及频道来识别有用的电视节目，以及通过允许依照个性化的偏好搜寻或分类有用的电视节目来方便对感兴趣节目的识别。

许多推荐工具已经被计划或建议用来推荐感兴趣的电视节目或其它项目。电视节目推荐工具例如将观众偏好应用于 EPG 以得到一组对于一特定观众可能是感兴趣的推荐节目。一般地，电视节目推荐工具使用隐含的或明显的技术，或是使用上述技术的一些组合来获得观众的偏好。隐含的电视节目推荐工具以不强迫别人接受的方式，根据从

观众的观看历史记录得到的信息生成电视节目推荐。另一方面，明显的电视节目推荐工具明确地询问观众有关他们对于节目属性，诸如名称、类型、演员、频道以及日期/时间的偏好，以得出观众简表并生成推荐。

虽然当前可用的推荐工具协助用户来识别感兴趣的项目，但是它们也受到许多限制，如果它们克服了这些限制，将会很大地改进这些推荐工具的便利性和性能。例如，为了成为综合性质的，明显的推荐工具的初始化非常冗长，需要每一个新用户回答有关在粗粒度级上指定他们的偏好的非常详细的调查。虽然隐含的电视节目推荐工具通过观察观看行为而不引人注目地得到一个简表，但是它们需要长的时间来变得准确。另外，这些隐含的电视节目推荐工具至少需要一最小数量的观看历史记录以便开始做出任何推荐。因此，在当第一次获得推荐工具时，这些隐含的电视节目推荐工具并不能够做出任何推荐。

因此，需要一种能够在可得到足够的个性化的观看历史记录之前，不引人注意地推荐诸如电视节目的项目的方法及装置。另外，需要一种根据第三方的观看习惯来为一给定的用户生成节目推荐的方法及装置。

总体上，公开了一种向用户推荐感兴趣的项目，诸如电视节目推荐的方法及装置。根据本发明的一个方面，在可得到用户的观看历史记录或购买历史记录之前-诸如在当用户第一次获得推荐器时生成推荐。最初，采用来自一个或多个第三方的观看历史记录或购买历史记录来向特定用户推荐感兴趣的项目。

处理第三方观看或购买历史记录以生成反映由有代表性的观众所选定项目的典型样式的定型简表（stereotype profile）。每个定型简表都是在某些方面彼此相类似的项目（数据点）的一个群（claster）。用户选择感兴趣的定型以便用最接近他或她自己的兴趣的项目来初始化他或她的简表。

聚类例程把第三方观看或购买历史记录（数据集）划分成群，以使得在一个群内的点（例如，电视节目）比其它任何群更接近该群的平均值（mean）。还公开了用于计算一个群的符号平均值的平均值计算例程。利用各个群的平均值，根据数据点至各个群之间的距离来把诸如电视节目的给定数据点分配给群。

所公开的距离计算例程根据给定电视节目与给定群的平均值之间的距离来评估电视节目与各个群的接近度。计算出的距离量度量化在样本数据集内的各种例子之间的差别以确定一个群的范围。采用值差量度（VDM）技术或其变更来计算两个电视节目之间的特征值之间的距离。根据已知的修改的 VDM（MVDM）技术，用下式给出对于特定特征的两个值之间的距离 δ：

$$\delta(V1, V2) = \sum | C1i/C1 - C2i/C2 |^r$$

其中 V1 和 V2 是对于考虑中的特征的两个可能的值。在说明性实施例的节目推荐环境中，感兴趣的类别为“被观看”和“未被观看”。一般地，所公开的距离计算例程在如果这些值对于所有的分类都以相同的相对频率出现时，就将这些值看作是相似的。

通过参照下面的详细描述以及附图将获得对本发明以及本发明的进一步的特征和优点的更完全理解。

图 1 是本发明的电视节目推荐器的示意框图；

图 2 是取自图 1 的示例性节目数据库的样本表；

图 3 是描述具体化本发明原理的图 1 的定型简表处理的流程图；

图 4 是描述具体化本发明原理的图 1 的聚类例程的流程图；

图 5 是描述具体化本发明原理的图 1 的平均值计算例程的流程图；

图 6 是描述具体化本发明原理的图 1 的距离计算例程的流程图；

图 7A 是取自示例性频道特征值出现表的一样本表，该示例性频道特征值出现表表示对于各个类别的各个频道特征值的出现数目；

图 7B 是取自示例性特征值对距离表的一样本表，该示例性特征值对距离表表示从图 7A 所示的示例性计数计算出的各个特征值对之间的距离；以及

图 8 是描述具体化本发明原理的图 1 的聚类性能评估例程的流程图。

图 1 说明了本发明的电视节目推荐器 100。如图 1 所示，该示例性电视节目推荐器 100 评估如在下面结合图 2 所论述的节目数据库 200 内的节目以识别特定观众感兴趣的节目。能够例如使用采用众所周知的屏上呈现技术的顶置终端/电视（未示出）来把一组推荐节目呈现给

观众。虽然这里是在电视节目推荐的上下文中说明了本发明，但是本发明能够应用于任何根据用户行为，诸如观看历史记录或购买历史记录而自动生成的推荐。

根据本发明的一个特征，电视节目推荐器 100 能够在用户的观看历史记录 140 可得到之前，诸如当用户第一次得到该电视节目推荐器 100 时生成电视节目推荐。如图 1 所示，电视节目推荐器 100 最初采用来自一个或多个第三方的观看历史记录 130 来推荐特定用户感兴趣的节目。一般地，该第三方观看历史记录 130 是基于具有代表大量人数的人口统计状况，诸如年龄、收入、性别及教育的一个或多个采样人数的观看习惯。

如图 1 所示，第三方观看历史记录 130 由一组被给定人数观看以及未被给定人数观看的节目组成。通过观察被该给定人数实际观看的节目来获得被观看的该组节目。通过例如随机采样节目数据库 200 内的节目来获得未被观看的该组节目。在一进一步的变更中，根据序列号为 No. 09/819，286、申请日为 2001 年 3 月 28 日、名称为“An Adaptive Sampling Technique for Selecting Negative Examples for Artifical Intelligence Applications”的美国专利申请的教导来获得未被观看的该组节目，该篇申请被转让给本发明的受让人并在此并入作为参考。

根据本发明的另一个特征，电视节目推荐器 100 处理第三方观看历史记录 130 以生成反映由有代表性的观众所观看的电视节目的典型样式的定型简表。如下面进一步论述的，定型简表是在某些方面彼此相似的电视节目（数据点）的群。因而，一给定的群对应于取自展示特定样式的第三方观看历史记录 130 的一特殊片段的电视节目。

根据本发明来处理第三方观看历史记录 130 以提供展示某些特定样式的节目群。此后，用户能够选择最相关的定型并因此用与他或她自己的兴趣最接近的节目来初始化他或她的简表。然后根据每个单独用户他们自己的记录样式以及给予节目的反馈，该定型的简表调整并向每个单独用户的特定的、个人观看行为发展。在一实施例中，当确定节目得分时，可以对取自用户自己的观看历史记录 140 的节目比取自第三方观看历史记录 130 的节目给予更高的加权。

