基于大数据整合与文本挖掘的中药生物分子信息文献系统关键技术模型*

江启煜 何晓华 刘秀峰 刘慧玲 王庆香

（广州中医药大学 广州 510405）

〔摘要〕 基于大数据整合和文本挖掘技术，建立具有多层次信息检索和知识推理发现功能的中药生物分子信息文献系统，介绍系统研发框架、数据库设计、特色与创新之处，指出系统能够为中药研究提供方便可靠的基础数据和分析支持，具有广阔应用前景。

〔关键词〕 大数据；文本挖掘；中药药理；生物分子信息；关键技术

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.02.011

Key Technology Model of the Traditional Chinese Medicine Biomolecular Information Literature System Based on Big Data Integration and Text Mining

JIANG Qiyu, HE Xiaohua, LIU Xiufeng, LIU Huiling, WANG Qingxiang, Guangzhou University of Chinese Medicine, Guangzhou 510405, China

〔**Abstract**〕 Based on big data integration and text mining technology, the Traditional Chinese Medicine (TCM) biomolecular information literature system with multilevel functions of information retrieval and knowledge reasoning discovery has been built. The paper introduces system development framework, design of database, features and innovative parts, points out that the system is able to provide TCM study with convenient and reliable basic data and analysis supports and has a broad application prospective.

〔**Keywords**〕 big data; text mining; pharmacology of Traditional Chinese Medicine (TCM); biomolecular information; key technology

1 引言

中医药的传承离不开现代化的发展，随着生命科学和中药药理的深入研究，巨量的蛋白、基因、通路等生物信息以及药物实验数据不断被发现，如何基于这些大数据建立中药生物分子信息数据平台，为中医药研究提供方便可靠的基础数据支撑和分析支持，具有重要的科研学术价值。这将有助于促进中药药理与生命科学交叉学科领域的深入研究，为中医药临床与生命科学研究之间提供关键纽带。本研究通过大数据[1-4]和文本挖掘[5-9]技术将多个国际著名生物信息数据库的数据、中药信息以及文献信息进行大规模整合，建立具有多层次信息检索和知识推理发现功能的中药生物分子信息文献系统，具有显著现实意义。

〔收稿日期〕 2018－10－10

〔作者简介〕 江启煜，讲师，发表论文20篇。

〔基金项目〕 广东省科技计划项目“基于大数据整合和文本挖掘的中药生物分子信息文献系统的研发”（项目编号：2017A0303 03070）；广东省普通高校青年创新人才项目（项目编号：2016KQNCX024）。

2 系统研发框架与数据库设计

2.1 系统研发技术框架

2.1.1 中药生物分子信息数据库的建立 此阶段主要完成《中药学》、TCMID[9]、Pubchem[10]、HIT[11]、Reactome[12]、KEGG[13]等数据的采集和整合，形成中药－成份－靶蛋白－生物通路的多层次中药生物分子信息数据库。建立的子库包括中药－功效数据库、中药－化学成份数据库、成份－靶蛋白数据库、靶蛋白－靶蛋白相互作用数据库、靶蛋白－通路数据库。

2.1.2 中药生物分子信息文献数据库的建立 通过文本挖掘技术，计算机程序自动从 CBM、Medkine 数据库对中药生物分子信息数据库中的数据进行检索并智能提取返回页面的文献信息，获取的文献信息包括作者、文题、刊名、出版年份、卷号(期号)、起止页码、文摘、关键词，加上检索词、数据库出处这两个字段创建中药生物分子信息文献数据库。

2.1.3 多层次数据信息与文献检索技术的设计与开发 采用双向大数据驱动检索策略分别创建线程向宏观和微观两个层次方向的数据子库同时检索，通过 ADO. NET 连接 SQL Server 数据库，在线程同步控制模块的协同下进行数据整合，将最终检索结果返回用户界面[14-19]。支持 3 种检索模式：单库、跨库、集群跨库检索。

