

(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
2024 年 4 月 4 日 (04.04.2024)



(10) 国际公布号
WO 2024/066903 A1

(51) 国际专利分类号:
G06F 40/289 (2020.01)

(21) 国际申请号: PCT/CN2023/116180

(22) 国际申请日: 2023 年 8 月 31 日 (31.08.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202211211885.0 2022年9月30日 (30.09.2022) CN

(71) 申请人: 上海寰通商务科技有限公司 (OVAL TECHNOLOGIES, INC.) [CN/CN]; 中国上海市静安区天目西路128号嘉里不夜城企业中心第一座19楼1904-05室, Shanghai 200000 (CN)。

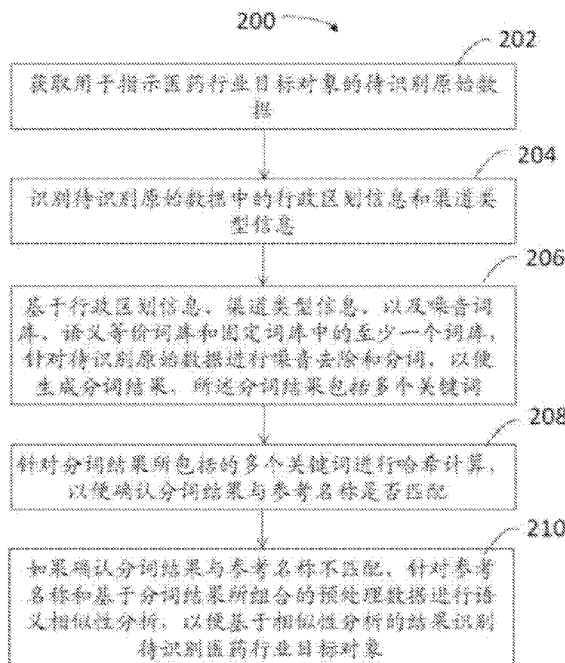
(72) 发明人: 姜金陆(JIANG, Jinlu); 中国上海市静安区静安区天目西路128号嘉里不夜城企业中心第一座19楼1904-05室, Shanghai 200000 (CN)。

(74) 代理人: 北京汇捷知识产权代理事务所(普通合伙) (BEIJING HUIJIE INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)); 中国上海市杨浦区四平路1388号同济联合广场C座302室, Shanghai 200092 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ,

(54) **Title:** METHOD AND DEVICE FOR RECOGNIZING PHARMACEUTICAL-INDUSTRY TARGET OBJECT TO BE RECOGNIZED, AND MEDIUM

(54) 发明名称: 识别待识别医药行业目标对象的方法、设备和介质



202 Acquire original data to be subjected to recognition, which is used for indicating a pharmaceutical-industry target object
204 Identify administrative-division information and channel type information from said original data
206 On the basis of the administrative-division information, the channel type information, and at least one of a noise lexicon, a semantic equivalence lexicon and a fixed lexicon, perform noise removal and word segmentation on said original data, so as to generate word segmentation results, wherein the word segmentation results include a plurality of keywords
208 Perform hash calculation on the plurality of keywords included in the word segmentation results, so as to determine whether the word segmentation results match reference names
210 If it is determined that the word segmentation results do not match the reference names, perform semantic similarity analysis on the reference names and preprocessed data that is obtained by means of combination based on the word segmentation results, so as to recognize, on the basis of a similarity analysis result, a pharmaceutical-industry target object to be recognized

[图2]

(57) **Abstract:** A method for recognizing a pharmaceutical-industry target object to be recognized. The method comprises: acquiring original data to be subjected to recognition, which is used for indicating a pharmaceutical-industry target object; identifying administrative-division information and channel type information from said original data; on the basis of the administrative-division information, the channel type information, and at least one of a noise lexicon, a semantic equivalence lexicon and a fixed lexicon, performing noise removal and word segmentation on said original data, so as to generate word segmentation results; performing hash calculation on a

LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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))。
— 含有关于援引加入遗漏项目或部分的信息(细则 20.6)

plurality of keywords included in the word segmentation results, so as to determine whether the word segmentation results match reference names; and in response to it being determined that the word segmentation results do not match the reference names, performing semantic similarity analysis on the reference names and preprocessed data that is obtained by means of combination based on the word segmentation results, so as to recognize a pharmaceutical-industry target object to be recognized. Therefore, a pharmaceutical-industry target object can be quickly and accurately recognized. Further disclosed are a device for recognizing a pharmaceutical-industry target object to be recognized, and a medium.

(57) 摘要: 一种用于识别待识别医药行业目标对象的方法, 该方法包括: 获取用于指示医药行业目标对象的待识别原始数据; 识别待识别原始数据中的行政区划信息和渠道类型信息; 基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库, 针对待识别原始数据进行噪音去除和分词, 以便生成分词结果; 针对分词结果所包括的多个关键词进行哈希计算, 以便确认分词结果与参考名称是否匹配; 以及响应于确认分词结果与参考名称不匹配, 针对参考名称和基于分词结果所组合的预处理数据进行语义相似性分析, 以便识别待识别医药行业目标对象。由此, 能够快速并且准确地针对医药行业目标对象进行识别。还公开了一种用于识别待识别医药行业目标对象的设备和介质。

说明书

发明名称: 识别待识别医药行业目标对象的方法、设备和介质 技术领域

[0001] 本公开的实施例总体涉及数据识别领域，并且更具体地涉及一种用于识别待识别医药行业目标对象的方法、计算设备和计算机存储介质。

背景技术

[0002] 传统的用于识别待识别医药行业目标对象(待识别医药行业目标对象例如但不限于医药分销领域中的机构)的方法通常包括：基于纯粹人工对未知的待识别医药行业目标对象进行识别；以及基于自然语言处理的简单分词技术对待识别医药行业目标对象进行识别两种方法。

[0003] 关于基于纯粹人工的识别方法，其虽然能够识别不规范的医药行业目标对象的原始数据，但是识别效率不高，并且存在因识别主体的经验差异而使得识别结果存在差异性，因此，难以适应大数据量的待识别医药行业目标对象的准确、快速识别，进而难以适应医药行业的服务平台对医药行业目标对象的识别需求。关于基于简单分词技术的识别方法，鉴于医药行业目标对象的原始数据表达得不规范，并且通常在内容和结构上存在明显的差异，加之医药行业没有现成的分词与匹配逻辑，因此使得目标对象的识别准确率相对较低。

[0004] 综上，传统的用于识别待识别医药行业目标对象的方法存在的不足之处在于：难以快速并且准确地针对医药行业目标对象进行识别。

[0005] 发明内容

[0006] 针对上述问题，本公开提供了一种用于识别待识别医药行业目标对象的方法、计算设备和计算机存储介质，能够快速并且准确地针对医药行业目标对象进行识别。

[0007] 根据本公开的第一方面，提供了一种用于识别待识别医药行业目标对象的方法，包括：获取用于指示医药行业目标对象的待识别原始数据；识别待识别原始数据中的行政区划信息和渠道类型信息；基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，针对待识别原始

数据进行噪音去除和分词，以便生成分词结果，分词结果包括多个关键词；针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配；以及响应于确认分词结果与参考名称不匹配，针对参考名称和基于分词结果所组合的预处理数据进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象。

[0008] 根据本公开的第二方面，提供了一种计算设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能够执行本公开的第一方面的方法。

[0009] 在本公开的第三方面中，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中计算机指令用于使计算机执行本公开的第一方面的方法。

[0010] 在一些实施例中，针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配包括：计算分词结果所包括的多个关键词的哈希值之和，以便生成分词结果哈希值之和；计算参考名称所包括的多个关键词的哈希值之和，以便生成参考名称哈希值之和；确认分词结果哈希值之和参考名称哈希值之和是否相等；以及响应于确认分词结果哈希值之和参考名称哈希值之和相等，确定分词结果与参考名称匹配。

[0011] 在一些实施例中，用于识别待识别医药行业目标对象的方法还包括：响应于确认分词结果与参考名称相匹配，将待识别医药行业目标对象识别为与参考名称相关联的目标对象。

[0012] 在一些实施例中，生成分词结果包括：基于所识别出的行政区划信息，获取待识别原始数据中除去行政区划信息之外的非行政区划数据；针对非行政区划数据，进行噪音词去除和等价词替换；以及基于固定词库，针对经由噪音词去除和等价词替换后的数据进行分割，以便生成与待识别原始数据对应的分词结果，分词结果包括多个关键词和指示分割位的多个预定标识符。

[0013] 在一些实施例中，用于识别待识别医药行业目标对象的方法还包括：用于识别待识别医药行业目标对象的方法识别待识别原始数据中的数字型的词；针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形

式的关键词；以及把分词结果所包括的多个关键词组合成不含地理信息的预处理数据，以用于与参考名称的匹配。

- [0014] 在一些实施例中，针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形式的关键词包括：将待识别原始数据中的大写中文数字和/或小写中文数字转换为阿拉伯数字；确定经转换的阿拉伯数字的位数是否大于或者等于预定位数阈值；响应于确定经转换的阿拉伯数字的位数大于或者等于预定位数阈值，去除经转换的阿拉伯数字；响应于确定经转换的阿拉伯数字的位数小于预定位数阈值，确定经转换的阿拉伯数字是否位于待识别原始数据的起始位置或者终止位置；响应于确定经转换的阿拉伯数字位于待识别原始数据的起始位置或者终止位置，确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据是否指示预定渠道类型；以及响应于确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据未指示预定渠道类型，去除经转换的阿拉伯数字。
- [0015] 在一些实施例中，渠道类型信息包括：渠道类型子分类名称、渠道类型分类名称和渠道类型分类序号。
- [0016] 在一些实施例中，识别待识别原始数据中的行政区划信息和渠道类型信息包括：确定分别与不同优先级顺序关联的多个关键词集合，每个关键词集合包括多个预定关键词；在多个关键词集合中，确定待识别原始数据中所包括的预定关键词所在的目标关键词集合；基于目标关键词集合所关联的优先级顺序，确定与待识别原始数据匹配的渠道类型子分类名称；以及基于所确定的渠道类型子分类名称，确定与待识别原始数据匹配的渠道类型分类名称和渠道类型分类序号。
- [0017] 在一些实施例中，针对待识别原始数据进行噪音去除和分词包括：确定多组关联词，每组关联词包括原始词和等价词，原始词和等价词在指示医药行业目标对象时具有一致的语义；为每组关联词确定关联的序号和所属分类，序号指示每组关联词的优先级；以及基于所确定的关联的序号，使用等价词替换和分割待识别原始数据，使得经由等价词替换和分割的数据中包括等价词和预定标示符，预定标示符指示分割位。

