

# (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  
2017年7月13日 (13.07.2017)

WIPO | PCT

(10) 国际公布号  
WO 2017/117805 A1

- (51) 国际专利分类号:  
G06F 17/30 (2006.01) G06F 11/34 (2006.01)
- (21) 国际申请号: PCT/CN2016/070499
- (22) 国际申请日: 2016年1月8日 (08.01.2016)
- (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)。
- (74) 代理人: 深圳市科冠知识产权代理有限公司 (SHENZHEN KEGUAN INTELLECTUAL PROPERTY AGENCY CO., LTD); 中国广东省深圳市南山区南海大道东华园 5 栋 303, Guangdong 518000 (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。

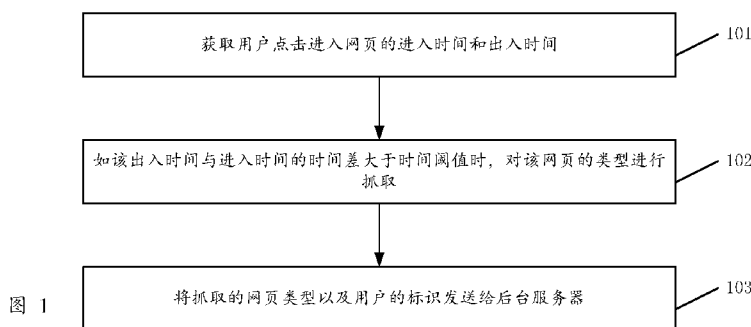
- (84) 指定国 (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), 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))。

(54) Title: WEB INFORMATION CAPTURING METHOD AND SYSTEM

(54) 发明名称: 网络信息的抓取方法及系统



101 ACQUIRE AN ACCESS TIME AND AN EXIT TIME AFTER A USER CLICKS TO ACCESS A WEBPAGE  
 102 IF A TIME DIFFERENCE BETWEEN THE EXIT TIME AND THE ACCESS TIME IS GREATER THAN A DURATION THRESHOLD, CAPTURE THE TYPE OF THE WEBPAGE  
 103 SEND THE CAPTURED WEBPAGE TYPE AND AN IDENTIFIER OF THE USER TO A BACKEND SERVER

(57) Abstract: A web information capturing method and system. The method comprises the following steps: acquiring an access time and an exit time after a user clicks to access a webpage (101); if a time difference between the exit time and the access time is greater than a duration threshold, capturing the type of the webpage (102); and sending the captured webpage type and an identifier of the user to a backend server (103). The method and system have an advantage of capturing web information accurately.

(57) 摘要: 一种网络信息的抓取方法及系统, 所述方法包括如下步骤: 获取用户点击进入网页的进入时间和出入时间 (101); 如该出入时间与进入时间的的时间差大于时间阈值时, 对该网页的类型进行抓取 (102); 将抓取的网页类型以及用户的标识发送给后台服务器 (103)。所述方法及系统具有网络信息抓取准确的优点。



WO 2017/117805 A1

## 网络信息的抓取方法及系统

### 技术领域

- [1] 本发明涉及互联网领域，尤其涉及一种网络信息的抓取方法及系统。

### 背景技术

- [2] 网络是由节点和连线构成，表示诸多对象及其相互联系。在数学上，网络是一种图，一般认为专指加权图。网络除了数学定义外，还有具体的物理含义，即网络是从某种相同类型的实际问题中抽象出来的模型。在计算机领域中，网络是信息传输、接收、共享的虚拟平台，通过它把各个点、面、体的信息联系在一起，从而实现这些资源的共享，网络是人类发展史来最重要的发明，提高了科技和人类社会的发展。
- [3] 现有的网络信息是海量的，而每个用户又是特定的人物，所以对于此种情况来说，网络信息的抓取非常重要，现有的网络信息的抓取没有任何的筛选，所以其网络信息抓取的效果不好。

### 对发明的公开

### 技术问题

- [4] 本申请提供一种网络信息的抓取方法。其解决现有技术的技术方案网络信息抓取的效果不好的缺点。

### 问题的解决方案

### 技术解决方案

- [5] 一方面，提供一种网络信息的抓取方法，所述方法包括如下步骤：
- [6] 获取用户点击进入网页的进入时间和出入时间；
- [7] 如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；
- [8] 将抓取的网页类型以及用户的标识发送给后台服务器。
- [9] 可选的，所述方法还包括：
- [10] 如该时间差小于时间阈值，不对该网络的类型进行抓取。

