(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106228023 B
(45)授权公告日 2018.08.28

(21)申请号 201610621176.8

(22)申请日 2016.08.01

(65)同一申请的已公布的文献号
申请公布号 CN 106228023 A

(43)申请公布日 2016.12.14

(73)专利权人 清华大学
地址 100084 北京市海淀区清华园1号

(72)发明人 金涛 王建民 徐啸

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 廖元秋

(51)Int.Cl.
*G16H 50/70*(2018.01)

(56)对比文件
CN 1582443 A,2005.02.16,
CN 101571890 A,2009.11.04,
CN 105808712 A,2016.07.27,
李劲松 等.临床路径的本体建模与实例验证.《中国数字医学》.2011,第36卷(第5期),第27-31页.
汤琼 等.一种基于数据挖掘的临床路径系统方案研究.《电脑知识与技术》.2011,第7卷(第28期),第6795-6796,6799页.

审查员 何俊伟

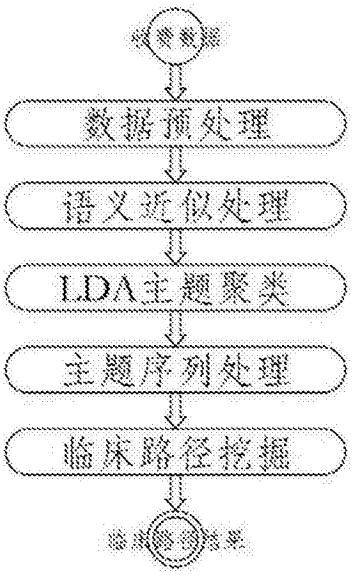
权利要求书2页 说明书6页 附图1页

(54)发明名称

# 一种基于本体和主题模型的临床路径挖掘方法

(57)摘要

本发明提出了一种基于本体和主题模型的临床路径挖掘方法。给定医院实际收费数据,通过五阶段处理,数据预处理、语义近似处理、主题聚类、主题序列处理、临床路径挖掘,最终得到符合医院实际的临床路径。本发明能够从医院实际的收费数据出发,通过一系列的数据处理,得到符合医院实际情况的疾病诊疗的临床路径,从而辅助制定符合医院实际的临床路径,该方法相比人工制定效率更高并且更客观全面。




CN 106228023 B

1.一种基于本体和主题模型的临床路径挖掘方法，其特征在于，根据给定医院实际收费数据，通过五阶段处理，数据预处理、语义近似处理、主题聚类、主题序列处理、临床路径挖掘，最终得到符合医院实际的临床路径；各阶段处理具体包括以下步骤：

(1)数据预处理：对获取的医院原始收费数据进行预处理并调整数据的格式；输入的原始医院收费数据，至少需要包括病人编号、收费项目名称、收费项目类别、使用总量、日期这几个字段；具体包括：

(1-1)对收费项目的数据进行处理，删除与疾病诊疗不相关的数据，删除与收费项目名称不相关的用语；

(1-2)对(1-1)删除后保留的数据进行规范化处理，具体过程为：

(1-2-1)同一病人同一天同样收费项目使用总量进行加和；

(1-2-2)同一病人同一天不同收费项目的使用总量进行归一化处理，将使用数量都归一化到区间[0,100]；

(1-3)对(1-2)规范化后的数据调整输出格式，具体过程为：

(1-3-1)将所有收费项目都分配一个唯一编号；

(1-3-2)将经过步骤(1-2)处理后的收费数据整理成如下输出格式“病人编号@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……”表示某一病人在某一天具体有哪些收费项目，同一收费项目编号重复次数为该收费项目在当日归一化后的数量；

(2)语义近似处理：根据阶段(1)得到的规定格式的数据中所有收费项目名称找出目的相似的医疗服务项目，基于收费项目的相似度输出指定格式内容，具体包括以下步骤：

(2-1)使用有道翻译API进行收费项目名称的中译英，删除翻译结果中的分类标签；

(2-2)基于Snomed CT计算不同收费项目之间的语义相似度，基于Intrinsic IC-based的方法度量不同项目之间的语义相似度；

(2-3)根据收费项目的相似度进行处理，将所有两两相似度值不小于0.8的收费项目对输出到一个must-links文件中，输出格式为“MERGE_收费项目编号收费项目编号”；其中收费项目编号为(1-3-1)得出的编号；