- [0018] 在一些实施例中，针对待识别原始数据进行噪音去除和分词包括：确定预处理数据和参考名称的重合部分；在预处理数据中删除重合部分，以便获得剩余部分；响应于确定满足第一预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，匹配可信度等级为第一等级指示待识别原始数据和参考名称之间相匹配，第一预定条件包括以下任一项：确定剩余部分所包括的字数小于或者等于第一字数阈值；确定剩余部分所包括的字数大于第二字数阈值并且剩余部分和重合部分关联有相同的渠道类型信息，第二字数阈值大于第一字数阈值；剩余部分所包括的字数大于第一字数阈值并且小于第二字数阈值并且剩余部分包含一对括号；剩余部分包含“原”或括号和“原”；剩余部分包含一对括号并且括号的字数小于第三字数阈值，第三字数阈值大于第一字数阈值并且小于第二字数阈值；确定预处理数据和参考名称存在重合部分，并且预处理数据和参考名称具有相同的渠道类型子分类。
- [0019] 在一些实施例中，针对分词结果和参考名称进行语义相似性分析还包括：响应于确定满足第二预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第二等级，第二预定可信度条件包括：确定预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称的渠道类型分类信息相同；响应于确定满足第三预定可信度条件，确定待识别原始数据和参考名称之间的不匹配，第三预定可信度条件包括：预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称渠道类型分类信息不同。
- [0020] 在一些实施例中，识别待识别原始数据中的行政区划信息和渠道类型信息包括：基于关于省份、城市、区县的全称、简称、曾用名和排除词来识别待识别的机构名称中的行政区划信息，行政区划信息包括省份信息、城市信息和区县信息；响应于确认所识别的区县信息或者城市信息没有指示唯一的区县或者城市，利用所识别的区县信息或者城市信息的下级行政区划信息、或者待识别目标对象的关联目标对象的行政区划信息来识别待识别原始数据中的行政区划信息。

[0021] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0022] 结合附图并参考以下详细说明，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中，相同或相似的附图标注表示相同或相似的元素。

[0023] 图1示出了用于实现根据本发明的实施例的用于识别待识别医药行业目标对象的方法的系统的示意图。

[0024] 图2示出了根据本公开的实施例的用于识别待识别医药行业目标对象的方法的流程图。

[0025] 图3示出了根据本公开的实施例的用于识别待识别原始数据中的行政区划信息和渠道类型信息的方法的流程图。

[0026] 图4示出了根据本公开的实施例的用于分割出待识别原始数据中的数字词形式的关键词的方法的流程图。

[0027] 图5示出了根据本公开的实施例的用于针对分词结果和参考名称进行语义相似性分析的方法的流程图。

[0028] 图6示出了根据本公开的实施例的用于生成分词结果的方法的流程图。

[0029] 图7示出了根据本公开的实施例的电子设备的框图。

具体实施方式

[0030] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0031] 在本文中使用的术语“包括”及其变形表示开放性包括，即“包括但不限于”。除非特别申明，术语“或”表示“和/或”。术语“基于”表示“至少部分

地基于”。术语“一个示例实施例”和“一个实施例”表示“至少一个示例实施例”。术语“另一实施例”表示“至少一个另外的实施例”。术语“第一”、“第二”等等可以指代不同的或相同的对象。下文还可能包括其他明确的和隐含的定义。

[0032] 如前文所描述，传统的、基于纯粹人工的识别方法，其识别效率不高，并且存在因识别主体的经验差异而使得识别结果存在差异性，因此，难以适应针对大数据量的待识别医药行业目标对象的准确而快速地识别，进而无法满足医药行业的服务平台对医药行业目标对象的识别需求。传统的、基于简单分词技术的识别方法因缺乏医药行业的分词与模式逻辑，因而使得针对目标对象的识别准确率相对较低。因此，传统的用于识别待识别医药行业目标对象的方法存在的不足之处在于：难以快速并且准确地针对医药行业目标对象进行识别。例如，传统的用于识别待识别医药行业目标对象的方法难以快速而准确地识别“三门峡医药有限责任公司”和“华润三门峡医药有限公司”。

[0033] 为了至少部分地解决上述问题以及其他潜在问题中的一个或者多个，本公开的示例实施例提出了一种用于识别待识别医药行业目标对象的方案，在本公开方案中，通过针对所获取的、用于指示医药行业目标对象的待识别原始数据进行行政区划信息和渠道类型信息的识别，以便基于所识别的行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，对待识别原始数据进行噪音去除和分词，以生成分词结果，本公开可以使得分词结果为经由噪音去除、经由语义等价词和/或固定词标准化后的分词结果，而且辅助渠道类型信息加以判断，因而能够克服医药行业目标对象的原始数据结构差异、表达不规范和容易混淆的问题。另外，本公开利用针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配；以及如果确认分词结果与参考名称不匹配，针对由分词结果所生成的预处理数据和参考名称进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象，本公开可以基于在标准化后的分词结果的基础上，先经由哈希计算准确识别分词结果与参考名称的匹配关系，在无法匹配的基础上利用语义相似性

分析结果识别待识别医药行业目标对象，因而本公开能够更为快速和准确地识别待识别医药行业目标对象。

[0034] 图1示出了用于实现根据本发明的实施例的用于识别待识别医药行业目标对象的方法的系统100的示意图。如图1中所示，系统100包括计算设备110和服务器130和网络140。计算设备110、服务器130可以通过网络140(例如，因特网)进行数据交互。

[0035] 服务器130，其例如可以将用于指示医药行业目标对象的待识别原始数据发送给计算设备110。

[0036] 关于计算设备110，其例如用于获取服务器130所提供的用于指示医药行业目标对象的待识别原始数据；以及识别待识别原始数据中的行政区划信息和渠道类型信息。计算设备110还可以基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，对待识别原始数据进行噪音去除和分词，以便生成分词结果；针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配；以及如果确认分词结果与参考名称不匹配，针对分词结果和参考名称进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象。计算设备110可以具有一个或多个处理单元，包括诸如GPU、FPGA和ASIC等的专用处理单元以及诸如CPU的通用处理单元。另外，在每个计算设备110上也可以运行着一个或多个虚拟机。在一些实施例中，计算设备110与医学影像成像设备110可以集成在一起，也可以是彼此分立设置。在一些实施例中，计算设备110例如包括待识别原始数据获取单元112、行政区划和渠道类型信息识别单元114、分词结果生成单元116、哈希计算单元118、待识别医药行业目标对象识别单元120。

[0037] 关于待识别原始数据获取单元112，其用于获取用于指示医药行业目标对象的待识别原始数据。

[0038] 关于行政区划和渠道类型信息识别单元114，其用于识别待识别原始数据中的行政区划信息和渠道类型信息。

- [0039] 关于分词结果生成单元116，其用于基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，针对待识别原始数据进行噪音去除和分词，以便生成分词结果，分词结果包括多个关键词。
- [0040] 关于哈希计算单元118，基于针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配。
- [0041] 关于待识别医药行业目标对象识别单元120，其用于如果确认分词结果与参考名称不匹配，针对参考名称和基于分词结果所组合的预处理数据进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象。
- [0042] 以下结合图2描述用于识别待识别医药行业目标对象的方法200。图2示出了根据本公开的实施例的用于识别待识别医药行业目标对象的方法200的流程图。方法200可由如图1所示的计算设备110执行，也可以在图7所示的电子设备700处执行。应当理解的是，方法200还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。
- [0043] 在步骤202，计算设备110获取用于指示医药行业目标对象的待识别原始数据。例如，计算设备110获取来自服务器130的关于医药分销领域中的未知机构的待识别原始数据。
- [0044] 关于待识别的医药行业目标对象，其例如而不限于是医药分销领域中的未知机构。例如，计算设备110需要识别某个医药分销领域中的未知的公司机构名称代表哪个标准的机构名称。应当理解，同一个医药行业目标对象(例如而不限于是同一医药商店)与不同的医药机构(例如经销商)之间有供货关系，这家医药行业目标对象可能在不同的医药机构(例如经销商)处其名称或者叫法不一致。
- [0045] 在步骤204，计算设备110识别待识别原始数据中的行政区划信息和渠道类型信息。
- [0046] 关于行政区划信息，其例如包括：省份、城市、区县三级行政机构的所属信息。
- [0047] 关于识别待识别原始数据中的行政区划信息的方法，其例如包括：计算设备110基于关于省份、城市、区县的全称、简称、曾用名和排除词来识别待识别的机构名称中的行政区划信息，行政区划信息包括省份信息、城市信息和区县信

息；如果确认所识别的区县信息或者城市信息没有指示唯一的区县或者城市，利用所识别的区县信息或者城市信息的下级行政区划信息、或者待识别目标对象的关联目标对象的行政区划信息来识别待识别原始数据中的行政区划信息。

具体而言，如果计算设备110确定待识别原始数据中所含有的省份信息包括省份全称、简称、省会城市或省会城市的曾用名，并且不包括关于省份或省会城市的排除名称，则确定识别省份信息；如果确定待识别原始数据中所含有的城市信息包括城市全称、简称或曾用名，并且不包含城市的排除名称，则确定识别城市信息；以及确定待识别原始数据中所含有的区县信息包括区县的全称、简称或曾用名，并且不包括区县的排除名称，则确定识别区县信息；如果计算设备110确定满足以下任一项，则确定识别出行政区划信息：确认识别省份信息、城市信息和区县信息；确认识别行政区划信息和区县信息、确认识别城市信息和区县信息；确认所识别区县信息或者城市信息指示唯一的区县或者城市。

[0048] 例如，如果计算设备110确定待识别的机构名称中同时含有省市区县三级行政机构、省和区县两级行政机构(例如，省+县/二级市场/区)、市和区县两级行政机构(例如，地市级+区县级)，就不需要后续的检测就可以直接识别出待识别的机构名称所属的行政区划信息，即认为已经准确找到待识别的机构名称所属的省市县。

[0049] 例如，如果计算设备110确定待识别的机构名称中包括区县的全称或者城市的全称，且该区县的名称或者城市的全称是唯一的，则确定识别出待识别的机构名称所属的行政区划信息。应当理解，全国的城市和县是唯一的，如果待识别的机构名称中包含了唯一的城市/县的全称或简称或曾用名，即认为能唯一识别到待识别的机构名称所属的省市县。

[0050] 如果计算设备110确认所识别的区县信息或者城市信息没有指示唯一的区县或者城市，利用所识别的区县信息或者城市信息的下级行政区划信息、或者待识别目标对象的关联目标对象的行政区划信息来识别待识别原始数据中的行政区划信息。例如，“通州区永顺镇果元村卫生室”和“通州区金沙镇本草药房”，其中，通州区并未指示唯一的区县，例如，北京包括通州、江苏省也包括通州。因此，可以借助于下级行政区划信息(例如，乡镇和区县关系)来识别待

识别的机构名称中的行政区划信息。例如，通过“通州”+“永顺”能够找到唯一行政区划关系“北京+通州+永顺”，则此时将定位到北京市通州区；同理，通过“通州”+“金沙”能够找到唯一行政区划关系“江苏省+南通市+通州区”。

[0051] 再例如，如下表一所示，待识别的目标对象(例如买方机构)的名称例如为“北沟卫生院”，单从“北沟卫生院”是无法找到其所属省市县区的地理信息或者行政区划信息的。计算设备110可以识别关联的目标对象(例如，卖方机构“华润烟台医药有限公司”)是山东烟台的，计算设备110可以在烟台地区找下游乡镇是否存在“北沟”，最终能够找到唯一的“蓬莱区”下有一个北沟镇。

