基于文档集的生物信息挖掘模型研究

孙红敏,姜楠楠,李 想
SUN Hongmin, JIANG Nannan, LI Xiang

东北农业大学 电信与信息学院,哈尔滨 150030
School of Electrical and Information, Northeast Agricultural University, Harbin 150030, China

SUN Hongmin, JIANG Nannan, LI Xiang. Research on biological information mining model based on document set. Computer Engineering and Applications, 2016, 52(24):102-106.

Abstract: As the quantity of literature increases dramatically, to get the information manully can't adapt to the speed of added literature. This paper proposes a new model of biological data mining, utilizing some tools of open source such as Stanford Parser, using some approaches such as natural language processing and statistics. It also analyzes its crucial technique. During the process to test the SBQTL(Soybean Quantitative Trait Loci) using this model, the precision and recall rate are 93.0% and 78.4% respectively. During the process to test the PubMed, the precision and recall rate are 94.3% and 80.0% respectively. So the problem that the researchers who are engaged in biomedicine can find the information they need from large quantity of literature quickly and efficiently is solved, and biologists can find closet information in biomedicine and verificate the newest science discovery. Thus, people can better understand the phenomenon of biomedicine.
Key words: text mining; Stanford Parser; text preprocessing; dependencies; information extraction

摘 要:针对生物医学文献的数量急剧增长,人工从文献中获取所需要的信息已不能适应生物医学文献数量迅速生长的需要。利用Stanford Parser等开源工具,采用自然语言处理技术、统计学等多种方法,提出了一种新型的生物信息挖掘模型,并对其关键技术进行分析。该模型在对全文文本SBQTL(Soybean Quantitative Trait Loci)测试中父母本信息提取的准确率和召回率分别为93.0%和78.4%;在对PubMed测试中,准确率和召回率分别为94.3%和80.0%。解决了生物医学研究者从海量文献中更有效、快速地找到所需信息的问题,以便生物学家发现隐藏的生物医学知识并验证得到新的科学发现,从而使人们对生物医学现象的认识得到了提高。
关键词:文本挖掘;Stanford Parser;文本预处理;依存关系;信息抽取
文献标志码:A **中图分类号:**TP311 doi:10.3778/j.issn.1002-8331.1505-0282

1 引言

近年来随着高通量生物技术的发展,生物医学的实验手段、研究方法均得到了巨大的改进,导致生物医学领域内实验数据的急速增长,数学、化学、统计学和计算机科学等领域的专家对数据的存储、传输、处理、理解与应用等一系列问题产生浓厚的兴趣并通过实验取得了大量成果。生物医学文献作为学术交流和成果展示的主要方式之一,其数目增长速度远远超过了其他学科领域[1]。例如,US National Library of Medicine提供的在线生物医学文献数据库PubMed是现代生物医学高价值文献存储和研究发展的代表资源。自1966年以来到2014年,已收录70多个国家,40多个语种的生物医学文献3 500万篇以上,成为医学和生物学科学研究的重要知识依据。

PubMed生物医学文献数目逐年增长示意图如图1所示。

面对如此大规模的、快速增长的科学文献数据,即便是领域内的专家也无法依赖手工方式快速地从中获取感兴趣的信息,做到完全掌握其领域研究的现状和未来发展的趋势。因此,采用文本挖掘技术从这座数据宝库中高效提取生物医学信息的需求变得非常迫切[2]。分析海量的生物医学数据最重要的是如何才能有效地利用这些文本中所蕴含的生物医学信息。通常采用在PubMed中或者互联网上对关键词进行检索,但是这只能找到与用户需求相关的文件列表,而不能从文本集合

作者简介:孙红敏(1971—),女,教授,硕士生导师,研究方向为农业信息技术,E-mail:sunhongmin111@126.com;姜楠楠(1989—),女,硕士研究生,研究方向为生物信息技术;李想(1992—),女,硕士研究生,研究方向为信息抽取。
收稿日期:2015-06-01 **修回日期:**2015-09-08 **文章编号:**1002-8331(2016)24-0102-05
CNKI**网络优先出版:**2015-09-29, http://www.cnki.net/kcms/detail/11.2127.TP.20150929.1107.040.html

