维普资讯 http://www.cqvip.com

电子病历数据预处理技术*)

庄 军[1,4] 郭 平[1] 周 杨[1] 白桂花[2] 王月毅[3]

(重庆大学计算机学院 重庆 400044)[1] (第三军医大学西南医院 重庆 400038)[2]

(重庆大学电气工程学院 重庆 400045)[3] (解放军第322医院信息科 大同 037006)[4]

摘 要 多年积累的电子病历是一项重要的不可再生资源,对其数据的有效处理、利用是一项非常必要和有意义的工作。本文研究了电子病历数据前期处理的相关技术,涉及从电子文档资料到基础数据库的转换,对存入数据库的数据实施数据清理和数据变换等。通过数据预处理可以消除数据中的噪声、不完整和不一致性,实现数据的规范化和有效压缩,从而形成高质量的数据,可使数据的再处理(统计、数学建模、数据挖掘等)更加有效。

关键词 电子病历,数据预处理,数据清理,数据规约

Computer-based Patient Record Data Pre-processing Technologies

ZHUANG Jun[1,4] GUO Ping[1] ZHOU Yang[1] BAI Gui-Hua[2] WANG Yue-Yi[3]

(College of Computer Science, Chongqing University, P. R. China, Chongqing 400044)[1]

(Southwest Hospital of Third Military Medical University, Chongqing 400038)[2]

(College of Electrical Engineering, Chongqing University, Chongqing 400044)[3]

(Department of Information of 322nd Hospital of PLA, Datong 037006)[4]

Abstract The computer-based patient record diabetes that accumulated by many years is not re-genesis, so the effective management and use are an important and meaningful work. In this paper, we study the pre-processing technologies of computer-based patient record data, which involve the electronic document converted into the recording of database, data cleaning, data transformation and data reduction. Based on the pre-processing technologies, we can solve the noise, the non-integrity and the inconsistency of data, so data can be classified and compressed. Finally, we can obtain high quality data, which is more effective on data processing.

Keywords CPR, Data pre-processing, Data cleaning, Data reduction

1 引言

近年来,医院信息系统得到了广泛的应用。随着应用的深入和市场化的要求,特别是医疗保障体制对医院信息需求的深化,如何处理和利用医院信息系统中大量的信息已经成为至关重要的问题。电子病历(Computer-based Patient Record,CPR)是用"电子"的方式保存个人终身的健康和保健信息[1]。电子病历替代纸张病历,作为健康医疗信息的主要资源可满足所有临床、法律和行政的需要。但目前,电子病历仅局限于病程记录的电子文本,简单地说每份电子病历记录着医生对患者从入院到出院的整个治疗行为过程。

电子病历中包含的医疗信息是相当丰富的。通过对电子病历中的大量信息的分析和应用,可以为医疗、教学、科研和医院管理提供包括病历检索、智能知识库、医疗质量统计、医疗评价、质量评估、健康评估、揭示诊疗过程中警示等相当广泛的服务功能[2]。实现这些服务的前提条件是必须运用数据预处理技术,对电子病历进行数据分析,提取出完整、一致和规范化的信息。本文研究了从电子病历中提取完整、一致和规范化信息的一般步骤,并结合实际的病历数据实现了从电子病历文本数据到数据库数据的转换和处理。

2 数据预处理的基本过程

医院信息系统通过多年的实际应用,已经积累了大量历史数据,这些电子病历文档是非常有意义和不可再生的资料。如果再经过有效的数据处理,结合医院信息系统中其它的相关信息,就有可能从中发现有效的、潜在有用的规律,从而为建立医学研究、疾病防控、医疗政策制定等提供有效决策支持。

目前使用的医院信息系统中的电子病历大多以文本形式出现,病人的信息分散在病历、病程记录、护理记录和实验检查等不同的文本、表格中,不同记录反映的是不同时间病人的临床平面信息,医生只能在浏览全部的病历资料后,经过分析才能得到疾病的发生、发展与转归的"立体化"信息[3]。因此有必要将分散的病人信息精确提取出来存放在数据库或数据仓库中,成为临床数据仓库的一部分。

