《智能信息处理》课程考试

基于知识图谱的医疗知识搜索

董雪松

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 15 日

基于知识图谱的医疗知识搜索

董雪松

（大连海事大学 信息科学技术学院, 大连 116026）

摘　要　互联网信息的爆发式增长，既为用户带来了丰富的信息知识，也为用户从中筛选所需知识造成了困难。传统的搜索引擎基于对互联网数据的全文索引，通过关键词匹配索引的方式为用户返回相关信息的链接，而不是直接明确的知识点，用户仍需从返回的大量冗余链接中查找并提炼自己所需的知识。如何从海量的、结构多样化的信息中有针对的为用户返回精确信息，已成为当前知识搜索的研究热点。知识图谱技术的兴起为该研究提供了新的解决思路。知识图谱能够以一种更直观的方式表达出现实世界中的实体的信息以及实体和概念之间的关联。本文将互联网文本数据作为语料资源，对其进行知识图谱构建的研究。本文基于构建的中文医疗领域知识图谱，设计实现了医疗知识搜索系统。通过对用户输入的自然语言进行句法分析和语义依存分析等处理，识别用户的搜索意图，借助知识图谱，以一种更加直观、精确的方式返回用户所需的知识。

关键词　知识图谱；序列标注；医疗知识搜索；

中图法分类号　TP36　　DOI 号

Medical knowledge search based on knowledge graph

Dong Xue Song

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract　The explosive growth of Internet information has not only brought users a wealth of information knowledge, but also caused difficulties for users to filter the required knowledge. Traditional search engines are based on the full-text index of Internet data and return links to relevant information for users through keyword matching indexing instead of direct and clear knowledge points. Users still need to find and refine themselves from the large number of redundant links returned. The required knowledge. How to return accurate information for users from the massive and diverse structure of information has become a current research hotspot in knowledge search. The rise of knowledge graph technology provides new solutions for this research. The knowledge graph can express the information of entities in the real world and the relationships between entities and concepts in a more intuitive way. This paper uses Internet text data as a corpus resource, and conducts research on the construction of knowledge graphs. Based on the constructed knowledge map of the Chinese medical field, this paper designs and implements a medical knowledge search system. Through the processing of syntactic analysis and semantic dependency analysis of the natural language input by the user, the user's search intention is recognized, and the knowledge graph is used to return the knowledge required by the user in a more intuitive and accurate way.

Key words　Knowledge graph; sequence labeling; medical knowledge search;

1 引言

随着信息技术和互联网的快速发展,越来越多的人通过网络渠道获取信息和分享资讯，海量的信息充斥网络。互联网作为人们获取知识的一个重要渠道，在享受其带来的丰富信息知识搜索便利的同时，也使得人们陷入了选择困难:如何快速从海量的、多模态化的、冗余的信息当中获取自己所需的知识，已成为当前需要解决的问题。

目前大多数用户查询相关知识是通过搜索引擎的渠道，其中健康与医疗主题的搜索排名占首位。传统的搜索引擎其原理是以文档关键词建立倒排索引，来为用户提供更高效的信息查询。虽然在效率和准确性上表现良好，但是局限性也十分明显:其返回的结果比较单一，不能对用户的查询输入和查询的目标进行理解，因此在搜索精度上存在明显缺陷，需要用户从返回结果中再查找信息。这是由于HTML 标签的数据形式缺乏语义，导致关键的信息难以有效抽取。直到语义 Web 的概念提出[1]，为如何解决互联网信息的语义问题提供了可行方案。Web 资源可以通过<主语，谓语，宾语>三元组形式的资源描述框架 RDF 标记语言来对语义进行描述。此外，还可以加入网络本体语言 OWL 来对资源的语义进一步表示。复杂的语义知识库既含有事实类知识，其背后还包含众多的如实体、关系、规则等的语义信息。因此，语义网可以为知识库构建、用户自然语言理解、知识推理和计算等方面提供强有力的支持。