[0052] 表一

卖方	买方	买 方 省份	买 方 城市	买 方 区县
华润烟台医药有限公司	北沟卫生院	山 东 省	烟 台 市	蓬 莱 区

细则 20.6,
01.02.2024

[0053] 在一些实施例中，计算设备110基于关于省份、城市、区县的全称、简称、曾用名和排除词来识别待识别的机构名称中的行政区划信息，行政区划信息包括省份信息、城市信息和区县信息省份、城市、区县。以下表二示例性示出了关于市、区县的全称、简称、曾用名和排除词。在表二中，关于省的全称、简称、曾用名和排除词未示出。

[0054] 表二

城市 全称	城市 简称	城市 排除 词 1	城市 排除 词 2	城市 排除 词 3	区县 全称	区县 简称	区县 排除 词 1	区县 排除 词 2	区县 排除 词 3
焦作市	焦作	焦作路	焦作街	焦作村	沁阳市	沁阳	沁阳路	沁阳街	沁阳村
安庆市	安庆	安庆路	安庆街	安庆村	望江县	望江	沁阳路	沁阳街	沁阳村

细则 20.6,
01.02.2024

[0055] 再例如，“三门医药有限责任公司”、“华润三门峡医药有限公司”和“夹江县医药公司三门市”等，存在一些容易混淆的区县简称。计算设备110可以基于关于省份、城市、区县的排除词辅助识别别的机构名称中的行政区划信息。例如，三门县，其简称是三门，排除词包括：三门峡、三门市、第三门、三门

店。通过采用上述手段，本发明可以准确识别容易混淆的行政区划信息，进而有利于提高识别目标对象的准确度。

[0056] 关于渠道类型信息，其例如包括：渠道类型子分类名称、渠道类型子分类名称和渠道类型分类序号。应当理解，医药流通行业机构数据分为三大类：经销商、医疗终端、零售终端，每一个类别下又有细分类，比如零售终端下又分为单体药店和连锁药店分店，机构名称通常含有渠道类型信息这类属性信息的，这些属性信息有助于提高待识别的医药行业目标对象的识别准确性，比如零售终端不能识别到医疗终端。因此，通过识别待识别原始数据中的渠道类型信息，有利于提高待识别的医药行业目标对象的识别精确度。

[0057] 关于识别待识别原始数据中的行政区划信息和渠道类型信息的方法，其例如包括：计算设备110确定分别与不同优先级顺序关联的多个关键词集合，每个关键词集合包括多个预定关键词；在多个关键词集合中，确定待识别原始数据中所包括的预定关键词所在的目标关键词集合；基于目标关键词集合所关联的优先级顺序，确定与待识别原始数据匹配的渠道类型子分类名称；以及基于所确定的渠道类型子分类名称，确定与待识别原始数据匹配的渠道类型分类名称和渠道类型分类序号。

[0058] 在步骤206，计算设备110基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，针对待识别原始数据进行噪音去除和分词，以便生成分词结果，分词结果包括多个关键词。

[0059] 关于对待识别原始数据进行噪音去除和分词的方法，其例如包括：确认经由噪音去除和归一化处理的预处理数据是否匹配参考名称的全称、别名和曾用名中的至少一项；如果确认经由噪音去除和归一化处理的预处理数据与参考名称的全称、别名和曾用名均不匹配，针对预处理数据进行分词，以便生成分词结果。如果预处理数据等于参考名称的别名或曾用名、或者预处理数据加上其上游名称等于参考名称或其别名、或者预处理数据加上其上游名称与参考名称或其别名同音字全，则计算设备110确定待识别原始数据与参考名称匹配，而不需要针对预处理数据进行分词。

- [0060] 关于生成分词结果的方法，其例如包括：基于所识别出的行政区划信息，获取待识别原始数据中除去行政区划信息之外的非行政区划数据；针对非行政区划数据，进行噪音词去除和等价词替换；以及基于固定词库，针对经由噪音词去除和等价词替换后的数据进行分割，以便生成与待识别原始数据对应的分词结果，分词结果包括多个关键词和指示分割位的多个预定标识符；识别待识别原始数据中的数字型的词；针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形式的关键词；以及把分词结果所包括的多个关键词组合成不含地理信息的预处理数据，以用于与参考名称的匹配。下文将结合图6详细说明用于针对分词结果和参考名称进行语义相似性分析的方法，在此，不再赘述。
- [0061] 关于分割出待识别原始数据中的数字词形式的关键词的方法，其例如包括：计算设备110将待识别原始数据中的大写中文数字和/或小写中文数字转换为阿拉伯数字；确定经转换的阿拉伯数字的位数是否大于或者等于预定位数阈值；响应于确定经转换的阿拉伯数字的位数大于或者等于预定位数阈值，去除经转换的阿拉伯数字；响应于确定经转换的阿拉伯数字的位数小于预定位数定阈值，确定经转换的阿拉伯数字是否位于待识别原始数据的起始位置或者终止位置；响应于确定经转换的阿拉伯数字位于待识别原始数据的起始位置或者终止位置，确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据是否指示预定渠道类型；以及响应于确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据未指示预定渠道类型，去除经转换的阿拉伯数字。下文将结合图4详细说明用于分割出待识别原始数据中的数字词形式的关键词的方法，在此，不再赘述。
- 关于针对非行政区划数据进行噪音词去除和等价词替换的方法，其例如包括：计算设备110确定多组关联词，确定多组关联词，每组关联词包括原始词和等价词，原始词和等价词在指示医药行业目标对象时具有一致的语义；为每组关联词确定关联的顺序号和所属分类，顺序号指示每组关联词的优先级；以及基于所确定的关联的顺序号，使用等价词替换和分割待识别原始数据，使得经由等价词替换和分割的数据中包括等价词和预定标示符，预定标示符指示分割位。

[0062] 在步骤208，计算设备110针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配。

[0063] 关于确认分词结果与参考名称是否匹配的方法，其例如包括：计算设备110计算分词结果所包括的多个关键词的哈希值之和，以便生成分词结果哈希值之和；计算参考名称所包括的多个关键词的哈希值之和，以便生成参考名称哈希值之和；确认分词结果哈希值之和参考名称哈希值之和是否相等；以及响应于确认分词结果哈希值之和参考名称哈希值之和相等，确定分词结果与参考名称匹配。通过采用被分词后的关键词的哈希值之和的算法逻辑，能够使得匹配结果不因关键词词序不同而影响。

[0064] 以下公式(1)示意性示出了确认分词结果与参考名称是否匹配的算法。

[0065] 在上述公式(1)中， $\text{ora_hash}(\text{key reference } i)$ 代表针对参考数据的分词结果所包括的第*i*个关键词所计算的哈希值。*i*代表关键词的序号。代表参考名称哈希值之和。*n*代表关键词的总数量，例如表三或表四中，关键词的总数量*n*是19。 $\text{ora_hash}(\text{key original } i)$ 代表针对待识别原始数据的分词结果所包括的第*i*个关键词所计算的哈希值。代表分词结果哈希值之和。

[0066] 例如，以下表三示意性示出了参考名称的分词结果。其中，参考名称例如为“阜阳市延生大药房零售连锁有限公司梦廉分店”，参考名称例如被分为表三中关键词1至关键词19的十九个关键词。表三中仅示意性示出其中的九个关键词。

[0067] 表三

参考名称	关键词1	关键词2	关键词3	关键词4	关键词5	关键词6	关键词7	关键词8	关键词9
阜阳延生大药房零售连锁有限公司梦廉分店	安徽省	阜阳市	/	延生	药	公司	梦廉	店	/
宝鸡百姓药业有限责任公司七十一店	陕西省	宝鸡市	/	百姓	公司	店	/	/	71

细则 20.6,
01.02.2024

[0068] 例如，以下表四示意性示出了待识别原始数据的分词结果。其中，待识别原始数据例如为“阜阳市延生大药房零售连锁公司(梦廉)”，待识别原始数据例如被分为表四中的中关键词1至关键词19的十九个关键词。表四中仅示意性示出其中的九个关键词。

[0069] 表四

待识别原始数据	关键词1	关键词2	关键词3	关键词4	关键词5	关键词6	关键词7	关键词8	关键词9
阜阳市延生大药房零售连锁公司(梦廉)	安	阜	/	延	药	公	梦	廉	店
柒拾壹宝鸡百姓药业有限公	陕	宝	/	百	公	店	/	/	71

细则 20.6,
01.02.2024

[0070] 上述待识别原始数据“阜阳市延生大药房零售连锁公司(梦廉)”通过去噪和分词，打破了词序，并且大小写被归一化，生成表四中的关键词1至关键词19的十九个关键词。待识别原始数据的分词结果的关键词1至关键词19所有关键词的哈希值之和(即，分词结果哈希值之和)等于参考名称之和(参考名称指标准的目标对象名称)，这样，计算设备110确定分词结果与参考名称匹配。

[0071] 以下示例性示出用于实现确认分词结果与参考名称是否匹配的算法的示意性程序代码。

```
[0072] select*
[0073] from(select a.collatejobdetailid, a.orgname, o.ovalmasterid asstdorgid, o.orgcode as
stdorgcode,
[0074] o.orgname as stdorgname,2as status,2as gradelevel, length(o.orgname)asorglen,
[0075] case when a.channelname=o.channel and substr(a.keyword05, -1)in(' - , ' 店
' , ' 药' , ' 生' , ' 诊' , ' 院' )then' 99%'
```


[0076] when a.channelname=o.channel then ' 98%' else ' 95%' end as grade, ' 拆词
全等推荐' as splitstatus_std

[0077] from collatejobdetaila,ovalmaster o

[0078] where a.jobid=v_jobid.....