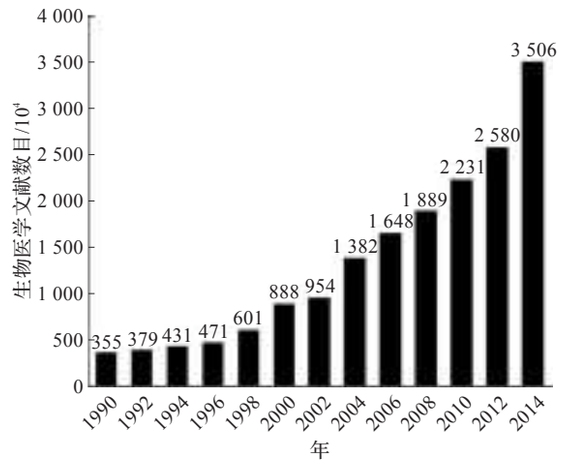

图1　PubMed生物医学文献数目逐年增长示意图

中直接获取用户感兴趣的信息。因此，从大规模生物医学文献中提供自动获取相关知识的有效工具具有十分重要的意义[3]。

国外对于文本挖掘的研究开展较早，在医学领域应用的成功例子是Swanson和Smalheise的工作，通过半自动地分析生物医学文献的题名，提出了一个完全创新、有价值的医学猜测："人体镁元素的缺乏可能导致偏头痛。"由于医学领域的sublanguage(次语言)效应及医学领域文献资源的重要作用，现今多项重要的文本挖掘项目都面向医学，其中两个著名的项目分别是：加州伯克利分校信息管理和系统系的Hearst M.A.主持的LIDNI Project和纽约大学的Linguistic String Project 70年代后的主要研究方向MLP(Medical Language Project)。其中，MLP把叙述性的临床医学文档转换成结构化的语义表示，在此基础上可进行各种有意义的医学归纳知识发现和研究[4]。

国内正式引入文本挖掘的概念并开展文本挖掘研究是从最近几年才开始的[5]。目前国内在生物医学文本挖掘领域的研究相对比较少，主要有哈尔滨工业大学和清华大学，均取得了一定成果。例如：哈工大研究人员主要致力于生物医学命名实体识别和关系的识别的研究，他们在综合多种统计学习方法进行多分类器融合的研究上取得了一定的成果，进一步提高了生物医学命名实体识别的精确率和召回率。在关系识别的研究上主要应用基于特征的机器学习方法并取得了一定的成果[3]。

本文针对生物医学文献，采用自然语言处理技术，设计并实现了一个基于文档集的生物信息挖掘模型，以满足计算机辅助信息获取的需求。

2　文本挖掘的关键技术分析

文本挖掘主要是从大量的、无结构的文档集中发现潜在的、可能的数据模式、内在联系、发展趋势等，抽取有用、新颖、可理解的、散布在文本文件中用户感兴趣的知识，并且利用这些知识更好地组织信息的过程[6]。

2.1　信息抽取

信息抽取是把文本中包含的信息进行结构化处理，变成表格一样的组织形式，便于利用数据库来存储，输入信息抽取系统的是原始文本，输出的是固定格式的信息点[7]。抽取出各种各样文档中的信息点，然后以统一的形式集成在一起[7]。信息以统一的形式集成在一起的好处是便于检查和比较以及对数据作自动化处理[8]。

信息抽取技术并不试图全面了解整篇文档，只是对文档中包含的部分相关信息进行分析。至于是哪些相关的信息，那将由系统设计时定下的领域范围而定[9]。本文主要抽取生物医学文献中的父母本信息。

2.2　基于Stanford Parser的语法分析和依存关系

Stanford Parser(斯坦福句法分析工具)是一款由Java实现的开源句法解析工具，主要基于优化的基于概率规则集和词汇化依存句法分析方法[10]，是一个词汇化的概率上下文无关语法分析器，同时也使用了依存分析。根据不同的语法可以输出不同的分析结果[11]。这些信息将为实体间的关系识别提供重要的参考信息[10]。以"Significant variation for seed yield was observed among the RILs in both the Chinese and Canadian environment"为例，经过Stanford Parser解析的语法树和依存句法分析结果分别如图2和表1所示。

