中医药知识图谱构建与应用*

阮 彤 孙程琳 王昊奋 方之家 殷亦超

（华东理工大学 上海 200237） （上海曙光医院 上海 200021）

〔摘要〕 在调研国内外通用和医疗行业专用知识图谱的基础上利用文本抽取、关系数据转换以及数据融合等技术，探索中医药知识图谱自动化构建方法与标准化流程，实现中医药知识图谱的智能应用，包括基于模板的中医药知识问答和基于知识图谱推理的辅助开药。

〔关键词〕 中医药知识图谱；知识问答；知识推理

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2016.04.002

Construction of Traditional Chinese Medicine Knowledge Graph and Its Application *RUAN Tong, SUN Cheng－lin, WANG Hao－fen, FANG Zhi－jia, East China University of Science and Technology, Shanghai 200237, China; YIN Yi－chao, Shanghai Shuguang Hospital, Shanghai 200021, China*

〔**Abstract**〕 Based on researches about domestic and foreign knowledge graphs used universally and specially for the medical industry, the paper explores methods and standard processes of constructing an automatic traditional Chinese medicine knowledge graph by use of technologies such as text extraction, conversion of relational data and data integration, etc. It aims to realize the intelligent application of this knowledge graph, including template－based questions and answers of traditional Chinese medicine knowledge and ancillary prescriptions inferred based on the knowledge graph.

〔**Keywords**〕 Traditional Chinese medicine knowledge graph; Question answering; Knowledge inference

1 引言

随着越来越多语义数据的发布，语义万维网数据源的数量激增，链接开放数据（Linked Open Data）规模越来越大。国内外的搜索引擎巨头纷纷以此为基础构建自己的知识图谱，来进一步提高搜索质量，实现语义搜索，如谷歌知识图谱（Google Knowledge Graph）、百度“知心”和搜狗的“知立方”。对于医疗领域来说，大量的医疗信息学领域的工作者发布了多个链接数据集，同时存在一些链接数据平台整合了其中大部分数据。较为知名的生物医疗数据集平台包括关联开放药物数据集（Linked Open Drug Data，LODD），Liked Life Data 和 Bio2RDF 等。

知识图谱由于具有知识语义化、数据易关联、易扩充等特性，也逐渐被国内医疗界接受。如中国中医科学院中医药信息研究所基于已有的中医药学语言系统[1-4]构建中医药知识图谱[2]。作为上海中医特色医院的曙光医院，拥有较多的中医药数据与

〔收稿日期〕 2016－01－05

〔作者简介〕 阮彤，副教授，发表论文37 篇。

〔基金项目〕 国家高技术研究发展计划“心血管疾病与肿瘤疾病中西医临床大数据处理分析与应用研究”（项目编号：2015AA020107）。

临床诊疗知识库。希望对此数据进行快速地整合和利用，形成临床中医药知识图谱，在此基础上未来形成基于语义技术的临床病例库。为此，本文设计了知识图谱的自动化构建流程，基于文本抽取技术、多策略学习[3]方法、关系数据到 RDF 转换(D2R)[4]等一系列信息技术，实现了中医药知识图谱的自动构建，在此基础上实现了基于模板的中医药知识问答和基于知识图谱推理的辅助开药。

2 相关工作

2.1 通用知识图谱的构建

对于通用知识图谱的构建而言，最主要的数据来源是互联网中的网页，通常采用自底向上的方式，先从网页中识别出实体数据，再从实体数据归纳出数据模式[5]。知识图谱通常存储在图数据库中。目前，国际上最知名的通用知识图谱为 DBpedia[6]，它是世界上最核心的通用知识图谱，以维基百科为数据源，支持 125 种语言。Yet Another Great Ontology[7-9]（YAGO）是一个轻量级的可扩展的本体（Ontology），具备很好的覆盖面和可靠的质量。其中包含的实体都经过了人工评估，其准确率达到 95% 以上，每一个事实都标注了置信度。国内相关工作起步较晚，Zhishi. me[10]致力于发布高质量的中文开放知识图谱，以 3 大中文百科（百度百科、互动百科和维基中文）为数据源，通过结构化信息抽取方法抽取出实体和实体信息。自学习中文本体[11]（Self - supervised Chinese Ontology，SSCO）也是中文领域较早的通用知识图谱之一。同样以中文百科为数据源，通过多种统计学习手段来提高自己的数据质量。SSCO 还通过属性的学习将一些实体普遍存在的属性进行归纳，提升为概念的属性，使得数据模式更加清晰。

