

(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
2016 年 9 月 22 日 (22.09.2016)



(10) 国际公布号
WO 2016/145675 A1

(51) 国际专利分类号:
G06N 3/02 (2006.01)

(21) 国际申请号: PCT/CN2015/075472

(22) 国际申请日: 2015 年 3 月 31 日 (31.03.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201510111904.6 2015 年 3 月 13 日 (13.03.2015) CN

(71) 申请人: 中国科学院声学研究所 (INSTITUTE OF ACOUSTICS, CHINESE ACADEMY OF SCIENCES) [CN/CN]; 中国北京市海淀区北四环西路 21 号, Beijing 100190 (CN)。上海尚恩华科网络科技股份有限公司 (SHANGHAI 3NTV NETWORK TECHNOLOGY CO. LTD.) [CN/CN]; 中国上海市闸北区共和新路 1481 号 1 号楼 408 室, Shanghai 200072 (CN)。

(72) 发明人: 王劲林 (WANG, Jinlin); 中国北京市海淀区北四环西路 21 号, Beijing 100190 (CN)。尤佳莉 (YOU, Jiali); 中国北京市海淀区北四环西路 21 号,

Beijing 100190 (CN)。盛益强 (SHENG, Yiqiang); 中国北京市海淀区北四环西路 21 号, Beijing 100190 (CN)。李超鹏 (LI, Chaopeng); 中国北京市海淀区北四环西路 21 号, Beijing 100190 (CN)。

(74) 代理人: 北京方安思达知识产权代理有限公司 (FASTA INTELLECTUAL PROPERTY LIMITED); 中国北京市海淀区中关村北二条 13 号院 1 号楼 202 室, Beijing 100190 (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: BIG DATA PROCESSING METHOD FOR SEGMENT-BASED TWO-GRADE DEEP LEARNING MODEL

(54) 发明名称: 一种基于分段的两级深度学习模型的大数据处理方法

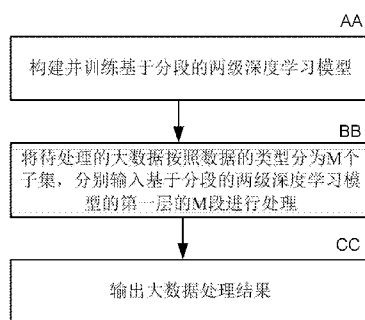


图 1

AA Constructing and training a segment-based two-grade deep learning model

BB Dividing big data to be processed into M sub-sets according to the type of the data and respectively inputting same into M segments of a first layer of the segment-based two-grade deep learning model for processing

CC Outputting a big data processing result

(57) Abstract: Provided is a big data processing method for a segment-based two-grade deep learning model. The method comprises: step 1), constructing and training a segment-based two-grade deep learning model, wherein the model is divided into two grades in a longitudinal level: a first grade and a second grade, each layer of the first grade is divided into M segments in a horizontal direction, and the weight between neuron nodes of adjacent layers in different segments of the first grade is zero; step 2), dividing big data to be processed into M sub-sets according to the type of the data and respectively inputting same into M segments of a first layer of the segment-based two-grade deep learning model for processing; and step 3), outputting a big data processing result. The segment-based two-grade deep learning model provided in the present invention can effectively reduce the scale of a model and shortens a training time of the model. The method of the present invention can increase the big data processing speed and shorten the processing time.

(57) 摘要:

[见续页]



WO 2016/145675 A1



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))。

本发明提供了一种基于分段的两级深度学习模型的大数据处理方法，所述方法包括：步骤 1)构建并训练基于分段的两级深度学习模型，该模型从纵向层次上分为两级：第一级和第二级；所述第一级的各层从横向上分为 M 段；第一级中不同段间相邻层的神经元节点间权重为 0；步骤 2)将待处理的大数据按照数据的类型分为 M 个子集，分别输入基于分段的两级深度学习模型的第一层的 M 段进行处理；步骤 3)输出大数据处理结果。本发明提出的基于分段的两级深度学习模型能够有效地减小了模型规模，缩短了模型的训练时间；本发明的方法能够提高大数据的处理速度，缩短处理时间。

