基于电子商务的数据处理方法与装置

技术领域

本公开涉及数据挖掘技术领域，具体而言，涉及一种基于电子商务的数据处理方法与装置。

背景技术

随着电商业务的发展，传统的“千人一面”搜索推荐系统已不能有效的满足用户需求，且我国幅员辽阔，各地域在气候、习俗、环境等方面存在较大的差异。

目前电商的搜索系统主要根据商品与用户搜索关键词的文本相关性、商品本身信息质量等维度对商品进行展示排序，不涉及地域特征；商品推荐系统则主要依据用户过往行为、平台促销活动、人工运营等方式确定推荐商品，也没有将地域特征纳入推荐因子。因此，在现有的数据处理模式下，往往存在着搜索结果不能精准的贴近用户需求等问题。例如北方空调大部分需冷暖模式，而在华南地区大部分只需制冷模式，当华南地区的用户搜索空调时很难获取到精准贴合需求的搜索结果。此外，不纳入地域特征的推荐，也会导致流量转换损失，甚至引起用户反感，例如某个时期防雾霾口罩在北方热销，但推荐系统却将该类产品推荐给海南等地的用户。最后，在地方性传统节假日期间，地方特产、服饰等具有区域性的高销量，不纳入地域特征的搜索推荐系统对此“无能为力”。

因此，需要一种能够对商品的地域特征进行挖掘的数据处理方法。

需要说明的是，在上述背景技术部分公开的信息仅用于加强对本公开的背景的理解，因此可以包括不构成对本领域普通技术人员已知的现有技术的信息。

发明内容

本公开的目的在于提供一种基于电子商务的数据处理方法与装置，用于从用户的搜索行为日志以及商品的物流信息中，通过对数据进行清理、集成、计算等处理，输出关键词的地域特征画像，给搜索、推荐、广告系统提供基础数据支撑。

根据本公开实施例的第一方面，提供一种基于电子商务的数据处理方法，包括：获取数据，数据包括用户搜索日志和物流信息；根据数据获取基于地域的关键词权重值降序排名；根据基于地域的关键词权重值降序排名获取关键词在各地域的特征值；根据特征值标注关键词对应的热点地域。

在本公开的一种示例性实施例中，获取基于地域的关键词权重值降序排名包括：根据搜索日志获取基于地域的关键词搜索 PV；根据物流信息获取基于地域的关键词商品数；基于地域将关键词搜索 PV 与第一系数的乘积和关键词商品数与第二系数的乘积相加作为关键词在地域的权重值；去除权重值低于阈值的关键词，基于地域对关键词按权重值进行降序排名。

在本公开的一种示例性实施例中，根据基于地域的关键词权重值降序排名获取关键词在各地域的特征值包括：获取地域的总权重值降序排名；获取基于全部地域的关键词权重值降序排名；对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数；基于每一关键词以及每一地域计算特征值：（一地域的一关键词的权重值/地域的总权重值）*（总地域数/关键词在地域排名前 N 的地域数）。

在本公开的一种示例性实施例中，标注关键词对应的热点地域包括：获取一关键词在各地域的特征值的方差；去除方差小于阈值的地域，获取剩余地域的方差降序排名；根据方差降序排名标注关键词对应的热点地域。

在本公开的一种示例性实施例中，获取数据包括去除数据中的爬虫数据、黑名单用户数据、黑名单 IP 数据、无法判断来源的数据以及长尾关键词。

根据本公开的一个方面，提供一种基于电子商务的数据处理装置，包括：数据清洗模块，设置为获取数据，数据包括用户搜索日志和物流信息；数据集成模块，设置为根据数据获取基于地域的关键词权重值降序排名；数据计算模块，设置为根据基于地域的关键词权重值降序排名获取关键词在各地域的特征值；数据标注模块，设置为根据特征值标注关键词对应的热点地域。

在本公开的一种示例性实施例中，数据集成模块包括：元素获取单元，设置为根据搜索日志获取基于地域的关键词搜索 PV，以及根据物流信息获取基于地域的关键词商品数；权重值计算单元，设置为基于地域将关键词搜索 PV 与第一系数的乘积和关键词商品数与第二系数的乘积相加作为关键词在地域的权重值；权重值排名单元，设置为去除权重值低于阈值的关键词，基于地域对关键词按权重值进行降序排名。

在本公开的一种示例性实施例中，数据计算模块包括：第一权重值计算单元，设置为获取地域的总权重值降序排名；第二权重值计算单元，设置为获取基于全部地域的关键词权重值降序排名；关键词筛选单元，设置为对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数；计算单元，设置为基于每一关键词以及每一地域计算特征值：（一地域的一关键词的权重值/地域的总权重值）*（总地域数/关键词在地域排名前 N 的地域数）。

在本公开的一种示例性实施例中，数据标注模块包括：方差计算单元，设置为获取一关键词在各地域的特征值的方差；地域排序单元，设置为去除方差小于阈值的地域，获取剩余地域的方差降序排名；地域标注单元，设置为根据方差降序排名标注关键词对应的热点地域。

在本公开的一种示例性实施例中，数据清洗模块设置为去除数据中的爬虫数据、黑名单用户数据、黑名单 IP 数据、无法判断来源的数据以及长尾关键词。

根据本公开的一个方面，提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现上述任意一项的方法步骤。

根据本公开的一个方面，提供一种电子设备，包括存储器；以及耦合到所属存储器的

处理器，处理器被配置为基于存储在存储器中的指令，执行如上述任意一项的方法。

本公开提供的数据处理方法与装置通过对搜索行为及物流信息进行数据清理、集成、特征值计算、热点地域标注等处理，能够真实准确的挖掘出关键词的地域特征，生成关键词地域特征画像，并通过数据滚动保证所挖掘数据的时效性，最终为搜索推荐等业务提供数据支持，有助于构建“千人千面”的个性化搜索推荐系统。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本公开。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本公开的实施例，并与说明书一起用于解释本公开的原理。显而易见地，下面描述中的附图仅仅是本公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 示意性示出本公开示例性实施例中数据处理方法的流程图。