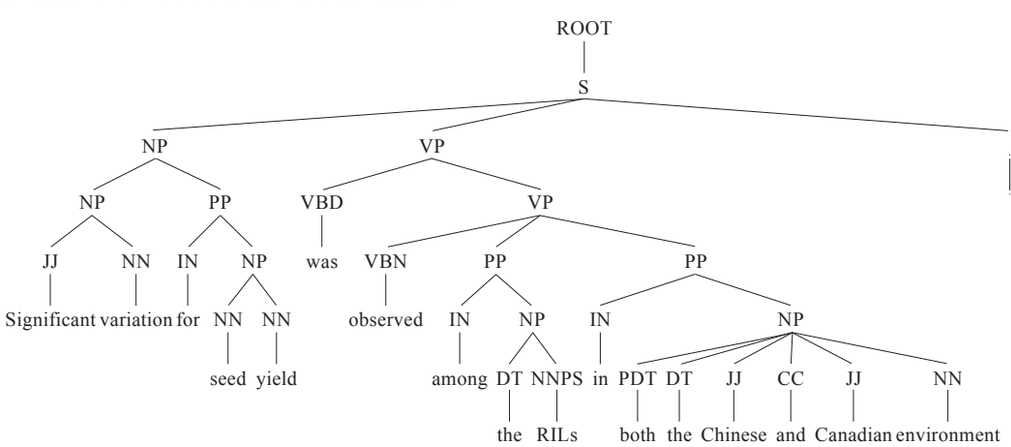

图2　Parser解析生成的语法分析树

表1　利用Parser解析句子的依存句法分析结果及说明实例

例句	依存句法分析结果	结果说明
Significant variation for seed yield was Observed among the RILs in both the Chinese and Canadian Environment	amod(variation-2,Significant-1)	表明variation中包含了Significant
	nsubj(observed-7,variation-2)	表明observed与variation的主谓关系
	nn(yield-5,seed-4)	表明seed yield为复合名词词组
	prep_for(variation-2,yield-5)	表明介词关系
	auxpass(observed-7,was-6)	表示助词关系
	det(RILs-10,the-9)	表示限定词
	prep_among(observed-7,RILs-10)	表明介词关系
	preconj(Chinese-14,both-12)	表示连词关系
	det(Chinese-14,the-13)	表示限定词
	prep_in(observed-7,Chinese-14)	表明介词关系
	amod(environment-17,Canadian-16)	表明environment中包含了Canadian
	prep_in(observed-7,environment-17)	表明介词关系
	conj_and(Chinese-14,environment-17)	表明Chinese和environment的并列关系

表1中间的依存句法分析结果中,括号内是句中的实例,括号前的amod、auxpass、nn等关键词标识了具体的依存关系,该图清晰地展示了句中的主、谓、宾模块[10]。而图2也清晰地展示了一个完整的复杂句中简单句的组成[10]。

3　基于文档集的生物信息挖掘模型研究

传统数据挖掘的对象是数据库、数据仓库等,其数据主要是结构化的,对于文本这类非结构化的数据而言,其挖掘效率比较低,挖掘结果也不尽人意。而由人工将文本这类非结构化数据转换成结构化数据的工作量是非常大的,而且面对海量的互联网信息,这种方式显然是不可取的。随着信息抽取技术的发展,其技术越来越成熟,其手段日渐多元化,也逐步开始应用于许多传统领域[8]。提出的基于文档集的生物信息挖掘模型(如图3所示)就是这方面的探索之一。本模型充分利用了文本挖掘技术的优势,有利于提高文本知识挖掘的准确率和召回率,真正得到用户想要的信息。