2.2 国内外医疗行业知识图谱

2.2.1 国外 医疗领域存在大量面向不同子领域的分类体系标准与数据，如何将这些生物医疗数据整合形成公开的链接数据，对应到已有的分类体系和标准中，是一直以来的研究热点。现有的医疗领域分类标准与医疗本体包括被广泛使用的疾病分类系统 ICD9 和 ICD10、医学主题词表 MeSH、临床医疗术语集 SNOMED - CT，还包括面向药物的命名系统 RxNorm、针对观测指标的编码系统 LOINC、基因本体 Gene Ontology 等。美国统一医学语言系统 UMLS 整合了生物医学领域包括 ICD - 10、MeSH、SNOMED CT、LOINC、RxNorm、Gene Ontology 在内的 100 多部受控词表，共收录了 300 多万个生物医学概念和 1 200 多万个概念名称。此外，UMLS 还提供了词表之间的映射结构，使这些不同的术语系统能够彼此转换。基于这些分类体系和标准，医学信息学领域的工作者发布了多个链接数据集。同时存在一些链接数据平台整合了其中大部分数据。Linked Open Drug Data 整合了 14 个数据集，包含超过 800 万的 RDF 三元组，和超过 37 万的 RDF 链接（2009 年 8 月统计）。对于单个链接数据集，DrugBank 是广为人知的一个医疗链接数据集。自其在 2006 年首次发布以来，已经被广泛用于模拟药物靶点发现、药物设计、药物对接或筛选、药物代谢预测、药物相互作用预测和药学教育等方面。

2.2.2 国内 中国中医科学院中医药信息研究所于 2002 年开始研制中医药学语言系统[1]，目前已发展为包含 12 万多个概念、60 余万术语以及 127 余万语义关系的大型语义网络，在此基础上构建了中医药知识图谱[2]。对如何从多个数据来源构建知识图谱没有给出具体描述，相关应用也停留在浏览与检索方面[2]。本文基于医院已有的信息系统以及临床知识库数据构造中医药知识图谱，提供了语义问答和辅助开方应用。

3 中医药知识图谱构建

3.1 中医药知识图谱剖析

图 1 展示了中医药知识图谱的一部分。中医药知识图谱可以看作是一张图 G，由中医药模式图 G_s、中医药数据图 G_d 以及 G_S 和 G_d 之间的关系 R 组成，即 $G = < G_s, G_d, R >$。模式图 $G_S = < N_s, P_s, E_s >$，其中 N_s 表示图中的类节点，P_s 表示属性边，E_s 表示由多条边连接的两个类之间的关系，$E_s \subseteq N_S \times P_s \times N_S$。数

据图 $G_d = < N_d, P_d, E_d >$，N_d 表示实例节点和字符节点，P_d 表示属性边，使用 E_d 表示由多条边连接的两个节点之间的关系。每条边和边两边的节点都表示一个三元组事实（主语、谓语、宾语）。例如图 1 中虚线部分，肝郁气滞证的基本方为柴胡疏肝散，其中边“基本方”表示属性，其主语是“肝郁气滞证”，宾语是“柴胡疏肝散加减”。模式图和数据图之间的关系 R 使用属性 rdf：type 表示，即 $R = \{(\text{实例}, rdf{:}type, \text{类}) \mid \text{实例} \in N_d, \text{类} \in N_s\}$。

