DOI: 10.3976/ j. issn. 1002 – 4026. 2011. 06. 022

中医药领域信息抽取技术的研究与应用

来建梅,曹慧,马金刚
(山东中医药大学,山东 济南 250355)

摘要:通过概述信息抽取技术在电子病历、中医文献和中医药网络资源三个方面的应用及研究现状,指出该技术有利于发挥中医药的利用价值,促进现代中医药的发展。但目前各项研究停留在理论和实验阶段,需提高系统的实用性,建立自动抽取规则,实现全自动的信息抽取。

关键词:中医药;信息抽取;结构化数据
中图分类号:TP391 **文献标识码**:A **文章编号**:1002-4026(2011)06-0088-04

The research and application of information extraction in the field of traditional Chinese medicine

LAI Jian-mei, CAO Hui, MA Jin-gang
(Institute of Science and Technology, Shandong University of Traditional Chinese Medicine, Jinan 250355, China)

Abstract: This paper summarized the application and current research situation of the technology of information extraction in the area of electronic medical record, medical materials and medical network resourse. Thus we drew the conclusion that the technology can help to make full use of Chinese medicine and can contribute to the development of modern Chinese medicine. However, the current status of this technology is still in the theoretical and experimental period. Therefore, it is quite necessary to improve the practicability of the system, to construct an automatic decimation rule, and to realize the automatic information extraction.

Key words: traditional Chinese medicine; information extraction; structured data

随着中医药领域的迅猛发展,中医药信息呈指数级的爆炸性增长趋势,各种中医药古籍种类、版本繁多,中医药网络资源分布散乱,缺乏统一的数据共享服务系统。多年来,由于人们对中医药认识的不足以及中医药知识普及手段的缺乏,使中医药信息的使用率较低,人们很难从中快速提取自己真正需要的信息,给资源的利用造成了很大的困难,因此检索及发现有价值的中医药信息已成为一项重要的任务。中医药文献对中医药学的发展一直起着积极的推动作用,至今仍是中医药临床、科研信息的重要来源。

当前,中医药信息还没有得到充分的挖掘和利用,利用信息抽取(information extraction)技术对中医药信息进行抽取,从海量信息中迅速找到自己真正需要的信息,将半结构化、非结构化的网页信息转化和提取为结构化的数据,形成统一的数据共享服务平台,将有助于避免资源的流失及研究的重复浪费,有利于发挥中

收稿日期:2011-09-20
基金项目:济南市科技局自主创新计划(200906007)
作者简介:来建梅(1987 –),女,硕士,研究方向为生物医学信息处理与分析。

医药信息的利用价值,促进现代中医药学的发展。

1 信息抽取技术概述

信息抽取的前身是文本理解,最早开始于 20 世纪 60 年代中期,主要是从自然语言文本中获取结构化信息的研究,这被看作是信息抽取技术的研究初始。从 20 世纪 80 年代末开始,信息抽取研究蓬勃开展起来,这主要得益于消息理解系列会议(Message Understanding Confernece,MUC)的召开,MUC 系列会议对信息抽取这一研究方向的确立和发展起到了巨大的推动作用[1]。近年来,信息抽取技术的研究与应用更为活跃,也越来越被广大信息处理研究者所青睐,主要内容是自动抽取文本语料中出现的实体、关系、事件等。

