

(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
2020 年 10 月 29 日 (29.10.2020)



(10) 国际公布号
WO 2020/215437 A1

(51) 国际专利分类号:
G06F 16/29 (2019.01)

(21) 国际申请号: PCT/CN2019/088771

(22) 国际申请日: 2019 年 5 月 28 日 (28.05.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910334422.5 2019 年 4 月 24 日 (24.04.2019) CN

(71) 申请人: 上海交通大学 (SHANGHAI JIAO TONG UNIVERSITY) [CN/CN]; 中国上海市闵行区东川路 800 号, Shanghai 200240 (CN)。

(72) 发明人: 过敏意 (GUO, Minyi); 中国上海市闵行区东川路 800 号, Shanghai 200240 (CN)。 姚斌 (YAO, Bin); 中国上海市闵行区东川路 800 号, Shanghai 200240 (CN)。 张建锋 (ZHANG, Jianfeng); 中国浙江省杭州市西湖区西斗门路 3 号天堂软件园 A 幢 10 楼 G 座, Zhejiang 311604 (CN)。 林昊 (LIN, Hao); 中国浙江省杭州市西湖区西斗门路 3 号天堂软件园 A 幢 10 楼 G 座, Zhejiang 311604

(CN)。 陈全 (CHEN, Quan); 中国上海市闵行区东川路 800 号, Shanghai 200240 (CN)。

(74) 代理人: 上海市汇业律师事务所 (SHANGHAI HUIYE LAW FIRM); 中国上海市长宁区延安西路 726 号华敏翰尊国际 5 层, Shanghai 200050 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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,

(54) Title: APPROXIMATE SEARCH METHOD FOR SPATIAL KEYWORD QUERY IN ELECTRONIC MAP

(54) 发明名称: 应用于电子地图的空间关键字查询的近似搜索方法

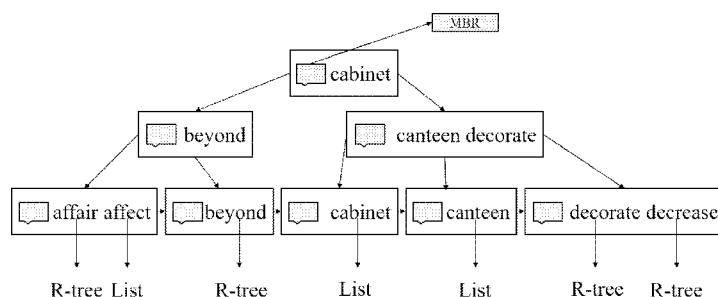


图 1

(57) Abstract: Disclosed is an approximate search method for spatial keyword query in an electronic map. The method comprises the following steps: S1: reading all of pieces of data in a data set, constructing an index, and executing step S2 for each keyword of a single piece of data; S2: comparing the frequency of the keyword with a frequency threshold; S3: using a mapping function to map the keyword to an R-tree, organizing spatial text objects comprising the keyword, and executing S5; S4: mapping the keyword to an object list, and executing S5; S5: for objects in a leaf node, calculating a smallest boundary rectangle of the objects, and using the smallest boundary rectangle to represent a spatial region thereof, and for a non-leaf node, acquiring smallest boundary rectangles of child nodes of the non-leaf node, forming a larger smallest boundary rectangle, and using the larger smallest boundary rectangle to represent a space coverage range of the node; S6: performing query searching for approximate Boolean range query (aBRQ) within a spatial framework constructed in S5; and S7: performing, within the spatial framework constructed in S5, performing query searching for k nearest neighbor queries comprised in an approximate keyword.



