

(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 12 月 3 日 (03.12.2015)



(10) 国际公布号
WO 2015/180622 A1

- (51) 国际专利分类号:
G06F 17/30 (2006.01)
- (21) 国际申请号: PCT/CN2015/079800
- (22) 国际申请日: 2015 年 5 月 26 日 (26.05.2015)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201410225991.3 2014 年 5 月 26 日 (26.05.2014) 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: 刘臻 (LIU, Liu); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。苏晓东 (SU, Xiaodong); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。常富洋 (CHANG,

Fuyang); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。王安滨 (WANG, Anbin); 中国北京市朝阳区酒仙桥路 6 号院 2 号楼, Beijing 100015 (CN)。秦吉胜 (QIN, Jisheng); 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: 北京市隆安律师事务所 (BEIJING LONGAN LAW FIRM); 中国北京市朝阳区建国门外大街 21 号北京国际俱乐部 188 室, Beijing 100020 (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,

[见续页]

(54) Title: METHOD AND APPARATUS FOR DETERMINING CATEGORICAL ATTRIBUTE OF QUERIED WORD IN SEARCH

(54) 发明名称: 一种确定搜索查询词类别属性的方法和装置

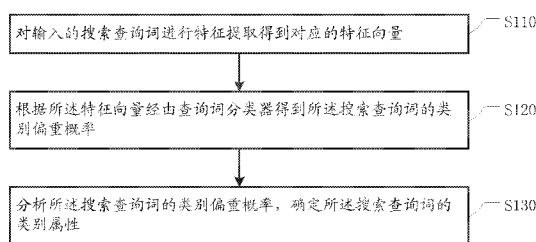


图 1 / FIG. 1

- S110 EXTRACTING A CHARACTERISTIC OF AN INPUT QUERIED WORD IN SEARCH TO OBTAIN A CORRESPONDING CHARACTERISTIC VECTOR
- S120 ACQUIRING A CATEGORICAL PREFERENCE PROBABILITY OF THE QUERIED WORD IN SEARCH BY USING A QUERIED WORD CLASSIFIER ACCORDING TO THE CHARACTERISTIC VECTOR, AND ANALYZING THE CATEGORICAL PREFERENCE PROBABILITY OF THE QUERIED WORD IN SEARCH
- S130 DETERMINING A CATEGORICAL ATTRIBUTE OF THE QUERIED WORD IN SEARCH

(57) Abstract: A method and an apparatus for determining a categorical attribute of a queried word in search. The method comprises: extracting a characteristic of an input queried word in search to obtain a corresponding characteristic vector (S110); acquiring a categorical preference probability of the queried word in search by using a queried word classifier according to the characteristic vector, and analyzing the categorical preference probability of the queried word in search (S120); and determining a categorical attribute of the queried word in search (S130). The categorical preference probability of the queried word in search is obtained by extracting the characteristic vector of the queried word in search and inputting the characteristic vector into the queried word classifier, and the categorical attribute of the queried word is determined by analyzing the categorical preference probability, so that a basis is provided for search, and the search accuracy is ensured, and basic characteristics can be provided for subsequent events such as search sequencing.

(57) 摘要: 一种确定搜索查询词类别属性的方法和装置。所述方法包括: 对输入的搜索查询词进行特征提取得到对应的特征向量 (S110), 根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率, 分析所述搜索查询词的类别偏重概率 (S120), 确定所述搜索查询词的类别属性

(S130)。通过提取搜索查询词的特征向量, 将该特征向量输入到查询词分类器得到所述搜索查询词的类别偏重概率, 并通过对类别偏重概率确定该查询词的类别属性, 从而为搜索提供基础依据, 保障搜索的准确性。并且能够为搜索排序等后续事件提供基础特征。



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, 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))。

一种确定搜索查询词类别属性的方法和装置

技术领域

本发明涉及计算机网络技术领域，具体涉及一种确定搜索查询词类别属性的方法和装置。

5 背景技术

在一次完整的搜索中，搜索引擎接收到用户输入的搜索查询词后通常会经过预处理搜索查询词、理解搜索查询词、检索文档、排序、展现等过程，整个过程需要在毫秒级时间内完成。而搜索查询词的分类对于理解查询词这一过程是十分重要的，它不仅能反映当前用户当前的兴趣意图，为本次检索提供依据，又能作为后续搜索引擎结果排序模型、广告 CTR 预估模型、自然语言模型的基础特征。

支持向量机 (SVM) 为机器学习领域中监督学习模型的一种，由 Vapnik 等人于 1995 年提出。最基本的 SVM 模型为“二元分类”模型，其学习方式为最大化间隔策略。对于简单的线性可分数据，通过“硬间隔最大化”函数学习出硬间隔支持向量机；对于近似线性可分的数据，通过“软间隔最大化”学习出软间隔支持向量机；对于完全线性不可分的数据，通过将数据映射到更高维空间，在高维空间学习出软间隔支持向量机，在这一过程中采用“核方法”可以隐式地将输入空间的内积映射到高维空间后再做内积，相当于在高维空间学习软间隔支持向量机。

Liblinear 是由台湾大学林智仁教授及其研究团队开发的线性 SVM 软件包，主要实现了线性多元分类与线性回归。Liblinear 考虑到大规模机器学习应用，它并没有引入“核方法”，而是假设数据线性或近似线性可分，直接训练线性分类器。经过多年发展，liblinear 在工业界被广泛的应用于大规模分类与回归问题的解决，其不仅在训练与预测的性能上远远优于 SVM，其准确率也达到令人满意的效果。从概率的角度，业界实际的文本处理项目中，通常采用布尔向量模型，特征数量往往少则几十万多则上亿，而获取到的训练数据仅仅覆盖特征空间中的一部分数据，因此其线性不可分概率便较小了。

