(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 109065174 B
(45)授权公告日 2022.02.18

(21)申请号 201810843072.0
(22)申请日 2018.07.27
(65)同一申请的已公布的文献号
申请公布号 CN 109065174 A
(43)申请公布日 2018.12.21
(73)专利权人 合肥工业大学
地址 230009 安徽省合肥市包河区屯溪路193号
(72)发明人 丁帅 蔡琼 潘金鑫 金行
(74)专利代理机构 北京久诚知识产权代理事务所(特殊普通合伙) 11542
代理人 余罡
(51)Int.Cl.
G16H 50/70(2018.01)

(56)对比文件
CN 106156272 A,2016.11.23
CN 102317786 A,2012.01.11
CN 107613520 A,2018.01.19
CN 103365978 A,2013.10.23
US 2007218519 A1,2007.09.20

审查员 王璐

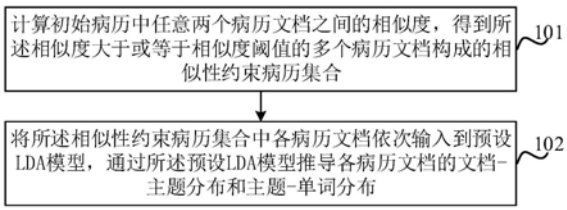
权利要求书2页 说明书11页 附图5页

(54)发明名称

考虑相似约束的病历主题获取方法及装置

(57)摘要

本发明提供了一种考虑相似约束的病历主题获取方法及装置。所述方法包括:计算初始病历中任意两个病历文档之间的相似度,得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合;将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型,通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。可见,本实施例中可以很好地模拟医生诊疗过程确定病历文本的思维过程,从而有利于获取主题的准确度。



CN 109065174 B

1.一种考虑相似约束的病历主题获取方法,其特征在于,所述方法包括:

计算初始病历中任意两个病历文档之间的相似度,得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合;

将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型,通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布;

通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布包括:

对所述相似性约束病历集合中各病历文档中每个单词随机赋予主题编号z;

重新扫描所述相似性约束病历集合,对每个单词按照$p(z_i = k|\vec{z}_{\neg i},\vec{w})$重新采样主题,使得到的新主题满足Gibbs Sampling收敛;其中,$p(z_i = k|\vec{z}_{\neg i},\vec{w})$表示单词$\vec{w}$分布为k的概率;

统计语料库中主题-单词共现频率矩阵,得到文档-主题分布和主题-单词分布;

所述预设LDA模型包括:

任意两个病历文档相似性约束采用主题分布距离$dis(\theta r^m, \theta r^n)$表示,公式为:

$$dis(\theta r^m,\ \theta r^n) = \frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln} d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m * L_n};$$

其中$\theta r^m = \{\theta_{m,1},\theta_{m,2},\cdots,\theta_{m,Lm}\}$,表示每个病历文档包括$L_m$个病程记录;$\theta_{m,Lm}$表示第$L_m$个病程记录的主题;$d(\theta_{m,Lm},\theta_{n,Ln})$表示为两个病程的主题向量之间的欧式距离;$\theta r^n = \{\theta_{n,1},\theta_{n,2},\cdots,\theta_{n,Ln}\}$,表示每个病历文档包括$L_n$个病程记录;$\theta_{n,Ln}$表示第$L_n$个病程记录的主题;

所述预设LDA模型还包括Gibbs-EM迭代函数,为:

$$p(z_i = k|\vec{z}_{\neg i},\vec{w},\alpha,\beta) \propto \frac{n_i^{(k)}+\beta_i-1}{\sum_{t=1}^{V}(n_t^{(k)}+\beta_t)-1} * \alpha_{mk};$$

$n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量;先验狄利克雷分布参数(α,β);病历集合单词总数V;预设LDA模型$\alpha_{mk} = \frac{e^{\mu_{m,k}}}{\sum_{k1=1}^{K} e^{\mu_{m,k1}}}$,$\mu_{mk}$代表病历文档m属于主题k的概率。

2.根据权利要求1所述的病历主题获取方法,其特征在于,计算初始病历中任意两个病历文档之间的相似度包括:

获取病历的多个相似性计算因素及各相似性计算因素的权重值;

分别计算任意两个病历文档关于各相似性计算因素的数值;

根据各相似性计算因素的数值和各相似性计算因素的权重值计算所述任意两个病历文档的相似度。

3.根据权利要求2所述的病历主题获取方法,其特征在于,所述相似性计算因素包括:性别属性的距离、年龄所属分段的距离、诊断结果的距离。

4.一种考虑相似约束的病历主题获取装置,其特征在于,所述装置包括:

病历集合获取模块,用于计算初始病历中任意两个病历文档之间的相似度,得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合;

主题分布推导模块,用于将所述相似性约束病历集合中各病历文档依次输入到预设

LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布；

所述主题分布推导模块包括：

主题编号单元，用于对所述相似性约束病历集合中各病历文档中每个单词随机赋予主题编号z；

主题迭代单元，用于重新扫描所述相似性约束病历集合，对每个单词按照$p(z_i = k|\vec{z}_{\neg i}, \vec{w})$重新采样主题，使得到的新主题满足GibbsSampling收敛；其中，$p(z_i = k|\vec{z}_{\neg i}, \vec{w})$表示单词$\vec{w}$分布为k的概率；

主题分布计算单元，用于统计语料库中主题-单词共现频率矩阵，得到文档-主题分布和主题-单词分布；

所述预设LDA模型包括：

任意两个病历文档相似性约束采用主题分布距离$dis(\theta r^m, \theta r^n)$表示，公式为：

$$dis(\theta r^m,\ \theta r^n) = \frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln} d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m * L_n};$$

其中$\theta r^m = \{\theta_{m,1}, \theta_{m,2}, \cdots, \theta_{m,Lm}\}$，表示每个病历文档包括$L_m$个病程记录；$\theta_{m,Lm}$表示第$L_m$个病程记录的主题；$d(\theta_{m,Lm}, \theta_{n,Ln})$表示为两个病程的主题向量之间的欧式距离；$\theta r^n = \{\theta_{n,1}, \theta_{n,2}, \cdots, \theta_{n,Ln}\}$，表示每个病历文档包括$L_n$个病程记录；$\theta_{n,Ln}$表示第$L_n$个病程记录的主题；

所述预设LDA模型还包括Gibbs-EM迭代函数，为：

$$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \beta) \propto \frac{n_i^{(k)} + \beta_i - 1}{\sum_{t=1}^{V}(n_t^{(k)} + \beta_t) - 1} * \alpha_{mk};$$