电视节目推荐器 100 可以具体化为任何计算设备，诸如个人计算

机或工作站，其含有诸如中央处理单元（CPU）的处理器115，以及诸如RAM和/或ROM的存储器120。电视节目推荐器100还可以具体化为例如在顶置终端或显示器（未示出）内的专用集成电路（ASIC）。另外，电视节目推荐器100可以具体化为任何可得到的电视节目推荐器，诸如从加利福尼亚桑尼维尔的Tivo有限公司商业地可购买到的Tivo™系统，或者是在序列号为No. 09/466，406、申请日为1999年12月17日、名称为"Method and Apparatus for Recommending Television Programming Using Decision Trees"的美国专利申请，序列号为No. 09/498，271、申请日为2000年2月4日、名称为"Bayesian TV Show Recommender"的美国专利申请，以及序列号为No. 09/627，139、申请日为2000年7月27日、名称为"Three-Way Media Recommendation Method and System"的美国专利申请，或它们的任何组合中描述的电视节目推荐器，每一种都在这里被并入作为参考，按照这里所修改的以完成本发明的特征和功能。

如图1所示以及在下面结合图2－8进一步论述的，电视节目推荐器100包括节目数据库200、定型简表过程300、聚类例程400、平均值计算例程500、距离计算例程600以及聚类性能评估例程800。一般地，节目数据库200可以具体化为众所周知的电子节目向导并可以为在给定时间间隔内可用的每个节目记录信息。定型简表过程300（i）处理第三方观看历史记录130以生成反映有代表性的观众所观看的电视节目的典型样式的定型简表；（ii）允许用户选择最为相关的定型并因此初始化他或她的简表；以及（iii）基于选定的定型生成推荐。

由定型简表过程300调用聚类例程400以把第三方观看历史记录130（数据集）划分成群，以使在一个群内的点（电视节目）比其它任何群更接近该群的平均值（质心）。聚类例程400调用平均值计算例程500以计算一个群的符号平均值。由聚类例程400调用距离计算例程600以根据在给定电视节目与给定群的平均值之间的距离来评估一电视节目与各个群的接近度。最后，聚类例程400调用聚类性能评估例程800以确定何时已满足用于创建群的停止标准。

图2是取自图1的节目数据库（EPG）200的样本表。如先前指出的，节目数据库200为在给定时间间隔内可用的各个节目记录信息。如图2所示，节目数据库200含有诸如记录205－220条的多条记录，

每一条记录都与一给定的节目有关。对于每个节目，节目数据库 200 分别在栏 240 及栏 245 内表示出与该节目有关的日期／时间以及频道。另外，分别在栏 250、255 和 270 内为各个节目标识出名称、类型以及演员。另外的众所周知的特征（未示出）-诸如节目的持续时间以及说明也能够包含在节目数据库 200 内。

图 3 是描述结合了本发明特征的定型简表过程 300 的示例性实现的流程图。如先前指出的，定型简表过程 300（i）处理第三方观看历史记录 130 以生成反映有代表性的观众所观看的电视节目的典型样式的定型简表；（ii）允许用户选择最为相关的定型并因此初始化他或她的简表；以及（iii）基于选定的定型生成推荐。注意，可以例如，在工厂内脱机执行对第三方观看历史记录 130 的处理，并且能够向用户提供安装了所生成的定型简表以由用户进行选择的电视节目推荐器 100。

因而，如图 3 所示，定型简表过程 300 一开始在步骤 310 期间收集第三方观看历史记录 130。此后，定型简表过程 300 在步骤 320 期间执行下面结合图 4 所论述的聚类例程 400 以生成相应于定型简表的节目群。如下面进一步论述的，该示例性的聚类例程 400 可以对观看历史记录数据集 130 采用一种无监督数据聚类算法，诸如“k-平均值”聚类例程。如先前指出的，聚类例程 400 把第三方观看历史记录 130（数据集）划分成群，以使一个群内的点（电视节目）比其它任何群更接近该群的平均值（质心）。

然后，定型简表过程 300 在步骤 330 期间把表征每个定型简表的一个或多个标签分配给每个群。在一示例性的实施例中，该群的平均值变成为对于整个群的有代表性的电视节目，并且该平均值节目的特征能够用于标记该群。例如，电视节目推荐器 100 能够被配置成使得类型对每个群是主要因素或是定义特征。

在步骤 340 期间，把被标记的定型简表呈现给每个用户以便选择最接近该用户的兴趣的定型简表。组成每个选定群的节目能够被视为那个定型的“典型观看历史记录”，并且能够被用来为每个群建造一定型简表。因而，在步骤 350 期间为用户生成观看历史记录，该记录由来自选定定型简表的节目组成。最后，在步骤 360 期间把在上一步骤生成的观看历史记录加到节目推荐器上以得到节目推荐。节目推荐

器可以具体化为任何常规的节目推荐器，诸如上面所涉及的那些推荐器，虽然在这里进行了修改，但是对于本领域内的那些普通技术人员来说是显而易见的。在步骤 370 期间程序控制终止。

图 4 是描述结合了本发明特征的聚类例程 400 的示例性实现的流程图。如先前指出的，由定型简表过程 300 在步骤 320 期间调用聚类例程 400 来把第三方观看历史记录 130（数据集）划分成群，以使一个群内的点（电视节目）比其它任何群更接近该群的平均值（质心）。一般地，到聚类例程集中于在一样本数据集内寻找例子分组的无监督任务。本发明使用 k-平均值聚类算法来把数据集划分成 k 个群。如下文论述的，聚类例程 400 的两个主要参数是（i）用于寻找最接近的群的距离量度，在下面结合图 6 进行论述；以及（ii）k，要创建的群的数目。

该示例性的聚类例程 400 采用动态值 k，具有这样的条件，即，当示例数据的进一步聚类在分类精度上没有产生任何改进时已经达到一稳定的 k。另外，群的大小被递增到空群所被记录的那个点。因此，当已经达到这些群的平常水平时，聚类停止。

如图 4 所示，聚类例程 400 一开始在步骤 410 期间建立 k 个群。该示例性的聚类例程 400 通过选择最小数目的群，比如说两个而开始。对于这一固定的数目，聚类例程 400 处理整个观看历史记录数据集 130 并且通过数次重复，到达可以被看作是稳定的两个群（即，没有节目将从一个群移到另一个群，即使该算法将经历另一次重复）。在步骤 420 期间用一个或多个节目来初始化当前的 k 个群。

在一示例性的实现中，在步骤 420 期间，用从第三方观看历史记录 130 中选出的一些种子节目来初始化这些群。可以随机地或是顺序地选择用于初始化这些群的节目。在顺序实现中，可以用从观看历史记录 130 内的第一个节目开始的那些节目来初始化这些群，或是用起始于观看历史记录 130 内的任意一点的那些节目来初始化这些群。在再一种变更中，初始化各个群的节目数目还可以被改变。最后，用一个或多个“假定的”节目来初始化这些群，这些“假定的”节目由从第三方观看历史记录 130 内的节目中随机选取的特征值组成。