可见现有的搜索查询词分类器的准确率仍有待提高。

发明内容

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搜索查询词分类方法和装置。

依据本发明的一个方面，提供了一种确定搜索查询词类别属性的方法，

5 其中，该方法包括：

对输入的搜索查询词进行特征提取得到对应的特征向量；

根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率；

10 分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性。

依据本发明的另一个方面，提供了一种确定搜索查询词类别属性的装置，该装置包括：

特征提取单元，适于对输入的搜索查询词进行特征提取得到对应的特征向量；

15 分类器，适于根据所述特征向量得到所述搜索查询词的类别偏重概率并发送给输出单元；

输出单元，适于分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性并输出。

20 依据本发明的一个方面，还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服务器上运行时，导致所述服务器执行根据上述任一个所述的确定搜索查询词类别属性的方法。

依据本发明的一个方面，还提供了一种计算机可读介质，其中存储了如上所述的计算机程序。

25 本发明这种对输入的搜索查询词进行特征提取得到对应的特征向量，根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率，分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性的技术方案，通过提取搜索查询词的特征向量，将该特征向量输入到查询词分类器得到所述搜索查询词的类别偏重概率，并通过分析类别偏重概率确定该查询词的类别属性，从而为搜索提供基础依据，保障搜索的准确性。并且能够为
30 搜索排序等后续事件提供基础特征。

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。

5

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
10 相同的参考符号表示相同的部件。在附图中：

图1示出了根据本发明一个实施例的一种确定搜索查询词类别属性的方法的流程图；

图2示出了根据本发明一个实施例的确定分类器的分类模型以及利用分类器确定搜索查询词的类别属性的方法的流程图；

15 图3示出了根据本发明一个实施例的一种确定搜索查询词类别属性的装置的结构图；

图4示出了根据本发明又一个实施例的一种确定搜索查询词类别属性的装置的结构图；

图5示意性地示出了用于执行根据本发明的方法的服务器的框图；以及

20 图6示意性地示出了用于保持或者携带实现根据本发明的方法的程序代码的存储单元。

具体实施例

下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示
25 了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

图1示出了根据本发明一个实施例的一种确定搜索查询词类别属性的方法的流程图。如图1所示，该方法包括：

30 步骤S110，对输入的搜索查询词进行特征提取得到对应的特征向量。

步骤 S120, 根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率。

步骤 S130, 分析所述搜索查询词的类别偏重概率, 确定所述搜索查询词的类别属性。

5 图 1 所示的技术方案, 通过提取搜索查询词的特征向量, 将该特征向量输入到查询词分类器得到所述搜索查询词的类别偏重概率, 并通过分析类别偏重概率确定该查询词的类别属性, 从而为搜索提供基础依据, 保障搜索的准确性。并且能够为搜索排序等后续事件提供基础特征。即根据步骤 S130 中确定的所述搜索查询词的类别属性对搜索结果进行排序。

10 在本发明的一个实施例中, 对查询词分类器的分类模型的获取过程进行了改进, 以克服单纯的人工标注数据不能满足需求, 以及由于训练的数据不均衡导致的分类模型倾斜等问题。在本发明的一个实施例中, 在对输入的搜索查询词进行特征提取之前还进行预处理以及缓存查询等操作, 以提高效率。为说明上述技术方案, 下面以图 2 所示的流程为例进行说明。

15 图 2 示出了根据本发明一个实施例的确定分类器的分类模型以及利用分类器确定搜索查询词的类别属性的方法的流程图。如图 2 所示, 该方法包括确定分类器的分类模型的步骤 S220~步骤 S224, 即线下的训练学习过程; 以及利用分类器确定搜索查询词的类别属性的步骤 S230~步骤 S238, 即线上的预测过程。

20 步骤 S220, 获取标注类别的标注数据。

本步骤中, 可以获取人工标注的标注数据。也可以预先对网址链接进行分类标注; 根据用户搜索指定搜索查询词后点击的网址链接, 建立所述指定搜索查询词与所点击网址链接的分类之间的对应关系, 得到标注数据。或者也可以结合上述两种方式获取标注数据。

25 SVM 的学习属于监督学习, 其训练过程依赖于大量标注数据。单纯人工标注数据已经不能满足要达到一定准确度的分类器对于大规模标注数据量的需求, 因此本实施例中提供了一种半自动化训练数据标注方法, 采用一种使用搜索引擎点击反馈间接标注训练数据的方法。具体地, 首先收集大量人工标注 host, 可以采用 ODP 的开放数据或手动标注 host, 标注过程即建立 host
30 到类别的对应关系, 然后用户搜索某词语后点击了某些 host, 根据第一步建立的 host 到类别的对应关系, 我们便通过 host 间接地建立了搜索词到类别的

对应关系。这里，host是指网站的主机，和网站的网址链接是一一对应的，因此其实是对网址链接的标注。

步骤 S222，从每个类别的标注数据中抽样一定数量的数据，得到该类别的抽样数据。

- 5 经由上述的半自动化方法得到的标注数据存在显著的数据不均衡问题。人工标注也可能存在数据不均衡的问题。用于分类模型的训练数据不均衡意味着分类面将会偏向于数据少的类别一侧，使得分类模型倾向于将输入实例类别判定为数据较多的一类，导致分类错误，而这种情况在多元分类中更加复杂。为在多元分类模型中减少甚至避免数据不均衡问题，本实施例中主要
- 10 采用随机抽样和类别惩罚权值的调节的方法。

随机抽样为：从每个类别的标注数据中抽样多于第一预设值，少于第二预设值数量的数据，得到该类别的抽样数据。即从原有训练数据中每个类别以等概率抽样最少 m 条数据，最多 n 条数据，这样可以从一定程度上减少数据的极不均衡问题。

