



(12) 发明专利申请

(10) 申请公布号 CN 103942693 A

(43) 申请公布日 2014. 07. 23

(21) 申请号 201310019559. 4

(22) 申请日 2013. 01. 18

(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资本大厦一座四  
层 847 号邮箱

(72) 发明人 宋超 冯景华 张一楠 陈超

(74) 专利代理机构 北京润泽恒知识产权代理有  
限公司 11319

代理人 苏培华

(51) Int. Cl.

G06Q 30/00 (2012. 01)

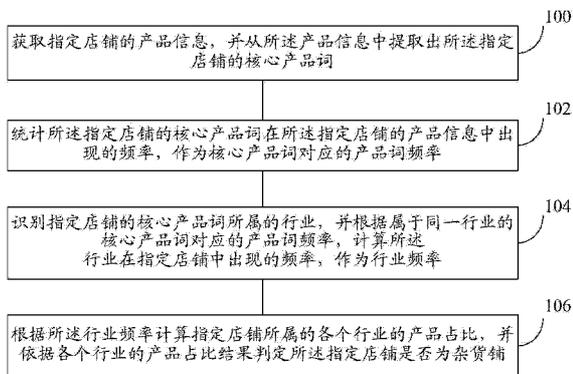
权利要求书3页 说明书12页 附图4页

(54) 发明名称

识别杂货铺的方法、装置及搜索店铺的方法、  
系统

(57) 摘要

本申请提供了一种识别杂货铺的方法及装置, 以实现准确识别杂货铺的目的, 避免因行业信息填写不准确或类目作弊导致的识别不准确的问题。其中一种识别杂货铺的方法包括: 获取指定店铺的产品信息, 并从所述产品信息中提取出所述指定店铺的核心产品词; 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率, 作为核心产品词对应的产品词频率; 识别所述指定店铺的核心产品词所属的行业, 并根据属于同一行业的核心产品词对应的产品词频率, 计算所述行业在所述指定店铺中出现的频率, 作为行业频率; 根据所述行业频率计算所述指定店铺所属的各个行业的产品占比, 并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。



1. 一种识别杂货铺的方法,其特征在于,包括:
  - 获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;
  - 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;
  - 识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率;
  - 根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。
2. 根据权利要求1所述的方法,其特征在于,所述从所述产品信息中提取出所述指定店铺的核心产品词,包括:
  - 从所述产品信息中提取出标题,并对所述标题进行切词,得到切词结果;
  - 将所述切词结果与核心产品词表进行匹配,匹配到的核心产品词作为所述指定店铺的核心产品词。
3. 根据权利要求1所述的方法,其特征在于,所述识别所述指定店铺的核心产品词所属的行业包括:
  - 统计核心产品词的行业点击率;
  - 将所述核心产品词的行业点击率与各行业的行业阈值进行匹配,判断所述核心产品词的行业点击率是否达到行业阈值;
  - 若所述核心产品词的行业点击率达到行业阈值,则判定该核心产品词属于该行业。
4. 根据权利要求1所述的方法,其特征在于,所述根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,包括:
  - 将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。
5. 根据权利要求1所述的方法,其特征在于,所述根据所述行业频率计算所述指定店铺所属的各个行业的产品占比包括:
  - 将所述行业的行业频率相加作为行业总频率;
  - 将所述行业的行业频率与所述行业总频率相除,相除的商作为该行业的产品占比。
6. 根据权利要求1所述的方法,其特征在于,所述依据各个行业的产品占比判定所述指定店铺是否为杂货铺包括:
  - 当指定店铺中有两个行业的产品占比超过阈值时,判定该店铺为杂货铺。
7. 一种搜索店铺的方法,其特征在于,包括:
  - 接收搜索关键词;
  - 查找与所述搜索关键词相匹配的店铺,得到候选店铺;
  - 将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出;
  - 所述杂货铺通过以下步骤识别:
    - 将所述候选店铺作为指定店铺,获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;
    - 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心

产品词对应的产品词频率；

识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率；

根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

8. 一种识别杂货铺的装置,其特征在于,包括：

提取模块,用于获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词；

产品词频率计算模块,用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率；

行业频率计算模块,用于识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率；

判定模块,用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

9. 根据权利要求8所述的装置,其特征在于,所述提取模块包括：

提取子模块,用于从所述产品信息中提取出标题,并对所述标题进行切词,得到切词结果；

匹配子模块,用于将所述切词结果与核心产品词表进行匹配,匹配到的核心产品词作为所述指定店铺的核心产品词。

10. 根据权利要求8所述的装置,其特征在于,所述行业频率计算模块包括：

行业识别子模块,用于统计核心产品词的行业点击率,并将所述核心产品词的行业点击率与各行业的行业阈值进行匹配,判断所述核心产品词的行业点击率是否达到行业阈值；若所述核心产品词的行业点击率达到行业阈值,则判定该核心产品词属于该行业；

计算子模块,用于将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。

11. 根据权利要求8所述的装置,其特征在于,所述判定模块包括：

产品占比计算子模块,用于将每个行业的行业频率相加作为行业总频率,并将每个行业的行业频率与所述行业总频率相除,相除的商作为该行业的产品占比；

判定子模块,用于当指定店铺中有两个行业的产品占比超过阈值时,判定该店铺为杂货铺。

12. 一种搜索店铺的系统,其特征在于,包括：

接收模块,用于接收搜索关键词；

查找模块,用于查找与所述搜索关键词相匹配的店铺,得到候选店铺；

排序模块,用于将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出；

所述杂货铺通过以下模块识别：

提取模块,用于将所述候选店铺作为指定店铺,获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词；

产品词频率计算模块,用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;