信息抽取是指从一段文本中抽取指定的事件、事实等信息,形成结构化的数据并填入一个数据库中供用户查询使用的过程。按照抽取操作所针对的对象不同,信息抽取可以分为自由文本信息抽取、结构化文本信息抽取、半结构化文本信息抽取和 Web 信息抽取。一般来说,信息抽取系统处理的对象是自然语言文本尤其是非结构化文本,但广义上讲,除了电子文本外,还能处理图像、视频等其他数据。目前大部分的信息抽取主要是针对 Web 上的资源,Web 信息抽取就是将 Web 作为信息源的一类信息抽取,从半结构化的 Web 文档中抽取数据。其核心就是将分散在 Internet 上的半结构化的 HTML 页面中的隐含的信息点抽取出来,并以更为结构化、语义更为清晰地形式表示,为用户在 Web 中查询数据、应用程序直接利用 Web 中的数据提供方便[2]。信息抽取系统的主要功能是从文本中抽取出特定的事实信息,比如,从病人的医疗记录中抽取出症状、诊断记录、检验结果、处方等信息。通常被抽取出来的信息以结构化的形式描述,可以直接存入数据库中,供用户查询以及进一步分析利用[3-6]。

2　信息抽取技术在中医药领域的应用

在中医药领域中,数据、信息和知识经常以文本的形式存在于科学文献、技术、管理报告以及病人的病历中,研究者、临床医生和管理者通过对文本的浏览和研读可以获取他们需要的信息,然而海量的文本却成为他们高效获得所需要信息的障碍。此时,计算机技术在中医药领域最大的效用就体现出来了,它可以将所需要的信息迅速抽取出来转化成结构化的数据。

在研究使用过程中,经常需要从病人的医疗记录中抽取出症状、发病时间、诊断记录和辩证结果等;从医书中抽取症状、病因、脉象和方剂等;从方剂中抽取方剂名称、方剂类别、成分、主治、功效、组成、炮制方法、用法用量等;从中药中抽取中药名称、中药类别、功效、主治、产地等。目前,这些用信息抽取技术都已经基本实现。信息抽取技术可以应用到中医药领域的各个方面。

2.1　信息抽取技术在电子病历中的应用

1968 年,Larry Weed 提出了使用面向问题的医学病历可以帮助医护人员快速明确病人所患疾病、针对疾病制定治疗方案并记录治疗结果[7]。随着信息技术的发展,电子病历作为医疗信息化建设的重要内容,在我国已经得到了长足的发展,并逐渐成为一种记录和管理患者信息的非常重要的现代化手段。医疗信息化和电子病历的发展虽然长久以来试图推动病历信息的结构化,但是由于临床信息的复杂性和灵活性,现有的结构化录入技术无法完全满足临床对于病历信息的要求,所以临床医生依然并将继续使用叙述性文本作为主要的形式来记录临床信息,包括临床医生书写的住院志、病程记录、会诊记录、手术记录以及各种医技科室发出的报告等。如果计算机可以准确获取这些重要的临床信息并应用于后续处理,必将提高医疗质量、降低医疗成本[8]。

由于电子病历种类繁多且内容复杂,实现完整病历的信息抽取非常困难。从 20 世纪 90 年代至今,在医学领域广泛使用的系统也有很多。美国纽约大学开展的由 SAGER 领导的项目 Linguistic String Project,开始于 20 世纪 60 年代中期并一直延续到 80 年代,对语言处理和医学语言处理产生了深远的影响。该项目的主要研究目的是建立一个大规模的英语计算语法,与之相关的应用是从医疗领域的 X 射线检查报告和医院的

出院记录中抽取信息格式,这种信息格式实际上就是现在所说的模板。LSP 系统是第一个在医疗领域应用的自然语言处理系统,它建立了 40 种表示病人文档相关临床信息的子类,以及 6 种语义连接子类,包括病人管理信息,除药物治疗以外的治疗信息,药物治疗,检查和检验结果,病人行为和状态[7]。

由哥伦比亚大学的 Carol Friedman 等人设计的 MEDLEE 系统也是一个很成功的医学信息抽取系统,作为临床信息系统(CIS)的一个独立模块在纽约长老会医院使用,它将文本形式的病历报告转换成编码数据以促进乳腺癌研究,有利于病人看护质量的提高[7]。这些项目都可以很好的被应用于电子病历之中,促进了电子病历的发展。信息抽取技术在电子病历中的成功,将克服临床决策支持、临床路径管理等前沿医疗信息发展所面临的诸多瓶颈问题,提升我国医疗信息技术产业的核心竞争力。