[0079] and a.keyword01=o.keyword01

[0080] and a.keyword02=o.keyword02

[0081] and a.keyword03=o.keyword03

[0082] and a.hashvalue=o.hashvalue

[0083] /*hashvalue

[0084] ora_hash(a.keyword04)+ora_hash(a.keyword05)+

[0085] ora_hash(a.keyword06)+ora_hash(a.keyword07)+

[0086] ora_hash(a.keyword08)+ora_hash(a.keyword09)+

[0087] ora_hash(a.keyword10)+ora_hash(a.keyword11)+

[0088] ora_hash(a.keyword12)+ora_hash(a.keyword19)=

[0089] ora_hash(o.keyword04)+ora_hash(o.keyword05)+

[0090] ora_hash(o.keyword06)+ora_hash(o.keyword07)+

[0091] ora_hash(o.keyword08)+ora_hash(o.keyword09)+

[0092] ora_hash(o.keyword10)+ora_hash(o.keyword11)+

[0093] ora_hash(o.keyword12)+ora_hash(o.keyword19)*/*

[0094])

[0095] order by orgname,orglen

[0096] 在步骤210, 如果计算设备110确认分词结果与参考名称不匹配, 针对参考名称
和基于分词结果所组合的预处理数据进行语义相似性分析, 以便基于相似性分
析的结果识别待识别医药行业目标对象。例如, 如果计算设备110确认分词结果
与参考名称相匹配, 将待识别医药行业目标对象识别为与参考名称相关联的目
标对象。

[0097] 关于针对分词结果和参考名称进行语义相似性分析的方法, 其例如包括: 确定
预处理数据和参考名称的重合部分; 在预处理数据中删除重合部分, 以便获得

剩余部分；响应于确定满足第一预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，匹配可信度等级为第一等级指示待识别原始数据和参考名称之间相匹配，第一预定条件包括以下任一项：确定剩余部分所包括的字数小于或者等于第一字数阈值；确定剩余部分所包括的字数大于第二字数阈值并且剩余部分和重合部分关联有相同的渠道类型信息，第二字数阈值大于第一字数阈值；剩余部分所包括的字数大于第一字数阈值并且小于第二字数阈值并且剩余部分包含一对括号；剩余部分包含“原”或括号和“原”；剩余部分包含一对括号并且括号的字数小于第三字数阈值，第三字数阈值大于第一字数阈值并且小于第二字数阈值；响应于满足第二预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第二等级，第二预定可信度条件包括以下任一项：确定预处理数据和参考名称存在重合部分，并且预处理数据和参考名称具有相同的渠道类型子分类；确定预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称的渠道类型分类信息相同；响应于满足第三预定可信度条件，确定待识别原始数据和参考名称之间的不匹配，第三预定可信度条件包括：预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称渠道类型分类信息不同。下文将结合图5详细说明用于针对分词结果和参考名称进行语义相似性分析的方法，在此，不再赘述。

[0098] 在一些实施例中，如果针对分词结果和参考名称的语义相似性分析或针对分词结果进行哈希计算均不能精确识别待识别的医药行业目标对象，那么计算设备110可以基于渠道类型信息调整分词结果和参考名称的语义相似性分析的权重，以便基于经调整后的权重针对分词结果和参考名称进行语义相似性分析。

[0099] 在上述方案中，通过针对所获取的、用于指示医药行业目标对象的待识别原始数据进行行政区划信息和渠道类型信息的识别，以便基于所识别的行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，对待识别原始数据进行噪音去除和分词，以生成分词结果，本公开可以使得分词结果为经由噪音去除、经由语义等价词和/或固定词标准化后的分词结果，而且辅助渠道类型信息加以判断，因而能够克服医药行业目标对象的原始

数据结构差异、表达不规范和容易混淆的问题。另外，本公开利用针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配；以及如果确认分词结果与参考名称不匹配，针对由分词结果所生成的预处理数据和参考名称进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象，本公开可以基于在标准化后的分词结果的基础上，先经由哈希计算准确识别分词结果与参考名称的匹配关系，在无法匹配的基础上利用语义相似性分析结果识别待识别医药行业目标对象，因而本公开能够更为快速和准确地识别待识别医药行业目标对象。

- [0100] 以下结合图3说明用于识别待识别原始数据中的行政区划信息和渠道类型信息的方法。图3示出了根据本公开的实施例的用于识别待识别原始数据中的行政区划信息和渠道类型信息的方法300的流程图。方法300可由如图1所示的计算设备110执行，也可以在图7所示的电子设备700处执行。应当理解的是，方法300还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。
- [0101] 在步骤302，计算设备110确定分别与不同优先级顺序关联的多个关键词集合，每个关键词集合包括多个预定关键词。
- [0102] 关于用于识别渠道类型分类的关键词集合，其例如包括用于识别连锁药店的关键词集合、用于识别单体药店的关键词集合、用于识别连锁公司的关键词集合、用于识别医院的关键词集合和用于识别卫生监督所的关键词集合。
- [0103] 以下表三示例性示意出用于识别连锁药店的关键词集合、用于识别单体药店的关键词集合。

[0104] 表五

优先级顺序	渠道类型分类序号	渠道类型分类名称	渠道类型子分类名称	关键词集合
3	114	终端药店	连锁药店分店	“%药房有限公司”、“%连锁%”、“%同仁堂%”、“%复美%”、“%大%参林%”、“%天天好%”、“%百姓%”、“%好药师%”、“%复星%”、“%

				海王%”、“%益丰%”、“%美罗%”
18	114	终端药店	单体药店	(“%零售中心%”、“%药%零售%公司%”、“%药%公司%零售%”、“%药%(%医院%) %”)、而非 “%连锁%店%”、“%公司%店%”

- [0105] 在步骤304，计算设备110在多个关键词集合中，确定待识别原始数据中所包括的预定关键词所在的目标关键词集合。例如，计算设备110确定待识别原始数据中包括的预定关键词“%零售中心%”，并且不包括“%连锁%店%”，则所包括的预定关键词所在的目标关键词集合为表五中第二行的关键词集合。
- [0106] 在步骤306，计算设备110基于目标关键词集合所关联的优先级顺序，确定与待识别原始数据匹配的渠道类型子分类名称。例如，表五中第二行的关键词集合所关联的优先级顺序为18，计算设备110基于该优先级顺序18确定与待识别原始数据匹配的渠道类型子分类名称为“单体药店”。应当理解，在识别渠道类型时，基于目标关键词集合所关联的优先级顺序，优先符合条件的，即认为“定位到与待识别原始数据匹配的渠道类型子分类名称”。
- [0107] 在步骤308，计算设备110基于所确定的渠道类型子分类名称，确定与待识别原始数据匹配的渠道类型分类名称和渠道类型分类序号。例如，计算设备110基于所确定的渠道类型子分类名称“单体药店”，确定与待识别原始数据匹配的渠道类型分类名称为“终端药店”和渠道类型分类序号为“114”。
- [0108] 在上述方案中，本公开能够准确确定待识别原始数据所属渠道类型，有利于基于准确的所属渠道类型提高识别医药行业目标对象的准确度。

- [0109] 以下结合图4说明用于分割出待识别原始数据中的数字词形式的关键词的方法。图4示出了根据本公开的实施例的用于分割出待识别原始数据中的数字词形式的关键词的方法400的流程图。方法400可由如图1所示的计算设备110执行，也可以在图7所示的电子设备700处执行。应当理解的是，方法400还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。
- [0110] 在步骤402，计算设备110将待识别原始数据中的大写中文数字和/或小写中文数字转换为阿拉伯数字。应当理解，在待识别原始数据中，可能会含有大小写数字、电话、邮编。这些数字可能出现在待识别目标对象的名称的前面、后面或者中间。例如，计算设备110可以将出现在待识别原始数据中的所有大小写中文数字统一变成阿拉伯数字，例如，将“一百五十一”、“一五一”、或者“壹佰五十一”最终都转换为阿拉伯数字“151”。通过利用上述手段，利于使得待识别原始数据中的数字型词进行归一化。
- [0111] 在步骤404，计算设备110确定经转换的阿拉伯数字的位数是否大于或者等于预定位数阈值。
- [0112] 关于预定位数阈值，其例如而不限于是6及以上的一个数。
- [0113] 如果计算设备110确定经转换的阿拉伯数字的位数大于或者等于预定位数阈值，在步骤406，去除经转换的阿拉伯数字。例如，如果确定经转换的阿拉伯数字的位数大于或者等于6(或者6位及以上)，无论经转换的阿拉伯数字在何处出现，都直接去掉经转换的阿拉伯数字。因为，该经转换的阿拉伯数字可能是电话、邮编等信息。
- [0114] 在步骤408，如果计算设备110确定经转换的阿拉伯数字的位数小于预定位数阈值，确定经转换的阿拉伯数字是否位于待识别原始数据的起始位置或者终止位置。
- [0115] 在步骤410，如果计算设备110确定经转换的阿拉伯数字位于待识别原始数据的起始位置或者终止位置，确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据是否指示预定渠道类型。例如，如果计算设备110确定经转换的阿拉伯数字小于预定位数阈值，并且出现在待识别目标对象名称的起始位置，则也可

以去除经转换的阿拉伯数字，因为该经转换的阿拉伯数字很可能是在提供目标对象名称时不小心加入的序号。

[0116] 如果计算设备110确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据未指示预定渠道类型，跳转至步骤406，去除经转换的阿拉伯数字。

[0117] 在步骤412，如果计算设备110确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据指示预定渠道类型，不去除经转换的阿拉伯数字。

[0118] 例如，如果计算设备110确定经转换的阿拉伯数字位于待识别原始数据的起始位置或者终止位置，并且紧随经转换的阿拉伯数字的不是药店类型或者医疗机构类型(渠道类型不是药店)，则可以去除阿拉伯数字；如果计算设备110确定与位于终止位置的阿拉伯数字位是在终止位置出现，并且紧随经转换的阿拉伯数字的是药店类型或者医疗机构类型且经转换的阿拉伯数字小于或者等于预定数字阈值，则不能去除阿拉伯数字。例如，以下表六中示例的“56店王志恒”，其中，阿拉伯数字“56”位于待识别原始数据的起始位置，并且紧随经转换的阿拉伯数字“56”的是药店类型，同时，阿拉伯数字“56”小于市级医药有限公司所关联的预定数字阈值，此时，计算设备110确定不能去除阿拉伯数字“56”。

[0119] 例如，以下表六中示例的“56店王志恒”，其中，阿拉伯数字“56”位于待识别原始数据的起始位置，并且紧随经转换的阿拉伯数字“56”的是药店类型，同时，阿拉伯数字“56”小于市级医药有限公司所关联的预定数字阈值，此时，计算设备110确定不能去除阿拉伯数字“56”。

[0120] 再例如，以下表六中示例的“骑鹤乡王楼卫生室50”或“骑鹤乡王楼卫生室1”，其中，阿拉伯数字“50”或“1”位于待识别原始数据的终止位置，并且紧随经转换的阿拉伯数字“50”或“1”是医疗机构类型，假设阿拉伯数字“50”大于乡级卫生室所关联的预定数字阈值，此时，计算设备110确定去除阿拉伯数字“50”；而阿拉伯数字“1”小于乡级卫生室所关联的预定数字阈值，此时，计算设备110确定不能去除阿拉伯数字“1”。

细则 20.6,
01.02.2024

[0121] 表六

卖方	买方	
华润烟台医药有	56 店王志恒	不能去除阿拉伯数
限公司		字
华润烟台医药有	56 人民路街道社区卫生	可以去除阿拉伯数
限公司	服务中心	字
华润烟台医药有	骑鹤乡王楼卫生室 1	不能去除阿拉伯数
限公司		字
华润烟台医药有	骑鹤乡王楼卫生室 50	可以去除阿拉伯数
限公司		字

[0122] 在上述方案中，本公开能够准确识别并剔除待识别原始数据中的数字型噪音，而利于准确分割出利于识别目标对象的数字型关键词。

[0123] 以下结合图5说明用于针对分词结果和参考名称进行语义相似性分析的方法。图5示出了根据本公开的实施例的用于针对分词结果和参考名称进行语义相似性分析的方法500的流程图。方法500可由如图1所示的计算设备110执行，也可以在图7所示的电子设备700处执行。应当理解的是，方法500还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。

