中医自然语言处理研究方法综述

柴 华 路海明 刘清晨

（清华大学信息技术研究院 北京 100084）

〔摘要〕 简要介绍自然语言处理在中医学中的应用，通过对相关文献的研究分析，阐述关联规则挖掘、聚类分析、信息抽取、机器学习等方法的特点与应用方向。总结构建中医知识网络的相关方法，基于构建知识网络的方法提出未来中医自然语言处理研究的新思路。

〔关键词〕 中医；自然语言处理；文本挖掘；知识网络；词向量

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2015.10.013

Overview of Research Methods for Natural Language Processing in Traditional Chinese Medicine *CHAI Hua*, *LU Hai－ming*, *LIU Qing－chen*, *Research Institute of Information Technology*, *Tsinghua University*, *Beijing 100084*, *China*

〔**Abstract**〕 The paper makes a brief introduction to the application of natural language processing in Traditional Chinese Medicine (TCM). Through research and analysis of relevant literatures, it describes the features and application directions of such methods as the association rule mining, clustering analysis, information extraction, machine learning, etc. It also summarizes methods related to the establishment of knowledge networks on TCM and proposes new ideas for future researches of natural language processing in TCM based on the establishment of knowledge networks.

〔**Keywords**〕 Traditional Chinese medicine; Natural language processing; Text mining; Knowledge network; Distributed representation

1 引言

数据挖掘是数据库知识发现（Knowledge Discovery in Databases, KDD）中的一个步骤，一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。近年来医疗数据挖掘发展迅速，然而目前医疗数据结构化处于起步阶段，更多的医疗数据仍然以自然语言文本形式出现，这些医学文本资料中的知识是不同地域、不同时代人们智慧的结晶，展现的是大量、未整理的文献资料以及诊疗记录，而这一点在中医学中尤为突出。

自然人的学习能力有限，因此学者们尝试通过自然语言处理（Natural Language Processing, NLP）辅助完成汇总中医知识的过程，将知识提炼出来，提取其中有用的诊疗信息，最终形成知识本体或者知识网络，从而为后续的各种文本挖掘任务提供标准和便利。NLP 属于人工智能的子领域，其核心目的是使得计算机能够理解和生成人类的自然语言，任务主要包括信息抽取、机器翻译、情感分析、摘要提取等，所用到的技术包括命名体识别、语义消歧、指代消解、词性标注、结构分析等。大量医学文本资料中含有的病史、诊断、治疗方法、药物等名词，给 NLP 的应用提供了可能性。利用 NLP 技术将隐藏在文本中的知识挖掘出来，对医学的发展具

〔修回日期〕 2015－05－28

〔作者简介〕 柴华，在读硕士研究生；路海明，研究员。

有重要意义，目前已有医学和生物学领域的相关研究[1]。同时 20 世纪 80 ~ 90 年代，一些医学本体数据库逐渐建立起来，如一体化医学信息系统、临床医学系统术语等，使得利用 NLP 挖掘医学知识的资料和工具更为充足。

2 中医学中的 NLP 方法

2.1 关联规则挖掘

2.1.1 概述 关联规则是数据挖掘的常用方法，核心在于分析类似“某些事情的发生引出另外一些事件的发生”的规则，包括简单关联、时序关联、数量关联、因果关联等，核心算法是以支持度和置信度作为判断标准，确定是否存在关联关系。著名的关联算法有 Apriori 算法及其改进算法 FP - growth，通过计算出频繁项集来表示规则前件和后件中的事项明显同时出现。

2.1.2 关联规则在中医学中的应用 主要是方剂的关联性挖掘，如任廷革等[2]尝试构建了中药方剂数据库，收集了近 2 000 年来约 10 万个方剂数据，共 100 万余条数据记录，而且给出了从中挖掘关联规则的方法[3]。王大阜[4]使用 Apriori 算法对所收集的方剂数据库进行关联分析，挖掘出了当归 = = > 生地（支持度 7.86%，置信度 78.57%）、白藓皮 = = > 土茯苓（支持度 7.14%，置信度 83.33%）等关联规则，将方剂中常用的搭配药物分析出来，对中药的配方循证起到指导作用。朱立成[5]对 445 例名医医案进行关联分析，挖掘出哮喘医案的病因、病位、证候与四诊信息的关联关系，病因、病位、证候、四诊信息与用药的关联关系，以及中药之间的关联关系。

