(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 106096609 A
(43)申请公布日 2016.11.09

(21)申请号 201610428913.2

(22)申请日 2016.06.16

(71)申请人 武汉大学
地址 430072 湖北省武汉市武昌区珞珈山武汉大学

(72)发明人 黄浩 钟林杌 李宗鹏 颜钱

(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
代理人 魏波

(51)Int.Cl.
G06K 9/32(2006.01)
G06K 9/34(2006.01)
G06K 9/72(2006.01)
G06Q 30/06(2012.01)

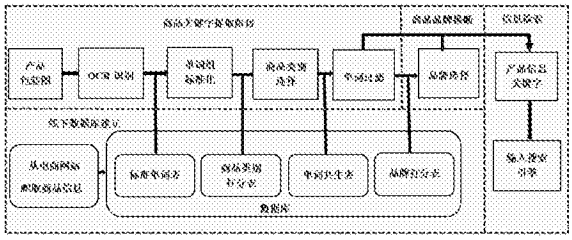
权利要求书4页 说明书8页 附图3页

(54)发明名称

一种基于OCR的商品查询关键字自动生成方法

(57)摘要

本发明公开了一种基于OCR的商品查询关键字自动生成方法，首先建立商品信息数据库。然后利用OCR技术提取产品包装图中的文字信息，获得包含产品信息的单词组。接着通过计算单词组与数据库中单词的相似性，矫正错误字符，完成单词组标准化。接着通过打分规则将得分最高的商品类别作为单词组所代表产品的类别。随后选择该商品类别对应的单词共生表并计算单词组中各单词的共生性得分来过滤掉无用单词。最后，通过该商品类别的品牌打分表和打分规则选择得分最高的品牌作为单词组代表产品的品牌名，将该品牌名结合过滤后的单词组作为商品查询关键字供用户检索使用。本发明计算效率高，对数据库的更新方便，极大地提高用户查询商品信息时的正确性。



CN 106096609 A

1.一种基于OCR的商品查询关键字自动生成方法，其特征在于：首先构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，并所有的表存入数据库中；然后基于商品类别打分表进行商品查询关键字自动生成；其中所述商品查询关键字自动生成包括以下步骤：

步骤1：利用OCR技术提取产品包装图中的全部可识别文字信息，并对返回的字符数据集进行预处理，去掉单个字符长度的单词和非数字、非字母的符号，形成包含产品信息的一个单词组；

步骤2：分别采用Levenshtein Distance和Damerau-Levenshtein Distance两种编辑距离方法，计算步骤1中获得的单词组中每个单词与数据库单词表中所有单词的相似性，并把两个相似性结果的调和平均值作为该单词对数据库单词表中每个单词的相似性值；将单词组中对数据库所有单词的相似性都低于给定阈值τ_s的单词丢弃；对于剩余的单词，使用数据库中与其相似性值最大的单词来替换，并保存各自的最大相似性值S_{max}，完成单词组的标准化工作；

步骤3：若标准化后的产品信息单词组中含有某一产品品牌，则直接将该品牌所在的商品类别作为单词组所代表产品的商品类别；

否则就根据标准化后的产品信息单词组对不同的商品类别进行打分，并且对于每个商品类别，记录单词组中只在该商品类别中出现的单词的个数，将得分最高的商品类别作为单词组所代表产品的类别；若所有商品类别的得分相同，则独占单词数最多的商品类别作为单词组所代表产品的类别；否则无法判断；

步骤4：对确定了商品类别的单词组选择相应的单词共生表，对于单词组中的每一个单词，计算其与单词组中其它单词的共生性得分；若单词组中每个单词的共生性得分均一致，不丢弃任何单词，否则认为得分低于给定的阈值τ_a的单词代表的是无用信息，丢弃该单词，完成单词过滤；

步骤5：若过滤后的商品信息单词组中含有某一产品品牌，将该品牌名结合过滤后的单词组作为商品查询关键字返回，商品查询关键字生成过程结束；否则通过过滤后的商品信息单词组和对应的品牌打分表对所有品牌的打分，选取得分最高的品牌作为该产品的品牌名，将该品牌名结合过滤后的单词组作为商品查询关键字返回。

2.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法，其特征在于：所述构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，是在电商网站上进行商品信息的爬取，在每一个商品类别下形成一个产品信息表；经过对每一个产品信息表的进一步处理生成产品名表、单词表、单词共生表和品牌打分表；综合所有的单词表形成一个商品类别打分表，将所有的表存入数据库中。

3.根据权利要求1或2所述的基于OCR的商品查询关键字自动生成方法，其特征在于：所述构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，其具体实现过程是：

步骤A.1：在电商网站上按照不同商品类别爬取产品的名称、品牌并建立产品信息表，所述产品信息表属性包括产品序号pid、产品品牌brand、产品名name；

步骤A.2：在每个商品类别下，对每个产品的产品名进行修剪，修剪规则为：(1)将大写字母全部转为小写字母；(2)将“/”两边的单词分开；(3)去除无用字符，无用字符包括不是

数字或英文字母表中的字母;(4)去除表示单位的单词;形成修剪后的产品名表;所述产品名表属性包括产品序号pid、修剪后的产品名prunedname;

步骤A.3:基于修剪后的产品名表,对于每个商品类别下出现的单词,统计每个单词的出现次数以及产品名中含有该单词的产品的pid,形成单词表,所述单词表属性为包括单词序号wid、单词word、单词数目num、产品序号pid;