- [11] 可选的，所述方法还包括：
- [12] 如该网页的点击次数超过次数阈值，则对该网络的类型进行抓取。
- [13] 第二方面，提供一种网络信息的抓取系统，所述系统包括：
- [14] 获取单元，用于获取用户点击进入网页的进入时间和出入时间；
- [15] 抓取单元，用于如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；
- [16] 发送单元，用于将抓取的网页类型以及用户的标识发送给后台服务器。
- [17] 可选的，所述系统还包括：
- [18] 放弃单元，用于如该时间差小于时间阈值，不对该网络的类型进行抓取。
- [19] 可选的，所述抓取单元，还用于如该网页的点击次数超过次数阈值，则对该网络的类型进行抓取。

## 发明的有益效果

### 有益效果

- [20] 本发明提供的技术方案能够依据用户点击网页的时间来确定是否抓取用户的网络信息，其对网络信息的筛选具有积极的作用，所以其具有网络信息抓取效果好的优点。

## 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网络信息的抓取方法的流程图；
- [23] 图2为本发明第二较佳实施方式提供的一种网络信息的抓取系统的结构图。

## 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网络信息的抓取方法，该方法如图1所示，包括如下步骤：

[26] 步骤S101、获取用户点击进入网页的进入时间和出入时间；

[27] 步骤S102、如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；

[28] 步骤S103、将抓取的网页类型以及用户的标识发送给后台服务器。

[29] 本发明提供的技术方案能够依据用户点击网页的时间来确定是否抓取用户的网络信息，其对网络信息的筛选具有积极的作用，所以其具有网络信息抓取效果好的优点。

[30] 可选的，上述方法在步骤S103之后还可以包括：

[31] 如该时间差小于时间阈值，不对该网络类型进行抓取。

[32] 可选的，上述方法在步骤S103之后还可以包括：

[33] 如该网页的点击次数超过次数阈值，则对该网络类型进行抓取。

[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网络信息的抓取系统，该系统如图2所示，包括：

[35] 获取单元201，用于获取用户点击进入网页的进入时间和出入时间；

[36] 抓取单元202，用于如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；

[37] 发送单元203，用于将抓取的网页类型以及用户的标识发送给后台服务器。

[38] 本发明提供的技术方案能够依据用户点击网页的时间来确定是否抓取用户的网络信息，其对网络信息的筛选具有积极的作用，所以其具有网络信息抓取效果好的优点。

[39] 可选的，上述系统还可以包括：

[40] 放弃单元204，用于如该时间差小于时间阈值，不对该网络类型进行抓取。

[41] 可选的，上述抓取单元202，还用于如该网页的点击次数超过次数阈值，则对

该网络的类型进行抓取。

- [42] 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。
- [43] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中沒有详细描述的部分，可以参见其他实施例的相关描述。
- [44] 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过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。
- [45] 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] 一种网络信息的抓取方法，其特征在于，所述方法包括如下步骤：  
获取用户点击进入网页的进入时间和出入时间；  
如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；  
将抓取的网页类型以及用户的标识发送给后台服务器。
- [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：  
如该时间差小于时间阈值，不对该网络类型进行抓取。
- [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：  
如该网页的点击次数超过次数阈值，则对该网络类型进行抓取。
- [权利要求 4] 一种网络信息的抓取系统，其特征在于，所述系统包括：  
获取单元，用于获取用户点击进入网页的进入时间和出入时间；  
抓取单元，用于如该出入时间与进入时间的时间差大于时间阈值时，对该网页的类型进行抓取；  
发送单元，用于将抓取的网页类型以及用户的标识发送给后台服务器。
- [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：  
放弃单元，用于如该时间差小于时间阈值，不对该网络类型进行抓取。
- [权利要求 6] 根据权利要求4所述的系统，其特征在于，  
所述抓取单元，还用于如该网页的点击次数超过次数阈值，则对该网络类型进行抓取。