2.1.3 局限性 关联分析挖掘出来的知识有限，仅仅考虑到了并发的情况，一般局限于某个术语与其他某个或某些术语共现频次较高类似的结果。大部分的应用建立在获取结构化数据的前提条件下，更多展现的是对结构化数据分析的能力。

2.2 聚类分析

2.2.1 概述 中医有阴阳五行学说，人体有五脏六腑之分，均彰显出可分类的特点，聚类分析应用于中医学中应当与中医自身的性质相契合。学者们利用聚类分析方法对中医文本挖掘进行研究，具体为症状分类和药物评价。

2.2.2 对症状的聚类 症状分类的语料多来自中医的诊断手稿，常见于从某一种特殊的疾病入手，利用诊断手稿对症状聚类，得出该疾病的表型特点。麻晓慧[6]利用有关胆道感染、胆石症病案共 739 例，将 92 项临床表型聚类，得到胆病症状的表现分类特点，归纳总结了胆病的症状群。袁世宏等[7]使用聚类分析方法寻找肾虚症状的自然类群，聚类的结果与中医理论的描述基本一致，为中医的科学性提供了很好的佐证。除症状之外，何裕民等[8]使用模糊聚类，得出体质的类型分类（强壮质、虚弱质、失调质）及若干亚型。

2.2.3 药物评价聚类 药物评价方向主要是利用聚类方法将类似性状或相同功效的药物聚在一起，应用中医药理论总结知识。何前锋等[9]对中药按照功效聚类，定义药物之间的相似性，对中药分类整理做出一定的贡献。

2.2.4 局限性 相比于信息抽取，聚类分析偏向整体性质，从宏观的角度对疾病、症状、药物做出分类整理，只能得到概括性的评价，无法挖掘出具体的诊疗方法信息。

2.3 信息抽取

2.3.1 概述 中医文献大都是以自然语言的方式描述的，而且纷繁复杂，医疗记录中蕴含着症状、诊断信息，医书中蕴含方剂、病理信息，药物典籍中蕴含组分、制作方法信息等，如果采用人工方法提取这些信息，耗费的人力、物力是难以估量的。然而，由于中医术语名词都包含在描述语言中，而且文献描述语言简练、逻辑简单，因此可以考虑使用信息抽取算法来自动获取结构化信息。

2.3.2 隐弥科夫模型为主的信息抽取 近年来，隐马尔科夫模型（Hidden Markov Model，HMM）在信息抽取领域中被广泛应用。顾铮等[10]利用 HMM 对中医古籍进行了信息抽取，将症状、病因、脉象和方剂看作模型的 4 种状态，然后利用命名实体识别结合人工标注的方法来从文献中提取相应的名词，最终计算出 HMM 相关参数，达到了信息抽取的目的。庄力[11]以中医临床诊疗数据面向普通公众便捷信息服务为目标，设计并实现了中医临床诊疗垂直搜索系统 TCMVSE，可以完成 Web 信息搜集、信息抽取、信息索引与检索等功能。

2.3.3 不足 信息抽取需要人工定义抽取的模板，而且经常面临数据缺失的情况，得到的结构化数据也属于缺失数据，给进一步分析带来一定的困难。但是作为将非结构化信息转化为结构化信息的最小损失手段之一，信息抽取在中医 NLP 研究中具有非常重要的地位。

