(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
2023 年 7 月 6 日 (06.07.2023)

WIPO | PCT

(10) 国际公布号
WO 2023/124191 A1

(51) 国际专利分类号:
G06F 16/906 (2019.01) *G06F 16/93* (2019.01)
G06F 16/901 (2019.01)

(21) 国际申请号: PCT/CN2022/116971

(22) 国际申请日: 2022 年 9 月 5 日 (05.09.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202111649231.1 2021年12月30日 (30.12.2021) CN

(71) 申请人: 之江实验室 (ZHEJIANG LAB) [CN/CN]; 中国浙江省杭州市余杭区中泰街道科创大道, Zhejiang 311121 (CN)。

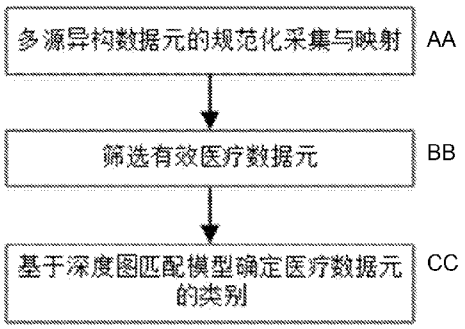
(72) 发明人: 李劲松 (LI, Jingsong); 中国浙江省杭州市余杭区文一西路1818号, Zhejiang 310023 (CN)。 辛然 (XIN, Ran); 中国浙江省杭州市余杭区文一西路1818号, Zhejiang 310023 (CN)。 杨宗峰 (YANG, Zongfeng); 中国浙江省杭州市余杭区文一西路1818号, Zhejiang 310023 (CN)。 周天舒 (ZHOU, Tianshu); 中国浙江省杭州市余杭区文一西路1818号, Zhejiang 310023 (CN)。 田雨 (TIAN, Yu); 中国浙江省杭州市余杭区文一西路1818号, Zhejiang 310023 (CN)。

(74) 代理人: 杭州求是专利事务所有限公司 (HANGZHOU QIUSHI PATENT OFFICE CO., LTD.); 中国浙江省杭州市西湖区古墩路701号绿城紫金广场C座1506室 / 刘静, Zhejiang 310030 (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, 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,

(54) **Title:** DEPTH MAP MATCHING-BASED AUTOMATIC CLASSIFICATION METHOD AND SYSTEM FOR MEDICAL DATA ELEMENTS

(54) 发明名称: 基于深度图匹配的医疗数据元自动化分类方法及系统



AA Standardized collection and mapping of multi-source heterogeneous data elements
BB Screen effective medical data element
CC Determine a type of the medical data element on the basis of a depth map matching model

图 1

(57) **Abstract:** Disclosed are a depth map matching-based automatic classification method and system for medical data elements. The present invention defines a minimum metadata information-based medical data element graph data model, and also causes a depth map matching model to be suitable for a condition of a local data swamp having extremely low metadata information, thereby achieving the purpose of completing automatic classification of data elements using the least metadata information, as well as ensuring that graph structure data collected under a graph data model standard is suitable for training the depth map matching model; a vector representation of a medical data element is calculated on the basis of a representation learning method, and an effective data element which may be mapped to a standard data model is quickly and automatically screened by means of classification of the vector representation; vector representation of a column vertex is calculated on the basis of a graph attention mechanism, and the depth map matching model is constructed to complete automatic classification of the medical data element. The method and the system of the present invention have good scalability, and can used to process various data swamp-to-data lake conversion problems.

WO 2023/124191 A1

[见续页]

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, 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))。

(57) 摘要：本发明公开了一种基于深度图匹配的医疗数据元自动化分类方法及系统，本发明定义了基于最小元数据信息的医疗数据元图数据模型，使得深度图匹配模型的效果同样适用于极低元数据信息的局部数据沼泽的情况，达到使用最少的元数据信息完成数据元自动化分类的目的，同时保证在图数据模型标准下采集的图结构数据适用于深度图匹配模型的训练；基于表示学习方法计算医疗数据元的向量表示，通过向量表示的分类，快速、自动化筛选有可能映射到标准数据模型的有效数据元；基于图注意力机制计算列顶点的向量表示，构建深度图匹配模型完成医疗数据元的自动化分类。本发明方法及系统具有良好的可拓展性，可应用于各类数据沼泽向数据湖转化问题的处理。

基于深度图匹配的医疗数据元自动化分类方法及系统

技术领域

本发明属于区域性医疗大数据中心、数据生产平台领域，尤其涉及一种基于深度图匹配的医疗数据元自动化分类方法及系统。

背景技术

随着医疗信息化的建设与发展，大数据与医疗服务的结合，促进了智慧医疗技术不断提升。目前，智慧医疗已经初具雏形，区域性医疗机构组成医联体或医共体并构建统一的医疗大数据中心已成后续智慧医疗数据治理体系发展的必然趋势。然而，医疗机构形态各异的信息平台、软件以及结构复杂的系统，导致不同机构平台之间无法实现数据的共享与交互，数据呈碎片化，形成数据孤岛。在区域性医疗机构间构建医疗大数据中心的过程中，时常发现机构内数据（尤其是历史久远的数据）缺乏管理，信息系统文档缺乏有效维护，字段备注丢失，文档质量低下，难以快速有效追溯数据血缘，形成局部的数据沼泽。传统的医疗大数据中心开发过程中，需要各医疗机构信息化部门和信息系统提供厂商相关负责人员配合医疗大数据中心的开发人员基于标准数据模型（如OMOP CDM）开发的数据接口（包括数据库视图、数据字典）完成数据发现、分类和数据关联映射任务，并完成人工分类和关联映射的数据存在标准数据模型对应的标准数据库中。数据来源的多样性，数据沼泽的密集和不可预知性普遍造成数据接口开发周期长、协调过程复杂、返工次数多等问题，耗费大量的人力物力财力，阻碍了区域性医疗大数据中心的快速自动化构建，同时为后续医疗数据的深度利用创造了很多困难。

医疗大数据中心开发过程中的数据发现、分类和数据关联映射任务，可以抽象为医疗数据元的筛选、分类任务和分类后的医疗数据元关联映射任务。首先，平台开发方案设计人员基于标准数据模型定义标准数据元分类体系和对应的数据接口规范。其后，开发人员通过规则查找和人工搜索筛选并确定与数据

接口规范匹配的数据元，这一过程称为数据发现，数据发现过程确定了平台开发过程中医疗机构数据湖内哪些数据元应该被采集；开发人员根据数据发现的结果开发数据接口，并通过完成数据采集工作。最后，开发人员将医疗机构数据湖内的多源异构的数据元按照标准数据元分类体系进行分类，整合并关联映射到标准数据元分类体系上。

现有技术缺点主要体现在以下两个方面：

1）医疗机构信息系统数量多、提供厂商来源各异，数据采集过程复杂，依赖大量人工，阻碍了医疗大数据中心的建设和大数据应用的有效开展。一家三甲级医疗机构的信息系统数量可以达到 100-300 之多，形成了一个巨大的数据湖。数据湖中数据量大，关系错综复杂，决定了数据接口开发阶段的数据发现工作需要依赖医疗机构信息化部门和信息系统提供厂商相关负责人员的长期配合，数据接口之间相互衔接，造成数据发现工作的人工成本大，耗费时间长。中间环节一旦出现故障，问题的排查过程非常复杂。很大程度上阻碍了医疗大数据中心的开发和大数据应用的有效开展。

2）医疗机构信息系统更迭频繁，历史系统文档维护困难、缺失严重等常见问题在医疗机构的数据湖内形成局部的数据沼泽，进一步增加了数据接口开发的难度。医疗数据包含病人诊疗过程中生成的诊疗数据和医疗机构运营过程中的观测数据，来源多样，关系复杂。随着医疗机构信息系统版本的更迭，历史数据沉睡在医疗机构数据湖中缺乏有效管理，形成局部的数据沼泽。医疗大数据中心的构建需要对这些历史数据进行整合，完成数据沼泽向数据湖的转化。由于医疗机构信息化部门和信息系统提供厂商相关负责人员更替频繁，历史系统文档丢失情况时有发生，面对文档丢失，数据接口开发人员只能依靠重复试错的方法对医疗机构数据湖中所有可能的数据进行人工筛选来完成数据发现，由于医疗机构信息系统的数量多，关联关系复杂，人工筛选的方法难以有效利用医疗机构数据湖的全局信息，耗时长，错误率高，大幅增加了数据发现工作的工作周期和难度。当数据湖内数据间的关联结构过于复杂超过人工能接受的

程度时，只能放弃对应数据接口的开发，使得对应类别的数据无法找到可关联映射的数据，造成该分类的数据丢失。

发明内容

医疗大数据中心的构建过程中，医疗机构局部数据沼泽普遍存在等问题导致数据接口开发时间长，维护困难。传统的解决方案依赖人工处理，难以大规模完成海量数据的数据发现、分类和关联映射问题。医疗机构数据湖内的多源异构的数据可以抽象为由未知分类的数据元组成的待筛选医疗数据元集合。过去的几年里，图神经网络的兴起与应用成功推动了图结构数据的深度学习范式的发展。

本发明利用基于图神经网络的深度图匹配算法，改进基于人工处理的数据元分类方法，最大程度降低对于信息系统数据文档的依赖，在只获取医疗机构数据湖内极少元数据信息的条件下，基于医疗数据文本语义实现有效数据元的快速筛选，实现医疗机构数据湖内数据的自动化数据发现，基于深度图匹配算法实现医疗数据元的快速分类，实现医疗机构数据湖内数据元向标准数据元分类体系的自动化分类和关联映射，大幅度提升医疗大数据中心开发过程中数据接口开发的效率。本发明提供的数据元分类方法具有良好的可拓展性，可应用于各类数据沼泽向数据湖转化问题的处理。

