文章编号：2095-2163(2019)02-0033-05 中图分类号：TP391.41 文献标志码：A

面向药物不良反应发现系统的多源数据融合研究

喻 捷

(东华大学 计算机科学与技术学院，上海 201620)

摘 要：为了支持面向 EHR 的药物不良反应发现系统，本文提出在已有不良反应发现系统的基础上，需要构建以人体系统器官分类的不良反应症状库。本文面向非结构化电子病历的药物不良反应发现系统，通过多特征提取的模式识别方法构筑多轴性表示的 ICD-10 数据集，从而将具有多轴性表示的 MedDRA 与 ICD-10 进行融合。然后在已融合的两大数据集的基础上，通过构建向量空间模型并将已融合数据集用 FAERS 数据集进行聚类，从而形成不良反应症状到疾病术语或医学术语的层级关联。

关键词：数据融合；MedDRA；ICD-10；FAERS；多轴性；层级关联

Multi-source data fusion for adverse drug reaction discovery system

YU Jie

(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 In order to support the discovery system of adverse drug reactions for EHR, this paper proposes to build a system of adverse reaction symptoms based on the classification of human body organs on the basis of existing adverse reaction discovery systems. In this paper, the unfavorable adverse drug reaction discovery system for unstructured electronic patient records is constructed by multi-feature extraction pattern recognition method to build a multi-axis representation of the ICD-10 data set, thereby merging MedDRA with multi-axis representation with ICD-10. Then, on the basis of the fused two large data sets, a vector space model is constructed and the fused data sets are clustered with the FAERS data set, thereby forming a hierarchical association of adverse reaction symptoms to disease terms or medical terms.

【Key words】 data fusion; MedDRA; ICD-10; FAERS; multi-axis; hierarchical correlation

0 引 言

随着医疗大数据领域的迅速发展，很多用于处理药物不良反应(drug adverse reaction，ADR)相关的电子病历的医院管理系统(Hospital Information System，HIS)，自身识别电子病历中所涵盖的药物不良反应术语的能力以及所使用的不良反应术语库的全面性都无法与当前的需求相匹配，尤其是电子健康记录(electronic health record，EHR)的发展，对于症状库的深度与广度提出了更高的要求。同时，如何从发现的不良反应术语，关联到发现其所对应的疾病术语或更高层级的医学术语，成为了学界亟待解决的问题。本文所研究的课题就是面向已有的药物不良反应发现与呈报系统，对多源异构数据源进行数据融合。区别于以往药物不良反应发现与呈报系统所使用到的单一的不良反应术语集，研究中融合了多个具有代表性的医学术语集，形成了层级的、多轴性的、更为全面的、按系统器官分类的症状库。本文的研究内容强化了药物不良反应发现与呈报系统的设计性能，使其具有从电子病历涵盖的不良反应信息中得到对应的疾病术语及医学术语表示的能力，并可通过筛选得到不良反应所涉及的系统器官类别。

多源异构是大数据的基本特征之一，多源数据融合成为了大数据分析处理的关键环节，多源数据融合也成为大数据领域重要的研究主题与热点方向[1]。本文所涉及的多源异构数据融合通过对相同领域但不同结构的数据集的融合，提高数据集的完备性，并进一步挖掘数据的潜在价值。

1 多源异构数据融合

数据融合按照一定准则综合分析、处理来自多个数据源的信息，从而获得比其各个组成部分都更为充分、准确的信息，在全面信息的基础上进行相应决策与估计，进而得出更为精确、可靠的结论[2]。

数据融合算法是数据融合的核心部分。目前，多源数据融合领域广泛运用的算法有基于 D-S 理论[3]、模糊集理论[4]、主题图[5]和语义规则[6]的数据融合算法。

本文所研究的面向药物不良反应发现与呈报系统的多源异构数据融合，融合的数据源分别为 FAERS (FDA Adverse Event Reporting System,

