

(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
2018 年 2 月 22 日 (22.02.2018)



(10) 国际公布号
WO 2018/032251 A1

(51) 国际专利分类号:
G06F 17/30 (2006.01)

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(21) 国际申请号: PCT/CN2016/095300

(22) 国际申请日: 2016 年 8 月 15 日 (15.08.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(72) 发明人: 及

(71) 申请人: 马 岩 (MA, Yan) [CN/CN]; 中国 广东省 深圳市 南山区 华侨城 假日湾 3 栋 3F, Guangdong 518000 (CN)。

(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, 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))。

(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,

(54) Title: METHOD AND SYSTEM FOR APPLYING SECURITY LEVEL TO DATA FETCHING OF BIG DATA

(54) 发明名称: 安全等级在大数据抓取中的应用方法及系统

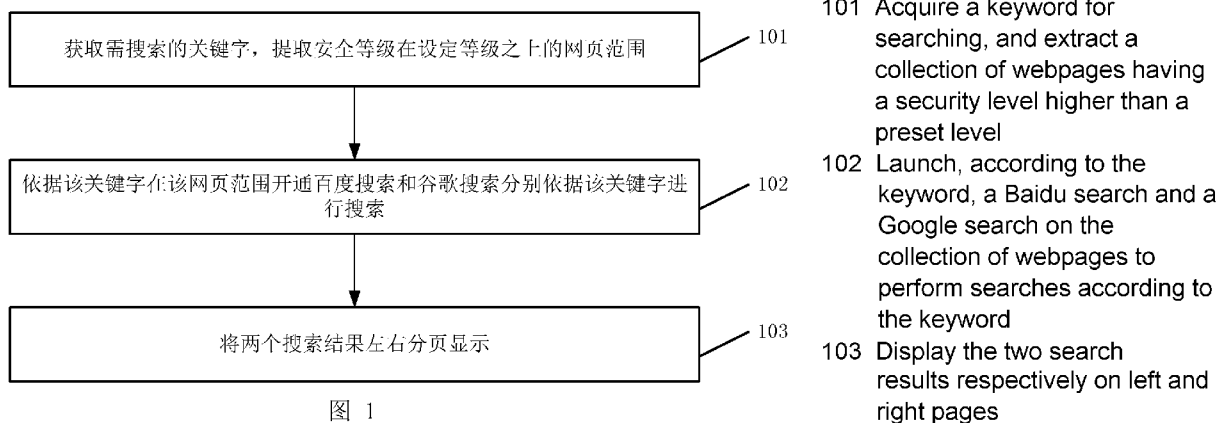


图 1

(57) Abstract: A method and system for applying a security level to data fetching of big data. The method comprises the following steps: acquiring a keyword for searching, and extracting a collection of webpages having a security level higher than a preset level (101); launching, according to the keyword, a Baidu search and a Google search on the collection of webpages to perform searches according to the keyword (102); and displaying the two search results respectively on left and right pages (103). The method of the present invention has the advantage of search effectiveness.

(57) 摘要: 一种安全等级在大数据抓取中的应用方法, 所述方法包括如下步骤: 获取需搜索的关键词, 提取安全等级在设定等级之上的网页范围(101); 依据该关键词在该网页范围开通百度搜索和谷歌搜索, 并分别依据该关键词进行搜索(102); 将两个搜索结果左右分页显示(103)。上述方法具有有效搜索的优点。



WO 2018/032251 A1

说明书

发明名称：安全等级在大数据抓取中的应用方法及系统

技术领域

- [1] 本发明涉及大数据领域，尤其涉及一种安全等级在大数据抓取中的应用方法及系统。

背景技术

- [2] 大数据（big data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产，现有的大数据有多样，例如网页数据，如何从网页数据中搜索出想要的数据是一个很值得研究的问题，现有的技术方案无法实现网页数据的有效搜索。

对发明的公开

技术问题

- [3] 本申请提供一种安全等级在大数据抓取中的应用方法。其解决现有技术的技术方案无法实现网页数据的有效搜索的缺点。

问题的解决方案

技术解决方案

- [4] 一方面，提供一种安全等级在大数据抓取中的应用方法，所述方法包括如下步骤：
- [5] 获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
- [6] 依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
- [7] 将两个搜索结果左右分页显示。
- [8] 可选的，所述方法还包括：
- [9] 如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。
- [10] 可选的，所述方法还包括：