此后，聚类例程 400 在步骤 430 期间启动平均值计算例程 500 以计算各个群的当前平均值，将在下面结合图 5 论述平均值计算例程

500。然后，聚类例程 400 在步骤 440 期间执行距离计算例程 600 以确定在第三方观看历史记录 130 内的各个节目与各个群之间的距离，将在下面结合图 6 论述距离计算例程 600。然后，在步骤 460 期间，把观看历史记录 130 内的各个节目分配给最接近的群。

在步骤 470 期间，执行测试以确定是否有节目已经从一个群移到了另一个群。如果在步骤 470 期间确定一节目已从一个群移到了另一个群，则程序控制返回到步骤 430 并按照上述方式继续，直到识别出一组稳定的群。而如果在步骤 470 期间确定没有节目从一个群移到了另一个群，则程序控制进到步骤 480。

在步骤 480 期间执行进一步的测试以确定是否已满足特定的性能标准，或是是否识别出空的群（总称为“停止标准”）。如果在步骤 480 期间确定尚未满足停止标准，则在步骤 485 期间递增 k 的值，并且程序控制返回到步骤 420 并按照上述方式继续。而如果在步骤 480 期间确定已满足停止标准，则程序控制终止。将在下面结合图 8 进一步论述该停止标准的评估。

该示例性的聚类例程 400 把节目只放到一个群内，从而创建所谓的“脆”（crisp）群。进一步的变更将会采用模糊聚类，其允许一特殊的例子（电视节目）部分地属于许多个群。在模糊聚类方法中，给电视节目分配加权，该加权表示了电视节目到群平均值有多近。该加权能够视该电视节目与群平均值之间的距离的二次方的倒数而定。与单个电视节目有关的所有群的加权的总和必须总计为 100%。

群的符号平均值的计算

图 5 是描述结合了本发明特征的平均值计算例程 500 的示例性实现的流程图。如先前指出的，由聚类例程 400 调用平均值计算例程 500 来计算一个群的符号平均值。对于数字数据，该平均值是最小化方差的一个值。把这一概念扩展到符号数据，能够通过寻找最小化群内方差的 x_μ 值来确定一个群的平均值（并因此确定此群的半径或范围）。

$$Var\ (J) = \Sigma_{i \in J}\ (x_i - x_\mu)^2 \qquad (1)$$

$$群半径\ R(J) = \sqrt{Var(J)} \qquad (2)$$

其中 J 是一个源自同一类（被观看或未被观看）的电视节目群，x_i 是对应演出 i 的符号特征值，x_μ 是来自 J 内的其中一个电视节目的特征值以使它最小化 Var (J)。

因此，如图 5 所示，平均值计算例程 500 一开始在步骤 510 期间识别当前处于一给定群 J 内的节目。对于正在考虑中的该当前的符号属性，在步骤 520 期间使用等式（1）来为每个可能的符号值 x_μ 计算群 J 的方差。在步骤 530 期间，将最小化该方差的符号值 x_μ 选作为平均值。

在步骤 540 期间执行测试以确定是否存在需要考虑的另外的符号属性。如果在步骤 540 期间确定了存在需要考虑的另外的符号属性，则程序控制返回到步骤 520 并按照上述方式继续。而如果在步骤 540 期间确定了没有需要考虑的另外的符号属性，则程序控制返回到聚类例程 400。

在计算上，J 内的每个符号特征值都被尝试作为 x_μ，并且最小化该方差的符号值变成为群 J 内的考虑中的符号属性的平均值。有两种可能的平均值计算类型，称为基于显示的平均值以及基于特征的平均值。

基于特征的符号平均值

这里论述的示例性平均值计算例程 500 为基于特征的，其中结果群平均值由从群 J 内的例子（节目）中抽取出的特征值组成，这是因为符号属性的平均值必须是符号属性的可能的值之一。然而需要注意，群平均值可以是“假定的”电视节目，这一点很重要。该假定节目的特征值可以包括从这些例子之一（比方说，EBC）抽取出的频道值，以及从这些例子中的另外一个（比方说，BBC 世界新闻，实际上它从未在 EBC 上播出）抽取出的名称值。因此，展示最小方差的任何一个特征值被选定用来代表那一个特征的平均值。对于所有特征位置，重复平均值计算例程 500，直到在步骤 540 期间确定了所有特征值（即，符号属性）已经被考虑。由此得到的结果假定节目被用来代表此群的平均值。

基于节目的符号平均值

在一进一步的变更中，在用于方差的等式（1）中，x_i 可以是电视节目 i 本身，以及类似地，x_μ 可以是群 J 内的、最小化群 J 内节目组上的方差的节目。在此情形中，这些节目之间的、而不是单独的特征值之间的距离是要被最小化的相关量度。另外，在此情形中的结果平均值不是假定的节目，而正是从集合 J 中选出的一个节目。在群 J 内

如此找到的、最小化群 J 内的所有节目上的方差的任何一个节目被用来代表此群的平均值。

使用多个节目的符号平均值

上面论述的示例性平均值计算例程 500 使用用于各个可能的特征的一个单独的特征值表征了一个群的平均值（不论是按照基于特征的实现，还是按照基于节目的实现）。然而已经发现，在平均值计算期间仅仅依靠用于各个特征的一个特征值常常会导致不适当的聚类，这是由于该平均值不再是这个群的代表性的群中心。换言之，可能不希望仅仅用一个节目来代表一个群，而是可以用表示平均值或是多个平均值的多个节目代表一个群。因此，在一进一步的变更中，可以用多个平均值或是对于各个可能特征的多个特征值来代表一个群。因而，在步骤 530 期间选择最小化方差的 N 个特征值（对应基于特征的符号平均值）或 N 个节目（对应基于节目的符号平均值），其中 N 是用来代表一个群的平均值的节目数。

节目与群之间的距离计算

如先前指出的，由聚类例程 400 调用距离计算例程 600 以根据给定的电视节目与给定群的平均值之间的距离来评估电视节目与各个群的接近度。计算出的距离量度量化样本数据集内的各种例子之间的差别以确定一个群的范围。为了能够聚类用户简表，必须计算在观看历史记录内的任何两个电视节目之间的距离。一般地，彼此接近的电视节目趋向于落入一个群内。存在许多相对简单的技术用来计算数字值向量之间的距离，诸如欧几里德距离，曼哈顿距离，以及马哈拉诺比斯距离。

然而现有的距离计算技术不能用在电视节目向量的情形中，这是因为电视节目主要是由符号特征值组成。例如，能够用下面的特征向量来表示两个电视节目，诸如 2001 年 3 月 22 日晚上 8 点 EBC 播出的“朋友”剧目，以及 2001 年 3 月 25 日晚上 8 点 FEX 播出的“西蒙一家”剧目：