$n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量；先验狄利克雷分布参数(α，β)；病历集合单词总数V；预设LDA模型$\alpha_{mk} = \frac{e^{\mu_{m,k}}}{\sum_{k1=1}^{K} e^{\mu_{m,k1}}}$，$\mu_{mk}$代表病历文档m属于主题k的概率。

5.根据权利要求4所述的病历主题获取装置，其特征在于，所述病历集合获取模块包括：

权重值获取单元，用于获取病历的多个相似性计算因素及各相似性计算因素的权重值；

因素数据计算单元，用于分别计算任意两个病历文档关于各相似性计算因素的数值；

相似度计算单元，用于根据各相似性计算因素的数值和各相似性计算因素的权重值计算所述任意两个病历文档的相似度。

6.根据权利要求5所述的病历主题获取方法装置，其特征在于，所述相似性计算因素包括：性别属性的距离、年龄所属分段的距离、诊断结果的距离。

考虑相似约束的病历主题获取方法及装置

技术领域

[0001] 本发明涉及数据挖掘技术领域，尤其涉及一种考虑相似约束的病历主题获取方法及装置。

背景技术

[0002] 目前，主题模型大多应用于在线社交媒体领域的网络舆情主题演化分析方面，有利于根据不同时间段的网络主题分布对网络舆情变化有效监控，甚至积极引导其发展方向。另外，主题模型在临床诊疗领域也有少量应用，目的在于分析病历文档中疾病-用药和疾病-症状之间的诊治规律，分析过程包括：将每个病历文档作为一个独立样本输入的模型中，通过大量数量的训练，得到最终的主题分析结果。

[0003] 然而，在实现本发明方案的过程中发明人发现：一方面，由于同一疾病的两个患者之间的病情发展具有相似性，导致医生对其做出的诊断方案会受到以前治疗相似患者的诊断方案的影响。另一方面，两个患者之间存在个体差异，例如体质、性别、年龄、病情阶段等，这样医生会根据不同患者出具不同的诊疗方案。实际诊疗过程中，可能存在身体条件和所患疾病具有相似的两个患者，这个他们的诊疗方案也存在相似部分。例如：糖尿病患者会同时患有多种糖尿病并发症，但相同的并发症的诊疗方案和病情发展应该具有相似性。

发明内容

[0004] 针对现有技术中的缺陷，本发明提供了一种考虑相似约束的病历主题获取方法及装置，用于解决相关技术中存在的技术问题。

[0005] 第一方面，本发明实施例提供了一种考虑相似约束的病历主题获取方法，所述方法包括：

[0006] 计算初始病历中任意两个病历文档之间的相似度，得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合；

[0007] 将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。

[0008] 可选地，计算初始病历中任意两个病历文档之间的相似度包括：

[0009] 获取病历的多个相似性计算因素及各相似性计算因素的权重值；

[0010] 分别计算任意两个病历文档关于各相似性计算因素的数值；

[0011] 根据各相似性计算因素的数值和各相似性计算因素的权重值计算所述任意两个病历文档的相似度。

[0012] 可选地，所述相似性计算因素包括：性别属性的距离、年龄所属分段的距离、诊断结果的距离。

[0013] 可选地，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布包括：

[0014] 对所述相似性约束病历集合中各病历文档中每个单词随机赋予主题编号z；

[0015] 重新扫描所述相似性约束病历集合，对每个单词按照$p(z_i = k|\vec{z}_{\neg i}, \vec{w})$重新采样主题，使得到的新主题满足GibbsSampling收敛；

[0016] 统计语料库中主题-单词共现频率矩阵，得到文档-主题分布和主题-单词分布。

[0017] 可选地，所述预设LDA模型包括：

[0018] 任意两个病历文档相似性约束采用主题分布距离dis(θr^m,θr^n)表示，公式为：

$$dis(\theta r^m,\ \theta r^n) = \frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln} d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m * L_n};$$

[0019] 其中θr^m=$\{\theta_{m,1},\theta_{m,2},\cdots,\theta_{m,Lm}\}$，表示每个病历文档包括$L_m$个病程记录；$\theta_{m,Lm}$表示第$L_m$个病程记录的主题；d($\theta_{m,Lm}$,$\theta_{n,Ln}$)表示为两个病程的主题向量之间的欧式距离；

[0020] 所述预设LDA模型还包括Gibbs-EM迭代函数，为：

[0021] $$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \beta) \propto \frac{n_i^{(k)}+\beta_i-1}{\sum_{t=1}^{V}(n_t^{(k)}+\beta_t)-1} * \alpha_{mk};$$

[0022] $n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量。

[0023] 第二方面，本发明实施例提供了一种考虑相似约束的病历主题获取装置，所述装置包括：

[0024] 病历集合获取模块，用于计算初始病历中任意两个病历文档之间的相似度，得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合；

[0025] 主题分布推导模块，用于将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。

[0026] 可选地，所述病历集合获取模块包括：

[0027] 权重值获取单元，用于获取病历的多个相似性计算因素及各相似性计算因素的权重值；

[0028] 因素数据计算单元，用于分别计算任意两个病历文档关于各相似性计算因素的数值；

[0029] 相似度计算单元，用于根据各相似性计算因素的数值和各相似性计算因素的权重值计算所述任意两个病历文档的相似度。

[0030] 可选地，所述相似性计算因素包括：性别属性的距离、年龄所属分段的距离、诊断结果的距离。

[0031] 可选地，所述主题分布推导模块包括：

[0032] 主题编号单元，用于对所述相似性约束病历集合中各病历文档中每个单词随机赋予主题编号z；

[0033] 主题迭代单元，用于重新扫描所述相似性约束病历集合，对每个单词按照$p(z_i = k|\vec{z}_{\neg i}, \vec{w})$重新采样主题，使得到的新主题满足GibbsSampling收敛；

[0034] 主题分布计算单元，用于统计语料库中主题-单词共现频率矩阵，得到文档-主题分布和主题-单词分布。

[0035] 可选地，所述预设LDA模型包括：

[0036] 任意两个病历文档相似性约束采用主题分布距离dis(θr^m,θr^n)表示，公式为：

$$dis(\theta r^m,\ \theta r^n)=\frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln}d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m*L_n};$$

[0037] 其中$\theta r^m=\{\theta_{m,1},\theta_{m,2},\cdots,\theta_{m,Lm}\}$，表示每个病历文档包括$L_m$个病程记录；$\theta_{m,Lm}$表示第$L_m$个病程记录的主题；$d(\theta_{m,Lm},\theta_{n,Ln})$表示为两个病程的主题向量之间的欧式距离；

