北京语言大学

硕士研究生学位论文

题 目：**文物领域知识图谱构建及应用研究**

姓　名：卢梦依

国　籍：中国

学　号：201721198599

院　系：信息科学学院

专　业：软件工程

研究方向：自然语言处理

导　师：刘鹏远　副教授

二〇二〇年5月

北京语言大学
BEIJING LANGUAGE AND CULTURE UNIVERSITY

论文原创性声明

本人郑重声明：所呈交的论文，是本人在导师指导下，独立进行的研究工作及取得的研究成果。尽我所知，除了文中已经注明引用和致谢的地方外，论文中不包含其他人或集体已经发表或撰写的研究成果，也不包含为获得北京语言大学或其他教育机构的学位或证书所使用过的材料。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

签　　名：_____

日　　期：_____

学位论文知识产权权属声明

本人郑重声明：本人所呈交论文，是在导师指导下所完成的，论文知识产权归属北京语言大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版本，允许论文被查询和借阅，将论文编入有关数据库进行检索等。本人离校后发表或使用学位论文或与该论文直接相关的学术论文获成果时，署名仍为北京语言大学。

签　　名：_____

导师签名：刘明宇

日　　期：_____

文物领域知识图谱的构建及应用研究

摘 要

文以载道，物传精神，文物是历史文化的载体，是中华文明源远流长和生生不息的实物见证，是传承弘扬中华优秀传统文化的历史根脉，也是重要的历史资源。近年来，随着国家文物保护和事业发展工作的大力推进，文物宣传工作趋于多样化，使大家对文物的关注明显提升，让文物活起来已蔚然成势，但其领域数字化资源丰富，数据海量，多源且异构，如何有效组织管理及呈现非结构化和半结构化文物数据成为研究重点。目前面向文物领域的知识图谱研究并不多，具体的应用多是博物馆为了专题文物的展陈而构建的小规模知识图谱，构建方式也大都依赖于人工构建，缺乏自动化方法。因此本文基于上述背景，从实际应用出发，设计并实现了完整的文物知识图谱构建和应用方案。

本文主要开展的工作有：1. 针对文物领域的本体构建。基于文物领域的特点以及实际应用场景的考量，结合本体构建方法七步法和循环获取法，提出了基于知识图谱应用场景的文物领域本体构建方法，2. 针对文物领域知识图谱的构建，提出了基于特征词集的文物属性知识抽取方法将 web 非结构化文本中具有的属性知识抽取出来，经验证，该方法能有效对文本中的属性进行抽取。3. 基于文物知识图谱的智能问答模型框架设计。针对问句的意图识别使用了 BERT 模型进行对问句的意图分类，能够更好的挖掘自然语言问句中的语义信息，经实验证明，能对问句的意图分类产生比较好的效果。针对问句的槽位提取，本文使用基于 Bi-LSTM 模型和字典匹配模型的结合完成了槽位提取功能，最后将问句的意图和属性信息进行转化为 SPARQL 语句，将其送到基于 Apache Fuseki 搭建的 SPARQL 查询检索 API 中，完成对问句的解析返回最终答案，经实验验证，该方法使回答问题的准确性得到了提升，可以有效对基础问答进行回答。

本文针对上述智能问答模型实现了一个构建了一个文物助手系统，还增加了语音问答、每日文物知识和博物馆资讯查询的功能，以微信公众号的展示形式为用户提供服务，可以实时对用户的自然语言问句进行准确回答，同时也在实际环境中对设计的系统进行了功能上和性能上的检验，结果表明该系统有良好的问稳定性，证明上述算法具有可行性。

关键词：文物 知识图谱 本体 智能问答 实体识别 属性链接

Research on Construction and Application of Knowledge Graph Based on Cultural Relics

Abstract

The text contains the Tao, the spirit of the material, and the cultural relic is the carrier of history and culture. It is a physical testimony of the long history and endlessness of Chinese civilization. It is the historical root of inheriting and promoting China's excellent traditional culture, and it is also an important historical resource. In recent years, with the vigorous promotion of national cultural relics protection and career development, cultural relics propaganda work has become more diversified, making everyone's attention to cultural relics significantly increased, and making cultural relics alive has become a trend, but its field is rich in digital resources and data. Massive, multi-source and heterogeneous, how to effectively organize and present unstructured and semi-structured cultural relics data has become the research focus. At present, there is not much research on knowledge graphs in the field of cultural relics. The specific applications are mostly small-scale knowledge graphs built by museums for the display of special cultural relics. Most of the construction methods rely on manual construction and lack of automated methods. Therefore, based on the above background and starting from practical application, this paper designs and implements a complete cultural relics knowledge graph construction and application scheme.