- 15 类别惩罚：在训练的过程中，对不同类别设置不同的惩罚因子；其中，对抽样后数据条数较少的类别设置较大的惩罚因子。即针对抽样后的数据，在训练过程中，通过施加不同类别的惩罚因子权重，对抽样后数据条数较少的类别设置较大的惩罚因子，可以避免分类面偏向少数据的类别。

步骤 S224，对各类别的抽样数据进行训练得到分类模型。

- 20 训练过程的特征数据提取和生成与预测的过程一致，包括分词、生成特征向量、生成libsvm格式数据。

训练。利用 liblinear 可以实现多元分类模型的训练，即采用 Liblinear 对各类别的抽样数据进行训练得到分类模型。

- 25 在原有 liblinear 实现中，利用 OpenMP 将训练过程改写为多分类并行训练，可以大大提高训练效率。多元分类的分类模型为一矩阵 M ， M 为 $m \times n$ 矩阵， m 为分类个数， n 为特征个数，矩阵的每个元素 $a(i, j)$ 表示第 j 个特征在第 i 个分类的分类权值，该权值为浮点型。

- 30 由于特征数少则几十万多则上百万，而实际的分类个数也有约600个，那么所得分类模型矩阵至少包含上亿个元素，原有liblinear训练输出的分类模型占用近4G磁盘空间。因此为提高在离线/在线预测时分类模型加载效率，在本发明的一个实施例中通过两种方法降低分类模型大小：

第一，通过减少分类模型中的分类权值的精度，来降低分类模型所占用的存储空间。例如，通过将权值的小数部分截取到6位，降低了一半的磁盘

存储。

第二，使用L1正则化训练模型。L1正则具有特征选择的效果，所得分类模型有大量的特征权值为0，同样可以降低磁盘存储。

步骤 S226，将分类模型输出到分类器。

5 通过上述过程得到了分类器的分类模型。下面是一次的在线预测过程。

步骤 S230，接收输入的搜索查询词。

步骤 S232，对输入的搜索查询词进行预处理。

由于搜索引擎搜索框中的输入词语各式各样，杂乱的信息势必干扰分类效果，因此需要对搜索词进行预处理。该过程实际为清洗词语的过程，包括
10 以下处理中的一种或多种：长词过滤、删除特殊字符和删除停用词。

步骤 S234，根据搜索查询词查询缓存，如果命中缓存，则直接输出所述搜索查询词的类别偏重概率；如果没有命中缓存，则执行步骤 S236。

这里，需要预设缓存，在预设的缓存中保存一定数量的搜索查询词和对应的类别偏重概率。每隔一段时间对缓存中的数据进行更新。

15 在预设的缓存中保存一定数量的搜索查询词和对应的类别偏重概率可以以为：

从已确定类别偏重概率的搜索查询词中找出被查询次数最多的前预定个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到所述缓存中；

20 或者，

针对不同的内容分发网络 CDN 分别预设缓存；对于每个 CDN 的缓存，从访问该 CDN 的已确定类别偏重概率的搜索查询词中，找出被查询次数最多的前预定个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到该 CDN 的缓存中。这种方式考虑了查询词访问的地域性
25 差异，以提高缓存命中率。

步骤 S236，对搜索查询词进行特征提取得到对应的特征向量。

本步骤具体包括：对输入的搜索查询词进行分词，利用分词后的结果构造 libsvm 格式的特征向量。

这是因为分类器的输入数据为词语的特征数据，该步骤通过将搜索查询
30 词转换为符合分类器输入格式的特征向量，主要包括分词和构造特征向量。利用分词后的结构构造 libsvm 格式的特征向量，该格式的向量为分类器的输

入。libsvm 特征向量格式采用稀疏表示的向量空间模型，其经过转换后的单查询词的特征空间维度为 60 万至 100 万维。

步骤 S238，将所述特征向量输入到分类器，分类器基于分类模型进行预测，得到所述搜索查询词的类别偏重概率后输出。

5 在本发明的实施例中，搜索查询词经过转换后生成libsvm格式的数据，假设特征向量为列向量 X ，分类模型矩阵为 M ，则对该词预测为第 i 类的概率为 $p_i = X * M_i$ 。其预测输出为该词被判别为每个类别下的概率值。除此之外，还可以包括其它一些数值信息，比如被预测为不同类别的概率的方差等。实践证明，这些输出在后期有关置信度的计算和按条件过滤的计算中非常有用。

汉语的一次多义现象很普遍，分类模型中也保留了这一特性，即同一查询词会有多个类别输出，而输出的标准则根据模型预测的该词在不同类别下的概率、方差等数值特征计算。比如通过模型预测，“三国演义”属于书籍、电视剧、工商业的概率分别为0.9、0.8、0.2。那么可以观察到，该词属于书籍、电视剧的概率要显著高于工商业的概率，因此我们认为该词可以属于书籍、电视剧。以上的识别方法可以通过依次计算排序后前 n 个概率的样本均值实现。这便是通过分类器在不同类别下的概率直接的关系得出。

15 本发明的上述技术方案可以实时分析用户查询词所属的类别，训练数据量大且有较高的准确度，最终训练的模型分类精度较高。类别涵盖面较广，能满足大部分查询词分类业务和机器学习模型的需求，属于互联网基础组件。

图3示出了根据本发明一个实施例的一种确定搜索查询词类别属性的装置的结构图。如图3所示，该确定搜索查询词类别属性的装置300包括：

25 特征提取单元301，适于对输入的搜索查询词进行特征提取得到对应的特征向量。

分类器302，适于根据所述特征向量得到所述搜索查询词的类别偏重概率并发送给输出单元。

输出单元303，适于分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性并输出。

30 图4示出了根据本发明又一个实施例的一种确定搜索查询词类别属性的装置的结构图。如图4所示，该确定搜索查询词类别属性的装置400包括：

特征提取单元401，适于对输入的搜索查询词进行特征提取得到对应的

特征向量。

分类器 402, 适于根据所述特征向量得到所述搜索查询词的类别偏重概率并发送给输出单元。

输出单元 403, 适于分析所述搜索查询词的类别偏重概率, 确定所述搜索查询词的类别属性并输出。

在本发明的一个实施例中, 图 4 所示的确定搜索查询词类别属性的装置 400 还包括:

标注数据获取单元 404, 适于获取标注类别的标注数据;

抽样单元 405, 适于从每个类别的标注数据中抽样多于第一预设值, 少于第二预设值数量的数据, 得到该类别的抽样数据;

训练单元 406, 适于对各类别的抽样数据进行训练得到所述分类器的分类模型。

在本发明的一个实施例中, 所述训练单元 406, 适于在训练的过程中, 对不同类别设置不同的惩罚因子; 其中, 对抽样后数据条数较少的类别设置较大的惩罚因子。

在本发明的一个实施例中, 所述训练单元 406 得到的所述分类模型为 $m \times n$ 的矩阵, m 为分类个数, n 为特征个数, 矩阵的每个元素 $a(i, j)$ 表示第 j 个特征在第 i 个分类的分类权值;

所述训练单元 406, 进一步适于通过减少分类模型中的分类权值的精度, 来降低分类模型所占用的存储空间; 和/或, 进一步适于使用 L1 正则化训练模型, 以降低分类模型所占用的存储空间。

在本发明的一个实施例中, 所述标注数据获取单元 404, 适于获取人工标注的标注数据; 和/或, 适于预先对网址链接进行分类标注, 根据用户搜索指定搜索查询词后点击的网址链接, 建立所述指定搜索查询词与所点击网址链接的分类之间的对应关系, 得到标注数据。

在本发明的一个实施例中, 所述训练单元 406, 适于采用 Liblinear 对各类别的抽样数据进行训练得到分类模型;

所述特征提取单元 401, 适于对输入的搜索查询词进行分词, 利用分词后的结果构造 libsvm 格式的特征向量。

在本发明的一个实施例中, 图 4 所示的确定搜索查询词类别属性的装置

400 还包括：预处理单元 407，适于对输入搜索查询词进行预处理；所述预处理包括以下处理中的一种或多种：长词过滤、删除特殊字符和删除停用词。

在本发明的一个实施例中，图 4 所示的确定搜索查询词类别属性的装置

5 400 还包括：

缓存单元 408，适于保存一定数量的搜索查询词和对应的类别偏重概率；

缓存查询单元 409，适于根据输入的搜索查询词查询缓存单元 408；如果命中缓存，则直接将所述输入的搜索查询词的类别偏重概率发送给所述输出单元 403；如果没有命中缓存，则将所述输入的搜索查询词发送给所述特征提取单元 401。

在本发明的一个实施例中，图 4 所示的确定搜索查询词类别属性的装置 400 还包括：缓存数据设置单元 410；

所述缓存数据设置单元 410，适于从已确定类别偏重概率的搜索查询词中找出被查询次数最多的前预定个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到所述缓存单元中；

或者，

针对不同的内容分发网络 CDN 分别设置缓存单元；

所述缓存数据设置单元 410，适于对每个 CDN 的缓存单元，从访问该 CDN 的已确定类别偏重概率的搜索查询词中，找出被查询次数最多的前预定个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到该 CDN 的缓存单元中。

在本发明的一个实施例中，图 4 所示的确定搜索查询词类别属性的装置 400 还包括：排序单元 411，适于根据确定的所述搜索查询词的类别属性对搜索结果进行排序。

综上所述，本发明这种对输入的搜索查询词进行特征提取得到对应的特征向量，根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率，分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性的技术方案，通过提取搜索查询词的特征向量，将该特征向量输入到查询词分类器得到所述搜索查询词的类别偏重概率，并通过分析类别偏重概率确定该查询词的类别属性，从而为搜索提供基础依据，保障搜索的准确性。

并且能够为搜索排序等后续事件提供基础特征。

需要说明的是：

在此提供的算法和显示不与任何特定计算机、虚拟系统或者其它设备固有相关。各种通用系统也可以与基于在此的示教一起使用。根据上面的描述，
5 构造这类系统所要求的结构是显而易见的。此外，本发明也不针对任何特定编程语言。应当明白，可以利用各种编程语言实现在此描述的本发明的内容，并且上面对特定语言所做的描述是为了披露本发明的最佳实施方式。

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
10 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：即所要求保护的本发明要求比在每个权利要求
15 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
20 把它们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
25 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
30

利要求书中，所要求保护的实施例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确定搜索查询词类别属性的装置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。

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确定搜索查询词类别属性的装置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 5 示出了可以实现根据本发明的确定搜索查询词类别属性的方法的服务器，例如搜索服务器。该服务器传统上包括处理器 510 和以存储器 520 形式的计算机程序产品或者计算机可读介质。存储器 520 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 520 具有用于执行上述方法中的任何方法步骤的程序代码 531 的存储空间 530。例如，用于程序代码的存储空间 530 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 531。这些程序代码可以从一个或者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 6 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 5 的服务器中

的存储器 520 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 531'，即可以由例如诸如 510 之类的处理器读取的代码，这些代码当由服务器运行时，导致该服务器执行上面所描述的方法中的各个步骤。

- 5 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，
10 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。
15 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
20 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、一种确定搜索查询词类别属性的方法，其中，该方法包括：

