(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 2017/117782 A1

(51) 国际专利分类号:
*G06F 17/27* (2006.01)

(21) 国际申请号: PCT/CN2016/070406

(22) 国际申请日: 2016 年 1 月 7 日 (07.01.2016)

(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: NETWORK INFORMATION WORD SEGMENTATION PROCESSING METHOD AND SYSTEM

(54) 发明名称 : 网络信息的分词处理方法及系统

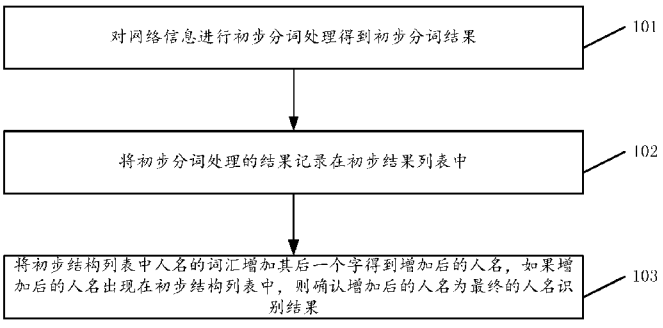


图 1

101 PERFORM A PRELIMINARY WORD SEGMENTATION PROCESSING ON NETWORK INFORMATION TO ACQUIRE A PRELIMINARY WORD SEGMENTATION RESULT

102 RECORD THE PRELIMINARY WORD SEGMENTATION PROCESSING RESULT IN A PRELIMINARY RESULT LIST

103 ADD A SUBSEQUENT WORD TO A NAME WORD LIST IN A PRELIMINARY STRUCTURE LIST TO ACQUIRE A COMBINED NAME WORD, AND IF THE COMBINED NAME WORD APPEARS IN THE PRELIMINARY STRUCTURE LIST, THEN DETERMINE THE COMBINED NAME WORD TO BE A FINAL NAME WORD IDENTIFICATION RESULT

(57) Abstract: A network information word segmentation processing method and system, the method comprising the following steps: performing a preliminary word segmentation processing on network information to acquire a preliminary word segmentation result (101); recording the preliminary word segmentation processing result in a preliminary result list (102); adding a subsequent word to a name word in a preliminary structure list to acquire a combined name word, and if the combined name word appears in the preliminary structure list, then determining the combined name word to be a final name word identification result (103). The method has a favorable word segmentation effect.

(57) 摘要: 一种网络信息的分词处理方法及系统，所述方法包括如下步骤：对网络信息进行初步分词处理得到初步分词结果（101）；将初步分词处理的结果记录在初步结果列表中（102）；将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果（103）。该方法具有分词效果好的优点。

WO 2017/117782 A1

# 网络信息的分词处理方法及系统

## 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分词处理方法及系统。

## 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、体的信息联系到一起，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] 现有的分词处理的方法对词汇的处理一般都是通过比对或符号等方式来处理，此方式对于正常的词汇处理没有问题，但是对于人名的处理来说，因为人名没有任何的特性，所以其处理会不准确。

## 对发明的公开

### 技术问题

[4] 本申请提供一种网络信息的分词处理方法。其解决现有技术的技术方案对人名识别不准确的缺点。

## 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分词处理方法，所述方法包括如下步骤：

[6] 对网络信息进行初步分词处理得到初步分词结果；

[7] 将初步分词处理的结果记录在初步结果列表中；

[8] 将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[9] 可选的，所述方法还包括：

[10] 将最终的人名识别结果替换初步结构列表中的人名的词汇。

[11] 可选的，所述方法还包括：

[12] 如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

[13] 第二方面，提供一种网络信息的分词处理系统，所述系统包括：

[14] 分词单元，用于对网络信息进行初步分词处理得到初步分词结果；

[15] 记录单元，用于将初步分词处理的结果记录在初步结果列表中；

[16] 校验单元，用于将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[17] 可选的，所述系统还包括：

[18] 更新单元，用于将最终的人名识别结果替换初步结构列表中的人名的词汇。

[19] 可选的，所述校验单元，还用于如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

## 发明的有益效果

### 有益效果

[20] 本发明提供的技术方案对网络信息进行初步分词处理后，将特定数量的词汇增加后面一个字后再次比对，从来达到对人名进行有效识别的优点。

## 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分词处理方法的流程图；

[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分词处理系统的结构图。

# 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分词处理方法，该方法如图1所示，包括如下步骤：

[26] 步骤S101、对网络信息进行初步分词处理得到初步分词结果；

[27] 上述步骤中的初步分词处理的方式可以有多种，例如百度分词处理方法，当然也可以为其他的现有技术的方法来进行初步分词处理。

[28] 步骤S102、将初步分词处理的结果记录在初步结果列表中；

[29] 步骤S103、将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[30] 本发明提供的技术方案对网络信息进行初步分词处理后，将特定数量的词汇增加后面一个字后再次比对，从来达到对人名进行有效识别的优点。