本发明的目的是通过以下技术方案来实现的：

本发明一方面公开了一种基于深度图匹配的医疗数据元自动化分类方法，该方法包括以下步骤：

(1)定义基于最小元数据信息的医疗数据元图数据模型；将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向所述医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；

(2)计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，由有效列顶点

集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；

(3)从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类。

进一步地，所述医疗数据元图数据模型采用有向属性图建模，图由顶点和边两种图元素构成；

所述顶点是由标签和对应标签的属性组构成的，标签代表顶点的类型，属性组代表标签拥有的一种或多种属性；所述顶点的本体信息包含顶点类型及每类顶点对应的属性信息，所述顶点类型包括数据库顶点、表顶点和列顶点，所述数据库顶点对应的属性信息包括数据库顶点索引和数据库类型信息，所述表顶点对应的属性信息包括表顶点索引，所述列顶点对应的属性信息包括列顶点索引、列数据类型信息和列向量表示；

所述边是由边类型和边属性构成的，每一条边均为有向边；所述边的本体信息包含边类型及每类边对应的属性信息，所述边类型包括起点为数据库顶点、终点为表顶点的父子关联，起点为表顶点、终点为列顶点的父子关联，以及起点和终点均为列顶点的外键，三种边类型对应的属性信息均为边索引。

进一步地，所述多源异构的数据元向医疗数据元图数据模型的映射，包括：

将来自多源异构的医疗数据从数据湖中采集，组成待筛选医疗数据元集合；

使用元数据采集工具对数据湖中存储的元数据进行抓取；

使用列向量生成器，对待筛选医疗数据元集合中各表各列中存储的数据进行遍历，利用列向量表示模型预测得到各表各列的列向量表示；

通过图数据关联映射，将采集的元数据和产生的列向量表示向医疗数据元图数据模型关联映射，得到待筛选医疗数据元图数据。

进一步地，所述列向量生成器以数据表中的单列作为一个数据元单位，使用列向量表示模型转化各列存储的数据，计算各列的向量表示；

所述列向量表示模型的训练包括：列向量表示模型的训练数据为存储在标准数据库中的人工完成医疗数据元分类、数据结构符合标准数据模型的列数据，记为标准分类列；标准分类医疗数据元图数据中的列顶点与对应标准分类列存在一一对应关系；

设标准分类医疗数据元图数据中列顶点集合为$C=\{c_{k,j}\}$，其中$c_{k,j}$表示列顶点集合对应的标准分类列中第k列，第j行的数据，$c_{k,j}=\{w_t\}_{t=1,2,...,m}$，$m$为第$j$行字符总数，$w_t$为构成数据$c_{k,j}$的字符；通过文本表示模型$h$计算得到字符$w_t$的初始向量表示$h(w_t)$；在标准分类医疗数据元图数据的列顶点$C_k$下随机抽取$n$行数据$\{c_{k,j}\}_{j=1,2,...,n}$，第$j$行数据的向量表示为$v(c_{k,j})=\frac{1}{m}\sum_{t=1}^{m}h(w_t)$，根据自注意力机制计算得到标准分类医疗数据元图数据中列顶点C_k下各行数据的相关性，得到列顶点C_k的列向量表示$H(C_k)$，计算公式为：

$$v(C_k)=\frac{1}{n}\sum_{j=1}^{n}v(c_{k,j})$$

$$H(C_k)=softmax\left(\frac{v(C_k)v(C_k)^T}{\sqrt{d_k}}\right)v(C_k)$$

其中$v(C_k)$为列顶点C_k的向量表示，d_k为$v(C_k)$的维度，*softmax* 为 softmax 函数；

所述列向量表示模型的预测包括：列向量表示模型的预测数据为数据湖中各数据库中各表各列所组成的待筛选医疗数据元集合，以列为遍历单元对待筛选医疗数据元集合进行遍历；使用列向量表示模型计算对列顶点每次随机抽样的列向量表示；对预测的多次随机抽样的列向量表示结果求平均值，作为所述列顶点最终的列向量表示。

进一步地，所述计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度，包括：

对于待筛选医疗数据元图数据中存储的列顶点C_k，在除去C_k的列顶点集合中随机抽取p个列顶点$\{C_t\}_{t=1,2,\ldots,p}$，通过计算列顶点C_k与抽取的列顶点的相关性，计算C_k在医疗数据元图数据模型中的重要度分数$Im(C_k)$，$Im(C_k)$定义为：

$$H(\{C_t\}) = \frac{1}{p}\sum_{t=1}^{p} H(C_t)$$

$$Im(C_k) = Importance\_score(C_k, \{C_t\}) = \frac{H(C_k)\cdot H(\{C_t\})}{\|H(C_k)\| \times \|H(\{C_t\})\|}$$

其中$Importance_score$为重要度函数。

进一步地，所述医疗数据元筛选模型的训练与预测具体为：

将根据标准数据元分类体系，人工分类和关联映射构建的标准分类医疗数据元集合转换为标准分类医疗数据元图数据，设标准分类医疗数据元图数据中存储的列顶点集合为$S = \{s_k\}$，设构建标准分类医疗数据元集合过程中被人工筛选排除的列对应的列顶点集合为$S' = \{s'_k\}$；

训练时从集合S中随机抽取q个列顶点作为正样本集合$\{s_t\}_{t=1,2,\ldots,q}$，从集合S'中随机抽取q个列顶点作为负样本集合$\{s'_t\}_{t=1,2,\ldots,q}$；设样本$(s_i, y_i)$的重要度分数为$Im(s_i)$，$s_i$表示第$i$个列顶点，$y_i \epsilon \{0,1\}$表示样本真实类别，则基于重要度分数计算医疗数据元筛选模型的损失函数$Loss$：

$$Loss = -\frac{1}{2*q}\sum_{i=1}^{2\times q}(y_i \log(Im(s_i)) + (1-y_i)\log(1-Im(s_i)))$$

所述医疗数据元筛选模型在预测时，通过计算阈值L'判断列顶点C_k对应的待筛选医疗数据元集合中的列是否为有效数据元，阈值L'计算公式：

$$L' = \frac{1}{1+e^{-Im(C_k)}}$$

若$L' \geq 0.5$，则说明列顶点C_k为有效列顶点，对应的列为有效数据元；

由筛选后的有效列顶点集合关联组成待分类医疗数据元图数据，对应的筛选后的列集合组成待分类医疗数据元集合。

进一步地，所述从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合，包括：

设由标准数据模型定义的标准数据元分类体系中所有标准分类集合为E，标准分类医疗数据元图数据中的列顶点集合为D，$D_i \in D$在标准数据元分类体系中的分类为$E_i \in E$；设待分类医疗数据元图数据中存储的列顶点集合为C；医疗数据元分类过程抽象为在D中找到与列顶点$C_k \in C$匹配度最高的列顶点D_i，从而确定列顶点C_k对应的列的分类为E_i；

对于列顶点$D_i \in D$，从D_i对应的列中随机抽取r_0个数据$(d_{i,j})_{j=1,2,\ldots,r_0}$，对于列顶点$C_k \in C$，从$C_k$对应的列中随机抽取$r_0$个数据$(c_{k,l})_{l=1,2,\ldots,r_0}$，则$D_i$和$C_k$的匹配度$match_1(D_i, C_k)$为：

$$match\_1(D_i, C_k) = \frac{1}{{r_0}^2} \sum_{j=1}^{r_0} \sum_{l=1}^{r_0} \frac{v(d_{i,j}) \cdot v(c_{k,l})}{\left\| v(d_{i,j}) \right\| \times \left\| v(c_{k,l}) \right\|}$$

其中$v(x)$代表数据x的向量表示，则D_i对应的种子顶点为与其匹配度最高的列顶点$\hat{C}_i$，即：

$$\hat{C}_i = arg \max_{C_k \in C} \left(match\_1(D_i, C_k) \right) 。$$

进一步地，所述基于种子顶点集合进行待分类医疗数据元图数据的子图切割，包括：

以$sch(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在父子关系的列顶点集合，以$ext(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在外键关系的列顶点集合，则基于种子顶点$\hat{C}_i$切割得到的子图$slice(\hat{C}_i)$为：

$$slice(\hat{C}_i) = sch(\hat{C}_i) \cup ext(\hat{C}_i)$$

以$N(D_i)$表示标准分类医疗数据元图数据中与D_i关联同一父顶点的列顶点集合，则深度图匹配模型的目标是从子图$slice(\hat{C}_i)$中搜索子图，使得搜索到的子

图中的列顶点与$N(D_i)$中的列顶点一一匹配，实现$slice(\hat{C}_i)$中列顶点对应的医疗数据元的分类。

进一步地，所述利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，包括：

根据图注意力机制，计算标准分类医疗数据元图数据中列顶点D_i的向量表示$V(D_i)$为：

$$V(D_i)=\sum_{D'\in N(D_i)} w(D',D_i)\cdot v'(D')$$

其中$v'(D')=\frac{1}{r_1}\sum_{j=1}^{r_1} v(d'_j)$，$(d'_j)_{j=1,2,\ldots,r_1}$为从列顶点$D'$对应的列中随机抽取$r_1$个数据；$w(D',D_i)$表示$N(D_i)$中的某一列顶点$D'$对于列顶点$D_i$的权重函数；

根据图注意力机制，计算待分类医疗数据元图数据的列顶点$\hat{C}_i$的向量表示$V(\hat{C}_i)$为：

$$V(\hat{C}_i)=\sum_{C'\in slice(\hat{C}_i)} w(C',\hat{C}_i)\cdot v'(C')$$

其中$v'(C')=\frac{1}{r_1}\sum_{j=1}^{r_1} v(c'_j)$，$(c'_j)_{j=1,2,\ldots,r_1}$为从列顶点$C'$对应的列中随机抽取$r_1$个数据；$w(C',\hat{C}_i)$表示$slice(\hat{C}_i)$中的某一列顶点$C'$对于列顶点$\hat{C}_i$的权重函数；