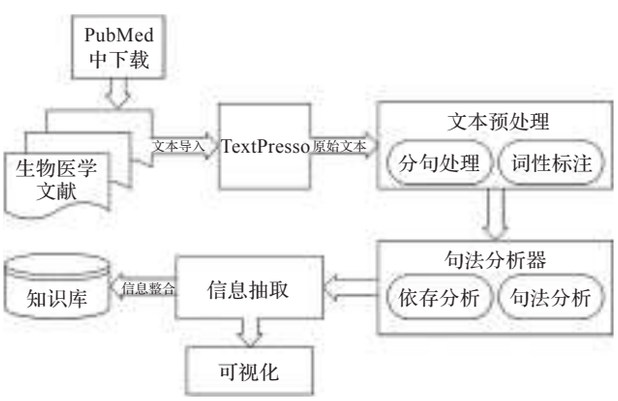

图3　基于文档集的生物信息挖掘模型

基于文档集的生物信息挖掘过程大致包括以下几个步聚:

(1)在网上下载预处理的文献。在网页上包含着大量有用的文本信息,首先要处理所有相关文档的自动检索,并且同时保证不相关的信息尽可能得少,有用信息的揭示使用网页爬行器来完成[8]。

(2)文本预处理。这是该模型研究中比较重要的一步。输入文本后,术语注解器会借助术语表用标准形式来识别这些术语[8]。得到内嵌在文本文档中的标准词汇后使用XML标记。然后定义一个规则,用这个规则进行搜索关键词,并在搜索到的关键词的位置进行标注。

(3)句法分析。利用斯坦福句法分析器(Stanford Parser)将分词及词性标注器的输出作为输入,为输入的语句生成其句法分析树及依存关系树[12]。为后续信息抽取做准备。

(4)信息抽取。根据句法分析树及依存关系对文本进行信息抽取。其主要任务是关联规则、分类规则、聚类规则等挖掘。

(5)结果显示。不同类型的知识(规则)可以采用不同的结果表示方式,比如:分类规则可以采用分类树,关联规则可以采用"蕴含—>"等形式表示。还可以将信息整合,形成知识库,最终进行可视化。

4　研究的方法与实现

信息抽取是从文本(Document)中抽取用户感兴趣的信息,并形成结构化(Structured)的数据[13]。本文父母本的信息抽取是综合运用自然语言处理、统计学等多种方法,结合生物医学研究的相关成果,从自然语言描述的生物医学文献中自动地抽取父母本信息。脚本由两个大部分组成,一个是句子的预处理过程,另一个是对句子进行分析进行父母本提取的过程。由两部分组成的系统流程如图4所示。

其基本分为以下三个过程:

(1)接收文档数据

在PubMed(文献数据库)中下载所需要的文献,下载后的文件为PDF格式,然后将其转换为txt格式。

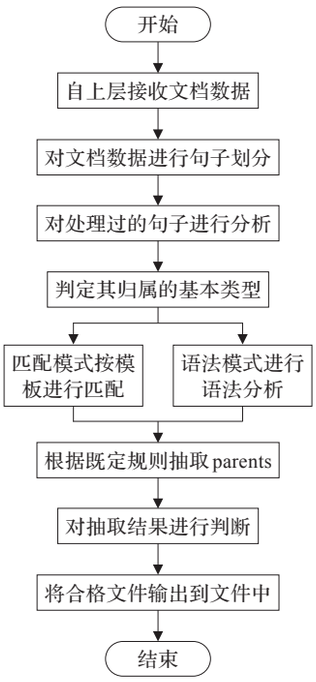

图4　信息提取的流程图

(2)文本预处理

对所下载的文献进行切割,切割成摘要、标题、作者、日期、出版社五部分,然后将所有数据导入到文献挖掘工具Textpresso上。然后定义一个规则,用这个规则进行搜索关键词,然后将搜索到的关键词的位置进行标注。

(3)提取的规则

提取的规则总共分为两个大的部分:语法提取和匹配提取。

语法提取的大致意思就是根据parser输出的结果的依赖关系来制定具体的提取方案。这种提取的结果精度很高,基本不会出现错误的现象。但是相对的这种提取方法对输入数据的要求过高,首先单句的长度不能高于60(这是不可调节因素),在这之后还要求转化出的数据不能出现错误并且作者必须按照语法来写。