步骤A.4:基于所有单词表,生成一个商品类别打分表,表中的每一项代表一个单词在对应的商品类别下的出现比例,计算公式如下:

$$P[i][j]=\frac{\frac{num_{ij}}{total\_num_i}}{\sum_{i=1}^{N_c}\frac{num_{ij}}{total\_num_i}}\quad (i\in\{1,2,...,N\},\ j\in\{1,2,3,...,N_c\})$$

其中N表示单词表包含的单词总数;N_c表示商品类别数目;P[i][j]表示单词i在商品类别j下的出现比例;num_{ij}表示单词i在商品类别j下出现的次数;$total_num_i$表示单词i在所有商品类别中出现的总次数;

步骤A.5:对于每一个商品类别各生成一个单词共生表ACM,其中的每一项代表对应的两个单词的共生性得分,计算公式如下:

$$ACM[i][j]=\frac{word\_num_{ij}}{word\_num_i}+\frac{pre\_next_{ij}}{word\_num_i}(i,j\in\{1,...,n\})$$

其中n为该商品类别包含的单词总数;ACM[i][j]表示单词i和单词j的共生性得分;$word_num_i$则表示单词i在该商品类别中出现的次数;$word_num_{ij}$表示该商品类别中单词i和单词j在修剪后的产品名中同时出现的次数;pre_next_{ij}表示单词i和单词j在修剪后的产品名中紧挨着出现次数;

步骤A.6:对于每一个商品类别各生成一个品牌打分表WordBrand,其中的每一项代表一个单词对一个品牌的贡献得分,计算公式如下:

$$WordBrand[i][j]=\sum_{k=1}^{N_{ij}}\frac{1}{namelength_k},\quad (i\in\{1,2,...,n\},\ j\in\{1,2,...,N_b\})$$

其中n表示该商品类别包含的单词总数;N_b表示该商品类别包含的品牌数目;WordBrand[i][j]表示单词i对品牌j的贡献得分;N_{ij}表示在某一商品类别中含有单词i且品牌为j的产品的个数,$namelength_k$表示含有单词i且品牌为j的产品k修剪后的产品名长度;

步骤A.7:将所有的表存入数据库中。

4.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法,其特征在于,步骤2中相似性的计算公式为:

$$\begin{cases}Similarity(s,w_i)=1-\dfrac{Ed(s,w_i)}{\max\{Length(s),Length(w_i)\}}\\ w_i\in W(i\in\{1,...,N\})\end{cases}$$

其中s为OCR返回的单词组中的一个单词;W为数据库单词表中所有的单词;N为数据库单词表包含的单词总数;Ed为编辑距离的计算方法;Similarity(s,w_i)表示单词组中的单词s与数据库中单词w_i的相似性;Length(s)表示单词s的长度;Length(w_i)表示单词w_i的长

度。

5.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法，其特征在于，步骤2中所述$\tau_s \in [0,1]$。

6.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法，其特征在于，步骤3中所述根据标准化后的产品信息单词组对不同的商品类别进行打分，商品类别打分规则为：若某单词只在一个商品类别中出现，则根据表1进行打分；

表1 单词只在一个商品类别中出现时该商品类别得分规则

该词在步骤3中统计的最大相似性 S_{max} 大于阈值 τ_{sc}		该词在步骤3中统计的最大相似性 S_{max} 小于阈值 τ_{sc}	
单词长度 L 不超过阈值 τ_L	单词长度 L 超过 τ_L	单词长度 L 不超过 τ_L	单词长度 L 超过 τ_L
该库得分+$grade$	该库得分+$grade$	该库得分+$grade$	该库得分+$grade$

若单词在多个商品类别中出现，每个商品类别的加分值为该单词在商品类别打分表中对应项的值乘以给定的数值C_m；其中对应项是该单词在商品类别的出现比例。

7.根据权利要求6所述的基于OCR的商品查询关键字自动生成方法，其特征在于：$\tau_{sc} \in [0,1]$，$\tau_L \in [1,15]$，$grade \in [1,100]$，$C_m \in [1,20]$。

8.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法，其特征在于，步骤4中所述共生性得分计算公式为：

$$app\_proportion_i = \frac{acm\_num_i}{stdWordNum - 1 - outlierNum} (i \in \{1,...,stdWordNum\})$$

；

其中stdWordNum表示单词组中的单词个数；$app_proportion_i$表示标准化后单词组中单词i的共生性得分；acm_num_i代表单词组中与单词i在单词共生表中对应的值大于给定阈值τ_c的单词个数；outlierNum表示单词组中与其它单词均不共生的单词个数。

9.根据权利要求8所述的基于OCR的商品查询关键字自动生成方法，其特征在于：所述$\tau_a \in [0,1]$，$\tau_c \in [0,1]$。

10.根据权利要求1所述的基于OCR的商品查询关键字自动生成方法，其特征在于，步骤5中所述对所有品牌的打分过程如下：

步骤5.1：根据步骤3确定的商品类别选择相应的品牌打分表WordBrand，根据该品牌打分表和过滤后的单词组对相应商品类型下的所有品牌进行打分；计算公式为：

$$score[k] = \sum_{i=1}^{N_f} WordBrand[indexOf(word_i)][k] (k \in \{1,...,N_b\})$$

其中N_b为该商品类别包含的品牌数目；score[k]为品牌k的得分；N_f为过滤后的单词组含有的单词总数；indexOf($word_i$)表示单词$word_i$在该商品类别单词表中的wid；

步骤5.2：给定不同的数值k，将单词组中任意k个单词组合，若该单词组合只在一个品牌中的出现，该品牌增加分值grade1；若在多个品牌中出现，则对应的多个品牌增加分值grade2。