列顶点$D'\in N(D_i)$和列顶点$C'\in slice(\hat{C}_i)$的匹配度$match_2(D',C')$为：

$$match\_2(D',C')=\frac{V(D')\cdot V(C')}{\|V(D')\|\times\|V(C')\|}$$

取与C'匹配度最高的列顶点$\hat{D}'$，即：

$$\hat{D}'=arg\max_{D'\in N(D_i)} match\_2(D',C')$$

待分类医疗数据元图数据中的列顶点C'对应的列的分类为$\hat{D}'$对应的标准数据元分类体系中的类别。

本发明另一方面公开了一种基于深度图匹配的医疗数据元自动化分类系统，该系统包括：

多源异构数据元的规范化采集与映射模块：定义基于最小元数据信息的医疗数据元图数据模型；将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向所述医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；

有效医疗数据元筛选模块：计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，对应的列为有效医疗数据元，由有效列顶点集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；

基于深度图匹配模型的医疗数据元分类模块：从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类。

本发明的有益效果是：

1）本发明只利用了医疗机构数据湖中存储的极少的元数据信息，使用医疗数据元图数据模型实现医疗机构内医疗数据元的规范化采集和待筛选、分类医疗数据元之间关系信息的充分利用。

2）本发明方法缩小了数据发现、分类和关联映射过程对医疗机构信息系统历史文档的依赖，历史文档的缺失、错误对于医疗数据元的分类结果影响较小。

3）本发明方法大幅度减少了人工对数据发现、分类和关联映射过程的干预，通过人工智能算法对待分类医疗数据元进行分类，为医疗大数据中心数据的实时更新和动态汇聚、深度利用需求中存在的医疗数据元自动化分类难题提供了启发式的解决方案。

附图说明

图 1 为本发明方法整体流程图；

图 2 为传统医疗数据元分类方法流程图；

图 3 为本发明提供的基于深度图匹配的医疗数据元自动化分类方法实现过程示意图；

图 4 为医疗数据元图数据模型的一个示例；

图 5 为多源异构数据元向医疗数据元图数据模型的映射示意图。

具体实施方式

为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图对本发明的具体实施方式做详细的说明。

在下面的描述中阐述了很多具体细节以便于充分理解本发明，但是本发明还可以采用其他不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本发明内涵的情况下做类似推广，因此本发明不受下面公开的具体实施例的限制。

以下首先对本发明中涉及的术语进行说明：

元数据：描述其它数据的数据。元数据是关于数据的数据，在某些时候不特指某个单独的数据，可以理解为是一组用来描述数据的信息组/数据组，该信息组/数据组中的一切数据、信息，都描述/反映了某个数据的某方面特征，则该信息组/数据组可称为一个元数据。元数据可以为数据说明其元素或属性（名称、大小、数据类型等），或其结构（长度、字段、数据列），或其相关数据（位于何处、如何联系、拥有者）。在日常生活中，元数据无所不在。只要有一类事物，就可以定义一套元数据。

数据元：可理解为数据的基本单元。卫生信息基本数据元规范和定义了医药卫生领域所有相关信息的唯一中文名称与代码，并且代码以字母、汉字、数字式的字符串形式表示。数据元列举并定义了特定语义环境中的一种信息资源。完整的数据元名称=对象类术语+特征类术语+表示类术语+（限定类术语）。

数据元与元数据的区别和联系：元数据不可能涵盖理解数据元所要表示的数据所必需的所有信息。数据元的相关信息是任何一个（组织的）元数据的一个完整的组成部分。元数据的每一个元素都是一个数据元，用符合数据元标准的元数据属性和描述方法来说明元数据。将元数据存储于一个库中，并使之条

理化就需要建模，建模就需要从数据元的注册系统中或库中获取元数据。元数据，它是以一种一致、标准的方式来表达的数据元。元数据与数据元字典格式均由行号、中文名称、英文名称、标识符（短语）、定义、约束／条件、最大出现次数、数据类型、数据的值域等属性组成。不同之处是数据元字典格式中另有语境和同义词名称等属性。

数据湖：数据湖是一种在系统或存储库中以自然格式存储数据的方法，它有助于以各种模式和结构形式配置数据，通常是对象块或文件。数据湖的主要思想是对企业中的所有数据进行统一存储，从原始数据（源系统数据的精确副本）转换为用于报告、可视化、分析和机器学习等各种任务的目标数据。国内一般把整个 HDFS 叫做数据仓库（广义），即存放所有数据的地方，而国外一般叫数据湖（data lake）。当数据湖缺乏管理的时候，就会形成数据沼泽。搭建数据湖容易，但是让数据湖发挥价值是很难的。最终数据湖只是一直往里面灌数据，而应用场景极少，没有输出或者极少输出，形成单向湖。大部分使用数据湖的企业在数据真的需要使用的时候，往往因为数据湖中的数据质量太差而无法最终使用。

图神经网络：在过去的几年中，神经网络的兴起与应用成功推动了模式识别和数据挖掘的研究。许多曾经严重依赖于手工提取特征的机器学习任务（如目标检测、机器翻译和语音识别），如今都已被各种端到端的深度学习范式彻底改变了。尽管传统的深度学习方法被应用在提取欧氏空间数据的特征方面取得了巨大的成功，但许多实际应用场景中的数据是从非欧式空间生成的，传统的深度学习方法在处理非欧式空间数据上的表现却仍难以使人满意。图中的每个数据样本（节点）都会有边与图中其他实数据样本相关，这些信息可用于捕获实例之间的相互依赖关系。图神经网络是应用于图结构数据（非欧式空间）上的神经网络。

深度图匹配：图匹配是人工智能中的一个经典问题，在若干领域都有重要的应用，比如计算机视觉中匹配 2D/3D 形状，生物信息学中匹配蛋白质网络，

社交网络中匹配不同网络当中的用户等。深度图匹配即基于图神经网络解决图匹配问题的方法。

如图 1 所示，本发明提供了一种基于深度图匹配的医疗数据元自动化分类方法，该方法包括以下步骤：

（1）多源异构数据元的规范化采集与映射，包括：

定义基于最小元数据信息的医疗数据元图数据模型；

将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；

（2）计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，由有效列顶点集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；

（3）从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类。

图 2 为传统医疗数据元分类方法流程图。以下参见图 3 详细描述本发明方法各部分的实现过程。

一、多源异构数据元的规范化采集与映射

1.1 医疗数据元图数据模型的定义

医疗机构数据汇聚形成数据湖，数据湖的数据具有多源异构的特性，包括医疗过程中对诊疗过程和医疗机构运营过程的观测数据，观测数据库的目的和设计各不相同。诊疗过程形成的电子病历旨在支持临床实践，而医疗机构运营数据则是为院内管理和医保报销流程构建的。每一种都是为了不同的目的而收集的，导致数据具有不同的逻辑组织和物理格式。

数据模型是数据库设计中用来对现实世界进行抽象的工具，通过建立标准统一的数据模型，定义数据结构、数据操作、数据约束，可以有效保证采集的数据质量和数据表征的标准可控，图数据模型是基于图数据库开发的数据模型。

由于数据湖中数据库类型不同，数据表、数据列间关系复杂。医疗机构内的观测数据时间跨度大，普遍存在数据库文档信息缺失的现象。为了使得本发明提及的深度图匹配模型的效果同样适用于极低元数据信息的局部数据沼泽的情况，达到使用最小的元数据信息完成数据元自动化分类的目的，同时保证在图数据模型标准下采集的图结构数据适用于深度图匹配模型的训练，本发明基于数据湖内数据库的最小元数据信息，定义了一种基于最小元数据信息的医疗数据元图数据模型，为医疗大数据中心建立过程中医疗数据元的自动化分类提供了一种启发式的解决方案。

图数据模型采用有向属性图来建模，图由两种图元素构成：顶点 Vertex 和边 Edge。其中顶点由标签和对应标签的属性组构成，标签代表顶点的类型，属性组代表标签拥有的一种或多种属性。顶点的本体信息包含顶点类型及每类顶点对应的属性信息。

本发明定义的医疗数据元图数据模型的顶点的本体信息如下表所示：

表 1 医疗数据元图数据模型的顶点的本体信息表

顶点标签	顶点描述	属性	属性说明
Database	数据库顶点	vid	数据库顶点索引
		DatabaseType	数据库类型信息
Table	表顶点	vid	表顶点索引
Column	列顶点	vid	列顶点索引
		type	列数据类型信息
		vector_embeddings	列向量表示

其中 vid 为图中每一顶点的唯一索引 id，可统一使用哈希散列编码。vector_embeddings 为列向量表示模型预测的列向量表示结果。

在图数据模型中，边由边类型和边属性构成，每一条边均为有向边，有向边表明一个顶点（起点 src）指向另一个顶点（终点 dst）的关联关系。边的本体信息包含边类型及每类边对应的属性信息。

本发明定义的医疗数据元图数据模型的边的本体信息如下表所示：

表 2 医疗数据元图数据模型的边的本体信息表

起点标签	终点标签	边类型	属性	属性说明
Database	Table	父子关联	eid	边索引
Table	Column	父子关联	eid	边索引
Column	Column	外键	eid	边索引

图 4 为医疗数据元图数据模型的一个示例。

1.2 多源异构数据元向医疗数据元图数据模型的映射

本发明的数据采集与关联映射过程，将来自多源异构的医疗数据从数据湖中采集，组成待筛选医疗数据元集合。使用元数据采集工具对数据湖中存储的元数据进行抓取。使用列向量生成器，对待筛选医疗数据元集合中各表各列中存储的数据进行遍历，利用列向量表示模型预测得到各表各列的列向量表示。最后通过图数据关联映射，将采集的元数据和产生的列向量表示向医疗数据元图数据模型关联映射，得到待筛选医疗数据元图数据。参见图 5，具体实现描述如下：