# 说明书附图

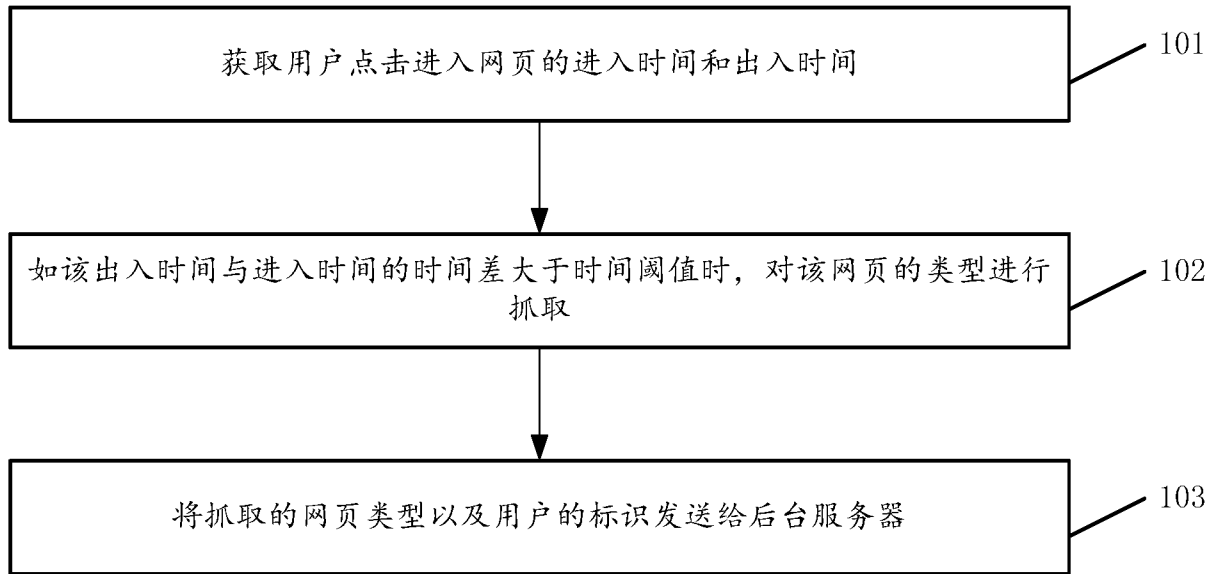


图 1

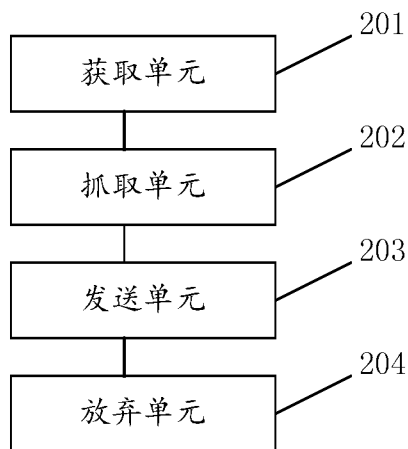


图 2

# INTERNATIONAL SEARCH REPORT

International application No.

**PCT/CN2016/070499**

## A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/30 (2006.01) i; G06F 11/34 (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)

WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: network, page, webpage, grab+, grasp+, time, brows+, user

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104199874 A (HARBIN ENGINEERING UNIVERSITY), 10 December 2014 (10.12.2014), description, paragraphs [0022]-[0042]	1-6
A	CN 104700289 A (CHINA UNITED NETWORK COMMUNICATIONS CORPORATION LIMITED), 10 June 2015 (10.06.2015), the whole document	1-6
A	JP 2004280501 A (NRI & NCC CO., LTD.), 07 October 2004 (07.10.2004), the whole document	1-6
A	CN 102347930 A (CHINA TELECOM CORPORATION LIMITED), 08 February 2012 (08.02.2012), the whole document	1-6
A	CN 102495874 A (JIANGSU SHIDEWEI NETWORK TECHNOLOGY CO., LTD.), 13 June 2012 (13.06.2012), the whole document	1-6

Further documents are listed in the continuation of Box C.

See patent family annex.

<p>* Special categories of cited documents:</p> <p>“A” document defining the general state of the art which is not considered to be of particular relevance</p> <p>“E” earlier application or patent but published on or after the international filing date</p> <p>“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)</p> <p>“O” document referring to an oral disclosure, use, exhibition or other means</p> <p>“P” document published prior to the international filing date but later than the priority date claimed</p>	<p>“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</p> <p>“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</p> <p>“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</p> <p>“&amp;” document member of the same patent family</p>
---	---

Date of the actual completion of the international search  
13 September 2016 (13.09.2016)

Date of mailing of the international search report  
**28 September 2016 (28.09.2016)**

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  
**ZHAO, Ting**  
Telephone No.: (86-10) **62414434**

**INTERNATIONAL SEARCH REPORT**  
Information on patent family members

International application No.