2.2 信息抽取技术在医学文献中的应用

中医药文献是医学的主要知识资源,实现文献的数字化是信息时代中医文献研究的必然要求。国内对中医药文献信息抽取研究相对较多,极大地促进了中医药的现代化进程,如从中药复方的临床文献进行复方名称的抽取[9];利用信息抽取技术从 Web 形式的中医药文献资料中抽取结构化中医临床诊疗信息的中医临床诊疗垂直搜索系统 TCMVSE[10]。目前的信息抽取技术在文献方面的研究大都集中于中医临床文献,且仅限于实体信息的抽取。

在长达 5000 年的中医历史中积累了海量的医案文献,而采用人工处理的方法已经不能适应现在的要求,采用计算机技术建立中医医案数据库,能有效整合古今医案信息资源。上海中医药大学伤寒论教研室利用结构化查询语言开发了"历代医案分析统计系统",这套系统能够按医家、朝代、方剂、病名、病症、病机、治法、药物类别进行查询,并经过统计获得相关症状、病机、治法、药物上的特点。2000 年 9 月,北京中医药大学的"北京中大安信科技发展有限公司"和中国科学院科技政策与管理科学研究所的"北京盘拓咨询有限公司"联合开发的"中医药基础数据库系统"为中医药历史文献数据的收集和整理提供了先进的支撑环境[11]。主要实现了中医理论数据库、中药数据库、针灸穴位及处方数据库、中医医案数据库、中医人名和书籍数据库、中医现代文献主题结构解析数据库等 9 个子系统。这些大型中医医案数据库的建立,为中医药文献资源的继承和创新提供了一个共享平台,充分并完善了中医古籍文献数据的挖掘和集成,加速了中医药资源的数字化进程。对提高中医药信息资源的管理水平和公众对中医药的认知度具有积极的促进作用[12]。

在中医药文献的信息抽取中,虽然信息抽取技术在中医药文献分析中有一定的研究,但整体仍处于起步阶段,大部分研究还停留在理论和实验层次。选择合适的抽取算法,将自然语言处理技术和机器学习的最新研究成果应用于大规模的中医药文献的信息抽取中以取代当前的抽取方法是该领域的研究方向。

2.3 信息抽取技术在中医药网络资源中的应用[13-16]

长期以来,我国的中医药网络资源分布散乱,许多数据资源分散在不同领域、不同单位、不同专家手里,不能有效地跨区域、跨机构共享,形成了严重的数据壁垒。针对网络上分布散乱的中医药资源,可以用基于 HTML 结构的信息抽取方法实现对中医药资源的抽取,将其转换成结构化的数据存储到数据库中。在进行信息抽取之前,先把 HTML 文件转换成解析树,这个解析过程反应其层次结构,接着半自动或者自动地生成抽取规则,并把它应用于这棵树上。具体过程见下图 1。

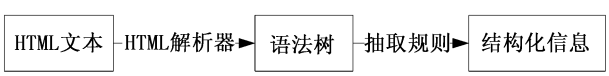

图 1 HTML 文件转换成解析树示意图

Fig. 1 Schematic diagram of a HTML file conversion into a parse tree

同一个中医药网站的所有网页的格式基本相同,不同的只是网页中的具体数据,所以可以通过 HTML 解析器把相应的 HTML 文本解析成语法树,对不同的网站生成不同的抽取规则,存放到模板库中。在进行信息抽取的过程中,自动地到模板库中匹配相应的模板,匹配后抽取其相应的数据资源,形成结构化信息。北京中医药大学在 1989 年完成了"中医方剂信息智能分析支援系统",收集了对 40 余万条方剂信息的解释,可产生 800 余万相关数据,并于 1997 年得到国家教育部博士点