知识图谱[2]是语义网的技术之一,已成为当前搜索引擎[3]技术发展的一个研究重点。Google 是这一概念的倡导者和先行者，期望通过知识图谱来刻画现实世界中各种实体和概念，以及它们之间的关联。知识图谱是将互联网文本中的知识进行抽取，以图的形式构建,用节点表示实体或概念 ，用节点之间的连线表示关系。知识图谱把各种实体和概念整合在一起，构建了一个关系网络，为研究者提供了“关系”的视角来分析和研究问题。知识图谱加速了语义搜索的发展，通过知识图谱，用户可以更准确的获得知识以及知识和知识之间的逻辑关系，用户获得的不只是通向知识的链接，还有知识本身。比如搜索“感冒了吃什么好”，系统能直接返回推荐的相关食物，而不是返回包含“感冒了吃什么好”几个关键字的网页。近几年来，医疗信息管理系统、电子病历档案等信息化设备工具在医疗单位迅速普及，医疗领域的知识问答社区、知识百科纷纷建立，都产生了大量的医疗数据。当前，人们对健康问题地关注日益增加，如何更好的利用海量的医疗信息资源为用户提供精准的医疗健康知识，已成为一个热点。知识图谱为此提供了一个新的方法和途径。针对海量医疗信息中的数据多源、冗余、内容分散等问题，抽取其中的知识实体并进行有效整合，为用户提供优质地知识问答、知识推理等服务。所以，基于知识图谱构建医疗知识搜索系统的研究具有重要的现实意义和显著的应用价值。

2 知识图谱技术

知识图谱技术[4]是本文构建知识搜索系统的基石，主要解决知识的获取与整合问题。知识图谱本质上是结构化的语义网络，以图的形式进行存储。存储结构由节点和节点联系组成。在知识图谱中，真实世界中的各种事物被抽象为一个个节点，而各种事物的相互关系则被抽象为节点间的连线。知识图谱提供了一种更为直观的方式观察真实世界中的关系网络。Google 最先将知识图谱应用于其搜索业务的优化。与传统的利用关键词搜索相比，利用知识图谱可以识别用户输入文本中的语义，发现信息中的实体和隐含联系，从而提供更高质量的搜索。用户不必浏览大量网页就可以准确定位和深度获取知识。

2.1 知识图谱的相关概念

本体(ontology)[5]是共享概念模型的显式说明，描述概念与概念间的关系;是语义 Web 的关键技术，用于 Web 网页添加语义。语义 Web 理念中的本体与知识图谱，二者密切相关。本体描述概念及概念间的关系，是大多数知识图谱的模式层，是知识图谱的概念模型和逻辑基础。知识图谱与本体的相同之处在于:二者都通过定义元数据以支持语义服务。不同之处在于:知识图谱更灵活，支持通过添加自定义的标签划分事物的类别。本体侧重概念模型的说明，能对知识表示进行概括性、抽象性的描述，强调的是概念以及概念之间的关系。大部分本体不包含过多的实例，本体实例的填充通常是在本体构建完成以后进行的。知识图谱更侧重描述实体关系，在实体层面对本体进行大量的丰富与扩充。可以认为，本体是知识图谱的抽象表达，描述知识图谱的上层模式;知识图谱是本体的实例化，是基于本体的

知识库。

知识图谱采用三元组描述事实，所使用的描述语言大多是已研发的本体语言，如 RDFS、OWL 等。知识图谱也可以通过 RDFS 或 OWL 定义规则用于知识推理。知识图谱的关键技术也与本体很相似，涉及:(1)知识图谱构建阶段的实体抽取、关系抽取、语义解析等机器学习和自然语言处理方法和算法，(2)用于知识图谱存储的知识表示、图数据库和知识融合等方法和技术，(3)知识图谱应用阶段的数据集成、知识推理等。