11.根据权利要求10所述的基于OCR的商品查询关键字自动生成方法，其特征在于：所述k∈[1,10]，grade1∈[1,30]，grade2∈[1,30]。

一种基于OCR的商品查询关键字自动生成方法

技术领域

[0001] 本发明属于信息检索技术领域，尤其涉及一种在OCR基础上的商品关键字自动生成方法。

背景技术

[0002] 互联网以及手持智能终端在过去的10年间经历了爆炸式的发展，这极大地丰富了人们的信息获取途径并改变了人们的生活方式，越来越多的人选择通过电商完成购物。借助各种电商网站上详细的产品信息以及其它购买者对商品的评价，人们可以更好地进行购物选择。但是当购物者在商场、书店等地购物时，查询商品的具体信息就变得较为困难。通常人们的做法是阅读产品包装并人为提取组织其中可能的关键字，之后再输入到搜索引擎中进行查询。但手工提取产品关键字的过程费时费力，而且对于购物者来说精确选择关键字较为困难，更为糟糕的是一些无用单词可能会干扰查询结果。

[0003] OCR(Optical Character Recognition，光学字符识别)能对图像中的文本信息进行分析识别处理，通过检测暗、亮的模式确定其形状，用字符识别方法将形状翻译成计算机文字。随着带有拍照功能的手持智能终端的广泛普及，利用OCR技术对拍摄的商品包装照片中的文字信息进行提取显得水到渠成。但是，OCR识别出来的信息存在大量噪音，且存在一些无用信息。如果不对这些信息进行进一步的分析，其结果很可能影响用户的使用。因此需要对OCR识别的信息进一步分析整合。

发明内容

[0004] 为了解决上述技术问题，本发明提供了一种基于OCR的商品查询关键字自动生成方法，在获取一张用手持智能终端拍摄的产品包装图后，OCR将会对该产品图片进行文字提取并返回一个包含大量噪音和无用信息的字符数据集，之后通过矫正错误字符(标准化)、选择商品类别、过滤无用信息、确定产品品牌四个过程最终生成合理的产品关键字。

[0005] 本发明所采用的技术方案是：一种基于OCR的商品查询关键字自动生成方法，其特征在于：首先构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，并所有的表存入数据库中；然后基于商品类别打分表进行商品查询关键字自动生成；其中所述商品查询关键字自动生成包括以下步骤：

[0006] 步骤1：利用OCR技术提取产品包装图中的全部可识别文字信息，并对返回的字符数据集进行预处理，去掉单个字符长度的单词和无用符号(非数字、非字母的符号)，形成包含产品信息的一个单词组；

[0007] 步骤2：分别采用Levenshtein Distance和Damerau-Levenshtein Distance两种编辑距离方法，计算步骤1中获得的单词组中每个单词与数据库单词表中所有单词的相似性，并把两个相似性结果的调和平均值作为该单词对数据库单词表中每个单词的相似性值；将单词组中对数据库所有单词的相似性都低于给定阈值τ_s的单词丢弃；对于剩余的单词，使用数据库中与其相似性值最大的单词来替换，并保存各自的最大相似性值S_{max}，完成

单词组的标准化工作；

[0008] 步骤3：若标准化后的产品信息单词组中含有某一产品品牌，则直接将该品牌所在的商品类别作为单词组所代表产品的商品类别；

[0009] 否则就根据标准化后的产品信息单词组对不同的商品类别进行打分，并且对于每个商品类别，记录单词组中只在该商品类别中出现的单词的个数，将得分最高的商品类别作为单词组所代表产品的类别；若所有商品类别的得分相同，则独占单词数最多的商品类别作为单词组所代表产品的类别；否则无法判断；

[0010] 步骤4：对确定了商品类别的单词组选择相应的单词共生表，对于单词组中的每一个单词，计算其与单词组中其它单词的共生性得分；若单词组中每个单词的共生性得分均一致，不丢弃任何单词，否则认为得分低于给定的阈值τ_a的单词代表的是无用信息，丢弃该单词，完成单词过滤；

[0011] 步骤5：若过滤后的商品信息单词组中含有某一产品品牌，将该品牌名结合过滤后的单词组作为商品查询关键字返回，商品查询关键字生成过程结束；否则通过过滤后的商品信息单词组和对应的品牌打分表对所有品牌的打分，选取得分最高的品牌作为该产品的品牌名，将该品牌名结合过滤后的单词组作为商品查询关键字返回。

[0012] 作为优选，所述构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，是在电商网站上进行商品信息的爬取，在每一个商品类别下形成一个产品信息表；经过对每一个产品信息表的进一步处理生成产品名表、单词表、单词共生表和品牌打分表；综合所有的单词表形成一个商品类别打分表，将所有的表存入数据库中。

[0013] 作为优选，所述构建所有商品的产品名表、单词表、单词共生表和品牌打分表，综合所有的单词表形成商品类别打分表，其具体实现过程是：

[0014] 步骤A.1：在电商网站上按照不同商品类别爬取产品的名称、品牌并建立产品信息表，所述产品信息表属性包括产品序号(pid)、产品品牌(brand)、产品名(name)；

[0015] 步骤A.2：在每个商品类别下，对每个产品的产品名进行修剪，修剪规则为：(1)将大写字母全部转为小写字母；(2)将“/”两边的单词分开，如cleanse/tone转为cleansetone；(3)去除无用字符(不是数字或英文字母表中的字母)(4)去除表示单位的单词；形成修剪后的产品名表；所述产品名表属性包括产品序号(pid)、修剪后的产品名(prunedname)；