图 2 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S104 的子流程图。

图 3 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S106 的子流程图。

图 4 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S108 的子流程图。

图 5 意性示出本公开一个示例性实施例中一种数据处理装置的方框图。

图 6 意性示出本公开一个示例性实施例中数据处理装置的工作流程示意图。

图 7 意性示出本公开一个示例性实施例中另一种数据处理装置的方框图。

具体实施方式

现在将参考附图更全面地描述示例实施方式。然而，示例实施方式能够以多种形式实施，且不应被理解为限于在此阐述的范例。相反，提供这些实施方式使得本公开将更加全面和完整，并将示例实施方式的构思全面地传达给本领域的技术人员。所描述的特征、结构或特性可以以任何合适的方式结合在一个或更多实施方式中。在下面的描述中，提供许多具体细节从而给出对本公开的实施方式的充分理解。然而，本领域技术人员将意识到，可以实践本公开的技术方案而省略特定细节中的一个或更多，或者可以采用其它的方法、组元、装置、步骤等。在其它情况下，不详细示出或描述公知技术方案以避免喧宾夺主而使得本公开的各方面变得模糊。

此外，附图仅为本公开的示意性图解，图中相同的附图标记表示相同或类似的部分，因而将省略对它们的重复描述。附图中所示的一些方框图是功能实体，不一定必须与物理或逻辑上独立的实体相对应。可以采用软件形式来实现这些功能实体，或在一个或多个硬件模块或集成电路中实现这些功能实体，或在不同网络和/或处理器装置和/或微控制器装置中实现这些功能实体。

下面结合附图对本公开示例实施方式进行详细说明。

图 1 示意性示出本公开示例性实施例中数据处理方法的流程图。

参考图 1，数据处理方法 100 可以包括：

步骤 S102，获取数据，数据包括用户搜索日志和物流信息。

步骤 S104，根据数据获取基于地域的关键词权重值降序排名。

步骤 S104，根据基于地域的关键词权重值降序排名获取关键词在各地域的特征值。

步骤 S106，根据特征值标注关键词对应的热点地域。

数据处理方法 100 主要涉及数据清洗、数据集成、关键词地域特征值计算、关键词画像等流程。整个计算流程全部采用分布式计算框架，从而可以提高海量数据处理能力和数据计算时效性。

本公开提供的数据处理方法与装置通过对搜索行为及物流信息进行数据清理、集成、特征值计算、热点地域标注等处理，能够真实准确的挖掘出关键词的地域特征，生成关键词地域特征画像，并通过数据滚动保证所挖掘数据的时效性，最终为搜索推荐等业务提供数据支持，有助于构建“千人千面”的个性化搜索推荐系统。

下面对数据处理方法 100 的各步骤进行详细说明。

在步骤 S102，获取用户搜索日志和物流信息数据包括从数据仓库中获取，也包括从系统实时日志流信息和实时物流信息中获取。步骤 S102 也可以称为数据清洗步骤，在此步骤中，输入的数据包括用户搜索日志和物流信息，输出的数据包括合法搜索日志和物流信息。对数据进行清洗的流程可以为去除爬虫数据、去除黑名单用户 ID 的数据、去除黑名单 IP 数据、去除无法判断来源的数据以及去除长尾关键词。其中，长尾关键词是指搜索频率低于阈值、搜索量波动较大的关键词。上述数据清洗流程的顺序以及内容仅为示例性的，本领域相关技术人员可以根据实际情况对数据进行清洗以及整理。

图 2 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S104 的子流程图。

参考图 2，步骤 S104 包括：

步骤 S1042，根据搜索日志获取基于地域的关键词搜索 PV。

步骤 S1044，根据物流信息获取基于地域的关键词商品数。

步骤 S1046，基于地域将关键词搜索 PV 与第一系数的乘积和关键词商品数与第二系数的乘积相加作为关键词在地域的权重值。

步骤 S1048，去除权重值低于阈值的关键词，基于地域对关键词按权重值进行降序排名。

步骤 S104 可以被称为数据集成步骤。在此步骤中，输入数据为步骤 S104 输出的搜索日志和物流信息数据，输出数据为基于地域的关键词权重值排序，例如格式为关键词-地域-权重值-序号的表格。

在步骤 S1042 中，可以从搜索日志中统计出格式为关键词-地域-搜索 PV 的列表，表示为一个地域的一个商品种类的搜索数量。

搜索 PV（PageView，页面浏览量）是用户使用搜索接口搜索关键词的次数，用户每使用一次搜索接口计一个 PV。地域是指根据搜索日志获取的用户 IP 所在地域，其具体可以为国家、地区、行政省等分类方式，也可以为其他可以用于区分地域的分类方式，本公开对此不作特殊限定。但是可以理解的是，本公开所提及的“地域”不论遵从哪一种分类方式，均保持为同一种分类方式。

在步骤 S1044 中，可以从物流信息中统计出格式为关键词-地域-商品数的列表，表示为一地域的一个商品种类的实际购买数量。

在步骤 S1046 中，可以将步骤 S1042 与步骤 S1044 的结果按比例求并集，基于地域将一个关键词的搜索 PV 与第一系数的乘积和商品数与第二系数的乘积相加作为该关键词在该地域的权重值，并输出格式为关键词-地域-权重值的列表。上述第一系数与第二系数可以相等也可以为不等，本公开对此不作特殊限定。例如，当关键词“毛巾”在地域“北京”的搜索 PV 为 10000，且发货到“北京”的“毛巾”数量为 1000 时，设置第一系数为0.2，第二系数为0.8，则关键词“毛巾”在地域“北京”的权重为10000*0.2+1000*0.8=2800。设置第一系数以及第二系数的目的是根据不同商品之间搜索-购买的比例来调节商品的权重值。例如“衣服”的搜索-购买比例往往明显大于“冰箱”的搜索-购买比例，此时通过设置系数对各商品的搜索-购买比例进行调整可以更真实反映出商品的实际权重。

在步骤 S1048 中，首先需要去除权重值低于阈值的数据，从而不再对关注度低的商品进行统计。阈值的数值可以自由设置。其次可以根据步骤 S1046 输出的列表按照权重值降序排序，输出格式为关键词-地域-权重值-序号的列表。

图 3 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S106 的子流程图。

参考图 3，步骤 S106 包括：

步骤 S1062，获取地域的总权重值降序排名。

步骤 S1064，获取基于全部地域的关键词权重值降序排名。

步骤 S1066，对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数。

步骤 S1068，基于每一关键词以及每一地域计算 TF-IDF 值：

（一地域的一关键词的权重值/地域的总权重值）*（总地域数/关键词在地域排名前 N 的地域数）。

步骤 S106 的输入数据为步骤 S104 输出的关键词-地域-权重值-序号数据，输出数据为格式为关键词-地域-权重值-TF-IDF 值的列表。

在步骤 S1062 中，统计基于全部关键词的各地域总权重值，输出格式为地域-权重值的列表。

在步骤 S1064 中，统计基于全部地域的各关键词总权重值，并对各关键词基于总权重值降序排列，输出格式为关键词-权重值-序号的列表。

在步骤 S1066 中，首先可以对各地域提取排名前 N 的关键词，输出格式为关键词-地

域-权重值的列表；然后根据步骤 S1064 输出的列表提取在全部地域排名前 xN 的关键词，输出格式为关键词-权重值的列表。其中 N 是自然数，x 为扩展系数，在一些实施例中，x 例如可以等于 10。获取以上两个列表后，对二者取交集，从而对每个地域获取权重值既在地域排名前 N 又在全部地域范围内排名前 xN 的关键词，并输出格式为关键词-地域-权重值的列表。