[0124] 在步骤502，计算设备110确定预处理数据和参考名称的重合部分。

[0125] 在步骤504，计算设备110在预处理数据中删除重合部分，以便获得剩余部分。

[0126] 例如，预处理数据为“宣化区田志东诊所1”和参考名称“宣化区田志东诊所”，预处理数据和参考名称的重合部分为“宣化区田志东诊所”。预处理数据中删除重合部分之后的剩余部分为“1”

[0127] 在步骤506，如果计算设备110确定满足第一预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，匹配可信度等级为第一等级指示待识别原始数据和参考名称之间相匹配，第一预定条件包括以下任一项：确定剩余部分所包括的字数小于或者等于第一字数阈值；确定剩余部分所包括的字数大于第二字数阈值并且剩余部分和重合部分关联有相同的渠道类型信息，第二字数阈值大于第一字数阈值；剩余部分所包括的字数大于第一字数阈值并且小于第二字数阈值并且剩余部分包含一对括号；剩余部分包含“原”或括号和“原”；或者剩余部分包含一对括号并且括号的字数小于第三字

数阈值，第三字数阈值大于第一字数阈值并且小于第二字数阈值；确定预处理数据和参考名称存在重合部分，并且预处理数据和参考名称具有相同的渠道类型子分类。

- [0128] 关于第一字数阈值，其例如而不限于是2。例如，上述剩余部分“1”所包括的字数小于第一字数阈值，确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，例如，匹配相似度为100%，即，待识别原始数据和参考名称之间相匹配。关于第二字数阈值，其例如而不限于是10。例如，预处理数据为“七台河市源福大药房(七台河市源鸿福医疗器械商店)”和参考名称“七台河市源福大药房”之间的重合部分为“宣化区田志东诊所”。预处理数据中删除重合部分之后的剩余部分为“(七台河市源鸿福医疗器械商店)”。剩余部分所包括的字数大于10，并且剩余部分和相同部分的渠道类型信息是一样的，则确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，例如，匹配相似度为98%，即，待识别原始数据和参考名称之间高度相似，因而相匹配。
- [0129] 例如，预处理数据为“获嘉县亢村镇亢北村卫生室(原亢北联合卫生室)”和参考名称“亢村镇亢北村卫生室”之间的重合部分为“亢村镇亢北村卫生室”。预处理数据中删除重合部分之后的剩余部分为“获嘉县(原亢北联合卫生室)”。剩余部分包含“原”或“(原”，则确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，例如，匹配相似度为99%，即，待识别原始数据和参考名称之间高度相似，因而相匹配。
- [0130] 关于第三字数阈值，其例如而不限于是4。例如，预处理数据为“NE中山市三乡镇湘鸿堂药品零售店(06)”和参考名称“中山市三乡镇湘鸿堂药品零售店”之间的重合部分为“中山市三乡镇湘鸿堂药品零售店”。预处理数据中删除重合部分之后的剩余部分包含一对括号，且括号的内容长度小于4个字符，则确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，例如，匹配相似度为96%，即，待识别原始数据和参考名称之间高度相似，因而相匹配。
- [0131] 例如，预处理数据“厦门湖里叮铃医生第一门诊部有限公司”和参考名称“厦门湖里叮铃医生第一门诊部”之间的重合部分为“厦门湖里叮铃医生第一门诊部”。二者有完整相互重合的部分，并且二者有相同的渠道子分类名称，则确

定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，例如，匹配相似度为70%，即，待识别原始数据和参考名称之间具有较高相似，因而相匹配。

[0132] 关于匹配可信度等级为第一等级，其例如指示匹配相似度处于70%至100%之间。

[0133] 在步骤508，如果计算设备110确定满足第二预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第二等级，第二预定可信度条件包括：确定预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称的渠道类型分类信息相同。

[0134] 例如，预处理数据为“淮南市潘集区芦集镇卫生院(城北村卫生室)”，参考名称为“芦集镇城北村卫生室”之间存在重合部分。预处理数据“淮南市潘集区芦集镇卫生院(城北村卫生室)”的分词结果经由结构重组后为“芦集卫生(城北卫生)”，分词结果经由结构重组后的“(城北卫生)”被参考名称“芦集城北卫生”所包含，并且预处理数据和参考名称的渠道类型分类信息相同，确定待识别原始数据和参考名称之间的匹配可信度等级为第二等级，例如，匹配相似度为65%，即，待识别原始数据和参考名称之间具有一定相似性。

[0135] 在一些实施例中，匹配可信度等级为第二等级可以被认为是预处理数据和参考名称相匹配。

[0136] 在步骤510，如果计算设备110确定满足第三预定可信度条件，确定待识别原始数据和参考名称之间的不匹配，第三预定可信度条件包括：预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称渠道类型分类信息不同。

[0137] 例如，预处理数据“华润青岛医药有限公司崂山路分店”和参考名称“华润青岛医药有限公司”之间存在重合部分。预处理数据“华润青岛医药有限公司崂山路分店”和参考名称“华润青岛医药有限公司”所关联的渠道类型信息不同，参考名称“华润青岛医药有限公司”是商业公司，预处理数据“华润青岛医药有限公司崂山路分店”是药店，二者不具有相似性。则确定待识别原始数据和参考名称之间的不匹配。

- [0138] 在通过采用上述手段，本公开能够在预处理数据与参考名称存在差别的情况下，依然能够快速并准确地识别待识别原始数据和参考名称之间是否匹配。
- [0139] 以下结合图6说明用于生成分词结果的方。图6示出了根据本公开的实施例的用于生成分词结果的方法600的流程图。方法600可由如图1所示的计算设备110执行，也可以在图7所示的电子设备700处执行。应当理解的是，方法600还可以包括未示出的附加框和/或可以省略所示出的框，本公开的范围在此方面不受限制。
- [0140] 在步骤602，计算设备110基于所识别出的行政区划信息，获取待识别原始数据中除去行政区划信息之外的非行政区划数据。
- [0141] 在步骤604，计算设备110针对非行政区划数据，进行噪音词去除和等价词替换。
- [0142] 关于等价词，其例如是在医学行业中可以视为等价识别的词，或者语义等价的词。例如，“歧河乡李庄卫生室”和“歧河乡李庄卫生所”，这上述两个名称中，在分词结构上只有一个“卫生室”和“卫生所”存在差别，现实中上述两个名称通常属于一个类型，在医学行业中可以视为等价识别，属于“等价词”。
- [0143] 等价词库包括数据庞大的等价词，这些词例如是经由人工标注或者基于机器学习而分类出来的。以下表六示意性示出了等价词库中的部分等价词。

[0144] 表六

顺序号	原始词	等价词	分类
126	医药贸易公司	%公司%	连锁类
126	药店有限公司	%公司%	连锁类
140	商店有限公司	%公司%	连锁类
61	药店连锁有限公司	%公司%	连锁类
209	国药控股国大药房	%公司%	连锁类
225	零售店	%店%	连锁类
226	直营部	%店%	连锁类
252	店面	%店%	连锁类
224	加盟连锁店	%店%	连锁类
230	加盟连锁	%店%	连锁类
235	加盟店	%店%	连锁类
240	连锁店	%店%	连锁类
245	分店	%店%	连锁类

[0145] 关于针对非行政区划数据进行等价词替换的方法，其例如包括计算设备110确定多组关联词，每组关联词包括原始词和等价词，原始词和等价词具有在指示医药行业目标对象时具有一致的语义；为每组关联词确定关联的顺序号和所属分类，顺序号指示每组关联词的优先级；以及基于所确定的关联的顺序号，使用等价词替换和分割待识别原始数据，使得经由等价词替换和分割的数据中包括等价词和预定标示符，预定标示符指示分割位。

[0146] 关于预定标示符，其例如而不限于是“%”，预定标示符表示相应位置处是一个分割位。在使用等价词替换和分割时是有优先顺序的，例如，计算设备110按照“顺序号”(例如表六中所示，一般而言，长度大的原始词优先级高)，使用等价词替换和分割待识别原始数据，比如，“苏州汇仁医药商店有限公司”中的原始词“商店有限公司”被等价词%公司%所替换和分割，经由等价词替换和分割的数据例如为“苏州汇仁医药%公司”。

[0147] 在步骤606，计算设备110基于固定词库，针对经由噪音词去除和等价词替换后的数据进行分割，以便生成与待识别原始数据对应的分词结果，分词结果包括多个关键词和指示分割位的多个预定标识符。关于固定词，其例如至少包括：省份、城市、区县的全称和简称，以及其他常规固定词组。例如，机构名

称“北京师范大学附属中学卫生站”中，师范大学、附属、中学、卫生站这些词属于固定词，不需要再拆分。在一些实施例中，计算设备110进行噪音词去除和等价词替换之后，基于ASCII码表，确认经由噪音词去除和等价词替换后的数据中的每一个字符的ASCII值是否处于第一预定数值范围(例如是48~57)之外的，以便将处于第一预定数值范围(例如是48~57)之外的字符全部去掉。采用上述手段的原因主要在于：通过噪音词去除和等价词替换可以去掉一批噪音词，但机构名称中也经常出现字母、横杠等内容，中国的医疗零售机构和医疗终端不含有大小写字母或其它英文符号，因此，除中文和数字外，其它符号需要去除，因而通过上述手段，本公开能够进一步过滤噪音词。

[0148] 关于针对经由噪音词去除和等价词替换后的数据进行分割的方法，其例如包括：计算设备110在完成噪音词、等价词替换完成后，用固定词去“分割”待识别原始数据。以“苏州汇仁医药商店有限公司”为例，经由等价词替换和分割后变成“苏州汇仁医药%公司”；通过固定词库被分割后变成“苏州%汇仁%医药%公司”，其中“苏州为地理信息，其会被单独截取并另外存储，剩余部分被逐个分离出来，以便生成与待识别原始数据对应的分词结果。例如，以下表七示例了待识别原始数据“苏州汇仁医药商店有限公司”的分析结果，表八示例了待识别原始数据“医药商店有限公司(苏州汇仁)”的分词结果。表九示例了待识别原始数据“苏州汇仁医药贸易有限公司”的分词结果。