行业频率计算模块,用于识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率;

判定模块,用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

## 识别杂货铺的方法、装置及搜索店铺的方法、系统

### 技术领域

[0001] 本申请涉及搜索技术,特别是涉及一种识别杂货铺的方法、装置及搜索店铺的方法、系统。

### 背景技术

[0002] 目前电子商务网站(简称电商网站)提供了便利的产品信息以及供应商获取途径,但是这些电商网站提供的供应商实力良莠不齐,用户无法快速从海量的信息中找到有实力、专业性强的供应商。而这类专业性强的供应商往往经营领域比较集中,专注于做某个特定领域的产品,而不是从事多个领域、每个领域做的都不够专业的杂货铺。因此,需要将这些从事多个领域、不够专业的杂货铺从大量的店铺信息中识别出来,以提高搜索准确率。

[0003] 电子商务网站发布的产品信息中包含产品所属的类目,现有的杂货铺识别方法是直接统计电子商务网站发布的类目信息,以类目代表行业,根据每个行业的占比识别出杂货铺。

[0004] 但是,如果发布的产品信息没有填写准确的类目,或者进行类目作弊,例如:在发布的时候将店铺中涉及多个行业的产品类目全部设置为某个行业,上述识别方法将不能准确识别出杂货铺店铺。

### 发明内容

[0005] 本申请提供了一种识别杂货铺的方法及装置,以实现准确识别杂货铺的目的,避免因行业信息填写不准确或类目作弊导致的识别不准确的问题。

[0006] 相应的,本申请还提供了一种搜索店铺的方法及系统,在搜索的时候降低杂货铺的排序,从而提高搜索准确率。

[0007] 为了解决上述问题,本申请公开了一种识别杂货铺的方法,包括:

[0008] 获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;

[0009] 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;

[0010] 识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率;

[0011] 根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0012] 可选地,所述从所述产品信息中提取出所述指定店铺的核心产品词,包括:

[0013] 从所述产品信息中提取出标题,并对所述标题进行切词,得到切词结果;

[0014] 将所述切词结果与核心产品词表进行匹配,匹配到的核心产品词作为所述指定店铺的核心产品词。

[0015] 可选地,所述识别所述指定店铺的核心产品词所属的行业包括:

- [0016] 统计核心产品词的行业点击率；
- [0017] 将所述核心产品词的行业点击率与各行业的行业阈值进行匹配，判断所述核心产品词的行业点击率是否达到行业阈值；
- [0018] 若所述核心产品词的行业点击率达到行业阈值，则判定该核心产品词属于该行业。
- [0019] 可选地，所述根据属于同一行业的核心产品词对应的产品词频率，计算所述行业在所述指定店铺中出现的频率，包括：
- [0020] 将所述属于同一行业的核心产品词对应的产品词频率进行加和，作为该行业在所述指定店铺中出现的频率。
- [0021] 可选地，所述根据所述行业频率计算所述指定店铺所属的各个行业的产品占比包括：
- [0022] 将所述行业的行业频率相加作为行业总频率；
- [0023] 将所述行业的行业频率与所述行业总频率相除，相除的商作为该行业的产品占比。
- [0024] 可选地，所述依据各个行业的产品占比判定所述指定店铺是否为杂货铺包括：
- [0025] 当指定店铺中有两个行业的产品占比超过阈值时，判定该店铺为杂货铺。
- [0026] 本申请还公开了一种搜索店铺的方法，包括：
- [0027] 接收搜索关键词；
- [0028] 查找与所述搜索关键词相匹配的店铺，得到候选店铺；
- [0029] 将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出；
- [0030] 所述杂货铺通过以下步骤识别：
- [0031] 将所述候选店铺作为指定店铺，获取指定店铺的产品信息，并从所述产品信息中提取出所述指定店铺的核心产品词；
- [0032] 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率，作为核心产品词对应的产品词频率；
- [0033] 识别所述指定店铺的核心产品词所属的行业，并根据属于同一行业的核心产品词对应的产品词频率，计算所述行业在所述指定店铺中出现的频率，作为行业频率；
- [0034] 根据所述行业频率计算所述指定店铺所属的各个行业的产品占比，并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。
- [0035] 本申请还公开了一种识别杂货铺的装置，包括：
- [0036] 提取模块，用于获取指定店铺的产品信息，并从所述产品信息中提取出所述指定店铺的核心产品词；
- [0037] 产品词频率计算模块，用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率，作为核心产品词对应的产品词频率；
- [0038] 行业频率计算模块，用于识别所述指定店铺的核心产品词所属的行业，并根据属于同一行业的核心产品词对应的产品词频率，计算所述行业在所述指定店铺中出现的频率，作为行业频率；
- [0039] 判定模块，用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占

比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0040] 可选地,所述提取模块包括:

[0041] 提取子模块,用于从所述产品信息中提取出标题,并对所述标题进行切词,得到切词结果;

[0042] 匹配子模块,用于将所述切词结果与核心产品词表进行匹配,匹配到的核心产品词作为所述指定店铺的核心产品词。

[0043] 可选地,所述行业频率计算模块包括:

[0044] 行业识别子模块,用于统计核心产品词的行业点击率,并将所述核心产品词的行业点击率与各个行业的行业阈值进行匹配,判断所述核心产品词的行业点击率是否达到行业阈值;若所述核心产品词的行业点击率达到行业阈值,则判定该核心产品词属于该行业;

[0045] 计算子模块,用于将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。

[0046] 可选地,所述判定模块包括:

[0047] 产品占比计算子模块,用于将每个行业的行业频率相加作为行业总频率,并将每个行业的行业频率与所述行业总频率相除,相除的商作为该行业的产品占比;

[0048] 判定子模块,用于当指定店铺中有两个行业的产品占比超过阈值时,判定该店铺为杂货铺。

[0049] 本申请还公开了一种搜索店铺的系统,包括:

[0050] 接收模块,用于接收搜索关键词;

[0051] 查找模块,用于查找与所述搜索关键词相匹配的店铺,得到候选店铺;

[0052] 排序模块,用于将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出;

[0053] 所述杂货铺通过以下模块识别:

[0054] 提取模块,用于将所述候选店铺作为指定店铺,获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;

[0055] 产品词频率计算模块,用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;

[0056] 行业频率计算模块,用于识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率;

[0057] 判定模块,用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0058] 与现有技术相比,本申请包括以下优点:

[0059] 本申请实施例提供的识别杂货铺的方法,首先从指定店铺的产品信息中提取出核心产品词,其次,统计核心产品词在指定店铺中的产品词频率,并基于用户的搜索点击行为挖掘出核心产品词对应的行业;最后,根据行业频率计算指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。由于本申请实施例是根据从产品信息中提取到的核心产品词识别出产品分布,再根据产品分布识别出所属的行业,而不是直接使用用户填写的行业信息,因此可以避免卖家行业分布信息填写不准确或

类目作弊导致的杂货铺识别不准确的问题,提高了识别杂货铺的准确率。

[0060] 本申请实施例提供的搜索店铺的方法,可以在搜索的时候将这些从事多个领域,不够专业的杂货铺从大量的店铺中识别出来,并降低其排序,从而提高搜索的准确率。

[0061] 当然,实施本申请的任一产品不一定需要同时达到以上所述的所有优点。

#### 附图说明

[0062] 图 1 是本申请实施例所述一种识别杂货铺的方法的流程图;

[0063] 图 2 是本申请实施例所述指定店铺的产品信息示意图;

[0064] 图 3 是本申请实施例所述指定店铺的一条产品信息示意图;

[0065] 图 4 是本申请实施例所述一种识别杂货铺的装置的结构框图;

[0066] 图 5 是本申请实施例所述一种搜索店铺的方法的流程图;

[0067] 图 6 是本申请实施例所述一种搜索店铺的系统的结构框图。

#### 具体实施方式

[0068] 为使本申请的上述目的、特征和优点能够更加明显易懂,下面结合附图和具体实施方式对本申请作进一步详细的说明。

[0069] 杂货铺是指从事多个领域,每个领域做的都不够专业的店铺。本申请就是将这些从事多个领域,不够专业的杂货铺从大量的店铺中识别出来。

[0070] 本申请是从卖家发布的产品信息的标题或其他商品描述信息中挖掘出核心产品词,并且基于用户的搜索点击行为挖掘出核心产品词对应的行业,最后统计行业的分布,根据行业分布识别出杂货铺。下面通过实施例进行详细说明。

[0071] 参照图 1,其示出了本申请实施例所述一种识别杂货铺的方法的流程图,本实施例具体可以包括以下步骤:

[0072] 步骤 100,获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;

[0073] 指定店铺是指本次要识别的某个店铺,没有特别指定,可以理解为是泛指某个待识别的店铺。

[0074] 如图 2 所示是一家指定店铺的产品信息示意图,店铺的每条产品信息通常包含标题、属性、类目、价格、图片、详情页面的描述信息等几个部分,产品信息是由店铺的卖家自行填写的。

[0075] 下面以图 3 所示的该指定店铺的一条产品信息为例进行说明。在图 3 所示的产品信息中,“新中长款 大码 毛衣 女装 宽松 休闲 蝙蝠..”是标题,¥25 是价格。行业信息没有展现出来,是用户从类目体系中选择的,例如图 3 的产品信息用户会指定到“毛衣”类目,属于“女装”行业。类目是一整个体系;例如“服装”下面有“女装”、“男装”、“童装”等类目,而“女装”类目下面又有“连衣裙”、“毛衣”、“牛仔裤”、“T 恤”、“羽绒服”、“皮衣”等类目。

[0076] 本实施例中可以采用以下方式从所述产品信息中提取出所述指定店铺的核心产品词:

[0077] 首先,从所述产品信息中提取出标题,并对所述标题进行切词,得到切词结果;

[0078] 所谓切词,是指将一个汉字序列切分成一个一个单独的词。例如,从图 3 所示的产品信息中提取出标题“新中长款 大码 毛衣 女装 宽松 休闲 蝙蝠..”,然后对标题进行切词,结果为“新中长款、大码、毛衣、女装、宽松、休闲、蝙蝠”。

[0079] 其次,将所述切词结果与核心产品词表进行匹配,匹配到的核心产品词作为所述指定店铺的核心产品词。

[0080] 核心产品词表记录了能够标识产品的词,可以通过训练模型获得,也可以通过经验人工标注。例如,“连衣裙”、“起重机”、“玩具”等能够标识产品的词都位于核心产品词表中,而“女式连衣裙”中的“女式”为产品修饰词,并不在核心产品词表中。

[0081] 将上述切词结果中的词与核心产品词表进行匹配,在核心产品词表中出现的词作为核心产品词。例如,上述切词结果中出现在核心产品词表中的词为“毛衣”,即图 3 所示的产品信息中,核心产品词为“毛衣”。