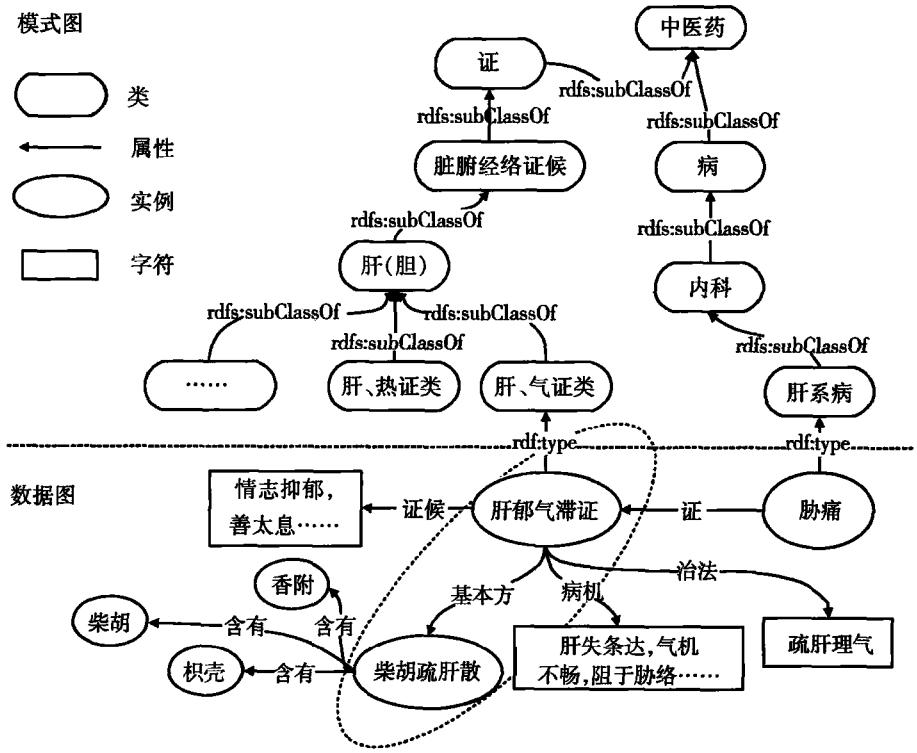

图 1 中医药知识图谱（部分）

3.2 中医药知识图谱构建流程

3.2.1 概述 如图 2 所示，中医药知识图谱的构建与使用流程主要分为以下 6 个阶段：（1）在领域专家的帮助下，根据领域知识创建中医药知识图谱的模式。（2）信息转化模块，将关系数据库中存储的中医药结构化信息转为 RDF 数据。（3）信息抽取模块，采用本文作者提出的多策略学习[3]的方法从半结构化或非结构化数据中抽取数据。所谓多策略学习，是指利用不同数据源之间的冗余信息，使用较易抽取的信息来辅助抽取那些不易抽取的信息。因此，通常利用已有的结构化数据，帮助生成如 Word 文档或是网页数据等半结构化数据的包装器，然后利用半结构化数据的抽取结果，在文本等非结构化数据上，应用远程监督学习的方法进行抽取。（4）中医药知识集成模块，对不同来源数据进行整合，进行模式对齐与实例匹配。进入知识图谱库中。（5）中医药知识反馈模块，由人工专家解决上一阶段产生的模式层、数据层冲突。（6）中医药知识服务模块，在构建好的中医药知识图谱上实现问答和辅助开药。如果中医药知识图谱的模式或是数据来源发生了改变，则需要进一步迭代这几个步骤。下面以曙光 1 期中医药知识图谱的构建过程为例，介绍中医药知识图谱的半自动化构建过程。目前所构建的中医药知识图谱主要包括疾病库、证库、症状库、中草药库和方剂库。这几个库之间是高度相关的。而利用知识图谱，可以方便地表达数据之间的关联。由于中医临证可以采用同病异治，异病同治等方法，因此，几个库之间的关联尤为重要。而利用知识图谱，可以方便地表达数据之间的关联。例如，胁痛与肝郁气滞证相关，胁痛加肝郁气滞证可以使用柴胡疏肝散进行治疗，而柴胡疏肝散是由香附、柴胡、枳壳等组成。其中胁痛存储在疾病库中，肝郁气滞证存储在证库中，柴胡疏肝散