- [11] 将推广的网页屏蔽。
- [12] 第二方面，提供一种安全等级在大数据抓取中的应用系统，所述系统包括：
- [13] 获取单元，用于获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
- [14] 搜索单元，用于依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
- [15] 分页单元，用于将两个搜索结果左右分页显示。
- [16] 可选的，所述系统还包括：
- [17] 分配单元，用于如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。
- [18] 可选的，所述系统还包括：
- [19] 屏蔽单元，用于将推广的网页屏蔽。

发明的有益效果

有益效果

- [20] 本发明提供的技术方案获取需搜索的关键字，依据该关键字开通百度搜索和谷歌搜索分别依据该关键字进行搜索，将两个搜索结果左右分页显示，所以其具有有效搜索的优点。

对附图的简要说明

附图说明

- [21] 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。
- [22] 图1为本发明第一较佳实施方式提供的一种安全等级在大数据抓取中的应用方法的流程图；
- [23] 图2为本发明第二较佳实施方式提供的一种安全等级在大数据抓取中的应用系统的结构图。

发明实施例

本发明的实施方式

- [24] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。
- [25] 请参考图1，图1是本发明第一较佳实施方式提出的一种安全等级在大数据抓取中的应用方法，该方法如图1所示，包括如下步骤：
- [26] 步骤S101、获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
- [27] 步骤S102、依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
- [28] 步骤S103、将两个搜索结果左右分页显示。
- [29] 本发明提供的技术方案获取需搜索的关键字，依据该关键字开通百度搜索和谷歌搜索分别依据该关键字进行搜索，将两个搜索结果左右分页显示，所以其具有有效搜索的优点。
- [30] 可选的，上述方法在步骤S103之后还可以包括：
- [31] 如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。
- [32] 可选的，上述方法在步骤S103之后还可以包括：
- [33] 将推广的网页屏蔽。
- [34] 请参考图2，图2是本发明第二较佳实施方式提出的一种安全等级在大数据抓取中的应用系统，该系统包括：
- [35] 获取单元201，用于获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
- [36] 搜索单元202，用于依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
- [37] 分页单元203，用于将两个搜索结果左右分页显示。
- [38] 本发明提供的技术方案获取需搜索的关键字，依据该关键字开通百度搜索和谷

歌搜索分别依据该关键字进行搜索，将两个搜索结果左右分页显示，所以其具有有效搜索的优点。

[39] 可选的，上述系统还可以包括：

[40] 分配单元204，用于如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。

[41] 可选的，上述系统还可以包括：

[42] 屏蔽单元205，用于将推广的网页屏蔽。

[43] 需要说明的是，对于前述的各个方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某一些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[44] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详细描述的部分，可以参见其他实施例的相关描述。

[45] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：闪存盘、只读存储器（英文：Read-Only Memory，简称：ROM）、随机存取器（英文：Random Access Memory，简称：RAM）、磁盘或光盘等。

[46] 以上对本发明实施例所提供的内容下载方法及相关设备、系统进行了详细介绍，本文中应用了具体个例对本发明的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本发明的方法及其核心思想；同时，对于本领域的一般技术人员，依据本发明的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本发明的限制。

权利要求书

- [权利要求 1] 一种安全等级在大数据抓取中的应用方法，其特征在于，所述方法包括如下步骤：
获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
将两个搜索结果左右分页显示。
- [权利要求 2] 根据权利要求1所述的方法，其特征在于，所述方法还包括：
如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。
- [权利要求 3] 根据权利要求1所述的方法，其特征在于，所述方法还包括：
将推广的网页屏蔽。
- [权利要求 4] 一种安全等级在大数据抓取中的应用系统，其特征在于，所述系统包括：
获取单元，用于获取需搜索的关键字，提取安全等级在设定等级之上的网页范围；
搜索单元，用于依据该关键字在该网页范围开通百度搜索和谷歌搜索分别依据该关键字进行搜索；
分页单元，用于将两个搜索结果左右分页显示。
- [权利要求 5] 根据权利要求4所述的系统，其特征在于，所述系统还包括：
分配单元，用于如两个搜索结果中具有相同的搜索结果，将相同的搜索结果在任一个分页显示。
- [权利要求 6] 根据权利要求4所述的系统，其特征在于，所述系统还包括：
屏蔽单元，用于将推广的网页屏蔽。