2.1.4 知识发现推理功能的研发 基于后台的生物信息大数据库，生成由巨量信息节点（中药、成份及靶蛋白）组成的复杂关联网络[20-22]，对于用户输入的若干关键词（中药、成份及靶蛋白），搜索生成以这些关键词为中心的关联子网络，即能发现这些关键词与其他生物信息之间的拓扑关联。该功能可以发现成份与目标靶点之间的机制路径、作用中介及其文献信息，对药物作用机制的研究或新药研发具有重要的揭示和启发作用。该系统主要包括多层次检索、推理分析两大模块，系统研发技术框架，见图 1。

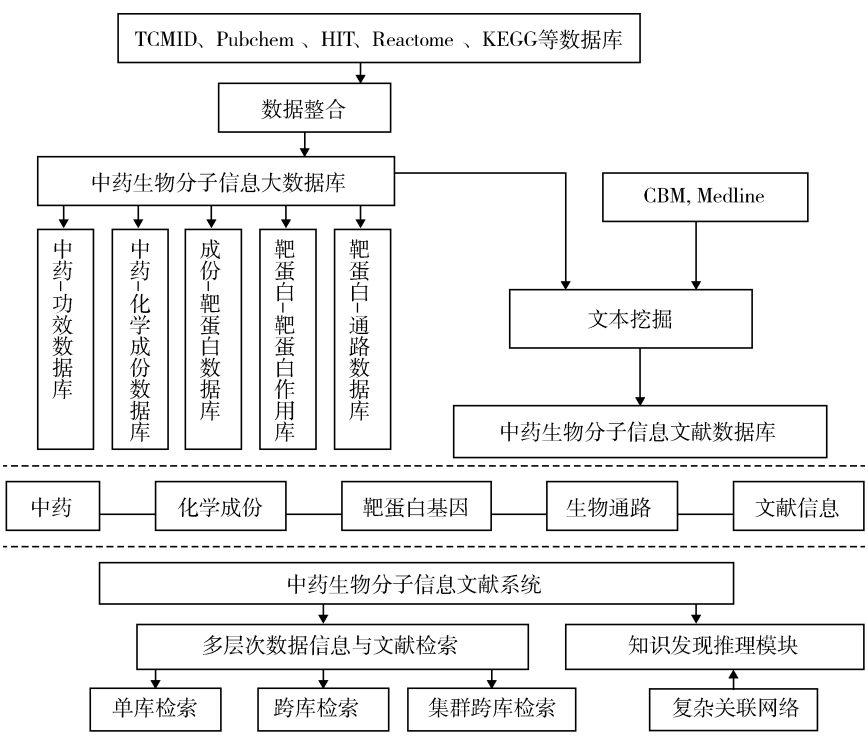

图 1 系统研发技术框架

2.2 数据库设计

2.2.1 基于大数据整合建立中药生物分子信息数据库 由以下子库构成：（1）中药－功效数据库。字段包括 ID、标准药名、同义药名、功效、性能、分类。数据来源为《中国药典》第 10 版以及《中药学》第 7 版教材。（2）中药－化学成份数据库。字段包括 ID 、标准药名、拉丁名、中药对应的化学成份名、化学成份对应的 CAS 号。数据来源为 TCMID 数据库（http：//www. megabionet. org/tcmid），HIT 数据库（http：//lifecenter. sgst. cn/hit/），Pubchem 数据库（pubchem. ncbi. nlm. nih. gov/）。（3）化学成份－靶蛋白数据库。字段包括 ID、化学成份 CAS 号、化学成份对应的靶蛋白全名、化学成份对应的靶蛋白 Symbol 号。数据来源为 TCMID 数据库、HIT 数据库以及 Pubchem 数据库。（4）靶蛋白－靶蛋白相互作用库。字段包括 ID 、靶蛋白 A 、靶蛋白 B 、相互作用主类型、相互作用子类型。数据来源为人类生物通路反应数据库 Reactome 中的 FIsInGene_ with_ annotations 子库（http：//www. reactome. org/）。（5）靶蛋白－通路数据库。字段包括 ID 、生物通路名、参与的基因集。数据来源为人类生物通路反应数据库 Reactome，国际基因组数据库 KEGG（http：//www. kegg. jp），国际蛋白质数据库 UniProt（http：//www. uniprot. org/），以及人类基因数据库 Genecards（http：//www. genecards. org）。上述所需的数据从各大数据库检索提取后按照目标子库的字段进行整合以及预处理后导入 SQL Server 2008 中，使用 SQL Server 2008 构建并储存。