①语法提取

语法提取针对那些没有特殊符号的且语法严谨的句子很有效。如表2所示。

②匹配提取

匹配提取的核心就是制作合理有效的模板。如表3所示。

5　实验结果与讨论

5.1　实验数据准备

(1)PubMed测试语料库

在PubMed中下载的831篇文档中抽取500篇进行分析。500篇数据中共计125条父母本信息,未提出的共18条,总计提出107条,其中有6条提取错误。为了减少错误提取率对脚本进行了部分调整,总共运行了三次,平均的正确率为94.3%,平均的回收率为80%。

(2)soybean QTL全文测试语料库

在SBQTL全文测试语料库中抽取了关于"soybean+QTL"MEDLINE文献200篇,200篇数据中共计51条父母本信息,未提出的共7条,总计提出43条,其中有3条提取错误。得到的准确率和召回率分别为93%和78.4%。

表2　语法提取

提取方法	模板	举例及说明
语法提取	prep_between/of(cross,*)	这样的模板可算是非常典型的 the cross between A and B或者是the cross of A and B这种类型的比较常见但是会出现一些变式
	prep_from(derived,*)	比较常见的就是parents derived from A and B
	prep_of(allele|s,*)	比较常见的就是parents derived from A with B
	prep_of(mapping,*)	mapping这个词出现的几率不是很高,为了提取率增高,也将它加入了模板,mapping of A and B
	prep_of(hybridization,*)	杂交这个单词可以更好地体现出父母本,hybridization of A and B

表3　匹配提取

提取方法	模板	举例及说明
匹配提取	(\w+)\?×\(\w+)	×不是基本字符parser不能识别,由于其特殊性所以决定把它作为模板,形式为A×B
	(\w+?-?\w+?-?\w+)\·\(\w+-?\w+?-?\w+?)	第二种不规则写法,既把×改写成了·,一样的道理parser不能识别,所以把其作为模板,形式为A·B
	'(.+)'\·\(\w+)	全角的引号,同理parser无法识别
	'(.+)'\and\'(.+)'	全角引号而引起parser分析错误,特殊处理
	(\w+\?\d?)\£\(\w+\?\d?)	文本在转换时出错,以£作为标记提取两边部分,形式为A£B
	(\w+?-?\w+?-?\w+)\?×\?(\w+\\d+?\ )	(\w+)\?×\(\w+)的变形,为了完整提取HS-100这种类型的数据
	(\w+)\(\w+-?\w+?-?\w+)\9\(\w+-?\w+?-?\w+?)\(\w+-?\w+?-?\w+?)?和(\w+)\(\w+.\w+)\9\(\w+-?\w+?-?\w+?)\?(\w+)?	前一个是通式,而后一个是针对性比较强的提取模板,总体上针对的是 A9B这种形式的具体处理和筛选过程

5.2 评测指标

信息抽取技术评价指标主要使用三个常用的统计学指标:召回率(*Recall*)、准确率(*Precision*)以及召回率和准确率的加权几何平均值(*F*-value)三方面进行测量[14]。召回率相当于是测量被正确抽取的父母本信息的比例,准确率是抽出的信息中正确的父母本信息所占的比率,而 *F* 值用于综合评价识别方法的有效性。计算公式如下:

$$召回率\ (R)=TA/AN$$

$$准确率\ (P)=TA/(TA+FA)$$

$$F\text{-value}\ (F)=2PR/(P+R)$$

其中,*TA* 代表从语料库中抽取到的正确父母本信息数目;*FA* 代表从语料库中抽取到的错误父母本信息数目;*AN* 代表语料库中存在的父母本信息数目。

5.3 实验结果

实验结果如表4所示。

表4 实验结果

	测试集	准确率/%	召回率/%	*F* 值
文献[6]	标准语料	75.4	82.1	0.78
文献[15]	标准语料	85.7	66.7	0.75
文献[16]	GENIA 语料	72.9	71.1	0.72
本文	SBQTL 语料	93.0	78.4	0.85
	PubMed 语料	94.3	80.0	0.87