（1）元数据采集工具

a)数据库适配：由于医疗机构内数据湖通常包含不同类型数据库，元数据采集工具需针对不同类型数据库开发数据库适配模块实现适配。

b)解析配置：由于最终的关联映射目标为医疗数据元图数据模型，采集信息配置为仅采集元数据中的表格列信息、血缘关系信息和各列的外键信息；对于主键、约束、索引、权限、触发器等常见元数据则不在采集范围之内。

c)元数据抓取：针对解析配置情况，对数据湖内的各数据库执行元数据抓取操作。

d)数据关联：针对数据库适配情况，将不同类型数据库的字段类型统一映射到图数据库数据类型上。如 oracle 数据库的 varchar2 类型和 MySQL 数据库的 varchar 类型统一映射为图数据库的 string 类型，其他类型数据库同理。

（2）列向量生成器

列向量生成器以数据表中的单列作为一个数据元单位，使用列向量表示模型转化各列存储的数据，计算各列的向量表示；

a)列向量表示模型的训练

列向量表示模型的训练数据为存储在标准数据库中的人工完成医疗数据元分类、数据结构符合标准数据模型的列数据，简称为标准分类列。

标准分类医疗数据元图数据中的列顶点与对应标准分类列存在一一对应关系。

获得医疗数据元图数据中列顶点向量表示的方法，是将对应医疗数据元集合中的列中存储的数据转化为文本数据，每列文本数据头尾分别加上[CLS]、[SEP]表示数据的开头和结束。

设标准分类医疗数据元图数据中列顶点集合为$C=\{c_{k,j}\}$，其中$c_{k,j}$表示列顶点集合对应的标准分类列中第k列，第j行的数据，$c_{k,j}=\{w_t\}_{t=1,2,\ldots,m}$，$m$为第$j$行字符总数，$w_t$为构成数据$c_{k,j}$的字符。通过文本表示模型$h$计算得到字符$w_t$的初始向量表示$h(w_t)$。文本表示模型$h$可以采用基于 Transformer 模型的深度双向语言表示模型（BERT 模型）。在标准分类医疗数据元图数据的列顶点C_k下随机抽取n行数据$\{c_{k,j}\}_{j=1,2,\ldots,n}$，第$j$行数据的向量表示为$v(c_{k,j})=\frac{1}{m}\sum_{t=1}^{m}h(w_t)$，根据自注意力机制（self-attention）计算得到标准分类医疗数据元图数据中列顶点C_k下各行数据的相关性，得到列顶点C_k的列向量表示$H(C_k)$，计算公式为：

$$v(C_k)=\frac{1}{n}\sum_{j=1}^{n}v(c_{k,j})$$

$$H(C_k)=softmax\left(\frac{v(C_k)v(C_k)^T}{\sqrt{d_k}}\right)v(C_k)$$

其中$v(C_k)$为列顶点C_k的向量表示，d_k为$v(C_k)$的维度，*softmax* 为 softmax 函数。

为获得更精确的列顶点向量表示，在积累了足够量的标准分类列作为训练数据的情况下，可以使用标准分类列数据对列向量表示模型进行进一步的迁移学习。以列为单位，随机覆盖对应列数据中 15%的字符，使用[MASK]标签替带被覆盖字符。使用列向量表示模型预测被覆盖字符进一步训练和更新模型，这样得到的列向量表示模型更加匹配筛选有效数据元的任务。

b)列向量表示模型的预测

列向量表示模型的预测数据为数据湖中各数据库中各表各列所组成的待筛选医疗数据元集合，以列为遍历单元对待筛选医疗数据元集合进行遍历。为避免待筛选医疗数据元集合中存在列数据量过大导致列向量生成器性能下降，在使用列向量表示模型计算列向量表示过程中，可以使用随机抽样的方式（如随机抽取单列 1000 个数据，抽取 100 次），使用列向量表示模型计算对列顶点C_k进行第 *s* 次抽样的列向量表示$H_s(C_k)$。对预测的共S次抽样的列向量表示结果求平均值，作为C_k最终的列向量表示$H(C_k)=\frac{1}{S}\sum_{s=1}^{S}H_s(C_k)$，存储$H(C_k)$在医疗数据元图数据模型列顶点$C_k$的 vector_embeddings 属性内。

（3）图数据关联映射

将计算得到的待筛选医疗数据元集合中各列的列向量表示，以及元数据采集结果，分别关联映射为医疗数据元图数据模型中顶点和边对应的对象，入库到以医疗数据元图数据模型为数据标准的待筛选医疗数据元图数据中，对应的映射关系如下表所示。

表 3 图数据关联映射表

序号	映射对象	对象属性	元数据信息
1	Database	顶点	医疗机构内数据库名称（编号）
2	Table	顶点	数据库内数据表名称（编号）
3	Column	顶点	数据表内列名称（编号）
4	Database-Table	边	数据库和数据表的从属关系
5	Table-Column	边	数据表和表内列的包含关系
6	Column-Column	边	数据库列外键，列间血缘关系

二、快速、自动化筛选有效医疗数据元

医疗机构内数据湖存储的信息类型繁多，相比于标准数据模型的数据覆盖范围，通常存在大量信息冗余，为了快速、自动化筛选有效医疗数据元，在进行医疗数据元自动化分类任务之前，可以对待筛选医疗数据元集合中的数据元进行筛选，降低数据元分类任务的复杂度。本发明提出如下快速、自动化筛选有效医疗数据元的方法，包括以下两个步骤：（1）计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度。（2）构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出其中的有效医疗数据元，组成待分类医疗数据元集合。

2.1 基于列顶点向量表示计算列顶点在医疗数据元图数据模型中的重要度

待筛选医疗数据元图数据中存储的列顶点与待筛选医疗数据元集合中的列存在一一对应关系。对于待筛选医疗数据元图数据中存储的列顶点C_k，在除去C_k的列顶点集合中随机抽取p个列顶点$\{C_t\}_{t=1,2,\ldots,p}$，通过计算列顶点C_k与抽取的列顶点的相关性，计算C_k在医疗数据元图数据模型中的重要度分数$Im(C_k)$，$Im(C_k)$定义为：

$$H(\{C_t\}) = \frac{1}{p}\sum_{t=1}^{p} H(C_t)$$

$$Im(C_k) = Importance\_score(C_k, \{C_t\}) = \frac{H(C_k)\cdot H(\{C_t\})}{\|H(C_k)\| \times \|H(\{C_t\})\|}$$

其中$Importance_score$为重要度函数。

2.2 医疗数据元筛选模型的训练与预测

将根据标准数据元分类体系，人工分类和关联映射构建的标准分类医疗数据元集合转换为标准分类医疗数据元图数据，设标准分类医疗数据元图数据中存储的列顶点集合为$S=\{s_k\}$，设构建标准分类医疗数据元集合过程中被人工筛选排除的列对应的列顶点集合为$S'=\{s'_k\}$。

训练时从集合S中随机抽取q个列顶点作为正样本集合$\{s_t\}_{t=1,2,...,q}$，从集合S'中随机抽取q个列顶点作为负样本集合$\{s'_t\}_{t=1,2,...,q}$；设样本(s_i,y_i)的重要度分数为$Im(s_i)$，s_i表示第i个列顶点，$y_i\epsilon\{0,1\}$表示样本真实类别，则基于重要度分数计算医疗数据元筛选模型的损失函数$Loss$：

$$Loss=-\frac{1}{2*q}\sum_{i=1}^{2*q}(y_i\log(Im(s_i))+(1-y_i)\log(1-Im(s_i)))$$

通过 Adam 算法更新重要度函数，更新医疗数据元筛选模型。

医疗数据元筛选模型在预测时，通过计算阈值L'判断列顶点C_k对应的待筛选医疗数据元集合中的列是否为有效数据元，阈值L'计算公式：

$$L'=\frac{1}{1+e^{-Im(C_k)}}$$

若$L'\geq 0.5$，则说明列顶点C_k为有效列顶点，对应的列为有效数据元。

最终由筛选后的有效列顶点集合关联组成待分类医疗数据元图数据，对应的筛选后的列集合组成待分类医疗数据元集合。

三、基于深度图匹配模型确定医疗数据元的类别

3.1 从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合

待分类医疗数据元图数据中存储的列顶点与待分类医疗数据元集合中的列存在一一对应关系。设由标准数据模型定义的标准数据元分类体系中所有标准分类集合为E，标准分类医疗数据元图数据中的列顶点集合为D，$D_i\in D$在标准数据元分类体系中的分类为$E_i\in E$；设待分类医疗数据元图数据中存储的列顶点集合为C。则医疗数据元分类过程可以抽象为在D中找到与列顶点$C_k\in C$匹配度

最高的列顶点D_i，从而确定列顶点C_k对应的列的分类为E_i，而医疗大数据中心开发过程中的数据分类与关联映射过程，可以抽象为为标准数据元分类体系的所有分类E_i找到匹配度最高的C_k。

以标准数据模型为数据标准的标准数据库中有些列的数据的格式或内容会比较统一，与之存在关联映射关系的标准分类医疗数据元集合的列的格式或内容也会比较统一。如果首先为这些列对应的顶点定位到在待分类医疗数据元图数据中对应的顶点（称为种子顶点），可以缩小深度图匹配模型的搜索空间，从而提高其效率。对于列顶点$D_i \in D$，从D_i对应的列中随机抽取r_0个数据$(d_{i,j})_{j=1,2,\ldots,r_0}$，对于待分类医疗数据元图数据中的列顶点$C_k \in C$，同样从$C_k$对应的列中随机抽取$r_0$个数据$(c_{k,l})_{l=1,2,\ldots,r_0}$，则$D_i$和$C_k$的匹配度$match_1(D_i, C_k)$为：

$$match\_1(D_i, C_k) = \frac{1}{{r_0}^2} \sum_{j=1}^{r_0} \sum_{l=1}^{r_0} \frac{v(d_{i,j}) \cdot v(c_{k,l})}{\left\| v(d_{i,j}) \right\| \times \left\| v(c_{k,l}) \right\|}$$