[0016] 步骤A.3：基于修剪后的产品名表，对于每个商品类别下出现的单词，统计每个单词的出现次数以及产品名中含有该单词的产品的pid，形成单词表，所述产品名表属性包括产品序号(pid)、修剪后的产品名(prunedname)；

[0017] 步骤A.4：基于所有单词表，生成一个商品类别打分表，表中的每一项代表一个单词在对应的商品类别下的出现比例，计算公式如下：

[0018] $$P[i][j]=\frac{\frac{num_{ij}}{total\_num_i}}{\sum_{i=1}^{N_c}\frac{num_{ij}}{total\_num_i}}\quad (i\in\{1,2,...,N\},\ j\in\{1,2,3,...,N_c\})$$

[0019] 其中N表示单词表包含的单词总数；N_c表示商品类别数目；P[i][j]表示单词i在商

品类别j下的出现比例;num_{ij}表示单词i在商品类别j下出现的次数;$total_num_i$表示单词i在所有商品类别中出现的总次数;

[0020] 步骤A.5:对于每一个商品类别各生成一个单词共生表ACM,其中的每一项代表对应的两个单词的共生性得分,计算公式如下:

[0021] $$ACM[i][j]=\frac{word\_num_{ij}}{word\_num_i}+\frac{pre\_next_{ij}}{word\_num_i}(i,j\in\{1,...,n\})$$

[0022] 其中n为该商品类别包含的单词总数;ACM[i][j]表示单词i和单词j的共生性得分;$word_num_i$则表示单词i在该商品类别中出现的次数;$word_num_{ij}$表示该商品类别中单词i和单词j在修剪后的产品名中同时出现的次数;pre_next_{ij}表示单词i和单词j在修剪后的产品名中紧挨着出现次数;

[0023] 步骤A.6:对于每一个商品类别各生成一个品牌打分表WordBrand,其中的每一项代表一个单词对一个品牌的贡献得分,计算公式如下:

[0024] $$WordBrand[i][j]=\sum_{k=1}^{N_{ij}}\frac{1}{namelength_k},\quad (i\in\{1,2,...,n\}\ ,j\in\{1,2,...,N_b\})$$

[0025] 其中n表示该商品类别包含的单词总数;N_b表示该商品类别包含的品牌数目;WordBrand[i][j]表示单词i对品牌j的贡献得分;N_{ij}表示在某一商品类别中含有单词i且品牌为j的产品的个数,$namelength_k$表示含有单词i且品牌为j的产品k修剪后的产品名长度;

[0026] 步骤A.7:将所有的表存入数据库中。

[0027] 作为优选,步骤2中相似性的计算公式为:

[0028] $$\begin{cases}Similarity(s,w_i)=1-\dfrac{Ed(s,w_i)}{\max\{Length(s),Length(w_i)\}}\\ w_i\in W(i\in\{1,...,N\})\end{cases}$$

[0029] 其中s为OCR返回的单词组中的一个单词;W为数据库单词表中所有的单词;N为数据库单词表包含的单词总数;Ed为编辑距离的计算方法;$Similarity(s,w_i)$表示单词组中的单词s与数据库中单词w_i的相似性;Length(s)表示单词s的长度;$Length(w_i)$表示单词w_i的长度;

[0030] 作为优选,步骤2中所述$\tau_s\in[0,1]$。

[0031] 作为优选,步骤3中所述根据标准化后的产品信息单词组对不同的商品类别进行打分,商品类别打分规则为:若某单词只在一个商品类别中出现,则根据表1进行打分;

[0032] 表1 单词只在一个商品类别中出现时该商品类别得分规则

[0033]

该词在步骤3中统计的最大相似性 S_{max} 大于阈值 τ_{sc}	该词在步骤3中统计的最大相似性 S_{max} 小于阈值 τ_{sc}

[0034]

单词长度 L 不超过阈值 τ_L	单词长度 L 超过 τ_L	单词长度 L 不超过 τ_L	单词长度 L 超过 τ_L
该库得分+$grade$	该库得分+ $grade$	该库得分+ $grade$	该库得分+ $grade$

[0035] 若单词在多个商品类别中出现，每个商品类别的加分值为该单词在商品类别打分表中对应项的值乘以给定的数值C_m；其中对应项是该单词在商品类别的出现比例。

[0036] 作为优选，$\tau_{sc}\in[0,1]$，$\tau_L\in[1,15]$，$grade\in[1,100]$，$C_m\in[1,20]$。

[0037] 作为优选，步骤4中所述共生性得分计算公式为：

[0038] $$app\_proportion_i=\frac{acm\_num_i}{stdWordNum-1-outlierNum}(i\in\{1,...,stdWordNum\})$$；

[0039] 其中stdWordNum表示单词组中的单词个数；$app_proportion_i$表示标准化后单词组中单词i的共生性得分；acm_num_i代表单词组中与单词i在单词共生表中对应的值大于给定阈值τ_c的单词个数；outlierNum表示单词组中与其它单词均不共生的单词个数。

[0040] 作为优选，所述$\tau_a\in[0,1]$，$\tau_c\in[0,1]$。

[0041] 作为优选，步骤5中所述对所有品牌的打分过程如下：

[0042] 步骤5.1：根据步骤3确定的商品类别选择相应的品牌打分表WordBrand，根据该品牌打分表和过滤后的单词组对相应商品类型下的所有品牌进行打分；计算公式为：

[0043] $$score[k]=\sum_{i=1}^{N_f}WordBrand[indexOf(word_i)][k](k\in\{1,...,N_b\})$$

[0044] 其中N_b为该商品类别包含的品牌数目；score[k]为品牌k的得分；N_f为过滤后的单词组含有的单词总数；indexOf($word_i$)表示单词$word_i$在该商品类别单词表中的wid；