The main work carried out in this article are: 1. Ontology construction in the field of cultural relics. Based on the characteristics of the field of cultural relics and the consideration of actual application scenarios, combined with the seven-step method of the ontology construction method and the loop acquisition method, a method of constructing the ontology of the cultural relics domain based on the application scenario of the knowledge graph is proposed. The cultural property attribute knowledge extraction method based on feature word sets extracts the attribute knowledge contained in the web unstructured text. After verification, this method can effectively extract the attributes in the text. 3. Intelligent question answering model framework design based on knowledge map of cultural relics. In order to identify the intention of the question, the BERT model is used to classify the intention of the question, which can better mine the semantic information in the natural language question. It has been experimentally proved that it can produce a better effect on the intention classification of the question. For the slot extraction of the question, this article uses the combination of the Bi-LSTM model and the

dictionary matching model to complete the slot extraction function, and finally converts the intent and attribute information of the question into a SPARQL sentence, which is sent to the Apache Fuseki-based In the built SPARQL query and retrieval API, the analysis of the question is completed and the final answer is returned. After experimental verification, this method improves the accuracy of answering questions and can effectively answer basic questions and answers.

This article implements a cultural relics assistant system for the above intelligent question answering model, and also adds functions such as voice question answering, daily cultural relics knowledge and museum information query. Natural language questions are answered accurately. At the same time, the designed system is tested in terms of function and performance. The results show that the system has good question stability and proves the feasibility of the above algorithm.

key words：Cultural relics, knowledge graph ,ontology intelligent, question answering ,entity recognition ,attribute link

目录

第一章 绪论

1.1 研究背景

文物是人类在社会活动中遗留下来的具有历史、艺术、科学价值的遗物和遗迹。它是人类宝贵的历史文化遗产。它承载的是上千年的灿烂文明，传承着中华优秀的历史文化，源远流长，生生不息。

随着《国家宝藏》《我在故宫修文物》等优秀的综艺节目播出，文物不再冷冰冰的躺在博物馆，对文物的了解再也不是长篇大论的文字表达，而是以一种通俗易懂，雅俗共赏的基调活跃在了大众面前，文物是点，历史是线，延展的文化是面，点线面精彩又立体，让观众充分领略中华文化的博大精深和美妙绝伦，摆除了文物枯燥无味的刻板印象，向大众普及了文物知识，同时也激发了大众了解文物的热情。快节奏的时代，生活的步伐不断加快，但是存于血液中的文化传承精神无论何时都不会忘记。

今人不见古时月，今月却可照古人。当我们讨论文物时我们在讨论什么？是信息，它是历史的实物见证，通过它，我们可以了解到那个时代的人们的生活习俗社会情况、环境背景、工艺技术、艺术审美，甚至是社会状况。如果摧毁了时代的成就与历史，那就跟从未存在过一样，浮尘而已。没有实物的见证，就算古人代代口口相传，记忆也会慢慢遗忘，况且口说无凭，外人无法信服，文物是一个国家文化自信的重要支撑，只有保护好文物，传播文物价值，才能真正实现文化自信。文物不只是物质的存在体，更是信息的综合体，文物从生产开始，到它的使用，埋藏着周遭环境和人群对它施加的影响，这些影响可以帮助我们更多的了解古代社会、人群乃至环境的情况，而这也是文物身上赋存的信息。这些信息肩负着延续中华优秀传统文化的历史根脉。

互联网的高速发展，给人们的生活带来非常多的便捷，越来越多的产品和信息进行数字化转型。虽然我们国家的文博资源十分丰富，但是真正用户使用的频率却并不高，具研究统计，我们国家的民众平均两年才进入博物馆一次，然而在处于欧美地区的国家，当地民众走进博物馆的次数平均达到三到五次。其中原因，一方面，社会的快速发展使大众的心态也逐渐日益浮躁，很难静下心来去专注一件事，更甚，随着手机应用端的蓬勃发展，大家的业余时间都花费在手机上，走进博物馆去感受文物的魅力成为一件大众兴趣度不高的事件;另一方面，博物馆的展陈方式略显呆板陈旧，同时游览博物馆时深奥晦涩的讲解让人感觉乏味单调。为了让博物馆更加“鲜活”起来，打破之前的刻板僵化印象，打开其中的“奇妙”，让文物真正活起来，同时是为了充分挖掘，阐释及传播文物价值，文博专家希望能够借助科技的手段，让文物资源活起来，推广文物知识，复原文物历史，不再