[0038] 所述预设LDA模型还包括Gibbs-EM迭代函数，为：

[0039] $$p(z_i=k|\vec{z}_{\neg i},\vec{w},\alpha,\beta)\propto\frac{n_i^{(k)}+\beta_i-1}{\sum_{t=1}^{V}(n_t^{(k)}+\beta_t)-1}*\alpha_{mk};$$

[0040] $n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量。

[0041] 由上述技术方案可知，本发明实施例中通过计算两个病历文档的相似度，可以从初始病历中筛选出大于或等于相似度阈值的多个病历文档，后续过程中利用多个病历文档构成的相似性约束病历集合作为主题分析文档。可见，本实施例中可以很好地模拟医生诊疗过程确定病历文本的思维过程，从而有利于获取主题的准确度。

附图说明

[0042] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些图获得其他的附图。

[0043] 图1为本发明一实施例提供的考虑相似约束的病历主题获取方法的流程示意图；

[0044] 图2为病历文档中病程记录；

[0045] 图3为男性患者糖尿病并发症数量分布图；

[0046] 图4为女性患者糖尿病并发症数量分布图；

[0047] 图5为相似度阈值分别为0.5和0.6时主题数量和相似度约束指示SIM之间关系的示意图；

[0048] 图6为相似度阈值分别为0.7和0.8时主题数量和相似度约束指示SIM之间关系的示意图；

[0049] 图7为主题数与交交互信息之间的关系；

[0050] 图8～图10是本发明一实施例提供的考虑相似约束的病历主题获取装置的框图。

具体实施方式

[0051] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0052] 图1为本发明一实施例提供的考虑相似约束的病历主题获取方法的流程示意图。参见图1，一种考虑相似约束的病历主题获取方法包括：

[0053] 101，计算初始病历中任意两个病历文档之间的相似度，得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合；

[0054] 102，将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。

[0055] 下面结合附图和实施例对考虑相似约束的病历主题获取方法的各步骤作详细描述。

[0056] 首先，介绍101，计算初始病历中任意两个病历文档之间的相似度，得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合的步骤。

[0057] 患者在住院治疗过程中，会产生各种检测记录，例如入院记录、出院记录、病程记录、会诊记录等。若直接计算检测记录之间的相似性，则会极大的增加计算量。为方便说明，本实施例中将处理前的检测记录称之为初始病历。

[0058] 为降低计算量，本实施例中仅考虑初始病历中入院诊断部分的相似性。在一实施例中，相似性即是计算任意两份初始病历的距离，并且病历相似性约束构建可理解为收集两两间距离小于某个阈值的病历集合。

[0059] 实际应用中，初始病历中还会包括某个病症的多种并发症，例如糖尿病会导致多种并发症，如表1所示。

[0060] 表1糖尿病患者并发症示例

[0061]

序号	姓名	性别	年龄	入院诊断
1	张一	男	40	1 型糖尿病伴多并发症
2	李一	女	55	糖尿病伴多并发症、糖尿病伴周围血管病变
3	陈一	女	57	2 型糖尿病、2 型糖尿病性酮症
4	金一	男	27	糖尿病、糖尿病性肾病
5	沐一	女	44	糖尿病、高血压病（3 级 很高危） 高脂血症
6	李二	男	66	2 型糖尿病伴多并发症、2 型糖尿病性酮症

[0062] 分析表1可知，不同年龄段的患者对糖尿病及其并发症表征存在差异；加之，不同年龄段患者对药剂的承受能力不同，导致在临床诊疗过程中会存在表征、用药等方面的不同。因此，在计算病历文档的相似性时需要考虑患者的基本信息，本实施例中将患者姓名和年龄纳入病历文档的相似性计算因素。

[0063] 在一实施例中，将相同性别之间性别属性的距离设置为1，不同性别之间性别属性的距离设置为0，如下式所示：

[0064] $$d(sex_i, sex_j) = \begin{cases} 1 & sex_i = sex_j \\ 0 & sex_i \neq sex_j \end{cases} \quad (1)$$

[0065] 其中，sex_i，sex_j表示为不同两个人的性别。

[0066] 在一实施例中，根据国际人口年龄结构将年龄划分为4个年龄段，分别为：少年，0～17岁，表示为1；青年，18～45岁，表示为2；中年，18～45岁，表示为3；老年，大于59岁，表示为4。这样，本实施例可以计算两个患者所属年龄段的距离，如下式表示：

[0067] $$d(age_i, age_j) = 1 - \frac{|flag_i - flag_j|}{3} \quad (2)$$

[0068] 其中，age_i，age_j表示为不同两个人的年龄，$flag_i$，$flag_j$表示不同年龄所属分段。并且，两个年龄所属分段越靠近则距离越小，所属分段越远则距离越大。

[0069] 考虑到初始病历中采用离散型的文本化描述，本实施例中采用Jaccard距离计算不同初始病历中诊断结果之间的距离，如下式所示：

[0070] $$d(dia_i, dia_j) = \frac{(dia_i \cap dia_j)}{(dia_i \cup dia_j)} \quad (3)$$

[0071] 其中，dia_i，dia_j表示病历i和病历j的出院诊断布尔向量空间，大量本文考虑糖尿病并发症之间的病症。

[0072] 例如：$dia_i = \{123\}$，$dia_j = \{234\}$，$dia_i \cap dia_j = \{2,3\}$；$dia_i \cup dia_j = \{1,2,3,4\}$，那么$d(dia_i, dia_j) = 2/4 = 0.5$。

[0073] 需要说明的是，本实施例中仅考虑了所述相似性计算因素包括：性别属性的距离、年龄所属分段的距离、诊断结果的距离的情况，在文本主题获取方法的应用场景改变时，相似性计算因素的具体组成也可以作相应的调整，调整后的方案同样落入本申请的保护范围。

[0074] 在确定出相似性计算因素后，分别设置权重调节调节参数μ_1，μ_2，μ_3，并计算任意两个初始病历之间的相似度，如下式所示：

[0075] $sim(T_i, T_j) = \mu_1 * d(sex_i, sex_j) + \mu_2 * d(age_i, age_j) + \mu_3 * d(dia_i, dia_j)$

[0076] (3)

[0077] $\mu_1 + \mu_2 + \mu_3 = 1$ (4)

[0078] $0 \leqslant \mu_1, \mu_2, \mu_3 \leqslant 1$ (5)

