(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 102411591 A
(43)申请公布日 2012.04.11

(21)申请号 201010292384.0

(22)申请日 2010.09.21

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼岛资本大厦一座四层 847 号邮箱

(72)发明人 顾海杰 苏宁军 代其锋 马海平 张金银 陈恩红

(74)专利代理机构 北京同达信恒知识产权代理有限公司 11291
代理人 郭润湘

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

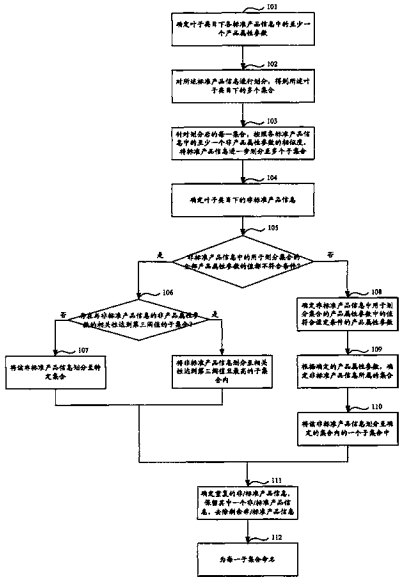
权利要求书 3 页 说明书 11 页 附图 3 页

(54)发明名称

一种信息处理的方法及设备

(57)摘要

本申请公开了一种信息处理的方法及设备，主要内容包括：在叶子类目的基础上，根据叶子类目下各标准产品信息中的至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，由于划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同，因此，最终得到的是产品信息细粒度划分的集合，买家用户以细粒度的产品信息集合为基础搜索、查询产品信息时，可以有效减少搜索、查询的时间、提高搜索、查询的准确性以及提高向买家用户推荐产品信息的准确度，且细粒度的产品信息集合也能够提高对产品信息进行操作的可用性，减少对产品信息进行操作时的运算量。



CN 102411591 A

1. 一种信息处理的方法，其特征在于，所述方法包括：

确定叶子类目下各标准产品信息中的至少一个产品属性参数；

利用确定的所述至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合；

其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

2. 如权利要求 1 所述的方法，其特征在于，确定叶子类目下各标准产品信息中的至少一个产品属性参数，具体包括：

确定叶子类目下各标准产品信息中的全部产品属性参数；

确定设定时长内每一产品属性参数作为搜索条件，在所述叶子类目下进行产品信息搜索的次数；

从所述全部产品属性参数中选择至少一个产品属性参数；

其中，选择的任一产品属性参数作为搜索条件进行产品信息搜索的次数达到第一阈值且该产品属性参数的值为离散型。

3. 如权利要求 2 所述的方法，其特征在于，对所述标准产品信息进行划分，得到所述叶子类目下的多个集合之后，所述方法还包括：

针对划分后得到的每一集合，确定集合中各标准产品信息中的至少一个非产品属性参数；

确定同一集合内各标准产品信息的至少一个非产品属性参数的相似度；

按照确定的相似度对同一集合内各标准产品信息做进一步划分，得到该集合下的多个子集合；

其中，划分在同一子集合内的任意两个标准产品信息的所述至少一个非产品属性参数之间的相似度达到第二阈值。

4. 如权利要求 3 所述的方法，其特征在于，所述方法还包括：

确定叶子类目下的非标准产品信息，所述非标准产品信息中的用于划分集合的至少一个产品属性参数的值不符合设定的条件；

得到集合下的多个子集合之后，所述方法还包括：

判断非标准产品信息中的用于划分集合的全部产品属性参数的值是否都不符合设定的条件；

若是，则确定该非标准产品信息中的至少一个非产品属性参数，并根据该非产品属性参数将该非标准产品信息划分至一个子集合内，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度达到第三阈值；

若否，则确定非标准产品信息中用于划分集合的产品属性参数中取值符合设定的条件的产品属性参数，并确定包含取值符合设定的条件的产品属性参数的集合，以及，根据非标准产品信息的非产品属性参数，将该非标准产品信息划分至确定的集合内的一个子集合中，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度最高。

5. 如权利要求 4 所述的方法，其特征在于，在非标准产品信息中的用于划分集合的全部产品属性参数的值不符合设定的条件，且该非标准产品信息的非产品属性参数与任一子

集合内的标准产品信息中的非产品属性参数的相似度低于第三阈值时，将该非标准产品信息划分至特定集合。

6. 如权利要求 4 或 5 所述的方法，其特征在于，所述方法还包括：

确定重复的标准产品信息，保留其中一个标准产品信息，去除剩余的标准产品信息；

确定重复的非标准产品信息，保留其中一个非标准产品信息，去除剩余的非标准产品信息；

所述重复的标准产品信息间的产品属性参数和非产品属性参数都相同，所述重复的非标准产品信息间的产品属性参数和非产品属性参数都相同。

7. 如权利要求 4 或 5 所述的方法，其特征在于，

确定每一子集合的标准产品信息和非标准产品信息中的产品属性参数和非产品属性参数；

统计确定该产品属性参数和非产品属性参数中的至少一个高频词；

利用确定的至少一个高频词为该子集合命名。

8. 如权利要求 1 所述的方法，其特征在于，得到所述叶子类目下的多个集合之后，所述方法还包括：

根据得到的各集合中的标准产品信息进行搜索或产品信息推荐。

9. 一种信息处理的设备，其特征在于，所述设备包括：

标准参数确定模块，用于确定叶子类目下各标准产品信息中的至少一个产品属性参数；

第一划分模块，用于利用确定的所述至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

10. 如权利要求 9 所述的设备，其特征在于，所述标准参数确定模块，包括：

属性参数确定子模块，用于确定叶子类目下各标准产品信息中的全部产品属性参数；

次数确定子模块，确定设定时长内每一产品属性参数作为搜索条件，在所述叶子类目下进行产品信息搜索的次数；

选择子模块，用于从所述全部产品属性参数中选择至少一个产品属性参数，其中，选择的任一产品属性参数作为搜索条件进行产品信息搜索的次数达到第一阈值且该产品属性参数的值为离散型。

11. 如权利要求 10 所述的设备，其特征在于，所述设备还包括：

非标准参数确定模块，用于针对划分后得到的每一集合，确定集合中各标准产品信息中的至少一个非产品属性参数；

相似度确定模块，用于确定同一集合内各标准产品信息的至少一个非产品属性参数的相似度；

第二划分模块，用于按照确定的相似度对同一集合内各标准产品信息做进一步划分，得到该集合下的多个子集合，其中，划分在同一子集合内的任意两个标准产品信息的所述至少一个非产品属性参数之间的相似度达到第二阈值。