[0149] 表七

关键词 1	关键词 2	关键词 3	关键词 4	关键词 5	关键词 6	关键词 7
江苏省	苏州市	-	汇仁	医药	公司	-

[0150] 表八

关键词 1	关键词 2	关键词 3	关键词 4	关键词 5	关键词 6	关键词 7
江苏省	苏州市	-	医药	公司	汇仁	-

[0151] 表九

关键词 1	关键词 2	关键词 3	关键词 4	关键词 5	关键词 6	关键词 7
江苏省	苏州市	-	汇仁	医药	公司	-

[0152] 在步骤608，计算设备110识别待识别原始数据中的数字型的词。

细则 20.6,
01.02.2024

细则 20.6,
01.02.2024

细则 20.6,
01.02.2024

- [0153] 在步骤610，计算设备110针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形式的关键词。上文已经结合图4说明了用于分割出待识别原始数据中的数字词形式的关键词的方法，在此，不再赘述。
- [0154] 在步骤612，计算设备110把分词结果所包括的多个关键词组合成不含地理信息的预处理数据，以用于与参考名称的匹配。
- [0155] 在一些实施例中，计算设备110可以把分割后的多个关键词组合成不含地理信息的预处理数据，以用于与参考名称的匹配。例如，以“苏州汇仁医药商店有限公司”为例，所组合成不含地理信息的预处理数据例如为“汇仁医药公司”，以用于与参考名称的匹配。
- [0156] 图7示意性示出了适于用来实现本发明实施例的电子设备700的框图。电子设备700可以是用于实现执行图2至图6所示的方法200至600。如图7所示，电子设备700包括中央处理单元(即，CPU 701)，其可以根据存储在只读存储器(即，ROM 702)中的计算机程序指令或者从存储单元708加载到随机访问存储器(即，RAM 703)中的计算机程序指令，来执行各种适当的动作和处理。在RAM 703中，还可存储电子设备700操作所需的各种程序和数据。CPU 701、ROM 702以及RAM 703通过总线704彼此相连。输入/输出接口(即，I/O接口705)也连接至总线704。
- [0157] 电子设备700中的多个部件连接至I/O接口705，包括：输入单元706、输出单元707、存储单元708，CPU 701执行上文所描述的各个方法和处理，例如执行方法200至600。例如，在一些实施例中，方法200至600可被实现为计算机软件程序，其被存储于机器可读介质，例如存储单元708。在一些实施例中，计算机程序的部分或者全部可以经由ROM 702和/或通信单元709而被载入和/或安装到电子设备700上。当计算机程序加载到RAM 703并由CPU 701执行时，可以执行上文描述的方法200至600的一个或多个操作。备选地，在其他实施例中，CPU 701可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行方法200至600的一个或多个动作。
- [0158] 需要进一步说明的是，本发明可以是方法、装置、系统和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于执行本发明的各个方面的计算机可读程序指令。

- [0159] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以但不限于电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括: 便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身, 诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如, 通过光纤电缆的光脉冲)、或者通过电线传输的电信号。
- [0160] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备, 或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令, 并转发该计算机可读程序指令, 以供存储在各个计算/处理设备中的计算机可读存储介质中。
- [0161] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码, 该编程语言包括面向对象的编程语言—诸如Smalltalk、C++等, 以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中, 远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机, 或者, 可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中, 通过利用计算机可读程序指令的状态信息来个性化定制电子电路, 例如可编程逻辑电

路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)，该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0162] 这里参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

[0163] 这些计算机可读程序指令可以提供给语音交互装置中的处理器、通用计算机、专用计算机或其它可编程数据处理装置的处理单元，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理单元执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0164] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0165] 附图中的流程图和框图显示了根据本发明的多个实施例的设备、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，该模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0166] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

[0167] 以上仅为本发明的可选实施例，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等效替换、改进等，均应包含在本发明的保护范围之内。

权利要求书

- [权利要求 1] 一种用于识别待识别医药行业目标对象的方法，包括：获取用于指示医药行业目标对象的待识别原始数据；识别待识别原始数据中的行政区划信息和渠道类型信息；基于行政区划信息、渠道类型信息、以及噪音词库、语义等价词库和固定词库中的至少一个词库，针对待识别原始数据进行噪音去除和分词，以便生成分词结果，所述分词结果包括多个关键词；针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配；以及响应于确认分词结果与参考名称不匹配，针对参考名称核基于分词结果所组成的预处理数据进行语义相似性分析，以便基于相似性分析的结果识别待识别医药行业目标对象。
- [权利要求 2] 根据权利要求1所述的方法，其中针对分词结果所包括的多个关键词进行哈希计算，以便确认分词结果与参考名称是否匹配包括：计算分词结果所包括的多个关键词的哈希值之和，以便生成分词结果哈希值之和；计算参考名称所包括的多个关键词的哈希值之和，以便生成参考名称哈希值之和；确认分词结果哈希值之和参考名称哈希值之和是否相等；以及响应于确认分词结果哈希值之和参考名称哈希值之和相等，确定分词结果与参考名称匹配。
- [权利要求 3] 根据权利要求1或2所述的方法，还包括：响应于确认分词结果与参考名称相匹配，将待识别医药行业目标对象识别为与参考名称相关联的目标对象。
- [权利要求 4] 根据权利要求1所述的方法，其中生成分词结果包括：基于所识别出的行政区划信息，获取待识别原始数据中除去行政区划信息之外的非行政区划数据；针对非行政区划数据，进行噪音词去除和等价词替换；以及基于固定词库，针对经由噪音词去除和等价词替换后的数据进行分割，以便生成与待识别原始数据对应的分词结果，分词结果包括多个关键词和指示分割位的多个预定标识符。

- [权利要求 5] 根据权利要求4所述的方法，还包括：识别待识别原始数据中的数字型的词；针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形式的关键词；以及把分词结果所包括的多个关键词组合成不含地理信息的预处理数据，以用于与参考名称的匹配。
- [权利要求 6] 根据权利要求2所述的方法，其中针对所识别的数字型的词进行归一化处理，以便分割出待识别原始数据中的数字词形式的关键词包括：将待识别原始数据中的大写中文数字和/或小写中文数字转换为阿拉伯数字；确定经转换的阿拉伯数字的位数是否大于或者等于预定位数阈值；响应于确定经转换的阿拉伯数字的位数大于或者等于预定位数阈值，去除经转换的阿拉伯数字；响应于确定经转换的阿拉伯数字的位数小于预定位数阈值，确定经转换的阿拉伯数字是否位于待识别原始数据的起始位置或者终止位置；响应于确定经转换的阿拉伯数字位于待识别原始数据的起始位置或者终止位置，确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据是否指示预定渠道类型；以及响应于确定与位于起始位置或者终止位置的阿拉伯数字位相邻的数据未指示预定渠道类型，去除经转换的阿拉伯数字。
- [权利要求 7] 根据权利要求2所述的方法，其中渠道类型信息包括：渠道类型子分类名称、渠道类型分类名称和渠道类型分类序号。
- [权利要求 8] 根据权利要求1所述的方法，其中识别待识别原始数据中的行政区划信息和渠道类型信息包括：确定分别与不同优先级顺序关联的多个关键词集合，每个关键词集合包括多个预定关键词；在多个关键词集合中，确定待识别原始数据中所包括的预定关键词所在的目标关键词集合；基于目标关键词集合所关联的优先级顺序，确定与待识别原始数据匹配的渠道类型子分类名称；以及基于所确定的渠道类型子分类名称，确定与待识别原始数据匹配的渠道类型分类名称和渠道类型分类序号。

- [权利要求 9] 根据权利要求1所述的方法，其中针对待识别原始数据进行噪音去除和分词包括：确定多组关联词，每组关联词包括原始词和等价词，原始词和等价词在指示医药行业目标对象时具有一致的语义；为每组关联词确定关联的顺序号和所属分类，顺序号指示每组关联词的优先级；以及基于所确定的关联的顺序号，使用等价词替换和分割待识别原始数据，使得经由等价词替换和分割的数据中包括等价词和预定标示符，所述预定标示符指示分割位。
- [权利要求 10] 根据权利要求1所述的方法，其中针对待识别原始数据进行噪音去除和分词包括：确定预处理数据和参考名称的重合部分；在预处理数据中删除重合部分，以便获得剩余部分；响应于确定满足第一预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第一等级，匹配可信度等级为第一等级指示待识别原始数据和参考名称之间相匹配，第一预定条件包括以下任一项：确定剩余部分所包括的字数小于或者等于第一字数阈值；确定剩余部分所包括的字数大于第二字数阈值并且剩余部分和重合部分关联有相同的渠道类型信息，第二字数阈值大于第一字数阈值；剩余部分所包括的字数大于第一字数阈值并且小于第二字数阈值并且剩余部分包含一对括号；剩余部分包含“原”或括号和“原”；剩余部分包含一对括号并且括号的字数小于第三字数阈值，第三字数阈值大于第一字数阈值并且小于第二字数阈值；确定预处理数据和参考名称存在重合部分，并且预处理数据和参考名称具有相同的渠道类型子分类。
- [权利要求 11] 根据权利要求10所述的方法，其中针对分词结果和参考名称进行语义相似性分析还包括：响应于确定满足第二预定可信度条件，确定待识别原始数据和参考名称之间的匹配可信度等级为第二等级，第二预定可信度条件包括：确定预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称的渠道类型分类信息相同；响应于确定满足第三预定可信度条件，确定待识

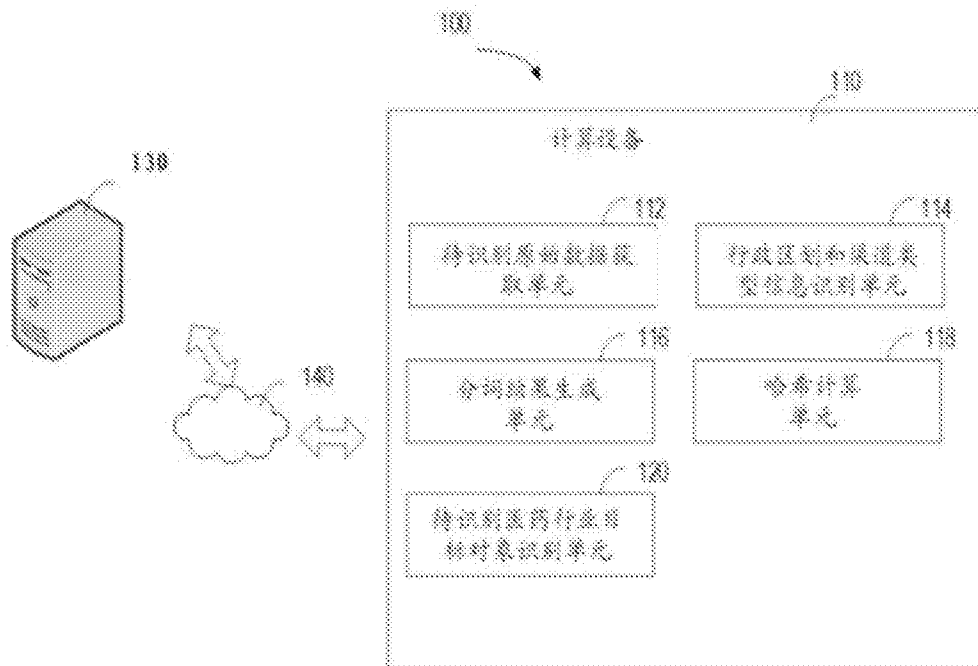
别原始数据和参考名称之间的不匹配，第三预定可信度条件包括：预处理数据和参考名称的分词结果经由结构重组后具有重合部分，并且预处理数据和参考名称渠道类型分类信息不同。

[权利要求 12] 根据权利要求1所述的方法，其中识别待识别原始数据中的行政区划信息和渠道类型信息包括：基于关于省份、城市、区县的全称、简称、曾用名和排除词来识别待识别的机构名称中的行政区划信息，行政区划信息包括省份信息、城市信息和区县信息；响应于确认所识别的区县信息或者城市信息没有指示唯一的区县或者城市，利用所识别的区县信息或者城市信息的下级行政区划信息、或者待识别目标对象的关联目标对象的行政区划信息来识别待识别原始数据中的行政区划信息。