**PCT/CN2016/070499**

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104199874 A	10 December 2014	None	
CN 104700289 A	10 June 2015	None	
JP 2004280501 A	07 October 2004	None	
CN 102347930 A	08 February 2012	None	
CN 102495874 A	13 June 2012	None	

<p>A. 主题的分类</p> <p>G06F 17/30(2006.01)i; G06F 11/34(2006.01)i</p> <p>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</p>																				
<p>B. 检索领域</p> <p>检索的最低限度文献(标明分类系统和分类号)</p> <p>G06F</p> <p>包含在检索领域中的除最低限度文献以外的检索文献</p> <p>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))</p> <p>WPI, EPODOC, CNPAT, CNKI, IEEE, GOOGLE: 网络, 网页, 抓取, 时间, 浏览, 用户, network, page, webpage, grab+, grasp+, time, brows+, user</p>																				
<p>C. 相关文件</p> <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 104199874 A (哈尔滨工程大学) 2014年 12月 10日 (2014 - 12 - 10) 说明书第[0022]-[0042]段</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>CN 104700289 A (中国联合网络通信集团有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>JP 2004280501 A (NRI &amp; NCC CO., LTD.) 2004年 10月 7日 (2004 - 10 - 07) 全文</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>CN 102347930 A (中国电信股份有限公司) 2012年 2月 8日 (2012 - 02 - 08) 全文</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>CN 102495874 A (江苏仕德伟网络科技股份有限公司) 2012年 6月 13日 (2012 - 06 - 13) 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	X	CN 104199874 A (哈尔滨工程大学) 2014年 12月 10日 (2014 - 12 - 10) 说明书第[0022]-[0042]段	1-6	A	CN 104700289 A (中国联合网络通信集团有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文	1-6	A	JP 2004280501 A (NRI & NCC CO., LTD.) 2004年 10月 7日 (2004 - 10 - 07) 全文	1-6	A	CN 102347930 A (中国电信股份有限公司) 2012年 2月 8日 (2012 - 02 - 08) 全文	1-6	A	CN 102495874 A (江苏仕德伟网络科技股份有限公司) 2012年 6月 13日 (2012 - 06 - 13) 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																		
X	CN 104199874 A (哈尔滨工程大学) 2014年 12月 10日 (2014 - 12 - 10) 说明书第[0022]-[0042]段	1-6																		
A	CN 104700289 A (中国联合网络通信集团有限公司) 2015年 6月 10日 (2015 - 06 - 10) 全文	1-6																		
A	JP 2004280501 A (NRI & NCC CO., LTD.) 2004年 10月 7日 (2004 - 10 - 07) 全文	1-6																		
A	CN 102347930 A (中国电信股份有限公司) 2012年 2月 8日 (2012 - 02 - 08) 全文	1-6																		
A	CN 102495874 A (江苏仕德伟网络科技股份有限公司) 2012年 6月 13日 (2012 - 06 - 13) 全文	1-6																		
<p><input type="checkbox"/> 其余文件在C栏的续页中列出。</p> <p><input checked="" type="checkbox"/> 见同族专利附件。</p>																				
<p>* 引用文件的具体类型:</p> <p>“A” 认为不特别相关的表示了现有技术一般状态的文件</p> <p>“E” 在国际申请日的当天或之后公布的在先申请或专利</p> <p>“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)</p> <p>“O” 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>“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</p> <p>“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</p> <p>“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</p> <p>“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</p> <p>“&amp;” 同族专利的文件</p>																				
<p>国际检索实际完成的日期</p> <p>2016年 9月 13日</p>		<p>国际检索报告邮寄日期</p> <p>2016年 9月 28日</p>																		
<p>ISA/CN的名称和邮寄地址</p> <p>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</p> <p>传真号 (86-10)62019451</p>		<p>授权官员</p> <p>赵婷</p> <p>电话号码 (86-10)62414434</p>																		

国际检索报告  
关于同族专利的信息

国际申请号

PCT/CN2016/070499

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	104199874	A	2014年 12月 10日	无	
CN	104700289	A	2015年 6月 10日	无	
JP	2004280501	A	2004年 10月 7日	无	
CN	102347930	A	2012年 2月 8日	无	
CN	102495874	A	2012年 6月 13日	无	