12. 如权利要求 11 所述的设备，其特征在于，所述设备还包括：

判断模块，用于判断非标准产品信息中的用于划分集合的全部产品属性参数的值是否

都不符合设定的条件，若是，则触发非标准参数确定模块，否则，触发标准参数确定模块，所述非标准产品信息中的用于划分集合的至少一个产品属性参数的值不符合设定的条件；

所述非标准参数确定模块，还用于确定该非标准产品信息中的至少一个非产品属性参数；

所述相似度确定模块，用于确定非标准产品信息的非产品属性参数与各子集合内的标准产品信息中的非产品属性参数的相似度；

所述标准参数确定模块，还用于确定非标准产品信息中用于划分集合的产品属性参数中符合设定的条件的产品属性参数；

所述第一划分模块，还用于确定包含所述符合设定的条件的产品属性参数的集合；

所述第二划分模块，还用于根据非产品属性参数，将该非标准产品信息划分至一个子集合内，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度达到第三阈值，或者，

根据非标准产品信息的非产品属性参数，将该非标准产品信息划分至第一划分模块确定的集合内的一个子集合中，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度最高。

13. 如权利要求 12 所述的设备，其特征在于，

所述第一划分模块，还用于在非标准产品信息中的用于划分集合的全部产品属性参数的值不符合设定的条件，且该非标准产品信息的非产品属性参数与任一子集合内的标准产品信息中的非产品属性参数的相似度低于第三阈值时，将该非标准产品信息划分至特定集合。

14. 如权利要求 12 或 13 所述的设备，其特征在于，所述设备还包括：

去重模块，用于确定重复的标准产品信息，保留其中一个标准产品信息，去除剩余的标准产品信息，以及，确定重复的非标准产品信息，保留其中一个非标准产品信息，去除剩余的非标准产品信息，所述重复的标准产品信息间的产品属性参数和非产品属性参数都相同，所述重复的非标准产品信息间的产品属性参数和非产品属性参数都相同。

15. 如权利要求 12 或 13 所述的设备，其特征在于，所述设备还包括：

命名模块，用于确定每一子集合的标准产品信息和非标准产品信息中的产品属性参数和非产品属性参数，统计确定该产品属性参数和非产品属性参数中的至少一个高频词，并利用确定的至少一个高频词为该子集合命名。

一种信息处理的方法及设备

技术领域

[0001] 本申请涉及计算机技术领域，尤其涉及一种信息处理的方法及设备。

背景技术

[0002] 随着计算机技术以及通信技术的不断发展，越来越多的用户在购物网站上搜索、查询、购买商品。用户在搜索、查询、购买商品之前，会浏览卖家用户在网站上发布的产品信息，所述卖家用户可以是企业实体、厂商或是个人经营者。

[0003] 网站服务器接收到的卖家用户上报的产品信息是海量信息，为了对接收到的产品信息所表示的产品进行分类以及有效地引导买家用户获得其想要的产品信息，网站服务器通常采用多级类目的方式来划分产品信息。多级类目体系一般有以下两个特征：

[0004] 特征 1：多级类目体系的架构相对稳定。

[0005] 架构相对稳定的多级类目体系一方面有助于卖家用户积累习惯，在向网站服务器上报产品信息时，按照多级类目体系的要求上报标准格式和内容的产品信息。另一方面有助于买家用户积累习惯，根据累计的经验在多级类目体系下快速搜索、查找想要获得的产品信息。

[0006] 特征 2：多级类目体系一般由网站服务器的运营人员人工运营。

[0007] 人工运营的方式可以将本领域的共有知识固定下来形成标准，有助于多级类目体系在各种网站内的推广使用。

[0008] 为了保持多级类目体系的上述两点特征，在通过多级类目的方式划分海量产品信息时，只能按照相对较粗的粒度划分产品信息，这是因为：由于产品信息的多种多样，如果将产品信息划分为较细的粒度，则多级类目的最底层叶子类目需要随着产品信息的改变而变化，不利于多级类目体系的稳定；且如果细粒度地划分产品信息，生成的多级类目的架构势必会非常庞大，增加了人工运营网站服务器的难度。

[0009] 例如：若某一叶子类目下是连衣裙的产品信息，针对其中的某一产品信息，在该产品信息中的产品材质由丝质修改为棉质时，该产品信息仍是该连衣裙叶子类目下的产品信息，叶子类目不发生变化。若多级类目体系划分的粒度更细，如某一叶子类目下是丝质连衣裙的产品信息，则当某一产品信息中的产品材质由丝质修改为棉质时，需要将该产品信息由丝质连衣裙的叶子类目改为棉质连衣裙的叶子类目，此时，叶子类目随着产品信息的改变而变化。同时，由于多级类目体系是树状的节点架构，因此，每增加一层子类目，多级类目体系中将增加大量的类目，使得多级类目的架构非常庞大。

[0010] 由于多级类目体系下的产品信息划分粒度较粗，因此，即使是多级类目体系中的最底层的叶子类目中包含的产品信息也依旧是海量的。在此情况下，买家用户通过多级类目体系搜索、查询产品信息时的查询时间较长，且查询的准确度较低，且网站服务器向买家用户推荐产品信息时，也只能以叶子类目为单位，向买家用户推荐叶子类目下的所有产品信息，使得推荐的产品信息差异很大，准确度不能满足买家用户的实际需求。除此以外，由于叶子类目中包含的产品信息量很大，属于同一叶子类目中的产品信息差异也很大，因此，

针对叶子类目下的产品信息的操作实现难度也较大。例如：在通过价格参数来自动抓取不安全的产品信息的操作中，一般认为极端价格很可能涉嫌假冒产品，假设 A 品牌的产品价格低于 100 元时表示该产品为假冒产品，而 B 品牌的同类型产品价格低于 20 元时表示该产品为假冒产品，如果某一产品的价格为 50 元，则通过价格参数的方式很难直接定位出价格为 50 元的产品是否为假冒产品，还必须结合该产品的其他信息来判断，而叶子类目下的产品信息众多，使得运算量非常大。