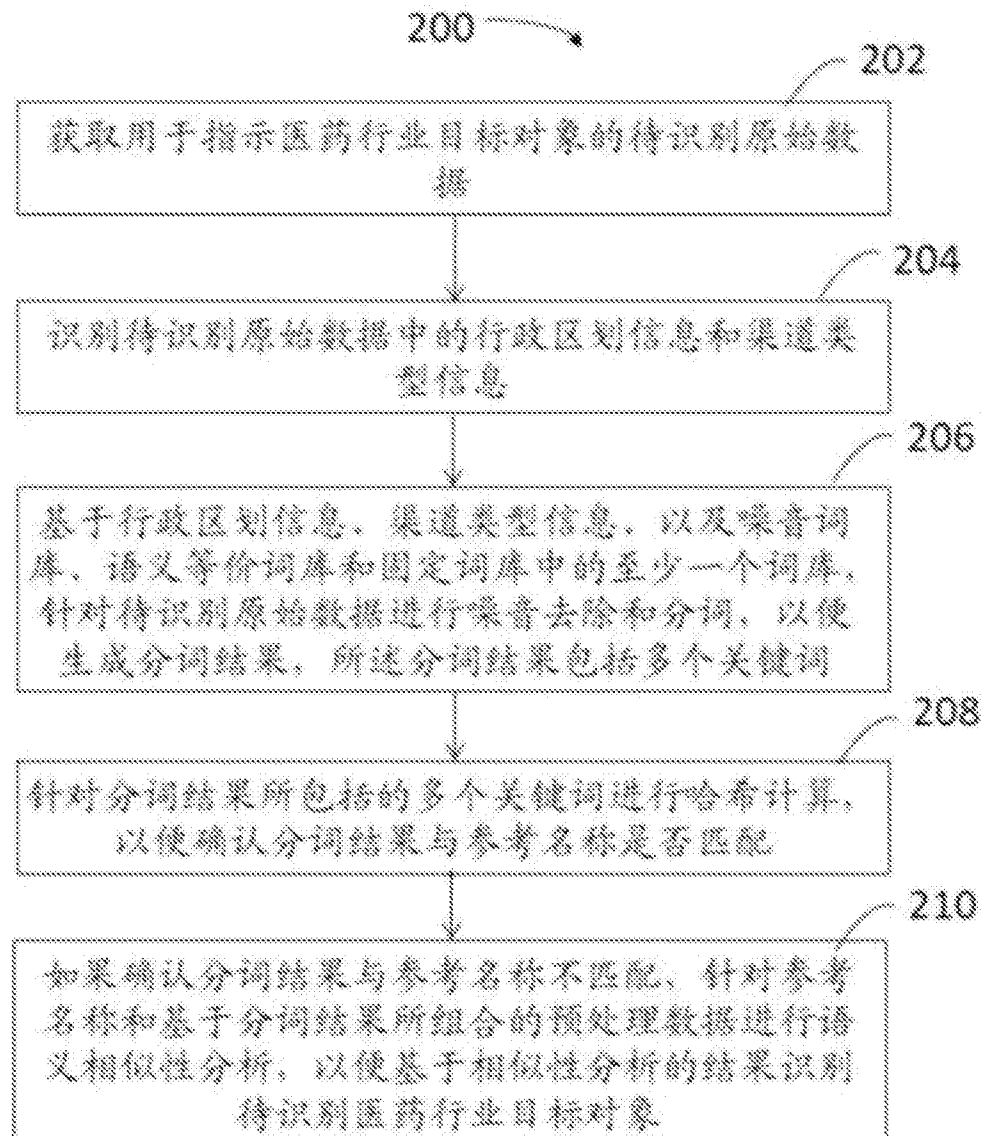
[权利要求 13] 一种计算设备，包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-12中任一项所述的方法。

[权利要求 14] 一种存储有计算机指令的非瞬时计算机可读存储介质，其中所述计算机指令用于使所述计算机执行权利要求1-12中任一项所述的方法。

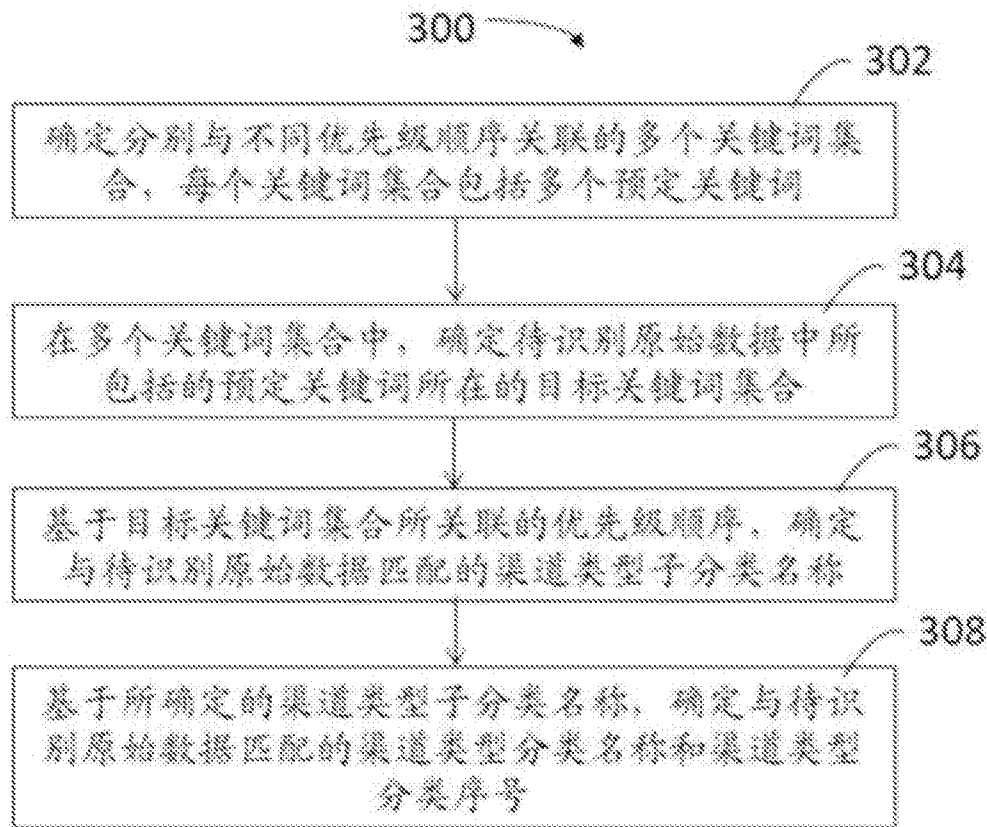
[图 1]



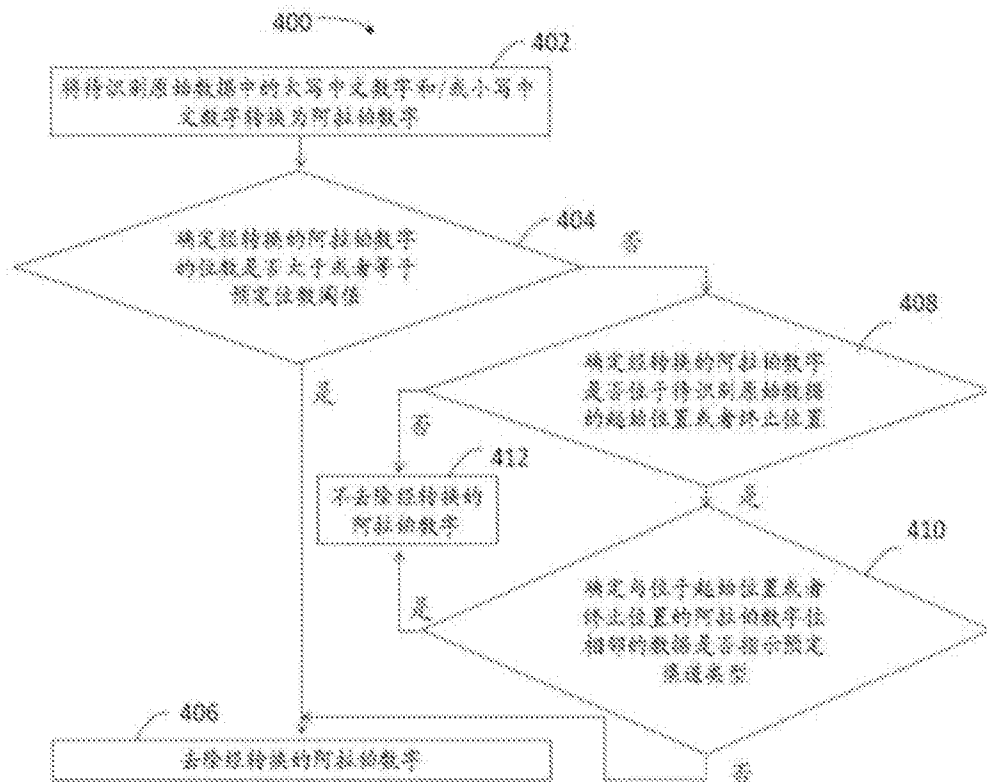
[图 2]



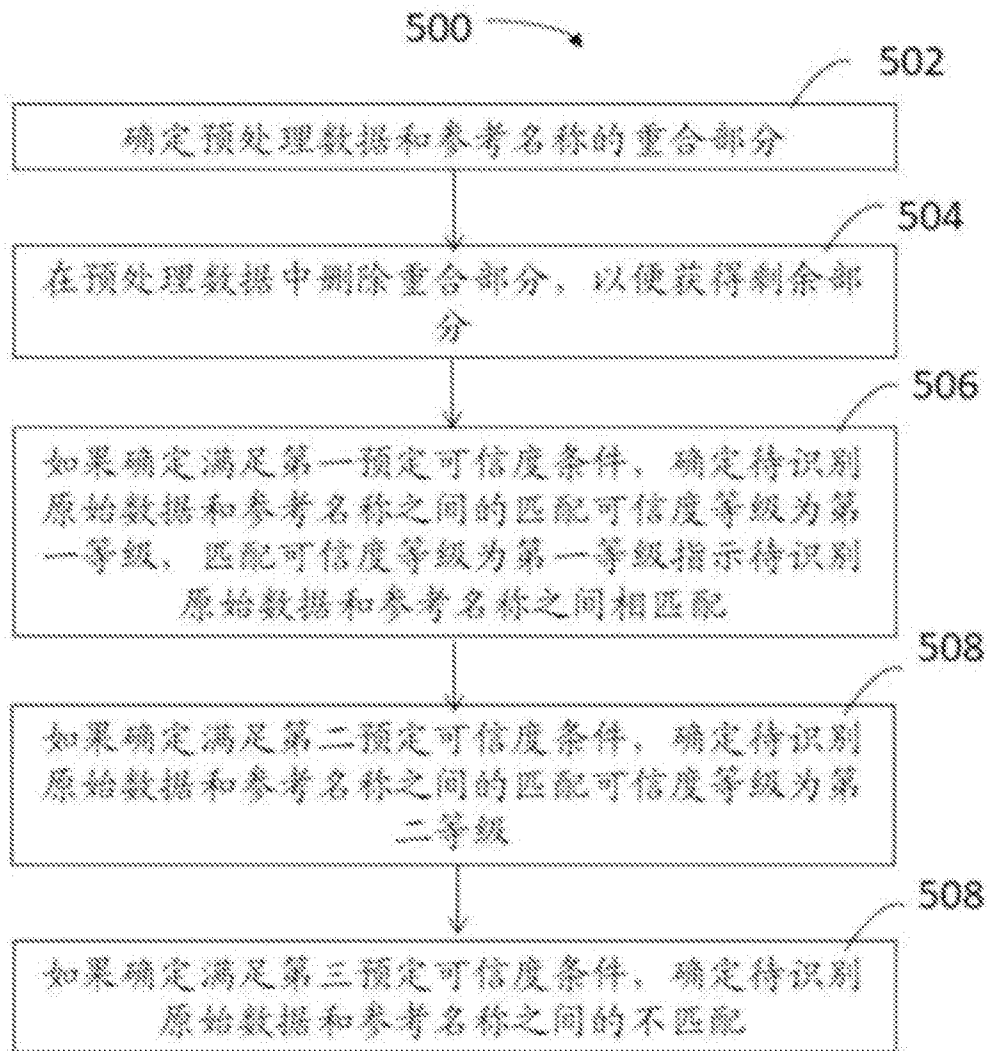
[图 3]