2.2.2 基于文本挖掘技术建立中药生物分子信息文献数据库 在具有上述结构化和层次化的中药生物分子信息数据的基础上构建相应的文献数据库是该系统的核心。主要构建药物－化学成份－蛋白（基因）信息的关联科技文献信息。文本挖掘方法为：将药物－化学成份－蛋白（基因）数据库中的每一项关联数据作为检索词，在中国生物医学文献数据库（http：//www. sinomed. ac. cn/zh/）以及 Medline（https：//www. ncbi. nlm. nih. gov/pubmed/）两大权威生物医学文献数据库中检索相关文献并提取以下信息：作者、文题、刊名、出版年份、卷号（期号）、起止页码、文摘、关键词。在上述信息的基础上加上检索词、数据库出处这两个字段导入 SQL Server 2008 中创建中药生物分子信息文献数据库。由于上述检索和信息提取过程非常繁冗，人工操作难以完成，所以本研究使用文本挖掘技术进行自动智能的检索和信息提取。文本挖掘过程为使用编程语言开发文本挖掘程序，以检索提取 Medline 数据库为例，分析其检索网址的构成以及检索结果网页的源代码，如其带检索词的检索网址为 https：//www. ncbi. nlm. nih. gov/pubmed/？term = 检索词。通过药物－化学成份－蛋白（基因）数据库中每一项数据之间的关联生成组合检索式，生成相应的网址，如 https：//www. ncbi. nlm. nih. gov/pubmed/？term = 成份 And 靶蛋白，即能检索化学成份与靶蛋白之间关联的科技文献。使用 webbrowsre 控件自动加载检索式相应网址，遍历 Document 对象的所有标签，如果标签的 Type 是“radio”并且 sid 等于原网页输出格式中为 Medline 格式的控件 sid，即使用 Click 方法模拟点击，使其输出 Medline 格式的结果，待 webbrowsre 控件重新处于 ReadyState 状态时，通过 innertext 方法读取整个 Document 对象中 HTML 内容，读取每个检索结果中的 FAU（作者）、TI（文题）、JT（刊名）、DP（出版时间）、PG（页码）、AB（摘要）、SO（出版信息）等字段。读取完毕后加上检索词、数据库出处这两个字段写入 SQL Server 2008 中，后期经过人工干预整理将能够建立中药－化学成份－蛋白（基因）文献信息数据库。

2.3 系统检索与推理模型

2.3.1 系统检索 由上述数据库可知各数据子库之间至少有一个字段是关联字段，因此在中药、化学成份小分子、蛋白基因、生物通路、文献信息任意一个数据子库输入检索信息，与其关联的其他数据子库的信息可检索获得。由于每个中药包含许多化学成份（小分子化合物），每个化学成份可能对应数千个靶蛋白（靶点），大量的靶蛋白又参与人体许多不同的生物通路，所以从宏观到微观层次的

检索数据量非常大。因此采用双向大数据驱动检索策略，即以某一信息子库层次为分界点，分别创建线程向宏观和微观两个层次方向的数据子库同时检索，通过 ADO. NET 连接 SQL Server 数据库，在线程同步控制模块的协同下进行数据整合，将最终检索结果返回用户界面。大数据双向多层次检索模块，见图 2。