其中$v(x)$代表数据x的向量表示，则D_i对应的种子顶点为与其匹配度最高的列顶点$\hat{C}_i$，即：

$$\hat{C}_i = arg \max_{C_k \in C} \left(match\_1(D_i, C_k) \right)$$

3.2 基于种子顶点集合进行待分类医疗数据元图数据的子图切割

以$sch(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在父子关系的列顶点集合，以$ext(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在外键关系的列顶点集合，则基于种子顶点$\hat{C}_i$切割得到的子图$slice(\hat{C}_i)$为：

$$slice(\hat{C}_i) = sch(\hat{C}_i) \cup ext(\hat{C}_i)$$

以$N(D_i)$表示标准分类医疗数据元图数据中与D_i关联同一父顶点的列顶点集合，则深度图匹配模型的目标是从子图$slice(\hat{C}_i)$中搜索合适的子图，使得搜索到的子图中的列顶点与$N(D_i)$中的列顶点一一匹配，从而实现$slice(\hat{C}_i)$中列顶点对应的医疗数据元的分类。

3.3 利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类

医疗数据元分类过程包括以下步骤：

（1）结合图注意力机制，分别计算标准分类医疗数据元图数据中列顶点D_i的向量表示$V(D_i)$和待分类医疗数据元图数据的列顶点$\hat{C}_i$的向量表示$V(\hat{C}_i)$；具体为：

根据图注意力机制，计算D_i的向量表示$V(D_i)$为：

$$V(D_i) = \sum_{D' \in N(D_i)} w(D', D_i) \cdot v'(D')$$

其中$v'(D') = \frac{1}{r_1}\sum_{j=1}^{r_1} v(d_j')$，$(d_j')_{j=1,2,\ldots,r_1}$为从列顶点$D'$对应的列中随机抽取$r_1$个数据；$w(D', D_i)$表示$N(D_i)$中的某一列顶点$D'$对于列顶点$D_i$的权重函数，具体计算方式为：

$$w(D', D_i) = \frac{\exp\left(tanh\left(v'(D') \cdot W_1 \cdot v'(D_i)\right)\right)}{\sum_{D'' \in N(D_i)} \exp\left(tanh\left(v'(D'') \cdot W_1 \cdot v'(D_i)\right)\right)}$$

其中$tanh(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)}$为非线性激活函数，$W_1$为训练得到的矩阵参数。

根据图注意力机制，计算$\hat{C}_i$的向量表示$V(\hat{C}_i)$为：

$$V(\hat{C}_i) = \sum_{C' \in slice(\hat{C}_i)} w(C', \hat{C}_i) \cdot v'(C')$$

其中$v'(C') = \frac{1}{r_1}\sum_{j=1}^{r_1} v(c_j')$，$(c_j')_{j=1,2,\ldots,r_1}$为从列顶点$C'$对应的列中随机抽取$r_1$个数据；$w(C', \hat{C}_i)$表示$slice(\hat{C}_i)$中的某一列顶点$C'$对于列顶点$\hat{C}_i$的权重函数，具体计算方式为：

$$w(C', \hat{C}_i) = \frac{\exp\left(tanh\left(v'(C') \cdot W_2 \cdot v'(\hat{C}_i)\right)\right)}{\sum_{C'' \in slice(\hat{C}_i)} \exp\left(tanh\left(v'(C'') \cdot W_2 \cdot v'(\hat{C}_i)\right)\right)}$$

其中$tanh(x) = \frac{1-\exp(-2x)}{1+\exp(-2x)}$为非线性激活函数，$W_2$为训练得到的矩阵参数。

（2）计算所有$D' \in N(D_i)$与$C' \in slice(\hat{C}_i)$的匹配度，基于匹配度计算得到列顶点C'的分类，对应得到待分类医疗数据元集合中C'对应列的分类结果。

标准分类医疗数据元图数据的列顶点D'和待分类医疗数据元图数据的列顶点C'的匹配度$match_2(D', C')$为：

$$match\_2(D', C') = \frac{V(D') \cdot V(C')}{\|V(D')\| \times \|V(C')\|}$$

取与C'匹配度最高的列顶点$\hat{D}'$，即：

$$\hat{D}' = arg \max_{D' \in N(D_i)} match\_2(D', C')$$

则说明待分类医疗数据元图数据中的列顶点C'对应的列的分类为$\hat{D}'$对应的标准数据元分类体系中的类别。

本发明实施例还提供一种基于深度图匹配的医疗数据元自动化分类系统，该系统包括：

多源异构数据元的规范化采集与映射模块：定义基于最小元数据信息的医疗数据元图数据模型；将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向所述医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；该模块的实现可以参考上述步骤一。

有效医疗数据元筛选模块：计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，对应的列为有效医疗数据元，由有效列顶点集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；该模块的实现可以参考上述步骤二。

基于深度图匹配模型的医疗数据元分类模块：从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数

据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类；该模块的实现可以参考上述步骤三。

本发明提出的基于深度图匹配的医疗数据元自动化分类方法及系统的关键点如下：

1）基于医疗机构内数据湖的最小元数据信息，定义了一种基于最小元数据信息的医疗数据元图数据模型，使得深度图匹配模型的效果同样适用于极低元数据信息的局部数据沼泽的情况，达到使用最少的元数据信息完成数据元自动化分类的目的，同时保证在图数据模型标准下采集的图结构数据适用于深度图匹配模型的训练。

2）基于表示学习方法计算医疗数据元的向量表示，通过向量表示的分类，快速、自动化筛选有可能映射到标准数据模型的有效数据元。

3）基于图注意力机制计算列顶点的向量表示，构建深度图匹配模型完成医疗数据元的自动化分类。

以上所述仅是本发明的优选实施方式，虽然本发明已以较佳实施例披露如上，然而并非用以限定本发明。任何熟悉本领域的技术人员，在不脱离本发明技术方案范围情况下，都可利用上述揭示的方法和技术内容对本发明技术方案做出许多可能的变动和修饰，或修改为等同变化的等效实施例。因此，凡是未脱离本发明技术方案的内容，依据本发明的技术实质对以上实施例所做的任何的简单修改、等同变化及修饰，均仍属于本发明技术方案保护的范围内。

权 利 要 求 书

1、一种基于深度图匹配的医疗数据元自动化分类方法，其特征在于，包括：

(1)定义基于最小元数据信息的医疗数据元图数据模型；将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向所述医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；所述医疗数据元图数据模型采用有向属性图建模，图由顶点和边两种图元素构成；

所述顶点是由标签和对应标签的属性组构成的，标签代表顶点的类型，属性组代表标签拥有的一种或多种属性；所述顶点的本体信息包含顶点类型及每类顶点对应的属性信息，所述顶点类型包括数据库顶点、表顶点和列顶点，所述数据库顶点对应的属性信息包括数据库顶点索引和数据库类型信息，所述表顶点对应的属性信息包括表顶点索引，所述列顶点对应的属性信息包括列顶点索引、列数据类型信息和列向量表示；

所述边是由边类型和边属性构成的，每一条边均为有向边；所述边的本体信息包含边类型及每类边对应的属性信息，所述边类型包括起点为数据库顶点、终点为表顶点的父子关联，起点为表顶点、终点为列顶点的父子关联，以及起点和终点均为列顶点的外键，三种边类型对应的属性信息均为边索引；

(2)计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，由有效列顶点集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；

(3)从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类。

2、根据权利要求 1 所述的方法，其特征在于，所述多源异构的数据元向医疗数据元图数据模型的映射，包括：

将来自多源异构的医疗数据从数据湖中采集，组成待筛选医疗数据元集合；

使用元数据采集工具对数据湖中存储的元数据进行抓取；

使用列向量生成器，对待筛选医疗数据元集合中各表各列中存储的数据进行遍历，利用列向量表示模型预测得到各表各列的列向量表示；

通过图数据关联映射，将采集的元数据和产生的列向量表示向医疗数据元图数据模型关联映射，得到待筛选医疗数据元图数据。

3、根据权利要求 2 所述的方法，其特征在于，所述列向量生成器以数据表中的单列作为一个数据元单位，使用列向量表示模型转化各列存储的数据，计算各列的向量表示；

所述列向量表示模型的训练包括：列向量表示模型的训练数据为存储在标准数据库中的人工完成医疗数据元分类、数据结构符合标准数据模型的列数据，记为标准分类列；标准分类医疗数据元图数据中的列顶点与对应标准分类列存在一一对应关系；

设标准分类医疗数据元图数据中列顶点集合为$C=\{c_{k,j}\}$，其中$c_{k,j}$表示列顶点集合对应的标准分类列中第k列，第j行的数据，$c_{k,j}=\{w_t\}_{t=1,2,..,m}$，$m$为第$j$行字符总数，$w_t$为构成数据$c_{k,j}$的字符；通过文本表示模型$h$计算得到字符$w_t$的初始向量表示$h(w_t)$；在标准分类医疗数据元图数据的列顶点$C_k$下随机抽取$n$行数据$\{c_{k,j}\}_{j=1,2,..,n}$，第$j$行数据的向量表示为$v(c_{k,j})=\frac{1}{m}\sum_{t=1}^{m}h(w_t)$，根据自注意力机制计算得到标准分类医疗数据元图数据中列顶点C_k下各行数据的相关性，得到列顶点C_k的列向量表示$H(C_k)$，计算公式为：

$$v(C_k)=\frac{1}{n}\sum_{j=1}^{n}v(c_{k,j})$$

$$H(C_k)=softmax\left(\frac{v(C_k)v(C_k)^T}{\sqrt{d_k}}\right)v(C_k)$$

其中$v(C_k)$为列顶点C_k的向量表示，d_k为$v(C_k)$的维度，*softmax* 为 softmax 函数；

所述列向量表示模型的预测包括：列向量表示模型的预测数据为数据湖中各数据库中各表各列所组成的待筛选医疗数据元集合，以列为遍历单元对待筛选医疗数据元集合进行遍历；使用列向量表示模型计算对列顶点每次随机抽样的列向量表示；对预测的多次随机抽样的列向量表示结果求平均值，作为所述列顶点最终的列向量表示。