作者简介：喻 捷(1994-)，男，硕士研究生，主要研究方向：数据挖掘、数据分析、文本分析。

收稿日期：2018-11-08

FAERS)的数据集、国际疾病分类(International Classification of Diseases, ICD)以及医学用语词典(MedDRA)。融合的目的在于构建多轴性的医学术语与疾病术语集合,并实现从不良反应术语到医学术语或疾病术语的一个一对多的层级性映射关联。这里将对此展开探讨分述如下。

1.1 多源数据集及其特征

FAERS 数据集来源于美国食品药品监督管理局(Food and Drug Administration)的药物不良反应报告系统,数据集包含的是用户提交到 FDA 的药物不良反应报告系统中的不良反应信息和用药错误信息。这个数据库是用来支持 FDA 的药物和生物制品安全监测系统的。本文所用到的 FAERS 数据集由 FAERS 数据库中的数据去重筛选后翻译得到,涵盖不良反应术语 8 000 条左右。

国际疾病分类是依据疾病的某些特征,按照规则将疾病分门别类,并用编码的方法来表示的系统。全世界通用的是第 10 次修订本《疾病和有关健康问题的国际统计分类》,称为 ICD-10。ICD-10 收入了疾病记录近 26 000 多条,主要包括 ICD-10 编码、手术码、疾病名称、拼音码。

医学用语词典(MedDRA)是由人用药物注册技术要求国际协调会(ICH)主办开发、在医药事务管理活动中使用的一套医学标准术语[7]。该术语集可广泛见于各种医学数据的编码、检索和分析,如不良事件、适用症与临床检查等场景。以本文用到的 MedDRA 21.0 版本为例,收录了疾病记录等 118 000 条左右,从上到下主要包括系统器官分类(System organ class, SOC)、位组语(High level group term, HLGT)、高位语(High level term, HLT)、首选语(Preferred term, PT)以及低位语(Lowest level term, LLT)这五层结构。

本文通过多源数据融合研发建立的以人体系统及器官分类的不良反应症状库,包含有 2 层。第一层为医学术语与疾病术语,第二层为不良反应术语。最终能够实现通过提取的不良反应信息,匹配不良反应信息所对应的医学术语与疾病术语,并得到涉及的人体系统及器官类。

1.2 多轴性融合

在医学上,医学术语或症状术语很多都涉及人体的多个系统或器官,比如缺铁性贫血就涉及到血液循环系统与内分泌系统。这种术语与系统或器官的一对多表示,更适合医学研究的需要。因此,本次研究引入带有多轴性的 MedDRA 数据集,参见表 1,即 MedDRA 的低位语与系统器官分类存在一对多的关系,MedDRA 中的医学术语对应一个或多个系统器官分类[8]。

表 1 缺铁性贫血在 MedDRA 中的多轴性表示

Tab. 1 Multiaxial expression of iron deficiency anemia in MedDRA

系统器官分类	位组语	高位语	首选语	低位语
血液及淋巴系统疾病	非溶血性贫血及骨髓抑制	各种贫血症	缺铁性贫血	缺铁性贫血
代谢及营养类疾病	铁及微量金属代谢异常	各种铁缺乏	缺铁性贫血	缺铁性贫血

此外,标准的不同,中西医学的不同等都有可能导致同一种疾病有多个不同的疾病名称。如西医疾病学中的蛛网膜下腔出血与中医中的脑中风表述的就是同一症状。为了丰富数据源中的疾病术语,避免出现同一种疾病的不同疾病名称的缺失,研究中又引入了 ICD-10 数据集作为 MedDRA 数据集的补充,但是 ICD-10 却不具有多轴性。对于 ICD-10 数据集,研究通过构建和 MedDRA 相同的多轴性表达方式,达到将 MedDRA 与 ICD-10 进行数据融合的目的。