一份完整的电子病历主要包括入院记录和病程记录两大部分。入院记录,记录了医生用医学术语对病人入院时病情状况的描述,主要有主诉、现病史、过去史、个人史、家庭史、体格检查、辅助检查和初步诊断、最后诊断等部分;病程记录则记录了病人在整个住院期间的各种检查结果、治疗过程和病情变化情况。因篇幅所限,略去电子病历的详细内容和格式。

*)基金项目:国家自然科学基金项目(编号:50378093)。庄 军 高级工程师,硕士研究生,主研方向:数据挖掘;郭 平 副教授,博士,主研方向:AI、GIS,数据挖掘。

维普资讯 http://www.cqvip.com

从电子病历中的信息转变到数据库中的数据需要运用数据预处理技术。数据预处理即是要消除电子病历中存在的数据不完整、冗余、不一致以及噪声等。图1是电子病历预处理的一个一般过程。其中的词典库存放医学专业各类术语——特征词[4]。数据抽取，又叫信息抽取，目的是从电子病历中扫描并抽取出与特征词相关的数据；数据清理操作是填充空缺值、平滑噪声数据，识别、删除孤立点；数据变换是通过规格化和聚集形成适合不同特定主题的数据集；数据规约操作是压缩现有的数据集，既能减少数据集的大小又不影响数据的特征和质量[5]。

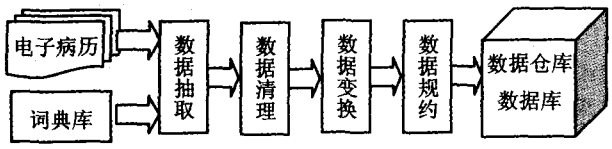

图1　电子病历的数据预处理过程

本文选取了某医院心内科关于冠状动脉粥样硬化性心脏病(或称冠心病)的2000余份电子病历做实验数据进行预处理技术研究，包括数据抽取、数据清理、数据变换、数据归约等。

3　电子病历数据预处理

下面以我们选取的冠心病电子病历为例，讨论数据预处理过程。

3.1　数据抽取

在数据抽取过程中，应用了一种针对中文非结构化数据清理问题的、基于数据挖掘技术的理论模型和抽取方法，其方法具有一定的通用性。例如在作词汇的匹配与合并时，根据中文非结构化数据的特点，应用了一种基于频繁模式挖掘和聚类的匹配与合并算法——C-Cleaner并进行重大的改进。针对具体的应用环境，设计并部分实现了一个数据抽取工具。

从电子病历史中抽数据，重要的是要建立医学专科领域的医学术语词典库，存放所谓的“特征词”。特征词就是在整个数据集中出现次数较少的或在匹配时应赋予较高权重的词或词的集合。做词法分析的工作量及难度都很大，特别是对于中文的处理与西文有很大不同，所以工具设计时回避了这个问题，而采用了直接对字符串进行模糊匹配的方法。匹配后再通过工具辅助用户做标准化和验证。也就是说先从现有数据中自动搜索出所有频繁模式(由于我们不做词法分析，因此无法保证搜索出来的都是词)，再由用户选择其中的非特征词保存起来，然后同样将每条词汇或语句减去这些非特词得到特征词。最后在具有多年临床经验的多名专家、教授指导下，完成特征词的验证和词库的建立工作。

C-Cleaner算法分成两部分，第一部分是基于频繁模式挖掘算法，改进的C-Cleaner算法将原来的Apriori算法换成SCG算法，SCG算法比Apriori算法有更高的效率；第二部分是聚类算法，选用了分级聚类算法。

改进的C-Cleaner算法描述如下：

```
输入:电子病历文档
输出:医学词汇的聚类和编码
步骤:
1 选取电子病历样本;
2 搜索样本,采用SCG算法获得频繁词freword();
3 根据搜索出来的频繁词,建非特征词词库gword();
4 计算样本的特征词feature();
5 计算聚类及编码Clustering();
6 对电子病历进行数据抽取;
7 end.
```