[0045] 步骤5.2：给定不同的数值k，将单词组中任意k个单词组合，若该单词组合只在一个品牌中的出现，该品牌增加分值grade1；若在多个品牌中出现，则对应的多个品牌增加分值grade2。

[0046] 作为优选，所述$k\in[1,10]$，$grade1\in[1,30]$，$grade2\in[1,30]$。

[0047] 本发明中提出的基于OCR的商品查询关键字自动生成技术，计算量很小，对于硬件要求很低，具有很高的效率；使用的数据库以及表格很小，更新方便；能够极大地提高用户查询商品信息时的正确性，改善用户的购物体验。

附图说明

[0048] 图1：本发明实施例的流程图。

[0049] 图2：本发明实施例的数据库示意图。

[0050] 图3：本发明实施例中样例产品的包装图。

[0051] 图4：本发明实施例中OCR返回的识别结果图。

[0052] 图5：本发明实施例中经过预处理的OCR识别结果图。

[0053] 图6：本发明实施例中商品类别得分与选择的商品类别结果图。

[0054] 图7：本发明实施例中单词组过滤后的结果图。

[0055] 图8:本发明实施例中选取的品牌以及最终生成的商品查询关键词结果图。

[0056] 图9:本发明实施例中利用生成的商品查询关键词在搜索引擎中查询的结果图。

具体实施方式

[0057] 为了便于本领域普通技术人员理解和实施本发明,下面结合附图及实施例对本发明作进一步的详细描述,应当理解,此处所描述的实施示例仅用于说明和解释本发明,并不用于限定本发明。

[0058] 请见图1,本发明提供一种基于OCR的商品查询关键字自动生成方法,其特征在于:首先构建所有商品的产品名表、单词表、单词共生表和品牌打分表,综合所有的单词表形成商品类别打分表,并所有的表存入数据库中;然后基于商品类别打分表进行商品查询关键字自动生成;

[0059] 构建所有商品的产品名表、单词表、单词共生表和品牌打分表,综合所有的单词表形成商品类别打分表,是在电商网站上进行商品信息的爬取,在每一个商品类别下形成一个产品信息表;经过对每一个产品信息表的进一步处理生成产品名表、单词表、单词共生表和品牌打分表;综合所有的单词表形成一个商品类别打分表,将所有的表存入数据库中;其具体实现过程是:

[0060] 步骤A.1,在亚马逊网站上按照不同商品类别(日用品,红酒,书籍)爬取产品的名称、品牌(其中书籍的品牌为作者名)并建立产品信息表(属性为:产品序号(pid)、产品品牌(brand)、产品名(name)),分别为commodity、wine、book。

[0061] 步骤A.2,在每个商品类别下,对每个产品的产品名进行修剪,修剪规则为:(1)将大写字母全部转为小写字母;(2)将“/”两边的单词分开,如cleanse/tone转为cleansetone;(3)去除无用字符(不是数字或英文字母表中的字母)(4)去除表示单位的单词;形成修剪后的产品名表;所述产品名表属性包括产品序号(pid)、修剪后的产品名(prunedname);形成3个修剪后的产品名表(属性为:产品序号(pid)、修剪后的产品名(prunedname)),分别为commodity_pruned、wine_pruned、book_pruned。

[0062] 步骤A.3,基于修剪后的产品名表,对于每个商品类别下出现的单词(即修剪后的产品名中含有的所有单词),统计每个单词的出现次数以及产品名中含有该单词的产品的pid,形成3个单词表(属性为单词序号(wid)、单词(word)、单词数目(num)、产品序号(pid)),分别为commodity_words、wine_words、book_words。数据库中的产品信息表,产品名表,单词表见图2。

[0063] 步骤A.4,基于数据库中的所有单词表,生成一个商品类别打分表,请见表2,表中的每一项代表一个单词在对应的商品类别下的出现比例,计算公式如下:

[0064] $$P[i][j]=\frac{\frac{num_{ij}}{total\_num_i}}{\sum_{i=1}^{N_c}\frac{num_{ij}}{total\_num_i}} \quad (i\in\{1,2,...,N\},\ j\in\{1,2,3\})$$

[0065] 其中N表示数据库单词表包含的单词总数;P[i][j]表示单词i在商品类别j下的出现比例;num_{ij}表示单词i在商品类别j下出现的次数;$total_num_i$表示单词i在三个商品类别中出现的总次数。

[0066] 表2 商品类别打分表的结构

[0067]

单词	commodity类别	wine类别	book类别
olay	1	0	0
with	0.8282208588957055	0.03680981595092025	0.13496932515337423
…	…	…	…

[0068] 步骤A.5,对于每一个商品类别各生成一个单词共生表ACM,其中的每一项代表对应的两个单词的共生性得分,计算公式如下:

[0069] $$ACM[i][j]=\frac{word\_num_{ij}}{word\_num_i}+\frac{pre\_next_{ij}}{word\_num_i}(i,j\in\{1,...,n\})$$

[0070] 其中n表示该商品类别包含的单词总数;ACM[i][j]表示单词i和单词j的共生性得分;$word_num_i$则表示单词i在该商品类别中出现的次数;$word_num_{ij}$表示该商品类别中单词i和单词j在修剪后的产品名中同时出现的次数;pre_next_{ij}表示单词i和单词j在修剪后的产品名中紧挨着出现的次数。

[0071] 步骤A.6,对于每一个商品类别各生成一个品牌打分表WordBrand,其中的每一项代表一个单词对一个品牌的贡献得分,计算公式如下:

[0072] $$WordBrand[i][j]=\sum_{k=1}^{N_{ij}}\frac{1}{namelength_k},\ (i\in\{1,2,...,n\}\ ,j\in\{1,2,...,N_b\})$$

[0073] 其中n表示该商品类别包含的单词总数;N_b表示该商品类别包含的品牌数目;WordBrand[i][j]表示单词i对品牌j的贡献得分;N_{ij}表示在某一商品类别中含有单词i且品牌为j的产品的个数,$namelength_k$表示含有单词i且品牌为j的产品k修剪后的产品名长度。

[0074] 然后进行商品查询关键字自动生成,具体包括以下步骤:

[0075] 步骤1:利用OCR技术提取产品包装图(图3)的全部可识别文字信息,识别结果如图4,并对返回的识别结果进行简单预处理,去掉单个字符长度的单词和一些无用符号(例如:“.”,“|”等),形成包含产品信息的一个单词组,预处理后结果如图5。

[0076] 步骤2:采用两种编辑距离方法Levenshtein Distance和Damerau-Levenshtein Distance,分别计算OCR返回的单词组中每个单词与数据库单词表中所有单词的相似性,并把两个相似性结果的调和平均值作为该单词对数据库单词表中每个单词的相似性值。将单词组中对数据库所有单词的相似性都低于阈值0.5的单词丢弃。对于剩余的单词,使用数据库中与其相似性值最大的单词来替换,并保存各自的最大相似性值S_{max},完成单词组的标准化工作。相似性的计算公式如下:

[0077] $$\begin{cases}Similarity(s,w_i)=1-\dfrac{Ed(s,w_i)}{\max\{Length(s),Length(w_i)\}}\\ w_i\in W(i\in\{1,...,N\})\end{cases}$$

[0078] 其中s为OCR返回的单词组中的一个单词;W为数据库单词表中的所有单词;N为数据库单词表所包含的单词总数;Ed为编辑距离的计算方法;Similarity(s,w_i)表示单词组中的单词s与数据库中单词w_i的相似性。

[0079] 步骤3:若标准化后的产品信息单词组中含有某一产品品牌(比如Olay、Nivea等),

则直接将该品牌所在的商品类别作为单词组所代表产品的商品类别，步骤3结束。否则就根据标准化后的单词组对不同的商品类别进行打分，并且对于每个商品类别，记录单词组中只在该商品类别中出现的单词的个数。将得分最高的商品类别作为单词组所代表产品的类别；若所有商品类别的得分相同，则独占单词数最多的商品类别作为单词组所代表产品的类别。对商品类别打分规则为：若某单词只在一个商品类别出现，根据表3打分；若单词在多个商品类别中出现，每个商品类别的加分值为该单词在商品类别打分表中对应项的值乘以常数5。三个商品类别得分及选择结果如图6所示。

[0080] 表3 单词只在一个商品类别中出现时该商品类别得分规则

[0081]

该词在步骤 3 中统计的最大相似性 S_{max} 大于 0.75		该词在步骤 3 中统计的最大相似性 S_{max} 小于 0.75	
单词长度 L 不超过 4	单词长度 L 超过 4	单词长度 L 不超过 4	单词长度 L 超过 4
该库得分+10	该库得分+55	该库得分+5	该库得分+20

[0082] 步骤4：对于确定了产品类别的单词组选择相应的单词共生表，对于单词组中的每一个单词，计算其与单词组中其它单词的共生性得分。若单词组中每个单词的共生性得分均一致，不丢弃任何单词。否则认为得分低于0.2的单词代表的是无用信息，丢弃该单词，完成单词过滤，单词组过滤后的结果如图7所示。共生性得分计算公式如下：

[0083] $$app\_proportion_i = \frac{acm\_num_i}{stdWordNum - 1 - outlierNum} (i \in \{1, ..., stdWordNum\})$$

[0084] 其中$app_proportion_i$为标准化后单词组中第i个单词的共生性得分；acm_num_i代表单词组中与第i个单词在单词共生表中对应的值大于0.05的单词个数；stdWordNum为单词组中的单词个数；outlierNum为单词组中与其它单词均不共生的单词个数。

[0085] 步骤5：若过滤后的产品信息单词组中含有某一产品品牌(比如Olay、Nivea等)，将该品牌名结合过滤后的单词组作为商品查询关键字返回，商品查询关键字生成过程结束。否则通过过滤后的产品信息单词组和对应的品牌打分表对所有品牌的打分，选取得分最高的品牌作为该产品的品牌名，将该品牌名结合过滤后的单词组作为商品查询关键字返回。选取的品牌以及最终生成的商品查询关键词结果如图8所示。利用生成的商品查询关键词在搜索引擎中查询的结果如图9所示(红框标出的为目标商品)。

[0086] 在步骤5中，对所有品牌的打分过程如下：

[0087] 步骤5.1，否则根据步骤3确定的商品类别选择相应的品牌打分表WordBrand，根据该品牌打分表和过滤后的单词组对相应商品类型下的所有品牌进行打分。计算公式为：

[0088] $$score[k] = \sum_{i=1}^{N_f} WordBrand[indexOf(word_i)][k] (k \in \{1, ..., N_b\})$$

[0089] 其中score[k]代表品牌k的得分；N_f为过滤后的单词组中含有的单词总数；N_b为相应商品类型下品牌的个数；$indexOf(word_i)$为单词$word_i$在该商品类别单词表中的wid。

[0090] 步骤5.2，令数值k分别等于1、2、3，根据表4对所有品牌打分。

[0091] 表4 不同k取值时的品牌得分规则

[0092]