名称：朋友	名称：西蒙一家
频道：EBC	频道：FEX
播出日期：2001-03-22	播出日期：2001-03-25
播出时间：2000	播出时间：2000

显然，已知的数字距离量度不能用来计算特征值“EBC”与“FEX”之间的距离。值差量度（VDM）是用于测量符号特征值域内的特征值之间的距离的现有技术。VDM技术考虑对于各个特征的每个可能的值的所有例子的总体分类相似性。使用这一方法，根据训练集内的例子而统计地导出一个定义所有特征值之间的距离的矩阵。对于用于计算符号特征值之间的距离的VDM技术的更为详细的论述，参见例如ACM通讯，29：12，1213-1228（1986）上刊载的由Stanfill与Waltz所著的“Toward Memory-Based Reasoning”一文，在此将其并入作为参考。

本发明采用VDM技术或其变更来计算两个电视节目之间的、或其它感兴趣的项目之间的特征值之间的距离。最初的VDM建议在两个特征值之间的距离计算中采用加权项，这使得距离量度为不对称的。修改的VDM（MVDM）省略了该加权项以使距离矩阵是对称的。对于用于计算符号特征值之间的距离的MVDM技术的更为详细的论述，参见例如马萨诸塞州，波士顿，Kluwer出版社（1993）的Machine Learning第10卷，57-58上刊载的由Cost与Salzberg所著的“A Weighted Nearest Neighbor Algorithm For Learning With Symbolic Features”一文，在此将其并入作为参考。

根据MVDM，用下式给出对于一特定特征的两个值，V1与V2之间的距离δ：

$$\delta(V1,V2)=\sum\left|C1i/C1-C2i/C2\right|^{r} \qquad (3)$$

在本发明的节目推荐环境中，变换MVDM等式（3）专门用来处理“被观看”和“未被观看”的类。

$$\delta(V1,V2)=\left|\frac{C1\_watched}{C1\_total}-\frac{C2\_watched}{C2\_total}\right|+\left|\frac{C1\_not\_watched}{C1\_total}-\frac{C2\_not\_watched}{C2\_total}\right| \qquad (4)$$

在等式（4）中，V1和V2是对于在考虑中的特征的两个可能的值。继续上面的例子，对于特征“频道”，第一个值V1等于“EBC”，第二个值V2等于“FEX”。这两个值之间的距离为这些例子被分类到的所有类别上的总和。对于本发明的该示例性节目推荐器实施例的有关类别为“被观看”和“未被观看”。C1i是V1（EBC）被分到类别i（i

等于意指被观看类别的1）的次数，而C1（C1_total）是V1出现于数据集内的总次数。值“r”为常数，通常被设置成1。

在如果这些值对于所有分类都以同一相对频率出现时，用等式（4）定义的该量度就将这些值看成是相似的。C1i/C1 项表示平均值余数将被分类为 i 的似然性，假定所讨论的这一特征具有值V1。因此，如果两个值对所有可能的分类都给出相似的似然性，则这两个值是相似的。等式（4）通过寻找在所有分类上的这些似然性的差异之和来计算两个值之间的总相似性。两个电视节目之间的距离为这两个电视节目向量的相应特征值之间的距离的和。

图7A是用于与特征“频道”有关的特征值的一部分距离表。图7A规划对于各个类别的各个频道特征值出现的数目。图7A所示的值是已经从示例性的第三方观看历史记录130中取出的。

图7B显示了利用MVDM等式（4）从图7A所示的示例性计数中计算出的各个特征值对之间的距离。直观地，EBC与ABS应该彼此“接近”，因为它们主要出现在被观看类别中，而不出现在未被观看类别中（ABS具有少的未被观看成分）。图7B用EBC与ABS之间的小的（非零）距离来确认了这一直觉。另一方面，ASPN主要出现在未被观看类别中并因此应当“远离”EBC与ABS，对于该数据集。图7B将EBC与ASPN之间的距离规划为1.895，处于最大的可能距离2.0之外。类似地，ABS与ASPN之间的距离具有1.828高的一个值。

因此，如图6所示，距离计算例程600一开始在步骤610期间识别第三方观看历史记录130内的节目。对于正在考虑中的当前节目，距离计算例程600在步骤620期间使用等式（4）来计算各个符号特征值到各个群平均值（用平均值计算例程500确定）的相应特征的距离。

在步骤630期间通过合计相应的特征值之间的距离来计算当前节目与群平均值之间距离。在步骤640期间执行测试以确定在该第三方观看历史记录130内是否有另外的要被考虑的节目。如果在步骤640期间确定了在该第三方观看历史记录130内有另外的要被考虑的节目，则在步骤650期间识别下个节目，并且程序控制进到步骤620并按照上面描述的方式继续。

而如果在步骤640期间确定了在该第三方观看历史记录130内没有另外的要被考虑的节目，则程序控制返回到聚类例程400。

如先前在题为“从多个节目导出的符号平均值”的小节内所论述的，可以用许多对于各个可能的特征的特征值来表征一个群的平均值（不论是在基于特征的实现中，还是在基于节目的实现中）。然后用距离计算例程 600 的变更来集中来自多平均值得出的结果以通过投票来达成一致的决定。例如，现在在步骤 620 期间计算一个节目的给定特征值与对于各种方法的各个相应的特征值之间的距离。最小距离结果被集中并用于投票，例如通过采用多数投票或专家的混合以便达成一致的决定。对于这些技术的更为详细的论述，参见例如在第 13 届图案识别国际会议会刊，第 II 卷，897-901，奥地利，维也纳（1996）上刊载的由 J. Kittler 等人所著的“Combing Classifiers”一文，在此将其并入作为参考。

停止标准

如先前指出的，聚类例程 400 调用图 8 所示的聚类性能评估例程 800 来确定何时已满足用于创建群的停止标准。该示例性聚类例程 400 采用动态 k 值，具有这样的条件，即，当示例数据的进一步聚类在分类精度上没有产生任何改进时已经达到一稳定的 k。另外，群的大小被递增到空群所被记录的那个点。因此，当已经达到这些群的平常水平时，聚类停止。

该示例性聚类性能评估例程 800 使用第三方观看历史记录 130 的节目子集（测试数据集）来测试聚类例程 400 的分类精度。对于该测试集内的每个节目，聚类性能评估例程 800 确定与其最接近的群（该群的平均值是最接近的），并比较该群的类别标签与考虑中的节目。匹配的类别标签的百分数转换为聚类例程 400 的精度。

因此，如图 8 所示，聚类性能评估例程 800 一开始在步骤 810 期间从第三方观看历史记录 130 收集节目子集以作为测试数据集。此后，在步骤 820 期间根据该群内被观看和未被观看的节目的百分数而把类别标签分配给各个群。例如，如果该群内的大多数节目都被观看了，则可以给这个群分配“被观看”标签。

在步骤 830 期间识别与测试集内的各个节目最接近的群并比较该指定的群的类别标签以确定该节目是否被实际观看。在其中用多个节目来代表一个群的平均值的实现中，可以采用平均值距离（到各个节目的）或投票方案。在程序控制返回到聚类例程 400 之前，在步骤 840