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)。

根据细则4.17的声明：

- 发明人资格(细则4.17(iv))

本国际公布：

- 包括国际检索报告(条约第21条(3))。

(57) 摘要：本发明公开了一种应用于电子地图的空间关键字查询的近似搜索方法，其包括如下步骤：**S1：**读取数据集的各条数据进行索引构建、针对单条数据的各个关键字分别跳转至步骤**S2**；**S2：**将关键字的频率与频率阈值进行比对；**S3：**采用映射函数将关键字映射到**R-tree**，并组织包含该关键字的空间文本对象，跳转至**S5**；**S4：**将关键字映射到对象列表中，跳转至**S5**；**S5：**对于叶子节点中的对象、计算其最小边界矩形来表示它们的空间区域，对于非叶子节点，收集其子节点的最小边界矩形并组成更大的最小边界矩形以代表该节点的空间覆盖范围；**S6：**在**S5**构建的空间框架内实现对**aBRQ(Approximate Boolean Range Query)**的查询搜索；**S7：**在**S5**构建的空间框架内实现对近似关键字包含的**k**最近邻查询的查询搜索。

应用于电子地图的空间关键字查询的近似搜索方法

技术领域

本发明属于定位技术领域，具体来说涉及一种基于 Spark 平台的应用于电子地图的空间关键字查询的近似搜索方法。

背景技术

近年来随着通信技术的发展和移动终端的广泛使用，基于位置的社会服务层出不穷。空间关键字查询是以用户的地理位置信息和多个查询关键字作为参数，返回和这些参数有着空间和文本相关度的空间对象。所谓近似关键字搜索，是指搜索给定距离度量下，与查询字符串相似的字符串。目前，主流的方案是基于空间优先的索引结构，然而，基于空间优先的结构对于平均具有数十个关键字的空间文本对象进行索引优化并不高效。因此，如何开发出一种基于文本匹配优先的空间关键字查询近似搜索方法，以避免无必要的文本匹配过程，降低匹配开销、以取得更好的搜索效果。是本领域技术人员需要研究的方向。以下为本申请中所涉及的字母缩写的注释：R-tree：B-tree向多维空间发展的另一种形式，它将空间对象按范围划分，每个结点都对应一个区域和一个磁盘页，非叶结点的磁盘页中存储其所有子结点的区域范围，非叶结点的所有子结点的区域都落在它的区域范围之内。IR-tree：以倒排索引和R-tree索引为基础，通过倒排索引解决文本相似度的计算模型。BFIR-tree：基于海量数据处理实现的IR-tree；CBFIR-tree：动态的BFIR-tree；S2I-V结构：对不同频率的关键字应被区别处理的模型结构；eBRQ：基于关键字包含的范围查询；aBRQ：基于近似关键字包含的k最近邻查询；false positive rate：误检率；。KNN算法：即临近算法，是数据挖掘分类技术中最简单的方法之一。I-Node：一个叶子R树节点，它存储了将每个关键字映射到空间关键字对象的倒排列表。

发明内容

本发明要解决的技术问题是提供了一种应用于电子地图的空间关键字查询的近似搜索方法，能够避免无必要的文本匹配过程，降低匹配开销、以取得更好的搜索效果。

其采用的技术方案如下：

一种应用于电子地图的空间关键字查询的近似搜索方法，其包括如下步骤：S1：读取数据集的各条数据进行索引构建、针对单条数据的各个关键字分别跳转至步骤S2；S2：将关键字的频率与频率阈值进行比对，若该频率值高于预设的频率阈值，则跳转至S3、若该

频率值低于预设的频率阈值，则跳转至S4；S3：采用映射函数将关键字映射到R-tree，并组织包含该关键字的空间文本对象，跳转至S5；S4：将关键字映射到对象列表中，跳转至S5；S5：对于叶子节点中的对象、计算其最小边界矩形来表示它们的空间区域，对于非叶子节点，收集其子节点的最小边界矩形并组成更大的最小边界矩形以代表该节点的空间覆盖范围；S6：在S5构建的空间框架内实现对aBRQ的查询搜索；S7：在S5构建的空间框架内实现对近似关键字包含的k最近邻查询的查询搜索。

优选的是，上述应用于电子地图的空间关键字查询的近似搜索方法中：步骤S5包括如下步骤：S61：对于非叶子节点，计算从查询点到该节点的MBR的距离，并检查它是否与给定圆形区域相交；S62：基于字符串匹配的剪枝能力选择可能的子节点，并在到达叶节点时访问相应的列表或R树以查找离给定点在给定范围内的空间文本对象；S63：获取当前的候选结果、对每个关键字的临时结果取交集作为最后解集。