(3)主题聚类：基于阶段(1)(2)的输出文件建立主题模型，将各个病人各天的所有收费项目视为一个文档，将收费项目视为一个单词，具体如下：

将阶段(1)中整理后的收费记录文件形如“病人编码@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……”、阶段(2)中整理得到的must-links文件，形如“MERGE_收费项目编号收费项目编号”作为输入，调用Tree-based(基于树)的LDA算法；输出两个概率矩阵作为主题模型，一个是各病人诊疗天文档到各主题的概率矩阵，反映了病人每天的诊疗是服务于哪些主题的；另一个是各主题到各收费项目的概率矩阵，反映了确定了诊疗主题后，需要采取哪些诊疗项目；

(4)主题序列处理：根据阶段(3)建立的主题模型对病人诊疗天文档进行主题标注，并针对每个病人将每天的诊疗主题串接起来形成诊疗主题序列，进而进行相应处理，具体包括以下步骤：

(4-1)根据阶段(3)中得到的病人诊疗天文档到各主题的概率矩阵，当某主题的概率大于最大概率设定比值，即将该主题赋予相应的病人诊疗天文档；

(4-2)针对某病人诊疗天文档,将(4-1)中得到的主题按概率从大到小排序,并以“-”连接,形成相应病人诊疗天文档的复合主题;

(4-3)针对(4-2)中得到的病人诊疗天文档的复合主题,进行计数统计;

(4-4)如果(4-3)中得到的某复合主题计数低于某一设定阈值,则进行低概率主题剪枝,具体操作为,将复合主题中最后一个主题直接剪除,若新的复合主题计数仍然过低,则继续进行剪枝直到计数满足所述阈值要求为止;

(4-5)针对某病人,将诊疗天文档按日期排序,并将相应复合主题串接在一起,形成相应病人的诊疗主题序列;

(4-6)针对(4-5)中得到的诊疗主题序列,判断各诊疗主题序列是否是其它病人诊疗主题序列的子序列,若是则直接移除;

(5)临床路径挖掘:将阶段(4)输出的诊疗主题序列整理成相应日志文件,对日志文件进行数据挖掘,得到最终的临床路径;具体包括以下步骤:

(5-1)将阶段(4)中得到的病人诊疗主题序列按照ProM日志文件要求整理成相应的日志文件;

(5-2)直接使用ProM中的启发式过程挖掘算法针对输入的病人诊疗主题序列日志文件进行挖掘,得到相应疾病的临床路径。

2.如权利要求1所述方法,其特征在于,所述步骤(1-1)具体处理过程为:

(1-1-1)删除与疾病诊疗不相关的数据记录;

(1-1-2)删除对收费项目名称的详细解释,使得不会造成对项目名称的误解;

(1-1-3)删除收费项目名称中的与项目无关的字样;

(1-1-4)统一同一收费项目的不同描述;

(1-1-5)删除收费项目名称中“一次性”字样;

(1-1-6)删除收费项目名称中“注射液”和“胶囊”字样;

(1-1-7)删除收费项目名称中“床旁”字样。

3.如权利要求1所述方法,其特征在于,所述步骤(2-2)具体包括:

具体过程如下:

(2-2-1)计算各收费项目在Snomed CT中的所有子孙叶子节点;

(2-2-2)计算各收费项目在Snomed CT中的所有祖先节点;

(2-2-3)计算给定两个收费项目在Snomed CT中的公共祖先节点;

(2-2-4)计算各收费项目以及公共祖先节点在Snomed CT中的内部信息量,计算公式为$-\log\left(\frac{|leaves(a)|+|subsumers(a)|}{max_leaves}\right)$,其中|leaves(a)|表示收费项目a在Snomed CT中所有子孙为叶子节点的总数,|subsumers(a)|表示收费项目a在Snomed CT中所有祖先节点的总数,max_leaves表示Snomed CT中所有叶子节点总数;