通过进一步筛选，可以对更有地域代表性的关键词进行统计，提高数据处理效率。

在步骤 S1066 中，根据步骤 S1062~S1064 的输出结果计算各关键词在各地域的特征值。

在本公开的一种示例性实施例中，上述特征值可以为 TF-IDF 值。

TF-IDF 值是指 TF*IDF。其中，TF(Term Frequency，词频)表示词条 t 在文档 d 中出现的频率。IDF(Inverse Document Frequency，逆向文件频率)表示包含词条 t 的文档越少，词条 t 的类别区分能力越强。

在本公开的一实施例中，计算 TF-IDF 值的公式可以被设置为：

（一地域的一关键词的权重值/该地域的总权重值）*（总地域数/该关键词在地域排名前 N 的地域数）　　（1）

上式涉及到的地域和关键词均为步骤 S1064 输出列表中存在的地域和关键词。其中，一地域的一关键词的权重值为根据步骤 S104 输出的关键词-地域-权重值-序号列表数据获取的在一个地域内一个关键词的总权重值；该地域的总权重值的数据来源为步骤 S1062 输出的地域-权重值的列表；总地域数为根据据步骤 S104 输出的关键词-地域-权重值-序号数据获取的地域数量，或者根据系统设置获取的地域数量；该关键词在地域排名前 N 的地域数为根据步骤S1066获取的关键词-地域-权重值的列表获取的与该关键词有关联的地域数量。

一地域的一关键词的权重值与该地域的总权重值的比值可以表示一关键词在一地域的出现频率，该比值越大越说明该关键词在该地域中出现频率高；总地域数与该关键词在地域排名前 N 的地域数的比值可以表示该关键词的出现频率是否有地域特殊性，该比值越大越说明该关键词的出现有地域特殊性。因此由式（1）可以得知：出现频率越大、地域特殊性越大的关键词的 TF-IDF 值越高，即对于该地域的地域特征越明显。

经过计算后，步骤 S1066 输出格式为关键词-地域-权重值-TF-IDF 值的列表。通过使用 TF-IDF 算法对关键词的地域特征进行计算，可以有效规避各区域绝对数据大小的影响，使本方法的计算结果更加准确。

在本公开的其他示例性实施例中，TF-IDF 算法也可以由空间向量余弦算法等算法替代，只要使用计算关键词显著特征的算法实施本方法的技术方案皆在本公开保护范围之内。

图 4 示意性示出本公开示例性实施例中数据处理方法 100 中步骤 S108 的子流程图。

参考图 4，步骤 S108 包括：

步骤 S1082，获取一关键词在各地域的特征值的方差。

步骤 S1084，去除方差小于阈值的地域，获取剩余地域的方差降序排名。

步骤 S1086，根据方差降序排名标注关键词对应的热点地域。

步骤 S108 的输入数据为步骤 S1066 输出的关键词-地域-权重值-特征值列表，输出格式为“关键词-热点地域 1.地域 2…地域 N”的列表。

在步骤 S1082 中，统计关键词在不同地域特征值的方差。此步骤主要目的是统计关键词在一个地域的地域特征是否与平均值相比有明显差异。

在步骤 S1084 中，对各方差进行处理。首先去除方差小于阈值的地域，即剔除地域特征接近平均值的地域。上述阈值的设置可根据实际情况调整。接下来可以将剩余地域按方差降序排序。

在步骤 S1086 中，根据方差降序排序对该关键词标注热点地域，即具有明显地域特征的地域。可以对热点地域的数量进行限定，也可以标记出所有方差在阈值以上的地域，本领域相关技术人员可以根据实际情况自行设置。

重复步骤 S108，即可对每个关键词标注其对应的热点地域。标注的结果可以以数据图表、地图等形式展现，也可以作为内部数据为搜索、推荐、广告系统等提供数据支持。

综上，数据处理方法 100 通过对搜索行为及物流信息进行数据清理、集成、特征值计算、热点地域标注等处理，能够真实准确的挖掘出关键词的地域特征，生成关键词地域特征画像，并通过数据滚动保证所挖掘数据的时效性，最终为搜索推荐等业务提供数据支持，有助于构建“千人千面”的个性化搜索推荐系统。

