专利技术交底书

发明名称	一种基于改进的 Single-pass 聚类算法的商品标签聚类方法			
技术联系人	方泉荣	邮　箱	fangquanrong@zallscg.com	
手　　机	13261607077		专利类型	☒发　　明 ☐实用新型
发明人	方泉荣			

一、 背景技术描述

（1） 本发明所属技术领域

本发明设计涉及一种基于改进的 single-pass 聚类算法的聚类方法。涉及图文识别、改进的 single-pass 聚类算法领域。

（2） 该行业的技术发展现状

201910606225.4

201711210195.2

（3） 要解决的技术问题

现有电商平台相关商品推荐的计算方案中，一般通过获取用户曾搜索过的关键词进行二次检索后，将具有相同关键词的商品进行推送。一而现有的解决方案中，存在以下问题：

1、推荐产品同质化严重。

2、商家为了促销，将无关商品附加上热销标签，误导消费者。

二、本发明的技术方案（重点部分）

（1）本发明采用的技术手段

1、基于 single-pass 聚类算法，添加图文相关性参数，实现商品标签聚类

（2）本发明采用的技术方案

本发明方案是一种基于改进的Single-pass聚类算法的商品标签聚类方法。其特征在，获取商品标题，进行文本预处理，得到有效的分词结果；商品图片识别后与分词结果进行对比并获得相关性参数，将相关性参数作为阈值引入single-pass聚类算法中，实现对商品标签的有效精准聚类。通过以下步骤实现所述聚类方法。

本发明的具体步骤如图1所示，包括：

步骤S1，通过爬虫技术遍历商品页，获取商品的标题作为文本集。

步骤S2，对文本集进行预处理，获得预处理后的文本集D。主要操作有以下步骤：

S21去除噪声干扰。将文本附带的符号、汉字拼音混合内容等非规范元素去除。

S22规定去除多余20个字的文本内容，因为其可能覆盖大量无意义分词。

S23通过NLPIR汉语分词系统等工具，实现对样本的中文分词

S24对分词后的结果去除停用词，停用词指不能表达商品含义的词语，如“你”、“特别”、“美观”等名词和形容词。

步骤S3，通过图文识别技术，识别商品图片特征，获得特征集$D_{cha}\{d_1,d_2,d_3...d_i\}$，其中$D_{cha}$是特征集合，d是特征，i是特征个数。如识别某品牌巧克力，特征集则包括d_1=零食，d_2=巧克力，d_3=某品牌。通过以下公式获得图文相关性参数P。

$$P=\frac{card(D)}{card(D_{cha})}$$

步骤S4，基于改进的single-pass聚类算法对集合D进行文本聚类。包括：

S41向集合D附加图文相关性参数P；

S42计算D中每项数据特征类别的聚类中心点，将所有小特征类别聚类成大特征类别；

S43对聚类中心点进行相似度计算，若相似度低于设定阈值，则列入待定类C_{und}，若相似度高于设定阈值，则进行相关性参数判断，若相关性参数在设定区间内，即证明图文无表述差异过大的现象，将该数据与上述聚类中心归为一类。相似度算法采用余弦相似度算法；若相关性参数不在设定区间内，则归入待定类；

S44 对待定类 C_{und} 重复步骤 S43，直到 $T(C_{und})=T+1(C_{und})$ 终止，此时得到最终聚类中心和聚类结果；

S45 对所属聚类结果进行人工筛查异常项。

（3）发明点

Single-pass 聚类算法是一种具有强的大规模数据处理能力和处理噪声能力，具有较强的发现任意形状簇的能力的文本聚类算法。其算法简单、只需要额外处理小簇的特点导致了其在实际场景中的应用强于其他文本聚类算法。

在电商平台的商品标签和相关商品推荐场景特征中，具有样本数大，商家制造的噪声因素多等特点。并且现有的相关商品推荐大多是同质化的商品，用户在购买商品后，对同质化的商品需求下降，常见的商品推荐方法转化率容易受到物品种类的影响，Single-pass 聚类算法能较好的解决相关问题。此外，本发明通过引入图文识别技术的结果，对 single-pass 算法进行了改进，实现了有效的降噪，并且一定程度弥补了 single-pass 聚类算法易产生多个小簇的缺点。

（4）本发明的技术效果

通过电商平台商品推荐实际场景的特征引入契合的模型，可以在小范围内弥补 single-pass 聚类算法的缺点。再通过改进的 single-pass 聚类算法解决现有商品推荐中商品同质化严重的情况，提高推荐商品转化率。

三、附图

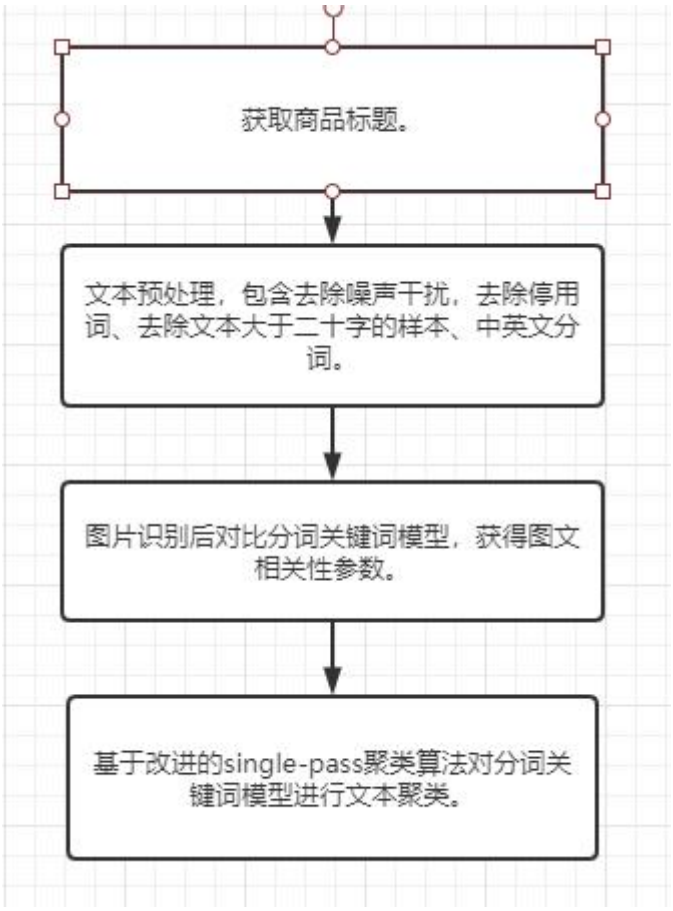

图 1

四、其它可替代方案（如有，请参照本发明的技术方案部分进行描述；如没有，则不写）