4、根据权利要求 3 所述的方法，其特征在于，所述计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度，包括：

对于待筛选医疗数据元图数据中存储的列顶点C_k，在除去C_k的列顶点集合中随机抽取p个列顶点$\{C_t\}_{t=1,2,\ldots,p}$，通过计算列顶点C_k与抽取的列顶点的相关性，计算C_k在医疗数据元图数据模型中的重要度分数$Im(C_k)$，$Im(C_k)$定义为：

$$H(\{C_t\}) = \frac{1}{p}\sum_{t=1}^{p} H(C_t)$$

$$Im(C_k) = Importance\_score(C_k, \{C_t\}) = \frac{H(C_k) \cdot H(\{C_t\})}{\|H(C_k)\| \times \|H(\{C_t\})\|}$$

其中$Importance_score$为重要度函数。

5、根据权利要求 1 所述的方法，其特征在于，所述医疗数据元筛选模型的训练与预测具体为：

将根据标准数据元分类体系，人工分类和关联映射构建的标准分类医疗数据元集合转换为标准分类医疗数据元图数据，设标准分类医疗数据元图数据中存储的列顶点集合为$S = \{s_k\}$，设构建标准分类医疗数据元集合过程中被人工筛选排除的列对应的列顶点集合为$S' = \{s'_k\}$；

训练时从集合S中随机抽取q个列顶点作为正样本集合$\{s_t\}_{t=1,2,\ldots,q}$，从集合S'中随机抽取q个列顶点作为负样本集合$\{s'_t\}_{t=1,2,\ldots,q}$；设样本$(s_i, y_i)$的重要度分

数为$Im(s_i)$，s_i表示第i个列顶点，$y_i \epsilon \{0,1\}$表示样本真实类别，则基于重要度分数计算医疗数据元筛选模型的损失函数$Loss$：

$$Loss = -\frac{1}{2*q}\sum_{i=1}^{2*q}(y_i \log(Im(s_i)) + (1-y_i)\log(1-Im(s_i)))$$

所述医疗数据元筛选模型在预测时，通过计算阈值L'判断列顶点C_k对应的待筛选医疗数据元集合中的列是否为有效数据元，阈值L'计算公式：

$$L' = \frac{1}{1+e^{-Im(C_k)}}$$

若$L' \geq 0.5$，则说明列顶点C_k为有效列顶点，对应的列为有效数据元；

由筛选后的有效列顶点集合关联组成待分类医疗数据元图数据，对应的筛选后的列集合组成待分类医疗数据元集合。

6、根据权利要求 1 所述的方法，其特征在于，所述从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合，包括：

设由标准数据模型定义的标准数据元分类体系中所有标准分类集合为E，标准分类医疗数据元图数据中的列顶点集合为D，$D_i \in D$在标准数据元分类体系中的分类为$E_i \in E$；设待分类医疗数据元图数据中存储的列顶点集合为C；医疗数据元分类过程抽象为在D中找到与列顶点$C_k \in C$匹配度最高的列顶点D_i，从而确定列顶点C_k对应的列的分类为E_i；

对于列顶点$D_i \in D$，从D_i对应的列中随机抽取r_0个数据$(d_{i,j})_{j=1,2,\ldots,r_0}$，对于列顶点$C_k \in C$，从$C_k$对应的列中随机抽取$r_0$个数据$(c_{k,l})_{l=1,2,\ldots,r_0}$，则$D_i$和$C_k$的匹配度$match_1(D_i, C_k)$为：

$$match\_1(D_i, C_k) = \frac{1}{{r_0}^2}\sum_{j=1}^{r_0}\sum_{l=1}^{r_0}\frac{v(d_{i,j})\cdot v(c_{k,l})}{\|v(d_{i,j})\| \times \|v(c_{k,l})\|}$$

其中$v(x)$代表数据x的向量表示，则D_i对应的种子顶点为与其匹配度最高的列顶点$\hat{C}_i$，即：

$\hat{C}_i = arg\max_{C_k \in C}(match\_1(D_i, C_k))$。

7、根据权利要求 6 所述的方法，其特征在于，所述基于种子顶点集合进行待分类医疗数据元图数据的子图切割，包括：

以$sch(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在父子关系的列顶点集合，以$ext(\hat{C}_i)$表示待分类医疗数据元图数据中与$\hat{C}_i$存在外键关系的列顶点集合，则基于种子顶点$\hat{C}_i$切割得到的子图$slice(\hat{C}_i)$为：

$$slice(\hat{C}_i) = sch(\hat{C}_i) \cup ext(\hat{C}_i)$$

以$N(D_i)$表示标准分类医疗数据元图数据中与D_i关联同一父顶点的列顶点集合，则深度图匹配模型的目标是从子图$slice(\hat{C}_i)$中搜索子图，使得搜索到的子图中的列顶点与$N(D_i)$中的列顶点一一匹配，实现$slice(\hat{C}_i)$中列顶点对应的医疗数据元的分类。

8、根据权利要求 7 所述的方法，其特征在于，所述利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，包括：

根据图注意力机制，计算标准分类医疗数据元图数据中列顶点D_i的向量表示$V(D_i)$为：

$$V(D_i) = \sum_{D' \in N(D_i)} w(D', D_i) \cdot v'(D')$$

其中$v'(D') = \frac{1}{r_1}\sum_{j=1}^{r_1} v(d'_j)$，$(d'_j)_{j=1,2,\ldots,r_1}$为从列顶点$D'$对应的列中随机抽取$r_1$个数据；$w(D', D_i)$表示$N(D_i)$中的某一列顶点$D'$对于列顶点$D_i$的权重函数；

根据图注意力机制，计算待分类医疗数据元图数据的列顶点$\hat{C}_i$的向量表示$V(\hat{C}_i)$为：

$$V(\hat{C}_i) = \sum_{C' \in slice(\hat{C}_i)} w(C', \hat{C}_i) \cdot v'(C')$$

其中$v'(C')=\frac{1}{r_1}\sum_{j=1}^{r_1}v(c_j')$，$(c_j')_{j=1,2,\ldots,r_1}$为从列顶点$C'$对应的列中随机抽取$r_1$个数据；$w(C',\hat{C}_i)$表示$slice(\hat{C}_i)$中的某一列顶点$C'$对于列顶点$\hat{C}_i$的权重函数；

列顶点$D'\in N(D_i)$和列顶点$C'\in slice(\hat{C}_i)$的匹配度$match_2(D',C')$为：

$$match\_2(D',C')=\frac{V(D')\cdot V(C')}{\|V(D')\|\times\|V(C')\|}$$

取与C'匹配度最高的列顶点$\hat{D}'$，即：

$$\hat{D}'=arg\max_{D'\in N(D_i)}match\_2(D',C')$$

待分类医疗数据元图数据中的列顶点C'对应的列的分类为$\hat{D}'$对应的标准数据元分类体系中的类别。

9、一种基于深度图匹配的医疗数据元自动化分类系统，其特征在于，包括：

多源异构数据元的规范化采集与映射模块：定义基于最小元数据信息的医疗数据元图数据模型；将医疗机构内数据湖中存储的多源异构的数据元组成待筛选医疗数据元集合，向所述医疗数据元图数据模型自动化映射，映射结果存储为待筛选医疗数据元图数据；所述医疗数据元图数据模型采用有向属性图建模，图由顶点和边两种图元素构成；

所述顶点是由标签和对应标签的属性组构成的，标签代表顶点的类型，属性组代表标签拥有的一种或多种属性；所述顶点的本体信息包含顶点类型及每类顶点对应的属性信息，所述顶点类型包括数据库顶点、表顶点和列顶点，所述数据库顶点对应的属性信息包括数据库顶点索引和数据库类型信息，所述表顶点对应的属性信息包括表顶点索引，所述列顶点对应的属性信息包括列顶点索引、列数据类型信息和列向量表示；

所述边是由边类型和边属性构成的，每一条边均为有向边；所述边的本体信息包含边类型及每类边对应的属性信息，所述边类型包括起点为数据库顶点、终点为表顶点的父子关联，起点为表顶点、终点为列顶点的父子关联，以及起点和终点均为列顶点的外键，三种边类型对应的属性信息均为边索引；

有效医疗数据元筛选模块：计算待筛选医疗数据元图数据中存储的各列顶点在医疗数据元图数据模型中的重要度；构建医疗数据元筛选模型，基于各列顶点的重要度计算各列顶点对应的列映射到标准数据模型的可能性，筛选出有效列顶点，对应的列为有效医疗数据元，由有效列顶点集合关联组成待分类医疗数据元图数据，有效列顶点对应的列集合组成待分类医疗数据元集合；

基于深度图匹配模型的医疗数据元分类模块：从待分类医疗数据元图数据中确定标准分类医疗数据元图数据的种子顶点集合；基于种子顶点集合进行待分类医疗数据元图数据的子图切割；利用深度图匹配模型完成对待分类医疗数据元图数据中列顶点的分类，从而得到列顶点对应的医疗数据元的分类。