对应于上述方法实施例，本公开还提供一种数据处理装置，可以用于执行上述方法实施例。

图 5 意性示出本公开一个示例性实施例中一种数据处理装置的方框图。

参考图 5，数据处理装置 500 可以包括：

数据清洗模块 502，设置为获取数据，数据包括用户搜索日志和物流信息。

数据集成模块 504，设置为根据数据获取基于地域的关键词权重值降序排名。

数据计算模块 506，设置为根据基于地域的关键词权重值降序排名获取关键词在各地域的特征值。

数据标注模块 508，设置为根据特征值标注关键词对应的热点地域。

在本公开的一种示例性实施例中，数据清洗模块 502 设置为去除数据中的爬虫数据、黑名单用户数据、黑名单 IP 数据、无法判断来源的数据以及长尾关键词。

在本公开的一种示例性实施例中，数据集成模块 504 包括：

元素获取单元 5042，设置为根据搜索日志获取基于地域的关键词搜索 PV，以及根据物流信息获取基于地域的关键词商品数。

权重值计算单元 5044，设置为基于地域将关键词搜索 PV 与第一系数的乘积和关键词商品数与第二系数的乘积相加作为关键词在地域的权重值。

权重值排名单元 5046，设置为去除权重值低于阈值的关键词，基于地域对关键词按权重值进行降序排名。

在本公开的一种示例性实施例中，数据计算模块 506 包括：

第一权重值计算单元 5062，设置为获取地域的总权重值降序排名。

第二权重值计算单元 5064，设置为获取基于全部地域的关键词权重值降序排名。

关键词筛选单元 5066，设置为对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数。

计算单元 5068，设置为基于每一关键词以及每一地域计算特征值：

（一地域的一关键词的权重值/地域的总权重值）*（总地域数/关键词在地域排名前 N 的地域数）。

在本公开的一种示例性实施例中，数据标注模块 508 包括：

方差计算单元 5082，设置为获取一关键词在各地域的特征值的方差。

地域排序单元 5084，设置为去除方差小于阈值的地域，获取剩余地域的方差降序排名。

地域标注单元 5086，设置为根据方差降序排名标注关键词对应的热点地域。

由于装置 500 的各功能已在其对应的方法实施例中予以详细说明，本公开于此不再赘述。

图 6 意性示出本公开一个示例性实施例中数据处理装置 500 的工作流程示意图。

参考图 6，数据清洗模块 502 从数据仓库中获取搜索行为数据以及物流信息数据，并将筛选后的数据发送给数据集成模块 504；数据集成模块 504 将筛选后的搜索行为数据以及物流信息数据集成为基于地域的关键词权重值列表，并将该列表输出给数据计算模块 506；数据计算模块 506 根据该列表计算关键词对应于地域的特征值，并将计算结果输出给数据标注模块 508；数据标注模块 508 对数据计算模块 506 输出的各关键词标注其对应的热点地域，并将标注结果发送给搜索系统、推荐系统、广告系统以及其他系统作为数据支持。

根据本公开的一个方面，提供一种数据处理装置，包括：

存储器；以及

耦合到所属存储器的处理器，处理器被配置为基于存储在存储器中的指令，执行如上述任意一项的方法。

该实施例中的装置的处理器执行操作的具体方式已经在有关该数据处理方法的实施例中执行了详细描述，此处将不做详细阐述说明。

图 7 是根据一示例性实施例示出的一种装置 700 的框图。装置 700 可以是智能手机、平板电脑等移动终端。

参照图 7，装置 700 可以包括以下一个或多个组件：处理组件 702，存储器 704，电源组件 706，多媒体组件 708，音频组件 710，传感器组件 714 以及通信组件 716。

处理组件 702 通常控制装置 700 的整体操作，诸如与显示，电话呼叫，数据通信，相机操作以及记录操作相关联的操作等。处理组件 702 可以包括一个或多个处理器 718 来执行指令，以完成上述的方法的全部或部分步骤。此外，处理组件 702 可以包括一个或多个模块，便于处理组件 702 和其他组件之间的交互。例如，处理组件 702 可以包括多媒体模块，以方便多媒体组件 708 和处理组件 702 之间的交互。

存储器 704 被配置为存储各种类型的数据以支持在装置 700 的操作。这些数据的示例包括用于在装置 700 上操作的任何应用程序或方法的指令。存储器 704 可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。存储器 704 中还存储有一个或多个模块，该一个或多个模块被配置成由该一个或多个处理器 718 执行，以完成上述任一所示方法中的全部或者部分步骤。

电源组件 706 为装置 700 的各种组件提供电力。电源组件 706 可以包括电源管理系统，一个或多个电源，及其他与为装置 700 生成、管理和分配电力相关联的组件。

多媒体组件 708 包括在装置 700 和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与触摸或滑动操作相关的持续时间和压力。

音频组件 710 被配置为输出和/或输入音频信号。例如，音频组件 710 包括一个麦克风（MIC），当装置 700 处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器 704 或经由通信组件 716 发送。在一些实施例中，音频组件 710 还包括一个扬声器，用于输出音频信号。

传感器组件 714 包括一个或多个传感器，用于为装置 700 提供各个方面的状态评估。例如，传感器组件 714 可以检测到装置 700 的打开/关闭状态，组件的相对定位，传感器组件 714 还可以检测装置 700 或装置 700 一个组件的位置改变以及装置 700 的温度变化。在一些实施例中，该传感器组件 714 还可以包括磁传感器，压力传感器或温度传感器。

通信组件 716 被配置为便于装置 700 和其他设备之间有线或无线方式的通信。装置 700 可以接入基于通信标准的无线网络，如 WiFi，2G 或 3G，或它们的组合。在一个示例性实施例中，通信组件 716 经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，通信组件 716 还包括近场通信（NFC）模块，以促进短程通信。例如，在 NFC 模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

在示例性实施例中，装置 700 可以被一个或多个应用专用集成电路（ASIC）、数字信

号处理器（DSP）、数字信号处理设备（DSPD）、可编程逻辑器件（PLD）、现场可编程门阵列（FPGA）、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

在本公开的一种示例性实施例中，还提供了一种计算机可读存储介质，其上存储有程序，该程序被处理器执行时实现如上述任意一项的数据处理方法。该计算机可读存储介质例如可以为包括指令的临时性和非临时性计算机可读存储介质。

本领域技术人员在考虑说明书及实践这里公开的发明后，将容易想到本公开的其它实施方案。本申请旨在涵盖本公开的任何变型、用途或者适应性变化，这些变型、用途或者适应性变化遵循本公开的一般性原理并包括本公开未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的，本公开的真正范围和构思由权利要求指出。

工业实用性

本公开提供的数据处理方法与装置通过对搜索行为及物流信息进行数据清理、集成、特征值计算、热点地域标注等处理，能够真实准确的挖掘出关键词的地域特征，生成关键词地域特征画像，并通过数据滚动保证所挖掘数据的时效性，最终为搜索推荐等业务提供数据支持，有助于构建“千人千面”的个性化搜索推荐系统。

权利要求

1. 一种基于电子商务的数据处理方法，其特征在于，包括：

获取数据，所述数据包括用户搜索日志和物流信息；

根据所述数据获取基于地域的关键词权重值降序排名；

根据所述基于地域的关键词权重值降序排名获取关键词在各地域的特征值；

根据所述特征值标注关键词对应的热点地域。

2. 如权利要求 1 所述的数据处理方法，其特征在于，所述获取基于地域的关键词权重值降序排名包括：

根据所述搜索日志获取基于地域的关键词搜索 PV；

根据所述物流信息获取基于地域的关键词商品数；

基于地域将所述关键词搜索 PV 与第一系数的乘积和所述关键词商品数与第二系数的乘积相加作为所述关键词在所述地域的权重值；

去除权重值低于阈值的关键词，基于地域对关键词按所述权重值进行降序排名。

3. 如权利要求 1 所述的数据处理方法，其特征在于，根据所述基于地域的关键词权重值降序排名获取关键词在各地域的特征值包括：

获取地域的总权重值降序排名；

获取基于全部地域的关键词权重值降序排名；

对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数；

基于每一关键词以及每一地域计算特征值：

（一地域的一关键词的权重值/所述地域的总权重值）*（总地域数/所述关键词在地域排名前 N 的地域数）。

4. 如权利要求 1 所述的数据处理方法，其特征在于，所述标注关键词对应的热点地域包括：

获取一关键词在各地域的特征值的方差；

去除方差小于阈值的地域，获取剩余地域的方差降序排名；

根据所述方差降序排名标注所述关键词对应的热点地域。

5. 如权利要求 1 所述的数据处理方法，其特征在于，所述获取数据包括去除所述数据中的爬虫数据、黑名单用户数据、黑名单 IP 数据、无法判断来源的数据以及长尾关键词。