将文物简单的陈列在博物馆，而是以多种形式多方面展示文物特点，挖掘文物背后的历史，以生动的讲述形式和多样的展示方式，跟着文物感受时代的风云涌动，朝代的更替变化，探索未知的历史记忆。中央也一直高度重视文化工作的推进，致力提升民族的文化自信，号召响应，各大博物馆都在对文物资源进行数字化整理，文物与现代科技正在进行深度的合作与融合。

全国文物普查统计，我国目前有不可移动文物 766722 处，可移动文物数量过亿，文物藏品 4138.9 万件/套。且文物领域是一个丰富的大领域，其所涉及的学科门类堪称包罗万象，面对如此海量的文物信息，如何对文物信息组织管理是重要的课题。且文物资源并非扁平的信息，并非单独存在的个体，文物是时代历史的融合，反应了时代的社会生活，往往与相关的人物，事件，艺术紧密勾连，互相联系。

基于文物的属性和特点，采用文物知识图谱来管理组织海量的文物信息是比较好的选择。知识图谱是一种基于图结构的语义知识框架，能够将知识合理有序组织起来，表现形式为由边和节点构成的的语义网络图，具体的来说，图结构中的“节点”代表为实体，用关系的表现形式为图里的“边”。而实体一般指的是物理世界中的实体指的是现实世界中的事物比如人、地名、概念、药物、公司等，关系则用来表达不同实体之间的某种联系。近年来，随着互联网的快速发展，海量数据也不断产生，与此同时大规模分布式计算力随之提升，机器从感知智能向认知智能迁跃，知识图谱的应用越来越广泛，在金融、医疗、电商等各种领域都已大显身手，图谱技术成为“兵家必争”之地。不其然，在文物领域也已有涉足，利用知识图谱对文物知识进行梳理组织，形象展示文物与文物历史的联系及文物知识整体结构，但目前都是由各大博物馆构建的小规模的知识图谱，应用于展馆的陈列及数字化展示，文物相关技术理论研究并不深。

由于文物知识的丰富性，涉及到文物本体，人物，历史，艺术，自然等，在不同领域都有不同的角度可利用，文物知识图谱不仅仅可利用于展示陈列，实际应用还可大有作为，如构建文物的问答系统，在我们浏览博物馆时，遇到感兴趣的文物，可对系统进行提问，这种个人定制化的解决方案和交互式的体验，可大大满足游客的获取知识的需求。或用于优化搜索引擎，在对文物进行搜索时，能更加结构化的全面输出文物的信息，使用户快速方便获取自己想要的信息；更者可用于艺术创意设计方面，如很多文物的设计十分具有艺术价值，颜色，花纹及材质等十分考究，对设计者具有参考价值。

然而，在文物知识图谱领域，国内对这方面的研究并不多，目前主要的研究方式是各大博物馆与科技公司进行联合完成对文物知识的数字化工作和主题知识库的构建，并且应用于博物馆的展览陈列和一些数字化的展示。考虑到文物知

识的应用不仅仅只局限于博物馆的展陈，还可用于构建的面向游客的问答系统，可用于回答游客的提问，部署到应用端，可以让用户足不出户就可以随时了解文物的信息；或者用于搜索引擎的改进，在搜索文物相关信息时可直接将知识进行结构化的展示，让用户可以一目了然。并且由于文物领域的专业性，缺乏相关理论方法及成熟的技术，整理数据难度高，人工成本大，因此如何构建文物领域的知识图谱及部署应用，也是本文研究重点。