一种基于分段的两级深度学习模型的大数据处理方法

技术领域

本发明涉及人工智能和大数据领域，特别涉及一种基于分段的两级深度学习模型的大数据处理方法。

背景技术

随着网络技术的快速发展，数据的容量和多样性快速增加，而处理数据的算法复杂度却难以改善，如何高效处理大数据已经成为一个紧迫的难题。在现有的依赖个人经验和手工操作来描述数据、标注数据、选择特征、提取特征、处理数据的方法，已经很难满足大数据快速增长的需求。随着人工智能技术的快速发展，特别是深度学习算法的研究突破，为解决大数据处理问题指明了一个值得探索的方向。

Hinton 等人在 2006 年提出了用于深度置信网的逐层初始化训练方法，这是深度学习方法的研究起点，该方法打破了持续了几十年的深度神经网络训练困难且效果不好的局面。此后，深度学习算法在图像识别、语音识别、自然语言理解等领域得到广泛的应用。深度学习是通过模拟人脑分层次的抽象，将底层数据逐层映射而获得更抽象的特征，由于它可以从大数据中自动提取特征，并通过海量的样本训练获得很好的处理效果，从而得到了广泛的关注。实际上，大数据的快速增长和深度学习的研究突破是相辅相成的，一方面大数据的快速增长需要一种高效处理海量数据的方法，另一方面深度学习模型的训练需要海量的样本数据。总之，大数据可以使深度学习的性能达到极致。

但是，现有的深度学习模型仍然存在很多严重的问题，例如：模型难以扩展、参数优化困难、训练时间过长、推理效率低下等。在 2013 年 Bengio 的一篇综述论文中，总结了目前深度学习所面临的挑战和难点，包括如何扩展现有的深度学习模型的规模，并应用到更大的数据集；如何减小参数优化困难；如何避免昂贵的推理和采样；以及如何解开变化因素等。

发明内容

本发明的目的在于克服在大数据应用中现有神经网络深度学习模型存在的上述问题，提出了基于分段的两级深度学习模型，通过对深度学习模型分级分段的处理，对跨段间的权重进行约束，提高了模型的扩展能力，基于该模型，本发明提出了一种基于分段的两级深度学习模型的大数据处理方法，该方法能够提高大数据的处理速度，缩短处理时间。

为了实现上述目的,本发明提出了一种基于分段的两级深度学习模型的大数据处理方法,所述方法包括:

步骤 1) 构建并训练基于分段的两级深度学习模型,该模型从纵向层次上分为两级:第一级和第二级;所述第一级的各层从横向上分为 M 段;其中, M 为多模态输入的模态个数;第一级中不同段间相邻层的神经元节点间权重为 0;

步骤 2) 将待处理的大数据按照数据的类型分为 M 个子集,分别输入基于分段的两级深度学习模型的第一层的 M 段进行处理;

步骤 3) 输出大数据处理结果。

上述技术方案中,所述步骤 1) 进一步包括:

步骤 101) 将深度为 L 层的深度学习模型从纵向层次上分为两级:第一级和第二级:输入层为第一层,输出层为第 L 层,第 L^* 层为划分层, $2 \leq L^* \leq L-1$;则从第一层到第 L^* 层间的所有层称为第一级,而第 L^*+1 层到第 L 层间的所有层称为第二级;

步骤 102) 将第一级内每层上的神经元节点从横向上分为 M 段:

设 L 层神经网络的输入宽度为 N ,即每层有 N 个神经元节点,第一级的神经元节点划分为 M 段,每一段的宽度为 D_m , $1 \leq m \leq M$, 且 $\sum_{m=1}^M D_m = N$,并在同一段中,任意两层的宽度都相同;

步骤 103) 将训练样本划分为 M 个子集,分别输入深度学习模型的第一层的 M 段;

步骤 104) 分别对第一级的 M 段子模型进行训练:

第一级中不同段间相邻层的神经元节点间权重为 0,即第 m 段的所有节点集合为 S_m ,第 $l-1$ 层的任意节点为 $s_{i^{(m)},l-1} \in S_m$, $2 \leq l \leq L^*$,而第 o 段第 l 层的任意节点 $s_{j^{(o)},l} \in S_o$, 且 $m \neq o$, 则节点 $s_{i^{(m)},l-1}$ 和 $s_{j^{(o)},l}$ 间权重 $w_{i^{(m)},j^{(o)},l} = 0$;