6. 一种基于电子商务的数据处理装置，其特征在于，包括：

数据清洗模块，设置为获取数据，所述数据包括用户搜索日志和物流信息；

数据集成模块，设置为根据所述数据获取基于地域的关键词权重值降序排名；

数据计算模块，设置为根据所述基于地域的关键词权重值降序排名获取关键词在各地域的特征值；

数据标注模块，设置为根据所述特征值标注关键词对应的热点地域。

7. 如权利要求 6 所述的数据处理装置，其特征在于，所述数据集成模块包括：

元素获取单元，设置为根据所述搜索日志获取基于地域的关键词搜索 PV，以及根据所述物流信息获取基于地域的关键词商品数；

权重值计算单元，设置为基于地域将所述关键词搜索 PV 与第一系数的乘积和所述关键词商品数与第二系数的乘积相加作为所述关键词在所述地域的权重值；

权重值排名单元，设置为去除权重值低于阈值的关键词，基于地域对关键词按所述权重值进行降序排名。

8. 如权利要求 6 所述的数据处理装置，其特征在于，所述数据计算模块包括：

第一权重值计算单元，设置为获取地域的总权重值降序排名；

第二权重值计算单元，设置为获取基于全部地域的关键词权重值降序排名；

关键词筛选单元，设置为对于各地域，获取权重值既在地域排名前 N 又在全部地域排名前 xN 的关键词，N 为自然数，x 为扩展系数；

计算单元，设置为基于每一关键词以及每一地域计算特征值：

（一地域的一关键词的权重值/所述地域的总权重值）*（总地域数/所述关键词在地域排名前 N 的地域数）。

9. 如权利要求 6 所述的数据处理装置，其特征在于，所述数据标注模块包括：

方差计算单元，设置为获取一关键词在各地域的特征值的方差；

地域排序单元，设置为去除方差小于阈值的地域，获取剩余地域的方差降序排名；

地域标注单元，设置为根据所述方差降序排名标注所述关键词对应的热点地域。

10. 如权利要求 6 所述的数据处理装置，其特征在于，所述数据清洗模块设置为去除所述数据中的爬虫数据、黑名单用户数据、黑名单 IP 数据、无法判断来源的数据以及长尾关键词。

11.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求 1-5 任一项所述的方法步骤。

摘要

本公开提供一种基于电子商务的数据处理方法与装置。数据处理方法包括：获取数据，所述数据包括用户搜索日志和物流信息；根据所述数据获取基于地域的关键词权重值降序排名；根据所述基于地域的关键词权重值降序排名获取关键词在各地域的特征值；根据所述特征值标注关键词对应的热点地域。本公开提供的基于电子商务的数据处理方法能够挖掘出关键词的地域特征。

（图 1）

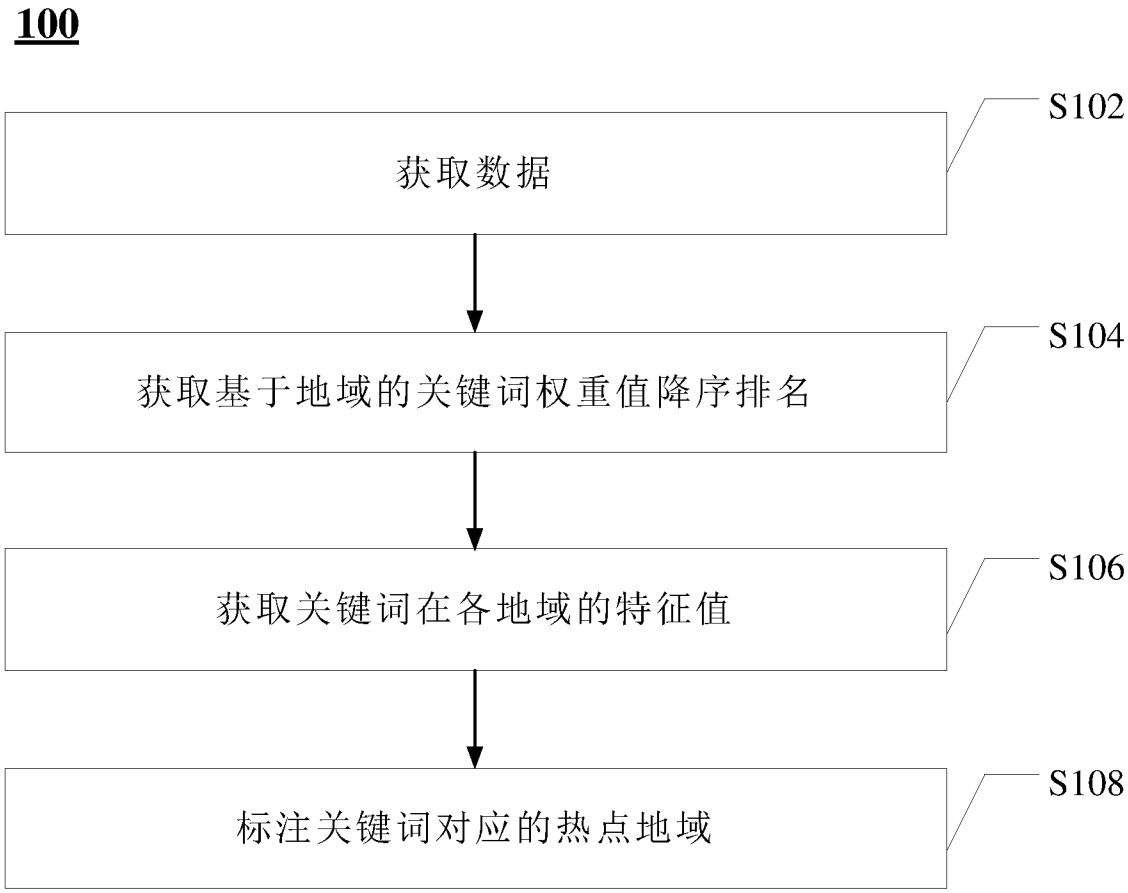
100
S102
获取数据
S104
获取基于地域的关键词权重值降序排名
S106
获取关键词在各地域的特征值
S108
标注关键词对应的热点地域