1.2 相关技术研究现状

1.2.1 语义网的研究现状

自 Web 的诞生，极大的改变了人们的生活，很多东西就直接能从网上获取，我们在网上购物，网上浏览新闻，娱乐方式也改变了，我们通过网络看电影，玩游戏，生活方方面面都被互联网所深深影响着，我们对互联网的依赖也越来越深，在网上发布信息的同时又搜索获取信息，这些信息像滚雪球，不断变大，最后如海量般潮水汹涌。面对如此海量的信息，如何准确全面快速获取到人们想要的信息成为巨大的挑战。传统的 Web 网络以网页文档为单位组织信息，这样的信息组织方式缺乏语义性，机器无法识别网页更细粒度的信息，无法挖掘其中信息的联系，使大量信息处于松散状态，不具备理解可读性。为了使机器能更加智能的理解网页的信息，让离散的信息相互连接，自动化处理集成多源数据，使整个网络互通互联，Berners-Lee 提出了语义网[1]。语义网的诞生使信息共享更加高效，机器之间的协同更加智能。与此同时，W3C 制定了一系列技术规范和发布相关的开发工具。基于语义的 Web 体系不断建设发展，在语义网的基础上，谷歌提出了知识图谱的概念。

语义体系结构如图所示，从底至上功能不断完备增强，其中核心构成是第三层的资源描述框架（RDF）以及第四层本体（Ontology）。

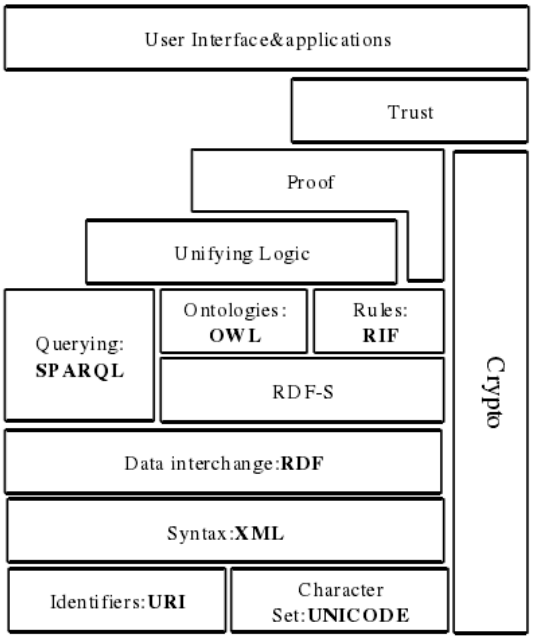


图 1 语义网层次结构

RDF 是语义网技术的基础，同时也是是语义网核心数据模型，在语义网中，所有数据都是以 RDF 格式表示，其十分具有灵活性。RDF 是一种图形结构，由一堆边相互连接节点组成，其中图的节点是椭圆形和矩形，边和节点都具有标签，标签为 URI 形式。在 RDF 中，有资源，字面量和空白这三种节点，资源节点是由椭圆形构成，任何有价值的信息可被成为资源，字面量是指属性值，用矩形表示，空白节点是指没有 URI 的资源。在使用过程中，RDF 引用三元组（主语，谓语，宾语）的形式表示。目前 RDF 格式主要有 RDF/XML,Turtle,Notation 3,N-Triple 及 JSON-LD 等[2]。

语义网中的本体的目的是明确定义，限定术语集合，使之规范表达，起到标准化作用，避免用多种表达方式指意同一个概念而产生的歧义和冗余，提高信息检索的准确度及快捷性。本体的目标是理解相关领域知识，制定领域内可广泛使用专业术语，其组织结构具层级结构化，规定类目的属性及之间的关系。具代表性的本体知识库有 WordNet,DBpedia，HowNet 等，常用的构建本体工具有 protégé[3]和 WebOnto 等。

1.2.2 知识图谱研究现状

在语义网的快速发展下，于2012年谷歌提出知识图谱的概念，并应用于优化搜索引擎，取得了良好的效果。在学业界和工业界都掀起一股热潮，并快速普及到各个领域。知识图谱基于图结构存储，用来描述现实世界的概念及相互关系。知识图谱通过对半结构或无结构数据进行处理加工并整合，组织成与RDF数据相似的三元组格式，表达实体与实体之间的关系，最后通过聚合大量知识，形成具