对照词典库从电子病历中抽取信息时，由于电子病历包含的信息量大，我们仅抽取了对冠状动脉硬化性心脏病诊断和治疗起关键作用的部分属性数据来进行分析和处理。这些数据包括以下几类：

1)病史类。病史类中的属性是为了说明病人来院前和当时的病情状况及相关情况。这部分属性主要在电子病历中的主述、现病史、过去史、个人史和家庭史中抽取。如：入院时症状、入院前病情状况、健康状况、生活习性和遗传病史等属性。

2)检查类。检查类中的属性反映来院后病人接受相关检查的结果。该部分属性主要从电子病历中的体检、检验和其它检查以及病程记录中抽取。如：血压、脉搏、呼吸、血常规、肝肾功、电解质、心谱、凝血三项、心电图、心脏彩超、冠状动脉照影等属性。此类属性大多会有多个分属性和多次检查值，因此需要建立这些属性间的关联以及属性值间的关联。

另外还有手术类(PTCA及冠状动脉内支架植入术)和诊断类数据等，在本文中不再一一叙述。

抽取出来的信息还需要编码转换后才能存入数据库中。对抽取出的数字类型的信息，如体温、呼吸、血压、脉搏等，可作为数值型字段直接存入数据库中；对于具有多选性的属性，我们将可选值进行编码，然后将编码值存入数据库。如下是对部分多选性属性的编码：

① 症状描述：是否胸闷，否为1，是为2；是否存在心前区压榨样疼痛，否为1，是为2；是否存在放射状疼痛(左肩部或/和左上臂部等处)，否为1，是为2；是否心悸(或心慌)，否为1，是为2；……

② 个人史描述：是否吸烟，否为1，是为2，已戒为3；是否饮酒，否为1，是为2，已戒为3；

③ 家庭病史描述：是否有冠心病遗传病病史，否为1，是为2；是否有高血压遗传病病史，否为1，是为2；是否有高血脂遗传病病史，否为1，是为2；是否有高胆固醇遗传病病史，否为1，是为2；

④ 实验检查值范围描述：正常为1，高为2，低为3。

对部分检查类数据(如心电图)、检验类数据(如血常规)等，它们由一个属性多个值或多个属性构成，其存储需要做特殊处理。如心电图，将根据心电图诊断的结果(主要参数)存入数据库，而不存储图形本身。

经数据抽取后，我们将电子病历中用于诊断和治疗的主要属性转换成了数据库中的数据，为后续的处理提供了方便。

3.2　数据清理

针对经过抽取后存入数据库的原始数据不完整和不一致的可以采用了填充空缺值、纠正非法值和纠正数据不一致性等数据清理方法进行处理。

(1)填充空缺值。对于没有数据值的属性项，对大部分数据项采取了根据该数据项的含义，定义一个缺省的空缺值，然后用它去替换缺少的空缺项的方法。

如未抽取到“高血压遗传病”时，将高血压遗传病字段置1，入院记录中没有实验检查记录的，从病程记录中抽取补充。如“血常规提示：N73.4%、L18.2%、HCT36.0%、MPV13.8fL、余正常。”在血常规表中填入抽取的对应数据，实验检查值范围的值分别填入2、3、3、2、1。

(2)纠正非法值。电子病历中的每种数据都是有一定的限定范围的，不在此范围的数据均视为非法数据。如果直接

维普资讯 http://www.cqvip.com

将这些数据作为再处理的输入，会大大地影响数据再处理结果和效率，因此，要对这一部分数据做纠正处理。考虑到电子病历的严肃性和经多人多次浏览，非法数据的出现往往是由于医生在书写病历时的疏忽造成的，例如，实验中发现数字“1”和字母“l”、数字“0”和字母“O”容易混淆，而且此类错误用肉眼很难发觉，因此尽可能恢复其正确性必须使用特定的方法。如在抽取“PAl36mg/L”时，首先分离字符、数字型字符和单位分别赋值给变量 s1、s2 和 s3，得 s1＝PAl，s2＝36，s3＝mg/L，从词典库的血脂电解质类中没有搜索到与 s1 完全匹配的记录，确定 s1 的最后一个字符为“l”，将“l”转换成数字字符“1”，再与变量 s2 相加后将字符转换成数字，得到“136”。从实验结果来看，非法数据的产生是由于数字“1”，被录成了字母“l”，使得实验检查结果 136 变成了 36。