期间确定匹配的类别标签的百分数。如果分类精度已达到预定的阈值，则聚类例程400将终止。

应当理解这里所示出并描述的实施例以及变更仅仅说明本发明的原理，可以由本领域的那些技术人员在不脱离本发明的范围和精神的情况下来实现各种修改。

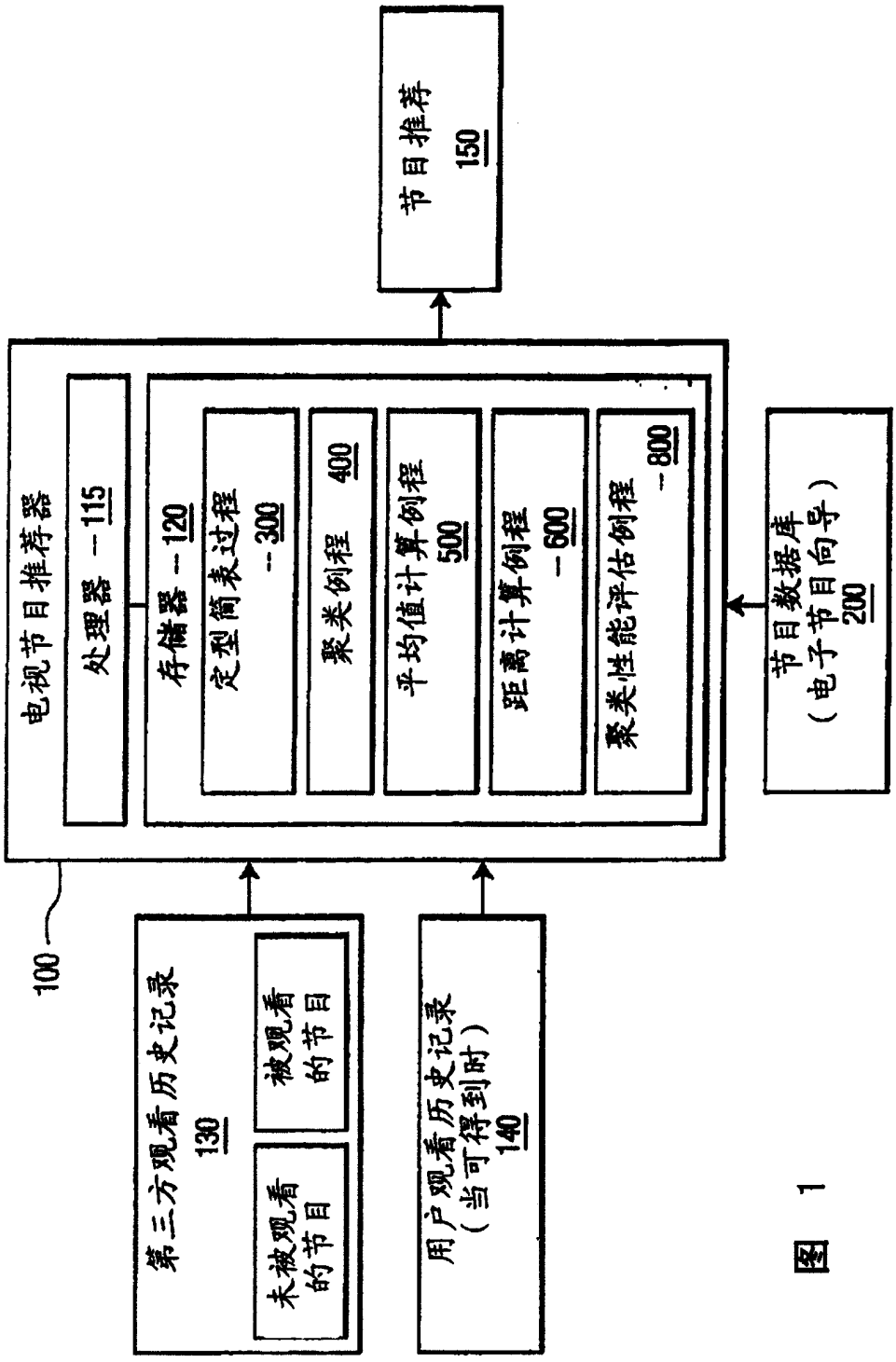

节目推荐
150
电视节目推荐器
处理器 115
存储器 120
定型简表过程
300
聚类例程 400
平均值计算例程
500
距离计算例程
600
聚类性能评估例程
800
节目数据库
（电子节目向导）
200
100
第三方观看历史记录
130
未被观看的节目
被观看的节目
用户观看历史记录
（当可得到时）
140

图 1

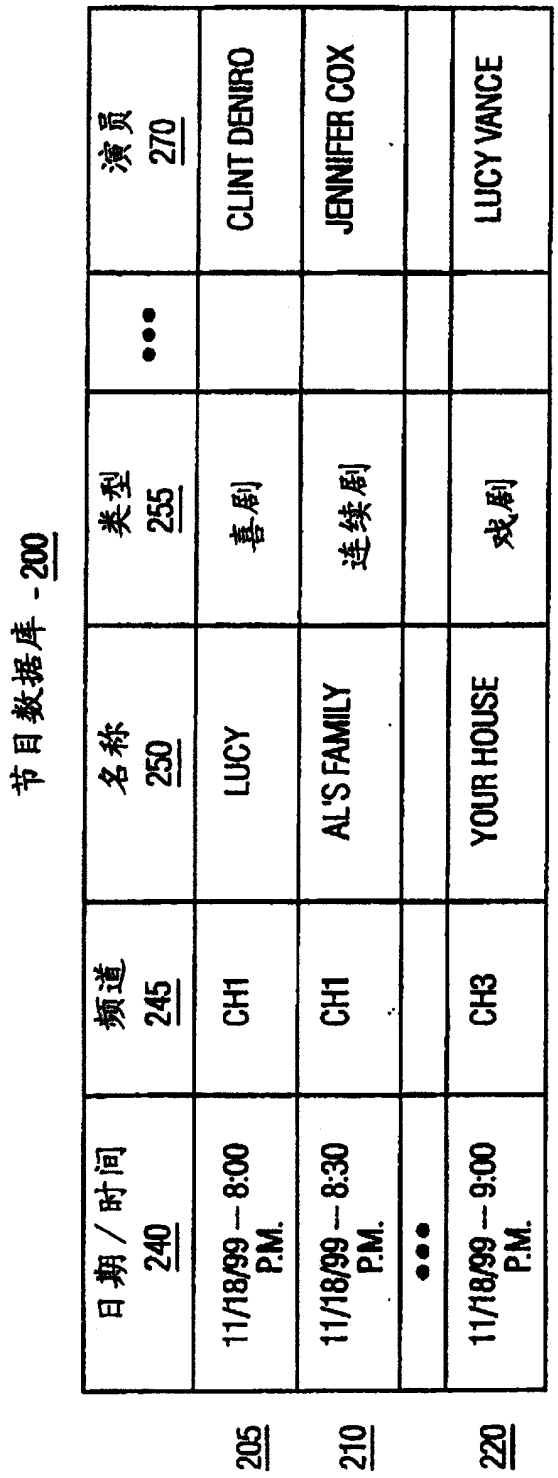

节目数据库 - 200
日期／时间 240
频道 245
名称 250
类型 255
•••
演员 270
11/18/99 — 8:00 P.M.
CH1
LUCY
喜剧
CLINT DENIRO
11/18/99 — 8:30 P.M.
CH1
AL'S FAMILY
连续剧
JENNIFER COX
•••
11/18/99 — 9:00 P.M.
CH3
YOUR HOUSE
戏剧
LUCY VANCE
205
210
220