存储在方剂库中，香附、柴胡、枳壳等存储在中草药库中。

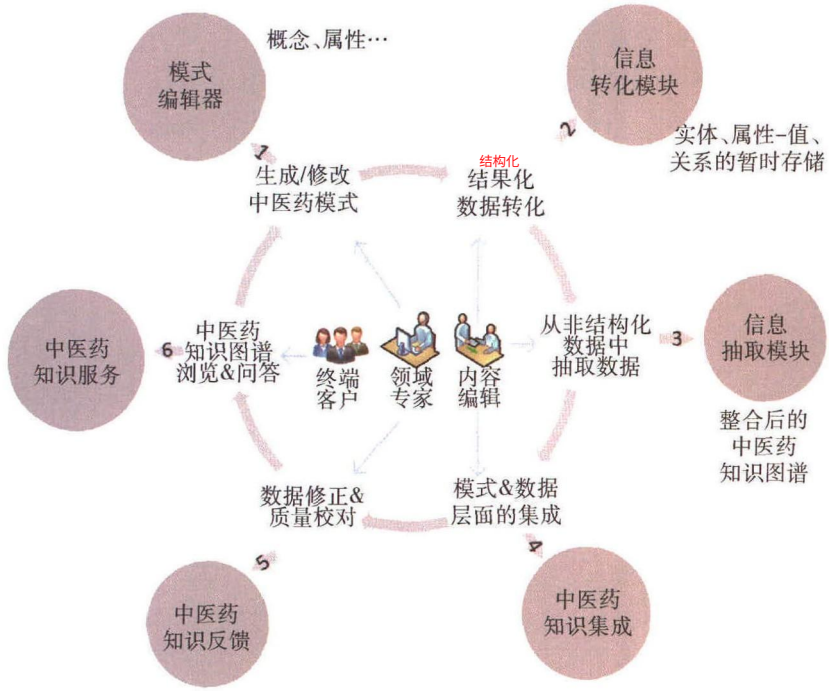

图2 中医药知识图谱构建流程

3.2.2 创建中医药知识图谱的模式 目前主要基于中医病证分类与代码（国家标准）构建中医疾病库和证库的模式。其中，疾病概念的属性有疾病名称、分类编码与描述，证概念的属性有证名称、证名分类编码和描述。方剂库的模式参考了中华中医药学会提供的诊疗指南，属性包括方剂名称、方剂成分与用量。中草药库的模式参考曙光医院数据库中存储的中草药信息及其对应的药品说明书，中草药概念的属性为药材名称、别称、药材编号、规格、服用方法、使用方法、药品代码和进药渠道。由于中医药知识图谱的存储采用三元组的方式，简化了模式更改的代价。项目过程中以及项目后期如果需要丰富模式，只要增删改相应的概念或属性即可。

3.2.3 关系数据库中结构化数据的转化 目前医院内部药品数据以关系数据库方式存储。是把关系数据库中的数据转换成为 RDF 形式的链接数据（Relational Database to RDF，D2R）。D2R 的先驱者 Christian Bizer 和 Andy Seaborne[4] 于 2004 年提出了一种用于描述关系数据库的数据模式与 RDF 模式及 OWL 映射关系的声明式语言 D2RQ，通过使用 D2RQ 进行描述后，用户可以把非 RDF 形式的数据（如关系数据库中的数据）看作虚拟的 RDF 数据，能够使用 RDF 数据查询语言（RDF Data Query Language，RDQL）进行查询。以曙光医院数据库中存储的中草药信息为例，为了将关系数据库中的数据转换成 < 主语，谓语，宾语 > 这样的三元组结构，中草药表的列名“药品名称”可以转化成 RDF 数据中的谓词，值映射为 RDF 宾语。例如，药品“000502”的药品名称可以用三元组 < 000502，药品名称，“羚羊角粉” > 表示。