[0082] 同理,可以从图 2 所示的指定店铺的产品信息中,取到核心产品词“毛衣”、“针织衫”、“蝙蝠衫”、“收纳盒”、“挂袋”。

[0083] 需要说明的是,本申请也可以采用其他的核心产品词提取方式,本申请的保护范围不应限定于上述实施例。

[0084] 步骤 102,统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;

[0085] 例如,在图 2 所示的指定店铺中,核心产品词“毛衣”出现了 4 次,所以该核心产品词“毛衣”对应的产品词频率为 4。同理,核心产品词“针织衫”对应的产品词频率为 3,核心产品词“蝙蝠衫”对应的产品词频率为 1,核心产品词“收纳盒”对应的产品词频率为 3,核心产品词“挂袋”对应的产品词频率为 1。如表 1 所示:

[0086]

核心产品词	频率
毛衣	4
针织衫	3
蝙蝠衫	1
收纳盒	3
挂袋	1

[0087] 表 1,指定店铺中各核心产品词对应的产品词频率

[0088] 步骤 104,识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算各行业在所述指定店铺中出现的频率,作为行业频率;

[0089] 每个核心产品词都有其对应的行业,例如,核心产品词“连衣裙”对应的行业是“女装”,核心产品词“橘子”对应的行业是“水果”。

[0090] 在本实施例中识别所述指定店铺的核心产品词所属的行业具体可以通过以下方式实现,当然,本申请的保护范围不限于此识别方式:

[0091] 首先,统计核心产品词的行业点击率;

[0092] 所述行业点击率是通过统计用户搜索该核心产品词时点击的行业来识别的。例如,用户在搜索“毛衣”时,搜索引擎根据用户的搜索词,返回到产品的搜索列表,用户在搜索列表中选择相关的产品,点击进入详情页面。在这个过程中大部分用户点击的产品都属于“服装”行业,因此可以通过统计大规模的用户点击行为,来得到行业点击率。

[0093] 其次,将所述核心产品词的行业点击率与各行业的行业阈值进行匹配,判断所述核心产品词的行业点击率是否达到行业阈值;若所述核心产品词的行业点击率达到行业阈值,则判定该核心产品词属于该行业。

[0094] 每个行业都有其对应的行业阈值,当核心产品词的行业点击率达到行业阈值的时候,判定该核心产品词属于该行业,当所述核心产品词的行业点击率未达到行业阈值,则判定该核心产品词不属于该行业,作为噪音进行去除。例如,“服装”行业的行业阈值为 0.2,用户在搜索“服装”的时的点击记录如下:连衣裙 100 词,毛衣 80 次,羽绒服 50 次、童裙 3 次,则核心产品词“连衣裙”的行业点击率为  $0.4292(100/(100+80+50+3) = 0.4292)$ ,核心产品词“毛衣”的行业点击率为  $0.3433(80/(100+80+50+3) = 0.4292)$ ,核心产品词“羽绒服”的行业点击率为  $0.2146(50/(100+80+50+3) = 0.4292)$ ,核心产品词“挂袋”的行业点击率为  $0.0129(3/(100+80+50+3) = 0.4292)$ 。核心产品词“连衣裙”、“毛衣”和“羽绒服”的行业点击率都大于或等于“服装”行业的行业阈值为 0.2,因此“连衣裙”、“毛衣”和“羽绒服”都属于“服装”行业,而“挂袋”的行业点击率小于“服装”行业的行业阈值为 0.2,因此“挂袋”并不属于“服装”行业,应作为点击噪音去除掉。

[0095] 同理,采用上述识别方法,可以得知图 2 所示的指定店铺中,核心产品词“毛衣”、“针织衫”和“蝙蝠衫”都属于“服装”行业,核心产品词“收纳盒”和“挂袋”属于“家居用品”行业。

[0096] 在本实施例中可以将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。

[0097] 具体地,可以将属于同一行业的核心产品词对应的产品词频率相加,结果作为该行业在所指定店铺中出现的频率,例如图 2 所示的指定店铺中,将属于同一行业“服装”的核心产品词“毛衣”对应的产品词频率 4、核心产品词“针织衫”对应的产品词频率 3 和核心产品词“蝙蝠衫”对应的产品词频率 1 相加,结果 8 作为“服装”行业在指定店铺中出现的频率,即在指定店铺中“服装”行业的行业频率为 8。同理,可以计算得到指定店铺中“家居用品”的行业频率为 4( $3+1 = 4$ ),如表 2 所示:

[0098]

行业	频率
服装	8
家居用品	4

[0099] 表 2,指定店铺中各行业的行业频率

[0100] 需要说明的是,本实施例是以简单的相加求和为例进行说明的,实际应用时也可以采用其他方式如加权求和的方式来实现,将属于同一行业的核心产品词对应的产品词频

率进行加权以后求和,作为该行业的行业频率。

[0101] 步骤 106,根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0102] 本实施例中可以采用以下公式计算各个行业的产品占比:

$$[0103] \quad Ratio_i = \frac{f_i}{\sum_{i=1}^n f_i}$$

[0104] 其中,  $Ratio_i$  表示产品占比;  $f_i$  表示某一行业在指定店铺中出现的频率,即行业频率;  $\sum_{i=1}^n f_i$  表示指定店铺所属的所有行业的行业频率之和,即下文所说的行业总频率。

[0105] 本实施例中根据所述行业频率计算所述指定店铺所属的各个行业的产品占比具体可以通过以下方式实现:

[0106] 首先,将所述行业的行业频率相加作为行业总频率;