图 2

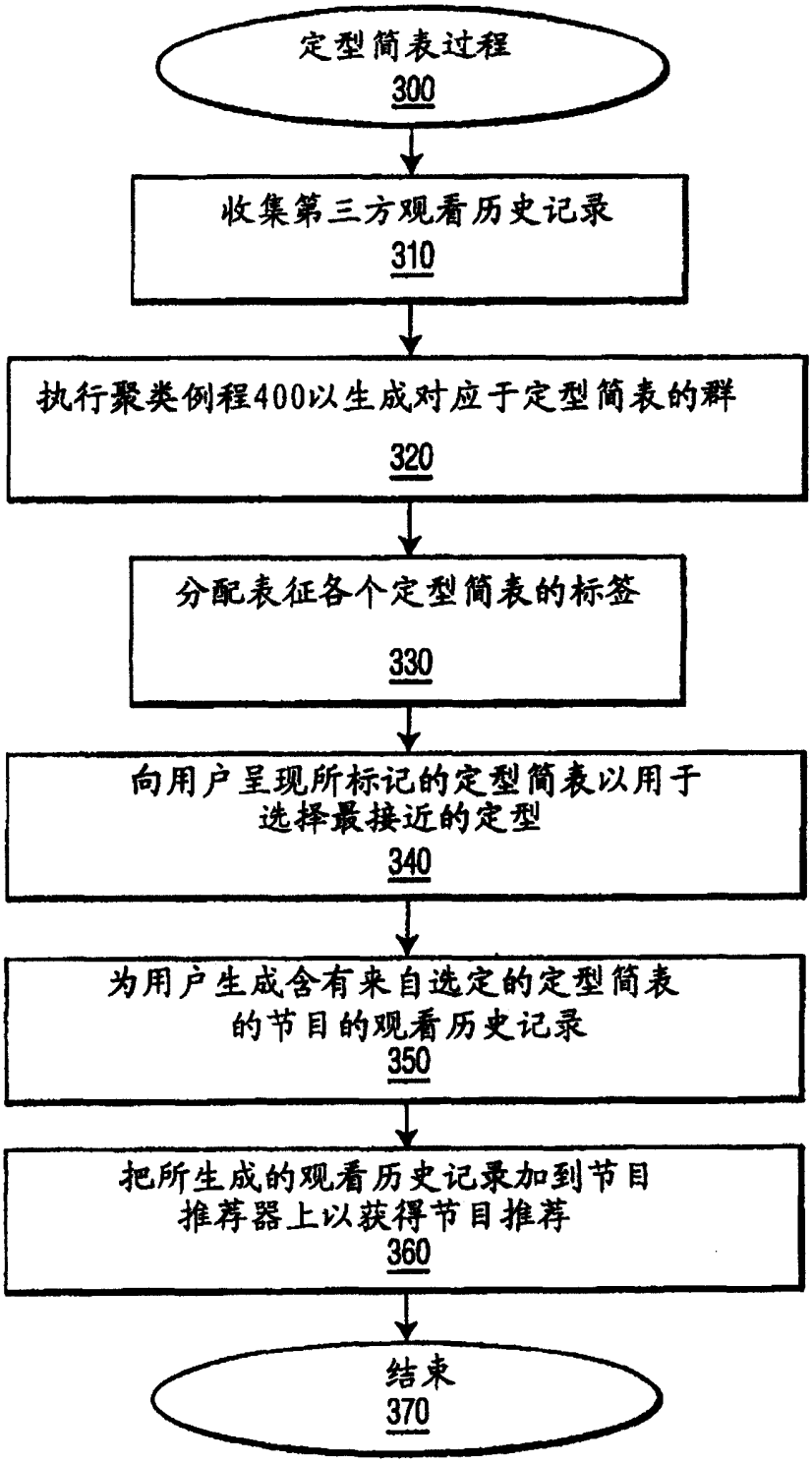

定型简表过程
300
收集第三方观看历史记录
310
执行聚类例程400以生成对应于定型简表的群
320
分配表征各个定型简表的标签
330
向用户呈现所标记的定型简表以用于选择最接近的定型
340
为用户生成含有来自选定的定型简表的节目的观看历史记录
350
把所生成的观看历史记录加到节目推荐器上以获得节目推荐
360
结束
370

