



(12)发明专利

(10)授权公告号 CN 106021362 B

(45)授权公告日 2018.04.13

(21)申请号 201610306218.9

(22)申请日 2016.05.10

(65)同一申请的已公布的文献号
申请公布号 CN 106021362 A

(43)申请公布日 2016.10.12

(73)专利权人 百度在线网络技术(北京)有限公司
地址 100085 北京市海淀区上地十街10号
百度大厦三层

(72)发明人 付立波 方高林 陈雨

(74)专利代理机构 北京品源专利代理有限公司
11332
代理人 孟金喆 胡彬

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

(56)对比文件

CN 104820718 A,2015.08.15,说明书第6-14段.

CN 104778281 A,2015.07.15,说明书第6-22段.

CN 103744970 A,2014.04.23,说明书第6-9,46段.

CN 101211355 A,2008.07.02,说明书第2页第4段,第4页第1,2段.

CN 102930296 A,2013.02.13,说明书第89-102段.

CN 103294817 A,2013.09.11,说明书第6-16段.

CN 103984746 A,2014.08.13,全文.

审查员 李燕东

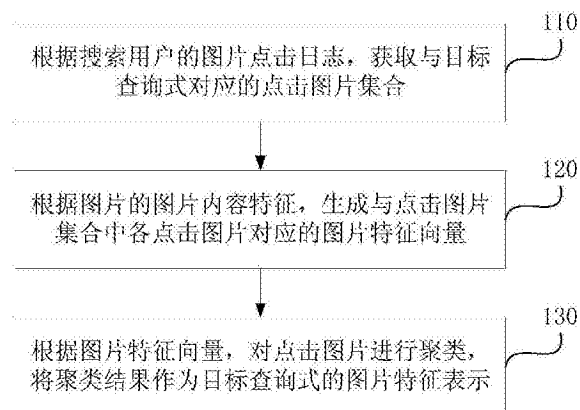
权利要求书3页 说明书12页 附图6页

(54)发明名称

查询式的图片特征表示的生成、图片搜索方法和装置

(57)摘要

本发明实施例公开了一种查询式的图片特征表示的生成、图片搜索方法和装置。所述查询式的图片特征表示的生成方法包括：根据搜索用户的图片点击日志，获取与目标查询式对应的点击图片集合；根据图片的图片内容特征，生成与点击图片集合中各点击图片对应的图片特征向量；根据所述图片特征向量，对各所述点击图片进行聚类，并将聚类结果作为所述目标查询式的图片特征表示。本发明的技术方案优化了现有的图片搜索技术，并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。



1. 一种查询式的图片特征表示的生成方法,其特征在於,包括:
根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合;
根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量;
根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示;
其中,根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示包括:
根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;
获取所述聚类簇集合中包括的异常聚类簇;
将滤除所述异常聚类簇后的所述聚类簇集合作为所述目标查询式的图片特征表示。
2. 根据权利要求1所述的方法,其特征在於,所述图片特征向量包括:类别分布向量、基于颜色分布直方图的颜色分布向量或者视觉词分布向量。
3. 根据权利要求2所述的方法,其特征在於,根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量包括:
将待处理点击图片的图片像素内容输入至类别分类器组中;其中,所述类别分类器组通过设定数量的、经过分类标注的训练数据训练生成;
获取所述类别分类器组输出的,由所述待处理点击图片分属于不同类别的概率值构成的第一向量;
将所述第一向量中包括的各概率值进行排序后,获取设定数量的有效概率值;
将由所述有效概率值所组成的第二向量作为与所述待处理点击图片对应的所述类别分布向量。
4. 根据权利要求1所述的方法,其特征在於,所述目标查询式具体包括:单一查询式或者满足设定相似度阈值条件的至少两个查询式。
5. 根据权利要求1所述的方法,其特征在於,获取所述聚类簇集合中包括的异常聚类簇包括:
获取与所述目标查询式的分类不匹配的聚类簇作为所述异常聚类簇。
6. 根据权利要求1所述的方法,其特征在於,获取所述聚类簇集合中包括的异常聚类簇包括:
将所述聚类簇集合中的各聚类簇按照包括的元素个数进行排序;
根据排序结果,保留设定数量的聚类簇作为有效聚类簇,将所述聚类簇集合中除去所述有效聚类簇之外的聚类簇作为所述异常聚类簇。
7. 根据权利要求1所述的方法,其特征在於,所述聚类算法为加权聚类算法;
在根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示之前,还包括:
根据各所述点击图片的权重参数,确定各点击图片的权重值;其中,所述权重参数包括与所述点击图片对应的点击率。
8. 根据权利要求1所述的方法,其特征在於,所述聚类簇集合中的聚类簇具体包括:质

心、聚类簇中包括的元素数量、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

9. 一种图片搜索方法,其特征在于,包括:

根据用户输入的查询式,确定与所述查询式对应的图片特征表示;其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成;

根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值;

根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户;

所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成,包括:

根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;

获取所述聚类簇集合中包括的异常聚类簇;

将滤除所述异常聚类簇后的所述聚类簇集合作为所述查询式的图片特征表示。

10. 根据权利要求9所述的方法,其特征在于,所述图片特征表示具体包括:聚类簇集合;

其中,所述聚类簇集合中的聚类簇具体包括:质心、聚类簇中包括的元素个数、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

11. 根据权利要求10所述的方法,其特征在于,根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值包括:

依次获取一个待排序图片作为当前比对图片,并获取所述当前比对图片的图片特征向量作为待比对向量,其中,所述待比对向量与所述聚类簇包括中各元素的图片特征向量的向量类型以及向量长度相同;

分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度,并将计算得到的最大相似度作为所述当前比对图片与所述查询式之间的相关性度量值;

返回执行获取一个待排序图片的操作,直至完成对全部待排序图片的处理。

12. 根据权利要求11所述的方法,其特征在于,具体的,分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度具体包括:

分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

13. 根据权利要求11所述的方法,其特征在于,分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度具体包括:

分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离;

将计算得到的各向量距离除以各聚类簇的标准方差的结果,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

14. 一种查询式的图片特征表示的生成装置,其特征在于,包括:

点击图片集合获取模块,用于根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合;

图片特征向量生成模块,用于根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量;

图片特征表示生成模块,用于根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示;

其中,所述图片特征表示生成模块具体用于:

根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;

获取所述聚类簇集合中包括的异常聚类簇;

将滤除所述异常聚类簇后的所述聚类簇集合作为所述目标查询式的图片特征表示。

15. 根据权利要求14所述的装置,其特征在于,所述聚类簇集合中的聚类簇具体包括:质心、聚类簇中包括的元素数量、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

16. 一种图片搜索装置,其特征在于,包括:

图片特征表示确定模块,用于根据用户输入的查询式,确定与所述查询式对应的图片特征表示;其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成;

相关性度量值计算模块,用于根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值;

图片搜索结果提供模块,用于根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户;

所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成,包括:

根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;

获取所述聚类簇集合中包括的异常聚类簇;

将滤除所述异常聚类簇后的所述聚类簇集合作为所述查询式的图片特征表示。

17. 根据权利要求16所述的装置,其特征在于,所述图片特征表示具体包括:聚类簇集合;

其中,所述聚类簇集合中的聚类簇具体包括:质心、聚类簇中包括的元素个数、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

18. 根据权利要求17所述的装置,其特征在于,所述相关性度量值计算模块具体用于:

依次获取一个待排序图片作为当前比对图片,并获取所述当前比对图片的图片特征向量作为待比对向量,其中,所述待比对向量与所述聚类簇包括中各元素的图片特征向量的向量类型以及向量长度相同;

分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度,并将计算得到的最大相似度作为所述当前比对图片与所述查询式之间的相关性度量值;

返回执行获取一个待排序图片的操作,直至完成对全部待排序图片的处理。

查询式的图片特征表示的生成、图片搜索方法和装置

技术领域

[0001] 本发明实施例涉及信息处理技术,尤其涉及一种查询式的图片特征表示的生成、图片搜索方法和装置。

背景技术

[0002] 图片搜索是指用户输入自然语言查询,例如,通过搜索引擎提供的文本输入框输入的查询式(也称为Query),从图片集合中查找并按照相关性等指标,返回排过序的图片结果给用户的信息检索过程。相关性是搜索引擎最主要的性能指标之一,它度量了返回结果与用户查询需求的相关程度。对于图片搜索引擎,返回的图片是无结构的像素格式,而用户输入的Query是文本格式,这是两种完全不同的信息格式,两者无法直接进行运算。

[0003] 目前主要通过下述三种方式描述图片搜索的相关性特征:1、文本匹配特征,通过图片周边文本与Query比较计算得到;2、分类匹配特征,对图片内容分类得到分类标签,再与Query比较计算得到;3、点击率特征,通过统计大量用户查询后的点击行为等得到特定图片与该Query的相关性度量。

[0004] 上述三种描述图片搜索相关性特性的方法均具有一定的局限性:

[0005] 特征文本匹配特征:图片的周边文本可能存在与图片内容的不一致性,很多情况下不能完整准确的描述图片的内容,因而会影响文本匹配特征的精准性;

[0006] 分类匹配特征:受限于类别体系的完整性和分类模型的准确性。一般的,类别体系越精细,分类难度越大,分类模型的准确性越差,且分类结果与Query文本的语义偏差越大,匹配越困难。但如果类别体系太粗略,与Query的匹配精度又不够。因此,该特征一般只起辅助作用;

[0007] 点击率特征:主要基于用户行为统计,一方面存在偏倚和噪声,另一方面有稀疏性,只有在高频Query下展现靠前且展现次数充足的图片,能够统计到较充分的点击,其它情况,统计不到点击,或者点击非常稀疏,缺少统计意义。

发明内容

[0008] 有鉴于此,本发明实施例提供一种查询式的图片特征表示的生成、图片搜索方法和装置,以优选现有的图片搜索技术,提高图片搜索结果与用户输入的查询式之间的相关度。

[0009] 在第一方面,本发明实施例提供了一种查询式的图片特征表示的生成方法,包括:

[0010] 根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合;

[0011] 根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量;

[0012] 根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示。

[0013] 在第二方面,本发明实施例还提供了一种图片搜索方法,包括:

[0014] 根据用户输入的查询式,确定与所述查询式对应的图片特征表示;其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成;

[0015] 根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值;

[0016] 根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户。

[0017] 在第三方面,本发明实施例提供了一种查询式的图片特征表示的生成装置,包括:

[0018] 点击图片集合获取模块,用于根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合;

[0019] 图片特征向量生成模块,用于根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量;

[0020] 图片特征表示生成模块,用于根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示。

[0021] 在第四方面,本发明实施例还提供了一种图片搜索装置,包括:

[0022] 图片特征表示确定模块,用于根据用户输入的查询式,确定与所述查询式对应的图片特征表示;其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成;

[0023] 相关性度量值计算模块,用于根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值;

[0024] 图片搜索结果提供模块,用于根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户。

[0025] 本发明实施例根据搜索用户的图片点击日志,获取与同一目标搜索式对应的点击图片集合,并生成与该点击图片集合对应的图片特征向量,基于该图片特征向量对各点击图片进行聚类,并将聚类结果作为所述目标搜索式的图片特征表示,当图片搜索引擎接收到用户输入的图片查询式后,通过确定与该图片查询式对应的图片特征表示,可以计算得到各待排序图片与所述图片查询式之间的相关性度量值,进而基于所述相关性度量值,将各待排序图片进行排序并将排序结果返回给所述用户,优化了现有的图片搜索技术,并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。

附图说明

[0026] 图1是本发明第一实施例的一种查询式的图片特征表示的生成方法的流程图;

[0027] 图2是本发明第二实施例的一种查询式的图片特征表示的生成方法的流程图;

[0028] 图3是本发明第三实施例的一种查询式的图片特征表示的生成方法的流程图;

[0029] 图4是本发明第四实施例的一种图片搜索方法的流程图;

[0030] 图5是本发明第五实施例的一种图片搜索方法的流程图;

[0031] 图6是本发明第六实施例的一种查询式的图片特征表示的生成装置的结构图;

[0032] 图7是本发明第七实施例的一种图片搜索装置的结构图。

具体实施方式

[0033] 为了使本发明的目的、技术方案和优点更加清楚,下面结合附图对本发明具体实施例作进一步的详细描述。可以理解的是,此处所描述的具体实施例仅仅用于解释本发明,而非对本发明的限定。

[0034] 另外还需要说明的是,为了便于描述,附图中仅示出了与本发明相关的部分而非全部内容。在更加详细地讨论示例性实施例之前应当提到的是,一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作(或步骤)描述成顺序的处理,但是其中的许多操作可以被并行地、并发地或者同时实施。此外,各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止,但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

[0035] 为了清楚、明白的描述本发明各实施例的内容,首先重点描述一下本发明的发明构思。

[0036] 如前所述,背景技术中介绍的三种描述图片搜索的相关性特征的方式,并未直接基于图片内容计算图片与查询式之间的相关性,而是通过图片的周边文本、图片的分类以及图片的点击率等与图片内容间接关联的特征计算图片与查询式之间的相关性。相区别的,在本申请中,发明人创造性的提出:建立查询式的图片特征表示,该图片特征表示中包括有与查询式的点击图片日志对应的图片的图片特征向量,通过将查询式的图片特征表示与各待排序的图片的图片特征向量进行相关度计算,可以得到查询式与各图片之间的相关性度量值,进而可以准确的实现针对用户输入的查询式进行图片搜索的目的。

[0037] 即:在某个查询式下一张图片的点击率(点击次数/展现次数)正相关于该图片与该查询式的相关性。在此基础上:

[0038] 1)通过对查询式下的点击图片聚类获得对该查询式的一种表示,这种表示与图片的表示格式相同,从而能够与图片直接进行匹配计算。这一步可以离线实现也可以在线实现。

[0039] 2)在在线检索阶段,使用前一步获得的表示,与该查询式下的每个待排序图片进行匹配计算,得到待排序图片与该查询式在这种表示上的相关性度量值,可以作为图片与查询式相关性的一维特征。

[0040] 第一实施例

[0041] 图1为本发明第一实施例提供的一种查询式的图片特征表示的生成方法的流程图,本实施例的方法可以由查询式的图片特征表示的生成装置来执行,该装置可通过硬件和/或软件的方式实现,并一般可集成于用于生成查询式的图片特征表示的服务器中。本实施例的方法具体包括:

[0042] 110、根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合。

[0043] 一般来说,当用户通过图片搜索引擎输入查询式后,会基于与该查询式对应的图片搜索结果点击查看一个或者多个图片。相应的,搜索引擎会将用户输入的查询式以及基于所述查询式点击图片信息存储于图片点击日志中。

[0044] 通过收集网络中不同用户的图片点击日志,并按照不同的查询式进行点击图片的汇总,可以获取与各目标查询式对应的点击图片集合。

[0045] 其中,所述点击图片集合中包括有用户基于所述目标查询式点击的一个或者多个图片。

[0046] 在本实施例中,所述目标查询式可以仅包括单一查询式;进一步的,考虑到用户点击的稀疏性,针对那些点击较少或者无点击的冷门查询式,可以根据语义相似性使得上述冷门查询式与其他查询式共享点击图片,相应的,所述目标查询式还可以包括满足设定相似度阈值条件的至少两个查询式。

[0047] 在一个具体的例子中,可以直接选取“生日贺卡”作为目标查询式,还可以通过语义相似度聚类的方式,将“生日贺卡”、“寿诞贺卡”以及“生辰贺卡”作为目标查询式。

[0048] 120、根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量。

[0049] 一般来说,根据图片的不同的图片内容特征,可以生成图片的不同类型的图片特征向量。

[0050] 在本实施例中,对图片特征向量的具体形式并不进行限定,只需要保证最后生成的任意图片特征向量之间可以计算距离(或者说相似度)即可。

[0051] 在一个具体的例子中,如果基于图片的低级视觉特征,所生成的图片特征向量可以为基于颜色分布直方图的颜色分布向量或者视觉词分布向量;如果基于图片的语义特征,所生成的图片特征向量可以为类别分布向量。

[0052] 130、根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示。

[0053] 考虑到与同一目标查询式对应的点击图片的数量可能很多,如果直接将全部点击图片所对应的图片特征向量作为所述目标查询式的图片特征表示,会使图片特征表示包含的信息量巨大,进而也会使基于该图片特征表示的图片排序过程的计算量复杂,排序的实时性变差。

[0054] 在本实施例中,考虑到虽然用户基于同样的目标查询式会点击不同的图片,但是很多图片之间是具有一定的相关性的,在一个例子中,基于目标查询式“范冰冰”的点击图片集合,虽然基于该查询式的搜索结果不同用户会点击查看不同的图片,但是用户的实际的查询需求是具有一定的相关性的,例如,有些用户具有点击查看范冰冰的剧照的需求,有些用户具有点击查看范冰冰的现场照的需求等。

[0055] 因此,可以根据所述图片特征向量,对各所述点击图片进行聚类,针对同一聚类簇中的图片,可以通过仅使用该聚类簇中的一个或者少数几个图片来代表同一聚类簇中全部图片的方式,或者将簇元素较少的聚类簇作为噪声进行删除等方式,对所述点击图片集合中各点击图片对应的图片特征向量进行优化处理,以优化所述目标查询式的图片特征表示中包括的图片特征向量的数量。

[0056] 在本实施例中,可以根据图片特征向量的不同形式,采取不同类型的聚类算法,典型的,可以采取K-means(K均值)聚类算法或者层次聚类算法等,可以采取加权聚类算法,也可以采取非加权聚类算法等,本实施例对此并不进行限制。

[0057] 在本实施例的一个优选的实施方式中,所述聚类算法可以为加权聚类算法;

[0058] 相应的,在根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示之前,还可以包括:

[0059] 根据各所述点击图片的权重参数,确定各点击图片的权重值。

[0060] 其中,所述权重参数可以为与所述点击图片对应的点击率,也可以为与所述点击

图片对应的图片热度,还可以为图片的质量特征参数等,本实施例对此并不进行限制。

[0061] 本发明实施例根据搜索用户的图片点击日志,获取与同一目标搜索式对应的点击图片集合,并生成与该点击图片集合对应的图片特征向量,基于该图片特征向量对各点击图片进行聚类,并将聚类结果作为所述目标搜索式的图片特征表示,当图片搜索引擎接收到用户输入的图片查询式后,通过确定与该图片查询式对应的图片特征表示,可以计算得到各待排序图片与所述图片查询式之间的相关性度量值,进而基于所述相关性度量值,将各待排序图片进行排序并将排序结果返回给所述用户,优化了现有的图片搜索技术,并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。

[0062] 第二实施例

[0063] 图2是本发明第二实施例的一种查询式的图片特征表示的生成方法的流程图。本实施例以上述实施例为基础进行优化,在本实施例中,将所述图片特征向量具体优化为:类别分布向量;

[0064] 相应的,将根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量具体优化为:将待处理点击图片的图片像素内容输入至类别分类器组中;其中,所述类别分类器组通过设定数量的、经过分类标注的训练数据训练生成;获取所述类别分类器组输出的,由所述待处理点击图片分属于不同类别的概率值构成的第一向量;将所述第一向量中包括的各概率值进行排序后,获取设定数量的有效概率值;将由所述有效概率值所组成的第二向量作为与所述待处理点击图片对应的所述类别分布向量。

[0065] 相应的,本实施例的方法具体包括:

[0066] 210、根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合。

[0067] 在本实施例中,根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合具体还可以包括:滤除所述图片点击日志中包括的噪声日志。

[0068] 这样设置的原因是:用户的图片点击日志里可能会存在大量噪声,例如有些不适宜的图片或者与相关图片反差过大的恶意图片等,反而会吸引用户点击,甚至出现在任何查询下都会引起点击。有些相关结果很多的查询式,用户在浏览到靠前的相关图片后,需求已经满足,后面的相关图片被点击的概率将会大幅下降。这两种行为都会造成点击/无点击、点击多少与相关性的失真,因此为了进一步提高所选取的正负样本图片的准确性,需要滤除所述图片点击日志中包括的噪声日志。

[0069] 识别和去除噪声日志是保证训练出的模型的准确性的必要操作。这里简单介绍两种方法:

[0070] 1) 点击查询式聚类法。汇总一张图片(包括重复图片、相似图片)被点击的所有查询式(以下称为点击查询式)并聚类,可以得到该图片所满足的主要需求类别,小众的类别可以认为是噪声从而去掉。与主要需求类别的距离都较远的点击查询式,可以作为为噪声点击。