发明内容

[0011] 本申请实施例提供一种信息处理的方法及设备，用以解决现有技术中存在的多级类目体系下产品信息划分粒度较大的问题。

[0012] 一种信息处理的方法，所述方法包括：

[0013] 确定叶子类目下各标准产品信息中的至少一个产品属性参数；

[0014] 利用确定的所述至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合；

[0015] 其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

[0016] 一种信息处理的设备，所述设备包括：

[0017] 标准参数确定模块，用于确定叶子类目下各标准产品信息中的至少一个产品属性参数；

[0018] 第一划分模块，用于利用确定的所述至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

[0019] 本申请有益效果如下：

[0020] 本申请实施例在叶子类目的基础上，根据叶子类目下各标准产品信息中的至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，由于划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同，因此，最终得到的是产品信息细粒度划分的集合，买家用户以细粒度的产品信息集合为基础搜索、查询产品信息时，可以有效减少搜索、查询的时间、提高搜索、查询的准确性以及提高向买家用户推荐产品信息的准确度，且细粒度的产品信息集合也能够提高对产品信息进行操作的可用性，减少对产品信息进行操作时的运算量。

附图说明

[0021] 图 1 为本申请中的信息处理设备在多级类目体系下的示意图；

[0022] 图 2 为本申请实施例一中信息处理方法示意图；

[0023] 图 3 为本申请实施例二中信息处理设备结构示意图；

[0024] 图 4 为本申请实施例二中信息处理设备中的标准参数确定模块结构示意图。

具体实施方式

[0025] 为了实现本申请目的，本申请在多类目体系的叶子类目基础上，对叶子类目下的

产品信息按照其产品属性参数作进一步划分，在每一个叶子类目下划分出多个细粒度的产品信息集合，使得在买家用户搜索、查询产品信息时，以细粒度的产品信息集合为基础，可以有效减少搜索、查询的时间、提高搜索、查询的准确性以及提高网站服务器向买家用户推荐产品信息的准确度，且细粒度的产品信息集合也能够提高对产品信息进行操作的可用性，减少对产品信息进行操作时的运算量；并且，本申请方案是在多类目体系的叶子类目基础上执行的，对多级类目体系的实质性内容没有改变，多级类目体系本身仍然具有上述的两点特征。

[0026] 本申请各实施例中涉及的信息处理设备可以应用在多级类目体系下，其架构如图 1 所示，在多级类目体系的最底层叶子类目下，信息处理设备对叶子类目中的产品信息进行处理，得到叶子类目下一层次的对产品信息做细粒度划分的集合和子集合。多级类目体系下的各叶子类目是相对独立的类目，本申请方案是要在每一叶子类目下继续构建细粒度的类目结构，因此，本实施例方案中可以采用分布式算法，将每一个叶子类目作为一个计算节点，通过计算机集群对多个叶子类目进行分布式操作，以加快本实施例中的信息处理方案。在图 1 所示的结构下，独立于多级类目体系下的一个信息处理设备可以对多级类目体系下的多个叶子类目中的产品信息进行处理。

[0027] 若将本申请方案中的信息处理方案处理后的集合和子集合中的标准产品信息应用于向买家用户推荐产品信息或是买家用户的产品信息搜索的场景，则用于向买家用户推送产品信息的服务器根据买家用户的需求（如买家用户输入的关键字或是买家用户在之前一段时间内的购买习惯），将集合或子集合中的产品信息推送给买家用户，或是搜索引擎从集合和子集合中进行关键字搜索，并将搜索得到的产品信息发送给买家用户。

[0028] 根据叶子类目下的产品信息的产品属性参数值是否符合设定的条件，可以将其划分为标准产品信息和非标准产品信息，标准产品信息的产品属性参数值符合设定的条件，而非标准产品信息的某些或全部产品属性参数值不符合设定的条件。所述设定的条件可以是产品属性参数值的取值是实际可用的取值。例如：某一叶子类目下包含的是针对连衣裙的产品信息，产品信息 1 中的价格参数值是 100 ～ 150 元区间，而产品信息 2 中的价格参数值是 ABC，并不是表示价格的正数数值，说明卖家用户在填写产品信息 2 中的价格参数值时可能出现失误，则将产品信息 2 看作是非标准产品信息。

[0029] 不论是标准产品信息还是非标准产品信息都包括产品属性参数和非产品属性参数，产品属性参数表示该产品信息对应产品本身的固定属性，包括但不限于：产品的价格、产品的材质、产品的品牌、产品的型号、产品的重量等；非产品属性参数表示卖家用户或是网站服务器为产品定义的属性，包括但不限于：产品信息所属的叶子类目、发布产品的标题、卖家用户 ID、产品的用途等。

[0030] 下面结合说明书附图对本申请实施例进行详细描述。

[0031] 实施例一：

[0032] 如图 2 所示，为本申请实施例一中信息处理方法示意图，所述方法包括以下步骤：

[0033] 步骤 101：确定叶子类目下各标准产品信息中的至少一个产品属性参数。

[0034] 假设本步骤中的叶子类目包含的是针对连衣裙的产品信息，包含的标准产品信息有 1000 个，其中 3 个标准产品信息（标准产品信息 1、标准产品信息 2 和标准产品信息 3）的内容如表 1 所示：每个标准产品信息中有 4 个产品属性参数，分别是连衣裙的价格、连衣

裙的材质、连衣裙的品牌、连衣裙的型号。

[0035]

标准产品信息	价格	材质	品牌	型号
标准产品信息 1	价格区间 1	材质 1	A	型号 1
标准产品信息 2	价格区间 1	材质 1	B	型号 2
标准产品信息 3	价格区间 2	材质 2	C	型号 3

[0036] 表 1

[0037] 若将表 1 中的 4 个产品属性参数都作为细粒度产品信息的划分条件，对叶子类目下的标准产品信息进行划分，则会由于划分条件过于复杂导致划分后得到的集合数量过多。例如：若叶子类目下的所有 1000 个标准产品信息中共包含 4 种连衣裙的价格区间、3 种连衣裙的材质、50 种连衣裙的品牌、5 种连衣裙的型号，则最多将会划分得到 4×3×50×5 = 3000 个集合。而在实际的处理过程中，产品属性参数的种类可能会更多，每种产品属性参数值的可选项也更多，因此，将全部产品属性参数都作为划分条件的话，运算量会比较大。对此，本步骤提出一种优化方案，从全部产品属性参数中选择部分产品属性参数来作为本步骤 101 中确定的产品属性参数用于作细粒度的产品信息划分。

