

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2015 年 6 月 11 日 (11.06.2015)



(10) 国际公布号
WO 2015/081748 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2014/087109
- (22) 国际申请日: 2014 年 9 月 22 日 (22.09.2014)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201310632676.8 2013 年 12 月 2 日 (02.12.2013) CN
201310632631.0 2013 年 12 月 2 日 (02.12.2013) CN
- (71) 申请人: 北京奇虎科技有限公司 (BEIJING QIHOO TECHNOLOGY COMPANY LIMITED) [CN/CN]; 中国北京市西城区新街口外大街 28 号 D 座 112 室 (德胜园区), Beijing 100088 (CN)。奇智软件 (北京) 有限公司 (QIZHI SOFTWARE (BEIJING) COMPANY LIMITED) [CN/CN]; 中国北京市朝阳区酒仙桥路 6 号院 2 号楼 B 座 2 层、3 层 301-306 室, Beijing 100015 (CN)。
- (72) 发明人: 陶哲 (TAO, Zhe); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。白明 (BAI, Ming); 中国北京市朝阳区酒仙桥路 6 号院 2

号楼, Beijing 100015 (CN)。韩玉刚 (HAN, Yugang); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。

- (74) 代理人: 北京智汇东方知识产权代理事务所 (普通合伙) (WISEAST INTELLECTUAL PROPERTY LAW FIRM); 中国北京市海淀区成府路 28 号优盛大厦 D 座-1111 室, Beijing 100083 (CN)。
- (81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE,

[见续页]

(54) Title: METHOD AND SYSTEM FOR IDENTIFYING CONTENT ATTRIBUTE OF PICTURE

(54) 发明名称: 图片内容属性识别方法和系统

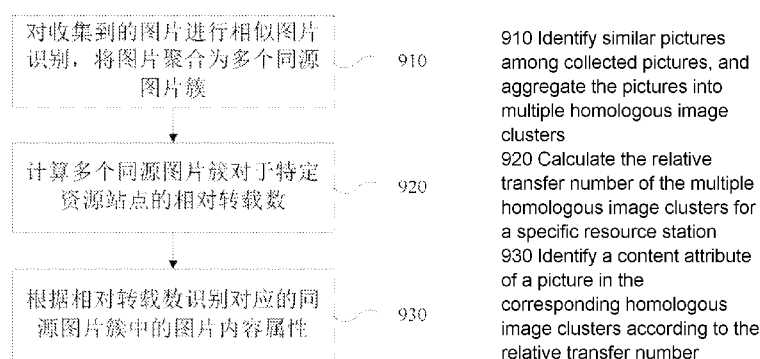


图 9 /Fig.9

(57) Abstract: A method and a system for identifying a content attribute of a picture. The method comprises: calculating the relative transfer number of multiple homologous image clusters for a specific resource station; and training and identifying a content attribute of a picture in the corresponding homologous image clusters according to the relative transfer number. The method and the system have the advantages that a content attributes of a picture can be identified according to data of the picture transferred or transmitted on a network, and the method and system are particularly used for determining whether the picture is an advertisement picture.

(57) 摘要: 一种图片内容属性识别方法和系统, 方法包括: 计算多个同源图片簇对于特定资源站点的相对转载数; 根据所述相对转载数训练识别对应的同源图片簇中的图片内容属性。优点在于, 根据图片在网络上被转载或传播的数据可以识别图片的内容属性, 尤其可以用于判断是否为广告图片。

WO 2015/081748 A1



IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,
RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD,
TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

图片内容属性识别方法和系统

技术领域

本发明涉及图像识别领域，具体涉及一种图片内容属性识别方法和系统。

5

背景技术

在网络中很多类型的资源站点上，都会出现一些广告图片，这些广告图片的种类非常丰富，其包括各类商品的广告（例如，关于奶粉、衣服的广告），和实体商店的广告，以及一些其他类型的广告。

10 这些广告图片不但会出现在商家的站点上，也会出现在其他资源站点的页面上，例如，在允许用户上传图片的社区（论坛、图片站等），会有一些用户上传广告图片。大量广告图片的存在，往往对用户造成干扰，甚至用户进行图片搜索时，也会出现与用户需求无关的广告图片。

从图片的图像内容角度来看，不同广告图片是没有特别多的相似点的，所以基于目前的图像识别技术，难以对图片的图片内容属性进行识别，即难以识别出哪些图片为广告图片，也就无法对广告图片进行针对性的处理，用户的体验必然受到广告图片的影响。

发明内容

20 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种图片内容属性识别方法和系统。

依据本发明的一个方面，提供了一种图片内容属性识别方法，其包括：计算多个同源图片簇对于特定资源站点的相对转载数；根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

25 可选地，所述计算多个同源图片簇对于特定资源站点的相对转载数之前，还包括：对收集到的图片进行相似图片识别，将图片聚合为多个同源图片簇。

可选地，根据所述相对转载数识别对应的同源图片簇中的图片内容属性，包括：

根据所述多个同源图片簇以及对应的相对转载数训练筛选器模型；

根据训练后的筛选器模型识别目标图片簇中的图片内容属性。

30 依据本发明的另一个方面，提供了一种图片内容属性识别系统，其包括：相对转载数计算模块，用于计算多个同源图片簇对于特定资源站点的相对转载数；图片内容属性识别模块，用于根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

根据本发明的又一个方面，提供了一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行根据权利要求 1-12 中的一个所述的图片内容属性识别方法。

35 根据本发明的再一个方面，提供了一种计算机可读介质，其中存储了如权利要求 23 所述的计算机程序。

根据本发明的图片内容属性识别方法和系统，利用了同源图片簇对于特定资源站点的相对转载数识别对应的同源图片簇中的图片内容属性，用于进行图片属性识别的数据选用相对转载数，这是因为相对转载数是能够反映图片在特定资源站点的站内站外比例的数据，而作为广告的图片的一个主要特点在于：在某一资源站点上转载的次数非常高，而在互联网范围内其他资源站点上转载的次数会相对地明显变少，因此相对转载数的大小可以用于区分别图片是否作为广告进行传播，利用相对转载数对图片的图片内容属性进行识别，可以准确地判断图片是否为广告图片。

进一步，将相似的图片聚合到同一同源图片簇中，对于一个同源图片簇而言，如果其中一张图片为广告图片，则其余图片也必然为广告图片，所以本发明的技术方案中以图片簇为单位进行图片内容属性的识别，以判断每个同源图片簇中的图片是否为广告图片。

进一步，利用了同源图片簇对于特定资源站点的相对转载数作为训练数据进行筛选器模型的训练，利用相对转载数进行的筛选器模型的训练，则得到的筛选器模型可以自行对图片的图片内容属性进行识别，准确地判断图片是否为广告图片。

上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

- 图 1 示出了根据本发明的一个实施例的图片内容属性识别方法的流程图；
- 图 2 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；
- 图 3 示出了根据本发明的一个实施例的图片内容属性识别方法的流程图；
- 图 4 示出了根据本发明的一个实施例的图片内容属性识别方法的部分流程图；
- 图 5 示出了根据本发明的一个实施例的图片内容属性识别方法的流程图；
- 图 6 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；
- 图 7 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；
- 图 8 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；
- 图 9 示出了根据本发明的一个实施例的图片内容属性识别方法的流程图；
- 图 10 示出了根据本发明的一个实施例的图片内容属性识别方法的部分流程图；
- 图 11 示出了根据本发明的一个实施例的图片内容属性识别方法的流程图；
- 图 12 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；

图 13 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；

图 14 示出了根据本发明的一个实施例的图片内容属性识别系统的框图；

图 15 示意性地示出了用于执行根据本发明的图片内容属性识别方法的计算设备的框图；以及

5 图 16 示意性地示出了用于保持或者携带实现根据本发明的图片内容属性识别方法的程序代码的存储单元。

具体实施方式

下面结合附图和具体的实施方式对本发明作进一步的描述。

10 为解决上述技术问题，本发明实施例提供了一种图片内容属性识别方法。图 1 示出了根据本发明一个实施例的图片内容属性识别方法的处理流程图。参见图 1，该方法包括：

步骤 S102、计算多个同源图片簇对于特定资源站点的相对转载数；

步骤 S104、根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

为支持图 1 所示的图片内容属性识别方法，本发明实施例还提供了一种图片内容属性识别系统。图 2 示出了根据本发明一个实施例的图片内容属性识别系统的结构示意图。参见图 2，该系统包括：

15 相对转载数计算模块 210，用于计算多个同源图片簇对于特定资源站点的相对转载数；

20 图片内容属性识别模块 220，与相对转载数计算模块 210 耦合，用于根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

图 1 所示流程指出，根据相对转载数能够识别图片内容属性，但是如何实现这一功能有不同的手段或方式，现以不同的实施例对不同的实现方式进行详细阐述。

实施例一

如图 3 所示，本发明的一个实施例（实施例一）提供了一种图片内容属性识别方法，其包括：步骤 310，计算多个同源图片簇对于特定资源站点的相对转载数，每个图片簇是对一组图片的聚合，例如，可以是相似度较高的一组图片，而相对转载数是一种能够反映同源图片簇的图片在特定资源站点站内站外的转载比例的数据，相对转载数的计算方式较多，本实施例中不对相对转载数的计算方式进行限制；步骤 320，根据多个同源图片簇以及对应的相对转载数训练筛选器模型，通过对广告图片的研究发现，广告图片有以下特点：广告图片生产成本低，很多广告图片都是商户花费金钱、花费时间制作的，因为广告图片的生产成本低，所以商户会将一张广告图片传播很多次，但是这些广告图片基本上只有商户会进行传播，而其他的用户则基本不会传播广告图片，广告图片在传播上的这种差别最终会体现在资源站点上的转载数上：在特定的资源站点上转载的次数非常多（商户故意传播），而在互联网其他站点上的转载的次数相对少的多（其他用户并不传播），也即广告图片在特定资源站点站内站外的转载比例会比较高，所以相对转载数可以作为区分广告图片和非广告图片的一种数据，而训练筛选器模型的工具包括但不