[0079] 最后，将相似度与相似度阈值τ作比较，筛选出相似度值大于或者等于相似度阈值的多个初始病历，并得到多个初始病历构成的相似性约束病历集合，记做$D = \{(T_i, T_j) | i, j \in [1, M]\}$。

[0080] 其次，介绍102，将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布的步骤。

[0081] 本实施例中，预设LDA模型是在现有的LDA模型基础上改进得到的。为方便技术人更好的理解预设LDA模型，先描述一下LDA模型的基本原理：

[0082] 潜在狄利克雷分布(Latent Dirichlet Allocation，LDA)是一种主题模型，其目的是寻找文档主题，包含文档、主题和单词三层结构，并且每篇文档都有各自主题相关的概率分布，而文档中单词是由不同主题分布抽样，如下式(6)所示：

[0083] $\sum p(单词|文档) = \sum p(单词|主题) * p(主题|文档)$ (6)

[0084] 利用LDA模型对病历文档建模，设有病历文档总数M，第m个病历文档中存在Nm个临床描述单词，每个单词表示为$\omega_{m,n}$，根据现有的词袋模型(bag of words)将文档和单词表示为文档-主题分布和主题-单词分布。在病历文本中主题可以理解为用药、观察、症状、手术等临床护理手段的统称，每个病历文本是多个主题的多项式分布，即每个病历文本是由临床护理过程中的多个步骤组合而成。

[0085] 相关技术中，LDA模型生成病历文本的步骤，如表2所示。

表 2 LDA 模型生成病历文本步骤

[0086]

1、现有 M 篇病历，每个病历含有 Nm 个单词，对于病历 m 中第 n 个临床描述单词$\omega_{m,n}$，设其主题为 zm,n；

2、生成先验分布参数为$\vec{\alpha}$的狄利克雷分布$\overrightarrow{\theta_m}$，作为病历 m 的文档-主题分布，从文档-主题分布$\overrightarrow{\theta_m}$中抽样得到主题 zm,n；

3、生成先验分布参数为$\vec{\beta}$的狄利克雷分布$\overrightarrow{\varphi_k}$，作为主题 k 的主题-单词分布；

4、当临床描述单词$\omega_{m,n}$主题为 zm,n 时，意味从主题-单词分布$\overrightarrow{\varphi_k}\left|k=z_{m,n}\right.$中抽样出临床描述单词$\omega_{m,n}$。

[0087] 可理解的是，由于每个主题是多个单词的多项式分布，对应每个临床护理步骤包含多个临床实际操作，并且文档-主题分布和主题-单词分布均符合狄利克雷参数为α和β先验分布，因此LDA模型能够很好模拟医生在诊疗过程中做出病历文本的思维过程。

[0088] 基于上述分析可知，LDA模型推理目的在于：通过当前测试文档集计算出LDA模型中的未知参数$(\vec{\alpha},\vec{\beta})$，并根据$(\vec{\alpha},\vec{\beta})$计算主题-单词分布和文档-主题分布。实际上，计算过程中可以直接推导出主题-单词分布和文档-主题分布，而无需计算$(\vec{\alpha},\vec{\beta})$。

[0089] 实际应用中，LDA模型的参数推理算法包括Gibbs抽样和EM变分两种。下面介绍两种方法。

[0090] 第一，Gibbs Sampling核心思想是马尔科夫蒙特卡洛(MCMC)方法，在每一次迭代过程中只改变一个维度的参数值，直到收敛输出待估计参数值。根据狄利克雷参数估计，推理可得到：

[0091] $$\hat{\theta}_{mk}=\frac{n_{m,\neg i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}\left(n_{m,\neg i}^{(k)}+\alpha_k\right)} \quad (7)$$

[0092] $$\hat{\varphi}_{kt}=\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}\left(n_{k,\neg i}^{(t)}+\beta_t\right)} \quad (8)$$

[0093] $$p(z_i=k|\vec{z}_{\neg i},\vec{w})\propto\hat{\theta}_{mk}*\hat{\varphi}_{kt} \quad (9)$$

[0094] 其中：$\hat{\theta}_{mk}$表示文档-主题分布，$\hat{\varphi}_{kt}$表示主题-单词分布，$p(z_i=k|\vec{z}_{\neg i},\vec{w})$表示单词$\vec{w}$分布为k的概率，i为一个数据对(m,n)，表示第m个文档中的第n个词。

[0095] 由于共有K个主题，因此需要进行K次迭代，采用训练步骤如表3所示：

[0096]

表 3 LDA 模型训练过程
1，随机进行主题初始化，对文本集每个文档中每个单词随机赋予主题编号 z； 2，重新扫描文本集，对每个单词按照$p(z_i = k\|\vec{z}_{\neg i}, \vec{w})$重新采样主题； 3，重复步骤 2，直到 Gibbs Sampling 收敛； 4，统计语料库中主题-单词共现频率矩阵，得到文档-主题分布和主题-单词分布。

[0097] 第二,EM变分算法在于寻找合适的参数,使得文本集中所观测到的主题-单词分布概率最大,类似于极大似然估计问题。EM变分算法分为两个迭代步骤：

[0098] 变分E-step考虑原步骤中后验概率p(w|α,β)公式求导困难,引入变分参数(γ,φ)求得近似后验概率分布q(θ,z|γ,φ)。

[0099] 变分M-step根据变分E-step的变分参数最大化近似函数L(γ,φ|α,β)。其中,先验狄利克雷分布参数(α,β)决定了主题-单词分布和文档-主题分布θ,w代表单词,z代表主题。

[0100] 由于LDA模型的迭代目标是最大化词语出现概率p(Z,W|α,β),这样可以有效满足糖尿病病程记录的数据特征,同时也会导致相似病历的主题分布出现较大差异,从而导致无法根据病历主题分布对病历进行有效的统计分析。

[0101] 为建立一个满足病历相似性约束的主题模型,本实施例中通过改变Gibbs抽样收敛条件策略来实现这一目标。

[0102] 考虑到每个病历中会同时存在多个按时间排序的病程记录,病历文档相似性计算应该考虑各病历文档中不同病程记录集合之间的相似性,即相似性约束病历集合D中各病历文档的不同病程记录集合的文档-主题分布尽可能相似。

[0103] 设T_m表示编号m的病历,包括L_m个病程记录,其病程记录的主题集合表示为$\theta r^m = \{\theta_{m,1}, \theta_{m,2}, \cdots, \theta_{m,Lm}\}$。存在两个病历文档的病程记录主题集合$\theta\gamma^m$,$\theta r^n$,可以利用两两主题分布距离均值计算病历相似性约束,如下：