1.3 层级性融合

本文所涉及的多源数据融合致力于构建从不良反应术语到医学术语或疾病术语的一个一对多的层级性映射关联。通过构建层级性关联,每一条不良反应术语都可以在疾病术语集或医学用语集中找到对应的一种或多种表示。在层级性关联中,每一条不良反应术语所涉及到的系统器官类别也可以表示为由其所对应的疾病术语或医学用语所涉及的系统器官类别。

如 FAERS 中提取的不良反应信息为血压升高,通过层级性融合,能够匹配出高血压心脏病、高血压性脑病等疾病,也能够匹配出撤退性高血压、反弹性高血压等医学用语,并且得到可能涉及的人体系统器官。设计运行结果详见表 2。

表 2 MedDRA、ICD-10 及 FAERS 的比较

Tab. 2 Comparison of MedDRA, ICD-10 and FAERS

名称	MedDRA	ICD-10	FAERS
结构	5 层结构:SOC, HLGT, HLT, PT, LLT	ICD-10 编码,手术码,疾病名称,拼音码	单个症状或不良反应术语
数量	118 000 余条	26 000 余条	8 000 余条
有无多轴性	有	无	无
有无系统器官层次上的分类	有	有	无
需要进行的融合方式	无	多轴性	层级性

2　相关工作

本文着眼于已有的不良反应发现系统，通过进行多轴性、层级性的多源数据融合，在原有的提取电子病历中的不良反应的基础上，通过提取的不良反应术语，找到对应的疾病术语及医学用语表示。同时，根据层级结构分析得到受不良反应影响的系统器官。在方法上，主要用到的是基于疾病术语特征提取的模式识别以及向量空间模型(Vector Space Model)。

本文所涉及的不良反应发现系统设计是在邓剑雄等人[9]提出的基于 HIS 的药品不良反应快速上报与智能搜索系统的基础上，融入了多源数据融合带来的不良反应到疾病术语与医学用语的层级性映射关联，实现对 HIS 系统的不良反应相关疾病报告功能。研究得到的面向药物不良反应发现与上报系统的系统结构如图 1 所示。

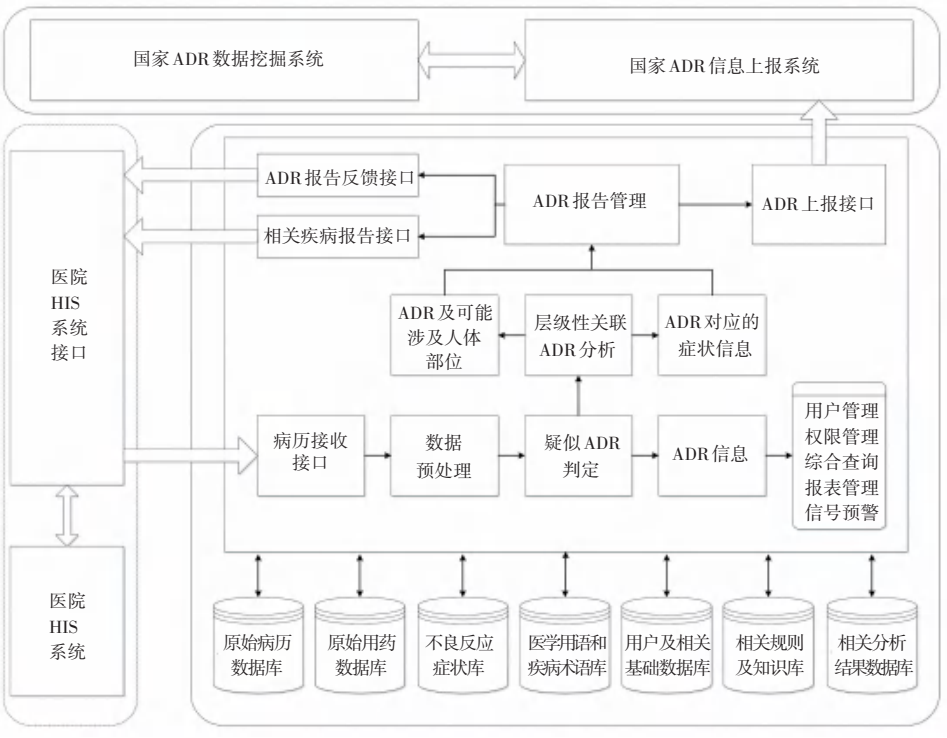

图 1　面向药物不良反应发现与上报系统的系统架构

Fig. 1　System architecture for adverse drug reaction discovery and reporting system