25

30

35

限于开源的 LIBSVM；步骤 330，根据训练后的筛选器模型识别目标图片簇中的图片内容属性，即识别目标图片簇中的图片是否为广告图片，有利于对广告图片进行过滤等处理，避免广告图片对用户的体验造成影响，假设目标图片簇为对应图片搜索请求的一组图片，则根据本实施例的技术方案，可以从其中识别出广告图片并进行过滤，从而将非广告图片作为搜索结果提供给用户，从而保证用户的使用体验。

在实际应用中，在本发明提出的相对转载数之外，还同时考虑到其他的特征，例如图片的长/宽，图片的大小，图片的清晰度，图片链接是否和网页同站，或图片跳转链接是否站外等特征，在训练筛选器时会根据多个同源图片簇各自对应的相对转载数，以及图片簇中的图片的长/宽，图片的大小，图片的清晰度，图片链接是否和网页同站，图片跳转链接是否站外中的一个或多个组合，先经过筛选器去学习和训练。在目标图片簇识别时，也会对应参照上述这些其他特征中的一个或多个来进行筛选并识别是否为广告图片。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 310 可以包括：对于多个同源图片簇中的一个同源图片簇，将同源图片簇中的图片在特定资源站点上的转载数，例如在图片站 A 上转载了 30 次，与在多个资源站点上的转载数相比较，例如在 10 个图片站（包括图片站 A）上共转载了 35 次，得到同源图片簇对于特定资源站点的相对转载数，多个资源站点包括特定资源站点，本实施例中提供了计算相对转载数的可行方式，且不对具体的比较方式进行限定，例如，取 $30/35$ 、 $30/(35-30)$ 作为相对转载数都是可以的。

如图 4 所示，本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 310 包括：步骤 311，计算特定资源站点上的图片的第一平均转载数，例如假设图片站 A 的第一平均转载数为 5；步骤 312，计算多个资源站点上的图片的第二平均转载数，例如假设 10 个图片站（包括图片站 A）的第二平均转载数为 20；步骤 313，取同源图片簇中的图片在特定资源站点上的转载数与第一平均转载数的第一差值，则第一差值实际上可反映同源图片簇的图片与其他图片在特定资源站点上的转载差异，差值越大则表示同源图片簇为广告图片的可能性越大，结合前述的实施例可知第一差值为 $30-5=25$ ，以及取同源图片簇中的图片在多个资源站点上的转载数与第二平均转载数的第二差值，则第二差值实际上可反映同源图片簇的图片与其他图片在多个资源站点上的转载差异，差值越大表示同源图片簇为广告图片的可能性越小，结合前述的实施例可知第二差值为 $35-20=15$ ，将第一差值和第二差值对比得到同源图片簇对于特定资源站点的相对转载数，本实施例中提供了另一种计算相对转载数的方式，且考虑到同源图片簇的图片与其他图片的转载差异，使得相对转载数能更好地反映图片是否为广告图片，本实施例中不对第一差值和第二差值对比方式进行限定，例如，取 $25/15$ ， $(25 \pm a)/(15 \pm b)$ 都是可以的， a 、 b 为常数。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 311 包括：取多个同源图片簇的图片中位于特定资源

站点上的多个图片，将多个图片的数量与多个图片对应的同源图片簇的数量进行对比，得到第一平均转载数，例如图片站 A 上有 100 张图片，该 100 张图片位于 20 个图片簇中，则第一平均转载数为 $100/20=5$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

5 本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 312 包括：将多个同源图片簇的图片的数量，与多个同源图片簇的数量进行比较，得到第二平均转载数，例如 10 个图片站（包括图片站 A）上有 1000 张图片，该 1000 张图片可聚类为 50 个图片簇，则第二平均转载数为 $1000/50=20$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

10 如图 5 所示，本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 310 之前，还包括：步骤 301，抓取多个资源站点上出现的图片链接（URL）；步骤 302，检测图片链接与同源图片簇的图片对应的链接是否相同，这反映了一张图片是否以不同的 URL 被转载，和/或检测图片链接对应的图片的校验信息与同源图片簇的图片的校验信息（包括但不限于 MD5 值）是否相同，
15 这反映了是否存在多张相同的图片，和/或检测图片链接对应的图片与同源图片簇的图片是否存在一个或多个相同的图像特征，这反映了多张图片是否相同，或由同一张图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等；步骤 303，根据检测结果，确定图片链接是否为同源图片簇的图片的转载，并统计同源图片簇的图片的转载数，则本实施例中提供了一种可全面统计图片转载数的技术方案。

20 本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，特定资源站点为多个同源图片簇中转载每个同源图片簇的图片最多的资源站点，转载图片最多次数的站点很可能为广告图片的商户进行传播的站点，该站点对应的转载数最能够有效地反映出图片是否为广告图片。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施
25 例的图片内容属性识别方法，每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征，则在本实施例的技术方案中，每个同源图片簇的图片相同，或可以同一图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等。

如图 6 所示，本发明的一个实施例提供了一种图片内容属性识别系统，用于支持实
30 施例一的图片内容属性识别方法，除包括图 2 中的各模块还包括其他模块。具体地，其包括：相对转载数计算模块 210，用于计算多个同源图片簇对于特定资源站点的相对转载数，每个图片簇是对一组图片的聚合，例如，可以是相似度较高的一组图片，而相对转载数是一种能够反映同源图片簇的图片在特定资源站点站内站外的转载比例的数据，相对转载数的计算方式较多，本实施例中不对相对转载数的计算方式进行限制；训练模块
35 230，用于将多个同源图片簇以及对应的相对转载数输入筛选器中训练筛选器模型。通过对广告图片的研究发现，广告图片有以下特点：广告图片生产成本高，很多广告图片都

是商户花费金钱、花费时间制作的，因为广告图片的生产成本高，所以商户会将一张广告图片传播很多次，但是这些广告图片基本上只有商户会进行传播，而其他用户则基本不会传播广告图片，广告图片在传播上的这种差别最终会体现在资源站点上的转载数上：在特定的资源站点上转载的次数非常多（商户故意传播），而在互联网其他站点上的转载的次数相对少的多（其他用户并不传播），也即广告图片在特定资源站点站内站外的转载比例会比较高，所以相对转载数可以作为区分广告图片和非广告图片的一种数据；筛选器 240，适于根据训练模块得到训练后的筛选器模型，并根据模型对目标图片簇进行筛选，本实施例中使用的筛选器包括但不限于开源的 LIBSVM；图片内容属性识别模块 220，用于根据筛选器 240 对目标图片簇进行筛选，识别目标图片簇中的图片内容属性，即识别目标图片簇中的图片是否为广告图片。

另外，实际应用中所述系统进一步包括：图片格式特征模块和/或图片链接特征模块；所述图片格式特征模块，适于提取同源图片簇以及目标图片簇中包含的图片的格式特征；所述图片链接特征模块，适于提取同源图片簇以及目标图片簇中包含的图片的链接特征；所述训练模块 230 进一步适于基于多个同源图片簇、对应的相对转载数以及对应的图片格式特征和/或图片链接特征，一同输入筛选器中训练筛选器模型；所述筛选器 240，进一步适于根据训练后的模型，结合目标图片簇对应的相对转载数以及对应的图片格式特征和/或图片链接特征，对目标图片簇进行筛选；所述图片内容属性识别模块 220，进一步用于根据所述筛选器基于目标图片簇对应的相对转载数以及对应的图片格式特征和/或图片链接特征对目标图片簇进行筛选，识别目标图片簇中的图片内容属性。

有利于对广告图片进行过滤等处理，避免广告图片对用户的体验造成影响，假设目标图片簇为对应图片搜索请求的一组图片，则根据本实施例的技术方案，可以从其中识别出广告图片并进行过滤，从而将非广告图片作为搜索结果提供给用户，从而保证用户的使用体验。

在实际应用中，在本发明提出的相对转载数之外，还考虑到其他的特征，例如图片的长/宽，图片的大小，图片的清晰度，图片链接是否和网页同站，或图片跳转链接是否站外等特征，同样先经过分类器去学习和训练。在目标图片簇识别时，也会考虑上述这些其他特征中的一个或多个来进行筛选并识别是否为广告图片。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，相对转载数计算模块 210 对于多个同源图片簇中的一个同源图片簇，将同源图片簇中的图片在特定资源站点上的转载数，例如在图片站 A 上转载了 30 次，与在多个资源站点上的转载数相比较，例如在 10 个图片站（包括图片站 A）上共转载了 35 次，得到同源图片簇对于特定资源站点的相对转载数，多个资源站点包括特定资源站点，本实施例中提供了计算相对转载数的可行方式，且不对具体的比较方式进行限定，例如，取 $30/35$ 、 $30/(35-30)$ 作为相对转载数都是可以的。

如图 7 所示，本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，还包括：第一平均转载数计算模块 250，用于