[0071] 2) 图片聚类法。汇总一个查询式(包括语义相同、相似的查询式)下的所有点击图片,对这些图片的分类结果或者分类表示进行聚类,可以得到满足该查询式需求的主要图片类别,小众的类别可以认为是噪声从而去掉。

[0072] 220、将待处理点击图片的图片像素内容输入至类别分类器组中。

[0073] 其中,所述类别分类器组通过设定数量的、经过分类标注的训练数据训练生成。

[0074] 在本实施例中,可以预先定义一个较全面精细的分类体系,例如千以上的类别,预先标注足够多的训练数据,训练一个类别分类器组。该类别分类器组的输入为图片像素内容,输出为图片属于每个类别的概率所构成的一个向量,即图片的类别分布向量。

[0075] 230、获取所述类别分类器组输出的,由所述待处理点击图片分属于不同类别的概率值构成的第一向量。

[0076] 240、将所述第一向量中包括的各概率值进行排序后,获取设定数量的有效概率值。

[0077] 250、将由所述有效概率值所组成的第二向量作为与所述待处理点击图片对应的所述类别分布向量,并将所述类别分布向量作为所述图片特征向量。

[0078] 在本实施例中,为了精简类别分布向量的向量长度,可以对所生成的类别分布向量可以做一些简化,例如,按概率大小排序后,只保留前R个类别和概率值,其中,R为大于1的整数。

[0079] 260、根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示。

[0080] 本实施例的技术方案基于图片的语义特征,生成图片的类别分布向量作为图片特征向量,并基于该类别分布向量生成与目标查询式对应的图片特征表示,使得最终生成的图片特征表示与查询式之间的契合度更高,进而可以进一步提高搜索引擎基于所述图片特征表示的图片排序结果更加符合搜索用户的实际需求。

[0081] 第三实施例

[0082] 图3是本发明第三实施例的一种查询式的图片特征表示的生成方法的流程图。本实施例以上述实施例为基础进行优化,在本实施例中,将根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示具体优化为:根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;获取所述聚类簇集合中包括的异常聚类簇;将滤除所述异常聚类簇后的所述聚类簇集合作为所述目标查询式的图片特征表示。

[0083] 相应的,本实施例的方法具体包括:

[0084] 310、根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合。

[0085] 320、根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量。

[0086] 330、根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合。

[0087] 在本实施例中,可以定义 $C_q = \{I_1^{cq}, I_2^{cq}, \dots, I_m^{cq}\}$ 为查询式q下的点击图片集合,其中,m为点击图片集合中包括的图片总数量,m为大于等于1的整数;依次标记每个图片 I_i^{cq} , $i \in [1, m]$ 的权重值 W_i , 其中,可以通过点击率来标记所述 W_i , 同时,使用 f_i 标记每个图片的图片特征向量,典型的,该图片特征向量 f_i 为一个定长向量。

[0088] 使用设定聚类算法对 f_i , $i \in [1, m]$ 进行聚类,生成聚类簇集合,其中,如果选取的聚类算法为带权聚类算法,则需要使用 W_i , $i \in [1, m]$ 对 f_i 进行加权处理。

[0089] 在本实施例的一个优选的实施方式中,所述聚类簇集合中的聚类簇具体包括:质

心、聚类簇中包括的元素数量、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

[0090] 如前例,假定最终将 $f_i, i \in [1, m]$ 聚为N类,N为大于1的整数,聚类后的第k个聚类簇 CL_k 可以表示为:

[0091] $CL_k = (f_k^c, n_k, \{f_{k1}, f_{k2}, \dots, f_{kn_k}\}, \sigma_k), k = 1, 2, \dots, N;$

[0092] 其中, f_k^c 为第k个聚类簇的质心(可以是真实质心,也可以是与真实质心接近的某个图片特征向量),同时, f_k^c 与 $f_i, i \in [1, m]$ 具有相同的向量格式, n_k 为第k个聚类簇中包括的元素数量, $\{f_{k1}, f_{k2}, \dots, f_{kn_k}\}$ 为第k个聚类簇中的元素, σ_k 为第k个聚类簇中元素的标准方差,用于度量该聚类簇中元素的聚集程度。(也可以使用其它的量如散度来度量上述元素的聚集程度)。

[0093] 此外,质心、标准方差(散度)的计算所使用的距离依 $f_i, i \in [1, m]$ 的向量形式而定,例如欧式距离、余弦距离等。聚类数量N的个数由实验决定。

[0094] 340、获取所述聚类簇集合中包括的异常聚类簇。

[0095] 在本实施例的一个优选的实施方式中,获取所述聚类簇集合中包括的异常聚类簇可以包括:

[0096] 获取与所述目标查询式的分类不匹配的聚类簇作为所述异常聚类簇。

[0097] 在本实施例的另一个优选的实施方式中,获取所述聚类簇集合中包括的异常聚类簇可以包括:将所述聚类簇集合中的各聚类簇按照包括的元素个数进行排序;根据排序结果,保留设定数量的聚类簇作为有效聚类簇,将所述聚类簇集合中除去所述有效聚类簇之外的聚类簇作为所述异常聚类簇。

[0098] 350、将滤除所述异常聚类簇后的所述聚类簇集合作为所述目标查询式的图片特征表示。

[0099] 需要重点说明的是:本实施例的技术方案可以离线实现,也可以在线实现。如果通过离线实现,则可以对每个目标查询式离线生成对应的图片特征表示,以Key-Value(键值对)大表的形式进行存储(典型的,Key为目标查询式,Value为图片特征表示),供在线图片搜索时调用,如果通过在线实现,则可以在用户在线搜索时,直接获取与用户当前输入的查询式对应的点击图片,在线聚类并生成相应的图片特征表示。

[0100] 本实施例的技术方案对于任何一个有一定数量的点击图片的目标查询式,都可以得到该目标查询式的一个图片特征表示,这个图片特征表示与图片的图片特征向量具有相同的格式,因而可以直接与图片的图片特征向量进行匹配计算。其次,基于点击图片聚类得到的图片特征表示,保证了这个图片特征表示既能真实反应目标查询式的实际需求,也能较好反应目标查询式需求的多样性,还可以从一定程度上起到过滤点击噪音的作用。

[0101] 另外,需要再次强调的是,本发明实施例与现有技术的区别主要在于:

[0102] 1) 本发明实施例所生成的查询式的图片特征表示突破了文本的限制,使得查询式有了图片内容层面上的表示,可以直接与图片内容进行匹配计算,这是与文本匹配特征和分类匹配特征相比的优点。

[0103] 2) 本发明实施例所生成的查询式的图片特征表示可以基于任意的图片内容特征。而分类匹配特征在图片分类的类别体系过小时,其分类结果的语义不精细,类别体系过大