[0104] $$dis(\theta r^m,\ \theta r^n) = \frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln} d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m * L_n} \qquad (10)$$

[0105] 其中,$d(\theta_{m,Lm},\theta_{n,Ln})$表示为两个病和向量之间的欧式距离,$dis(\theta r^m,\theta r^n)$越大表示相似度越低。

[0106] 最大目标函数可以修改为：

[0107] $$L(\alpha,\beta) = \log(p(Z,W|\alpha,\beta)) - \gamma\sum_{(\theta r^m,\theta r^n)\in D} dis(\theta r^m,\ \theta r^n) \qquad (11)$$

[0108] 本实施例中采用Gibbs-EM迭代方法进行LDA模型推导,将其将文档-主题分布α_m修改为正态分布μ_m,得到预设LDA模型：

[0109] $$\alpha_{mk} = \frac{e^{\mu m,k}}{\sum_{k1=1}^{K} e^{\mu m,k1}} \qquad (12)$$

[0110] 其中,μ_{mk}代表病历文档m属于主题k的概率,既然认为μ_m服从标准正态分布,则改进最大目标函数如下表达：

[0111] $$L(\mu) = \log(p(Z, W|\mu, \beta)) + \log p\left(\mu\middle|\left(0, N(0,1)\right)\right) - \gamma * \sum_{(\theta r^m, \theta r^n)\in D} dis\left(\theta r^m,\ \theta r^n\right) \quad (13)$$

[0112] 另外，本实施例中在采样过程中先固定文档主题分布α_m，那么Gibbs-EM迭代函数表达式为：

[0113] $$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \beta) \propto \frac{n_i^{(k)} + \beta_i - 1}{\sum_{t=1}^{V}(n_t^{(k)} + \beta_t) - 1} * \alpha_{mk} \quad (14)$$

[0114] 其中，$n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量，由于采用正态分布来代替原来的α，所以公式(14)可以用随机梯度下降方法进行推导，模型训练过程如表4：

[0115]

表 4 Gibbs-EM 模型训练过程

输入：M 篇文档，K 个主题，文档集合单词总数 V，常数$\beta = 0.1$

输出：“文档-主题”分布$\hat{\theta}_m^k$，“主题-单词”分布$\hat{\theta}_k^t$；

1，随机进行参数初始化，对 M 篇(所有)文档中每篇文档中每个单词随机赋予主题编号 z，对 M 篇(所有)文档中随机赋予主题参数 μ；

2，进行 E-step 迭代：固定参数θ_m^k，根据$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \beta)$对单词主题进行采样，对文本进行迭代直到收敛；

3，进行 M-Step 迭代：

(1)计算目标函数$L(\mu)$;

(2)通过随机梯度下降方法迭代$\mu_{m,k}$:

[0116]

For $m=1:M$:

For $k=1:K$

$$\mu_{m,k(n+1)}^t = \mu_{m,k(n)}^t - \sigma * \frac{\partial(L(\mu))}{\mu_{m,k(n)}^t}$$

End

End

(3)重新计算$L(\mu)$，若目标函数不收敛，则调回步骤(1)；若目标函数收敛，则记下最有$\mu_{m,k}$，M-step 结束；

4，计算$\hat{\theta}_m^k$和$\hat{\theta}_k^t$，若收敛则输出；否则返回步骤 2，重复 EM 步骤。

[0117] 之后，将相似性约束病历集合中各病历文档依次输入到预设LDA模型，通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。

[0118] 至此，本发明实施例中在分析文本挖掘对医疗诊断的影响以及潜在狄利克雷主题模型的建模过程和推理方法的基础上，设计了基于病历相似度约束的预设LDA模型。该预设LDA模型不仅仅考虑到不同病历文档之间的相似性约束，而且确定了医疗文本主题建模目标、推理过程和模型相关度量指标，从而可以从设LDA模型能够清晰反映各个诊疗阶段的侧重点以及病情演化过程，有利于提升病历主题挖掘的科学性、有效性和准确性。

[0119] 下面采用LDA模型和本申请的预设LDA模型(后续称Medical Record Similarity based Latent Dirichlet Allocation,MRS-LDA)来对比实验来说明本发明实施例提供的一种考虑相似约束的病历主题获取方法的有效性和优越性。

[0120] 其中，初始病历采用安徽医科大学第一附属医院内分泌科患者的住院病历，包括2015年至2017年总共1294个糖尿病患者的住院记录，每份病历文档主要包括入院记录、病程记录(如图2所示)、会诊记录和出院记录等。其中男女患者病历文档个数比例648:646，大致相同。

[0121] 参见图3和图4，在安徽医科大学第一附属医院接诊的糖尿病患者中，根据入院诊断判别不同年龄段以及不同性别的患者在同时患有的并发症数量上有明显区别。其中，老年人同时患有的糖尿病并发症数量较其他年龄段的人数大大增加，中年人多同时患有3到5种并发症，青年人会有糖尿病发生，但并没有更多的并发症出现，幼儿患糖尿患者数较少。

[0122] 本实施例中选取入院记录中患者的性别、年龄和入院诊断作为病历相似性约束计算数据基础，利用医生在患者住院期间的病程记录进行相关主题分析。在实验过程中，还可以做如下处理，包括：

[0123] (1)使用python爬虫方法，从1294位患者HTML格式的病历文档中分割入院记录、出院记录、病程记录等各个阶段的文本记录，同时分离出需要的患者信息、诊断结果以及病程记录文本。

[0124] (2)构建词典和停止词词库。本发明的研究内容是和医学相关的一些症状、药品、治疗方式等单词，除此之外病历文本中包含了大量对本文无关的单词，在统计了各个单词在病历中出现的频次之后，手动提取了12599个单词作为停止词添加到停止词库。同时，添加了ICD10中国的疾病名称作为补充的特征添加到了词典中。

[0125] (3)利用python中的jieba分词作为分词工具，使用上述词典和停止词库进行分词和去除停止词操作。

[0126] 考虑到在病历文档主题挖掘中，主题数量对文本主题建模的影响以及相似度阈值不同带来的相似病历数量不同，本实施例中将相似度阈值和主题数量为调节参数，病历相似度阈值τ取值范围为0.5～0.8，主题数量K＝7,10,13,15,20,30，在以上各个参数下分别计算模型的PMI-Score和病历相似性约束。