[31] 可选的，上述方法在步骤S103之后还可以包括：

[32] 将最终的人名识别结果替换初步结构列表中的人名的词汇。

[33] 可选的，上述方法步骤S103之后还可以包括：

[34] 如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

[35] 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分词处理系统，该系统包括：

[36] 分词单元201，用于对网络信息进行初步分词处理得到初步分词结果；

[37] 上述分词单元201中的初步分词处理的方式可以有多种，例如百度分词处理方法，当然也可以为其他的现有技术的方法来进行初步分词处理。

[38] 记录单元202，用于将初步分词处理的结果记录在初步结果列表中；

[39] 校验单元203，用于将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[40] 本发明提供的技术方案对网络信息进行初步分词处理后，将特定数量的词汇增加后面一个字后再次比对，从来达到对人名进行有效识别的优点。

[41] 可选的，上述系统还可以包括：

[42] 更新单元204，用于将最终的人名识别结果替换初步结构列表中的人名的词汇。

[43] 可选的，上述校验单元203，还用于如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

[44] 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。

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

[46] 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。

[47] 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] 一种网络信息的分词处理方法，其特征在于，所述方法包括如下步骤:

对网络信息进行初步分词处理得到初步分词结果;

将初步分词处理的结果记录在初步结果列表中;

将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[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:

将最终的人名识别结果替换初步结构列表中的人名的词汇。

[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:

如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

[权利要求 4] 一种网络信息的分词处理系统，其特征在于，所述系统包括:

分词单元，用于对网络信息进行初步分词处理得到初步分词结果;

记录单元，用于将初步分词处理的结果记录在初步结果列表中;

校验单元，用于将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果。

[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:

更新单元，用于将最终的人名识别结果替换初步结构列表中的人名的词汇。

[权利要求 6] 根据权利要求4所述的系统，其特征在于，

所述校验单元，还用于如增加后的人名未出现在初步结构列表中，则初步结构列表中人名为最终的人名识别结果。

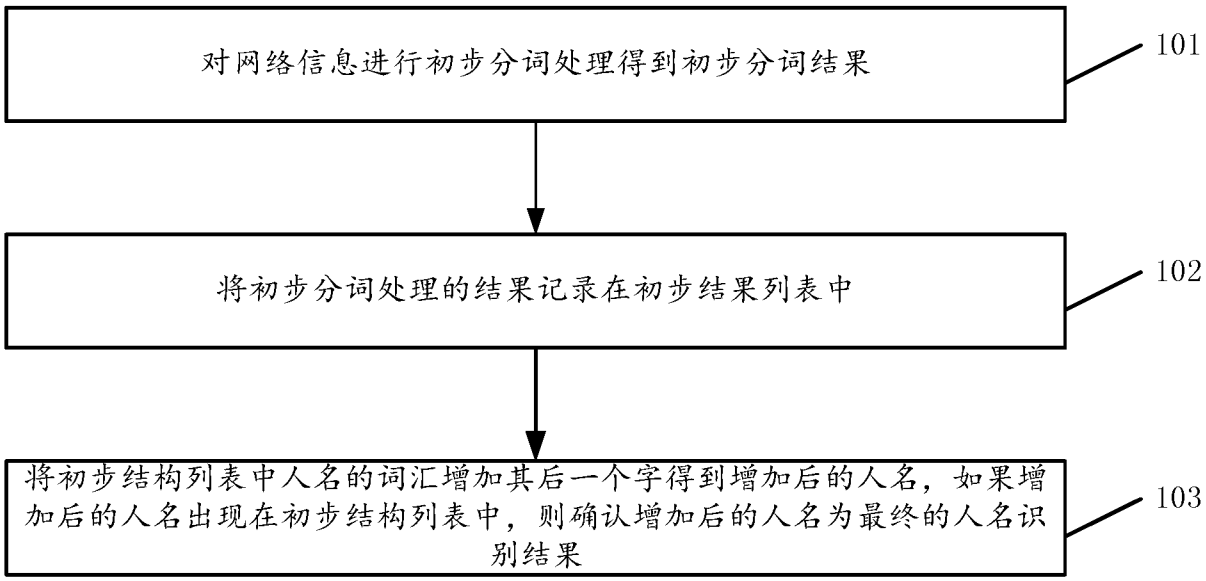
对网络信息进行初步分词处理得到初步分词结果
101
将初步分词处理的结果记录在初步结果列表中
102
将初步结构列表中人名的词汇增加其后一个字得到增加后的人名，如果增加后的人名出现在初步结构列表中，则确认增加后的人名为最终的人名识别结果
103