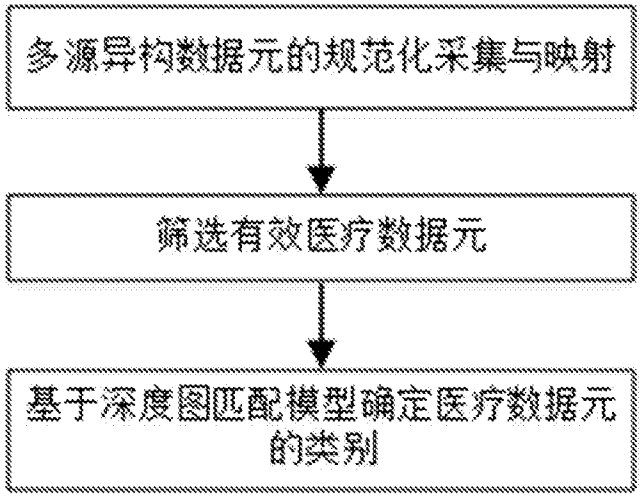


图 1

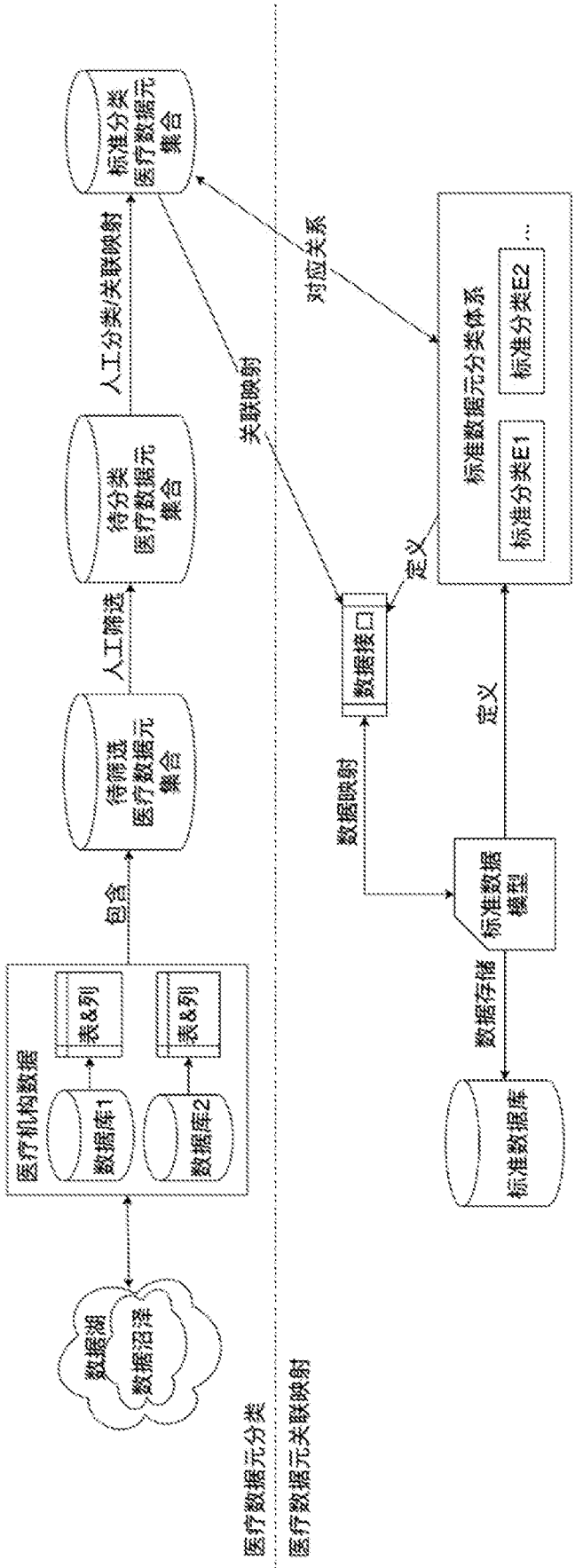

医疗机构数据
数据库1
表&列
数据库2
表&列
数据湖
数据沼泽
包含
待筛选
医疗数据元
集合
人工筛选
待分类
医疗数据元
集合
人工分类/关联映射
标准分类
医疗数据元
集合
关联映射
对应关系
数据接口
数据映射
定义
标准数据元分类体系
标准分类E1
标准分类E2
...
定义
标准数据
模型
数据存储
标准数据库
医疗数据元分类
医疗数据元关联映射

图 2

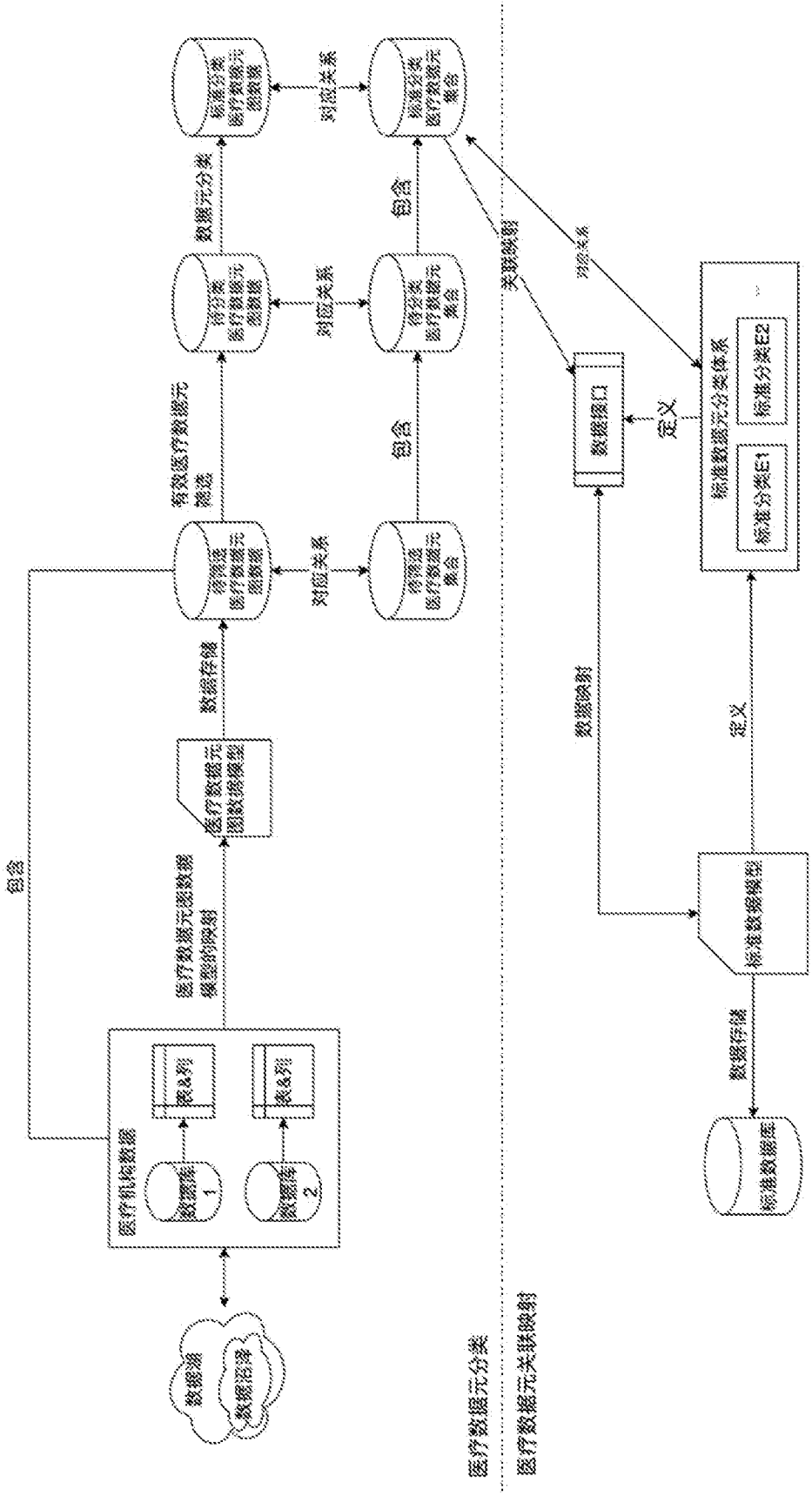

医疗数据元分类
医疗数据元关联映射
数据源
数据治理
医疗机构数据
数据库1
数据库2
表&列
表&列
医疗数据元源数据模型的映射
医疗数据元规范数据模型
数据存储
有效医疗数据元数据
有效医疗数据元集合
对应关系
包含
数据元分类
标准分类医疗数据元数据
标准分类医疗数据元集合
关联映射
关联关系
数据接口
定义
标准数据元分类体系
标准分类E1
标准分类E2
数据映射
标准数据模型
数据存储
标准数据库