[0127] 参见图5和图6，MRS-LDA模型与LDA模型在不同的主题参数和不同相似度下相似度约束结果对比，其中横坐标为主题数量K，纵坐标为相似度约束指标SIM。对比分析MRS-LDA模型在病历相似度约束上有明显优势。当主题相似度阈值一致时，伴随主题数量的增加，病历相似度约束有着不明显的下降，但MRA-LDA模型较LDA模型在病历相似度约束指标方面仍然存在较大优势。

[0128] 参见图7,在不同主题参数和不同相似度阈值下,MRS-LDA模型与LDA模型交交互信息(PIM-Score)的结果对比,其中横坐标为主题数量K,纵坐标为度量指标PIM-Score。在主题数量K＝15时,MRS-LDA模型在PIM-Score度量指标上优于LDA模型,并且在病历相似度阈值为0.5时比LDA模型好。

[0129] 通过对比实验,MRS-LDA模型在相似性约束度量指标上有很好的表现,在相同病历相似性阈值和主题数量下,MRS-LDA模型得出的相似病历的主题分布之间的距离更小,能够更好的描述相似病历之间的存在的关联。也就是说,本发明中在构造目标函数时添加了病历相似这一约束条件,可以使相似病历之间的主题分布较为接近,能够适用于病历主题挖掘的使用场景,且准确度较高。

[0130] 第二方面,本发明实施例提供了一种考虑相似约束的病历主题获取装置,参见图8,所述装置包括:

[0131] 病历集合获取模块801,用于计算初始病历中任意两个病历文档之间的相似度,得到所述相似度大于或等于相似度阈值的多个病历文档构成的相似性约束病历集合;

[0132] 主题分布推导模块802,用于将所述相似性约束病历集合中各病历文档依次输入到预设LDA模型,通过所述预设LDA模型推导各病历文档的文档-主题分布和主题-单词分布。

[0133] 可选地,参见图9,所述病历集合获取模块801包括:

[0134] 权重值获取单元901,用于获取病历的多个相似性计算因素及各相似性计算因素的权重值;

[0135] 因素数据计算单元902,用于分别计算任意两个病历文档关于各相似性计算因素的数值;

[0136] 相似度计算单元903,用于根据各相似性计算因素的数值和各相似性计算因素的权重值计算所述任意两个病历文档的相似度。

[0137] 可选地,所述相似性计算因素包括:性别属性的距离、年龄所属分段的距离、诊断结果的距离。

[0138] 可选地,参见图10,所述主题分布推导模块802包括:

[0139] 主题编号单元1001,用于对所述相似性约束病历集合中各病历文档中每个单词随机赋予主题编号z;

[0140] 主题迭代单元1002,用于重新扫描所述相似性约束病历集合,对每个单词按照$p(z_i = k|\vec{z}_{\neg i}, \vec{w})$重新采样主题,使得到的新主题满足GibbsSampling收敛;

[0141] 主题分布计算单元1003,用于统计语料库中主题-单词共现频率矩阵,得到文档-主题分布和主题-单词分布。

[0142] 可选地,所述预设LDA模型包括:

[0143] 任意两个病历文档相似性约束采用主题分布距离$dis(\theta r^m, \theta r^n)$表示,公式为:

$$dis(\theta r^m,\ \theta r^n) = \frac{d\left(\sum_{lm=1}^{Lm}\sum_{ln=1}^{Ln} d(\theta_{m,Lm},\theta_{n,Ln})\right)}{L_m * L_n};$$

[0144] 其中$\theta r^m = \{\theta_{m,1}, \theta_{m,2}, \cdots, \theta_{m,Lm}\}$,表示每个病历文档包括$L_m$个病程记录;$\theta_{m,Lm}$表示第$L_m$个病程记录的主题;$d(\theta_{m,Lm}, \theta_{n,Ln})$表示为两个病程的主题向量之间的欧式距离;

[0145] 所述预设LDA模型还包括Gibbs-EM迭代函数，为：

[0146] $$p(z_i = k|\vec{z}_{\neg i}, \vec{w}, \alpha, \beta) \propto \frac{n_i^{(k)}+\beta_i-1}{\sum_{t=1}^{V}(n_t^{(k)}+\beta_t)-1} * \alpha_{mk};$$

[0147] $n_i^{(k)}$代表相似性约束病历集合中主题为k的单词i的数量。

[0148] 需要说明的是，本发明实施例提供的考虑相似约束的病历主题获取装置与上述方法是一一对应的关系，上述方法的实施细节同样适用于上述装置，本发明实施例不再对上述系统进行详细说明。

[0149] 本发明的说明书中，说明了大量具体细节。然而，能够理解，本发明的实施例可以在没有这些具体细节的情况下实践。在一些实例中，并未详细示出公知的方法、结构和技术，以便不模糊对本说明书的理解。

[0150] 最后应说明的是：以上各实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述各实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分或者全部技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的范围，其均应涵盖在本发明的权利要求和说明书的范围当中。