除了本体之外，与知识图谱相关的概念还有知识地图和科学知识图谱。知识地图(knowledge map)将特定组织内的知识索引通过“地图”的形式串联在一起,揭示相关知识资源的类型、特征以及相互关系。知识地图的主要功能在于实现知识的快速检索、共享和再重用，充分有效地利用知识资源。知识地图是关于知识的来源的知识。知识并非存储在知识地图中，而是存储在知识地图所指向的知识源中。知识地图指向的知识源包含数据库、文件以及拥有丰富隐性知识的专家或员工。有的企业应用知识地图来揭示知识的结构，实现对知识及其相关知识的检索。另外，知识地图在文献学中也有应用，即科学知识图谱。科学知识图谱(mapping knowledge domain)是用来显示知识演化进程和知识结构的图形化与序列化的知识谱系。

2.2 知识图谱的架构

知识图谱的架构[6]可以从逻辑结构和技术结构两个角度来进行阐述。

知识图谱在逻辑结构上分为数据层和模式层。在数据层中，知识以“实体-关系-实体”或“实体-属性-值”三元组的形式存储。所有的三元组相互关连组成了关系网络，构成了知识的图谱。模式层是对知识进行规范整合。因本体库可以通过对规则、约束进行定义从而实现知识的规范化，所以常用本体库对模式层进行管理。

2.3 知识图谱的构建

知识图谱的构建[7]方式一般分为两种:自顶向下的方式和自底向上的方式。自顶向下的构建方式是基于本体构建的方式,以结构化程度高的百科类等网站为数据源，从中抽取本体和规则约束，填充到知识库中;而自底向上的构建方式是直将收集的数据通过模式识别、制定规则等方式，从中识别实体、属性以及关系，然后加入到知识图谱中。

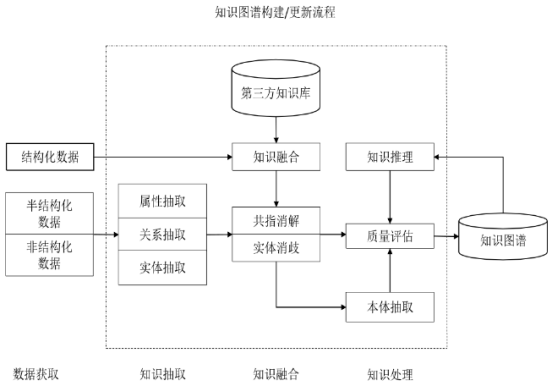

图 1 知识图谱的技术架构图

图 1 为知识图谱的技术架构图，方框内知识的抽取、融合与处理三步骤是知识图谱构建的核心。由图可见，结构化的数据因为规范化程度较高，可以较为容易的从中抽取知识;半结构和非结构化数据规范性较差，难以直接获取知识，因此需要借助属性抽取、关系抽取、实体抽取等一系列操作提取出知识的实体和关联，然后存入知识库中。知识图谱的构建过程是不断循环往复的，可将迭代的环节大致分为三个阶段:知识抽取、知识融合以及知识处理。

3 索引及索引技术

图形数据库解决了知识图谱的存储问题，最终对用户的搜索查询问题，展示对应的查询结果。当涉及到大批量的数据查询时，传统的查询方法在查询效率上很难满足当前的应用需求，尤其在搜索应用上表现更为突出。通常的解决方案是构建全文索引对检索进行优化，提升查询效率。这里采用 Lucene[8]全文索引框架对医疗知识搜索进行优化，以提供更好的用户查询体验。

3.1 全文检索框架 Lucene

Lucene 不是一个完整的搜索框架应用，是一个基于 JAVA 语言开发的全文检索工具包，能为开发者的程序中嵌入全文搜索的功能。

目前已经有很多应用集成 Lucene 实现搜索功能。例如 Eclipse，在帮助菜单中提供基于 Lucene 的全文搜索。Lucene 的索引是建立在文本类型数据上，因此只要能转换成文本类型数据的格式都可以用 Lucene 提供索引服务。例如从网络上采集的 HTML 文本数据，首先将 HTML 转换成文本文件，然后由 Lucene 处理创建索引，最后将创建的索引保存至磁盘或内存中。当用户进行搜索查询时，能通过索