图 3

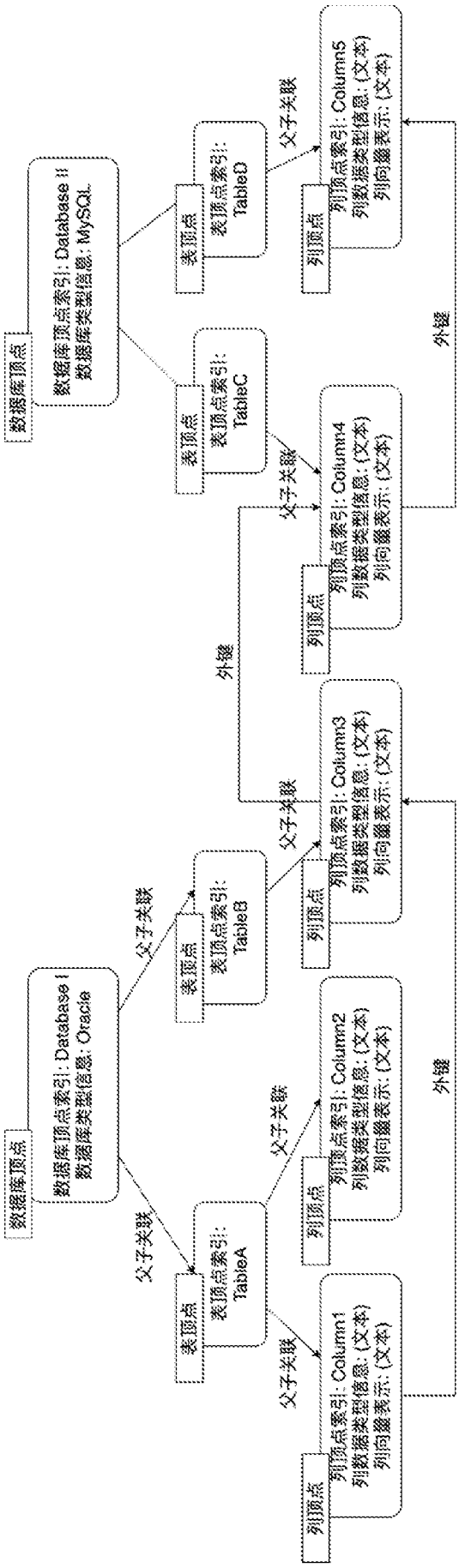

数据库顶点
数据库顶点索引: Database I
数据库类型信息: Oracle
数据库顶点
数据库顶点索引: Database II
数据库类型信息: MySQL
表顶点
表顶点索引: TableA
表顶点
表顶点索引: TableB
表顶点
表顶点索引: TableC
表顶点
表顶点索引: TableD
列顶点
列顶点索引: Column1
列数据类型信息: (文本)
列向量表示: (文本)
列顶点
列顶点索引: Column2
列数据类型信息: (文本)
列向量表示: (文本)
列顶点
列顶点索引: Column3
列数据类型信息: (文本)
列向量表示: (文本)
列顶点
列顶点索引: Column4
列数据类型信息: (文本)
列向量表示: (文本)
列顶点
列顶点索引: Column5
列数据类型信息: (文本)
列向量表示: (文本)
父子关联
父子关联
父子关联
父子关联
父子关联
父子关联
外键
外键
外键

图 4

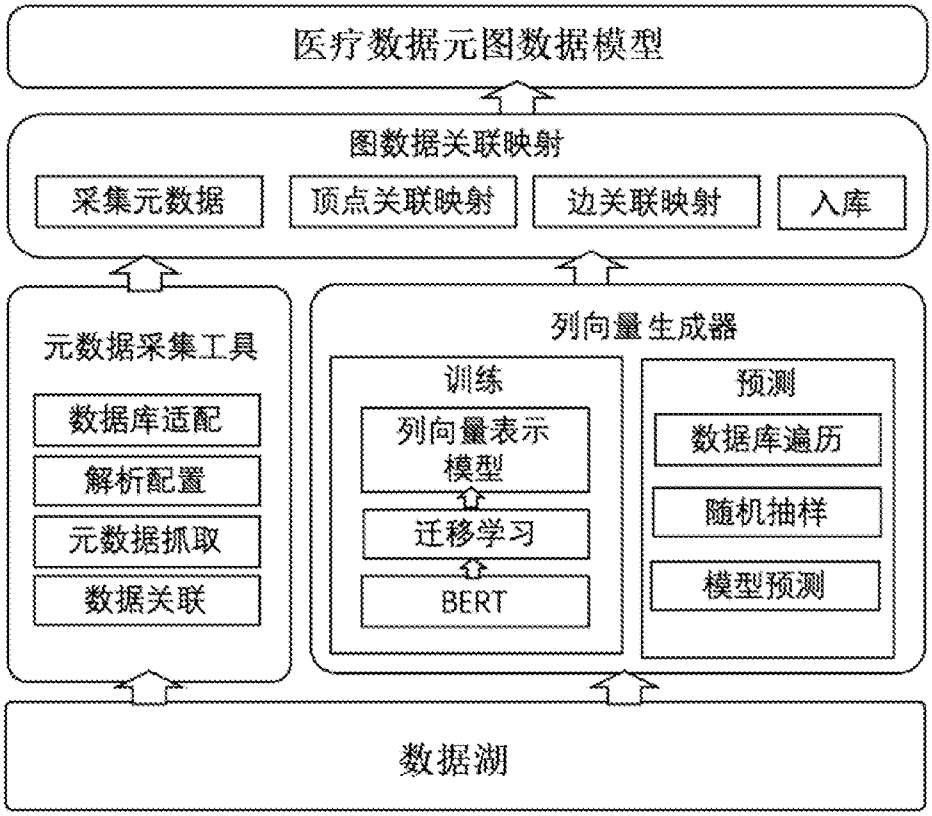

医疗数据元图数据模型
图数据关联映射
采集元数据
顶点关联映射
边关联映射
入库
元数据采集工具
数据库适配
解析配置
元数据抓取
数据关联
列向量生成器
训练
列向量表示模型
迁移学习
BERT
预测
数据库遍历
随机抽样
模型预测
数据湖

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/116971

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/906(2019.01)i; G06F 16/901(2019.01)i; G06F 16/93(2019.01)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)

CNTXT, ENTXT, ENTXTC, DWPI, CNKI: 深度图, 医疗, 数据, 分类, 元数据, 筛选, 数据元, 数据仓库, 数据湖, 匹配, 映射, depth image, depth map, medical, data, classify, metadata, select, choose, data element, data warehouse, database, data lake, match, mapping

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 114003791 A (ZHEJIANG LAB) 01 February 2022 (2022-02-01) claims 1-10, and description, paragraphs 2-20	1-9
A	CN 109471945 A (SUN YAT-SEN UNIVERSITY) 15 March 2019 (2019-03-15) entire document	1-9
A	US 2021089880 A1 (IBM) 25 March 2021 (2021-03-25) entire document	1-9
A	CN 110349639 A (ZHEJIANG LAB) 18 October 2019 (2019-10-18) entire document	1-9
A	CN 110021439 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 16 July 2019 (2019-07-16) entire document	1-9
A	CN 113656604 A (ZHEJIANG LAB) 16 November 2021 (2021-11-16) entire document	1-9
A	WO 2017152802 A1 (CHEN KUAN) 14 September 2017 (2017-09-14) entire document	1-9

☑ 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
19 October 2022	**28 October 2022**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2022/116971

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 109948680 A (HEFEI UNIVERSITY OF TECHNOLOGY) 28 June 2019 (2019-06-28) entire document	1-9

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2022/116971

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	114003791	A	01 February 2022	CN	114003791	B	08 April 2022
CN	109471945	A	15 March 2019	CN	109471945	B	23 November 2021
US	2021089880	A1	25 March 2021	WO	2021059081	A1	01 April 2021
				CN	114514530	A	17 May 2022
CN	110349639	A	18 October 2019	WO	2020233256	A1	26 November 2020
				JP	2022508350	A	19 January 2022
				JP	7093593	B2	30 June 2022
				CN	110349639	B	04 January 2022
CN	110021439	A	16 July 2019	WO	2020177230	A1	10 September 2020
				US	2021257066	A1	19 August 2021
				JP	2021532499	A	25 November 2021
				SG	11202008485	A1	29 October 2020
CN	113656604	A	16 November 2021	CN	113656604	B	22 February 2022
WO	2017152802	A1	14 September 2017	CN	105808712	A	27 July 2016
CN	109948680	A	28 June 2019	CN	109948680	B	11 June 2021

国际检索报告

国际申请号
PCT/CN2022/116971

A. 主题的分类

G06F 16/906(2019.01)i; G06F 16/901(2019.01)i; G06F 16/93(2019.01)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））

CNTXT, ENTXT, ENTXTC, DWPI, CNKI：深度图，医疗，数据，分类，元数据，筛选，数据元，数据仓库，数据湖，匹配，映射，depth image, depth map, medical, data, classify, metadata, select, choose, data element, data warehouse, database, data lake, match, mapping

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 114003791 A (之江实验室) 2022年2月1日 (2022-02-01) 权利要求1-10，说明书第2-20段	1-9
A	CN 109471945 A (中山大学) 2019年3月15日 (2019-03-15) 全文	1-9
A	US 2021089880 A1 (IBM) 2021年3月25日 (2021-03-25) 全文	1-9
A	CN 110349639 A (之江实验室) 2019年10月18日 (2019-10-18) 全文	1-9
A	CN 110021439 A (平安科技深圳有限公司) 2019年7月16日 (2019-07-16) 全文	1-9
A	CN 113656604 A (之江实验室) 2021年11月16日 (2021-11-16) 全文	1-9
A	WO 2017152802 A1 (CHEN KUAN) 2017年9月14日 (2017-09-14) 全文	1-9

☑ 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
2022年10月19日	2022年10月28日
ISA/CN的名称和邮寄地址 中国国家知识产权局(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 王雪莲 电话号码 62411747

国际检索报告

国际申请号
PCT/CN2022/116971

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 109948680 A (合肥工业大学) 2019年6月28日 (2019 - 06 - 28) 全文	1-9

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2022/116971

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	114003791	A	2022年2月1日	CN	114003791	B	2022年4月8日
CN	109471945	A	2019年3月15日	CN	109471945	B	2021年11月23日
US	2021089880	A1	2021年3月25日	WO	2021059081	A1	2021年4月1日
				CN	114514530	A	2022年5月17日
CN	110349639	A	2019年10月18日	WO	2020233256	A1	2020年11月26日
				JP	2022508350	A	2022年1月19日
				JP	7093593	B2	2022年6月30日
				CN	110349639	B	2022年1月4日
CN	110021439	A	2019年7月16日	WO	2020177230	A1	2020年9月10日
				US	2021257066	A1	2021年8月19日
				JP	2021532499	A	2021年11月25日
				SG	11202008485	A1	2020年10月29日
CN	113656604	A	2021年11月16日	CN	113656604	B	2022年2月22日
WO	2017152802	A1	2017年9月14日	CN	105808712	A	2016年7月27日
CN	109948680	A	2019年6月28日	CN	109948680	B	2021年6月11日