2.4 机器学习

医学中机器学习应用比较广泛的是针对结构化数据的分类方法，基于自然语言处理的方向相对较少，机器学习方法应用于文献的分类较为广泛，与文本知识挖掘为不同的研究方向，故不做赘述。中医方面，一些学者尝试使用机器学习技术就某个具体问题提出解决思路，取得一定的成效。孙燕[12]尝试利用支持向量机及相关改进算法对《伤寒论》进行方证分析和量化研究，针对特定药材量化分析并且应用支持向量机对《伤寒论》按照八法训练分类，给出了一些结果。晏峻峰等[13]利用粗糙集理论对中医诊断证素推理规则的获取、症状辨证素的量表制定等证素辨证研究的关键问题进行了研究，主要对症状的诊断和互相之间的关系做出一些系统性探讨。徐蕾[14]提出将决策树方法应用于中医证候学研究的思路，说明决策树方法在中医诊断辨证中的应用前景。卢延鑫等[15]通过词性标注规则提取名词并应用支持向量机对其分类，判定是否为致病因素并与流行病学专家给出的评测结果对比，得到了最高 80% 的准确率。

3 构建中医知识网络的方法

3.1 基于规则推理的知识网络

3.1.1 概述 基于语义理解构建知识网络，即在语义理解的基础上，进一步挖掘语义关系形成的网络关系，基于一定规则人工构建得出。绝大多数该方向的研究都是基于本体（Ontology）实现的，所构建出来的网络属于语义网络。语义网络具有简单、丰富、易读等特点而被广泛使用，著名的一体化医学语言系统（Unified Medical Language System，UMLS）就是基于语义网络而设计的框架，随后有很多学者基于 UMLS 发表了多篇文章。

3.1.2 中医药学语言系统 中国中医科学院信息所 2002 年借鉴 UMLS 的结构并成功应用于中医药领域，构建了中医药学语言系统（Traditional Chinese Medical Language System，TCMLS），至今已收录约 12 万个概念、30 万个术语和 127 万条语义关系，成功应用于文本挖掘、资源检索相关领域[16]。图 1 给出了 TCMLS 的概念结构，可以通过该层次结构为中医 NLP 提供结构化标准，使得文本挖掘研究有章可循、有据可依。图 2 举例刻画中风 - 牛黄清心丸语义网络关系，清晰地描述中药与症状之间的关系，系统还可以进一步延伸，给出多层次、多维度的网络关系示意图。TCMLS 作为一个面向中医药领域的规范化顶层本体，为中医药学语言系统中的所有概念提供了一体化的框架，对于中医药学语言系统的规范化和国际化具有重要意义[17]。