图 1

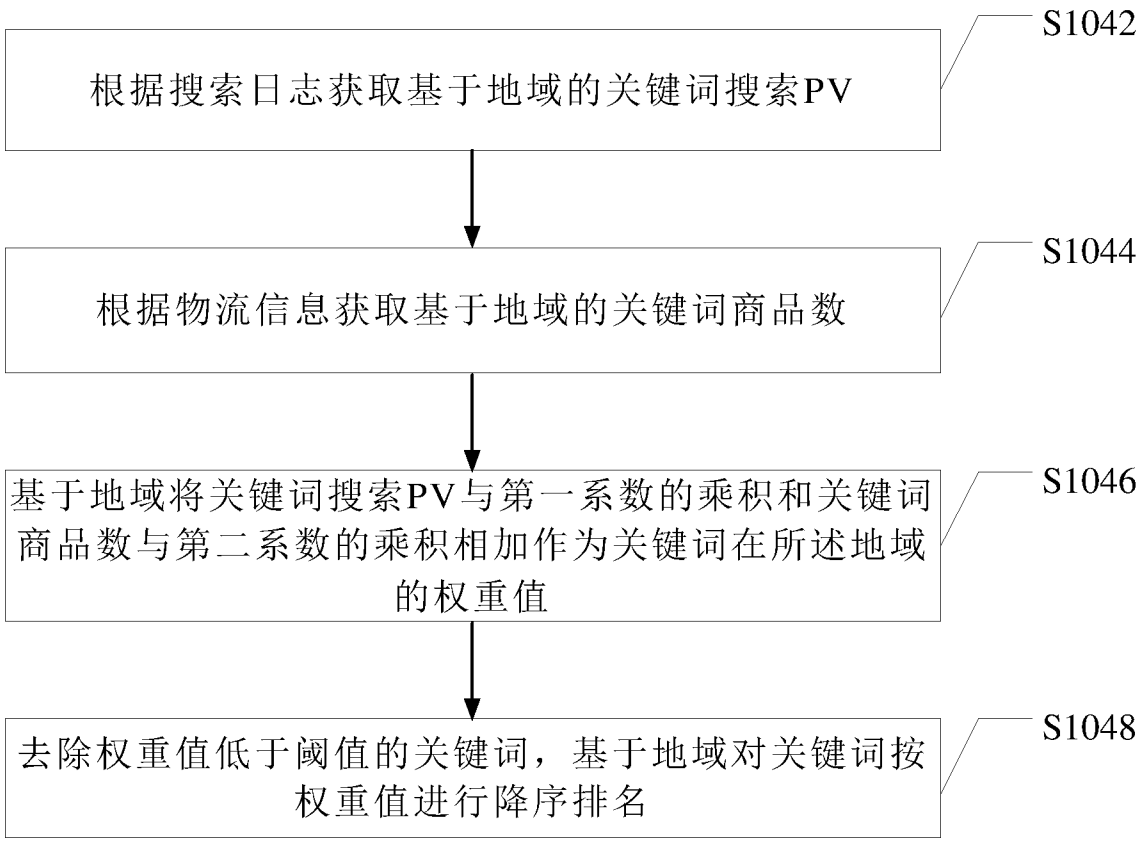
S104
S1042
根据搜索日志获取基于地域的关键词搜索PV
S1044
根据物流信息获取基于地域的关键词商品数
S1046
基于地域将关键词搜索PV与第一系数的乘积和关键词商品数与第二系数的乘积相加作为关键词在所述地域的权重值
S1048
去除权重值低于阈值的关键词，基于地域对关键词按权重值进行降序排名