计算特定资源站点上的图片的第一平均转载数，例如假设图片站 A 的第一平均转载数为 5；第二平均转载数计算模块 260，用于计算多个资源站点上的图片的第二平均转载数，例如假设 10 个图片站（包括图片站 A）的第二平均转载数为 20；相对转载数计算模块 210 取同源图片簇中的图片在特定资源站点上的转载数与第一平均转载数的第一差值，则第一差值实际上可反映同源图片簇的图片与其他图片在特定资源站点上的转载差异，差值越大则表示同源图片簇为广告图片的可能性越大，结合前述的实施例可知第一差值为 $30-5=25$ ，以及取同源图片簇中的图片在多个资源站点上的转载数与第二平均转载数的第二差值，则第二差值实际上可反映同源图片簇的图片与其他图片在多个资源站点上的转载差异，差值越大表示同源图片簇为广告图片的可能性越小，结合前述的实施例可知第二差值为 $35-20=15$ ，将第一差值和第二差值对比得到同源图片簇对于特定资源站点的相对转载数，本实施例中提供了另一种计算相对转载数的方式，且考虑到同源图片簇的图片与其他图片的转载差异，使得相对转载数能更好地反映图片是否为广告图片，本实施例中不对第一差值和第二差值对比方式进行限定，例如，取 $25/15$ ， $(25 \pm a) / (15 \pm b)$ 都是可以的，a、b 为常数。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，第一平均转载数计算模块 250 取多个同源图片簇的图片中位于特定资源站点上的多个图片，将多个图片的数量与多个图片对应的同源图片簇的数量进行对比，得到第一平均转载数，例如图片站 A 上有 100 张图片，该 100 张图片位于 20 个图片簇中，则第一平均转载数为 $100/20=5$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，第二平均转载数计算模块 260 将多个同源图片簇的图片的数量，与多个同源图片簇的数量进行比较，得到第二平均转载数，例如 10 个图片站（包括图片站 A）上有 1000 张图片，该 1000 张图片可聚类为 50 个图片簇，则第二平均转载数为 $1000/50=20$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

如图 8 所示，本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，还包括：图片链接抓取模块 270，用于抓取多个资源站点上出现的图片链接（URL）；图片链接检测模块 280，用于检测图片链接与同源图片簇的图片对应的链接是否相同，这反映了一张图片是否以不同的 URL 被转载，和/或检测图片链接对应的图片的校验信息与同源图片簇的图片的校验信息（包括但不限于 MD5 值）是否相同，这反映了是否存在多张相同的图片，和/或检测图片链接对应的图片与同源图片簇的图片是否存在一个或多个相同的图像特征，这反映了多张图片是否相同，或由同一张图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等；图片转载数统计模块 290，用于根据检测结果，确定图片链接是否为同源图片簇的图片的转载，并统计同源图片簇的图片的转载数，则本实施例中提供了一种可全面统计图片转载数的技术方案。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，特定资源站点为多个同源图片簇中转载每个同源图片簇的图片最多的资源站点，转载图片最多次数的站点很可能为广告图片的商户进行传播的站点，该站点对应的转载数最能够有效地反映出图片是否为广告图片。

5 本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征，则在本实施例的技术方案中，每个同源图片簇的图片相同，或可以同一图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等。

10 实施例二

如图9所示，本发明的一个实施例中提供了另一种图片内容属性识别方法，其包括：步骤910，对收集到的图片进行相似图片识别，将图片聚合为多个同源图片簇，本实施例中将相似的图片聚合到同一同源图片簇中，对于一个同源图片簇而言，如果其中一张图片为广告图片，则其余图片也必然为广告图片，所以本实施例中以图片簇为单位进行图
15 片内容属性的识别，以判断每个同源图片簇中的图片是否为广告图片，基于目前的图像识别技术可以识别相似图片，本实施例不对相似图片的识别技术进行限定；步骤920，计算多个同源图片簇对于特定资源站点的相对转载数，相对转载数是一种能够反映同源图片簇的图片在特定资源站点站内站外的转载比例的数据，相对转载数的计算方式较多，本实施例中不对相对转载数的计算方式进行限制；步骤930，根据相对转载数识别对应的
20 同源图片簇中的图片内容属性，通过对广告图片的研究发现，广告图片有以下特点：广告图片生产成本低，很多广告图片都是商户花费金钱、花费时间制作的，因为广告图片的生产成本低，所以商户会将一张广告图片传播很多次，但是这些广告图片基本上只有商户会进行传播，而其他的用户则基本不会传播广告图片，广告图片在传播上的这种差别最终会体现在资源站点上的转载数上：在特定的资源站点上转载的次数非常多（商户故意传播），而在互联网其他站点上的转载的次数相对少的多（其他用户并不传播），也即广告图片在特定资源站点站内站外的转载比例会比较高，所以相对转载数可以作为区分广告图片和非广告图片的一种数据，因此本实施例的技术方案能够识别同源图片簇中的图片是否为广告图片，有利于对广告图片进行过滤等处理，避免广告图片对用户的
25 体验造成影响，假设同源图片簇为对应图片搜索请求的一组图片，则根据本实施例的技术方案，可以从其中识别出广告图片并进行过滤，从而将非广告图片作为搜索结果提供给用户，从而保证用户的使用体验。

在实际应用中，在本发明提出的相对转载数之外，还同时考虑到其他的特征，例如图片的长/宽，图片的大小，图片的清晰度，图片链接是否和网页同站，或图片跳转链接是否站外等特征。在同源图片簇识别时，也会考虑上述这些其他特征中的一个或多个来
35 进行筛选并识别是否为广告图片。在实际应用的另一实施例中，也可以先经过SVM模型去学习训练，将相对转载数以及上述其他特征中的一个或多个的组合作为参数对分类

器进行训练，并且在最后识别时也使用训练后的 SVM 模型及对应特征作为参数去识别。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 920 可以包括：对于多个同源图片簇中的一个同源图片簇，将同源图片簇中的图片在特定资源站点上的转载数，例如在图片站 A 上转载了 30 次，与在多个资源站点上的转载数相比较，例如在 10 个图片站（包括图片站 A）上共转载了 35 次，得到同源图片簇对于特定资源站点的相对转载数，多个资源站点包括特定资源站点，本实施例中提供了计算相对转载数的可行方式，且不对具体的比较方式进行限定，例如，取 $30/35$ 、 $30/(35-30)$ 作为相对转载数都是可以的。

如图 10 所示，本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 920 包括：步骤 921，计算特定资源站点上的图片的第一平均转载数，例如假设图片站 A 的第一平均转载数为 5；步骤 922，计算多个资源站点上的图片的第二平均转载数，例如假设 10 个图片站（包括图片站 A）的第二平均转载数为 20；步骤 923，取同源图片簇中的图片在特定资源站点上的转载数与第一平均转载数的第一差值，则第一差值实际上可反映同源图片簇的图片与其他图片在特定资源站点上的转载差异，差值越大则表示同源图片簇为广告图片的可能性越大，结合前述的实施例可知第一差值为 $30-5=25$ ，以及取同源图片簇中的图片在多个资源站点上的转载数与第二平均转载数的第二差值，则第二差值实际上可反映同源图片簇的图片与其他图片在多个资源站点上的转载差异，差值越大表示同源图片簇为广告图片的可能性越小，结合前述的实施例可知第二差值为 $35-20=15$ ，将第一差值和第二差值对比得到同源图片簇对于特定资源站点的相对转载数，本实施例中提供了另一种计算相对转载数的方式，且考虑到同源图片簇的图片与其他图片的转载差异，使得相对转载数能更好地反映图片是否为广告图片，本实施例中不对第一差值和第二差值对比方式进行限定，例如，取 $25/15$ ， $(25 \pm a)/(15 \pm b)$ 都是可以的，a、b 为常数。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 921 包括：取多个同源图片簇的图片中位于特定资源站点上的多个图片，将多个图片的数量与多个图片对应的同源图片簇的数量进行对比，得到第一平均转载数，例如图片站 A 上有 100 张图片，该 100 张图片位于 20 个图片簇中，则第一平均转载数为 $100/20=5$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 922 包括：将多个同源图片簇的图片的数量，与多个同源图片簇的数量进行比较，得到第二平均转载数，例如 10 个图片站（包括图片站 A）上有 1000 张图片，该 1000 张图片可聚类为 50 个图片簇，则第二平均转载数为 $1000/50=20$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

如图 11 所示，本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，步骤 920 之前，还包括：步骤 911，抓取多个

资源站点上出现的 URL；步骤 912，检测图片链接与同源图片簇的图片对应的链接是否相同，这反映了一张图片是否以不同的 URL 被转载，和/或检测图片链接对应的图片的校验信息与同源图片簇的图片的校验信息（包括但不限于 MD5 值）是否相同，这反映了是否存在多张相同的图片，和/或检测图片链接对应的图片与同源图片簇的图片是否存在一个或多个相同的图像特征，这反映了多张图片是否相同，或由同一张图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等；步骤 913，根据检测结果，确定图片链接是否为同源图片簇的图片的转载，并统计同源图片簇的图片的转载数，则本实施例中提供了一种可全面统计图片转载数的技术方案，其中步骤 911 与步骤 910 的顺序不限。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，特定资源站点为多个同源图片簇中转载每个同源图片簇的图片最多的资源站点，转载图片最多次数的站点很可能为广告图片的商户进行传播的站点，该站点对应的转载数最能够有效地反映出图片是否为广告图片。

本发明的另一实施例提出一种图片内容属性识别方法，与上述实施例相比，本实施例的图片内容属性识别方法，每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征，则在本实施例的技术方案中，每个同源图片簇的图片相同，或可以同一图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等。