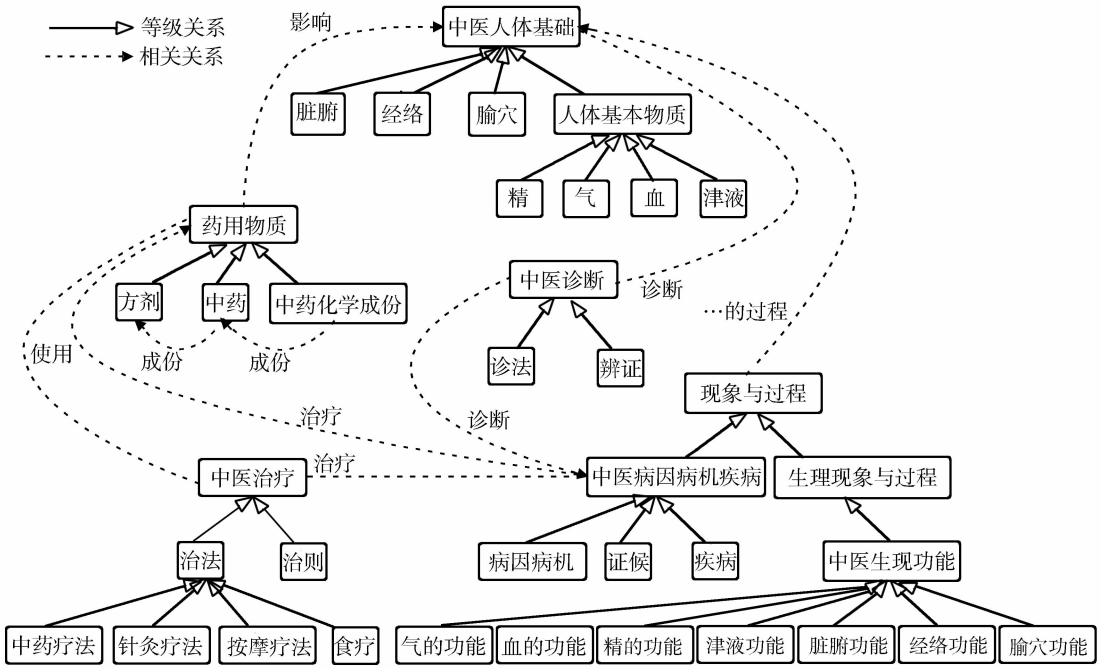

图 1 TCMLS 结构

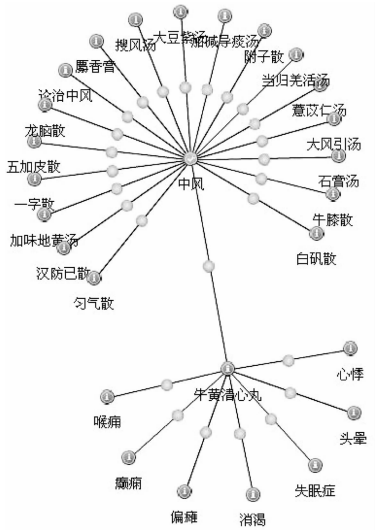

图 2 中风 – 牛黄清心丸语义网络关系

3.2 基于概率统计的知识网络

3.2.1 词向量模型的产生与发展 近年来，基于概率统计的自然语言处理模型出现了许多种，最具代表性的是 N – gram 模型[18]和最大熵模型[19]。2003 年 Bengio 等[20]将 N – gram 算法与神经网络算法结合起来，构建了一个 3 层神经网络来训练词汇的表示（Distributed Representation），相比于词袋模型（Bag of Words），神经网络训练得到的词向量维度较低，而且其中的关联信息也能够体现出来；更为重要的是，词向量的表示方法有可能解决自然语言与神经网络的代沟——维数灾难。Tomas Mikolov 完成了 Word2vec 代码，将 N – gram 算法改进为 Skip – gram 算法，同时对神经网络训练方式进一步改进，使用了层次 Softmax 简化运算，大大提高了算法速度，由于其代码简洁明了，被人们广泛传播学习[21]。

3.2.2 词向量模型在中医领域的应用 Word2vec 在医学领域中的应用在国外也是初有尝试。Miñarro – Giménez 等[22]通过 Word2vec 获取语言学上的一些规律信息，与其他已经公开发布的成果相比较，只得到不到 50% 的准确率，指出未来的研究中应当将 Word2vec 与医学本体相关知识结合起来，采用半监督的方法学习医学中的知识网络。Miñarro – Giménez 等[23]利用 National Drug File – Reference Terminology（NDF – RT）本体来评价 Word2vec 的效果，得到的准确率同样不高。Word2vec 应用于中医研究尚未见公开发布的，因此笔者将其应用于中医文献以及中医相关论文书籍中，以得到一些具有高研究价值的结果。为了能够清晰地展现高维数据的