[0038] 具体的选择产品属性参数的操作需要满足以下两方面要求：

[0039] 一方面，选择的产品属性参数应该是对外最能反映该产品信息的参数，即买家用户常用于搜索、查询的参数，以便于根据选择的产品属性参数进行划分后得到的集合能够提高用户搜索、查询的准确性；另一方面，选择的产品属性参数的值最好是离散型的，以减少在后续划分集合时由于标准产品信息中的产品属性参数的取值可能性过多导致运算量大的问题。

[0040] 为了满足以上两方面要求，本步骤中选择用户集合划分的产品属性参数的条件是：产品属性参数作为搜索条件，在之前的一段设定时长内，在叶子类目下进行产品信息搜索的次数需要达到第一阈值，且该产品属性参数的值为离散型。

[0041] 按照上述条件选择出至少一个产品属性参数后用于后续的集合划分操作。仍以表 1 所示的产品信息中的产品属性参数为例，若叶子类目下的所有 1000 个标准产品信息中共包含 4 种连衣裙的价格区间、3 种连衣裙的材质、50 种连衣裙的品牌、5 种连衣裙的型号，则可以将其中连衣裙的价格的值的 4 个区间看作离散型；连衣裙的材质的取值为 3 种不同的材质，可以看作是离散型；连衣裙的品牌的取值为 50 种不同的品牌，可以看作是离散型；连衣裙的型号的取值为 5 种不同的型号，可以看作是离散型。若买家用户最常用价格和材质为搜索条件进行产品信息的搜索、查询，则将产品的价格、产品的材质作为本步骤中确定的用于集合划分的产品属性参数。

[0042] 步骤 102：对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

[0043] 在本步骤中，若用于划分集合的产品属性参数是产品的价格和产品的材质，其中，产品的价格有 4 种价格区间，产品的材质有 3 种不同的材质，则划分后应该得到以下 12 个

集合：

[0044] 集合 1：价格区间 1+ 材质 1；集合 2：价格区间 1+ 材质 2；集合 3：价格区间 1+ 材质 3；集合 4：价格区间 2+ 材质 1；集合 5：价格区间 2+ 材质 2；集合 6：价格区间 2+ 材质 3；集合 7：价格区间 3+ 材质 1；集合 8：价格区间 3+ 材质 2；集合 9：价格区间 3+ 材质 3；集合 10：价格区间 4+ 材质 1；集合 11：价格区间 4+ 材质 2；集合 12：价格区间 4+ 材质 3。

[0045] 划分在同一集合中的任意两个标准产品信息中的价格区间和材质都相同，根据表 1 所示的各标准产品信息的产品属性参数描述，标准产品信息 1 和标准产品信息 2 应划分在集合 1，标准产品信息 3 应划分在集合 5，任意两个集合中满足类与类不相似的原则。

[0046] 由于在步骤 101 中选择用于细粒度产品信息划分的产品属性参数时，充分考虑到了买家用户的搜索、查询习惯以及产品属性参数的值为离散型，同时忽略未选择的产品属性参数，使得本步骤的划分结果不仅能够正确体现买家用户的使用习惯，还减少了划分时的运算量。

[0047] 通过以上步骤 101 和步骤 102 的方案，对多级类目体系的叶子类目中的产品信息作了细粒度的划分，得到了细粒度划分产品信息的集合，使得买家用户通过集合内的细粒度产品进行搜索、查询时可以有效减少等待时间、提高准确性以及提高网站服务器向买家用户推荐产品信息的准确性，有助于提高对产品信息的各项操作的可行性。例如，在将连衣裙的叶子类目划分为上述 12 个集合后，可以根据买家用户的搜索条件在相应的集合内进行搜索，由于集合内的产品信息数量远远小于叶子类目下的产品信息数量，因此，可以大大减少买家用户搜索等待时间，且保证搜索、查询的准确性；同时，在向买家用户推荐产品信息时，根据向买家用户推荐特定价位以及材质的产品信息，使得推荐的产品信息接近买家用户实际需求；另外，仍以通过价格参数来自动抓取不安全的产品信息的操作为例，通过上述方式划分细粒度的产品信息集合后，可以根据待测的价格快速定位出相应的集合，进而在定位出的集合中查询是否存在假冒产品，相比于在叶子类目下的操作，可以有效地减少操作的运算量，提高操作执行的有效性。

[0048] 在上述步骤 101 和步骤 102 的优选方案中，根据最能反映产品信息以及取值为离散型的产品属性参数作为划分集合的产品属性参数，可以进一步根据划分后得到的集合来提高用户搜索、查询的准确性，以及减少在划分集合时由于标准产品信息中的产品属性参数的值的可能性过多导致运算量大的问题。

[0049] 步骤 103：针对划分后的每一集合，按照集合中各标准产品信息中的至少一个非产品属性参数的相似度，将集合内的标准产品信息进一步划分至多个子集合。

[0050] 本步骤是实现本申请目的的优选步骤，在叶子类目下划分出多个集合后，再针对每一个集合作进一步的划分，可以在步骤 102 的基础上得到更细粒度的标准产品信息的子集合。

[0051] 本步骤的具体做法为：

[0052] 首先，针对划分后得到的每一集合，确定集合中各标准产品信息中的至少一个非产品属性参数。

[0053] 仍以步骤 101 和步骤 102 中涉及的连衣裙产品信息为例，假设本步骤中确定的非产品属性参数是卖家用户在网站服务器上发布产品的标题，标准产品信息 1 和标准产品信息 2 划分在集合 1 中，标准产品信息 1 内发布产品的标题中的关键词为：连衣裙、动物图案，

标准产品信息 2 内发布产品的标题中的关键词为 :连衣裙、条纹图案。

[0054] 然后,确定同一集合内各标准产品信息的至少一个非产品属性参数的相似度。

[0055] 由于以发布产品的标题为非产品属性参数,因此,本步骤中需要运算同一集合内各发布产品的标题之间的相似度。如通过 K- 中心点等聚类算法运算相似度。

[0056] 最后,将同一集合内各标准产品信息做进一步划分,得到该集合下的多个子集合,使得划分在同一子集合内的两两标准产品信息中的所述至少一个非产品属性参数之间的相似度达到第二阈值。