3.2.4 数据抽取 中医药知识图谱中，文本数据来源主要有 3 方面：（1）国家标准，例如 95 版证名分类标准。（2）曙光医院现有的以 Word 方式存储的临床知识库。（3）中医行业网站和百科。对于前两种数据，构建了针对 Office 软件的一些包装器，利用字体、术语出现等规则进行抽取。对于中医行业网站半结构化信息，构造了面向不同网站的包装器。对于网站、百科下的纯文本，利用远程监督学习方法进行抽取，抽取到的结果可以进一步迭代。例如柴胡的别名为地熏，将其作为种子，在百度百科中抽取到含有柴胡和地熏的句子，如“柴胡别名地熏、山菜、菇草、柴草”，从中抽取到柴胡的其他别名“山菜”、“菇草”、“柴草”等。

3.2.5 知识库模式与数据的集成 由于从多源数据中抽取数据，不同的数据源之间会存在重复或冲突，例如在百度百科中柴胡的别称为“地熏、山

菜、菇草、柴草”，中药查询网上柴胡的别称为“红柴胡、南柴胡、地熏、茈胡、山菜、茹草、柴草”，对数据源可信度进行评分，基于数据来源以及在不同来源中出现的次数，对数据项进行排序，补充到相应属性值字段中。

上面流程基本上完成了中医药知识图谱的构建，在抽取的过程中，当数据出现冲突，或是对数据源质量无法确认等情况下，系统将问题转发到专家，由专家进行最后评定，决定取舍。

4 应用

4.1 基于中医药知识图谱的问答

4.1.1 问答引擎架构 基于知识图谱的问答是一个研究热点，语义问答的实现有多种方法，包括基于模板的语义搜索[12]、子图匹配的方法[13]、语义解析的方法[14-15]等。基于模板的问答由于对于词汇和句法都具有比较好的可扩充性，适合面向领域的知识图谱的问答。因此，本文基于模板方法实现了中医药问答系统。图 3 为基于中医药知识图谱的语义问答引擎架构，输入为自然语言，输出为问题答案。语义问答的实现主要分为 3 个部分：基于知识图谱的分词、模板匹配和模板的翻译执行。