[0107] 例如,将表 2 所示的服装行业的行业频率 8 和家居用品的行业频率 4 相加,结果 12 作为行业总频率。

[0108] 其次,将所述行业的行业频率与所述行业总频率相除,相除的商作为该行业的产品占比。

[0109] 例如,将服装行业的行业频率 8 与行业总频率 12 相除,相除的商 0.67 ( $8/12 = 0.67$ ) 作为服装行业的产品占比;将家居用品行业的行业频率 4 与行业总频率 12 相除,相除的商 4 ( $4/12 = 0.33$ ) 作为服装行业的产品占比。

[0110] 本实施例中所述依据各个行业的产品占比判定所述指定店铺是否为杂货铺包括:当指定店铺中有两个行业的产品占比超过阈值时,判定该店铺为杂货铺。

[0111] 例如,在本实施例中,阈值设为 20%,在图 2 所示的指定店铺中,服装行业的产品占比为 0.67,即 67%,已经超过上述设定的阈值 20%;家居用品行业的产品占比为 0.33,即 33%,同样超过上述设定的阈值 20%,显然图 2 所示的指定店铺中,有两个行业的产品占比都超过了预设的阈值,因此判定图 2 所示指定店铺为杂货铺。

[0112] 综上所述,本申请实施例首先从指定店铺的产品信息中提取出核心产品词,其次,统计核心产品词在指定店铺中的产品词频率,并基于用户的搜索点击行为挖掘出核心产品词对应的行业;最后,根据行业频率计算指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。由于本申请实施例是根据从产品信息中提取到的核心产品词识别出产品分布,再根据产品分布识别出所属的行业,而不是直接使用用户填写的行业信息,可以避免卖家行业分布信息填写不准确或类目作弊导致的杂货铺识别不准确的问题,提高了识别杂货铺的准确率。

[0113] 基于上述识别杂货铺的方法的实施例的描述,本申请提供了相应的识别杂货铺的装置实施例,具体如下:

[0114] 参照图 4,其示出了本申请实施例所述一种识别杂货铺的装置的结构框图,本实施例具体可以包括以下模块:提取模块 10、产品词频率计算模块 12、行业频率计算模块 14 和判定模块 16,其中:

[0115] 提取模块 10,用于获取指定店铺的产品信息,并从所述产品信息中提取出所述指

定店铺的核心产品词；

[0116] 本实施例中提取模块 10 具体可以包括以下子模块：

[0117] 提取子模块，用于从所述产品信息中提取出标题，并对所述标题进行切词，得到切词结果；

[0118] 以图 3 所示的该指定店铺的一条产品信息为例进行说明，在图 3 所示的产品信息中，“新中长款 大码 毛衣 女装 宽松 休闲 蝙蝠..”是标题。

[0119] 所谓切词，是指将一个汉字序列切分成一个一个单独的词。例如，从图 3 所示的产品信息中提取出标题“新中长款 大码 毛衣 女装 宽松 休闲 蝙蝠..”，然后对标题进行切词，结果为“新中长款、大码、毛衣、女装、宽松、休闲、蝙蝠”。

[0120] 匹配子模块，用于将所述切词结果与核心产品词表进行匹配，匹配到的核心产品词作为所述指定店铺的核心产品词。

[0121] 核心产品词表记录了能够标识产品的词，可以通过经验人工标注。例如，“连衣裙”、“起重机”、“玩具”等能够标识产品的词都位于核心产品词表中，而“女式连衣裙”中的“女式”为产品修饰词，并不在核心产品词表中。

[0122] 将上述切词结果中的词与核心产品词表进行匹配，在核心产品词表中出现的词作为核心产品词，上述切词结果中出现在和核心产品词表中的词为“毛衣”，即图 3 所示的产品信息中，核心产品词为“毛衣”。

[0123] 同理，匹配子模块可以从图 2 所示的指定店铺的产品信息中，取到核心产品词“毛衣”、“针织衫”、“蝙蝠衫”、“收纳盒”、“挂袋”。

[0124] 产品词频率计算模块 12，用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率，作为核心产品词对应的产品词频率；

[0125] 例如，在图 2 所示的指定店铺中，核心产品词“毛衣”出现了 4 次，所以产品词频率计算模块 12 计算得到该核心产品词“毛衣”对应的产品词频率为 4。同理，产品词频率计算模块 12 计算得到核心产品词“针织衫”对应的产品词频率为 3，核心产品词“蝙蝠衫”对应的产品词频率为 1，核心产品词“收纳盒”对应的产品词频率为 3，核心产品词“挂袋”对应的产品词频率为 1。

[0126] 行业频率计算模块 14，用于识别所述指定店铺的核心产品词所属的行业，并根据属于同一行业的核心产品词对应的产品词频率，计算所述行业在所述指定店铺中出现的频率，作为行业频率；

[0127] 每个核心产品词都有其对应的行业，例如，核心产品词“连衣裙”对应的行业是“女装”，核心产品词“橘子”对应的行业是“水果”。

[0128] 本实施例中行业频率计算模块 14 具体可以包括以下子模块：

[0129] 行业识别子模块，用于统计核心产品词的行业点击率，并将所述核心产品词的行业点击率与各行业的行业阈值进行匹配，判断所述核心产品词的行业点击率是否达到行业阈值；若所述核心产品词的行业点击率达到行业阈值，则判定该核心产品词属于该行业；

[0130] 所述行业点击率是通过统计用户搜索该核心产品词时点击的行业来识别的。例如，用户在搜索“毛衣”时，搜索引擎根据用户的搜索词，返回到产品的搜索列表，用户在搜索列表中选择相关的产品，点击进入详情页面。在这个过程中大部分用户点击的产品都属于“服装”行业，因此可以通过统计大规模的用户点击行为，来得到行业点击率。