在上述约束条件下,利用深度神经网络学习算法分别对第一级的 M 段子模型进行训练;

步骤 105) 对第二级的各层进行训练;

步骤 106) 通过深度神经网络学习算法对各层的网络参数进行全局微调,直到各层的网络参数达到最优值。

上述技术方案中, L^* 的取值方法为:在 L^* 的取值区间内通过交叉验证方法确定一个最优值。

本发明的优点在于:

1、本发明提出的基于分段的两级深度学习模型能够有效地减小了模型规模,缩短

了模型的训练时间；

2、本发明提出的大数据处理方法支持多源异构或多模态的大数据的并行输入，提高了大数据的处理速度，缩短了处理时间。

附图说明

图 1 为本发明的基于分段的两级深度学习模型的大数据处理方法的流程图；

图 2 为基于分段的两级深度学习模型的示意图。

具体实施方式

以下结合附图对本发明的方法做进一步详细的描述。

如图 1 所示，一种基于分段的两级深度学习模型的大数据处理方法，包括：

步骤 1) 构建并训练基于分段的两级深度学习模型；包括：

步骤 101) 将深度为 L 层的深度学习模型从纵向上分为两级：第一级和第二级：

输入层为第一层，输出层为第 L 层，第 L^* 层为划分层，其中 $2 \leq L^* \leq L-1$ ，则从第一层到第 L^* 层间的所有层称为第一级，而第 L^*+1 层到第 L 层间的所有层称为第二级。

L^* 的取值方法为：在 L^* 的取值区间内通过交叉验证方法确定一个最优值。

步骤 102) 将第一级内每层上的神经元节点从横向上分为 M 段；其中， M 为多模态输入的模态个数；

如图 2 所示，设 L 层神经网络的输入宽度为 N ，即每层有 N 个神经元节点，第一级的神经元节点划分为 M 段，每一段的宽度为 D_m ， $1 \leq m \leq M$ ，且 $\sum_{m=1}^M D_m = N$ ，并在同一段中，任意两层的宽度都相同；

步骤 103) 将训练样本划分为 M 个子集，分别输入深度学习模型的第一层的 M 段；

步骤 104) 分别对第一级的 M 段子模型进行训练；

第一级中不同段间相邻层的神经元节点间权重为 0，即第 m 段的所有节点集合为 S_m ，第 $l-1$ 层的任意节点为 $s_{i^{(m)}, l-1} \in S_m$ ， $2 \leq l \leq L^*$ ，而第 o 段第 l 层的任意节点 $s_{j^{(o)}, l} \in S_o$ ，且 $m \neq o$ ，则节点 $s_{i^{(m)}, l-1}$ 和 $s_{j^{(o)}, l}$ 间权重 $w_{i^{(m)}, j^{(o)}, l} = 0$ ；

在上述约束条件下，利用深度神经网络学习算法分别对第一级的 M 段子模型进行训练。

步骤 105) 对第二级的各层进行训练；

步骤 106) 通过深度神经网络学习算法对各层的网络参数进行全局微调，直到各层

的网络参数达到最优值；

所述深度神经网络学习算法为 BP 算法。

步骤 2) 将待处理的大数据按照数据的类型分为 M 个子集，分别输入基于分段的两级深度学习模型的第一层的 M 段进行处理；

步骤 3) 输出大数据处理结果。

最后所应说明的是，以上实施例仅用以说明本发明的技术方案而非限制。尽管参照实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，对本发明的技术方案进行修改或者等同替换，都不脱离本发明技术方案的精神和范围，其均应涵盖在本发明的权利要求范围当中。

权 利 要 求

1、一种基于分段的两级深度学习模型的大数据处理方法，所述方法包括：

步骤 1) 构建并训练基于分段的两级深度学习模型，该模型从纵向层次上分为两级：第一级和第二级；所述第一级的各层从横向上分为 M 段；其中，M 为多模态输入的模态个数；第一级中不同段间相邻层的神经元节点间权重为 0；