图 3

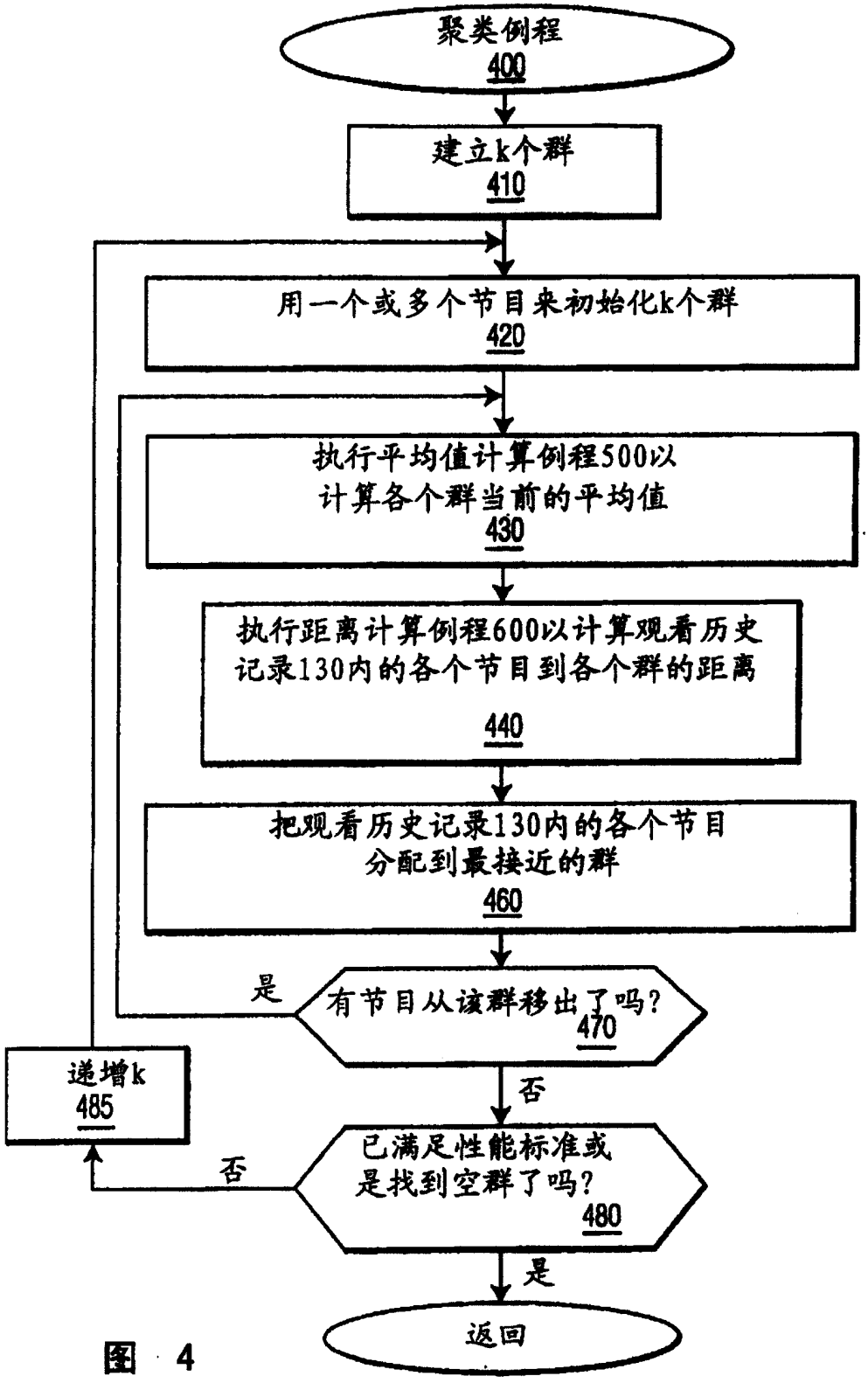

聚类例程
400
建立k个群
410
用一个或多个节目来初始化k个群
420
执行平均值计算例程500以
计算各个群当前的平均值
430
执行距离计算例程600以计算观看历史
记录130内的各个节目到各个群的距离
440
把观看历史记录130内的各个节目
分配到最接近的群
460
是
有节目从该群移出了吗?
470
否
递增k
485
否
已满足性能标准或
是找到空群了吗?
480
是
返回

图 4

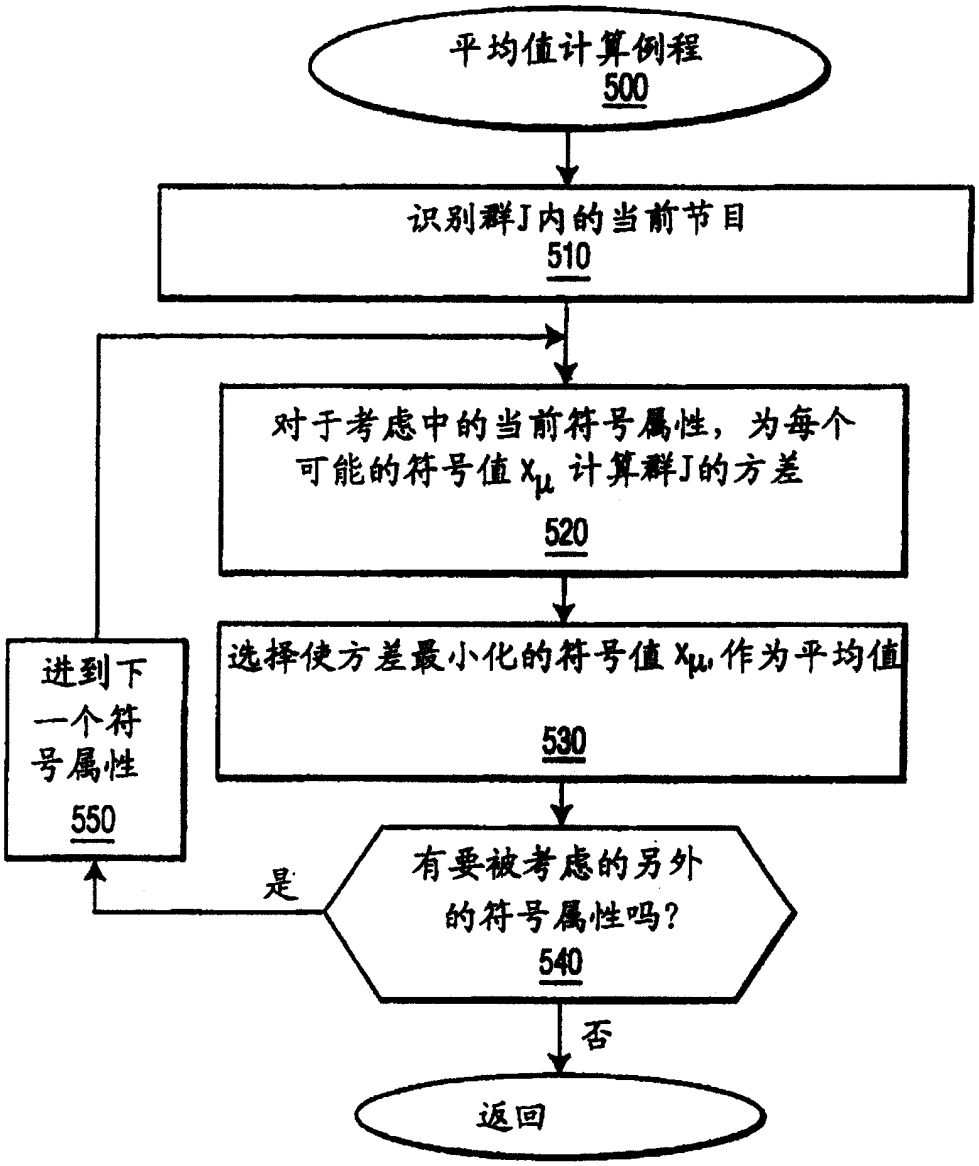

平均值计算例程
500
识别群J内的当前节目
510
对于考虑中的当前符号属性，为每个可能的符号值 x_μ 计算群J的方差
520
选择使方差最小化的符号值 x_μ 作为平均值
530
有要被考虑的另外的符号属性吗?
540
是
进到下一个符号属性
550
否
返回