如图 12 所示，本发明的一个实施例中提供了一种图片内容属性识别系统，用于支持实施例二提供的图片内容属性识别方法，除包括图 2 中的各模块还包括其他模块。具体地，其包括：图片聚合模块 1210，用于对收集到的图片进行相似图片识别，将图片聚合为多个同源图片簇，本实施例中将相似的图片聚合到同一同源图片簇中，对于一个同源图片簇而言，如果其中一张图片为广告图片，则其余图片也必然为广告图片，所以本实施例中以图片簇为单位进行图片内容属性的识别，以判断每个同源图片簇中的图片是否为广告图片，基于目前的图像识别技术可以识别相似图片，本实施例不对相似图片的识别技术进行限定；相对转载数计算模块 210，用于计算多个同源图片簇对于特定资源站点的相对转载数，相对转载数是一种能够反映同源图片簇的图片在特定资源站点站内站外的转载比例的数据，相对转载数的计算方式较多，本实施例中不对相对转载数的计算方式进行限制；图片内容属性识别模块 220，用于根据相对转载数识别对应的同源图片簇中的图片内容属性。通过对广告图片的研究发现，广告图片有以下特点：广告图片生产成本低，很多广告图片都是商户花费金钱、花费时间制作的，因为广告图片的生产成本高，所以商户会将一张广告图片传播很多次，但是这些广告图片基本上只有商户会进行传播，而其他的用户则基本不会传播广告图片，广告图片在传播上的这种差别最终会体现在资源站点上的转载数上：在特定的资源站点上转载的次数非常多（商户故意传播），而在互联网其他站点上的转载的次数相对少的多（其他用户并不传播），也即广告图片在特定资源站点站内站外的转载比例会比较高，所以相对转载数可以作为区分广告图片和非

广告图片的一种数据，因此本实施例的技术方案能够识别同源图片簇中的图片是否为广告图片，有利于对广告图片进行过滤等处理，避免广告图片对用户的体验造成影响，假设同源图片簇为对应图片搜索请求的一组图片，则根据本实施例的技术方案，可以从其中识别出广告图片并进行过滤，从而将非广告图片作为搜索结果提供给用户，从而保证用户的使用体验。

另外，实际应用中所述系统进一步包括：图片格式特征模块和/或图片链接特征模块；所述图片格式特征模块，适于提取同源图片簇中包含的图片的格式特征；所述图片链接特征模块，适于提取同源图片簇中包含的图片的链接特征；图片内容属性识别模块 220，进一步适于根据同源图片簇对应的相对转载数以及对应的图片格式特征和/或图片链接特征对同源图片簇进行筛选，识别同源图片簇中的图片内容属性。

在实际应用中，在本发明提出的相对转载数之外，还考虑到其他的特征，例如图片的长/宽，图片的大小，图片的清晰度，图片链接是否和网页同站，或图片跳转链接是否站外等特征。在同源图片簇识别时，也会考虑上述这些其他特征中的一个或多个来进行筛选并识别是否为广告图片。在实际应用的另一实施例中，也可以先经过 SVM 模型去学习和训练，将相对转载数以及上述其他特征中的一个或多个的组合作为参数对分类器进行训练，并且在最后识别时也使用训练后的 SVM 模型去识别。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，相对转载数计算模块 210 对于多个同源图片簇中的一个同源图片簇，将同源图片簇中的图片在特定资源站点上的转载数，例如在图片站 A 上转载了 30 次，与在多个资源站点上的转载数相比较，例如在 10 个图片站（包括图片站 A）上共转载了 35 次，得到同源图片簇对于特定资源站点的相对转载数，多个资源站点包括特定资源站点，本实施例中提供了计算相对转载数的可行方式，且不对具体的比较方式进行限定，例如，取 $30/35$ 、 $30/(35-30)$ 作为相对转载数都是可以的。

如图 13 所示，本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，还包括：第一平均转载数计算模块 1240，用于计算特定资源站点上的图片的第一平均转载数，例如假设图片站 A 的第一平均转载数为 5；第二平均转载数计算模块 1250，用于计算多个资源站点上的图片的第二平均转载数，例如假设 10 个图片站（包括图片站 A）的第二平均转载数为 20；相对转载数计算模块 210 取同源图片簇中的图片在特定资源站点上的转载数与第一平均转载数的第一差值，则第一差值实际上可反映同源图片簇的图片与其他图片在特定资源站点上的转载差异，差值越大则表示同源图片簇为广告图片的可能性越大，结合前述的实施例可知第一差值为 $30-5=25$ ，以及取同源图片簇中的图片在多个资源站点上的转载数与第二平均转载数的第二差值，则第二差值实际上可反映同源图片簇的图片与其他图片在多个资源站点上的转载差异，差值越大表示同源图片簇为广告图片的可能性越小，结合前述的实施例可知第二差值为 $35-20=15$ ，将第一差值和第二差值对比得到同源图片簇对于特定资源站点的相对转载数，本实施例中提供了另一种计算相对转载数的方式，且考虑到同源图片簇的

图片与其他图片的转载差异，使得相对转载数能更好地反映图片是否为广告图片，本实施例中不对第一差值和第二差值对比方式进行限定，例如，取 $25/15$ ， $(25 \pm a)/(15 \pm b)$ 都是可以的， a 、 b 为常数。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，第一平均转载数计算模块 240 取多个同源图片簇的图片中位于特定资源站点上的多个图片，将多个图片的数量与多个图片对应的同源图片簇的数量进行对比，得到第一平均转载数，例如图片站 A 上有 100 张图片，该 100 张图片位于 20 个图片簇中，则第一平均转载数为 $100/20=5$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，第二平均转载数计算模块 250 将多个同源图片簇的图片的数量，与多个同源图片簇的数量进行比较，得到第二平均转载数，例如 10 个图片站（包括图片站 A）上有 1000 张图片，该 1000 张图片可聚类为 50 个图片簇，则第二平均转载数为 $1000/50=20$ ，本实施例的技术方案中提供了一种快速高效得到平均转载数的方式。

如图 14 所示，本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，还包括：图片链接抓取模块 1260，用于抓取多个资源站点上出现的 URL；图片链接检测模块 1270，用于检测图片链接与同源图片簇的图片对应的链接是否相同，这反映了一张图片是否以不同的 URL 被转载，和/或检测图片链接对应的图片的校验信息与同源图片簇的图片的校验信息（包括但不限于 MD5 值）是否相同，这反映了是否存在多张相同的图片，和/或检测图片链接对应的图片与同源图片簇的图片是否存在一个或多个相同的图像特征，这反映了多张图片是否相同，或由同一张图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等；图片转载数统计模块 1280，用于根据检测结果，确定图片链接是否为同源图片簇的图片的转载，并统计同源图片簇的图片的转载数，则本实施例中提供了一种可全面统计图片转载数的技术方案。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，特定资源站点为多个同源图片簇中转载每个同源图片簇的图片最多的资源站点，转载图片最多次数的站点很可能为广告图片的商户进行传播的站点，该站点对应的转载数最能够有效地反映出图片是否为广告图片。

本发明的另一实施例提出一种图片内容属性识别系统，与上述实施例相比，本实施例的图片内容属性识别系统，每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征，则在本实施例的技术方案中，每个同源图片簇的图片相同，或可以同一图片修改得到，本实施例中的图像特征包括但不限于轮廓特征、颜色特征、直方图特征等。

在此处所提供的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、

结构和技术，以便不模糊对本说明书的理解。

类似地，应当理解，为了精简本公开并帮助理解各个发明方面中的一个或多个，在上面对本发明的示例性实施例的描述中，本发明的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而，并不应将该公开的方法解释成反映如下意图：即所
5 要求保护的本发明要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说，如下面的权利要求书所反映的那样，发明方面在于少于前面公开的单个实施例的所有特征。因此，遵循具体实施方式的权利要求书由此明确地并入该具体实施方式，其中每个权利要求本身都作为本发明的单独实施例。

本领域那些技术人员可以理解，可以对实施例中的设备中的模块进行自适应性地改变并且把它们设置在与该实施例不同的一个或多个设备中。可以把实施例中的模块或单元或组件组合成一个模块或单元或组件，以及此外可以把它分成多个子模块或子单元或子组件。除了这样的特征和/或过程或者单元中的至少一些是相互排斥之外，可以采用
10 任何组合对本说明书（包括伴随的权利要求、摘要和附图）中公开的所有特征以及如此公开的任何方法或者设备的所有过程或单元进行组合。除非另外明确陈述，本说明书（包括伴随的权利要求、摘要和附图）中公开的每个特征可以由提供相同、等同或相似目的的替代特征来代替。

此外，本领域的技术人员能够理解，尽管在此所述的一些实施例包括其它实施例中
所包括的某些特征而不是其它特征，但是不同实施例的特征的组合意味着处于本发明的范围之内并且形成不同的实施例。例如，在下面的权利要求书中，所要求保护的实施例
20 的任意之一都可以以任意的组合方式来使用。

本发明的各个部件实施例可以以硬件实现，或者以在一个或者多个处理器上运行的软件模块实现，或者以它们的组合实现。本领域的技术人员应当理解，可以在实践中使用微处理器或者数字信号处理器（DSP）来实现根据本发明实施例的图片内容属性识别系统中的一些或者全部部件的一些或者全部功能。本发明还可以实现为用于执行这里所
25 描述的方法的一部分或者全部的设备或者装置程序（例如，计算机程序和计算机程序产品）。这样的实现本发明的程序可以存储在计算机可读介质上，或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到，或者在载体信号上提供，或者以任何其他形式提供。