3　问题定义

首先，对本文研究问题进行定义，用 $L_1 = \{M_1, M_2, M_3, \ldots, M_n\}$ 来表示医学用语数据集，用 $L_2 = \{S_1, S_2, S_3, \ldots, S_n\}$ 来表示疾病术语数据集，用 $L_3 = \{A_1, A_2, A_3, \ldots, A_n\}$ 来表示不良反应术语集，$Label = \{lab_1, lab_2, lab_3, \ldots, lab_n\}$ 表示系统或器官类别。那么医学用语集和疾病术语集中的每一条记录都可以表示为 $< symptom, label >$ 的形式。对于 $M_i \in L_1, M_i.label$ 表示所属的系统器官类别标签，$M_i.symptom$ 表示医学用语，同理也可以表示 L_2。此外，L_1 所具有的多轴性可以表示为对于 $M_i, M_j \in L_1, i \neq j$，存在 $M_i.symptom = M_j.symptom$ 且 $M_i.label \neq M_j.label$。对于 $S_i \in L_2$，研究尝试通过实体链接的方式来继承 L_1 的多轴性。

本文通过对数据的预处理，构建了 $Label$ 以及 L_1, L_2, L_3，其中对于 $A_i \in L_3, A_i.Label$ 为空且为集合类型，$A_i.symptom$ 为不良反应术语，同时 A_i 还包含 $A_i.set$，用来存放层级性映射关联中满足条件的所有医学用语或疾病术语，并将对应的系统器官标签存入 $Label$ 集合中。

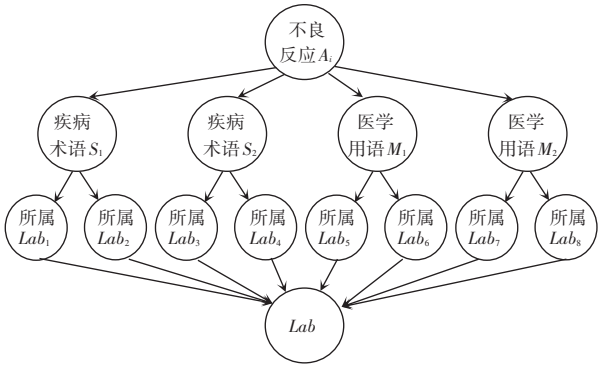

图 2　不良反应在层级性映射关联中的形式及 *Lab* 获取

Fig. 2　The form of adverse reactions in hierarchical mapping correlation and *Lab* acquisition

4 基于多轴性的分类设计

本文的多轴性融合是为 ICD-10 引入多轴性表示,从而与 MedDRA 融合,其实质是基于 ICD-10 疾病术语中涵盖的医学特征词语的模式识别。与常规的分类问题所不同的是,通常分类的特征选择都是从原始特征中挑选出最有代表性、分类性能好的特征,而对 ICD-10 引入多轴性需要提取多个分类明显的特征,对多个特征分别进行分类决策,最终可得对于 ICD-10 的每条疾病术语都属于一个或多个系统器官类别的运行结果,具体即如图 3 所示。