对输入的搜索查询词进行特征提取得到对应的特征向量；

5 根据所述特征向量经由查询词分类器得到所述搜索查询词的类别偏重概率；

分析所述搜索查询词的类别偏重概率，确定所述搜索查询词的类别属性。

2、如权利要求 1 所述的方法，其中，该方法还包括获得所述查询词分类器的分类模型的如下步骤：

获取标注类别的标注数据；

从每个类别的标注数据中抽样多于第一预设值，少于第二预设值数量的数据，得到该类别的抽样数据；

对各类别的抽样数据进行训练得到分类模型。

15 3、如权利要求 1-2 任一项所述的方法，其中，所述对各类别的抽样数据进行训练得到分类模型包括：

在训练的过程中，对不同类别设置不同的惩罚因子；其中，对抽样后数据条数较少的类别设置较大的惩罚因子。

4、如权利要求 1-3 任一项所述的方法，其中，所述分类模型为 $m \times n$ 的矩阵， m 为分类个数， n 为特征个数，矩阵的每个元素 $a(i, j)$ 表示第 j 个特征在第 i 个分类的分类权值；

所述对各类别的抽样数据进行训练得到分类模型还包括：

通过减少分类模型中的分类权值的精度，来降低分类模型所占用的存储空间；和/或，使用 $L1$ 正则化训练模型，以降低分类模型所占用的存储空间。

25 5、如权利要求 1-4 任一项所述的方法，其中，所述获取标注类别的标注数据包括：

获取人工标注的标注数据；

和/或，

30 预先对网址链接进行分类标注；根据用户搜索指定搜索查询词后点击的网址链接，建立所述指定搜索查询词与所点击网址链接的分类之间的对应关

系，得到标注数据。

6、如权利要求 1-5 中任一项所述的方法，其中，

所述对各类别的抽样数据进行训练得到分类模型包括：采用 Liblinear 对各类别的抽样数据进行训练得到分类模型；

5 所述对输入的搜索查询词进行特征提取得到对应的特征向量包括：对输入的搜索查询词进行分词，利用分词后的结果构造 libsvm 格式的特征向量。

7、如权利要求 1-6 任一项所述的方法，其中，在所述对输入的搜索查询词进行特征提取之前该方法还包括：对输入的搜索查询词进行预处理；

所述预处理包括以下处理中的一种或多种：长词过滤、删除特殊字符和
10 删除停用词。

8、如权利要求 1-7 任一项所述的方法，其中，该方法还包括：预设缓存，在所述缓存中保存一定数量的搜索查询词和对应的类别偏重概率；

在所述对输入的搜索查询词进行特征提取之前，该方法还包括：根据输入的搜索查询词查询缓存；如果命中缓存，则直接输出所述输入的搜索查询
15 词的类别偏重概率；如果没有命中缓存，则执行所述对输入的搜索查询词进行特征提取的步骤以及后续步骤。

9、如权利要求 1-8 任一项所述的方法，其中，所述预设缓存，在所述缓存中保存一定数量的搜索查询词和对应的类别偏重概率包括：

从已确定类别偏重概率的搜索查询词中找出被查询次数最多的前预定
20 个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到所述缓存中；

或者，

针对不同的内容分发网络 CDN 分别预设缓存；对于每个 CDN 的缓存，从访问该 CDN 的已确定类别偏重概率的搜索查询词中，找出被查询次数最
25 多的前预定个数的搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到该 CDN 的缓存中。

10、如权利要求 1-9 任一项所述的方法，其中，该方法进一步包括：根据确定的所述搜索查询词的类别属性对搜索结果进行排序。

11、一种确定搜索查询词类别属性的装置，其中，该装置包括：
30 特征提取单元，适于对输入的搜索查询词进行特征提取得到对应的特征

向量;

分类器,适于根据所述特征向量得到所述搜索查询词的类别偏重概率并发送给输出单元;

输出单元,适于分析所述搜索查询词的类别偏重概率,确定所述搜索查询词的类别属性并输出。

12、如权利要求 11 所述的装置,其中,该装置进一步包括:

标注数据获取单元,适于获取标注类别的标注数据;

抽样单元,适于从每个类别的标注数据中抽样多于第一预设值,少于第二预设值数量的数据,得到该类别的抽样数据;

10 训练单元,适于对各类别的抽样数据进行训练得到所述分类器的分类模型。

13、如权利要求 11 或 12 所述的装置,其中,

所述训练单元,适于在训练的过程中,对不同类别设置不同的惩罚因子;其中,对抽样后数据条数较少的类别设置较大的惩罚因子。

15 14、如权利要求 11-13 任一项所述的装置,其中,

所述训练单元得到的所述分类模型为 $m \times n$ 的矩阵, m 为分类个数, n 为特征个数,矩阵的每个元素 $a(i, j)$ 表示第 j 个特征在第 i 个分类的分类权值;

20 所述训练单元,进一步适于通过减少分类模型中的分类权值的精度,来降低分类模型所占用的存储空间;和/或,进一步适于使用 L1 正则化训练模型,以降低分类模型所占用的存储空间。

15、如权利要求 11-14 任一项所述的装置,其中,

25 所述标注数据获取单元,适于获取人工标注的标注数据;和/或,适于预先对网址链接进行分类标注,根据用户搜索指定搜索查询词后点击的网址链接,建立所述指定搜索查询词与所点击网址链接的分类之间的对应关系,得到标注数据。

16、如权利要求 11-15 任一项所述的装置,其中,

所述训练单元,适于采用 Liblinear 对各类别的抽样数据进行训练得到分类模型;

30 所述特征提取单元,适于对输入的搜索查询词进行分词,利用分词后的

结果构造 libsvm 格式的特征向量。

17、如权利要求 11-16 任一项所述的装置，其中，该装置还包括：预处理单元，适于对输入的搜索查询词进行预处理；

所述预处理包括以下处理中的一种或多种：长词过滤、删除特殊字符和
5 删除挺用词。

18、如权利要求 11-17 任一项所述的装置，其中，该装置还包括：

缓存单元，适于保存一定数量的搜索查询词和对应的类别偏重概率；

缓存查询单元，适于根据输入的搜索查询词查询缓存单元；如果命中缓存，则直接将所述输入的搜索查询词的类别偏重概率发送给所述输出单元；
10 如果没有命中缓存，则将所述输入的搜索查询词发送给所述特征提取单元。