[0131] 每个行业都有其对应的行业阈值,当核心产品词的行业点击率达到行业阈值的时候,判定该核心产品词属于该行业,当所述核心产品词的行业点击率未达到行业阈值,则判定该核心产品词不属于该行业,作为噪音进行去除。例如,“服装”行业的行业阈值为 0.2,用户在搜索“服装”的时的点击记录如下:连衣裙 100 词,毛衣 80 次,羽绒服 50 次、童裙 3 次,则核心产品词“连衣裙”的行业点击率为  $0.4292(100/(100+80+50+3) = 0.4292)$ ,核心产品词“毛衣”的行业点击率为 0.3433,核心产品词“羽绒服”的行业点击率为 0.2146,核心产品词“挂袋”的行业点击率为 0.0129。核心产品词“连衣裙”、“毛衣”和“羽绒服”的行业点击率都大于或等于“服装”行业的行业阈值为 0.2,因此“连衣裙”、“毛衣”和“羽绒服”都属于“服装”行业,而“挂袋”的行业点击率小于“服装”行业的行业阈值为 0.2,因此“挂袋”并不属于“服装”行业,应作为点击噪音去除掉。

[0132] 同理,采用上述识别方法,可以得知图 2 所示的指定店铺中,核心产品词“毛衣”、“针织衫”和“蝙蝠衫”都属于“服装”行业,核心产品词“收纳盒”和“挂袋”属于“家居用品”行业。

[0133] 计算子模块,用于将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。

[0134] 在本实施例中可以将所述属于同一行业的核心产品词对应的产品词频率进行加和,作为该行业在所述指定店铺中出现的频率。

[0135] 具体地,计算子模块可以将属于同一行业的核心产品词对应的产品词频率相加,结果作为该行业在所示指定店铺中出现的频率,例如图 2 所示的指定店铺中,将属于同一行业“服装”的核心产品词“毛衣”对应的产品词频率 4、核心产品词“针织衫”对应的产品词频率 3 和核心产品词“蝙蝠衫”对应的产品词频率 1 相加,结果 8 作为“服装”行业在指定店铺中出现的频率,即在指定店铺中“服装”行业的行业频率为 8。同理,可以计算得到指定店铺中“家居用品”的行业频率为  $4(3+1 = 4)$ 。

[0136] 需要说明的是,本实施例是以简单的相加求和为例进行说明的,实际应用时计算子模块也可以采用加权求和的方式来实现,将属于同一行业的核心产品词对应的产品词频率进行加权以后求和,作为该行业的行业频率。

[0137] 判定模块 16,用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0138] 本实施例中判定模块 16 具体可以包括以下子模块:

[0139] 产品占比计算子模块,用于将每个行业的行业频率相加作为行业总频率,并将每个行业的行业频率与所述行业总频率相除,相除的商作为该行业的产品占比;

[0140] 例如,产品占比计算子模块将表 2 所示的服装行业的行业频率 8 和家居用品的行业频率 4 相加,结果 12 作为行业总频率。然后,产品占比计算子模块将服装行业的行业频率 8 与行业总频率 12 相除,相除的商  $0.67(8/12 = 0.67)$  作为服装行业的产品占比;将家居用品行业的行业频率 4 与行业总频率 12 相除,相除的商  $4(4/12 = 0.33)$  作为服装行业的产品占比。

[0141] 判定子模块,用于当指定店铺中有两个行业的产品占比超过阈值时,判定该店铺为杂货铺。

[0142] 例如,在本实施例中,阈值设为 20%,在图 2 所示的指定店铺中,服装行业的产品

占比为 0.67, 即 67%, 已经超过上述设定的阈值 20%; 家居用品行业的产品占比为 0.33, 即 33%, 同样超过上述设定的阈值 20%, 显然图 2 所示的指定店铺中, 有两个行业的产品占比都超过了预设的阈值, 因此判定子模块判定图 2 所示指定店铺为杂货铺。

[0143] 本申请实施例中提取模块 10 从指定店铺的产品信息中提取出核心产品词, 然后产品词频率计算模块 12 统计核心产品词在指定店铺中的产品词频率, 行业频率计算模块 14 基于用户的搜索点击行为挖掘出核心产品词对应的行业; 判定模块 16 根据行业频率计算指定店铺所属的各个行业的产品占比, 并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。由于本申请实施例是根据从产品信息中提取到的核心产品词识别出产品分布, 再根据产品分布识别出所属的行业, 而不是直接使用用户填写的行业信息, 可以避免卖家行业分布信息填写不准确或类目作弊导致的杂货铺识别不准确的问题, 提高了识别杂货铺的准确率。

[0144] 对于上述识别杂货铺的装置实施例而言, 由于其与方法实施例基本相似, 所以描述的比较简单, 相关之处参见图 1 所示识别杂货铺的方法实施例的部分说明即可。

[0145] 基于上述识别杂货铺的方法的实施例的描述, 本申请提供了相应的搜索店铺的方法实施例, 可以在搜索的时候降低杂货铺的排序, 从而提高搜索准确率, 具体如下:

[0146] 参照图 5, 其示出了本申请实施例所述一种搜索店铺的方法的流程图, 本实施例具体可以包括以下步骤:

[0147] 步骤 200, 接收搜索关键词;

[0148] 搜索关键词是指买家在搜索店铺的时候输入的关键词, 例如, 女装。

[0149] 步骤 202, 查找与所述搜索关键词相匹配的店铺, 得到候选店铺;

[0150] 候选店铺中包含与搜索关键词相关的产品, 例如搜索关键词为女装, 与女装相关的产品有连衣裙、女式毛衣和女式羽绒服等, 如果一家店铺中包含连衣裙、女式毛衣或女式羽绒服等产品, 则该店铺为候选店铺。