时,分类的准确率又较低,且与查询式的文本(或者类别标签)的匹配难度急剧增大(即所谓的图片与文本的语义鸿沟)。与分类匹配特征相比,本发明避免了这些问题。

[0104] 3) 本发明实施例所生成的查询式的图片特征表示来自于查询式下的点击图片,然后可以与任意图片进行匹配计算,将用户点击行为所包含的图片对查询式的相关性度量推广到任意的无点击图片或稀疏点击图片。而点击率特征只适用于查询式之下统计到有效点击的图片。这是本发明相比较于点击率特征的优点。

[0105] 综上,本发明在查询式的图片特征表示、查询式与图片内容的匹配、点击推广等几个方面的能力比现有技术及各种融合和变种的能力更强,对所期望解决的问题解决得更彻底。

[0106] 第四实施例

[0107] 图4为本发明第四实施例提供的一种图片搜索方法的流程图,本实施例的方法可以由图片搜索装置来执行,该装置可通过硬件和/或软件的方式实现,并一般可集成于图片搜索引擎所在的服务器中。本实施例的方法具体包括:

[0108] 410、根据用户输入的查询式,确定与所述查询式对应的图片特征表示。

[0109] 在本实施例中,所述查询式具体是指用户通过图片搜索引擎输入的,文本形式的查询式,例如:“生日贺卡”。

[0110] 其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成。

[0111] 在本实施例的一个优选的实施方式中,所述图片特征表示具体包括:聚类簇集合;

[0112] 其中,所述聚类簇集合中的聚类簇具体包括:质心(可以是真实质心,也可以是与真实质心接近的某个图片特征向量)、聚类簇中包括的元素个数、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

[0113] 同时,所述质心与聚类簇中各元素的图片特征向量具有相同的向量格式。

[0114] 420、根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值。

[0115] 其中,可以选取所述图片特征表示中各聚类簇的所述质心,与各待排序图片对应的图片特征向量计算相似度,并根据计算结果确定所述查询式与各待排序图片之间的相关性度量值;也可以选取所述图片特征表示中设定数量的聚类簇中包括的设定数量图片特征向量,与各待排序图片对应的图片特征向量计算相似度,并根据计算结果确定所述查询式与各待排序图片之间的相关性度量值等,本实施例对此并不进行限制。

[0116] 430、根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户。

[0117] 本发明实施例的技术方案在图片搜索引擎接收到用户输入的图片查询式后,通过确定与该图片查询式对应的图片特征表示,可以计算得到各待排序图片与所述图片查询式之间的相关性度量值,进而基于所述相关性度量值,将各待排序图片进行排序并将排序结果返回给所述用户,优化了现有的图片搜索技术,并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。

[0118] 第五实施例

[0119] 图5是本发明第五实施例的一种图片搜索方法的流程图。本实施例以上述实施例

为基础进行优化,在本实施例中,将根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值具体优化为:依次获取一个待排序图片作为当前比对图片,并获取所述当前比对图片的图片特征向量作为待比对向量,其中,所述待比对向量与所述聚类簇包括中各元素的图片特征向量的向量类型以及向量长度相同;分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度,并将计算得到的最大相似度作为所述当前比对图片与所述查询式之间的相关性度量值;返回执行获取一个待排序图片的操作,直至完成对全部待排序图片的处理。

[0120] 相应的,本实施例的方法具体包括:

[0121] 510、根据用户输入的查询式,确定与所述查询式对应的图片特征表示。

[0122] 520、依次获取一个待排序图片作为当前比对图片,并获取所述当前比对图片的图片特征向量作为待比对向量。

[0123] 其中,所述待比对向量与所述聚类簇包括中各元素的图片特征向量的向量类型以及向量长度相同。

[0124] 530、分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度,并将计算得到的最大相似度作为所述当前比对图片与所述查询式之间的相关性度量值。

[0125] 在本实施例的一个优选的实施方式中,分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度具体可以包括:

[0126] 分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

[0127] 在本实施例的另一个优选的实施方式中,分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度具体可以包括:

[0128] 分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离;将计算得到的各向量距离除以各聚类簇的标准方差的结果,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

[0129] 在一个具体的例子中,与用户输出的查询式 q 对应的图片特征表示 $\text{Exp}_q = \{CL_1, CL_2, \dots, CL_K\}$, $CL_i, i \in [1, K]$,代表第 i 个聚类簇;

[0130] $CL_i = (f_i^c, n_i, \{f_1, f_{i2}, \dots, f_{in_i}\}, \sigma_i), i = 1, 2, \dots, K;$

[0131] f_i^c 为第 i 个聚类簇的质心(可以是真实质心,也可以是与真实质心接近的某个图片特征向量), n_i 为第 i 个聚类簇中包括的元素数量, $\{f_1, f_{i2}, \dots, f_{in_i}\}$ 为第 i 个聚类簇中的元素, σ_i 为第 i 个聚类簇中元素的标准方差,用于度量该聚类簇中元素的聚集程度。(也可以使用其它的量如散度来度量上述元素的聚集程度)。

[0132] 通过上述图片特征表示 Exp_q 说明该查询式 q 可能有 K 种有效需求。待排序图片只要满足任何一种需求,都可以认为与该查询式 q 相关。所以相关性度量值 ReI 可以定义为:

[0133] $\text{ReI}(q, f_j) = \max_i \text{dist}(f_j, f_i^c), i = 1, 2, \dots, K;$

[0134] f_j 为待排序图片的图片特征向量, f_i^c 为簇 CL_i 的质心, $\text{dist}(\cdot)$ 为距离定义,例如欧式距离或者余弦距离等。

[0135] 考虑到不同聚类簇中图片的聚集程度不同,为了使得相关性度量值 ReI 在数值上

可比,可以使用聚类簇的标准方差或者散度对ReI数值进行正则化。使用标准方差进行正则化示例如下:

$$[0136] \quad \text{ReI}_{\text{norm}}(q, f_j) = \max_i \frac{\text{dist}(f_j, f_i^c)}{\sigma_i}, i=1, 2, \dots, K。$$

[0137] 540、判断是否完成对全部待排序图片的处理:若是,执行550;否则,返回执行520。

[0138] 550、根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户。

[0139] 通过实验表明,使用本实施例的技术方案显著提升了图片搜索的相关性。评估表明,增加该图片特征表示之后,图片搜索在随机查询式上的搜索结果满意度提升15%以上,即搜索结果可感知变好的查询式数量与搜索结果可感知变差的查询式数量之差,占随机抽样查询式的比例在15%以上,效果非常显著。

[0140] 第六实施例

[0141] 图6是本发明第六实施例的一种查询式的图片特征表示的生成装置的结构图。如图6所示,所述装置包括:

[0142] 点击图片集合获取模块61,用于根据搜索用户的图片点击日志,获取与目标查询式对应的点击图片集合。

[0143] 图片特征向量生成模块62,用于根据图片的图片内容特征,生成与所述点击图片集合中各点击图片对应的图片特征向量。

[0144] 图片特征表示生成模块63,用于根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示。

[0145] 本发明实施例根据搜索用户的图片点击日志,获取与同一目标搜索式对应的点击图片集合,并生成与该点击图片集合对应的图片特征向量,基于该图片特征向量对各点击图片进行聚类,并将聚类结果作为所述目标搜索式的图片特征表示,当图片搜索引擎接收到用户输入的图片查询式后,通过确定与该图片查询式对应的图片特征表示,可以计算得到各待排序图片与所述图片查询式之间的相关性度量值,进而基于所述相关性度量值,将各待排序图片进行排序并将排序结果返回给所述用户,优化了现有的图片搜索技术,并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。

[0146] 在上述各实施例的基础上,所述图片特征向量可以包括:类别分布向量、基于颜色分布直方图的颜色分布向量或者视觉词分布向量。

[0147] 在上述各实施例的基础上,图片特征向量生成模块具体可以用于:

[0148] 将待处理点击图片的图片像素内容输入至类别分类器组中;其中,所述类别分类器组通过设定数量的、经过分类标注的训练数据训练生成;

[0149] 获取所述类别分类器组输出的,由所述待处理点击图片分属于不同类别的概率值构成的第一向量;

[0150] 将所述第一向量中包括的各概率值进行排序后,获取设定数量的有效概率值;

[0151] 将由所述有效概率值所组成的第二向量作为与所述待处理点击图片对应的所述类别分布向量。

[0152] 在上述各实施例的基础上,所述目标查询式具体可以包括:单一查询式或者满足设定相似度阈值条件的至少两个查询式。

[0153] 在上述各实施例的基础上,图片特征表示生成模块具体可以包括:

[0154] 聚类簇集合生成单元,用于根据与各点击图片对应的图片特征向量,以及与所述图片特征向量相匹配的聚类算法,对各所述点击图片进行聚类,生成聚类簇集合;

[0155] 异常聚类簇获取单元,用于获取所述聚类簇集合中包括的异常聚类簇;

[0156] 图片特征表示获取单元,用于将滤除所述异常聚类簇后的所述聚类簇集合作为所述目标查询式的图片特征表示。

[0157] 在上述各实施例的基础上,异常聚类簇获取单元具体可以用于:

[0158] 获取与所述目标查询式的分类不匹配的聚类簇作为所述异常聚类簇。

[0159] 在上述各实施例的基础上,异常聚类簇获取单元具体可以用于:

[0160] 将所述聚类簇集合中的各聚类簇按照包括的元素个数进行排序;

[0161] 根据排序结果,保留设定数量的聚类簇作为有效聚类簇,将所述聚类簇集合中除去所述有效聚类簇之外的聚类簇作为所述异常聚类簇。

[0162] 在上述各实施例的基础上,所述聚类算法可以为加权聚类算法;

[0163] 相应的,所述装置还可以包括:权重值确定模块,用于在根据所述图片特征向量,对各所述点击图片进行聚类,并将聚类结果作为所述目标查询式的图片特征表示之前,根据各所述点击图片的权重参数,确定各点击图片的权重值;其中,所述权重参数包括与所述点击图片对应的点击率。

[0164] 在上述各实施例的基础上,所述聚类簇集合中的聚类簇具体可以包括:质心、聚类簇中包括的元素数量、聚类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

[0165] 本发明实施例所提供的查询式的图片特征表示的生成装置可用于执行本发明任意实施例提供的查询式的图片特征表示的生成方法,具备相应的功能模块,实现相同的有益效果。

[0166] 第七实施例

[0167] 图7是本发明第七实施例的一种图片搜索装置的结构图。如图7所示,所述装置包括:

[0168] 图片特征表示确定模块71,用于根据用户输入的查询式,确定与所述查询式对应的图片特征表示;其中,所述查询式对应的图片特征表示通过对与所述查询式对应的点击图片集合中各图片的图片特征向量聚类生成。

[0169] 相关性度量值计算模块72,用于根据所述图片特征表示,计算所述查询式与各待排序图片之间的相关性度量值。

[0170] 图片搜索结果提供模块73,用于根据计算得到的所述相关性度量值对各所述待排序图片进行排序,并将与排序结果对应的图片搜索结果提供给所述用户。

[0171] 本发明实施例的技术方案在图片搜索引擎接收到用户输入的图片查询式后,通过确定与该图片查询式对应的图片特征表示,可以计算得到各待排序图片与所述图片查询式之间的相关性度量值,进而基于所述相关性度量值,将各待排序图片进行排序并将排序结果返回给所述用户,优化了现有的图片搜索技术,并可以大大提高图片搜索结果与用户输入的查询式之间的相关度。

[0172] 在上述各实施例的基础上,所述图片特征表示具体可以包括:聚类簇集合;

[0173] 其中,所述聚类簇集合中的聚类簇具体包括:质心、聚类簇中包括的元素个数、聚

类簇中各元素的图片特征向量以及聚类簇中元素的标准方差。

[0174] 在上述各实施例的基础上,所述相关性度量值计算模块具体可以用于:

[0175] 依次获取一个待排序图片作为当前比对图片,并获取所述当前比对图片的图片特征向量作为待比对向量,其中,所述待比对向量与所述聚类簇包括中各元素的图片特征向量的向量类型以及向量长度相同;

[0176] 分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的相似度,并将计算得到的最大相似度作为所述当前比对图片与所述查询式之间的相关性度量值;

[0177] 返回执行获取一个待排序图片的操作,直至完成对全部待排序图片的处理。

[0178] 在上述各实施例的基础上,所述相关性度量值计算模块具体还可以用于:

[0179] 分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

[0180] 在上述各实施例的基础上,所述相关性度量值计算模块具体还可以用于:

[0181] 分别计算所述待比对向量与所述聚类簇集合中的各聚类簇的质心之间的向量距离;

[0182] 将计算得到的各向量距离除以各聚类簇的标准方差的结果,作为所述待比对向量与所述聚类簇集合中的各聚类簇的相似度。

[0183] 本发明实施例所提供的图片搜索装置可用于执行本发明任意实施例提供的图片搜索方法,具备相应的功能模块,实现相同的有益效果。

[0184] 显然,本领域技术人员应该明白,上述的本发明的各模块或各步骤可以通过如上所述的服务器实施。可选地,本发明实施例可以用计算机装置可执行的程序来实现,从而可以将它们存储在存储装置中由处理器来执行,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等;或者将它们分别制作成各个集成电路模块,或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样,本发明不限制于任何特定的硬件和软件的结合。

[0185] 以上所述仅为本发明的优选实施例,并不用于限制本发明,对于本领域技术人员而言,本发明可以有各种改动和变化。凡在本发明的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