[0057] 若通过 K- 中心点等聚类算法运算相似度时,将相似度达到第二阈值的各标准产品信息划分在同一子集合,不同子集合之间,非产品属性参数的相似度较低。

[0058] 本实施例也不限于通过产品信息的其他非产品属性参数之间的相似度来划分子集合,如通过产品用途划分等。

[0059] 通过上述步骤 101 ～步骤 103 的方案,完成了针对标准产品信息的细粒度划分,进一步地,还可以对非标准产品信息作细粒度的划分。

[0060] 步骤 104 :确定叶子类目下的非标准产品信息。

[0061] 本步骤也是实现本申请目的的优选步骤,虽然叶子类目下的非标准产品信息中的部分或全部产品属性参数的值不符合设定的条件,但如果这些值是卖家用户上报时填写失误或是其他人为误差造成的,则该非标准产品信息也应该真实表示一个产品的相关信息,且该非标准产品信息也具有搜索、查询、向用户推荐或是用于其它操作的意义,因此,本优选步骤就是在已划分标准产品信息的集合和子集合后,进一步对非标准产品信息的划分。

[0062] 步骤 105 :判断非标准产品信息中的用于划分集合的全部产品属性参数的值是否都不符合设定的条件,若是,则执行步骤 106 ;否则执行步骤 108。

[0063] 非标准产品信息的某些或全部产品属性参数值不符合设定的条件,在本步骤中,需要判断在步骤 101 中确定的用于划分集合的产品属性参数是否是不符合设定的条件的参数。例如,在步骤 101 中确定的用户划分集合的产品属性参数为产品的价格、产品的材质,则在本步骤中将确定非标准产品信息中的产品的价格和产品的材质这两种产品属性参数是否都不符合设定的条件。

[0064] 需要说明的是,若在步骤 101 中将产品信息的全部产品属性参数都用于划分集合,则只要有产品属性参数值是不符合设定的条件的产品信息就定义为非标准产品信息 ;若在步骤 101 中设定部分产品属性参数用于划分集合,则在本步骤中涉及的非标准产品信息是指用于划分集合的至少一个产品属性参数值为不符合设定的条件的产品信息。

[0065] 特殊地,若存在某一非标准产品信息,该非标准产品信息中取值不符合设定的条件的产品属性参数并未用于划分集合,如用于划分集合的产品属性参数为产品的价格、产品的材质,但该非标准产品信息中取值不符合设定的条件的产品属性参数是产品型号,则该非标准产品信息可以作为标准产品信息的特例,在步骤 102 中进行划分 ;也可以仍旧作为非标准产品信息,在后续步骤中划分。

[0066] 步骤 106 :确定非标准产品信息中的至少一个非产品属性参数,并判断是否存在与非标准产品信息的非产品属性参数的相关性达到第三阈值的子集合,若存在,则将非标准产品信息划分至相关性达到第三阈值且最高的子集合内,并跳转至步骤 111 ;否则,执行步骤 107。

[0067] 由于非标准产品信息中的产品属性参数无法用于集合划分，因此，本步骤中利用非标准产品信息的非产品属性参数来判断该非标准产品信息应该属于哪一子集合。

[0068] 仍以用于划分集合的产品属性参数为产品的价格、产品的材质为例，用于划分子集合的非产品属性参数为发布产品的标题，则在本步骤中，某一非标准产品信息的产品的价格和产品的材质的值不符合设定的条件，确定该非标准产品信息的发布产品的标题，通过 K- 中心点等聚类算法运算该发布产品的标题与每一子集合中的各标准产品信息的发布产品的标题的相似度，为减少运算量，可以与每一子集合中的一个标准产品信息的发布产品的标题进行相似运算，查找出相似度达到第三阈值的标准产品信息所在的子集合，并将该非标准产品信息划分为相似度最高的标准产品信息所在的子集合。

[0069] 步骤 107 ：将该非标准产品信息划分至特定集合，并跳转至步骤 111。

[0070] 在本实施例中，由于存在某些非标准产品信息的用于划分集合的产品属性参数的值不符合设定的条件且非产品属性参数不与任何集合内的子集合相关，则为这一类非标准产品信息单独设置一个特定集合。

[0071] 该特定集合与步骤 102 中划分的集合满足类与类之间不相似的原则。

[0072] 步骤 108 ：确定非标准产品信息中用于划分集合的产品属性参数中的值符合设定的条件的产品属性参数。

[0073] 在本步骤中，由于非标准产品信息中部分用于划分集合的产品属性参数中的值符合设定的条件，因此，可以利用值符合设定的条件的这部分产品属性参数来划分非标准产品信息。

[0074] 例如 ：若在步骤 101 中确定的用于划分集合的产品属性参数为产品的价格、产品的材质，而某一非标准产品信息的产品的价格的值不符合设定的条件，但产品的材质的值符合设定的条件，则可以利用产品材质这一产品属性参数来划分非标准产品信息。

[0075] 步骤 109 ：根据步骤 108 确定的产品属性参数，确定非标准产品信息所属的集合。

[0076] 在本步骤中，根据确定的产品属性参数的值以及各集合中的标准产品信息中的该产品属性参数的值，查找出与该非标准产品信息的产品属性参数的值相同的集合。

[0077] 确定的集合数量可能不止一个，在此情况下，可以将确定的所有集合执行后续步骤。例如 ：在步骤 101 中示例中产生了 12 个集合，假设本步骤中的产品属性参数是产品的材质，取值为材质 1，则本步骤确定的该非标准产品信息所属的集合可能为集合 1、集合 4、集合 7 以及集合 10 中的一个。

[0078] 步骤 110 ：根据非标准产品信息的非产品属性参数，将该非标准产品信息划分至步骤 109 确定的集合内的一个子集合中。

[0079] 通过步骤 109 确定了非标准产品信息应该属于的集合，但还不能最终确定该非标准产品信息应该属于哪一集合中的哪一子集合，因此，本步骤进一步采用非产品属性信息对非标准产品信息作进一步划分，将非标准产品信息划分至非产品属性参数的相似度最高的子集合中。

[0080] 在步骤 106 中划分非标准产品信息时，与划分至的子集合之间的相似度需要达到第三阈值，进而再选择相似度最高的子集合，而在本步骤中只需要从步骤 109 确定的集合中选择相似度最高的子集合即可，这是因为 ：在步骤 106 中，非标准产品信息的用于划分集合的产品属性参数都不符合设定的条件，因此，如果只选择与非标准产品信息的非产品属