19、如权利要求 11-18 任一项所述的装置，其中，该装置还包括：缓存数据设置单元；

所述缓存数据设置单元，适于从已确定类别偏重概率的搜索查询词中找出被查询次数最多的前预定个数的搜索查询词，将该预定个数的搜索查询词
15 和相应的类别偏重概率对应保存到所述缓存单元中；

或者，

针对不同的内容分发网络 CDN 分别设置缓存单元；

所述缓存数据设置单元，适于对每个 CDN 的缓存单元，从访问该 CDN 的已确定类别偏重概率的搜索查询词中，找出被查询次数最多的前预定个数的
20 搜索查询词，将该预定个数的搜索查询词和相应的类别偏重概率对应保存到该 CDN 的缓存单元中。

20、如权利要求 11-19 任一项所述的装置，其中，该装置进一步包括：

排序单元，适于根据确定的所述搜索查询词的类别属性对搜索结果进行排序。

21、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服务器上运行时，导致所述服务器执行根据权利要求 1-10 中的任一个所述
25 的确定搜索查询词类别属性的方法。

22、一种计算机可读介质，其中存储了如权利要求 21 所述的计算机程序。

1/3

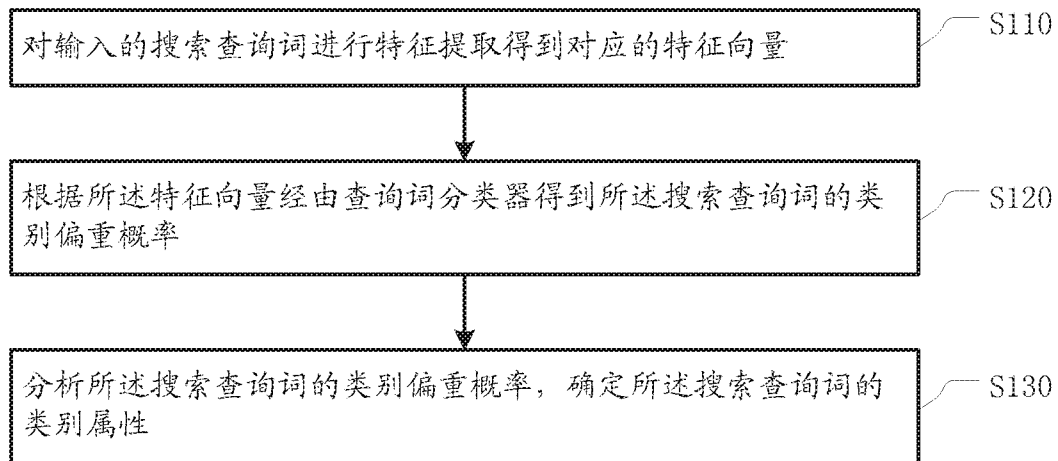


图 1

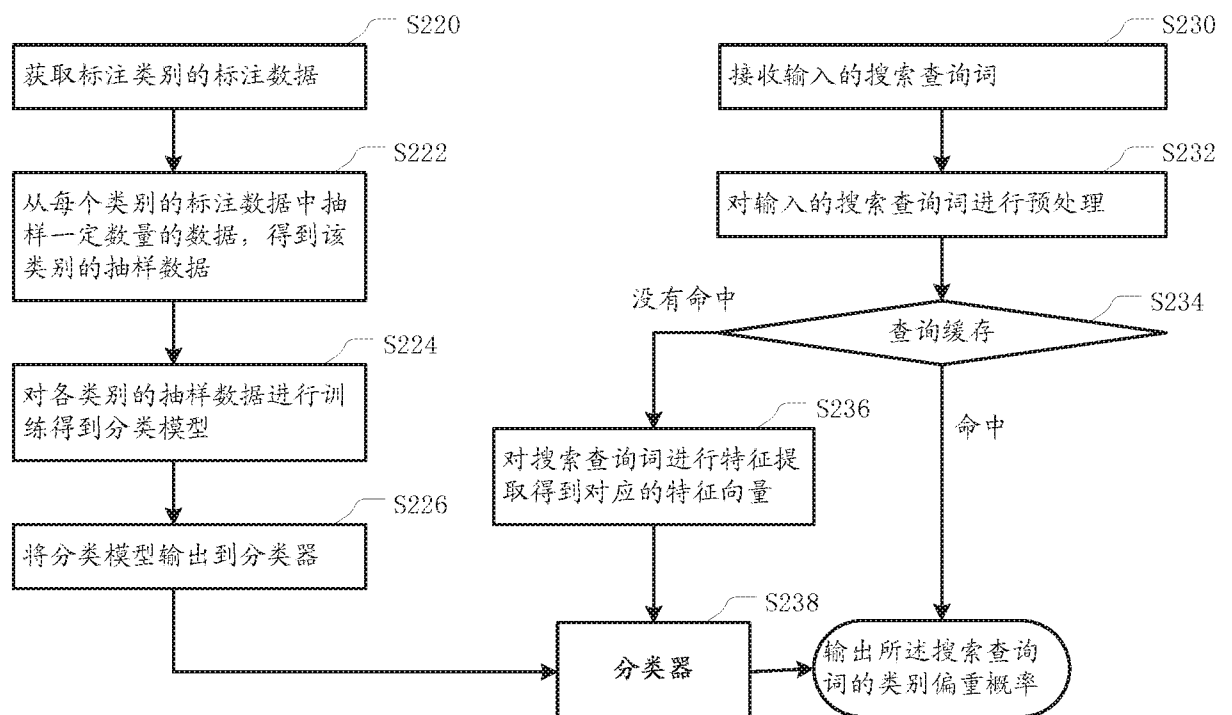


图 2

2/3

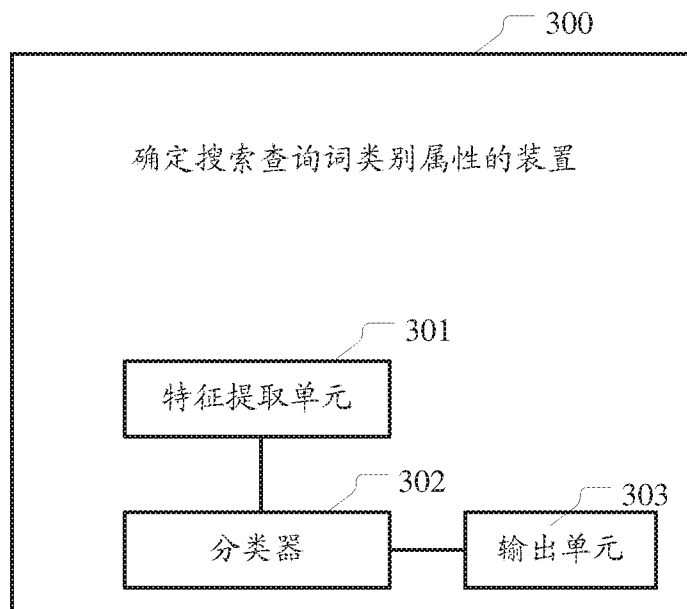


图 3

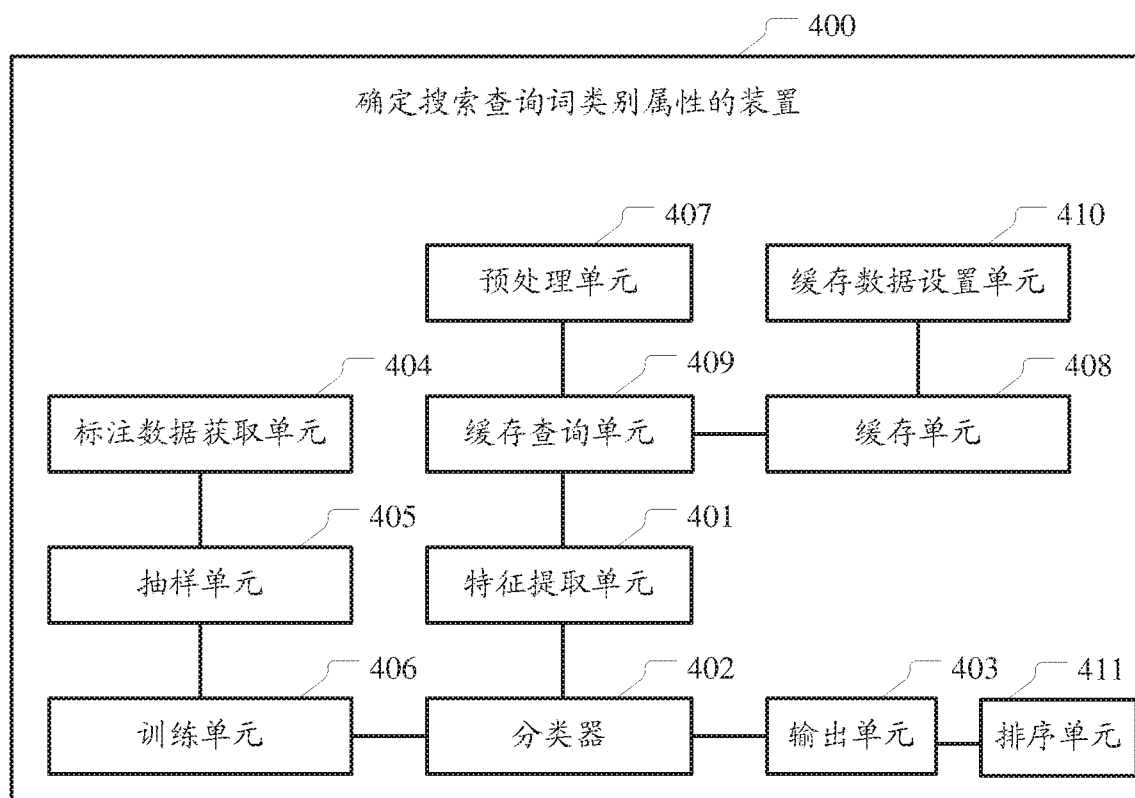


图 4

3/3

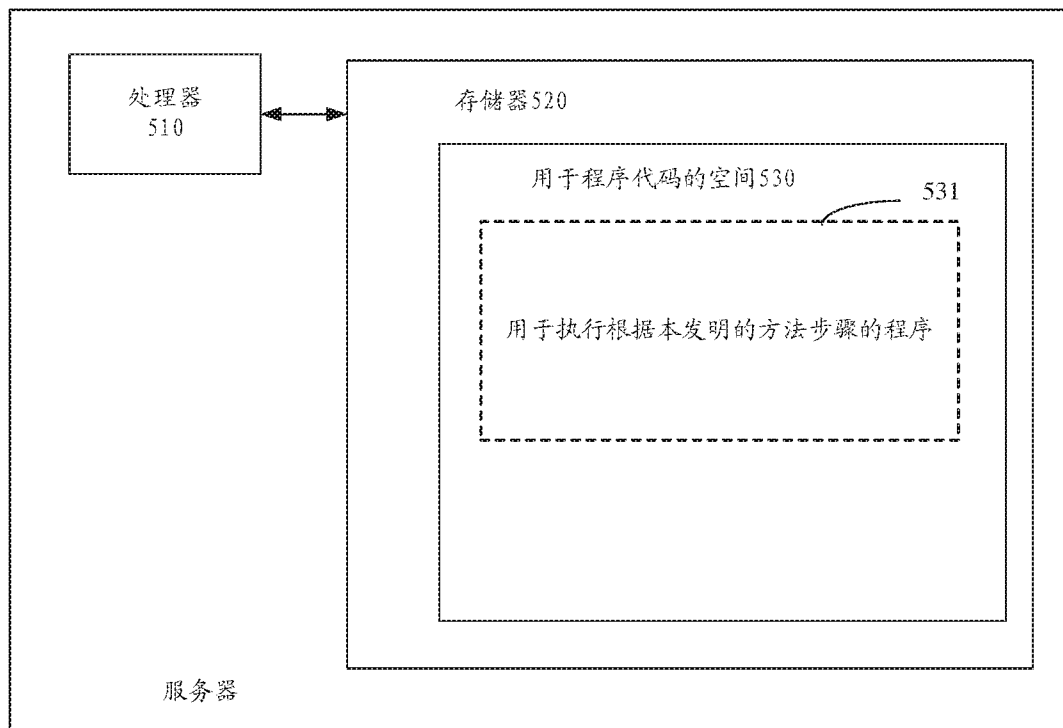


图 5

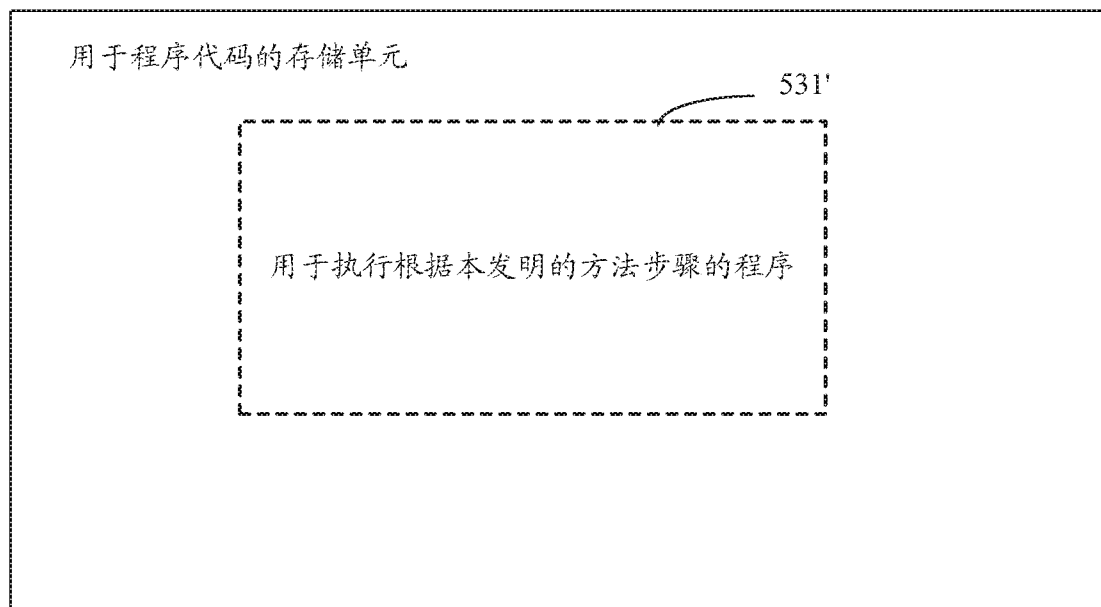


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/079800

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, WPI, EPODOC, CNKI, GOOGLE, ISI: characteristic vector, word, keyword, search, retrieve, classify, characteris, category, probability