更优选的是，上述应用于电子地图的空间关键字查询的近似搜索方法中，步骤S7包括如下步骤：S71：读取待查询关键字的数目，若关键字为单关键字，跳转至S72，若关键字为多关键字，跳转至S73；S72：将每个与关键字相似的字符串对应的k个最近邻的并集中找到全局的kNN集合；S73：扩展LBAK树以支持近似关键字包含的k最近邻查询。

通过采用上述技术方案：本发明以混合索引的方式将空间匹配和文本匹配的剪枝能力进行结合，从而避免了很多不必要的文本匹配过程，降低匹配开销、字符串的冗余存储，降低了存储空间成本。

附图说明

下面结合附图与具体实施方式对本发明作进一步详细的说明：

图1为实施例1的工作流程示意图。

图2为本发明中查询百分比逐渐增加时对本方案查询区域的影响示意图；

图3为本发明中关键字数目逐渐增加时对本方案查询区域的影响示意图。

具体实施方式

为了更清楚地说明本发明的技术方案，下面将结合各个实施例作进一步描述。

如图1所示为实施例1：

一种应用于电子地图的空间关键字查询的近似搜索方法，其包括如下步骤：

S1：读取数据集的各条数据进行索引构建、针对单条数据的各个关键字分别跳转至步骤S2；

S2: 将关键字的频率与频率阈值进行比对, 若该频率值高于预设的频率阈值, 则跳转至 S3; 若该频率值低于预设的频率阈值, 则跳转至 S4;

S3: 采用映射函数将关键字映射到 R-tree, 并组织包含该关键字的空间文本对象, 跳转至 S5;

S4: 将关键字映射到对象列表中, 跳转至 S5;

S5: 对于叶子节点中的对象、计算其最小边界矩形来表示它们的空间区域, 对于非叶子节点, 收集其子节点的最小边界矩形并组成更大的最小边界矩形以代表该节点的空间覆盖范围;

S6: 对 aBRQ 的查询算法: 假设一个 aBRQ $Q = \{Q_s = (\tau, \epsilon), Q_t = \{(s_1, h_1), \dots, (s_n, h_n)\}\}$, 对于每个关键字访问一次 Bed-first-tree (算法 1 行 2)。在访问 Bed-first-tree 的时候, 对于非叶子节点, 首先计算从查询点到该节点的 MBR 的距离, 并检查它是否与给定圆形区域相交。然后利用字符串匹配的剪枝能力来选择可能的子节点。如果到达叶节点, 访问相应的列表或 R 树以查找离给定点在给定范围内的空间文本对象, 然后获得当前的候选结果。对每个关键字的临时结果取交集即可获得最后的解集;

S7: 对近似关键字包含的 k 最近邻查询的查询算法: 对于具有单个关键字的近似关键字包含的 k 最近邻查询, 简单地找到为每个与其相似的字符串对应的 k 个最近邻 (称为 $\{K_1, K_2, \dots, K_n\}$)。然后在所有可能的 $\{K_1, K_2, \dots, K_n\}$ 的并集中找到全局的 kNN 集合。对于多个关键字, 情况则比较复杂, 因为特定关键字的 k 最近邻可能不包含其他的查询关键字。在这种情况下, 扩展 LBAK 树以支持近似关键字包含的 k 最近邻查询。由于 LBAK 树与 BFIR-TREE 类似, 从而可以快速修改后者的 eBKQ 算法来支持近似关键字包含的 k 最近邻查询。

在上述实施例中: 步骤 S2 中所述映射函数具有以下属性:

属性 1: 对于给定的 2 个字符串 s_i 和 s_j , 如在相对于给定的字符串长度为线性的时间内比较 $\phi(s_i)$ 和 $\phi(s_j)$ 的大小, 则称 ϕ 函数为可高效比较的。基于属性 1, 我们可以基于 B+-tree 有效的对字符串进行组织, 处理字符串更新操作。