例如，图 15 示出了可以实现根据本发明的图片内容属性识别方法的计算设备。该计算设备传统上包括处理器 1510 和以存储器 1520 形式的计算机程序产品或者计算机可读
30 介质。存储器 1520 可以是诸如闪存、EEPROM（电可擦除可编程只读存储器）、EPROM、硬盘或者 ROM 之类的电子存储器。存储器 1520 具有用于执行上述方法中的任何方法步骤的程序代码 1531 的存储空间 1530。例如，用于程序代码的存储空间 1530 可以包括分别用于实现上面的方法中的各种步骤的各个程序代码 1531。这些程序代码可以从一个或者
35 多个计算机程序产品中读出或者写入到这一个或者多个计算机程序产品中。这些计算机程序产品包括诸如硬盘，紧致盘（CD）、存储卡或者软盘之类的程序代码载体。这样

的计算机程序产品通常为如参考图 16 所述的便携式或者固定存储单元。该存储单元可以具有与图 15 的计算设备中的存储器 1520 类似布置的存储段、存储空间等。程序代码可以例如以适当形式进行压缩。通常，存储单元包括计算机可读代码 1531’，即可以由例如诸如 1510 之类的处理器读取的代码，这些代码当由计算设备运行时，导致该计算设备执行上面所描述的方法中的各个步骤。

本文中所称的“一个实施例”、“实施例”或者“一个或者多个实施例”意味着，结合实施例描述的特定特征、结构或者特性包括在本发明的至少一个实施例中。此外，请注意，这里“在一个实施例中”的词语例子不一定全指同一个实施例。

应该注意的是上述实施例对本发明进行说明而不是对本发明进行限制，并且本领域技术人员在不脱离所附权利要求的范围的情况下可设计出替换实施例。在权利要求中，不应将位于括号之间的任何参考符号构造成对权利要求的限制。单词“包含”不排除存在未列在权利要求中的元件或步骤。位于元件之前的单词“一”或“一个”不排除存在多个这样的元件。本发明可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。在列举了若干装置的单元权利要求中，这些装置中的若干个可以是通过同一个硬件项来具体体现。单词第一、第二、以及第三等的使用不表示任何顺序。可将这些单词解释为名称。

此外，还应当注意，本说明书中使用的语言主要是为了可读性和教导的目的而选择的，而不是为了解释或者限定本发明的主题而选择的。因此，在不偏离所附权利要求书的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。对于本发明的范围，对本发明所做的公开是说明性的，而非限制性的，本发明的范围由所附权利要求书限定。

权 利 要 求

1、一种图片内容属性识别方法，其包括：

计算多个同源图片簇对于特定资源站点的相对转载数；

5 根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

2、根据权利要求 1 所述的图片内容属性识别方法，其中，所述计算多个同源图片簇对于特定资源站点的相对转载数之前，还包括：

对收集到的图片进行相似图片识别，将图片聚合为多个同源图片簇。

3、根据权利要求 1 所述的图片内容属性识别方法，其中，根据所述相对转载数识别
10 对应的同源图片簇中的图片内容属性，包括：

根据所述多个同源图片簇以及对应的相对转载数训练筛选器模型；

根据训练后的筛选器模型识别目标图片簇中的图片内容属性。

4、根据权利要求 1 所述的图片内容属性识别方法，其中，所述计算多个同源图片簇对于特定资源站点的相对转载数的步骤包括：

15 对于所述多个同源图片簇中的一个同源图片簇，将所述同源图片簇中的图片在特定资源站点上的转载数，与在多个资源站点上的转载数相比较，得到所述同源图片簇对于所述特定资源站点的相对转载数，所述多个资源站点包括所述特定资源站点。

5、根据权利要求 4 所述的图片内容属性识别方法，其中，所述将所述同源图片簇中的图片在所述特定资源站点上的转载数，与在多个资源站点上的转载数相比较的步骤包
20 括：

计算所述特定资源站点上的图片的第一平均转载数；

计算所述多个资源站点上的图片的第二平均转载数；

取所述同源图片簇中的图片在所述特定资源站点上的转载数与所述第一平均转载数的第一差值，以及取所述同源图片簇中的图片在所述多个资源站点上的转载数与所述第二平均转载数的第二差值，将所述第一差值和所述第二差值对比得到所述同源图片簇对于所述特定资源站点的相对转载数。
25

6、根据权利要求 5 所述的图片内容属性识别方法，其中，所述计算所述特定资源站点上的图片的第一平均转载数的步骤包括：

取所述多个同源图片簇的图片中位于所述特定资源站点上的多个图片，将所述多个
30 图片的数量与所述多个图片对应的同源图片簇的数量进行对比，得到所述第一平均转载数。

7、根据权利要求 5 所述的图片内容属性识别方法，其中，所述计算所述多个资源站点上的图片的第二平均转载数的步骤包括：

将所述多个同源图片簇的图片的数量，与所述多个同源图片簇的数量进行比较，得
35 到所述第二平均转载数。

8、根据权利要求 4 所述的图片内容属性识别方法，其中，在所述将所述同源图片簇

中的图片在特定资源站点上的转载数，与在多个资源站点上的转载数相比较的步骤之前，还包括：

抓取所述多个资源站点上出现的图片链接；

5 检测所述图片链接与所述同源图片簇的图片对应的链接是否相同，和/或检测所述图片链接对应的图片的校验信息与所述同源图片簇的图片的校验信息是否相同，和/或检测所述图片链接对应的图片与所述同源图片簇的图片是否存在一个或多个相同的图像特征；

根据检测结果，确定所述图片链接是否为所述同源图片簇的图片的转载，并统计所述同源图片簇的图片的转载数。

10 9、根据权利要求4所述的图片内容属性识别方法，其中，

所述特定资源站点为所述多个同源图片簇中转载每个同源图片簇的图片最多的资源站点。

10、根据权利要求1至9中任一项所述的图片内容属性识别方法，其中，

15 每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征。

11、根据权利要求1至9任一项所述的方法，其中，所述方法进一步包括：提取所述同源图片簇中包含的图片的格式特征和/或图片的链接特征，根据所述多个同源图片簇、对应的相对转载数，以及对应包含的图片的格式特征训练筛选器模型；根据训练后的筛选器模型，基于所述相对转载数以及目标图片簇中包含的图片的格式特征和/或图片
20 的链接特征，来识别目标图片簇中的图片内容属性。

12、根据权利要求1至9任一项所述的方法，其中，所述图片的格式特征包括但不限于以下中的一种或几种组合：图片的长/宽，图片的大小，图片的清晰度。

13、根据权利要求1至9任一项所述的方法，其中，所述图片的链接特征包括但不限于以下中的一种或几种组合：图片链接是否和网页同站，图片跳转链接是否站外。

25 14、一种图片内容属性识别系统，其包括：

相对转载数计算模块，用于计算多个同源图片簇对于特定资源站点的相对转载数；

图片内容属性识别模块，用于根据所述相对转载数识别对应的同源图片簇中的图片内容属性。

15、根据权利要求14所述的图片内容属性识别系统，其中，还包括：

30 图片聚合模块，用于对收集到的图片进行相似图片识别，将图片聚合为多个同源图片簇。

16、根据权利要求14所述的图片内容属性识别系统，其中，还包括：

训练模块，用于将所述多个同源图片簇以及对应的相对转载数输入筛选器中训练筛选器模型；

35 筛选器，适于根据所述训练模块得到训练后的筛选器模型，并根据所述模型对目标图片簇进行筛选；

所述图片内容属性识别模块，还用于根据所述筛选器对目标图片簇进行筛选，识别目标图片簇中的图片内容属性。

17、根据权利要求 14 所述的图片内容属性识别系统，其中，

所述相对转载数计算模块对于所述多个同源图片簇中的一个同源图片簇，将所述同源图片簇中的图片在特定资源站点上的转载数，与在多个资源站点上的转载数相比较，得到所述同源图片簇对于所述特定资源站点的相对转载数，所述多个资源站点包括所述特定资源站点。

18、根据权利要求 17 所述的图片内容属性识别系统，其中，还包括：第一平均转载数计算模块，用于计算特定资源站点上的图片的第一平均转载数；第二平均转载数计算模块，用于计算多个资源站点上的图片的第二平均转载数；相对转载数计算模块取同源图片簇中的图片在特定资源站点上的转载数与第一平均转载数的第一差值，以及取同源图片簇中的图片在多个资源站点上的转载数与第二平均转载数的第二差值，将第一差值和第二差值对比得到同源图片簇对于特定资源站点的相对转载数。

19、根据权利要求 18 所述的图片内容属性识别系统，其中，第一平均转载数计算模块取多个同源图片簇的图片中位于特定资源站点上的多个图片，将多个图片的数量与多个图片对应的同源图片簇的数量进行对比，得到第一平均转载数。

20、根据权利要求 18 所述的图片内容属性识别系统，其中，第二平均转载数计算模块将多个同源图片簇的图片的数量，与多个同源图片簇的数量进行比较，得到第二平均转载数。

21、根据权利要求 14 至 20 任一项所述的图片内容属性识别系统，其中，还包括：图片链接抓取模块，用于抓取多个资源站点上出现的图片链接；图片链接检测模块，用于检测图片链接与同源图片簇的图片对应的链接是否相同，和/或检测图片链接对应的图片的校验信息与同源图片簇的图片的校验信息是否相同，和/或检测图片链接对应的图片与同源图片簇的图片是否存在一个或多个相同的图像特征；图片转载数统计模块，用于根据检测结果，确定图片链接是否为同源图片簇的图片的转载，并统计同源图片簇的图片的转载数。