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 104050240 A (BEIJING QIHOO TECHNOLOGY CO., LTD. et al.), 17 September 2014 (17.09.2014), claims 1-10, and description, paragraphs [0136]-[0145]	1-22
X	CN 103123636 A (BEIJING BAIDU NETWORK TECHNOLOGY CO., LTD.), 29 May 2013 (29.05.2013), description, paragraphs [0074]-[0116]	1-22
X	CN 103106262 A (SINA.COM TECHNOLOGY (CHINA) CO., LTD.), 15 May 2013 (15.05.2013), claims 1-7, and description, paragraphs [0004]-[0041]	1-22
A	CN 103810264 A (XI'AN UNIVERSITY OF TECHNOLOGY), 21 May 2014 (21.05.2014), the whole document	1-22

☐ 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 03 August 2015 (03.08.2015)	Date of mailing of the international search report 02 September 2015 (02.09.2015)
Name and mailing address of the ISA/CN: 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 WANG, Fang Telephone No.: (86-10) 82245496

International application No.
PCT/CN2015/079800

Form PCT/ISA/210 (patent family annex) (July 2009)

A. 主题的分类 G06F 17/30 (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, WPI, EPDOC, CNKI, GOOGLE, ISI: 词语, 搜索, 查询, 检索, 类目, 类别, 分类, 特征向量, 概率, word, keyword, search, retrieve, classify, characteris, category, probability		
C. 相关文件		
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 104050240 A (北京奇虎科技有限公司 等) 2014年 9月 17日 (2014 - 09 - 17) 权利要求1-10, 说明书第[0136]-[0145]段	1-22
X	CN 103123636 A (北京百度网讯科技有限公司) 2013年 5月 29日 (2013 - 05 - 29) 说明书第[0074]-[0116]段	1-22
X	CN 103106262 A (新浪网技术中国有限公司) 2013年 5月 15日 (2013 - 05 - 15) 权利要求1-7, 说明书第[0004]-[0041]段	1-22
A	CN 103810264 A (西安理工大学) 2014年 5月 21日 (2014 - 05 - 21) 全文	1-22
☐ 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 2015年 8月 3日		国际检索报告邮寄日期 2015年 9月 2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 王芳 电话号码 (86-10)82245496

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/079800

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	104050240	A	2014年 9月 17日	无	
CN	103123636	A	2013年 5月 29日	无	
CN	103106262	A	2013年 5月 15日	无	
CN	103810264	A	2014年 5月 21日	无	