[0151] 步骤 204, 将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出;

[0152] 杂货铺是指从事多个领域, 每个领域做的都不够专业的店铺。本实施例需要在搜索的时候将这些从事多个领域, 不够专业的杂货铺从大量的店铺中识别出来, 并降低其排序, 从而提高搜索的准确率。

[0153] 本实施例中所述杂货铺具体可以通过以下步骤识别:

[0154] 步骤 100, 将所述候选店铺作为指定店铺, 获取指定店铺的产品信息, 并从所述产品信息中提取出所述指定店铺的核心产品词;

[0155] 步骤 102, 统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率, 作为核心产品词对应的产品词频率;

[0156] 步骤 104, 识别所述指定店铺的核心产品词所属的行业, 并根据属于同一行业的核心产品词对应的产品词频率, 计算所述行业在所述指定店铺中出现的频率, 作为行业频率;

[0157] 步骤 106, 根据所述行业频率计算所述指定店铺所属的各个行业的产品占比, 并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0158] 对于上述识别杂货铺的步骤而言, 由于在识别杂货铺的方法实施例中已经进行了

详细的描述,相关之处参见图 1 所示的识别杂货铺的方法实施例的说明即可,本实施例在此不做赘述。

[0159] 基于上述搜索店铺的方法实施例的描述,本申请提供了相应的搜索店铺的系统实施例,具体如下:

[0160] 参照图 6,其示出了本申请实施例所述一种搜索店铺的系统结构框图,本实施例具体可以包括以下模块:

[0161] 接收模块 20,用于接收搜索关键词;

[0162] 查找模块 22,用于查找与所述搜索关键词相匹配的店铺,得到候选店铺;

[0163] 排序模块 24,用于将所述候选店铺中识别为杂货铺的候选店铺排在未识别为杂货铺的候选店铺之后并输出;

[0164] 本实施例中所述杂货铺具体可以通过以下模块识别:

[0165] 提取模块 10,用于将所述候选店铺作为指定店铺,获取指定店铺的产品信息,并从所述产品信息中提取出所述指定店铺的核心产品词;

[0166] 产品词频率计算模块 12,用于统计所述指定店铺的核心产品词在所述指定店铺的产品信息中出现的频率,作为核心产品词对应的产品词频率;

[0167] 行业频率计算模块 14,用于识别所述指定店铺的核心产品词所属的行业,并根据属于同一行业的核心产品词对应的产品词频率,计算所述行业在所述指定店铺中出现的频率,作为行业频率;

[0168] 判定模块 16,用于根据所述行业频率计算所述指定店铺所属的各个行业的产品占比,并依据各个行业的产品占比判定所述指定店铺是否为杂货铺。

[0169] 对于上述识别杂货铺的模块而言,由于在识别杂货铺的装置实施例中已经进行了详细的描述,相关之处参见图 4 所示的识别杂货铺的装置实施例的说明即可,本实施例在此不做赘述。

[0170] 杂货铺是指从事多个领域,每个领域做的都不够专业的店铺。本实施例提供的搜索店铺的系统,可以在搜索的时候将这些从事多个领域,不够专业的杂货铺从大量的店铺中识别出来,并降低其排序,从而提高搜索的准确率。

[0171] 本说明书中的各个实施例均采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似的部分互相参见即可。

[0172] 本领域技术人员易于想到的是:上述各个实施例的任意组合应用都是可行的,故上述各个实施例之间的任意组合都是本申请的实施方案,但是由于篇幅限制,本说明书在此就不一一详述了。

[0173] 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。

[0174] 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。

[0175] 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。

[0176] 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。

[0177] 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。

[0178] 以上对本申请所提供的一种识别杂货铺的方法、装置及搜索店铺的方法、系统,进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的一般技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