22、根据权利要求 14 至 20 任一项所述的图片内容属性识别系统，其中，特定资源站点为多个同源图片簇中转载每个同源图片簇的图片最多的资源站点。

23、根据权利要求 14 至 20 任一项所述的图片内容属性识别系统，其中，每个同源图片簇的图片对应同一源图片，且每个同源图片簇的图片与其对应的源图片具有一个或多个相同的图像特征。

24、一种计算机程序，包括计算机可读代码，当所述计算机可读代码在计算设备上运行时，导致所述计算设备执行根据权利要求 1-13 中的任一个所述的图片内容属性识别方法。

25、一种计算机可读介质，其中存储了如权利要求 24 所述的计算机程序。

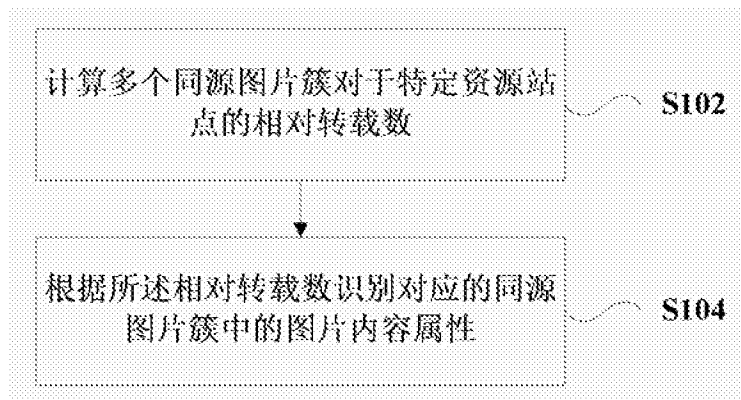


图 1

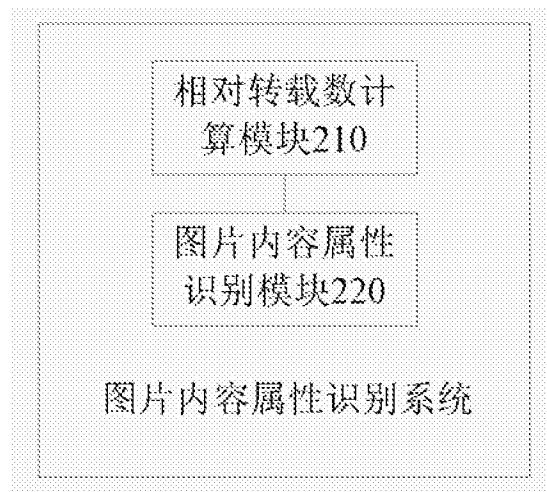


图 2

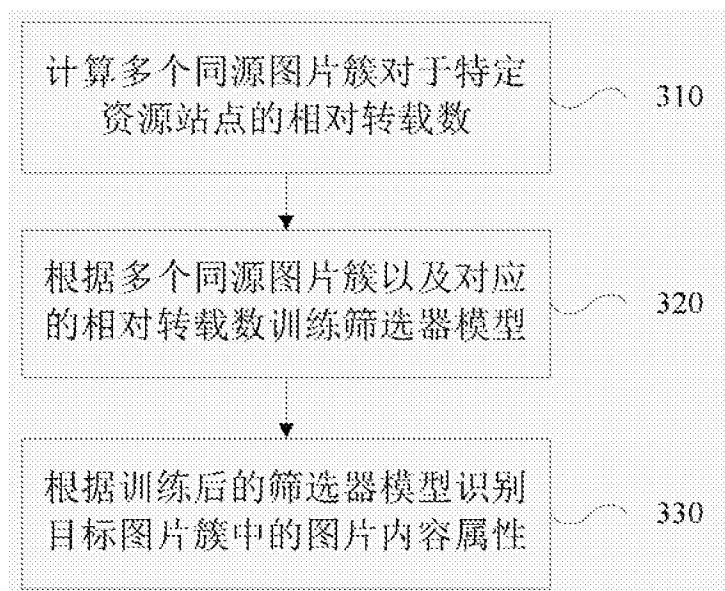


图 3

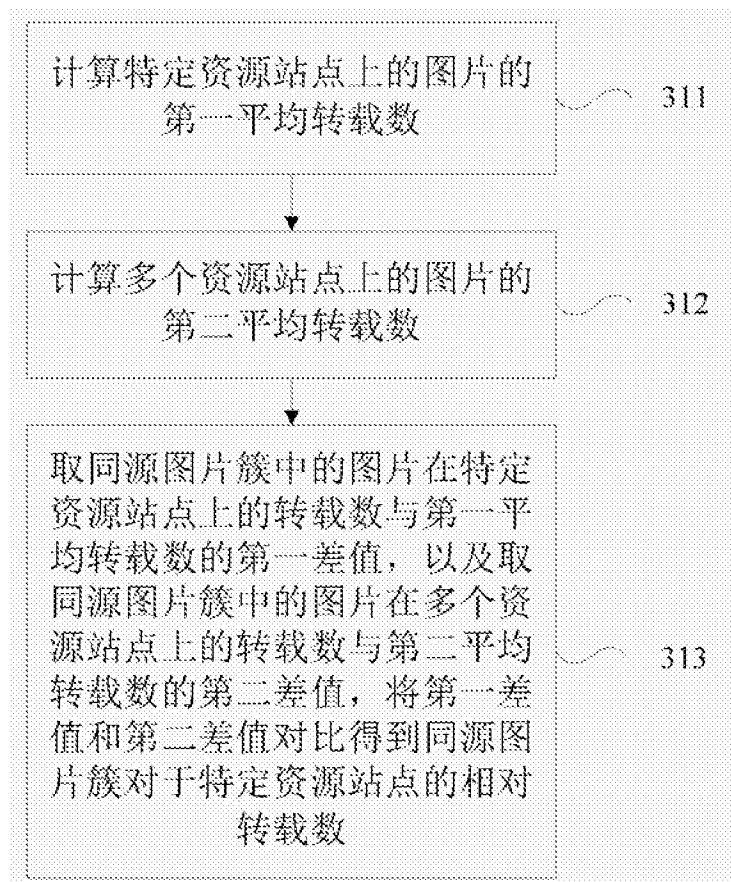


图 4

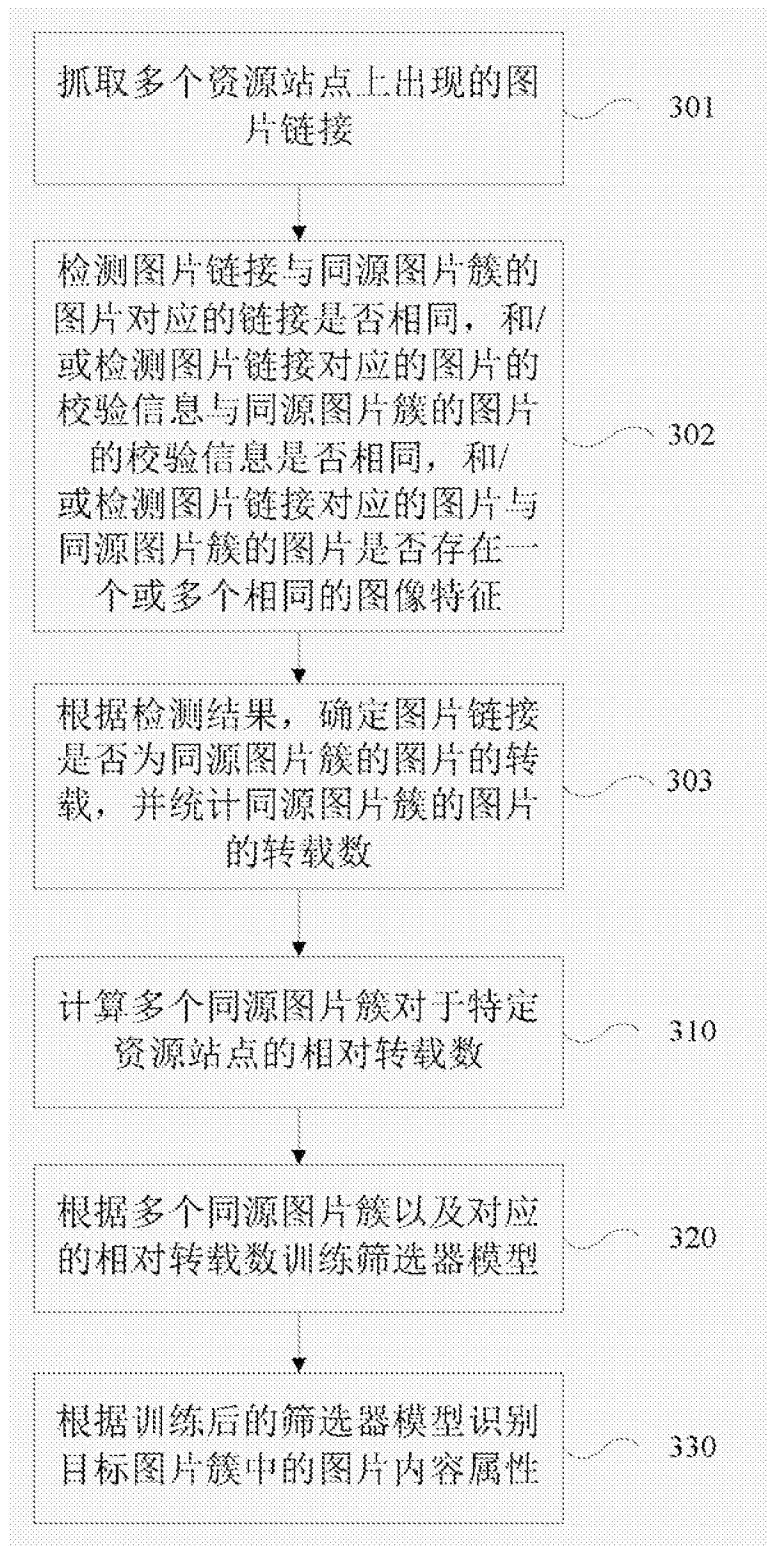


图 5



图 6



图 7



图 8

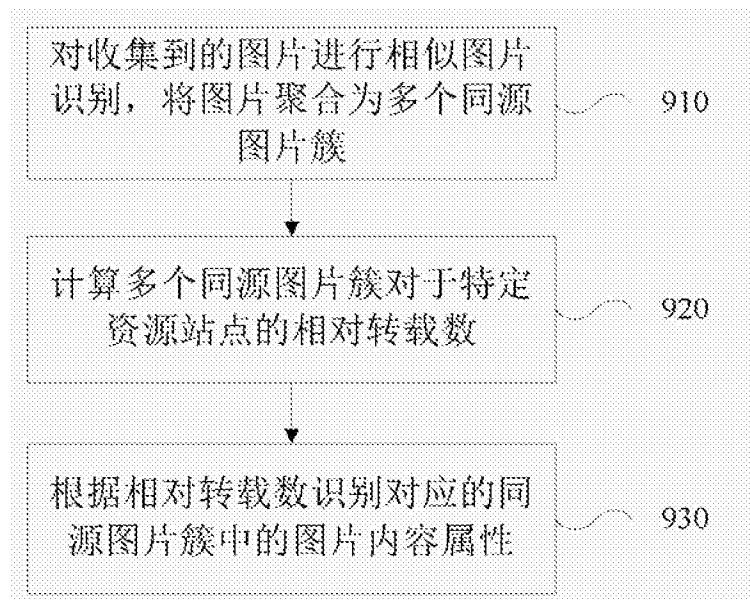


图 9

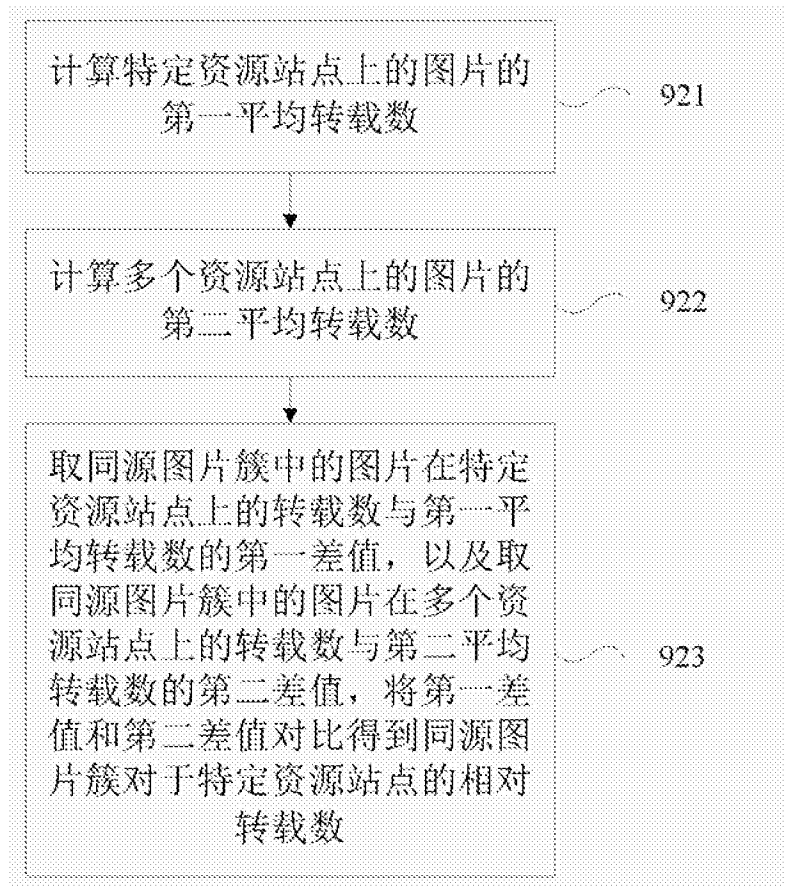


图 10

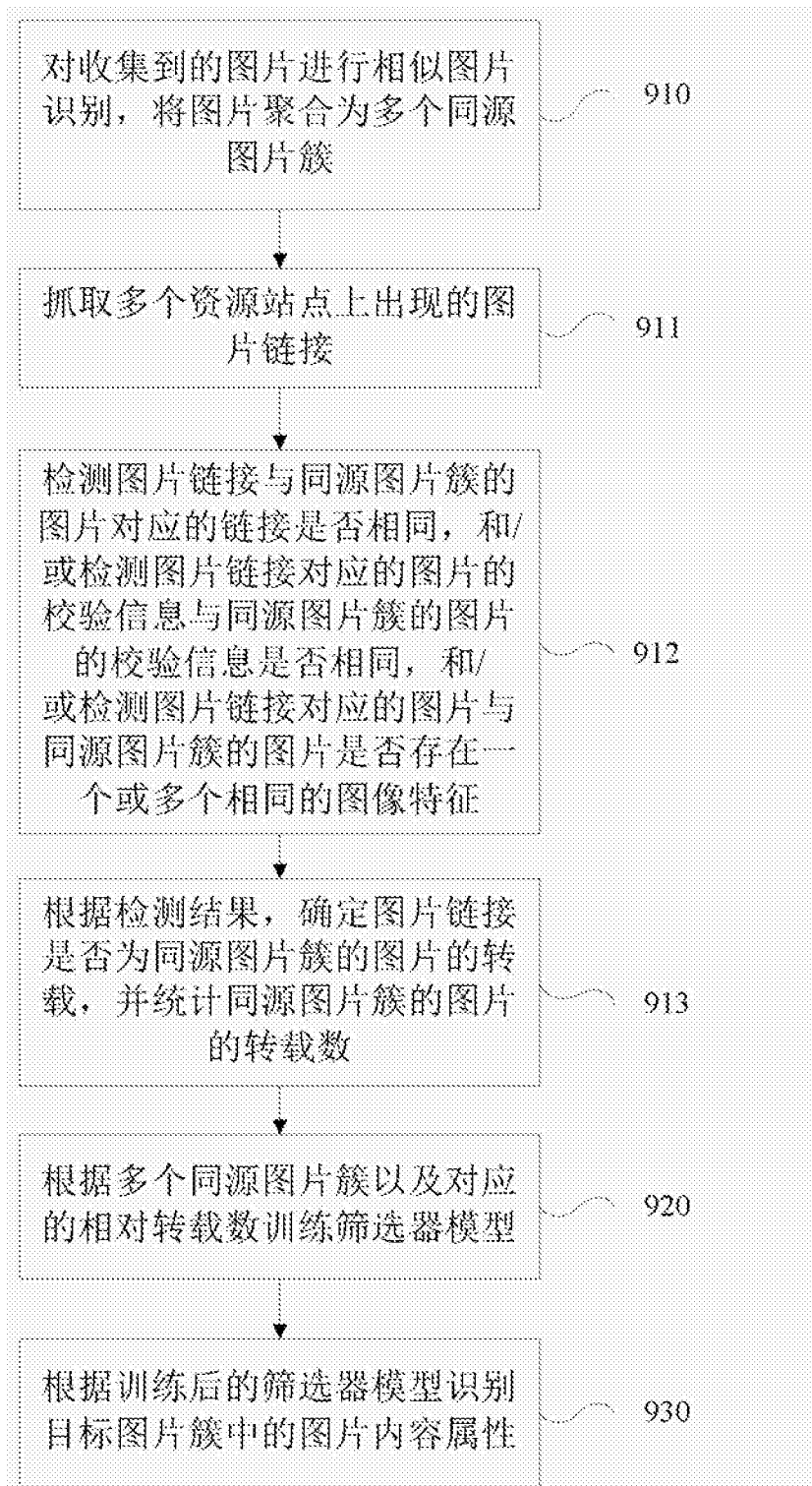


图 11



图 12



图 13

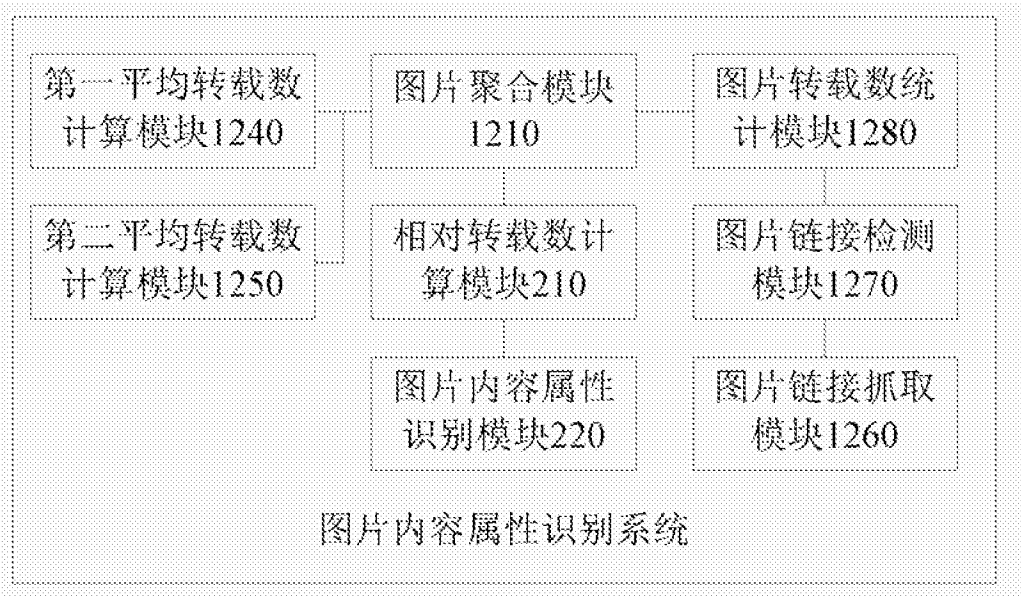


图 14

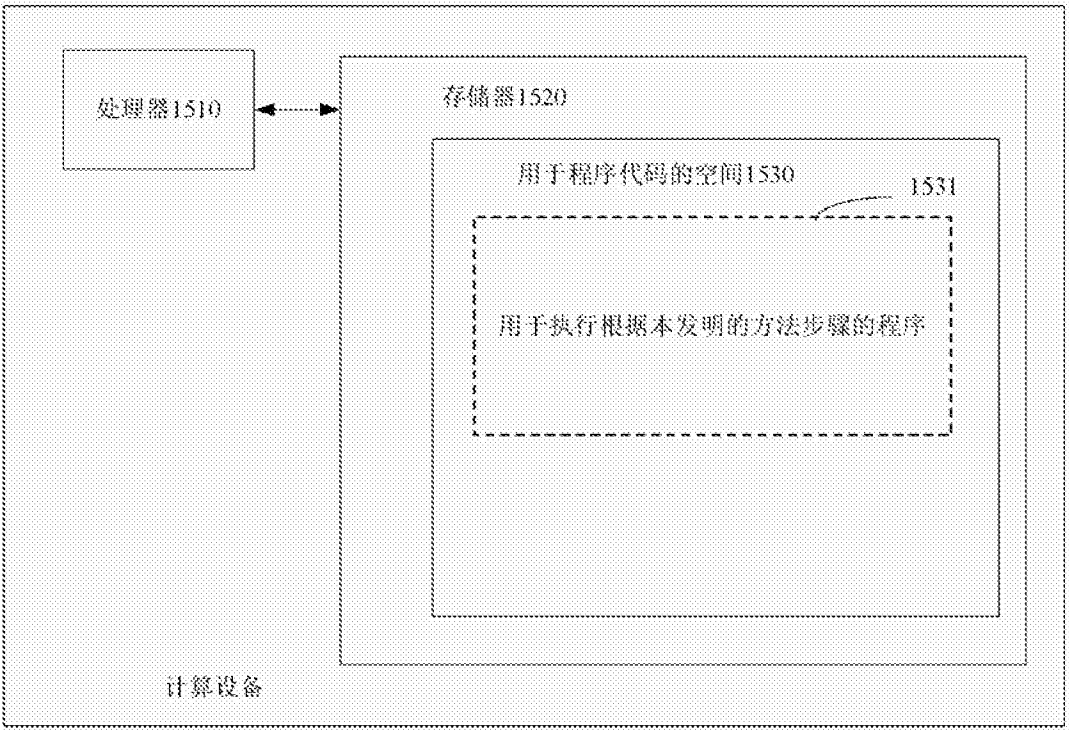


图 15

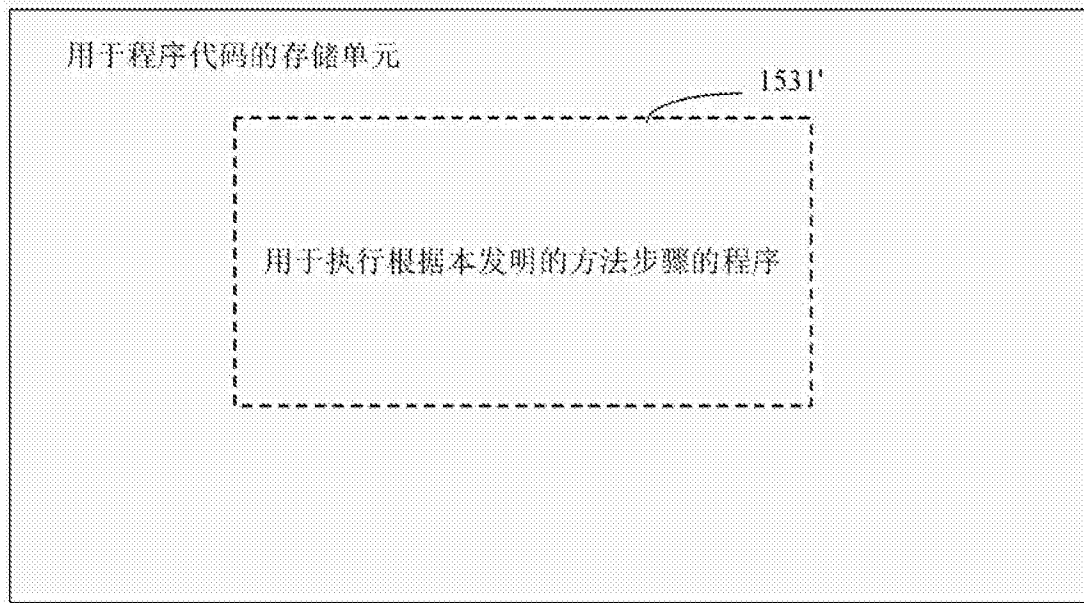


图 16

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2014/087109

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, IEEE: image?, picture?, photo+, advertisement, recogni+, distinguish, judge, transshipment, release, appear, amount, website, site, URL, forum

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 101071433 A (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) 14 November 2007 (14.11.2007) the abstract	1-3, 10, 12, 14-16, 23-25
A	CN 102419777 A (IFENG (BEIJING) INFORMATION TECHNOLOGY CO., LTD.) 18 April 2012 (18.04.2012) the whole document	1-25