属性 2: 给定字符串 q 和字符串区间 $[s_i, s_j]$, 如果映射函数 ϕ 可以高效地返回 q 和字符串区间的最小编辑距离, 则称 ϕ 为下界确定性映射。Bed-Tree 同时存储了所有字符串中的最小和最大字符串 $s_{\min}$ 和 $s_{\max}$, 它们表示在给定映射下 Bed-Tree 中字符串的边界。在根节点 n 中的字符串表示为 $\{s_1, \dots, s_k\}$, 其候选字符串区间: $\{(s_{\min}, s_1), (s_1, s_2), \dots, (s_{k-1}, s_k), (s_k, s_{\max})\}$ 。通过计算每个字符串区间到给定查询字符串的编辑距离下界, 我们可以排除不可能成为最终解的子节点。对于幸存的字符串区间, 我们访问相应的子节点,

并递归地遍历 Bed-Tree 直到到达叶子节点，在叶子节点，我们将计算查询关键字和叶子节点中的各个字符串之间的编辑距离，以验证候选者得到最终解。

比对实验

实验在具有两种配置的由 17 个节点组成的集群上执行：(1) 8 台具有 6 核 Intel Xeon E5-2603 v3 1.60GHz 处理器和 20GB RAM 的机器；(2) 2 台配备 6 核 Intel Xeon E5-2620 2.00GHz 处理器和 56GB RAM 的机器；(3) 7 台配备 6 核 Intel Xeon E5-2609 1.90GHz 处理器和 16GB RAM 的机器。我们选择一台类型 (2) 的机器作为主节点，其他机器作为从节点。每个从节点使用 15GB 内存和所有可用的 6 个内核进行后续计算。所有节点在 Ubuntu 14.04.2 LTS 系统上运行，并安装有 Hadoop 2.4.1 和 Spark 1.3.0。在两个真实的海量数据集上进行相关实验。

如图 2-3 所示：

当查询区域较小时，Bed-first-tree 稍微优于 LBAK-Tree。而当查询区域扩大时，Bed-first-tree 更具可扩展性和稳定性。随着查询关键词数量的增加，两者的表现都先增长，然后持续下降。通常，当只有一个查询关键字时，会检索出许多结果，而当分配两个查询关键字时，很多节点成为无关解，因此访问节点的次数大大减少，导致搜索开销的降低。然而，当查询关键字的数量继续增加时，字符串搜索的开销将成为主要问题。因此，当关键字大小大于 2 时，其性能出现下降趋势。因此，本发明的技术方案适用于大众点评等基于地理位置的服务应用。

以上所述，仅为本发明的具体实施例，但本发明的保护范围并不局限于此，任何熟悉本领域技术的技术人员在本发明公开的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。本发明的保护范围以权利要求书的保护范围为准。

权 利 要 求 书

1、一种应用于电子地图的空间关键字查询的近似搜索方法，其特征在于，包括如下步骤：

S1：读取数据集的各条数据进行索引构建、针对单条数据的各个关键字分别跳转至步骤 S2；

S2：将关键字的频率与频率阈值进行比对，若该频率值高于预设的频率阈值，则跳转至 S3、若该频率值低于预设的频率阈值，则跳转至 S4；

S3：采用映射函数将关键字映射到 R-tree，并组织包含该关键字的空间文本对象，跳转至 S5；

S4：将关键字映射到对象列表中，跳转至 S5；

S5：对于叶子节点中的对象、计算其最小边界矩形来表示它们的空间区域，对于非叶子节点，收集其子节点的最小边界矩形并组成更大的最小边界矩形以代表该节点的空间覆盖范围；