性参数相关性最高的子集合，而不为相关性设置最低的门限值（即第三阈值），则可能出现非标准产品信息与任一子集合的相关性都很低，但仍旧选择一个子集合作为非标准产品信息归属的子集合，导致该非标准产品信息和同一子集合内的其他标准产品信息的相关性低。而在步骤 110 中，由于在步骤 109 中确定的集合是通过非标准产品信息的一个取值符合设定的条件的产品属性参数来确定的，因此，在步骤 110 中用于选择的集合与非标准产品信息有一定的相关性，进而从中选择的子集合与非标准产品信息的相关性也较高。

[0081] 步骤 111 ：确定重复的标准产品信息，并保留其中一个标准产品信息，去除剩余的标准产品信息，以及确定重复的非标准产品信息，并保留其中一个非标准产品信息，去除剩余的非标准产品信息。

[0082] 若两个标准产品信息间的产品属性参数和非产品属性参数都相同，则定义这两个标准产品信息是重复的标准产品信息。

[0083] 若两个非标准产品信息间的产品属性参数和非产品属性参数都相同，则定义这两个非标准产品信息是重复的非标准产品信息。

[0084] 步骤 111 是实现本申请目的的优选步骤，由于卖家用户在网站服务器上可能出现重复发布产品信息的情况，为了减少对重复产品信息进行处理所占用的资源，步骤 111 对标准产品信息和非标准产品信息进行去重操作。

[0085] 步骤 111 的去重操作可以在步骤 101 之前，或是步骤 101 ～步骤 110 之间的任意时刻执行，或是在步骤 110 之后执行。

[0086] 步骤 112 ：为每一子集合命名。

[0087] 在执行到步骤 110 时，实际上已经自动对产品信息进行了划分，得到了以集合为单位或是子集合为单位的标准产品单元 (standard product unit，SPU)。本步骤作为本实施例的优选步骤，是为每一个 SPU 进行命名或者打标签，也就是标识每一 SPU 代表的内容。

[0088] 本步骤的具体做法是 ：

[0089] 首先确定每一子集合的标准产品信息和非标准产品信息中的产品属性参数和非产品属性参数 ；

[0090] 然后统计确定该产品属性参数和非产品属性参数中的至少一个高频词 ；

[0091] 最后将该高频词用于为该子集合命名。

[0092] 后续买家用户搜索、查询以及向买家用户推荐产品信息时，可以通过子集合名称中涉及的高频词作为关键字进行搜索或推荐。

[0093] 步骤 112 可以在步骤 110 之后且步骤 111 之前执行。

[0094] 实施例二 ：

[0095] 本申请实施例二提供一种与实施例一属于同一发明构思下的信息处理的设备，如图 3 所示，所述设备包括标准参数确定模块 11 和第一划分模块 12，其中 ：标准参数确定模块 11 用于确定叶子类目下各标准产品信息中的至少一个产品属性参数 ；第一划分模块 12 用于利用确定的所述至少一个产品属性参数对所述标准产品信息进行划分，得到所述叶子类目下的多个集合，其中，划分在同一集合内的标准产品信息中的所述至少一个产品属性参数的值都相同。

[0096] 如图 4 所示，所述标准参数确定模块具体包括属性参数确定子模块 21、次数确定子模块 22 和选择子模块 23，其中 ：属性参数确定子模块 21 用于确定叶子类目下各标准产

品信息中的全部产品属性参数；次数确定子模块 22 确定设定时长内每一产品属性参数作为搜索条件，在所述叶子类目下进行产品信息搜索的次数；选择子模块 23 用于从所述全部产品属性参数中选择至少一个产品属性参数，其中，选择的任一产品属性参数作为搜索条件进行产品信息搜索的次数达到第一阈值且该产品属性参数的值为离散型。

[0097] 所述设备还包括非标准参数确定模块 13、相似度确定模块 14 和第二划分模块 15，其中：非标准参数确定模块 13 用于针对划分后得到的每一集合，确定集合中各标准产品信息中的至少一个非产品属性参数；相似度确定模块 14 用于确定同一集合内各标准产品信息的至少一个非产品属性参数的相似度；第二划分模块 15 用于按照确定的相似度对同一集合内各标准产品信息做进一步划分，得到该集合下的多个子集合，其中，划分在同一子集合内的任意两个标准产品信息的所述至少一个非产品属性参数之间的相似度达到第二阈值。

[0098] 所述设备还包括判断模块 16，用于判断非标准产品信息中的用于划分集合的全部产品属性参数的值是否都不符合设定的条件，若是，则触发非标准参数确定模块 13，否则，触发标准参数确定模块 11，所述非标准产品信息中的用于划分集合的至少一个产品属性参数的值不符合设定的条件。

[0099] 根据判断模块 16 的触发，非标准参数确定模块 13 和标准参数确定模块 11 各自的运行过程如下：

[0100] 在非标准产品信息中的用于划分集合的全部产品属性参数的值都不符合设定的条件时：

[0101] 所述非标准参数确定模块 13 还用于确定该非标准产品信息中的至少一个非产品属性参数；所述相似度确定模块 14 用于确定非标准产品信息的非产品属性参数与各子集合内的标准产品信息中的非产品属性参数的相似度；所述第二划分模块 15 还用于根据非产品属性参数，将该非标准产品信息划分至一个子集合内，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度达到第三阈值。

[0102] 特殊地，非标准产品信息中的用于划分集合的全部产品属性参数的值不符合设定的条件，且该非标准产品信息的非产品属性参数与任一子集合内的标准产品信息中的非产品属性参数的相似度低于第三阈值时，所述第一划分模块 12 还用于在非标准产品信息中的用于划分集合的全部产品属性参数的值不符合设定的条件，且该非标准产品信息的非产品属性参数与任一子集合内的标准产品信息中的非产品属性参数的相似度低于第三阈值时，将该非标准产品信息划分至特定集合。

[0103] 在非标准产品信息中的用于划分集合的全部产品属性参数的值不全不符合设定的条件时：

[0104] 所述标准参数确定模块 11 还用于确定非标准产品信息中用于划分集合的产品属性参数中值符合设定的条件的产品属性参数；所述第一划分模块 12 还用于确定包含所述值符合设定的条件的产品属性参数的集合；所述第二划分模块 15 还用于根据非标准产品信息的非产品属性参数，将该非标准产品信息划分至第一划分模块 12 确定的集合内的一个子集合中，其中，该非标准产品信息的非产品属性参数与划分至的子集合内的标准产品信息中的非产品属性参数的相似度最高。