(3)纠正不一致数据。原始数据中除了缺失数据外，还存在一些不一致的数据。实际上，某些数据项间存在一定的相关性，可以用这种相关性来查找并纠正这些不一致性的数据，见数据归约处理部分的实例。

3.3 数据变换

数据变换将数据转换成统一的格式，以适合数据的再处理。在抽取出来的数据中，涉及到要转换的数据大多是表现形式上的差异，如：N73.4%、L18.2%、HCT36.0%、MPV13.8fL，数据保存时转换成：0.734、0.182、0.36、13.8。再如：“地高辛片 0.25mg 1/日”，数据保存时转换成：

表 1 曾服用药品表

drugName	dosage	dosageUnit	frequ	frequUnit
地高辛	0.25	mg	1	日

另外，有部分数据因单位的不同而需要用函数转换成同一种单位的表现形式，主要包括血压、空腹血糖、总胆固醇、甘油三酯、高密度脂蛋白胆固醇和低密度脂蛋白胆固醇等。如部分数据的转换函数：

甘油三酯：甘油三酯的单位是 mmole/L，但是有时该数据项是以 mg/dl 计量的，如果该数据大于 25，就认为属于这种情况，转换函数为：$f(V)=V/88.5$

高密度脂蛋白胆固醇：高密度脂蛋白胆固醇的单位是 mmole/L，但是有时该数据项是以 mg/dl 计量的，如果该数据大于 5，就认为属于这种情况，转换函数为：$f(V)=V/38.6$。

在海量数据上进行复杂的数据分析和处理将花费很长的时间，有时导致处理无法完成。数据归约技术可以得到小数据集的归约表示，但仍保持了原数据的完整性。数据归约有维归约、数据压缩、数值归约和概念分层等方法。

概念分层技术通过收集并用较高层的概念替换较低层的概念来定义数值属性的一个离散化。概念分层可以用来归约数据，通过这种概化尽管细节丢失了，但概化后的数据更有意义、更容易理解，并且所需的空间比原数据少。概念分层技术涉及到的部分主要有年龄段、个人史、病史时间属性、各组实验检查数据等，如在填充空缺值部分提及的血常规检查结果。病史时间属性归约成病史时间长和短两种。

维归约技术是通过删除不相关的属性(或维)来减少数据量。属性子集的选择可以用基本子集选择的启发式方法，这种方法主要包括逐步向前选择、逐步向后删除、向前选择和向后删除的结合和判定树归纳技术。本文采用逐步向后删除的技术，即每一步，删除掉在属性集中的属性。

维归约技术涉及到的维主要包括有症状、病史、个人史、辅助检查结果等。下面说明部分属性的维归约方法：

(1)症状部分。症状的属性包括是否胸闷 symptom_1、是否存在心前区压榨样疼痛 symptom_2、是否存在放射状疼痛 symptom_3、是否心悸(或心慌) symptom_4、是否胸痛 symptom_5、是否气促 symptom_6、其它 symptom_7 等。针对冠心病的特征症状的典型症状：心前区压榨样疼痛或放射状疼痛，可以将上述属性和病史 disease、病史时间 disease_his、症状时间 symptom_his 一起归约成一个典型症状属性 symptom_0，其属性值 1、2、3 分别表示典型、不典型和其它。为了纠正数据的不一致性，归约后的新属性值并不完全等同于属性 symptom_2 或 symptom_3 的值，可利用其它属性的值修正这些数据的不一致性。修正的方法如下：

```
if (disease_*='冠心病' and disease_his_*>0) then symptom_0
  =1//典型
  else if symptom_*=2 then symptom_0=1//典型
       else if (symptom_2=2 or symptom_3=2) then symptom_0
            =1//典型
            else if (symptom_5=2 and symptom_his>0) then
                 symptom_0=1//典型
                 else if (symptom_7=2 and symptom_his_7>
                      0) then symptom_0=2//不典型
                      else symptom_0=2 //其它
```