有层次结构语义知识库。知识图谱是语义网的实际应用，本质上是语义网的进一步的延伸和升华。

自谷歌发布知识图谱之后，知识图谱的概念越来越流行，很多公司也快速布局了知识图谱系统，将知识图谱作为基础架构的一部分，利用知识图谱技术来帮助自己保持领先地位。 知识图谱和图数据库在各种行业已大显身手，从银行，汽车行业，石油和天然气到制药和医疗保健，零售，出版，媒体等。尽管这些公司针对不同的领域使用了知识图谱，但最终的结果却是相同的：解决了数据孤岛问题，使大量的数据互相连接，使用并最终重用它们。近年来，知识图谱技术发展越来越成熟，很多公司将自己的研究成果公布出来供大众使用，如微软的Probase[4]、百度的“知心”[5]、搜狗知立方、YAGO[6]、wikidata[7]等通用领域的知识图谱。另外有很多学者在深入研究知识图谱技术，很多高校投入精力财力去研究知识图谱，如复旦大学的CN-DBpedia[8]，目前为中文最大的百科知识图谱，整理了来自百度百科，中文维基百科等来源的数据，经过数据清洗，抽取，知识融合等操作后，形成了高质量规范化且结构化的RDF格式数据，数据涵盖了众多领域，包括科技、教育、金融、娱乐、军事等多个领域，共包含了实体知识900多万，三元组条目8000多万，目前还在不断开发和建设中，现已广泛被各界研究者使用。Zhishi.mi是上海交通大学研究的国内最早的知识库系统，整合了中文百科、互动百科、中文维基百科三大百科的页面知识，利用制定的规则将结构化知识提取出来，对实体进行对齐融合后，将实体进行链接，共包含了超过了1000多万个实体和1个多亿三元组条目。使用者可用SPARQL进行查询，且结果以HTML的形式表示。清华大学以另一角度构建了知识图谱，考虑到其他语言的知识图谱与中文可进行知识共享，为了解决这个问题，清华大学构建了一个跨语言的大型中英文知识图谱XLORE，它分别从中英文的网页百科信息进行提取结构化知识，对齐中英文的等价实体，挖掘实体之间更深的联系，它共包含了1000多万个中英文的双语实体，60多万双语概念和5万多属性。

中文知识图谱不仅在通用领域有很大的发展，在很多垂直领域也受到广泛关注。在医学领域，孟祥龙等人[12]面向中药炮制构建了知识图谱并进行了可视化，在法律领域方面，H.Lian等人[9]构建了一个基于法律专业知识和社会媒体公众认知之间的联系的知识图谱，挖掘并展示了法律知识及与公众社会认知心理之间的关系。在农业领域，李嘉锐等人[10]提出了基于水稻本体的构建框架，采用了神经网络对方法对水稻实体进行半自动抽取，能够有效地提高本体构建效率与质量。另外余菜花[13]在中国低碳研究领域、程赛琰[14]电子政务研究领域、谢靖[15]文学学科领域、李伟平[16]在体育领域、辛伟[17]在军事心理学领域都利用了知识图谱对学科进行研究，都产生了非常好的效果。李明鑫等人[11]利用CNKI上的论文（截至2015年）研究分析了知识图谱论文在各个学科的分布情况，图为统计结果，通过图可以看出知识图谱在各个领域都有着广泛的使用。