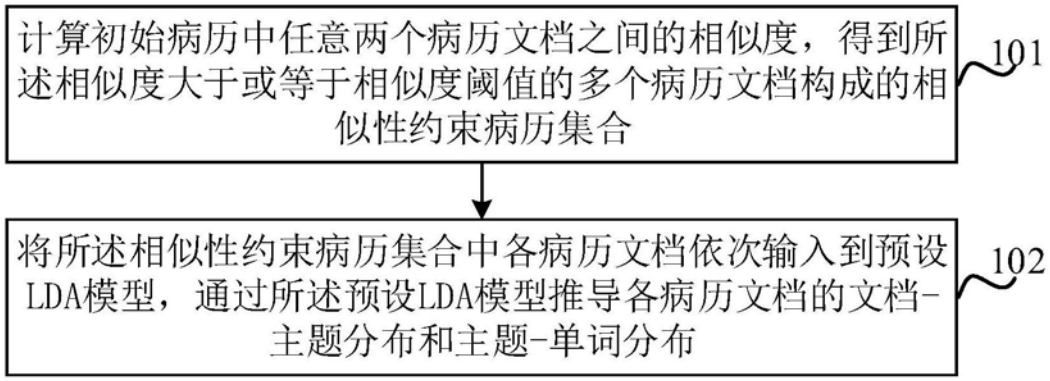


图1

安徽医科大学第一附属医院

病 程 记 录

姓名： XX 科别:内分泌科一病区 病区:内分泌一病区 床号:0105 住院号:2007035380

2015-05-20 14:54 YY 主任医师查房记录

今随 YY 主任医师查房，患者无特殊不适主诉，查体同前。辅检：2015.5.17甲功三项未见明显异常；2015.5.18尿微量白蛋白/肌酐：尿微量白蛋白 261.00mg/l，肌酐 16.3mmol/l，尿A/C 16.00mg/mmol；糖化血红蛋白 11.90%，2015.5.19 24h尿蛋白：24小时尿蛋白 0.29g/24H,尿量 1.6001；肝胆胰脾+泌尿系+双下肢血管+颈部血管B超：双侧椎动脉发育不对称，双下肢足背动脉扫及颗点光斑形成，随访； YY 查房后指出：现患者诊断考虑：现患者诊断考虑：1型糖尿病、糖尿病周围血管病变、糖尿病伴肾并发症、糖尿病伴眼并发症。治疗上予以门冬胰岛素联合甘精胰岛素调整血糖，硫辛酸抗氧化、胰激肽原酶联合银杏达莫改善循环，腺苷钴胺联合维生素B1营养神经；现患者为I型糖尿病，血糖波动大，今日予以加测三餐前血糖，防止发生低血糖，建议患者使用胰岛素泵便于血糖调整，待待眼科会诊能否行转科手术治疗白内障，病情继观。

医师签名： ZZ

第 1 页

图2

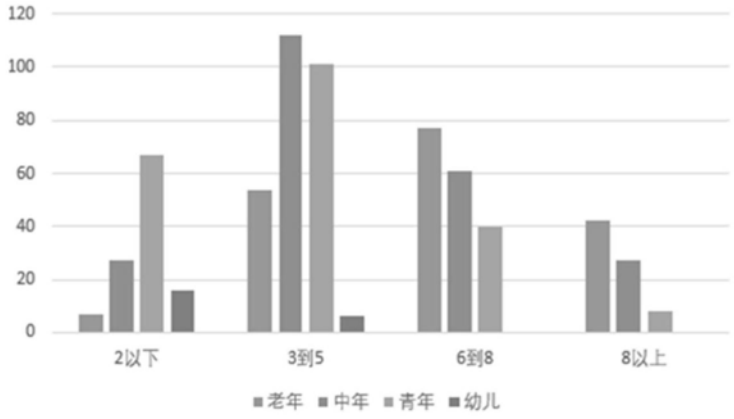

120
100
80
60
40
20
0
2以下
3到5
6到8
8以上
老年 中年 青年 幼儿

图3

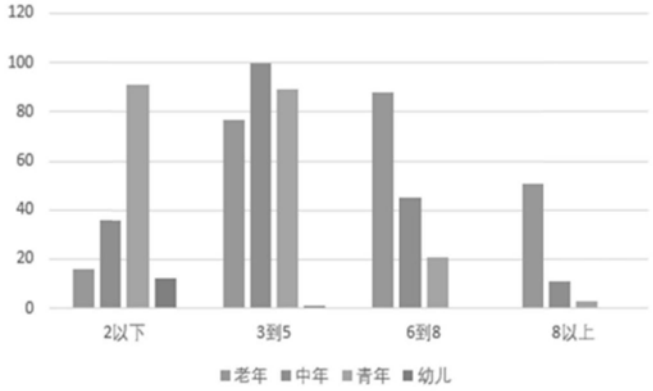

120
100
80
60
40
20
0
2以下
3到5
6到8
8以上
老年 中年 青年 幼儿

图4

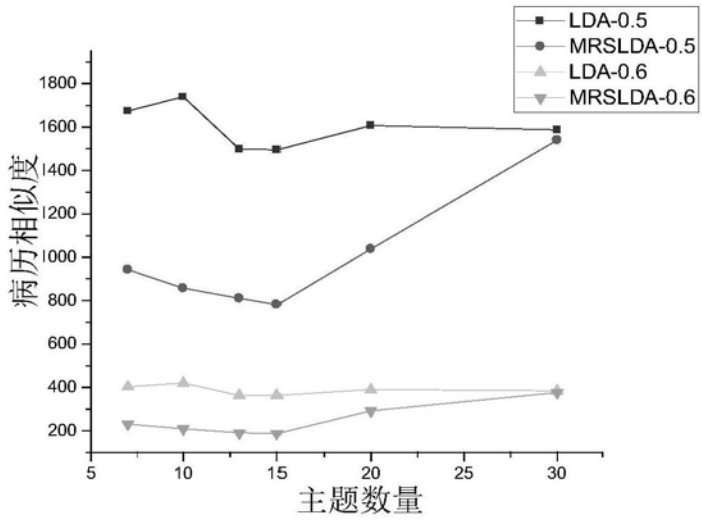

LDA-0.5
MRSLDA-0.5
LDA-0.6
MRSLDA-0.6
病历相似度
1800
1600
1400
1200
1000
800
600
400
200
5
10
15
20
25
30
主题数量