图 5

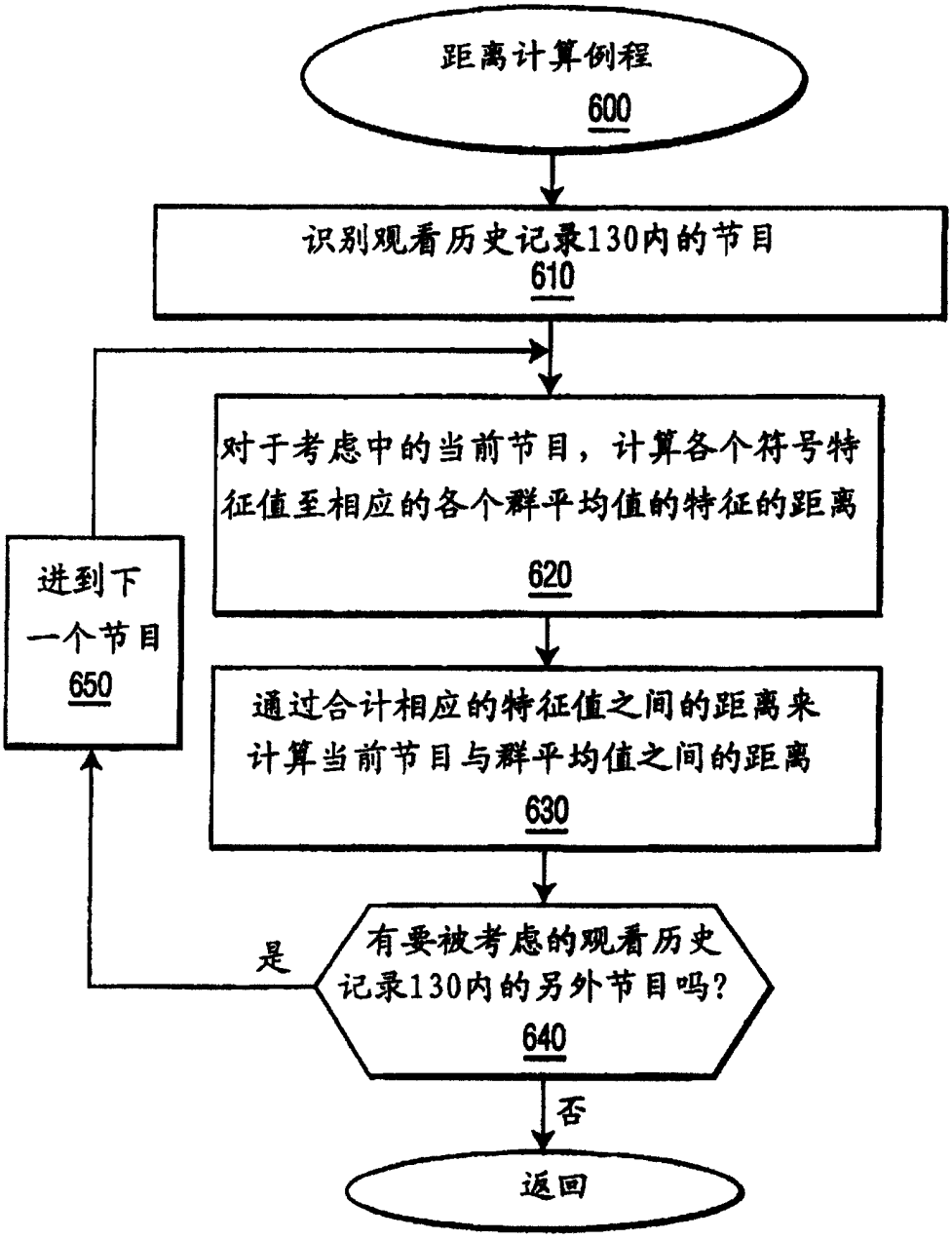

距离计算例程
600
识别观看历史记录130内的节目
610
对于考虑中的当前节目，计算各个符号特征值至相应的各个群平均值的特征的距离
620
通过合计相应的特征值之间的距离来计算当前节目与群平均值之间的距离
630
有要被考虑的观看历史记录130内的另外节目吗?
640
是
进到下一个节目
650
否
返回

图 6

频道特征值出现表 -700

	类别	
特征值	被观看	未被观看
EBC	353	0
ASPN	1	18
ABS	145	5

图 7A

特征值对距离表 -750

	EBC	ASPN	ABS
EBC	0	1.895	0.066
ASPN	1.895	0	1.828
ABS	0.066	1.828	0

图 7B

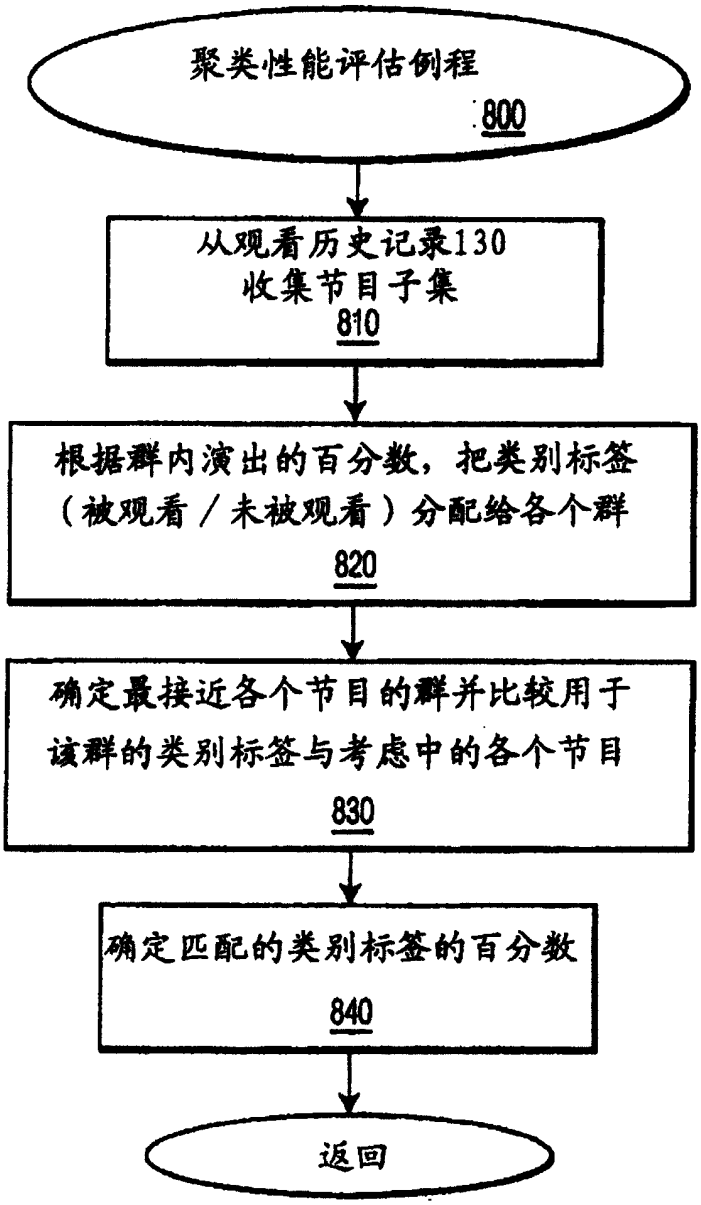

聚类性能评估例程
800
从观看历史记录130
收集节目子集
810
根据群内演出的百分数，把类别标签
（被观看／未被观看）分配给各个群
820
确定最接近各个节目的群并比较用于
该群的类别标签与考虑中的各个节目
830
确定匹配的类别标签的百分数
840
返回

图 8