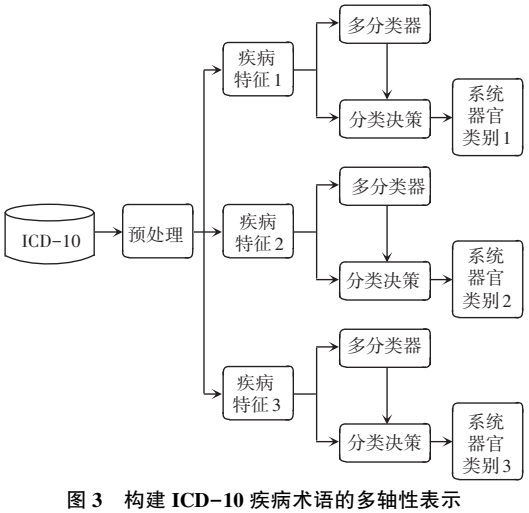

图 3 构建 ICD-10 疾病术语的多轴性表示

Fig. 3 Construction of multiaxial representation of ICD-10 disease terms

在特征的选择上,常见的医学术语特征有发病部位、病因、病理等。如鼻窦恶性肿瘤,按发病部位属于耳鼻喉,按病理属于恶性肿瘤。

5 基于层级映射的融合设计

本文的层级融合是构建以 FAERS 数据集为底层,多轴性融合后的 MedDRA 与 ICD-10 数据集为顶层的 2 层结构。研究通过构建词向量空间,并以 FAERS 数据集为对象进行聚类,来完成层级性映射关联。这里,设计给出的症状库层级性映射关联模型则如图 4 所示。

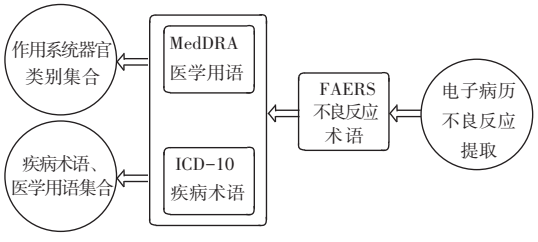

图 4 症状库层级性映射关联模型

Fig. 4 Hierarchical mapping correlation model of symptom library

5.1 词向量空间模型

层级性关联融合所涉及到的词典库包含了 FAERS 数据集的所有不良反应术语。因此,通过构建词向量空间模型,FAERS 数据集的不良反应术语都可以用 MedDRA 或 ICD-10 中的医学用语或疾病术语的夹角余弦值表示。

向量空间模型把对文本内容的处理简化为向量空间中的向量运算,并且是以空间上的相似度表达语义的相似度。对于 MedDRA、ICD-10 及 FAERS 数据集,研究拟将构建词向量空间模型,再通过计算夹角余弦值来评估相似度。

5.2 Skip-gram 模型

研究选择了 Skip-gram 模型作为生成数据源对应的词向量的模型。Skip-gram 是一种根据当前词语来预测上下文的词语模型。相对于根据上下文的词语预测当前词语出现的概率的模型,Skip-gram 在理解低频词上有比较好的效果,这点在本文的课题研究中显得尤为重要,很多在电子病历中频繁出现的不良反应术语在数据源中却属于低频词。Skip-gram 的输入层是一个词向量,投影层直接将输入层的词向量传递给输出层,整体的研发设计架构则如图 5 所示。

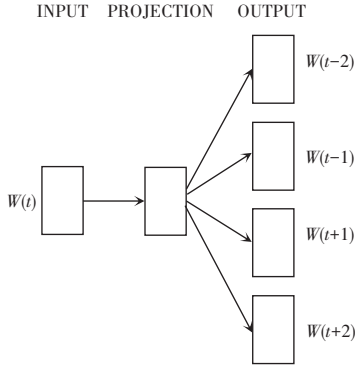

图 5 Skip-gram 模型

Fig. 5 Skip-gram model

6 结束语

本文是针对为医院提供的药物不良反应发现与呈报系统,通过对系统的症状库进行多源异构数据融合,形成了层级的按系统器官分类的症状库,并且在对于症状库的描述上更为全面,能够反映出症状库中的术语所涉及的多个系统器官类。本文虽然采用了神经网络语言模型中适宜于处理低频词的 Skip-gram 模型,但在低频词的层级性关联上仍然有待于提高。

(下转第 41 页)