引文件快速找到相应的文档。

下图 2 描述了运用 Lucene 构建搜索应用的流程。构建索引是建立搜索引擎的重要环节。这一环节的重要性在于:通过建立文档关键词的倒排索引，可以快速查找到相应的文档。若没有索引，需要将所有的本文文档全部遍历，检查其中是否有指定的关键词。整个过程会消耗大量的时间，因此无法满足对用户快速响应的需求。

索引可以被看作一种特殊的目录结构。类似于字典可以通过拼音或偏旁来查找字的页码，索引可以根据关键词，进而查找到对应的文档所在的位置。Lucene 采用的是倒排索引机制。其原理对于所维护的词/短语文件中的每个词/短语，在索引表中都会有一项记录该词/短语和对应的位置。这样能够对用户的输入查询快速的返回搜索结果。

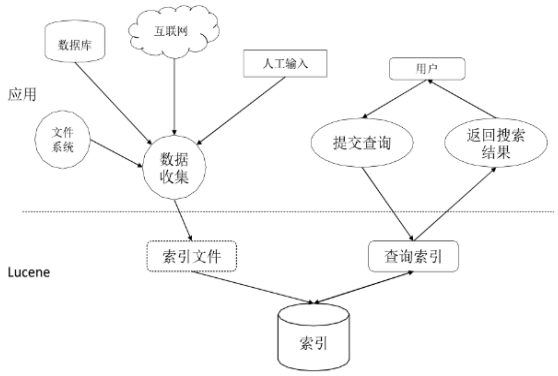

图 2　Lucene 构建搜索应用程序的流程

索引的建立一般分为四个步骤。

首先是获取内容。数据来源包括文件系统、数据库、Web 网页等，Lucene 本身不具备采集数据的功能，这一功能可以借助很多开源的爬虫得以实现。

然后是建立文档，从数据源获取原始数据后，需要把要索引的数据转化成文本，包括文档中的域值对。

再次是进行文档分析，Lucene 通过对关键词搜索实现查询，因此需要将文档通过文档分析进行分词操作。

最后是建立文档索引，建立单词与文档的映射，并通过建立倒排索引的方式记录单词出现的频次与位置，便于查找。

建立好倒排索引[9]文件后，就可以通过索引进行搜索。搜索引擎首先会对输入语句进行关键词解析，查找索引表，找到索引表中关键词的文档位置，最后返回相应的文档。

4　基于知识图谱的医疗知识搜索

为了实现基于知识图谱的医疗知识搜索，本文设计一个基于多个数据源的知识搜索系统,数据源包括图数据库存储的知识图谱信息和索引文档存储的知识问答信息。

4.1　系统总体框架

图 3 所示是整个医疗知识搜索系统的框架。当用户输入查询问句时，系统首先对该语句进行处理。处理内容包括实体识别、句法分析、语义分析等，构建对应的语法分析树。然后将处理结果提交到知识搜索分发模块，查找与处理结果匹配的搜索模板，明确用户的搜索意图;识别搜索意图后，根据意图转发到相应的核心服务模块进行处理;最后，将请求的处理结果进行整合、排序、推荐。

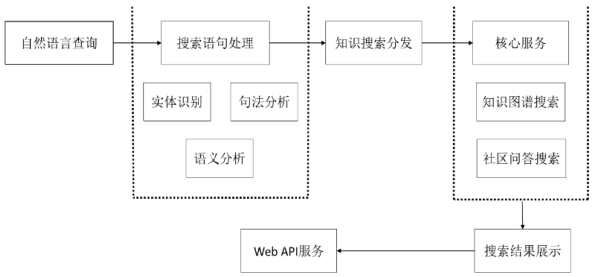

图 3　系统总体框架

4.2　基于知识图谱的搜索服务

基于知识图谱的搜索服务是以知识图谱作为底层的数据支持，核心在于将用户无结构化的自然语言查询语句转化为结构化的知识图谱查询语句,即从用户的自然语言查询中抽取出其中的实体和关系。本文采用预定义模板的方法来处理针对知识图谱的搜索服务。基于模板的问答在词汇和句法上具有较好的可扩充性，因此适合面向领域的知识图谱搜索。但因模板的定义需要过多人工干预，且模板容易存在遗漏，因此将基于语义的抽取作为补充。