若单词组中有只在一个品牌中出现的单词	若单词组中单词之间两两组合形成的组合出现在品牌名称中	若单词组中单词之间三三组合形成的组合出现在品牌名称中
出现一次则该品牌得分+5	出现一次则该品牌得分+1； 若组合只在一个品牌中出现，出现一次则该品牌得分+7	出现一次则该品牌得分+1； 若组合只在一个品牌中出现，出现一次则该品牌得分+9

[0093] 应当理解的是，本说明书未详细阐述的部分均属于现有技术。

[0094] 应当理解的是，上述针对较佳实施例的描述较为详细，并不能因此而认为是对本发明专利保护范围的限制，本领域的普通技术人员在本发明的启示下，在不脱离本发明权利要求所保护的范围情况下，还可以做出替换或变形，均落入本发明的保护范围之内，本发明的请求保护范围应以所附权利要求为准。

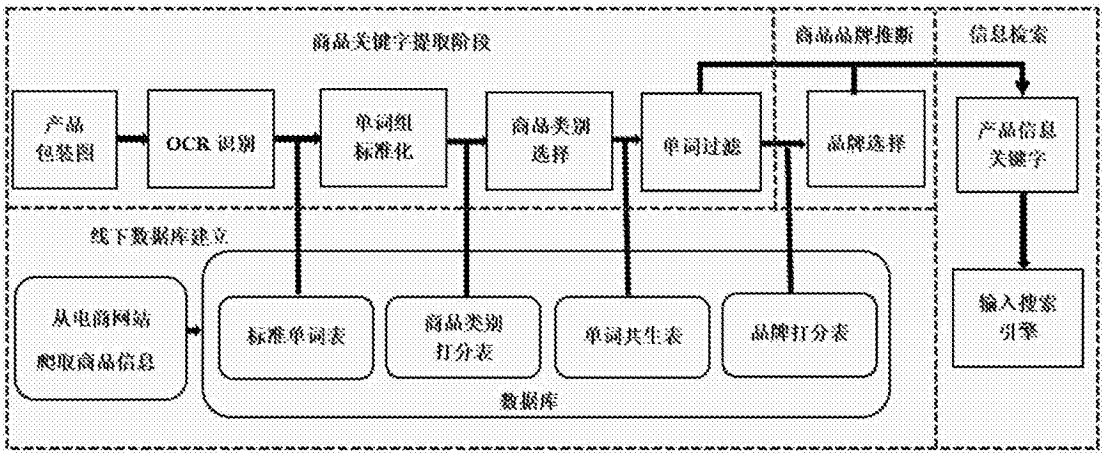


图1

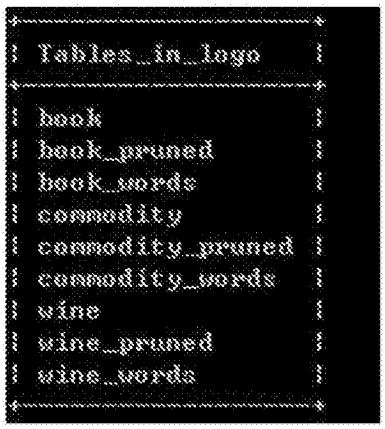


图2

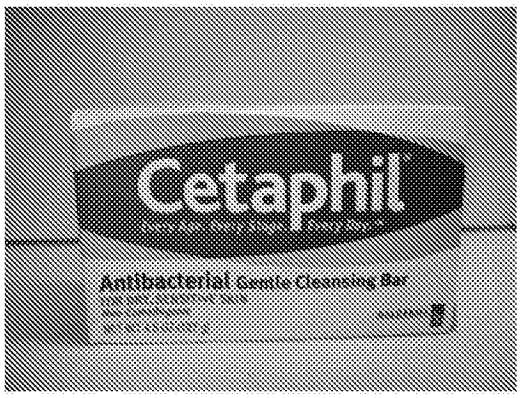


图3

```
-----OCR is done, and the results are as follows-----
i
c t p '
every age. every stage. every day. "'
antibacterial gentle cleansing bar
l for dry,dsensiitive skin
?q§??non-come ogenc
i;-  01021 g) qqelilah i
```

图4

```
-----preprocessed OCR results are as follows-----
every age every stage every day antibacterial gentle cleansing bar for dry dsensiitive skin non come ogenc 01021 qqelilah
```

图5

```
commodity grade: 295.2836627533091wine grade: 2.3785454438390756book grade: 192.33779080285183
choose commodity!
```

图6

```
-----Recommended keywords-----
        age day antibacterial gentle cleansing bar for dry sensitive skin adella
```

图7

```
-----brand-----
    cetaphil
-----final keywords before go Google -----
        age day antibacterial gentle cleansing bar for dry sensitive skin adella  cetaphil
```