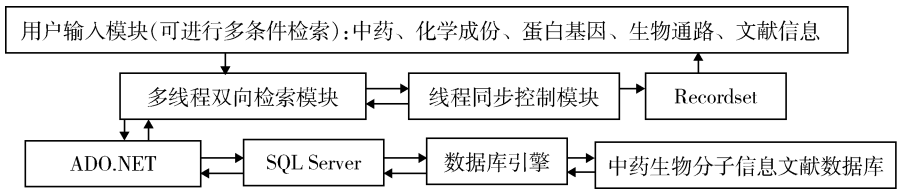

图 2 大数据双向多层次检索模块

2.3.2 知识发现推理 与西药相比，中药对疾病的作用具有多成份、多靶点特性，也就是说药物的功效是对多个靶点共同调控的综合结果。然而化学成份与目标靶点或者多靶点之间往往不一定是直接作用的，因此成份与目标靶点或者多靶点之间的作用路径对中药机制的研究或相关的药物实验设计具有重要的揭示和启发作用。知识发现推理功能示例，见图 3。如用户要分析化学成份 E 与基因 H 之间的关联路径，按照常规检索只能分别获取化学成份 E 以及基因 H 单独的信息。但由于 E、F、G、H 之间在本系统数据库中存在关联性，通过本系统的知识发现推理功能能够生成成份 E 与基因 F、G、H 之间的拓扑网络。通过系统的网络图输出即可知道化学成份 E 与基因 H 之间的作用路径有两条，一是作用于基因 F 而调控基因 H，二是作用于基因 G 而调控基因 H。基因 F 和基因 G 就是本系统发现的化学成份 E 到基因 H 的作用中介。具体而言，通过该系统的中药生物分子信息大数据库中的化学成份 - 靶蛋白数据库以及靶蛋白 - 靶蛋白相互作用数据库，检索用户输入的两个或多个目标信息与其他信息间的所有关联生成两两关联对，给定一个阈值 N，对于每个关联对，重新检索上述数据库，根据存在的关联将每个关联对扩展为多个 N 元关联信息链，在所有 N 元关联信息链中检索同时存在用户输入的两个或多个目标的信息链，将所有这些满足条件的信息链中的关联通过 html5 语言中的 < Canvas > </ Canvas > 画布对象功能在页面上将关联网络图画出，可通过集群跨库检索精确获取相关目标关联文献信息。

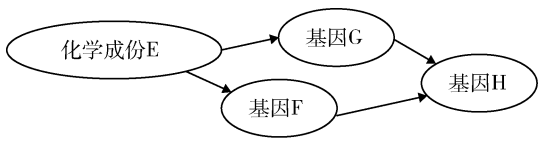

图 3 知识发现推理功能示例

3 系统特色与创新

3.1 专业精准的结构化数据

与现有的数据检索系统平台相比，目前较缺乏专业精准和结构化的中药生物分子信息检索平台。而本系统基于多个著名的国际生物学数据库，将大量的药物、化学成份、蛋白、基因、生物通路等生物信息大数据进行多层次的结构化关联整合，形成从宏观到微观的药物 - 化学成份 - 蛋白（基因）- 生物通路的多层次中药生物分子关联信息数据。

3.2 相关科技文献数据库

虽然目前的文献检索系统平台种类繁多，但大多是某个领域的大范围文献数据库。用户需要根据自身感兴趣知识逐步进行多次组合检索和筛选，得出最终的相关文献。而本系统基于整合的多层次中药生物分子关联信息数据生成内在的知识（检索词）关联复杂网络，除能够直接检索某个知识的相关科技文献外，还能通过关联功能检索与该知识密切相关的其他知识点的科技文献。

3.3 知识发现推理功能

知识发现推理能力是目前大部分检索系统平台

所缺乏的，是本系统的重要创新点之一。系统不但能够检索信息数据，而且能够智能推理知识数据之间的关联，这是大数据技术和人工智能算法在中医药数据库应用的未来发展趋势。对用户输入的两个检索词，本系统能够基于生物信息中的蛋白相互作用数据库发现推理这两个检索词之间的内在关联，形成完整的关联路径。该推理功能有助于中药药理机制的发现以及多成份、多靶点相互作用路径的深入研究。