S6：在 S5 构建的空间框架内实现对基于近似关键字包含的 k 个最近邻查询的查询搜索；

S7：在 S5 构建的空间框架内实现对近似关键字包含的 k 最近邻查询的查询搜索。

2、如权利要求 1 所述应用于电子地图的空间关键字查询的近似搜索方法，其特征在于，步骤 S6 包括如下步骤：

S61：对于非叶子节点，计算从查询点到该节点的 MBR 的距离，并检查它是否与给定圆形区域相交；

S62：基于字符串匹配的剪枝能力选择可能的子节点，并在到达叶节点时访问相应的列表或 R 树以查找离给定点在给定范围内的空间文本对象；

S63：获取当前的候选结果、对每个关键字的临时结果取交集作为最后解集。

3、如权利要求 1 所述应用于电子地图的空间关键字查询的近似搜索方法，其特征在于，步骤 S7 包括如下步骤：

S71：读取待查询关键字的数目，若关键字为单关键字，跳转至 S72，若关键字为多关键字，跳转至 S73；

S72：将每个与关键字相似的字符串对应的 k 个最近邻的并集中找到全局的 kNN 集合；

S73：扩展 LBAK 树以支持近似关键字包含的 k 最近邻查询。

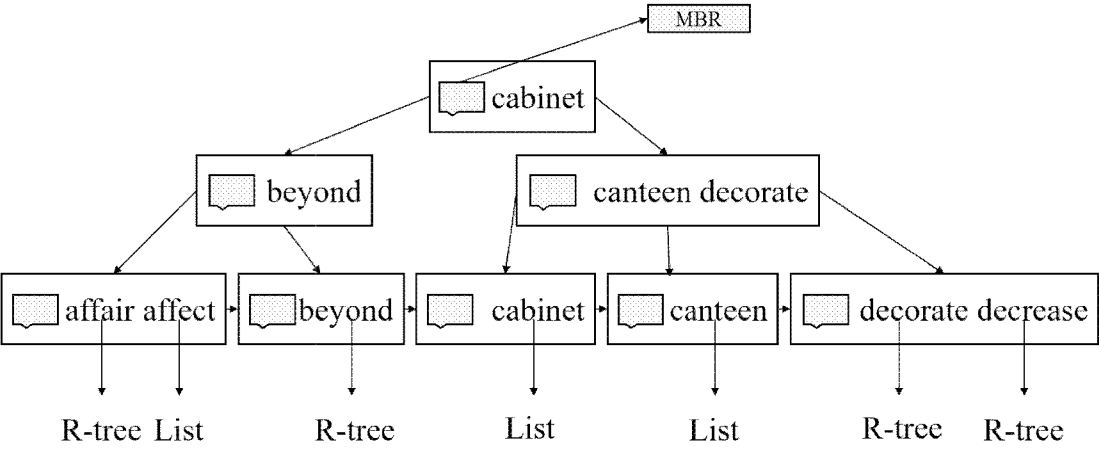


图 1

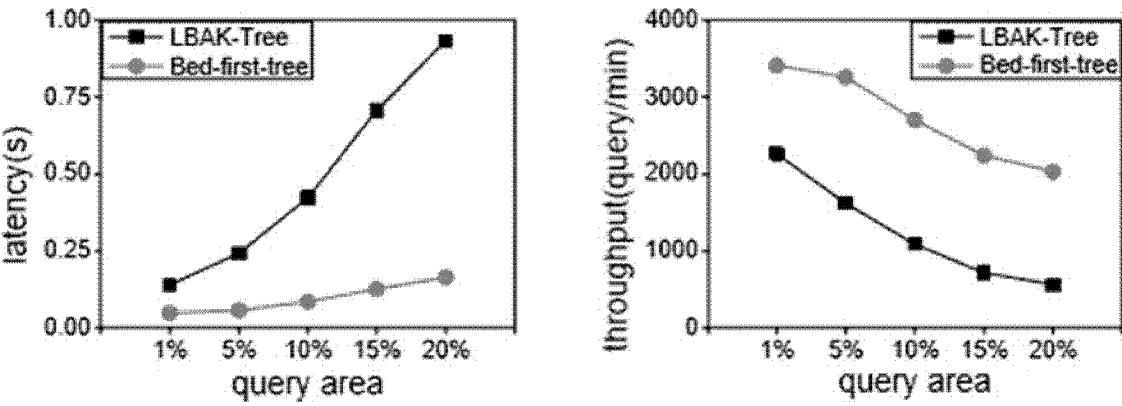


图 2

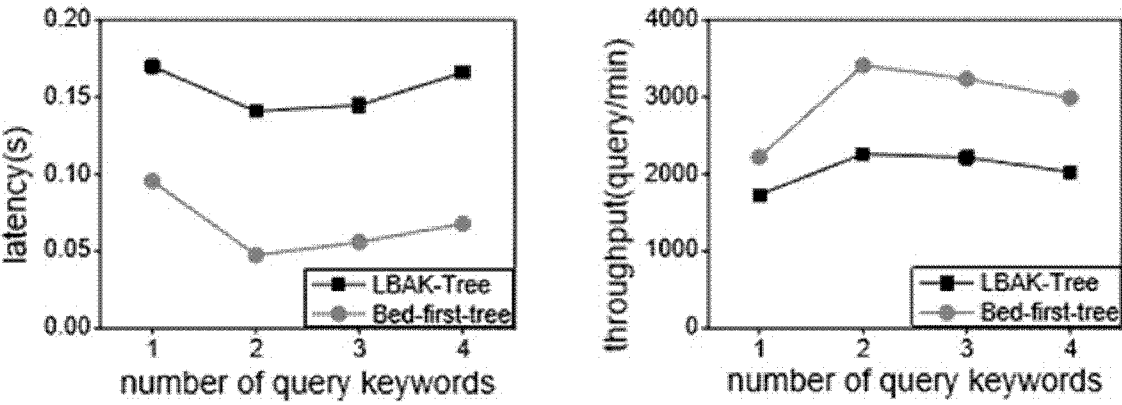


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/088771

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/29(2019.01)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)

CNABS; CNTXT; VEN; USTXT; WOTXT; EPTXT; CNKI: 电子地图, 空间, 关键字, 查询, 近似, 搜索, 索引, 频率, 边界, 最近邻, 剪枝, electronic map, space, keyword, query, similar, search, index, frequency, border, nearest neighbor, pruning

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104376112 A (SOOCHOW UNIVERSITY) 25 February 2015 (2015-02-25) description, paragraphs 6-13, 33-36, 40-45, 49-52, 56, and 57	1-3
A	CN 102214215 A (CHEN, Jiaheng) 12 October 2011 (2011-10-12) description, paragraphs 8-100	1-3
A	US 2009037403 A1 (MICROSOFT CORPORATION) 05 February 2009 (2009-02-05) entire document	1-3



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“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)

“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

“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

“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

“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