步骤 2) 将待处理的大数据按照数据的类型分为 M 个子集，分别输入基于分段的两级深度学习模型的第一层的 M 段进行处理；

步骤 3) 输出大数据处理结果。

2、根据权利要求 1 所述的基于分段的两级深度学习模型的大数据处理方法，其特征在于，所述步骤 1) 进一步包括：

步骤 101) 将深度为 L 层的深度学习模型从纵向层次上分为两级：第一级和第二级：输入层为第一层，输出层为第 L 层，第 L^* 层为划分层， $2 \leq L^* \leq L-1$ ；则从第一层到第 L^* 层间的所有层称为第一级，而第 L^*+1 层到第 L 层间的所有层称为第二级；

步骤 102) 将第一级内每层上的神经元节点从横向上分为 M 段：

设 L 层神经网络的输入宽度为 N，即每层有 N 个神经元节点，第一级的神经元节点划分为 M 段，每一段的宽度为 D_m ， $1 \leq m \leq M$ ，且 $\sum_{m=1}^M D_m = N$ ，并在同一段中，任意两层的宽度都相同；

步骤 103) 将训练样本划分为 M 个子集，分别输入深度学习模型的第一层的 M 段；

步骤 104) 分别对第一级的 M 段子模型进行训练：

第一级中不同段间相邻层的神经元节点间权重为 0，即第 m 段的所有节点集合为 S_m ，第 $l-1$ 层的任意节点为 $s_{i^{(m)}, l-1} \in S_m$ ， $2 \leq l \leq L^*$ ，而第 o 段第 l 层的任意节点 $s_{j^{(o)}, l} \in S_o$ ，且 $m \neq o$ ，则节点 $s_{i^{(m)}, l-1}$ 和 $s_{j^{(o)}, l}$ 间权重 $w_{i^{(m)}, j^{(o)}, l} = 0$ ；

在上述约束条件下，利用深度神经网络学习算法分别对第一级的 M 段子模型进行训练；

步骤 105) 对第二级的各层进行训练；

步骤 106) 通过深度神经网络学习算法对各层的网络参数进行全局微调，直到各层的网络参数达到最优值。

3、根据权利要求 2 所述的基于分段的两级深度学习模型的大数据处理方法，其特征在于，所述 L^* 的取值方法为：在 L^* 的取值区间内通过交叉验证方法确定一个最优值。