PX	CN 103617261 A (QIHU 360 SOFTWARE CORPORATION LIMITED) 05 March 2014 (05.04.2014) claims 1 to 10	1-25
PX	CN 103617262 A (QIHU 360 SOFTWARE CORPORATION LIMITED) 05 March 2014 (05.03.2014) claims 1 to 10	1-25

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 04 December 2014	Date of mailing of the international search report 24 December 2014
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer REN, Rui Telephone No. (86-10) 62413079

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2014/087109

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 101071433 A	14 November 2007	CN 101071433 B	18 August 2010
CN 102419777 A	18 April 2012	CN 102419777 B	02 October 2013
CN 103617261 A	05 March 2014	None	
CN 103617262 A	05 March 2014	None	

A. 主题的分类 G06F17/30 (2006.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																	
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNPAT, CNKI, WPI, EPODOC, IEEE:图像, 图象, 图片, 相片, 照片, 广告, 识别, 分类, 鉴别, 判断, 转载, 转发, 发表, 发布, 出现, 数量, 个数, 次数, 张数, 网站, 站点, 网址, 论坛, image?, picture?, photo+, advertisement, recogni+, distinguish																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 101071433 A (腾讯科技深圳有限公司) 2007年 11月 14日 (2007 - 11 - 14) 摘要</td> <td>1-3, 10, 12, 14-16, 23-25</td> </tr> <tr> <td>A</td> <td>CN 102419777 A (凤凰在线北京信息技术有限公司) 2012年 4月 18日 (2012 - 04 - 18) 全文</td> <td>1-25</td> </tr> <tr> <td>PX</td> <td>CN 103617261 