“&” document member of the same patent family

Date of the actual completion of the international search

11 December 2019

Date of mailing of the international search report

07 January 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Authorized officer

Facsimile No. (86-10)62019451

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/088771

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	104376112	A	25 February 2015	CN	104376112	B	14 September 2018
CN	102214215	A	12 October 2011	None			
US	2009037403	A1	05 February 2009	WO	2009017926	A3	19 March 2009
				WO	2009017926	A2	05 February 2009

国际检索报告

国际申请号

PCT/CN2019/088771

A. 主题的分类

G06F 16/29(2019.01)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))

CNABS;CNTXT;VEN;USTXT;WOTXT;EPTXT;CNKI: 电子地图, 空间, 关键字, 查询, 近似, 搜索, 索引, 频率, 边界, 最近邻, 剪枝, electronic map, space, keyword, query, similar, search, index, frequency, border, nearest neighbor, pruning

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
A	CN 104376112 A (苏州大学) 2015年 2月 25日 (2015 - 02 - 25) 说明书第6-13、33-36、40-45、49-52、56-57段	1-3
A	CN 102214215 A (陆嘉恒) 2011年 10月 12日 (2011 - 10 - 12) 说明书第8-100段	1-3
A	US 2009037403 A1 (MICROSOFT CORPORATION) 2009年 2月 5日 (2009 - 02 - 05) 全文	1-3

☐ 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

2019年 12月 11日

国际检索报告邮寄日期

2020年 1月 7日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

授权官员

邢丽超

电话号码 86-(20)-28950382

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/088771

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	104376112	A	2015年 2月 25日	CN	104376112	B	2018年 9月 14日
CN	102214215	A	2011年 10月 12日	无			
US	2009037403	A1	2009年 2月 5日	WO	2009017926	A3	2009年 3月 19日
				WO	2009017926	A2	2009年 2月 5日