(2-2-5)依据公式$sim(a,b)=\frac{2\times IIC(MICA(a,b))}{IIC(a)+IIC(b)}$计算各收费项目之间的语义相似度值,其中a、b表示需要计算相似度值的收费项目名称,IIC(MICA(a,b))表示a、b在Snomed CT中公共祖先内部信息量的最大值,IIC(a)、IIC(b)分别表示a、b在Snomed CT中的内部信息量。

# 一种基于本体和主题模型的临床路径挖掘方法

## 技术领域

[0001] 本发明属于计算机数据挖掘领域，特别涉及一种基于本体和主题模型的临床路径挖掘方法。

## 背景技术

[0002] 临床路径表明了疾病的诊疗工作是如何一步步开展的，反映了各科医生和护士是如何分工协作共同开展疾病诊疗工作的，同时也反映了各种医疗资源是如何一步步被消耗的。临床路径管理可以规范化疾病的诊疗过程，提高医疗质量，提高病人满意度，同时有效的控制医疗资源的消耗和医疗费用的增长。

[0003] 虽然临床路径管理受到世界各国的广泛关注，但实践效果并不理想。有文献对23个国家临床路径实践情况进行了调研，结论显示，进入临床路径管理的病人占比很小，并且大多限于急诊治疗。也有文献对我国临床路径实施的概况和成因进行了分析，结论指出，我国实施临床路径的医院数量少，区域分布不平衡，覆盖病种数量较少，病种较单一。有文献对限制我国目前临床路径实用的原因进行了归纳，结论认为制定个性化、更具体的符合医院实际情况的临床路径有利于推动临床路径管理在我国医院的实用化。临床路径虽然取决于疾病的诊疗指南规范，但由于医疗活动的高度实践性，医疗活动的具体开展必然高度依赖于具体的医院、医护人员和具体医疗资源的投入。所以如果考虑这些具体的医疗实践情况，制定出来的临床路径将具有更好的可执行性。

[0004] 临床路径的制定是一个非常复杂耗时的过程，需要各科专业人士通力合作，并且临床路径在具体医院的实施必须充分考虑实施医院的各种具体情况，如果完全依靠人为研讨制定临床路径必定存在以下问题：

[0005] (1)速度慢，需要各科专业人士充分沟通研讨，方能制定出实践性强的临床路径；目前国家规范仅给出300多种疾病的临床路径，而我国目前广泛使用的ICD-10疾病编码已有10000多种，如果再考虑并存症、并发症的组合情况，疾病数目非常大，单靠专家组研讨确定，费时费力；

[0006] (2)更新不及时，新药物、新技术、新方案、新证据不断出现，人为研讨不能及时更新；

[0007] (3)易出纰漏，由于疾病的诊疗复杂，人为研讨容易遗漏掉一些情况的考虑。

[0008] 由于医疗信息化的发展，医院已经积累了大量疾病诊疗相关数据，这些数据反映了疾病的实际诊疗过程，如果能从这些数据出发，通过数据技术处理，可以得到符合医院实际情况的疾病诊疗过程，对于医院临床路径的制定会有很好的辅助作用。

[0009] 本发明涉及的相关公开技术分别介绍如下：

[0010] 本体描述了特定领域中的概念(术语)以及概念之间的相互关系。比如对同一事物，可以有不同的概念(术语)，即同义词。在本发明中涉及的主要基于本体计算不同术语之间的语义相似度，即不同术语在多大程度上意思相近。

[0011] Snomed CT(系统化临床医学术语集)是由国际健康标准开发组织(IHTSDO)维护的

临床术语，被认为是世界上最全面、适用语言最多的临床术语集。Snomed CT包括有三大核心组件：概念、描述、关系，其中：

[0012] 概念，表示临床思想、活动、实体，有一个唯一的数字标识。

[0013] 描述，有三类：

[0014] (1)fully specified name，表示唯一无歧义的概念术语，并带有一个语义标签，比如“疾病”；

[0015] (2)preferred term，表示在多个描述中，针对相应的语种，优先选取的描述；

[0016] (3)synonyms，表示相同的临床概念。

[0017] 关系，用于表达不同概念之间的相关关系，有IS-A关系(表示从属关系，即某一概念是另一概念的子概念)和属性关系(表示某一概念是另一个概念的一个属性)。