图 2

S106

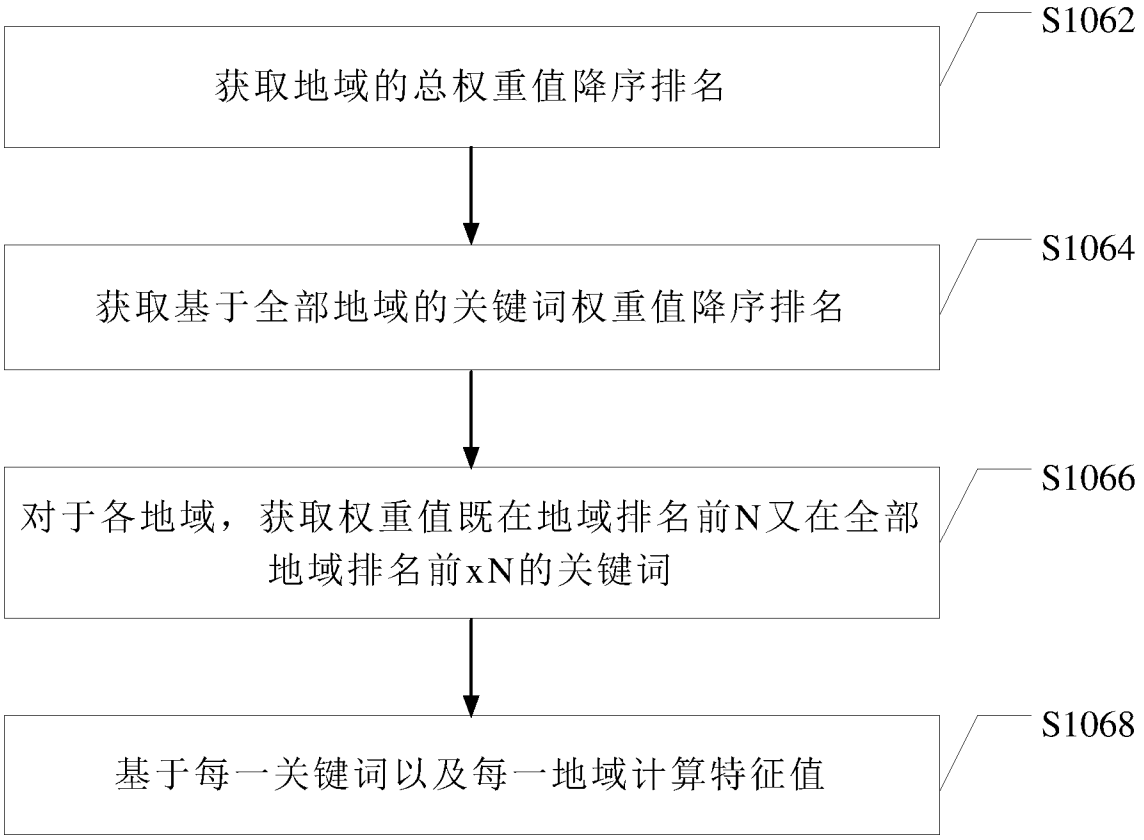

图 3

S108

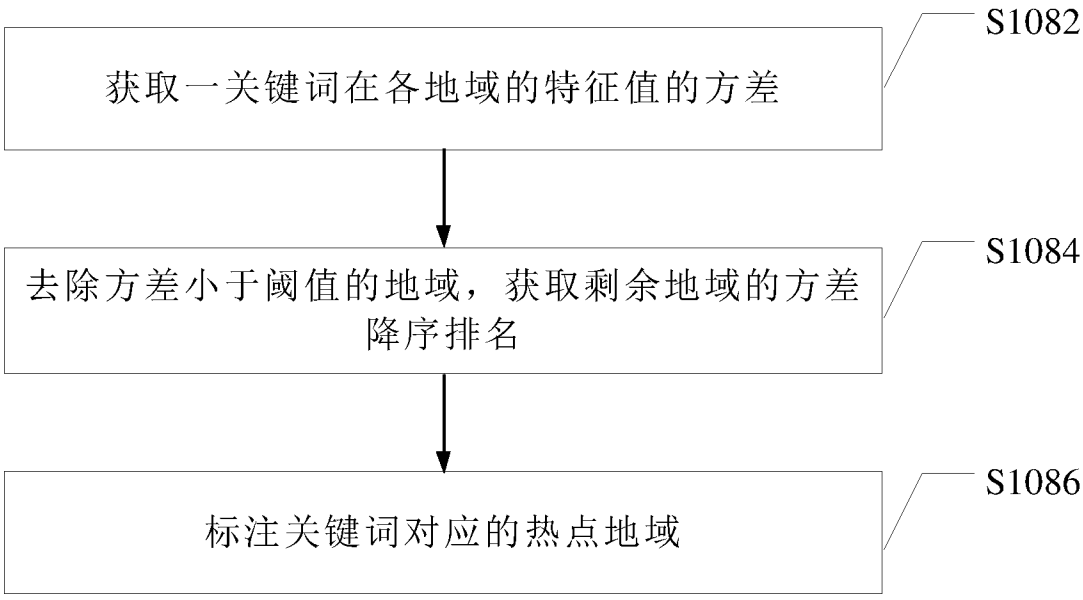

图 4

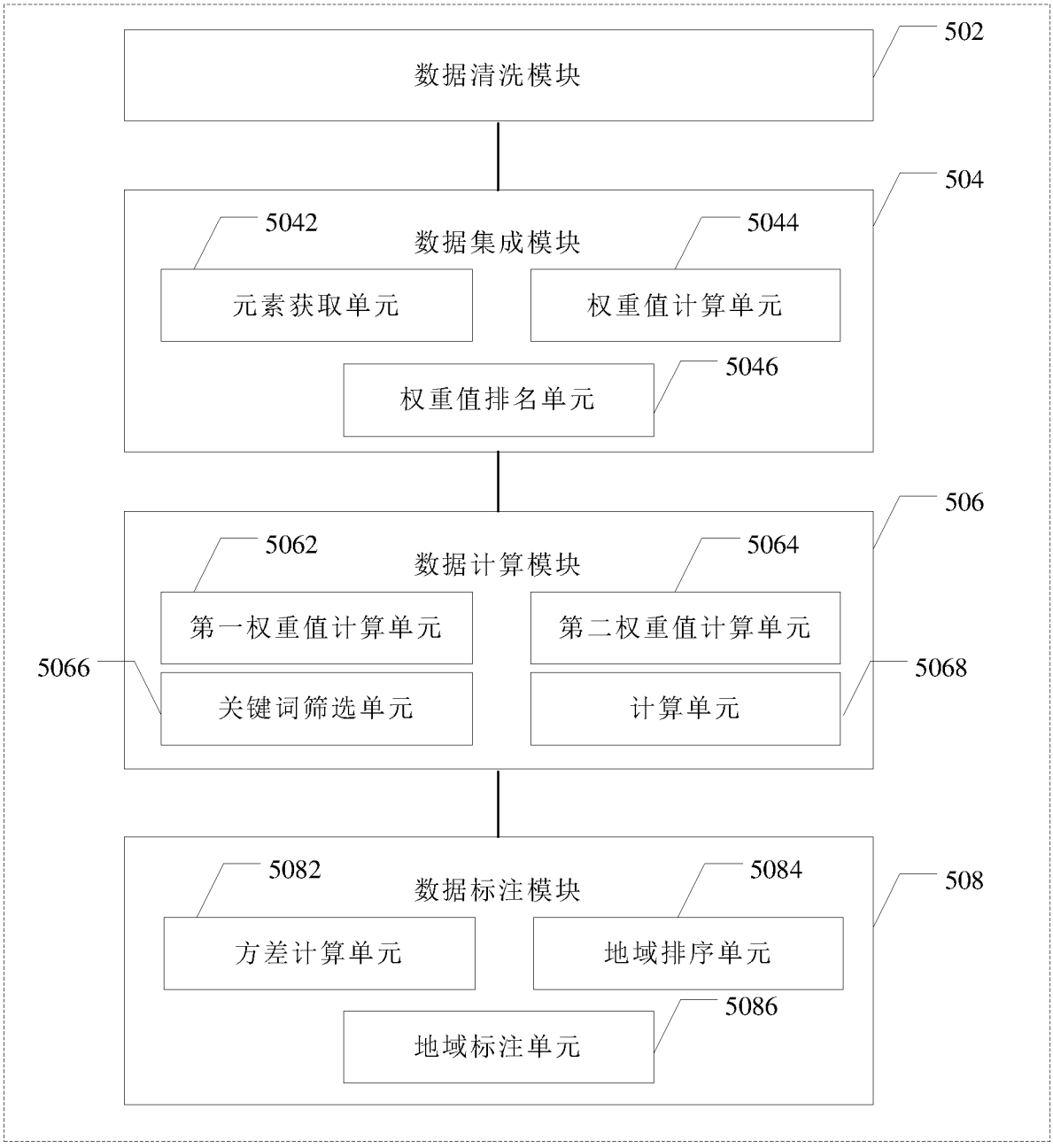
500
502
数据清洗模块
504
5042
数据集成模块
5044
元素获取单元
权重值计算单元
5046
权重值排名单元
506
5062
数据计算模块
5064
第一权重值计算单元
第二权重值计算单元
5066
5068
关键词筛选单元
计算单元
5082
数据标注模块
5084
508
方差计算单元
地域排序单元
5086
地域标注单元