情况，使用了 t－SNE[24] 可视化方法做出二维展现。

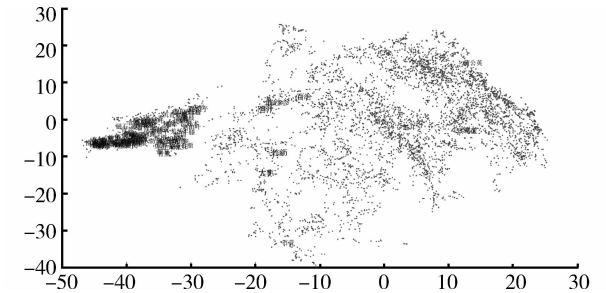

图 3 Word2vec 以中医文献为语料训练出来的 t－SNE 可视化展现结果（黑色词汇为中药名词）

从图 3 中可以看出，使用 Skip－gram 算法训练出来的词向量网络中，中药词汇几乎聚集在了一起，形成了孤岛，这是通过 t－SNE 降维到 13 维之后取差异性最大的前两个维给出的结果。通过词向量训练，能够将相近的词汇聚集在一起，聚类分析便是其应用方向之一。更为重要的是通过训练的过程，将中医相关术语量化为 100～200 维的向量，形成量化之后的知识网络，如果与中医药本体结合起来，将对图 3 给出的关系网络进一步量化，从而得到可信度更高的知识网络。

使用词向量方法还有很多值得研究的地方，如何改进训练过程的规则和参数使得知识网络更为接近真实结果，以及如何将已有的知识作为人工干预加入到训练过程中，都是需要进一步深化研究的课题。

4 结语

综合以上各种方法来看，NLP 在中医学研究过程中是非常有效的工具，通过信息抽取、量化分析，将中医文本知识转化为结构化数据，通过聚类分析和机器学习方法对结构化的数据进一步分析挖掘，可以完成对中医知识的总结整理，进一步有可能发现新的知识。未来对中医自然语言处理的研究有两条明确的路可走：一是语义理解，即选取具体的问题，在局部范围内理解文本中的知识，将其以结构化的方式展现出来，最后利用一些数据分析或者机器学习方法对结构化的信息加以处理，给出具体问题的解决思路。二是概率统计，网络中文本语料的增加使得概率统计 NLP 被广泛使用，词向量就是概率统计 NLP 的产物之一，虽然忽略了上下文的含义，但是通过大量文本的挖掘，可使知识逐渐浮现出来，这一点与大数据研究的思维完全相符。然而，词向量直接应用于医药文本得到的准确率始终不高，以医学本体为评价标准，正确率都比较低。因此，将概率统计的方法与本体知识结合才是最优的解决方法。通过本体和词向量构建中医领域知识网络，将进一步对中医领域的知识做一总体的整理，挖掘中医概念之间的关系，为中医诊疗提供更为实用的信息；如果能够大力开展该方向的课题研究，结合当前热门的大数据挖掘相关方法，最终可能引发中医历史性的革命，使得中医迅速并且持续地发展壮大。

参考文献

1 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.
2 任廷革，刘晓峰.“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）：90－91.
3 任廷革，刘晓峰，张帆，等. 计算技术对中医方剂知识的挖掘［J］. 科技导报，2010，28（15）：31－35.
4 王大阜. 关联规则在中医方剂数据集市中的应用［J］. 贵州大学学报：自然科学版，2006，23（3）：317－319.
5 朱立成，林色奇，薛汉荣，等. 名中医哮喘医案 445 例关联规则分析［J］. 江西中医学院学报，2008，19（5）：83－87.
6 麻晓慧，王泓午，何裕民. 胆病症状学聚类研究［J］. 中国中医基础医学杂志，2005，6（12）：59－61.
7 袁世宏，王米渠，王天芳，等. 聚类分析对肾虚症状的探索性研究［J］. 北京中医药大学学报，2006，29（4）：254－257.
8 何裕民，王莉，石凤亭，等. 体质的聚类研究［J］. 中国中医基础医学杂志，1996，2（5）：7－9.
9 何前锋，周雪忠，周忠眉，等. 基于中药功效的聚类分析［J］. 中国中医药信息杂志，2004，11（6）：561－562.
10 顾铮，顾平. 信息抽取技术在中医研究中的应用［J］. 医学信息（西安上半月），2007，20（1）：27－30.
11 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.
12 孙燕. 基于机器学习技术的《伤寒论》方证分析方法研