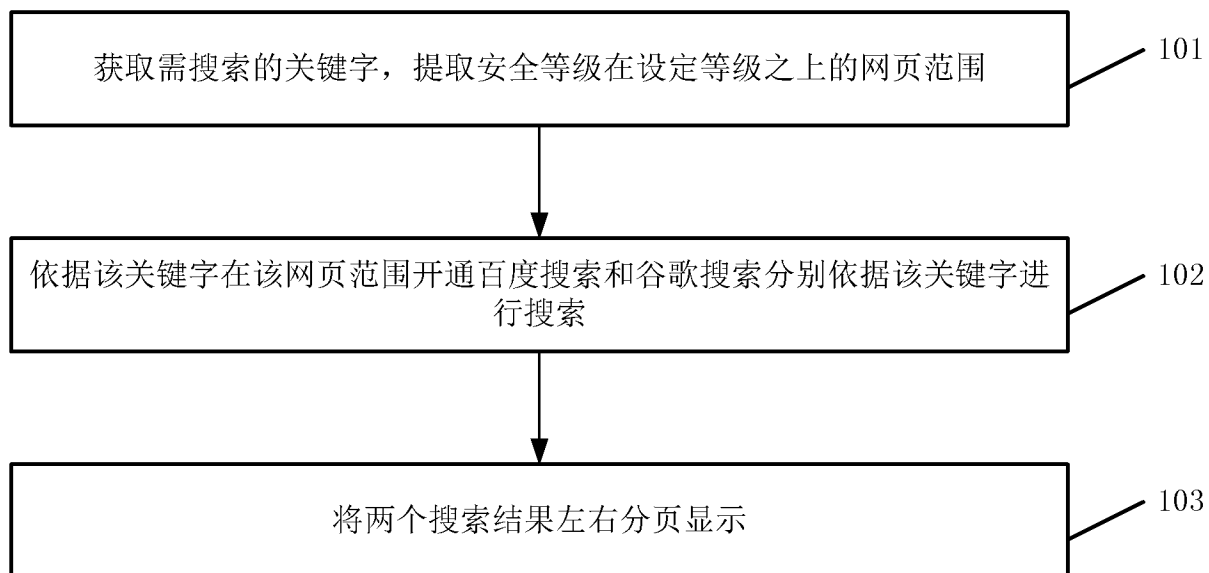


图 1

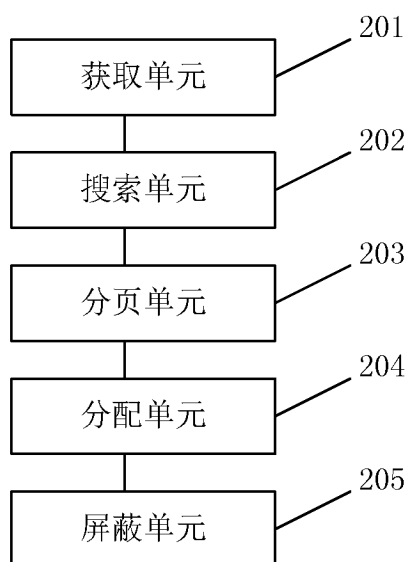


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/095300

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: 马岩, 网页, 搜索, 查询, 查找, 检索, 抓取, 爬取, 爬虫, 谷歌, 百度, 搜狗, 多平台, 多搜索引擎, 导航, 关键词, 关键字, 分页, 左屏, 右屏, 分屏, 左右, 安全等级, html, web, webpage, search+, hunt, capture, baidu, google, multi+, platform?, engine?, keyword?, left, right, page, pagination, windowing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
E	CN 106227888 A (MA, Yan) 14 December 2016 (14.12.2016), claims 1-6	1-6
X	CN 105683966 A (SHENZHEN BOXINNUODA TRADING CONSULTATION CO., LTD.) 15 June 2016 (15.06.2016), description, paragraphs [0025]-[0029]	1, 2, 4, 5
Y	CN 105683966 A (SHENZHEN BOXINNUODA TRADING CONSULTATION CO., LTD.) 15 June 2016 (15.06.2016), description, paragraphs [0025]-[0029]	3, 6
Y	CN 105849730 A (MA, Yan) 10 August 2016 (10.08.2016), description, paragraphs [0025]-[0033]	3, 6
A	CN 102004782 A (BEIJING SOGOU TECHNOLOGY DEVELOPMENT CO., LTD.) 06 April 2011 (06.04.2011), entire document	1-6