图 1

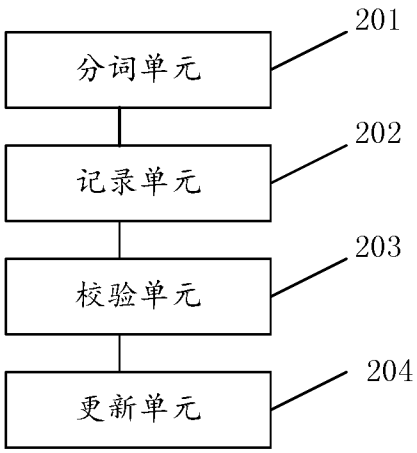
分词单元
201
记录单元
202
校验单元
203
更新单元
204

图 2

# INTERNATIONAL SEARCH REPORT

International application No.
**PCT/CN2016/070406**

**A. CLASSIFICATION OF SUBJECT MATTER**

G06F 17/27 (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; GOOGLE; CNKI: word segmentation, segmentation, name, correct, amendment, add

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | CN 104182423 A (EAST CHINA NORMAL UNIVERSITY), 03 December 2014 (03.12.2014), description, paragraphs [0004] and [0097]-[0102] | 1-6 |
| A | CN 102033879 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.), 27 April 2011 (27.04.2011), the whole document | 1-6 |
| A | CN 101950284 A (BEIJING FEINNO COMMUNICATION TECH CO., LTD.), 19 January 2011 (19.01.2011), the whole document | 1-6 |
| A | US 2007021956 A1 (QU, Yan et al.), 25 January 2007 (25.01.2007), the whole document | 1-6 |

☐ 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 | Date of mailing of the international search report |
|---|---|
| 11 July 2016 (11.07.2016) | **26 July 2016 (26.07.2016)** |
| Name and mailing address of the ISA/CN:<br>State Intellectual Property Office of the P. R. China<br>No. 6, Xitucheng Road, Jimenqiao<br>Haidian District, Beijing 100088, China<br>Facsimile No.: (86-10) 62019451 | Authorized officer<br>**XIONG, Yue**<br>Telephone No.: (86-10) **82245487** |

Form PCT/ISA/210 (second sheet) (July 2009)

# INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
**PCT/CN2016/070406**

| Patent Documents referred in the Report | Publication Date | Patent Family | Publication Date |
|---|---|---|---|
| CN 104182423 A | 03 December 2014 | None | |
| CN 102033879 A | 27 April 2011 | CN 102033879 B | 18 February 2015 |
| CN 101950284 A | 19 January 2011 | CN 101950284 B | 08 May 2013 |
| US 2007021956 A1 | 25 January 2007 | None | |

# 国际检索报告

国际申请号
PCT/CN2016/070406

A. 主题的分类

G06F 17/27(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;EPODOC;GOOGLE;CNKI:分词，人名，修正，增加，加上，segmentation,name,correct,amendment,add

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 |
|---|---|---|
| X | CN 104182423 A (华东师范大学) 2014年 12月 3日 (2014-12-03)<br>说明书第4，97-102段 | 1-6 |
| A | CN 102033879 A (腾讯科技深圳有限公司) 2011年 4月 27日 (2011-04-27)<br>全文 | 1-6 |
| A | CN 101950284 A (北京新媒传信科技有限公司) 2011年 1月 19日 (2011-01-19)<br>全文 | 1-6 |
| A | US 2007021956 A1 (QU, YAN 等) 2007年 1月 25日 (2007-01-25)<br>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 | 国际检索报告邮寄日期 |
|---|---|
| 2016年 7月 11日 | 2016年 7月 26日 |

| ISA/CN的名称和邮寄地址 | 受权官员 |
|---|---|
| 中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路6号 100088 | 熊跃 |
| 传真号 (86-10)62019451 | 电话号码 (86-10)82245487 |

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/070406

| 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 | 104182423 | A | 2014年 12月 3日 | 无 | | | |
| CN | 102033879 | A | 2011年 4月 27日 | CN | 102033879 | B | 2015年 2月 18日 |
| CN | 101950284 | A | 2011年 1月 19日 | CN | 101950284 | B | 2013年 5月 8日 |
| US | 2007021956 | A1 | 2007年 1月 25日 | 无 | | | |