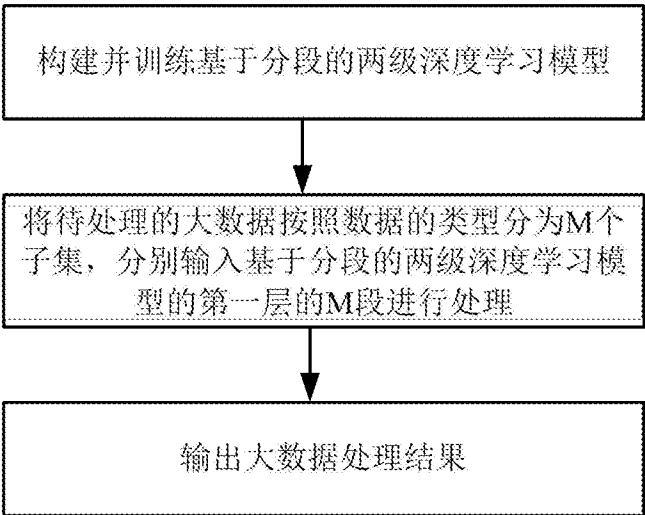


图 1

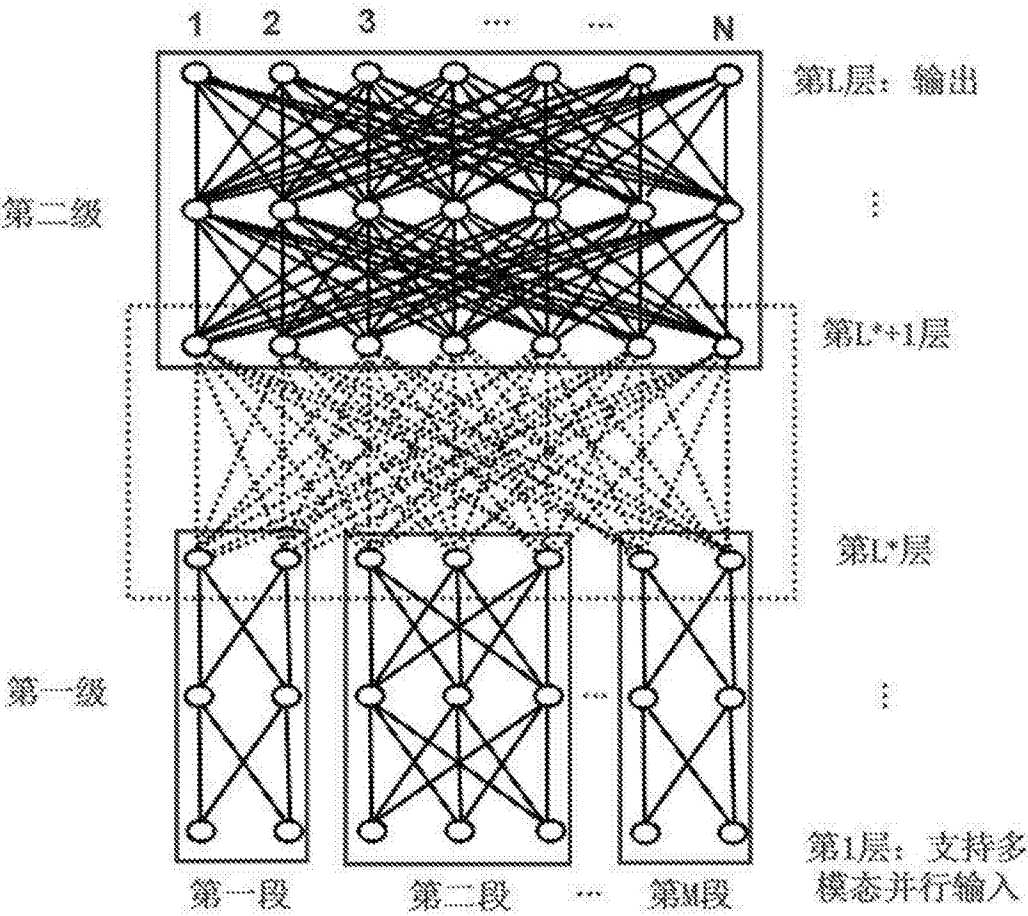


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2015/075472

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/02 (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)

G06N 3/-, G06F 17/-

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)

CNXTX, CNABS, DWPI, CNKI: nerve cell, depth, two-stage, deep, learning, big data, neural network, segment+, layer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 103945533 A (JINAN JIAKE ELECTRONIC TECHNOLOGY CO., LTD.), 23 July 2014 (23.07.2014), the whole document	1-3
A	WO 2014205231 A1 (UNIV MICHIGAN), 24 December 2014 (24.12.2014), the whole document	1-3
A	CN 104102929 A (HARBIN INSTITUTE OF TECHNOLOGY), 15 October 2014 (15.10.2014), the whole document	1-3

☐ 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 20 October 2015 (20.10.2015)	Date of mailing of the international search report 18 December 2015 (18.12.2015)
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 YI, Jian Telephone No.: (86-10) 62089283

International application No.
PCT/CN2015/075472

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

A. 主题的分类 G06N 3/02 (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) G06N 3/-, G06F 17/- 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)) CNTXT, CNABS, DWPI, CNKI: 神经元, 深度, 大数据, 学习, 分段, 两级, 神经网络, deep, learning, big data, neural network, segment+, layer		
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 103945533 A (济南嘉科电子技术有限公司) 2014年 7月 23日 (2014 - 07 - 23) 全文	1-3
A	WO 2014205231 A1 (UNIV MICHIGAN) 2014年 12月 24日 (2014 - 12 - 24) 全文	1-3
A	CN 104102929 A (哈尔滨工业大学) 2014年 10月 15日 (2014 - 10 - 15) 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 2015年 10月 20日		国际检索报告邮寄日期 2015年 12月 18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		授权官员 伊健 电话号码 (86-10) 62089283

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2015/075472

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	103945533	A	2014年 7月 23日	无	
WO	2014205231	A1	2014年 12月 24日	无	
CN	104102929	A	2014年 10月 15日	无	