注：* 代表该类属性的所有后缀数字。

(2)病史部分。病史部分的属性包括已知诊断 disease_1、disease_2、disease_3、disease_4 、disease_5。文中取与冠心病有关的主要诊断，病史属性归约成典型病史属性 disease_0，其属性值 1、2、3 分别表示明确、无关和相关。属性 disease_0 的逻辑关系如下：

```
if (disease_*='冠心病' or disease_*='心肌梗死') then disease_
  0=1//明确
  else if (disease_?='高血压' or disease_?='糖尿病' or dis-
       ease_?='高血脂'等)
       then disease_0=3  //相关
       else disease_0=2  //无关
```

注：* 代表该类属性的所有后缀数字。

4 数据预处理实验

经过预处理过的数据即可进行再处理了，而预处理后的数据可以很好地保证电子病历中的各种主要特征。当然也可根据不同目的改变数据预处理方法，本文仅提出了一些比较基本的方法已供参考。

数据预处理的部分效果见表 2 和表 3，其中表 2 是反映从电子病历中抽取出的信息，表 3 反映的是经过数据预处理后的信息。为了更好地说明预处理的效果，本文将多个数据库的部分属性放在同一张表中。表 2 和表 3 中(1)表示凝血 INR、(2)表示随机指血糖、(3)表示血清前清蛋白、(4)表示心梗标志物 MYO。

结束语 本文从大量电子病历的管理、存储、处理的角度考虑，对电子病历资料实施了有效的数据抽取、数据清理、数据变换和数据归约等的预处理，消除了其间的噪声、不完整性、不一致性、相关性等影响数据再处理效果的问题，经预处理后的数据可以直接作为再处理(如数学建模、数据挖掘等)的输入，为实现多种服务功能提供了可能。另外，本文针对确诊为冠心病的相关电子病历所做的预处理研究仅是对病历资料数据处理的一种尝试，如何更准确获取数据并且合理、充分地利用电子病历资料中包含的信息为医疗服务还需要做进一步的研究工作。

维普资讯 http://www.cqvip.com

表 2 未做数据预处理的数据

age	Symp1	Symp2	Symp3	Smk1	SmkT	Drk1	hth	fHis1	fHis2	Pres	Chek1
48			2	1		1	1		2	112/57	1.40(1)
74	2			1		1	1			135/65	8.8 (2)
92	2			1		1	1			181/92	168 (3)
72	2			2	60		1			105/78	36 (4)
51		2		1		2	1			90/65	
51		2		1		1	1			170/98	176 (3)

表 3 预处理后的数据

Age0	Symp0	Smk0	hth	fHis0	Pres0	Chek1
2	1	1	1	1	1	2 (1)
1	2	1	1	2	1	2(2)
1	2	1	1	2	3	3 (3)
1	2	2	1	2	1	3 (4)
2	2	2	1	2	1	1
2	2	1	1	2	3	3 (3)

参 考 文 献

1 万志红，范千云. 电子病历系统[J] 中国医药卫生信息，2002，22(1)：246～250

2 刘安滨. 再谈电子病历[J]. 中国医院管理，2003，23(4)：37～38

3 李少龙，吴一龙，汪建平. 立体化电子病历与询证医学的研究. 询证医学，2003，3(2)：74～77

4 Basili R, Pazienza M T. Lexical Acquisition and Information Extraction [C]. *Information Etraction: A Multidisciplinary Approach to an Emergin Information Technology*, 1997, 1299: 44～72

5 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques [M]. USA: Morgan Kaufmann Publishers, 2001. 70～95