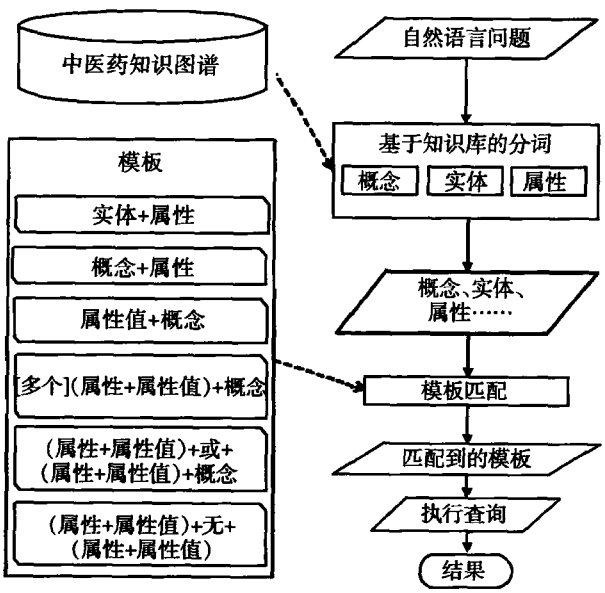

图 3 基于中医药知识图谱的问答引擎架构

4.1.2 对自然语言问答进行分词 基于条件随机场（Conditional Random Fields，CRF）等统计模型的方法速度不能满足问答应用的需要，因此，本文基于中医药知识图谱数据，采用传统的正向最大匹配的原则实现分词和实体识别。在分词的同时确定这些词在知识图谱中的类型，即判定这个词是一个概念、实体还是属性。然后基于事先定义的中医药领域的一些语义模板，将问答匹配到这些模板。模板实质上是中医药知识图谱的子图，例如模板“实体 + 属性”表示知识图谱的一个节点和一条边。模板匹配过程分为两步：第 1 步，根据解析到的实体以及类型确定匹配到的候选模板；第 2 步，判断候选模板与候选实体的结合能否构成知识图谱上的子图，在多个候选模板中找到匹配率最高的模板。确定好模板之后，将模板翻译成语义网络上的标准查询语言 SPARQL，在图数据库上执行。例如，模板“实体 + 属性”的 SPARQL 模板为“SELECT * where {实体 属性 ?O}”。

4.1.3 关于自然语言查询与系统模板的匹配示例 （1）柴胡疏肝散的组成？胁痛都有什么证？匹配到的模板是“实体 + 属性”，可用于查看方剂和疾病的详细信息。（2）肝、气证类包含什么证？匹配到的模板是“概念 + 属性”，可用于查看某类证包含的具体证信息。（3）患有肝郁气滞证的疾病？匹配到的模板是“属性值 + 概念”，可用于查看含有某证的所有疾病信息，因为不同的疾病可能关联相同的证。（4）情志抑郁、乳房胀痛并且脉弦的证？匹配到的模板是“［多个］（属性 + 属性值）+概念”，用户输入多个症状名称，系统可匹配出所有可能的证。（5）患有胁痛或者患有黄疸的所有证？匹配到的模板是“（属性 + 属性值）+或 +（属性 + 属性值）”，用于查看肝胆病胁痛和黄疸的所有证信息。（6）患有胁痛且不含有脉弦的证？匹配到的模板是“（属性 + 属性值）+无 +（属性 + 属性值）”，用于查看疾病胁痛的所有证中，不含有症状脉弦的证信息。

4.2 基于中医药知识图谱的辅助开药

项目将中医药知识图谱数据与推理技术结合起来，用于中医药辅助开方中。利用 Drools 引擎以及团队开发的转换程序，将图谱中存储的数据自动地转换成推理引擎适用的推理规则，这些规则结合医生工作站传来的病人事实数据，可以辅助医生进行开方。中医的药方包含基础药方和经验药方。基础药方是由疾病和证确定的，而经验药方则需要根据

病人的症状确定。当医生诊断得到病人所患疾病、证以及症状后，通过知识图谱的推理引擎即可得到推荐的药方。如图4所示，知识库中存储的事实为：肝郁气滞证的基础药方为柴胡12g，香附15g，枳壳12g，陈皮6g，川芎15g，白芍15g，甘草6g。肝郁气滞证的证候及经验药方为：气郁化火，胁肋掣痛，心急烦躁，口苦口干，尿黄便秘，舌红苔黄，脉象弦数，去川芎，加牡丹皮12g，栀子12g，黄连3g，川楝子9g，延胡索12g以清肝理气，活血止痛。将肝郁气滞证的药方自动化地转化为推理引擎Drools中的规则（人－证－肝郁气滞证）（人－症状－气郁化火 or 胁肋掣痛 or 心急烦躁 or 口苦口干 or 尿黄便秘 or 舌红苔黄 or 脉象弦数）——>（基本方－加－牡丹皮12g，栀子12g……）（基本方－减－川芎）。当输入的病人患有胁痛并被诊断为肝郁气滞证，根据规则只要使用基本方即可。但是，对于还有“口苦口干”症状的病人，根据规则还需要去除川芎，增加牡丹皮等中草药。