图5

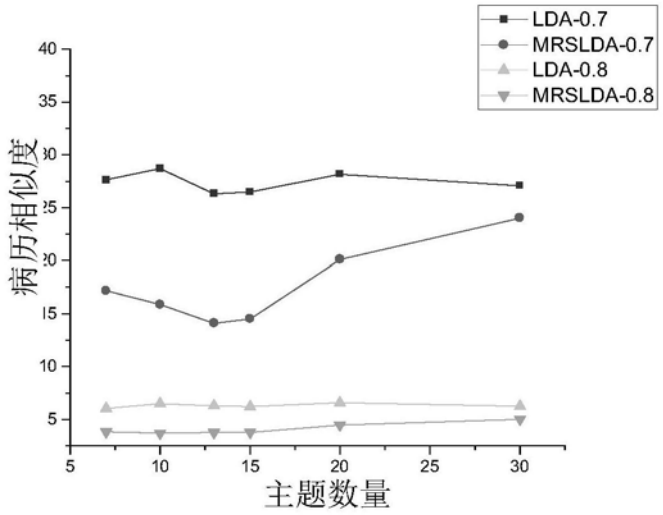

LDA-0.7
MRSLDA-0.7
LDA-0.8
MRSLDA-0.8
病历相似度
主题数量

图6

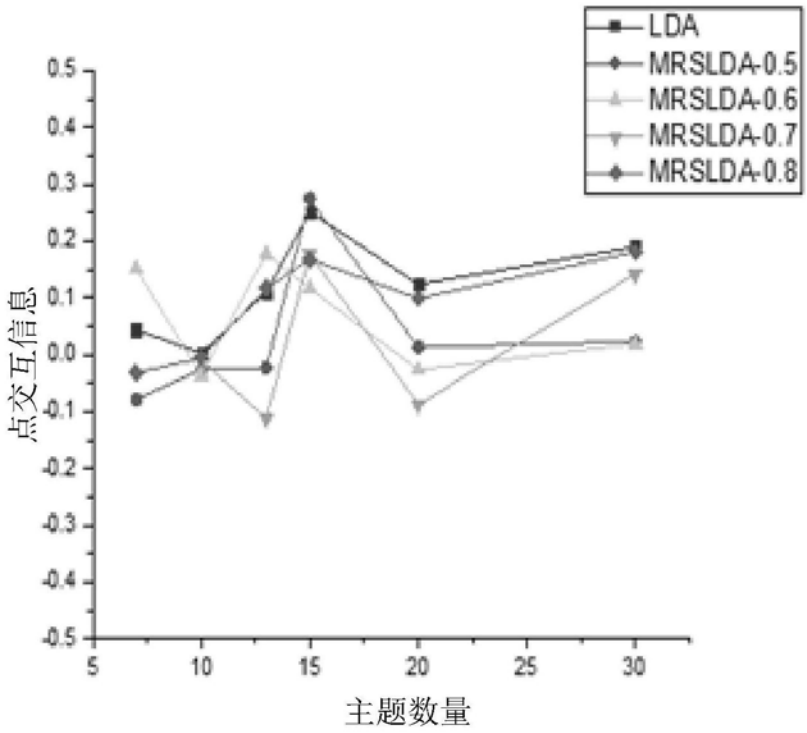

LDA
MRSLDA-0.5
MRSLDA-0.6
MRSLDA-0.7
MRSLDA-0.8
点交互信息
主题数量

图7

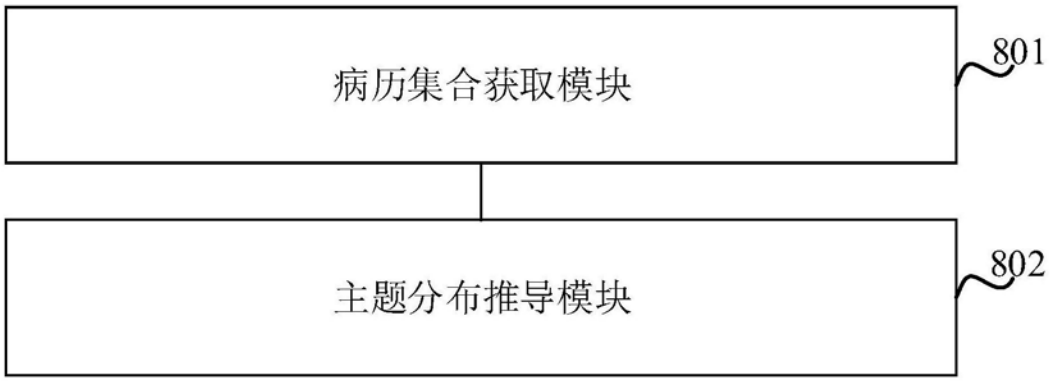

病历集合获取模块
801
主题分布推导模块
802

图8

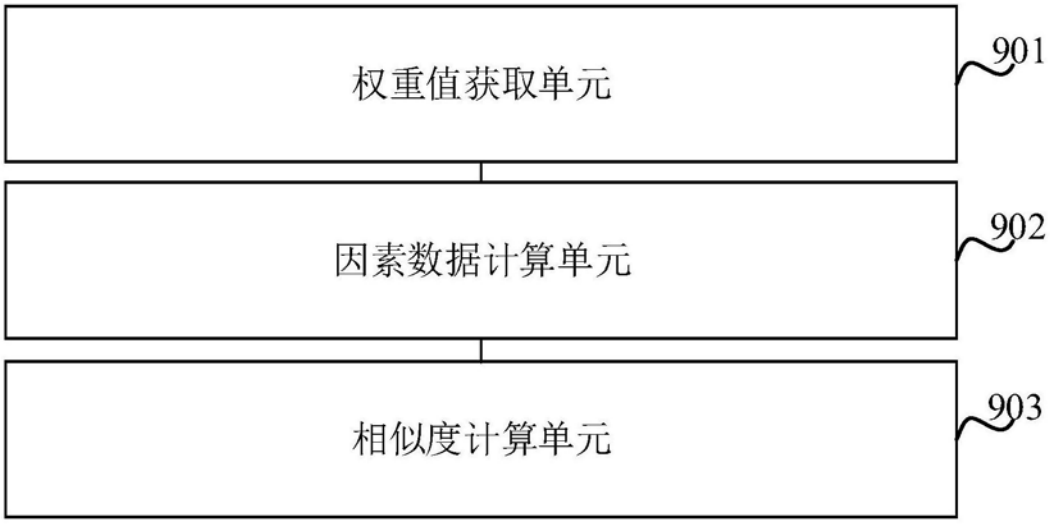

权重值获取单元
901
因素数据计算单元
902
相似度计算单元
903

图9

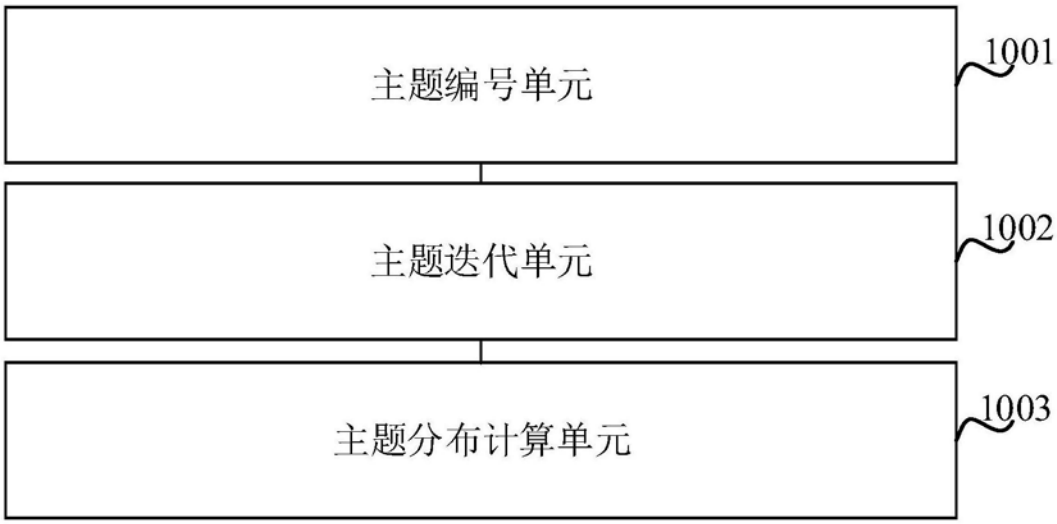


图10