图8

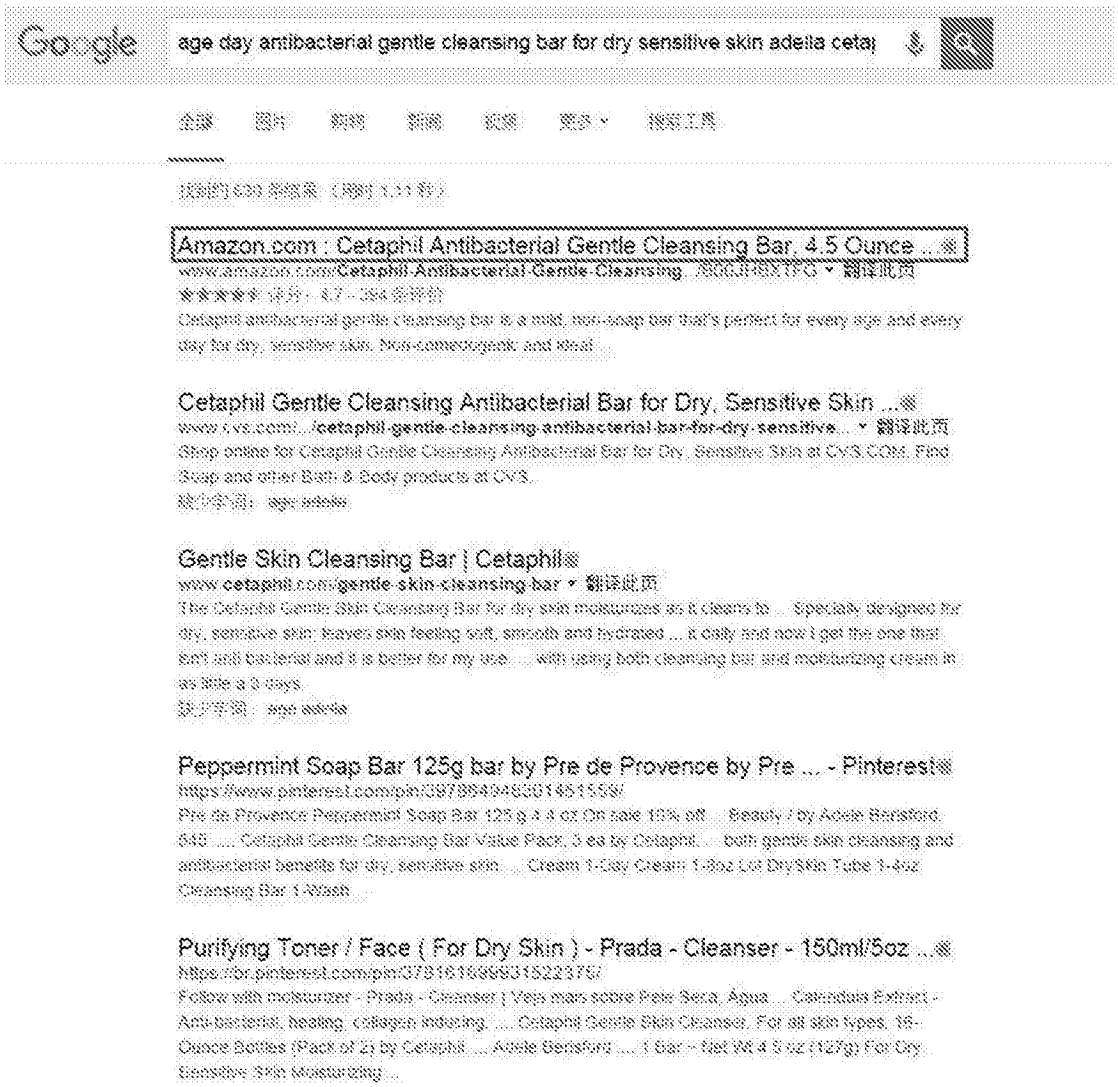

Google
age day antibacterial gentle cleansing bar for dry sensitive skin adella cetaj
全部 图片 购物 新闻 视频 更多 搜索工具
Amazon.com : Cetaphil Antibacterial Gentle Cleansing Bar, 4.5 Ounce ...
www.amazon.com/Cetaphil-Antibacterial-Gentle-Cleansing
Cetaphil antibacterial gentle cleansing bar is a mild, non-soap bar that's perfect for every age and every day for dry, sensitive skin. Non-comedogenic and ideal ...
Cetaphil Gentle Cleansing Antibacterial Bar for Dry, Sensitive Skin ...
www.cvs.com/.../cetaphil-gentle-cleansing-antibacterial-bar-for-dry-sensitive...
Shop online for Cetaphil Gentle Cleansing Antibacterial Bar for Dry, Sensitive Skin at CVS.COM. Find Soap and other Bath & Body products at CVS.
Gentle Skin Cleansing Bar | Cetaphil
www.cetaphil.com/gentle-skin-cleansing-bar
The Cetaphil Gentle Skin Cleansing Bar for dry skin moisturizes as it cleans to ... Specially designed for dry, sensitive skin; leaves skin feeling soft, smooth and hydrated ... it daily and now I get the one that isn't anti bacterial and it is better for my face ... with using both cleansing bar and moisturizing cream in as little a 3 days.
Peppermint Soap Bar 125g bar by Pre de Provence by Pre ... - Pinterest
https://www.pinterest.com/pin/397864946301451559/
Pre de Provence Peppermint Soap Bar 125 g 4.4 oz On sale 10% off ... Beauty / by Adele Beresford. 645 ... Cetaphil Gentle Cleansing Bar Value Pack, 3 ea by Cetaphil ... both gentle skin cleansing and antibacterial benefits for dry, sensitive skin. ... Cream 1-Day Cream 1-8oz Lot Dry/Skin Tube 1-4oz Cleansing Bar 1-Wash ...
Purifying Toner / Face (For Dry Skin) - Prada - Cleanser - 150ml/5oz ...
https://br.pinterest.com/pin/378161696903152237S/
Follow with moisturizer - Prada - Cleanser | Veja mais sobre Pele Seca, Água ... Calendula Extract - Anti-bacterial, healing. Collagen inducing. ... Cetaphil Gentle Skin Cleanser. For all skin types. 16-Ounce Bottles (Pack of 2) by Cetaphil. ... Adele Beresford. ... 1 Bar ~ Net Wt 4.5 oz (127g) For Dry Sensitive Skin Moisturizing ...

图9