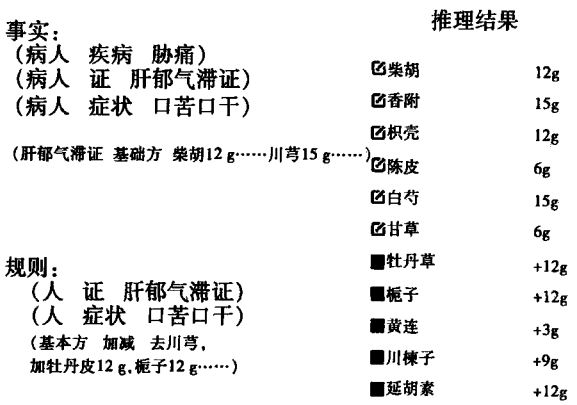

图4 辅助开方过程

5 结语

本项目提出中医药知识图谱的半自动化构建流程，使用该流程构建中医药知识图谱，实现基于中医药知识的智能应用：中医药知识问答和辅助开药。下一步的工作将利用中医药知识图谱，对电子病历进行自动化解析和标注，形成基于知识图谱技术的临床病例库。

参考文献

1 贾李蓉，于彤，崔蒙，等. 中医药学语言系统研究进展[J]. 中国数字医学，2014，(10)：57－59，62.

2 贾李蓉，刘静，于彤，等. 中医药知识图谱构建[J]. 医学信息学杂志，2015，(8)：51－53，59.

3 Tong Ruan, Yeli Lin, Haofen Wang, et al. A multi－strategy learning approach to competitor identification [J]. JIST, 2014, 8943: 197－212.

4 Bizer, Christian and Andy Seaborne. D2RQ－treating non－RDF databases as virtual RDF graphs [C]. Hiroshima: Proceedings of the 3rd International Semantic Web Conference (ISWC2004), 2004.

5 S Amit. Introducing the Knowledge Graph: things, not Strings [EB/OL]. [2015－12－20]. https://googleblog.blogspot.com/2012/05/introducing－knowledge－graph－things－not.html.

6 S Auer, C Bizer, G Kobilarov, et al. DBpedia: a nucleus for a web of open data [C]. Proc. of the 6th Int. The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference, 2007: 722－735.

7 F M Suchanek, G Kasneci, G Weikum. YAGO: a core of semantic knowledge unifying wordNet and wikipedia [C]. Proceedings of the 16th International Conference on World Wide Web, 2007: 697－706.

8 Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell [M]. Essex, UK: Elsevier Science Publishers Ltd, 2013: 28－61.

9 J Biega, E Kuzey, F M Suchanek. Inside YAGO2s: a transparent information extraction architecture [C]. New York: Proc of the 22th International Conference on World Wide Web, 2013: 325－328.

10 X Niu, X Sun, H Wang, et al. Zhishi. me: weaving Chinese linking open data [C]. Proceedings of the 10th International Conference on the Semantic Web, 2011: 205－220.

11 胡芳槐. 基于多种数据源的中文知识图谱构建方法研究[D]. 上海：华东理工大学，2015.

12 Weiguo Zheng, Lei Zou, Xiang Lian, et al. How to Build Templates for RDF Question/Answering: an uncertain graph similarity join approach [C]. Proceedings of SIGMOD Conference, 2015: 1809－1824.

13 Lei Zou, Ruizhe Huang, Haixun Wang, et al. Natural Language Question Answering over RDF: a graph data driven approach [C]. Proceedings of SIGMOD Conference, 2014: 313－324.

14 Wen－tau Yih, Ming－Wei Chang, Xiaodong He, et al. Semantic Parsing via Staged Query Graph Generation: question answering with knowledge base [C]. Proceedings of ACL, 2015: 1321－1331.

15 Qingqing Cai, Alexander Yates. Large－scale Semantic Parsing via Schema Matching and Lexicon Extension [C]. Proceedings of ACL, 2013: 423－433.