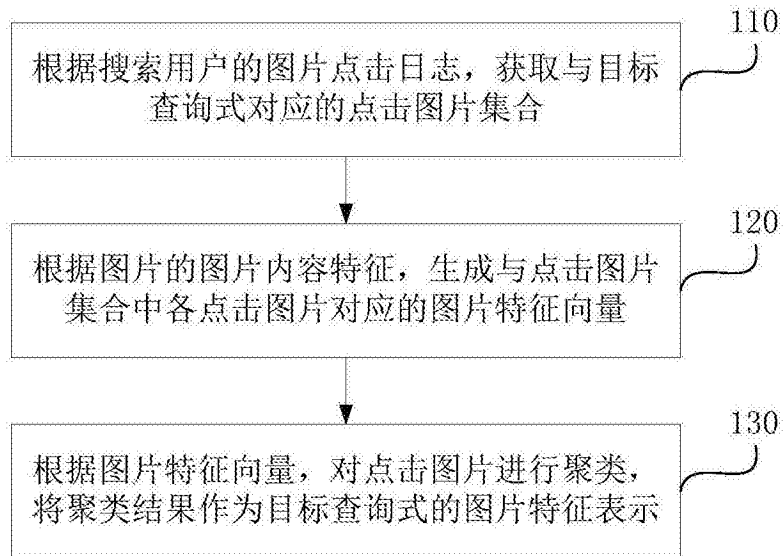


图1

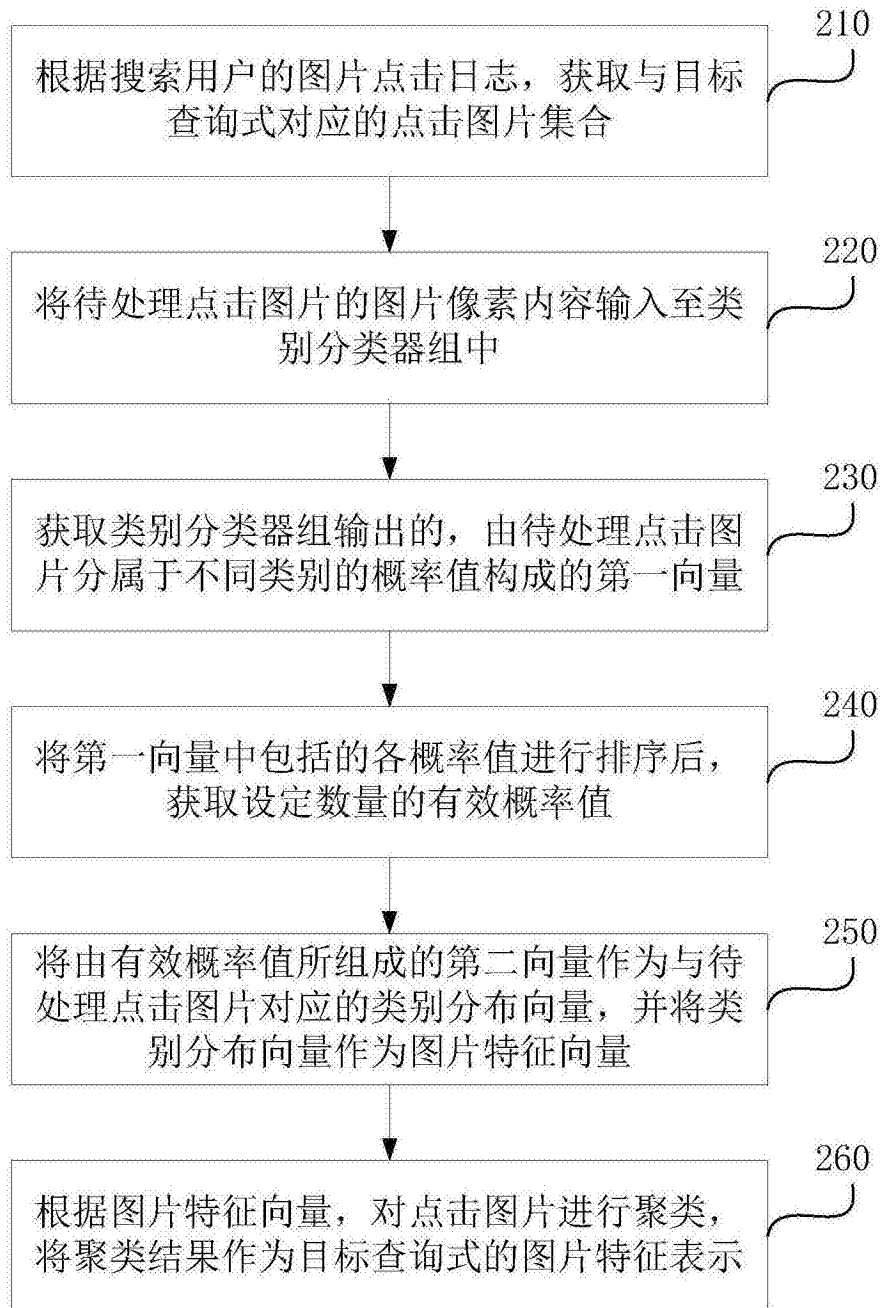


图2

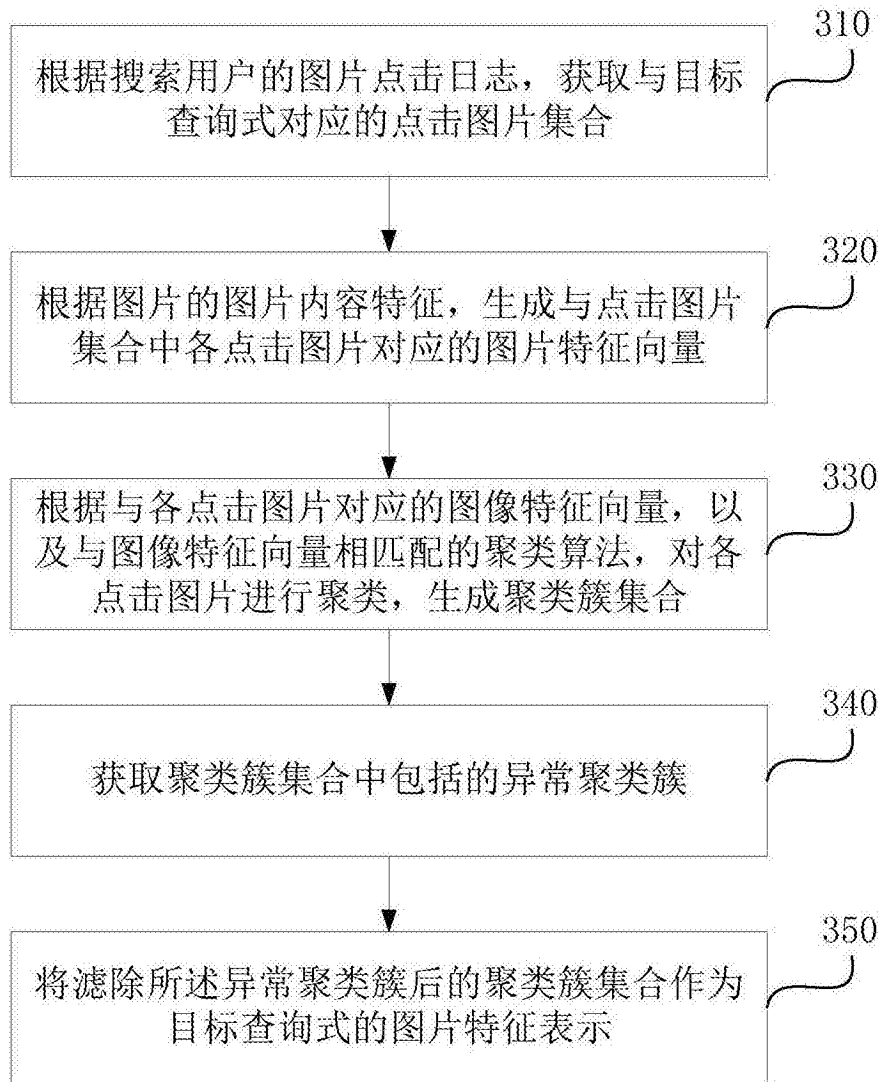


图3