[0105] 所述设备还包括去重模块 17，用于确定重复的标准产品信息，保留其中一个标准产品信息，去除剩余的标准产品信息，以及，确定重复的非标准产品信息，保留其中一个非标准产品信息，去除剩余的非标准产品信息，所述重复的标准产品信息间的产品属性参数和非产品属性参数都相同，所述重复的非标准产品信息间的产品属性参数和非产品属性参数都相同。

[0106] 所述设备还包括命名模块 18，用于确定每一子集合的标准产品信息和非标准产品信息中的产品属性参数和非产品属性参数。统计确定该产品属性参数和非产品属性参数中的至少一个高频词，并利用确定的至少一个高频词为该子集合命名。

[0107] 本实施例二中的信息处理设备还可以包括能够执行实施例一中涉及的各项功能的模块。

[0108] 通过本申请实施例提供的方法及设备，可以在叶子类目的基础上，以产品属性参数为条件划分细粒度的产品信息集合，进一步地，还以非产品属性参数为条件划分更加细粒度的产品信息子集合，使得在买家用户搜索、查询产品信息时，以细粒度的产品信息集合为基础，可以有效减少搜索、查询的时间、提高搜索、查询的准确性以及提高网站服务器向买家用户推荐产品信息的准确度，且细粒度的产品信息集合也能够提高对产品信息进行操作的可用性，减少对产品信息进行操作时的运算量；并且，在集合和子集合的划分过程中，对多级类目体系的实质性内容没有改变，遵守现有的多级类目体系的特点，且充分考虑买家用户搜索习惯，将买家用户常用于搜索的产品属性参数用于划分集合；另外，对于非标准产品信息中的产品属性参数的值不符合设定的条件的各种情况给出了对应的划分手段，使得各非标准产品信息能够尽可能地划分至相关性高的子集合中；最后，本申请方案还对产品信息进行去重以及对划分后的子集合进行命名，有效地减少了对重复产品信息进行处理所占用的资源，以及方便管理员根据子集合的名称对子集合进行管理和以子集合的名称为关键字查找子集合内的产品信息。

[0109] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

[0110] 本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0111] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0112] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

[0113] 尽管已描述了本申请的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0114] 显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请的这些修改和变型属于本申请权利要求及其等同技术的范围之内，则本申请也意图包含这些改动和变型在内。