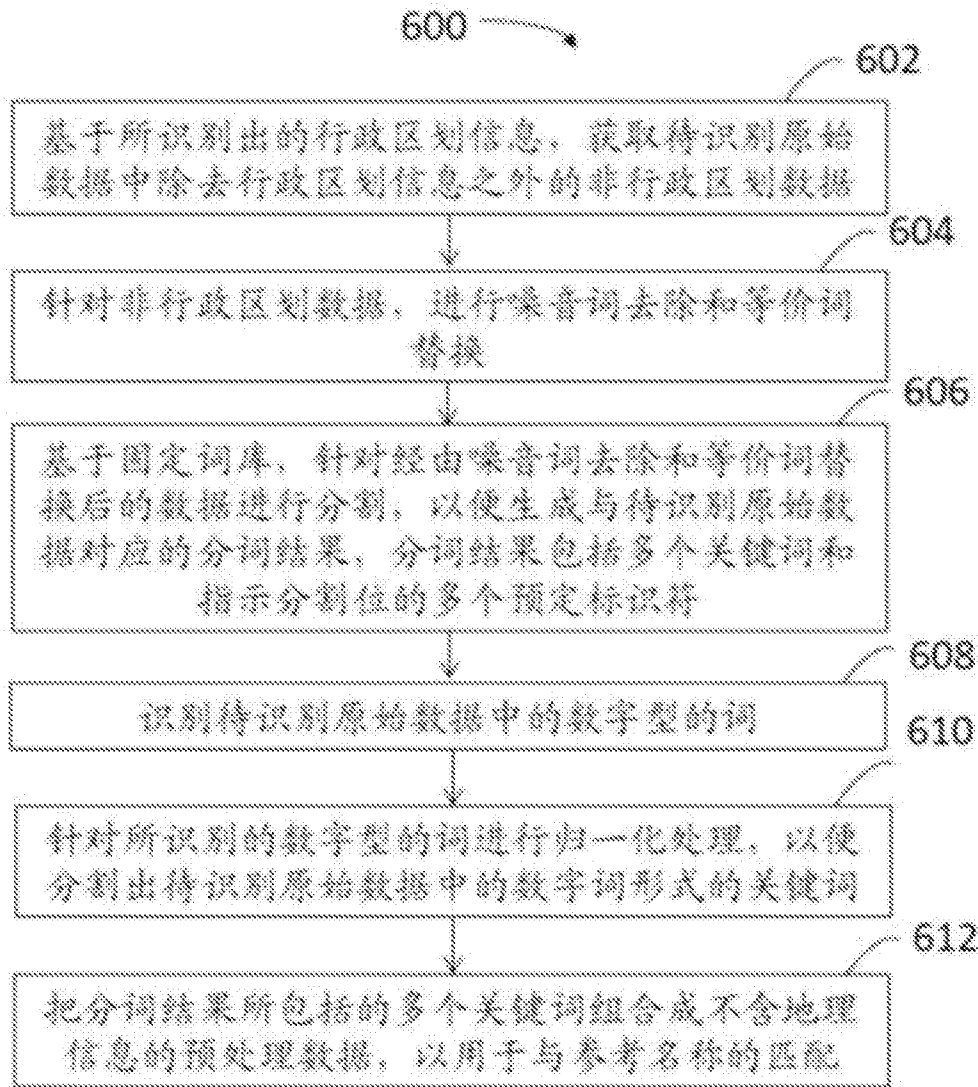
[图 4]



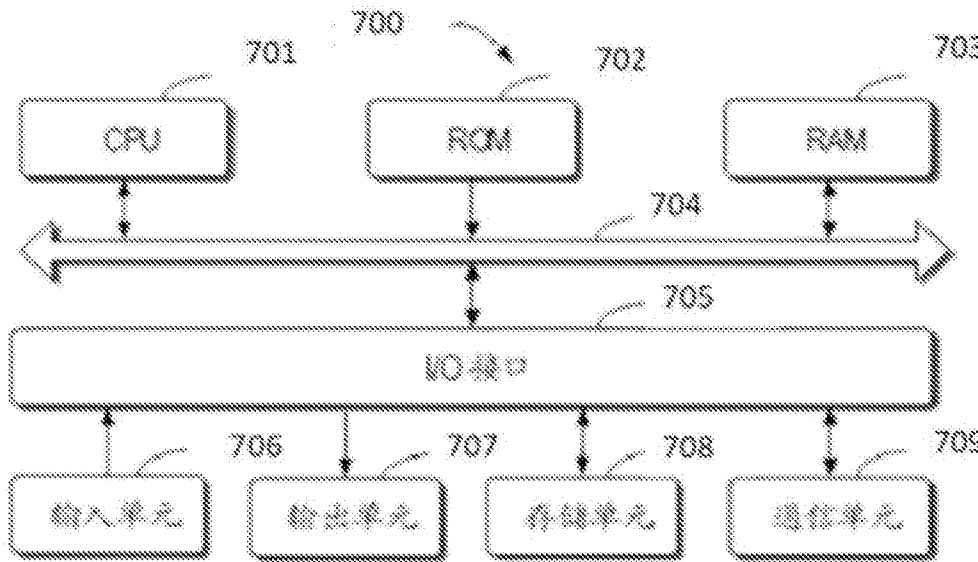
[图 5]



[图 6]



[图 7]



细则 26,
14.09.2023

细则 26,
14.09.2023

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/116180

A. CLASSIFICATION OF SUBJECT MATTER

G06F40/289(2020.01)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)

IPC: 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; DWPI; VEN; USTXT; EPTXT; WOTXT; CNKI; IEEE: 医药, 区域, 类型, 分词, 哈希, 匹配, 语义, 相似, 数字, 位置, 阈值, medicine, area, type, text segmentation, hash, match, semantics, similar, number, position, threshold

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 115730595 A (SHANGHAI HUANTONG BUSINESS TECHNOLOGY CO., LTD.) 03 March 2023 (2023-03-03) claims 1-14, and description, paragraphs 29-179	1-14
Y	CN 111899821 A (GUANGZHOU WONDFO BIOTECH CO., LTD.) 06 November 2020 (2020-11-06) description, paragraphs 49-238	1-5, 7-14
Y	CN 114911999 A (CHINA TELECOM CORP., LTD.) 16 August 2022 (2022-08-16) description, paragraphs 86-160	1-5, 7-14
Y	CN 111160012 A (SHANGHAI KINGSTAR WINNING SOFTWARE SCIENCE AND TECHNOLOGY CO., LTD.) 15 May 2020 (2020-05-15) description, paragraphs 29-45	1-5, 7-14
A	CN 112329750 A (BEIJING FOSAFER INFORMATION TECHNOLOGY CO., LTD.) 05 February 2021 (2021-02-05) entire document	1-14



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

“D” document cited by the applicant in the international application

“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

25 November 2023

Date of mailing of the international search report

06 December 2023

Name and mailing address of the ISA/CN

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

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/116180

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017132564 A1 (INTERNATIONAL BUSINESS MACHINES CORP.) 11 May 2017 (2017-05-11) entire document	1-14

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2023/116180

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	115730595	A	03 March 2023	CN	115730595	B	21 July 2023
CN	111899821	A	06 November 2020	None			
CN	114911999	A	16 August 2022	None			
CN	111160012	A	15 May 2020	None			
CN	112329750	A	05 February 2021	CN	112329750	B	20 April 2021
US	2017132564	A1	11 May 2017	US	10373103	B2	06 August 2019

A. 主题的分类 G06F40/289(2020.01)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) IPC: 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;DWPI;VEN;USTXT;EPTXT;WOTXT;CNKI;IEEE;医药, 区域, 类型, 分词, 哈希, 匹配, 语义, 相似, 数字, 位置, 阈值, medicine, area, type, text segmentation, hash, match, semantics, similar, number, position, threshold		
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
PX	CN 115730595 A (上海寰通商务科技有限公司) 2023年3月3日 (2023 - 03 - 03) 权利要求第1-14项, 说明书第29-179段	1-14
Y	CN 111899821 A (广州万孚生物技术股份有限公司) 2020年11月6日 (2020 - 11 - 06) 说明书第49-238段	1-5、7-14
Y	CN 114911999 A (中国电信股份有限公司) 2022年8月16日 (2022 - 08 - 16) 说明书第86-160段	1-5、7-14
Y	CN 111160012 A (上海金仕达卫宁软件科技有限公司) 2020年5月15日 (2020 - 05 - 15) 说明书第29-45段	1-5、7-14
A	CN 112329750 A (北京远鉴信息技术有限公司) 2021年2月5日 (2021 - 02 - 05) 全文	1-14
A	US 2017132564 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2017年5月11日 (2017 - 05 - 11) 全文	1-14
☐ 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 “D” 申请人在国际申请中引证的文件 “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 2023年11月25日	国际检索报告邮寄日期 2023年12月6日	
ISA/CN的名称和邮寄地址 中国国家知识产权局 中国北京市海淀区蓟门桥西土城路6号 100088	授权官员 戴诚 电话号码 (+86) 020-28958945	

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/116180

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	115730595	A	2023年3月3日	CN 115730595 B	2023年7月21日
CN	111899821	A	2020年11月6日	无	
CN	114911999	A	2022年8月16日	无	
CN	111160012	A	2020年5月15日	无	
CN	112329750	A	2021年2月5日	CN 112329750 B	2021年4月20日
US	2017132564	A1	2017年5月11日	US 10373103 B2	2019年8月6日