[0018] 美国版Snomed CT在国际版的基础上增加了一些概念，2015年9月美国版Snomed CT在2015年7月国际版的基础上增加了991个新概念。美国版Snomed CT的开发旨在使其成为美国首要的电子健康记录、科研数据库、临床试验数据库中临床信息的编码术语。其发行有两种格式，本发明中采用了美国版Release Format 2(RF2)格式。

[0019] 在Snomed CT中，临床术语是按层次结构组织的，被分为19个不同的类。需要注意的是Snomed CT是多对一的层次结构，即一个概念可以有多个父节点概念，Snomed CT的概念关系结构构成了一个有向无环图。本发明中仅考虑概念之间的IS-A关系，基于Snomed CT概念的IS-A关系结构构成的有向无环图计算不同概念之间的语义相似度。

[0020] 已经有大量的研究致力于基于Snomed CT本体结构计算术语之间的相似度，有文献通过对已有生物医疗领域基于本体进行语义相似度量研究工作的比较，认为IC-based(基于信息量)的度量方法相比而言更可靠。由于缺乏必要的语料库，本发明使用Intrinsic IC-based(基于内在固有信息量)的度量方法。其基本思路是基于一个本体结构，因为各概念在该本体中的层次位置反映了该概念所含的信息量，故可以基于概念在本体中的层次位置计算概念之间的语义相似度。

[0021] 主题模型是一种统计模型，用于发现一系列文档中的抽象主题。如果一篇文档是围绕某个主题展开的，必然会有一些词语频繁出现。当然一篇文档也可能有多个主题，并且各主题所占比例有所不同，取决于相应词语出现的频次。故主题模型包括两个重要方面，一方面，给定一篇文档，可以以不同概率归类为某一主题；另一方面，给定一个主题，不同词语对该主题有不同概率的贡献度。常用的主题建模算法为LDA算法。

[0022] 本发明使用了过程挖掘算法，过程挖掘算法解决的问题是从给定的事件日志中挖掘出能产生这些事件记录的过程模型。目前，开源工具ProM提供了很多可以直接使用的过程挖掘算法。

[0023] 由于医疗领域的复杂性，直接使用已有过程挖掘算法进行临床路径挖掘，极易得到一团乱麻状的模型。根据已有文献的结论，启发式算法相比而言能够更好的处理实际数据，并且能很好的应对日志的不完备性和噪声。故本发明直接使用ProM工具中的启发式算法进行挖掘。

## 发明内容

[0024] 本发明的目的是为克服已有方法的不足之处，提出一种基于本体和主题模型的临

床路径挖掘方法。本方法能够从医院实际的收费数据出发，通过一系列的数据处理，得到符合医院实际情况的疾病诊疗的临床路径，从而辅助制定符合医院实际的临床路径，该方法相比人工制定效率更高并且更客观全面。

[0025] 本发明提出的一种基于本体和主题模型的临床路径挖掘方法，其特征在于，根据给定医院实际收费数据，通过五阶段处理，数据预处理、语义近似处理、主题聚类、主题序列处理、临床路径挖掘，最终得到符合医院实际的临床路径；各阶段处理具体包括以下步骤：

[0026] (1)数据预处理：对获取的医院原始收费数据进行预处理并调整数据的格式；输入的原始医院收费数据，至少需要包括病人编号、收费项目名称、收费项目类别、使用总量、日期这几个字段；具体包括：

[0027] (1-1)对收费项目的数据进行处理，删除与疾病诊疗不相关的数据，删除与收费项目名称不相关的用语；

[0028] (1-2)对(1-1)删除后保留的数据进行规范化处理，具体过程为：

[0029] (1-2-1)同一病人同一天同样收费项目使用总量进行加和；

[0030] (1-2-2)同一病人同一天不同收费项目的使用总量进行归一化处理，将使用数量都归一化到区间[0,100]；

[0031] (1-3)对(1-2)规范化后的数据调整输出格式，具体过程为：

[0032] (1-3-1)将所有收费项目都分配一个唯一编号；

[0033] (1-3-2)将经过步骤(1-2)处理后的收费数据整理成如下输出格式“病人编号@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……”表示某一病人在某一天具体有哪些收费项目，同一收费项目编号重复次数为该收费项目在当日归一化后的数量；