4 结语

中药药理研究是中医药现代化的重要途径，通过大数据和文本挖掘技术整合并挖掘巨量的中药生物分子数据信息以及文献数据，以计算机检索系统的形式将其有机组织联系起来，建立中药生物分子信息文献系统，具有科学性和创新性。该系统功能涵盖中药功效性能等一般信息、有效化学成份，靶点及相关生物通路信息、中药生物分子文献信息查询以及靶点相互作用路径推理等多方面，极大方便科研人员对中药生物分子数据的搜集与分析，有助于推进中药药理科学研究，为中医药研究提供方便可靠的基础数据支撑。因此中药生物分子信息文献系统具有广阔的应用前景。

参考文献

1 张建英，何建成. 大数据在中医学中应用的可行性分析与展望［J］. 中华中医药杂志，2017，32（1）：17－20.

2 何伟. 大数据时代与中医药学术创新［J］. 中医杂志，2014，55（23）：1981－1984.

3 许海玉，刘振明，付岩，等. 中药整合药理学计算平台的开发与应用［J］. 中国中药杂志，2017，42（18）：3633－3638.

4 夏于芬，梁光平. 大数据背景下的中药现代化［J］. 亚太传统医药，2015，11（21）：1－3.

5 胡双，陆涛，胡建华. 文本挖掘技术在药物研究中的应用［J］. 医学信息学杂志，2013，34（8）：49－53.

6 吕婷，姜友好. 文本挖掘在生物医学领域中的应用及其系统工具［J］. 中华医学图书情报杂志，2010，19（4）：56－64.

7 任郭珉. 基于文本挖掘的药用植物数据库的建立及网络药理学分析［D］. 北京：北京协和医学院，2014.

8 郭洪涛. 文本挖掘技术在中医药文献研究中的应用［J］. 中医学报，2013，28（8）：1151－1152.

9 Xue R，Fang Z，Zhang M，et al. TCMID：traditional Chinese medicine integrative database for herb molecular mechanism analysis［J］. Nucleic Acids Res，2013，41（1）：1089－1095.

10 Wang Y，Suzek T，Zhang J，et al. PubChem BioAssay：2014 update［J］. Nucleic Acids Res，2014，42（1）：1075－1082.

11 Ye H，Ye L，Kang H，et al. HIT：linking herbal active ingredients to targets［J］. Nucleic Acids Res，2011，39（2）：1055－1059.

12 Croft D，Mundo AF，Haw R，et cal. The Reactome Pathway Knowledgebase［J］. Nucleic Acids Res，2014，42（3）：472－477.

13 Kanehisa M，Goto S，Sato Y，et al. Data，Information，Knowledge and Principle：back to metabolism in KEGG［J］. Nucleic Acids Res，2014，42（3）：199－205.

14 江凌，杨平利，杨梅，等. 基于 ADO. NET 技术访问 SQL Server 数据库的编程实现［J］. 现代电子技术，2014，37（8）：95－98.

15 叶安胜. 基于. NET 架构的 WEB 数据库访问技术研究与应用［D］. 成都：电子科技大学，2004.

16 宋阳，严平，曹彤. 基于 ASP、SQL Server 2000 实现的 Web 文献检索系统及其查询优化［J］. 计算机应用与软件，2006（10）：25－28.

17 张贝克，焦迪楠，马昕，等. Net 平台下知识网络系统及其搜索引擎的设计与实现［J］. 微型机与应用，2011，30（8）：4－7.

18 李畅. 基于模糊查询技术的文件检索系统研究［D］. 天津：天津大学，2012.

19 葛强. 基于大型数据库的智能搜索与摘要提取技术研究［D］. 成都：电子科技大学，2015.

20 王忠，陈寅萤，张盈颖，等. 多组分多靶点中药药理作用机制研究中的问题和解决策略［J］. 中国实验方剂学杂志，2018，24（5）：1－6.

21 严蓓华，杨铭，陈佳蕾，等. 复杂网络在中医药方面的研究和应用［J］. 中国实验方剂学杂志，2012，18（7）：276－280.

22 吕庆莉. 数据挖掘与复杂网络的融合及其在中医药领域应用［J］. 中草药，2016，47（8）：1430－1436.