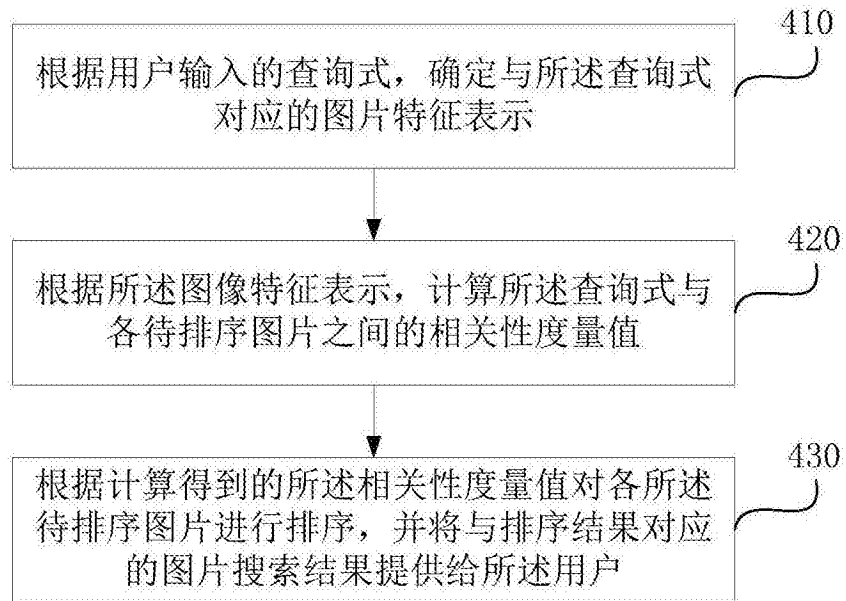


图4

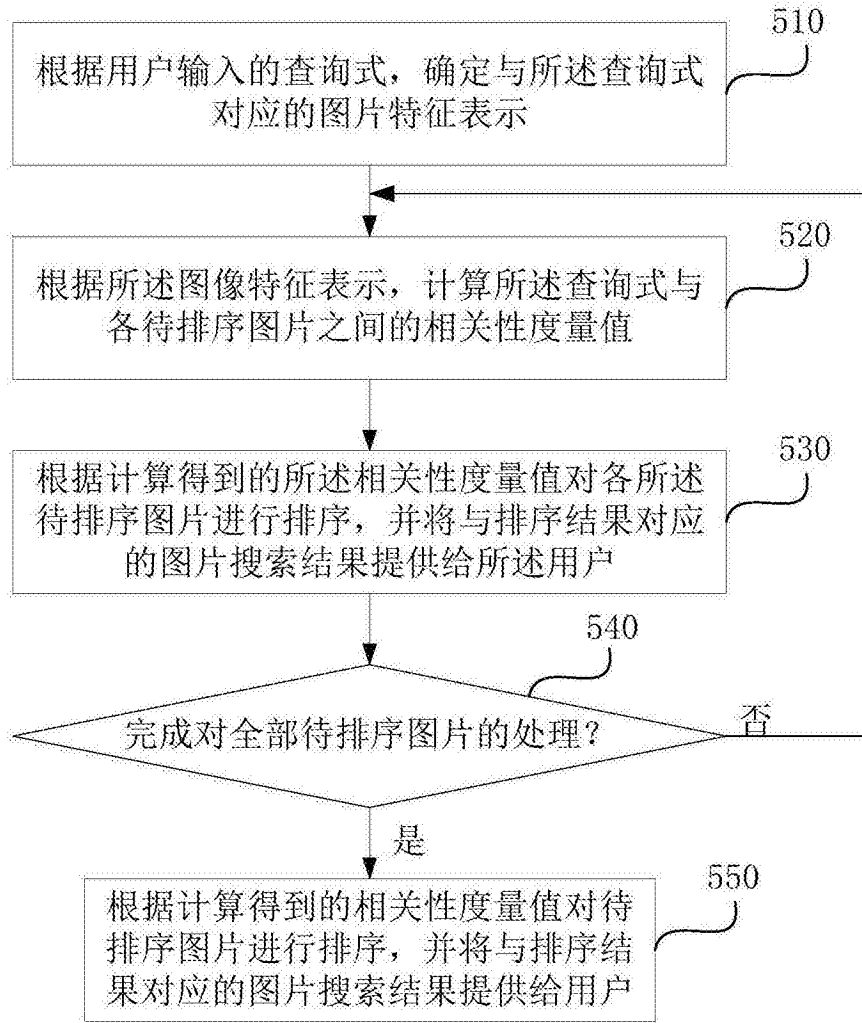


图5



图6

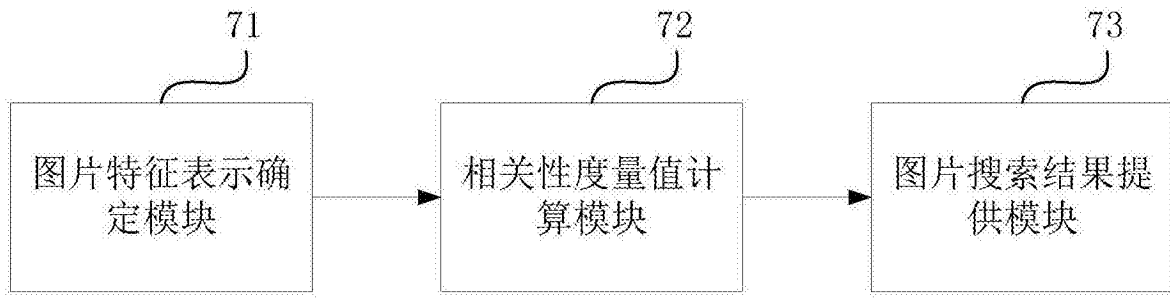


图7