[0034] (2)语义近似处理：根据阶段(1)得到的规定格式的数据中所有收费项目名称找出目的相似的医疗服务项目，基于收费项目的相似度输出指定格式内容，具体包括以下步骤：

[0035] (2-1)使用有道翻译API进行收费项目名称的中译英，删除翻译结果中的分类标签；

[0036] (2-2)基于Snomed CT计算不同收费项目之间的语义相似度，基于Intrinsic IC-based的方法度量不同项目之间的语义相似度；

[0037] (2-3)根据收费项目的相似度进行处理，将所有两两相似度值不小于0.8的收费项目对输出到一个must-links文件中，输出格式为“MERGE_收费项目编号收费项目编号”；其中收费项目编号为(1-3-1)得出的编号；

[0038] (3)主题聚类：基于阶段(1)(2)的输出文件建立主题模型，将各个病人各天的所有收费项目视为一个文档(病人诊疗天文档)，将收费项目视为一个单词，具体如下：

[0039] 将阶段(1)中整理后的收费记录文件形如“病人编码@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……”、阶段(2)中整理得到的must-links文件，形如“MERGE_收费项目编号收费项目编号”作为输入，调用Tree-based(基于树)的LDA算法；输出两个概率矩阵作为主题模型，一个是各病人诊疗天文档到各主题的概率矩阵，反映了病人每天的诊疗是服务于哪些主题的；另一个是各主题到各收费项目的概率矩阵，反映了确定了诊疗主题后，需要采取哪些诊疗项目；

[0040] (4)主题序列处理：根据阶段(3)建立的主题模型对病人诊疗天文档进行主题标

注，并针对每个病人将每天的诊疗主题串接起来形成诊疗主题序列，进而进行相应处理，具体包括以下步骤：

[0041] (4-1)根据阶段(3)中得到的病人诊疗天文档到各主题的概率矩阵，当某主题的概率大于最大概率设定比值，即将该主题赋予相应的病人诊疗天文档；

[0042] (4-2)针对某病人诊疗天文档，将(4-1)中得到的主题按概率从大到小排序，并以“-”连接，形成相应病人诊疗天文档的复合主题；

[0043] (4-3)针对(4-2)中得到的病人诊疗天文档的复合主题，进行计数统计；

[0044] (4-4)如果(4-3)中得到的某复合主题计数低于某一设定阈值，则进行低概率主题剪枝，具体操作为，将复合主题中最后一个主题直接剪除，若新的复合主题计数仍然过低，则继续进行剪枝直到计数满足所述阈值要求为止；

[0045] (4-5)针对某病人，将诊疗天文档按日期排序，并将相应复合主题串接在一起，形成相应病人的诊疗主题序列；

[0046] (4-6)针对(4-5)中得到的诊疗主题序列，判断各诊疗主题序列是否是其它病人诊疗主题序列的子序列，若是则直接移除；

[0047] (5)临床路径挖掘：将阶段(4)输出的诊疗主题序列整理成相应日志文件，对日志文件进行数据挖掘，得到最终的临床路径；具体包括以下步骤：

[0048] (5-1)将阶段(4)中得到的病人诊疗主题序列按照ProM日志文件要求整理成相应的日志文件；

[0049] (5-2)直接使用ProM中的启发式过程挖掘算法针对输入的病人诊疗主题序列日志文件进行挖掘，得到相应疾病的临床路径。

[0050] 本发明提出的基于本体和主题模型的临床路径挖掘方法，其优点是：

[0051] (1)从医院实际数据出发，挖掘得到的临床路径更符合医院的实际情况，可作为医院临床路径制定者的参考，相对于人为研讨制定，该方法更为客观全面；

[0052] (2)通过医院历史数据挖掘得到的临床路径是医院实际执行的临床路径，通过和国家规范的对比，有利于临床路径管理者发现差异，从而采取相应的措施；

[0053] (3)采用计算机挖掘方法得到临床路径，针对没有国家临床路径规范指导的疾病诊疗很有意义；

[0054] (4)人类对于疾病的认知不断发展，新技术、新资源、新方案不断出现，通过针对数据的挖掘处理得到临床路径，能及时的更新临床路径，更好的实施循证医学。

## 附图说明

[0055] 图1是本发明基于本体和主题模型的临床路径挖掘方法的流程框图。

## 具体实施方式

[0056] 本发明提出的一种基于本体和主题模型的临床路径挖掘方法，根据给定医院实际收费数据，通过五阶段处理，数据预处理、语义近似处理、主题聚类、主题序列处理、临床路径挖掘，最终得到符合医院实际的临床路径；各阶段处理具体包括以下步骤：