☒ 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 May 2017	Date of mailing of the international search report 09 June 2017
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 TANG, Yingyan Telephone No. (86-10) 62414436

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/095300

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105117476 A (LIU, Minkai) 02 December 2015 (02.12.2015), entire document	1-6

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/095300

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 106227888 A	14 December 2016	None	
CN 105683966 A	15 June 2016	None	
CN 105849730 A	10 August 2016	None	
CN 102004782 A	06 April 2011	None	
CN 105117476 A	02 December 2015	None	

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, CNKI, WPI, EPODOC, IEEE: 马岩, 网页, 搜索, 查询, 查找, 检索, 抓取, 爬取, 爬虫, 谷歌, 百度, 搜狗, 多平台, 多搜索引擎, 导航, 关键词, 关键字, 分页, 左屏, 右屏, 分屏, 左右, 安全等级, html, web, webpage, search+, hunt, capture, baidu, google, multi+, platform?, engine?, keyword?, left, right, page, pagination, windowing																							
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>E</td> <td>CN 106227888 A (马岩) 2016年 12月 14日 (2016 - 12 - 14) 权利要求1-6</td> <td>1-6</td> </tr> <tr> <td>X</td> <td>CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段</td> <td>1-2, 4-5</td> </tr> <tr> <td>Y</td> <td>CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段</td> <td>3, 6</td> </tr> <tr> <td>Y</td> <td>CN 105849730 A (马岩) 2016年 8月 10日 (2016 - 08 - 10) 说明书第[0025]-[0033]段</td> <td>3, 6</td> </tr> <tr> <td>A</td> <td>CN 102004782 A (北京搜狗科技发展有限公司) 2011年 4月 6日 (2011 - 04 - 06) 全文</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>CN 105117476 A (刘珉恺) 2015年 12月 2日 (2015 - 12 - 02) 全文</td> <td>1-6</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	E	CN 106227888 A (马岩) 2016年 12月 14日 (2016 - 12 - 14) 权利要求1-6	1-6	X	CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段	1-2, 4-5	Y	CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段	3, 6	Y	CN 105849730 A (马岩) 2016年 8月 10日 (2016 - 08 - 10) 说明书第[0025]-[0033]段	3, 6	A	CN 102004782 A (北京搜狗科技发展有限公司) 2011年 4月 6日 (2011 - 04 - 06) 全文	1-6	A	CN 105117476 A (刘珉恺) 2015年 12月 2日 (2015 - 12 - 02) 全文	1-6
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
E	CN 106227888 A (马岩) 2016年 12月 14日 (2016 - 12 - 14) 权利要求1-6	1-6																					
X	CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段	1-2, 4-5																					
Y	CN 105683966 A (深圳市博信诺达经贸咨询有限公司) 2016年 6月 15日 (2016 - 06 - 15) 说明书第[0025]-[0029]段	3, 6																					
Y	CN 105849730 A (马岩) 2016年 8月 10日 (2016 - 08 - 10) 说明书第[0025]-[0033]段	3, 6																					
A	CN 102004782 A (北京搜狗科技发展有限公司) 2011年 4月 6日 (2011 - 04 - 06) 全文	1-6																					
A	CN 105117476 A (刘珉恺) 2015年 12月 2日 (2015 - 12 - 02) 全文	1-6																					
☐ 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 2017年 5月 3日		国际检索报告邮寄日期 2017年 6月 9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) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10) 62019451		受权官员 唐楹琰 电话号码 (86-10) 62414436																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/095300

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	106227888	A	2016年 12月 14日	无	
CN	105683966	A	2016年 6月 15日	无	
CN	105849730	A	2016年 8月 10日	无	
CN	102004782	A	2011年 4月 6日	无	
CN	105117476	A	2015年 12月 2日	无	