（上接第 125 页）

LLQ 这 3 种代理缓存布局方法所不能容纳的流媒体片段数目，或称为“未被缓存的片段数”。

在一定的存储容量下，不能容纳的流媒体片段越少，说明代理缓存的利用率越高，代理缓存能够直接响应来自用户请求的机会越大，而片段的播放时延也越短。如图 7，8 和图 9 所示，在 S-Narrow、Uniform 和 S-Wide3 种不同的异构环境下，我们的布局方法 LQHR 所容纳的流媒体片段数目与 LLQ(LCQ)相比，最大超额分别达到了 20(76)，29(61)和 26(34)。LLQ 的性能曲线在上述 3 幅图中并不是始终随着磁盘空间的增加而降低的。这是因为：LLQ 这种布局方法中所能够选择的 4 种传送速率之间的差额是呈指数率增长的，当归一化磁盘容量在 0.2 到 0.3 之间变化时，相同的一块存储空间分配给已有的缓存片段来增加其传送速率比分配给一个新的片段更能满足用户对观看质量的请求。

虽然 LQHR 采用了比 LLQ 更细致的搜索步长来搜索合适的传送速率，但是在 S-Narrow、Uniform 和 S-Wide 这 3 种不同的异构环境下，在代理缓存的布局结果中出现的不同的传送速率的平均值分别为 4.0，4.3 和 4.7，并不会为现有的代理缓存系统增加很多的系统开销。

小结 在网络边缘对流媒体的热点片段进行缓存是流媒体分发研究领域的一个热点。异构环境下不同种类的接入用户对流媒体的播放质量具有不同的要求。为了使代理缓存最大限度地满足全体用户对收看质量的请求，本文首先提出了质量命中率的定义来衡量代理缓存对提高用户观看质量的贡献，然后提出了一种基于质量命中率的流媒体代理缓存布局方法。仿真结果表明，该布局方法不仅能使全体用户的平均质量命中率最大，而且有效地利用了代理缓存的磁盘空间，显著降低了片段的播放时延。

后续工作将在此基础上，围绕流媒体代理缓存在无线蜂窝网中的应用展开。

参 考 文 献

1 Costa C, Cunha I, Borges A, et al. Analyzing client interactivity in streaming media. In: Proc. 13th Int World Wide Web Conf., New York, NY, May 2004. 534～543

2 Sen S, Rexford J, Towsley D. Proxy prefix caching for multimedia streams. In: Proc. IEEE INFOCOM, Mar. 1999. 1310～1319

3 Chae Y, Go K, Buddhikot M M, et al. Silo, rainbow and caching token: schemes for scalable tolerant streaming caching. IEEE Journal on Selected Areas in Communications, 2002, 20(7): 1328～1344

4 Chen S, Shen B, Wee S, et al. Adaptive and lazy segmentation based proxy caching for streaming media delivery. In: Proc. of 13th Int Workshop on NOSSDAV, Monterey, CA, 2003. 22～31

5 Wu K, Yu P S, Wolf J L. Segmentation of multimedia streams for proxy caching. IEEE Transactions on Multimedia, 2004, 6(5): 770～780

6 Liu W, Chou C T, Yang Z K, et al. Popularity-wise proxy caching for interactive streaming media. In: Proc. 29th IEEE LCN, 2004. 250～257

7 Rejaie R, Yu H, Handley M, et al. Multimedia proxy caching mechanism for quality adaptive streaming applications in the Internet. In: Proc. IEEE INFOCOM, Mar. 2000. 980～989

8 Kangashariju J, Hartanto F, Reisslein M, et al. Distributing layered encoded video through caches. IEEE Transactions on Computers, June 2002, 51(6): 622～636

9 Tang X, Zhang F, Chanson S T. Streaming media caching algorithms for transcoding proxies. In: Proc. 31st Int Conf. on Parallel Processing, Aug. 2002. 287～295

10 Zine M, Schmitt J, Steinmetz R. Layer-encoded video in scalable adaptive streaming. IEEE Transactions on Multimedia, Feb. 2005, 7(1): 75～84

11 Liu J, Chu X, Xu J. Proxy cache management for fine-grained scalable video streaming. In: IEEE INFOCOM, Mar. 2004. 1490～1500