[0057] (1)数据预处理：对获取的医院原始收费数据进行预处理并调整数据的格式；输入的原始医院收费数据，至少需要包括病人编号、收费项目名称、收费项目类别、使用总量、日

期这几个字段(表示具体哪个病人在哪一天使用了哪些医疗服务);具体包括:

[0058] (1-1)对收费项目的数据进行处理,删除与疾病诊疗不相关的数据,删除与收费项目名称不相关的用语;具体处理过程为:

[0059] (1-1-1)删除与疾病诊疗不相关的数据记录,比如将收费项目类别为床位费、采暖费、其它费、各种"自费"的收费记录删除;

[0060] (1-1-2)删除对收费项目名称的详细解释,比如"鼻饲管置管(注食、注药、十二指肠灌注按2元/次收取)",括号中的详细描述了使用场景以及收费依据,舍弃括号内的内容不会造成对项目名称的误解;

[0061] (1-1-3)删除收费项目名称中的"进口"和"国产"字样;

[0062] (1-1-4)统一同一收费项目的不同描述,比如"12通道动态心电图"和"十二通道心电图检查",统一为"12通道动态心电图";

[0063] (1-1-5)删除收费项目名称中"一次性"字样;

[0064] (1-1-6)删除收费项目名称中"注射液"和"胶囊"字样;

[0065] (1-1-7)删除收费项目名称中"床旁"字样;

[0066] (1-2)对(1-1)删除后保留的数据进行规范化处理,具体过程为:

[0067] (1-2-1)同一病人同一天同样收费项目使用总量进行加和;

[0068] (1-2-2)同一病人同一天不同收费项目的使用总量进行归一化处理,将使用数量都归一化到区间[0,100];

[0069] (1-3)对(1-2)规范化后的数据调整输出格式,具体过程为:

[0070] (1-3-1)将所有收费项目都分配一个唯一编号;

[0071] (1-3-2)将经过步骤(1-2)处理后的收费数据整理成如下输出格式"病人编号@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……"表示某一病人在某一天具体有哪些收费项目,同一收费项目编号重复次数为该收费项目在当日归一化后的数量;

[0072] (2)语义近似处理:根据阶段(1)得到的规定格式的数据中所有收费项目名称找出目的相似的医疗服务项目,基于收费项目的相似度输出指定格式内容,具体包括以下步骤:

[0073] (2-1)使用有道翻译API进行收费项目名称的中译英,删除翻译结果中的分类标签,比如"[有化]"、"[无化]"等;

[0074] (2-2)基于Snomed CT计算不同收费项目之间的语义相似度,基于Intrinsic IC-based的方法度量不同项目之间的语义相似度,具体过程如下:

[0075] (2-2-1)计算各收费项目在Snomed CT中的所有子孙叶子节点;

[0076] (2-2-2)计算各收费项目在Snomed CT中的所有祖先节点;

[0077] (2-2-3)计算给定两个收费项目在Snomed CT中的公共祖先节点;

[0078] (2-2-4)计算各收费项目以及公共祖先节点在Snomed CT中的内部信息量,计算公式为$-\log\left(\frac{|leaves(a)|}{max_leaves}\right)$,其中|leaves(a)|表示收费项目a在Snomed CT中所有子孙为叶子节点的总数,|subsumers(a)|表示收费项目a在Snomed CT中所有祖先节点的总数,max_leaves表示Snomed CT中所有叶子节点总数;

[0079] (2-2-5)依据公式$sim(a,b)=\frac{2\times ic(lca(a,b))}{ic(a)+ic(b)}$计算各收费项目之间的语义相似度值,其中a、b