究［D］. 北京：北京中医药大学, 2007.
13 晏峻峰, 朱文锋. 粗糙集理论在中医证素辨证研究中的应用［J］. 中国中医基础医学杂志, 2006, 12（2）: 90 - 93.
14 徐蕾, 贺佳, 孟虹, 等. 决策树技术及其在医学中的应用［J］. 数理医药学杂志, 2004, 17（2）: 161 - 164.
15 卢延鑫, 姚旭峰, 王松旺. 利用自然语言处理技术提取致病因素信息研究［J］. 医学信息学杂志, 2013, 34（3）, 55 - 58.
16 贾李蓉, 于彤, 李海燕, 等. 中医药学语言系统的语义网络框架概述［C］. 北京：中国中医药信息大会, 2014.
17 于彤. 中医药学语言系统的语义网络框架［EB/OL］.［2013 - 07 - 15］. http://www.tcmkd.com/ontologies/tcmls/.
18 Brown P F, Desouza P V, Mercer R L, et al. Class - based N - gram Models of Natural Language［J］. Computational Linguistics, 1992, 18（4）: 467 - 479.
19 Berger A L, Pietra V J D, Pietra S A D. A Maximum Entropy Approach to Natural Language Processing［J］. Computational Linguistics, 1996, 22（1）: 39 - 71.
20 Bengio Y, Ducharme R, Vincent P, et al. A Neural Probabilistic Language Model［J］. The Journal of Machine Learning Research, 2003,（3）: 1137 - 1155.
21 Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space［J］. arXiv preprint arXiv: 1301.3781, 2013.
22 Miñarro - Giménez J A, Marín - Alonso O, Samwald M. Applying Deep Learning Techniques on Medical Corpora from the World Wide Web: a prototypical system and evaluation［J］. arXiv preprint arXiv: 1502.03682, 2015.
23 Miñarro - Giménez J A, Marín - Alonso O, Samwald M. Exploring the Application of Deep Learning Techniques on Medical Text Corpora［J］. Studies in Health Technology and Informatics, 2013,（205）: 584 - 588.
24 Van der Maaten L, Hinton G. Visualizing Data Using t - SNE［J］. Journal of Machine Learning Research, 2008,（9）: 2579 - 2605.

（上接第 57 页）
点所含资源的相似度决定查询的转发路径。如果当前节点所含的资源和查询的相似度小于设定的阈值，那么该节点所属簇集内的节点拥有和查询相关资源的可能性也较小——因为根据节点面向兴趣转移的拓扑连接调整，同一簇集内维护相似资源的节点。因此，借助历史反馈信息，将查询路由给需求连接的服务节点。反之，查询可能已经被发送到一个由一组查询主题相关的资源所在节点构成的社区中，当前簇集内包含与查询相关的大部分资源对象。因此根据 SAICA 算法的策略，对于给定的查询，目标社区的定位基本上可以在一个跳数内完成，有效控制了消息数量和搜索路径长度，从而提升了系统的整体搜索性能。

5 结语

本文通过分析提出了基于兴趣的层次化拓扑构建方法和基于兴趣簇的具有自适应能力的搜索算法，解决了医疗物联网中资源发现服务存在的弊端。下一步的工作是对其进行模拟实验，分析其同 ONS 的查询效率问题，以及不在兴趣范围内的节点的解决方案。

参考文献

1 孔宁. 物联网资源寻址关键技术研究［D］. 北京：中国科学院计算机网络信息中心, 2008.
2 黄宇, 金蓓弘. 非结构化 P2P 系统 Overlay 优化技术综述［J］. 小型微型计算机系统, 2008, 29（2）: 238 - 243.
3 李占波, 张哲. 基于 DHT - P2P 新型的 ONS 解析机制［J］. 计算机工程与应用, 2013, 49（3）: 91 - 94.
4 苏森. 无结构 P2P 网络中基于语义和节点存储能力的搜索关键技术研究.［D］. 北京：北京邮电大学, 2011.
5 周晓波, 周健, 卢汉成, 等. 一种基于层次化兴趣的非结构化 p2p 拓扑形成模型［J］. 软件学报, 2007, 18（12）: 3131 - 3138.
6 钱宁, 吴国新. 无结构化 P2P 网络资源搜索机制研究综述［J］. 计算机科学, 2010, 37（4）: 10 - 11.