学科专项基金的支持,用 Web_db 技术,将方剂数据库移植到 Oracle 7 for UNIX 平台,在 Internet 网上实验性地实现了方剂数据库的查询和分析处理[11]。浙江大学设计的针对数字图书馆的中医药信息服务的中医药信息服务系统,实现了相关的查询、推荐、方剂展示以及对比阅读等功能[17]。

利用信息抽取技术,整合和抢救离散的中医药数据资源,形成统一的结构化信息,能有助于避免资源的流失和研究的重复浪费,也为挖掘和搜寻分散的数据资源提供了方便。

3　总结与展望

信息抽取技术经过最近十几年的发展,已经成为处理数据资源的重要技术,在中医药领域也得到了较为广泛的应用,国内对信息抽取在中医药临床、方剂和中医候诊等方面的应用进行了大量的研究,极大地促进了中医药领域的现代化进程,利用信息抽取技术能够相对准确、快速的抽取我们需要的中医药信息,并且,目前中医药领域也存在很多已经投入临床使用和正在研究的医学语言处理系统。但要实现一个全自动的信息抽取引擎尚存在一定的困难,都或多或少的需要一些人为参与才能达到一定的满意度。本文详细介绍了信息抽取技术在电子病历、中医文献和中医药网络资源三个方面的应用,及具体研究现状,但均处于起步阶段,大部分停留在理论和实验的层次。因此在未来的发展过程中,我们需要建立更加智能地、自动地信息抽取系统,提高系统的实用性,建立自动抽取规则,实现全自动的信息抽取,更好的促进中医药领域的发展。

参考文献:

[1]李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用,2003,(10):1 - 5.

[2]陈钊,张冬梅. Web 信息抽取技术综述[J]. 计算机应用研究,2010,27(12):4401 - 4405.

[3]APPLET D E, ISRAEL D J. Introduction to Information Extraction Technology. Artificial Intelligence Communications, 1999, 12 (3):161 - 172.

[4]张素香. 信息抽取中关键技术的研究[D]. 北京:北京邮电大学,2007.

[5]刘迁,焦慧, 贾惠波. 信息抽取技术的发展现状及构建方法的研究[J]. 计算机应用研究, 2007, 24(7) 6 - 9.

[6]GAIZAUSKAS R, WILKS Y. Information Extraction: Beyond Document Retrieval[J]. Journal of Documentation, 1998, 54(1):70 - 105.

[7]李莹. 文本病历信息抽取方法研究[D]. 杭州:浙江大学,2009.

[8]李毅,保鹏飞, 薛万国. 中文电子病历的信息抽取研究[J]. 生物医学工程学杂志,2010, 27(4): 757 - 762.

[9]周雪忠. 文本挖掘在中医药中的若干研究[D]. 杭州:浙江大学,2004.

[10]庄力. 中医临床诊疗垂直搜索系统研究[D]. 北京:北京交通大学,2009.

[11]任廷革,刘晓峰,高剑波,等. "中医药基础数据库系统"介绍[J]. 中国中医药信息杂志,2001,8(11):90 - 92.

[12]胡雪琴,周昌乐,李绍滋. 中医医案数据库的数据基础研究[J]. 计算机工程与应用,2008,44(35): 220 - 223.

[13]肖春,周建龙. 生物医学领域中的文本信息抽取技术与系统综述[J]. 计算机应用研究, 2007, 24 (9):1 - 7.

[14]车万翔,刘挺,李生. 实体关系自动抽取[J]. 中文信息学报,2005,19(2):1 - 6.

[15]郑长松,傅彦,佘莉. 基于模板的 Web 信息自动提取方法[J]. 计算机应用研究,2009,26(2): 570 - 572,582.

[16]时达明,林鸿飞,赵晶. 基于模板化的 Blog 信息抽取[J]. 计算机工程与应用,2008,44(9):156 - 162.

[17]杨艳. 一种非结构化数据中医知识抽取与关联的方法[D]. 杭州:浙江大学,2010.