表示需要计算相似度值的收费项目名称，IIC(MICA(a,b))表示a、b在Snomed CT中公共祖先内部信息量的最大值，IIC(a)、IIC(b)分别表示a、b在Snomed CT中的内部信息量；

[0080] (2-3)根据收费项目的相似度进行处理，将所有两两相似度值不小于0.8的收费项目对输出到一个must-links(表示必然在同样主题中出现)文件中，输出格式为“MERGE_收费项目编号收费项目编号”；其中收费项目编号为(1-3-1)得出的编号；

[0081] (3)主题聚类：基于阶段(1)(2)的输出文件建立主题模型，将各个病人各天的所有收费项目视为一个文档(病人诊疗天文档)，将收费项目视为一个单词，具体如下：

[0082] 将阶段(1)中整理后的收费记录文件形如“病人编码@日期,收费项目编号收费项目编号收费项目编号收费项目编号收费项目编号……”、阶段(2)中整理得到的must-links文件，形如“MERGE_收费项目编号收费项目编号”作为输入，调用Tree-based(基于树)的LDA算法；输出两个概率矩阵作为主题模型，一个是各病人诊疗天文档到各主题的概率矩阵，反映了病人每天的诊疗是服务于哪些主题的；另一个是各主题到各收费项目的概率矩阵，反映了确定了诊疗主题后，需要采取哪些诊疗项目；

[0083] (4)主题序列处理：根据阶段(3)建立的主题模型对病人诊疗天文档进行主题标注，并针对每个病人将每天的诊疗主题串接起来形成诊疗主题序列，进而进行相应处理，具体包括以下步骤：

[0084] (4-1)根据阶段(3)中得到的病人诊疗天文档到各主题的概率矩阵，当某主题的概率大于最大概率一定比值(比如0.5)，即将该主题赋予相应的病人诊疗天文档；

[0085] (4-2)针对某病人诊疗天文档，将(4-1)中得到的主题按概率从大到小排序，并以“-”连接，形成相应病人诊疗天文档的复合主题；

[0086] (4-3)针对(4-2)中得到的病人诊疗天文档的复合主题，进行计数统计；

[0087] (4-4)如果(4-3)中得到的某复合主题计数低于某一阈值(比如所有病人诊疗天文档总数的10％)，则进行低概率主题剪枝，具体操作为，将复合主题中最后一个主题(概率最低)直接剪除，若新的复合主题计数仍然过低，则继续进行剪枝直到计数满足所述阈值要求为止；

[0088] (4-5)针对某病人，将诊疗天文档按日期排序，并将相应复合主题串接在一起，形成相应病人的诊疗主题序列；

[0089] (4-6)针对(4-5)中得到的诊疗主题序列，判断各诊疗主题序列是否是其它病人诊疗主题序列的子序列(子序列中出现的所有主题都能在父序列中找到，并且出现先后顺序一致)，若是则直接移除；

[0090] (5)临床路径挖掘：将阶段(4)输出的诊疗主题序列整理成相应日志文件，对日志文件进行数据挖掘，得到最终的临床路径。具体包括以下步骤：

[0091] (5-1)将阶段(4)中得到的病人诊疗主题序列按照ProM日志文件要求整理成相应的日志文件；

[0092] (5-2)直接使用ProM中的启发式过程挖掘算法针对输入的病人诊疗主题序列日志文件进行挖掘，得到相应疾病的临床路径。

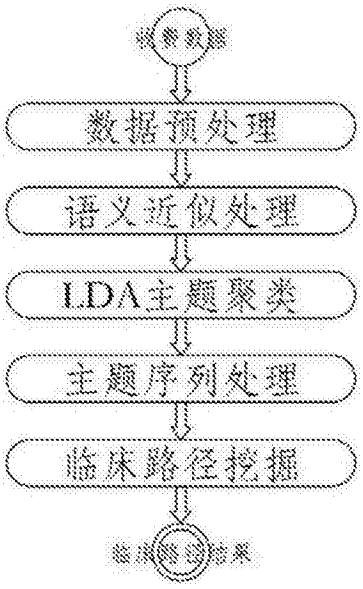



图1