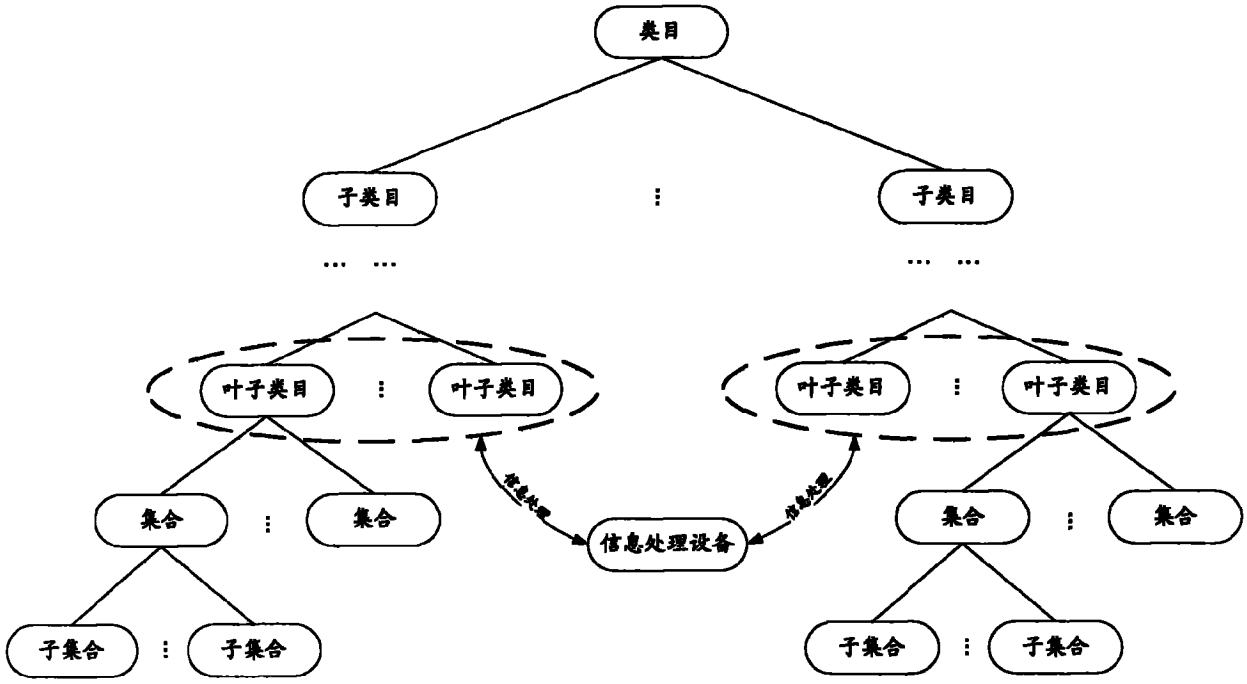

类目
子类目
子类目
叶子类目
叶子类目
叶子类目
叶子类目
集合
集合
集合
集合
信息处理设备
子集合
子集合
子集合
子集合

图 1

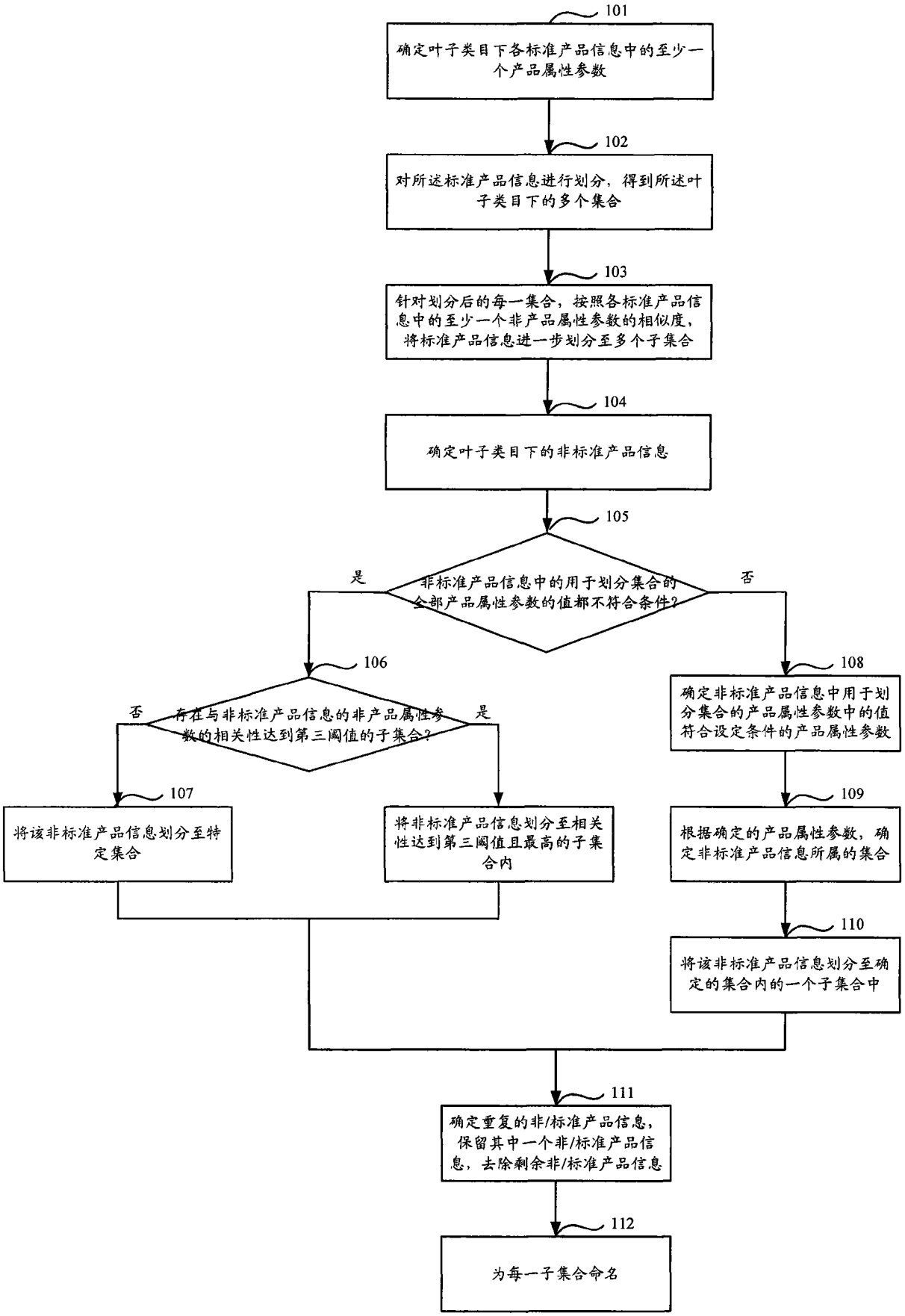

101
确定叶子类目下各标准产品信息中的至少一个产品属性参数
102
对所述标准产品信息进行划分，得到所述叶子类目下的多个集合
103
针对划分后的每一集合，按照各标准产品信息中的至少一个非产品属性参数的相似度，将标准产品信息进一步划分至多个子集合
104
确定叶子类目下的非标准产品信息
105
非标准产品信息中的用于划分集合的全部产品属性参数的值都不符合条件？
是
否
106
存在与非标准产品信息的非产品属性参数的相关性达到第三阈值的子集合？
否
是
107
将该非标准产品信息划分至特定集合
将非标准产品信息划分至相关性达到第三阈值且最高的子集合内
108
确定非标准产品信息中用于划分集合的产品属性参数中的值符合设定条件的产品属性参数
109
根据确定的产品属性参数，确定非标准产品信息所属的集合
110
将该非标准产品信息划分至确定的集合内的一个子集合中
111
确定重复的非/标准产品信息，保留其中一个非/标准产品信息，去除剩余非/标准产品信息
112
为每一子集合命名

图 2

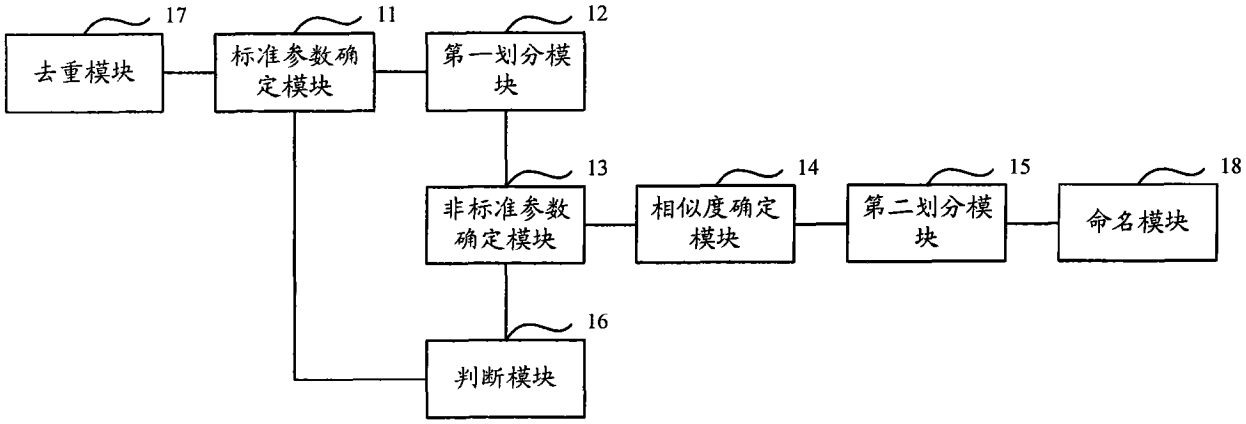

17
去重模块
11
标准参数确定模块
12
第一划分模块
13
非标准参数确定模块
14
相似度确定模块
15
第二划分模块
18
命名模块
16
判断模块

图 3

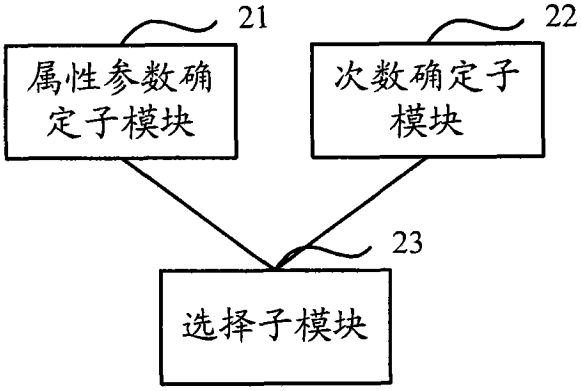

21
属性参数确定子模块
22
次数确定子模块
23
选择子模块

图 4