图 5

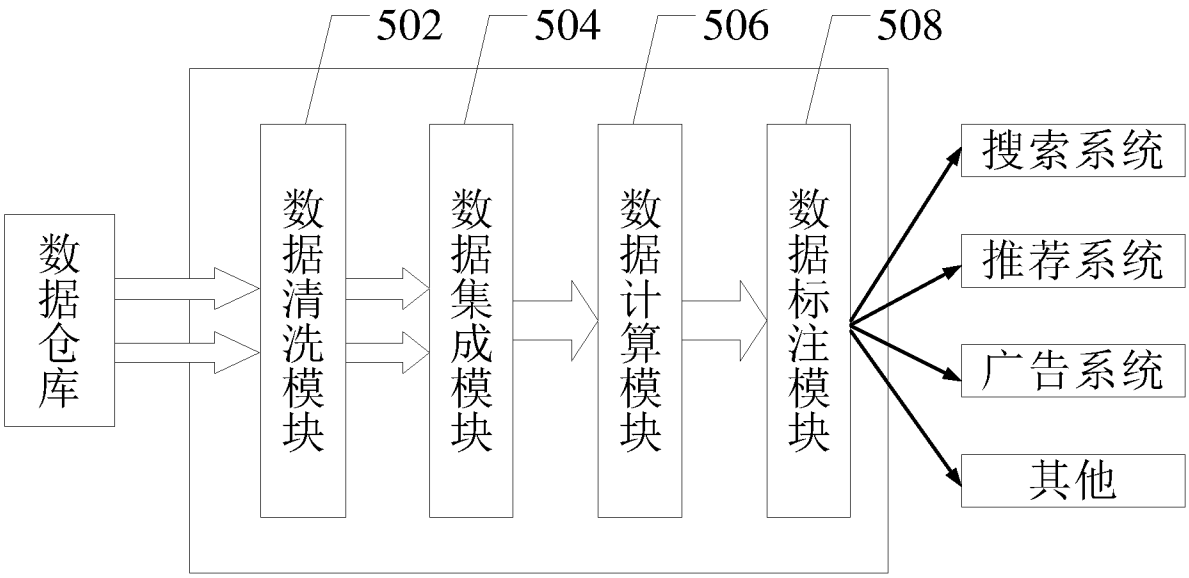
502
504
506
508
数据仓库
数据清洗模块
数据集成模块
数据计算模块
数据标注模块
搜索系统
推荐系统
广告系统
其他

图 6

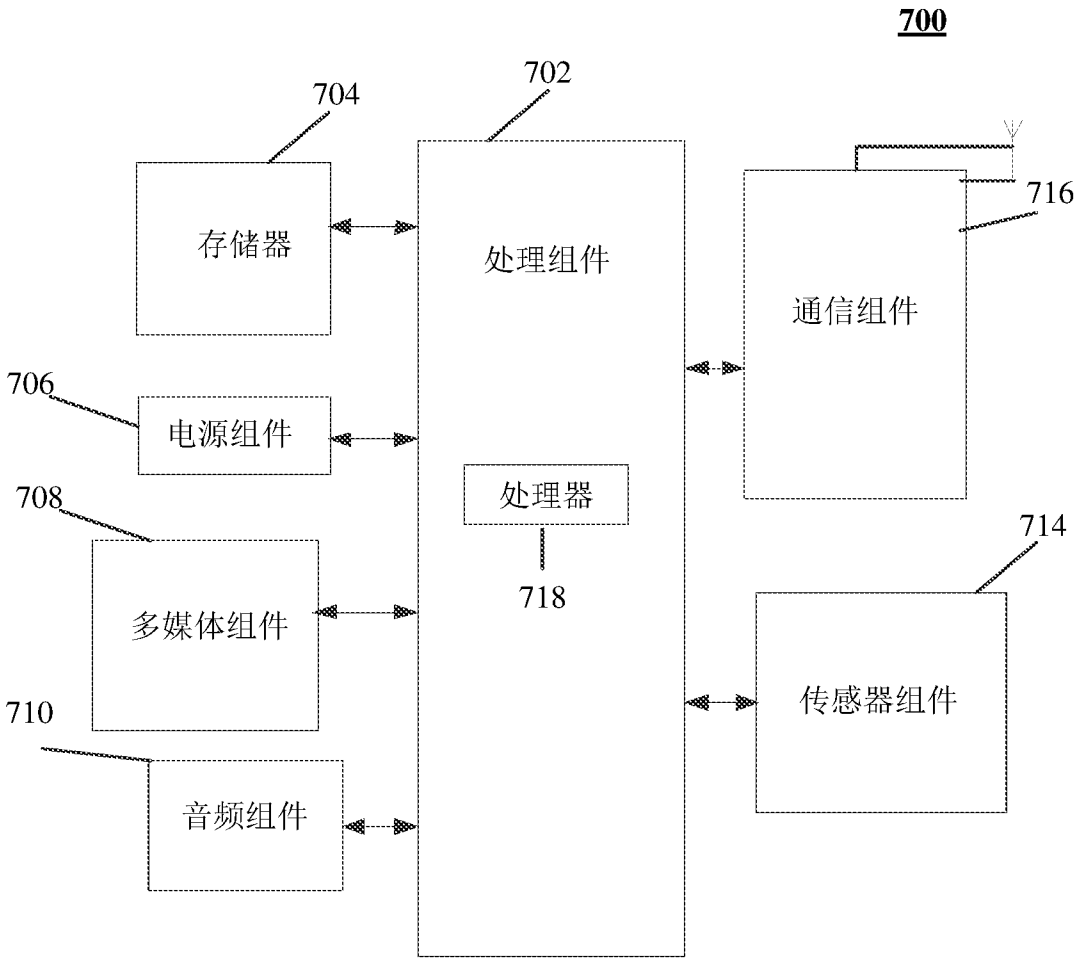
700
702
704
存储器
处理组件
716
通信组件
706
电源组件
处理器
718
708
多媒体组件
714
传感器组件
710
音频组件

图 7