在搜索语句处理模块中，已经对用户的自然语言查询语句进行了预处理，对其中的实体、关系以及语法依存进行了识别。然后通过比对分词后每个词的词性等信息，确定这个词是一个实体、属性或是概念。然后与预定义的医疗领域的模板进行匹配。模板实质上可以看做所构建知识图谱[10]中的一系列子图。其匹配流程如下:

1）根据识别到的实体及其类型确定匹配的候选模板

2）判断候选实体和模板是否能够成知识图谱的一个子图，并在多个候选模板中找到匹配度最高

的模板。

3）确定模板后，将模板翻译成对应的 Cypher 查询语言，在 neo4j 上执行。以模板“实体+属性”为例，对应的查询语言为*match (x:实体类型{属性名:属性值}）return x"。

5 总结

随着网络信息的爆炸式增长，以及人们对医疗健康问题的逐渐重视，如何从海量的信息中针对用户的知识获取需求提供精准且高质量的知识，以及清晰的知识网络展示，成为了研究者热衷的研究方向。知识图谱为解决这一问题提供了强有力的支撑。本文针对用户这一需求，构建了医疗领域知识图谱[11]，并提供相关的知识搜索服务。

本文借助构建的中文医疗知识图谱，设计并实现了一个医疗知识搜索系统。对于用户输入的自然语言形式的问句，针对用户不同的搜索意图，采用基于知识图谱的知识搜索和面向社区问答的知识搜索方案，为用户提供更精准、更直观的知识搜索结果。

参 考 文 献

[1]王则栋,张磊,李滕飞,王宇璐,王抵修.支持 Web 公式语义化的数据库设计[J].科学技术创新,2021(27):109-110.

[2]Aidan Hogan, et al.Knowledge Graphs.Morgan & Claypool Publishers,2021.

[3]陈娟.基于 Java 的搜索引擎的研究与设计[J].电子技术与软件工程,2021(21):8-9.

[4]刘巍,陈霄,陈静,周觐,张斌.知识图谱技术研究[J].指挥控制与仿真,2021,43(06):6-13.

[5]顾丹阳,李明倩,权冀川,刘勇,罗晨.基于本体的主战武器装备知识图谱构建[J].指挥控制与仿真,2021,43(06):14-20.

[6]周丽娜,马志强.基于知识图谱的网络信息体系智能参考架构设计[J].中国电子科学研究院学报,2018,13(04):378-383.

[7]Xiang Gao,Wenjing Cui,Leijiang Yao,Yajie Zhou,Guanhua Wang. Construction of Knowledge Map of Continuous Fiber Reinforced Ceramic Matrix Composites[P]. Computing and Pattern Recognition,2019.

[8]Liana Diesendruck,Rob Kooper,Luigi Marini,Kenton McHenry. Using Lucene to index and search the digitized 1940 US Census[J]. Concurrency and Computation: Practice and Experience,2014,26(13):

[9]姜琨,朱磊,宋省身,杨岳湘.倒排索引压缩算法研究综述[J].小型微型计算机系统,2020,41(04):715-723.

[10]黄伟, 李莉,and 徐彭娜."医疗知识图谱的自动问答系统分析研究." 福建电脑 37.11(2021):100-103. doi:10.16707/j.cnki.fjpc.2021.11.024.

[11]覃晓,廖兆琪,施宇 & 元昌安.(2020).知识图谱技术进展及展望. 广西科学院学报 (03),242-251. doi:10.13657/j.cnki.gxkxyxb.20201027.009.