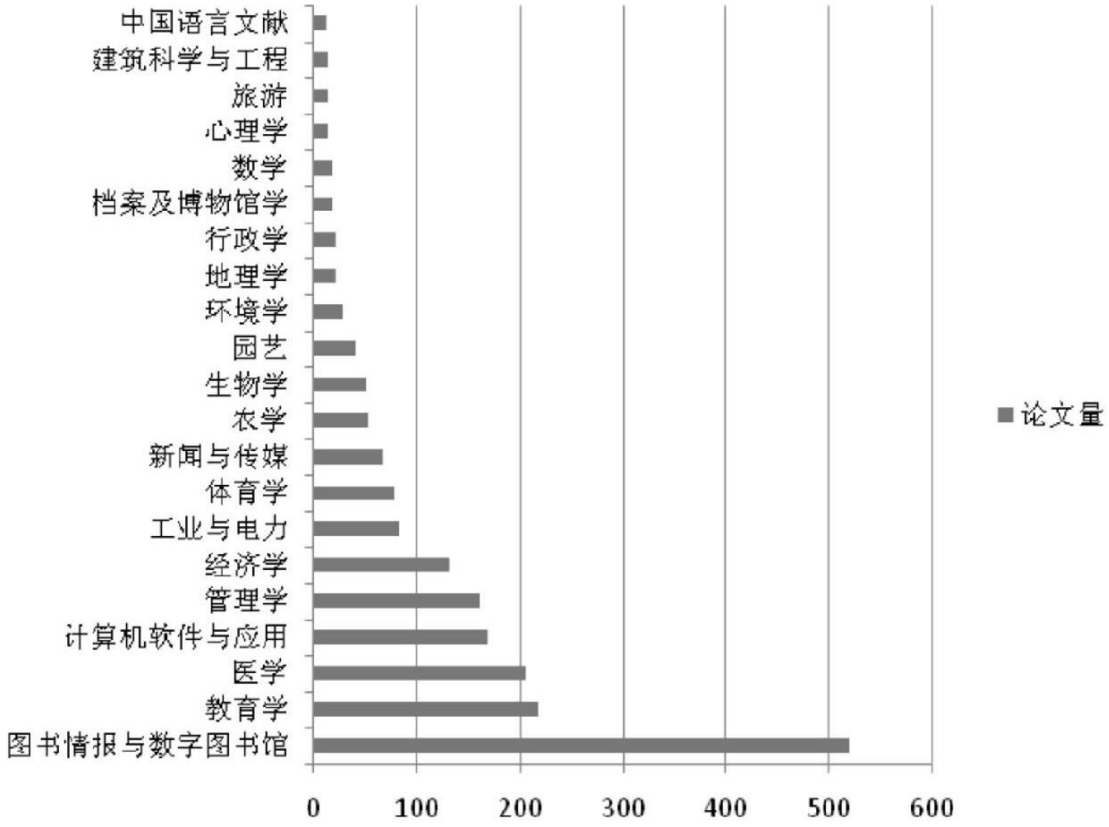


图2 CNKI知识图谱论文在各个学科领域的分布情况

1.2.3 文物领域知识图谱研究现状

随着我国文化事业，文物保护及传播事业的不断发展，国民对于文物的兴趣也逐渐高涨，如何处理海量的多源异构的文物信息成为研究重点。

在2006年Berners-Lee提出了链接开放数据概念，在此启发下，文化领域的数字资源及知识图谱构建开始进行了深入研究[18]。随着知识图谱相关的技术不断

成熟，文物领域的知识图谱也不断涌现，世界各国出现了具有代表性的文物知识图谱项目。文物知识图谱大部分使在知识图谱元数据模型上进行构建，主要有：DC[19],DC terms[20],SKOS[21],CIDOC-CRM[22],FRBR[23],EDM[24] 等 。 其 中CIDOC-CRM模型和EDM模型在国外较为广泛使用。

CRM（Conceptual Reference Model）是文化遗产信息领域的本体概念模型。由国际博协登记著录委员会（CIDOC）提出，耗费多年时间建设，梳理、集成和融合了多源异构的文物信息，构成了规范可扩展的本体知识库，使文物信息领域中的概念明确定义且可共享，描述了文物本体之间相互关系。该模型提供了一个通用的语义框架，在这个框架体系下，任何文物信息可相互映射，可对其进行扩展，并可以定义的概念和结构对文物本体和对应的关系进行描述。CIDOC-CRM模型目前共定义了90个实体，149种属性，其中包含了特定文物实体的时间，地点，人物，历史事件，人文艺术等信息。目前有很多国家的博物馆使用CIDOC-CRM模型，将博物馆的数据映射该模型。例如有大英博物馆的开放数据项目[25]，俄罗斯文化遗迹云平台[26]，波兰数字国家博物馆，徐悲鸿博物馆绘画藏品项目[27]等。在2011年大英博物馆基于CIDOC-CRM本体模型构建了知识图谱项目，共包含了1亿左右数据。并在此基础上构建了语义检索系统，与世界范围内的知识图谱相连接，同时大英博物馆还推出了虚拟研究环境ResearchSpace,作为艺术研究课题，并发布了2500多条三元组条目数据供进行研究。

Europeana是由欧盟组织联合发起的数字智慧博物馆项目，共27个国家参与其中，将各个国家的博物馆，图书馆和档案馆的数据资源进行了融合和整理，构建了大规模的语义知识本体模型，共包含了千万文物本体和几百万的外部连接及两亿多数据记录，这些资源类型具多样化，有文本、图片、音频、视频、动画等形式，其中，该项目的文物资源以EDM元数据模型进行描述，并提供了对外开放的数据接口。以EDM元数据模型为基础的知识图谱项目有：荷兰博物馆知识图谱项目[28]、西班牙文博物馆数字化项目[29]等。