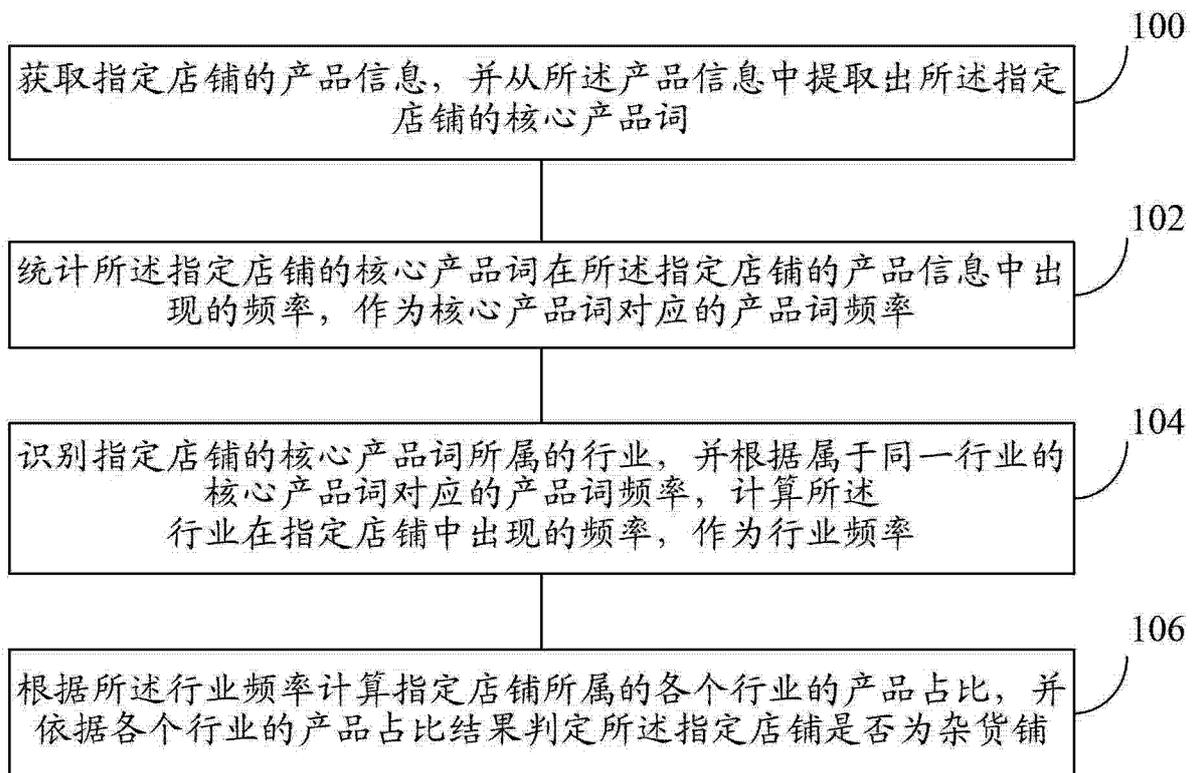


图 1

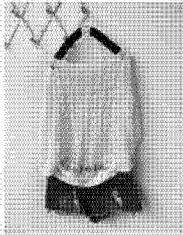
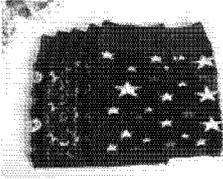
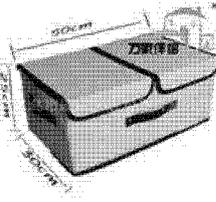
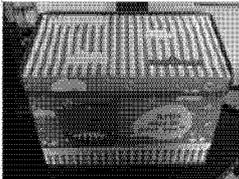
 <p>¥13.00 特价2012新款大码毛衣宽松一字领镂空蝙蝠衫针..</p>	 <p>¥25.00 新中长款 大码 毛衣 女装 宽松 休闲 蝙蝠 ..</p>	 <p>¥21.00 特价 复古 彩色条纹 宽松 镂空 套头 针织衫 ..</p>	 <p>¥41.00 2012秋冬韩版 菠萝针织外套 加绒加厚短款毛..</p>
 <p>¥36.00 秋装新款2012韩版女装V领蝙蝠袖宽松毛衣撞色..</p>	 <p>¥20.00 2012秋冬 韩版女装 镂空 爱心 撞色条纹 ..</p>	 <p>¥45.00 2011秋冬女装新款韩版宽松中长款条纹开衫加绒..</p>	 <p>¥20.00 2012热卖秋冬新款韩版A字裙针织毛线裙毛绒裙..</p>
 <p>¥13.50 新款 愤怒的小鸟收纳箱 小鸟箱 杂物收纳箱 玩..</p>	 <p>¥7.00 无纺布双盖毛衣收纳整理箱 收纳盒 收纳用品 厂..</p>	 <p>¥12.00 防水透气运动型 车用椅背袋 置物袋 汽车收纳袋 ..</p>	 <p>¥20.00 现货小房子收纳箱 新款小房子收纳盒 新款收纳箱..</p>

图 2

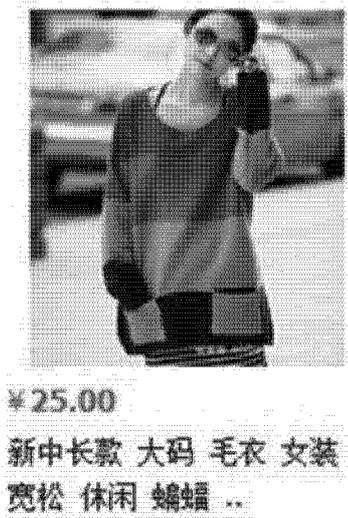


图 3



图 4

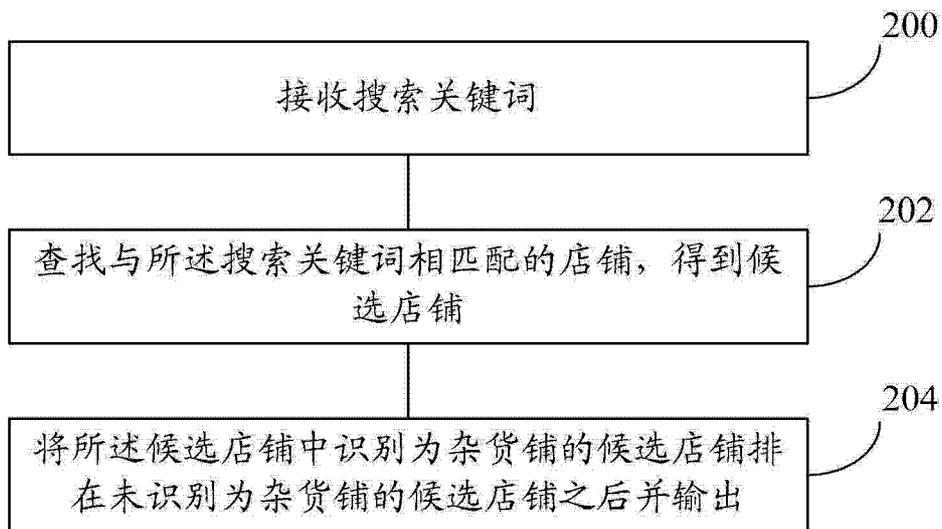


图 5



图 6