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10</td> <td>1-25</td> </tr> <tr> <td>PX</td> <td>CN 103617262 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10</td> <td>1-25</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 101071433 A (腾讯科技深圳有限公司) 2007年 11月 14日 (2007 - 11 - 14) 摘要	1-3, 10, 12, 14-16, 23-25	A	CN 102419777 A (凤凰在线北京信息技术有限公司) 2012年 4月 18日 (2012 - 04 - 18) 全文	1-25	PX	CN 103617261 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10	1-25	PX	CN 103617262 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10	1-25
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
X	CN 101071433 A (腾讯科技深圳有限公司) 2007年 11月 14日 (2007 - 11 - 14) 摘要	1-3, 10, 12, 14-16, 23-25															
A	CN 102419777 A (凤凰在线北京信息技术有限公司) 2012年 4月 18日 (2012 - 04 - 18) 全文	1-25															
PX	CN 103617261 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10	1-25															
PX	CN 103617262 A (北京奇虎科技有限公司等) 2014年 3月 05日 (2014 - 03 - 05) 权利要求1-10	1-25															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																	
国际检索实际完成的日期 2014年 12月 04日		国际检索报告邮寄日期 2014年 12月 24日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451		授权官员 任蕊 电话号码 (86-10)62413079															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2014/087109

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	101071433	A	2007年 11月 14日	CN 101071433 B	2010年 8月 18日
CN	102419777	A	2012年 4月 18日	CN 102419777 B	2013年 10月 02日
CN	103617261	A	2014年 3月 05日	无	
CN	103617262	A	2014年 3月 05日	无	