文献[6]在百度、搜狐等主流网站上选取30篇文档作为语料库进行预处理后,使用词条频度方法提取特征词,利用模型和改进算法在语料上获得了75.4%的准确率和82.1%的召回率。AbGene 系统是比较成功的生物医学命名实体识别系统之一[15],曾被多个研究者作为命名实体识别组件用于关系抽取研究当中。该系统使用7 000个手工标注命名实体类别的句子作为贝叶斯模型的训练语料,并采用手工统计规则作为后处理,同时使用命名实体所在的上下文来帮助校正识别错误。该系统达到了85.7%的准确率和66.7%的召回率。Tzong-han Tsai等[16]使用条件随机域模型结合丰富的特征集合和后处理过程在BIONLP2004测试语料上获得了69.1%的准确率和71.3%的召回率。

本文提出基于信息抽取的文本挖掘模型在对全文文本SBQTL测试中的父母本信息提取,准确率、召回率和 *F*-value 分别为93.0%、78.4%和0.85;在对PubMed测试中,准确率、召回率和 *F*-value 分别为94.3%、80.0%和0.87。

5.4 实验结果讨论

针对于召回率相对较低的问题,仔细分析了原因,文献中部分未识别的父母本如表5所示。发现其中占比例最高的未提取的信息中混杂了特殊字符的表达。例如<(MIN) and BLT531>,<from a cross between Essex×Forrest>,<derived from a cross between G.max ssp.max and ssp.soja Values>。除此之外,还有一些含有数字的信息未能提取出来,诸如<derived from N87-984-16·TN93-99 P1 N87-984-16,P2 TN93-99>,<derived from 'S08-80'·PI464925B>这样的信息也未能识别。

表5 文献中未识别的父母本及其原因

父母本	未识别原因
G.soja accession	混杂了特殊字符
Essex×Forrest	
"Resistant"and"Susceptible"	
BLT531 is Minsoy(MIN)	
N87-984-16·TN93-99 P1 N87-984-16,P2 TN93-99	含有数字的信息
(Williams 82)	
(IT182932)	
'S08-80'	
(C.arietinumICC4958#C.reticulatum PI489777)	
Williams and Williams 82	
PI4070305	
PI4070305	
IA2008	
QBTL29 and QBTL531	

这些问题为今后的研究工作指明了方向,考虑在该模型中加入语法信息形成的规则进行数据噪声过滤[17]。相信信息提取的召回率和准确率会有进一步的提高。此外,本文仅仅提取了父母本信息的情况,在今后的工作中,会对所有有用信息的提取提出相应办法。

6 结束语

提取有用的生物文献中的信息是生物医学文献文本挖掘的关键环节。本文提出了一种新型的生物信息挖掘模型,它实现简单,不需要训练数据。脚本运行在安装python的环境中,理论上对操作系统的要求不高。脚本中未使用任何操作系统的API,所以跨平台能力可以保证。该模型在对全文文本SBQTL测试中的父母本信息提取,准确率、召回率和 *F*-value 分别为93.0%、78.4%和0.85;在对PubMed语料库测试中,准确率、召回率和 *F*-value 分别为94.3%、80.0%和0.87。在分析未提取的信息后,提出将文本预处理以及语法规则等自然语言处理技术融入该模型作为未来工作的研究方向。

参考文献:

[1] 黄娟.基于文本挖掘技术的蛋白质相互作用预测方法研究[D].长沙:中南大学出版社,2009.

[2] 虞欢欢.基于机器学习的蛋白质相互作用关系抽取的研究[D].苏州:苏州大学出版社,2010.

[3] 王浩畅,赵铁军.生物医学文本挖掘技术的研究与进展[J].中文信息学报,2008,22(3):89-98.

[4] 周雪忠,吴朝晖.文本知识发现:基于信息抽取的文本挖掘[J].计算机科学,2003,30(1):63-66.

(下转188页)