然而，在文物领域，知识图谱在国内的研究并不多。浙江大学的林炀平[30]构建了面向创意设计的文物知识图谱，但是他提出的在本体构建的方法是以全手工的方式构建的，对文物知识没有使用自动化的抽取方法，人工成本大。西北大学的邱超[31]提出了一种面向文物领域的知识图谱自动化构建方案，并提出了结合规则和ELM的文物知识抽取算法，但是构建的知识本体简单，很多文物信息没有融入本体中，如文物纹饰、文物寓意、制作工艺、使用材料等信息都忽略掉了，并且采用基于规则的方法自动抽取，对问句的语义分析不够深刻，覆盖度不全，很容易漏掉一些语义通常出现但是问法不常见的问句。浙江大学的张娜[32]也是主要研究自动化构建文物知识图谱方案，她提出了一种基于半监督学习的文物关

系自动抽取算法，选择Tri-traning改进算法完成三元组的抽取工作，通过先对每种关系对文本数据进行人工标注，获得少量种子模板然后对模型进行协同训练，然后通过协同训练扩展关系种子模板，从而完成抽取工作。这种方法依赖于专家定义的关系模式，且存在部分尚未发现的文物关系。面向具体应用构建的知识图谱有天津大学张万如等人[33]构建的“养心殿知识图谱”，主要为故宫92周年“发现·养心殿”为主题用于数字体验的展览而构建的。万达信息科技联合上海博物馆，以“董其昌数字人文”为主题构建了书画专题的知识图谱[34]。总体来看，在文物领域的研究项目并不多，构建的知识图谱大多是小规模且面向展示应用的，总而言之，文物领域的研究还需要向更深更广的层面进行。

1.2.4 知识图谱的应用

在过去的十年左右的时间里，知识图谱早已潜入我们的日常生活中，无论是通过语音助手（例如Siri或小米小爱），直观的搜索结果，还是通过在线商店推荐者的个性化购物体验，我们每天都在不断与知识图谱进行交互。知识图谱在不同应用场景下发挥重要的作用，在语义搜索情境下，知识图谱可以提供高质量的结构化背景知识和常识知识，使用户在搜索时能更加快速且准确全面获取到想要的信息，如图搜索韩红，与韩红相关的网页在左侧，右侧还会给出韩红全面的个人信息，返回这种结构化的知识卡片，增强了用户的搜索体验。这种基于语义的搜索方式相比之间传统的基于关键字的方式，查询结果更加简洁更智能化，有效地提高了搜索结果的质量。智能问答也是一个重要的应用，基于知识图谱的问答系统可以用准确简介的自然语言回答用户的问题，大大减少了人工成本，提高效率。

图3 知识图谱在搜索引擎的应用

除了上述的领域外，知识图谱在其他领域也有十分重要的场景落地实现。比如在推荐领域，将知识图谱存在的知识三元组信息作为推荐系统外部知识引入，在冷启动或者用户信息稀疏的情况下可以把知识图谱中的信息作为辅助信息加入其中，从而缓解这些难题。应用方式主要三种：依次学习、联合学习、交替学习。

知识图谱还可以在应用在金融领域中，如使用于反欺诈的场景中。由于在欺诈案件中通常是会卷入比较复杂的关系网络中，而且涉及多种不同的数据源，引入知识图谱可以将多种数据源进行融合并且将其涉及到的关系进行梳理，并且在领域专家的指导下制定相关规则对数据进行判断是否具有一致性，可以识别出关系网络中潜在的欺诈风险。

知识图谱还可用于股票情报分析，通过将公司年报、公司公告、券商的研究报告、新闻报告等多种非结构文本信息作为构建知识图谱的数据源，然后利用知识图谱的相关技术提取出公司的股东、客户、子公司、供应商、合作伙伴和竞争对手等信息，然后构建基于公司的知识图谱。在发生某个经济事件时，通过该图谱可以辅助用户做出更好的投资决策。

由于知识图谱具有较强的推理功能，可以挖掘出知识之间隐藏的联系，且具有知识可解释性，能够输出更加全面的知识的优点，知识图谱的应用越来越广泛，

在各大领域都有其身影，特别是在工业界，很多公司都开始部署了自己的知识图谱系统，支撑业务场景使之发挥更好的性能。相信在未来，知识图谱会成为更炙手火热的研究工具。

1.3 本文研究内容

本文主要研究了如何将文物的基本信息和蕴含的历史、文化、艺术元素，表达成一个知识图谱并在此基础构建一个可供实际使用的文物智能问答系统，具体研究内容如下：

（1）研究面向文物领域的本体构建方法，提出了一种结合了七步法和循环获取法的文物领域本体构建方法，加入了迭代优化的思想，最后，我们通过该本体构建方法完成了文物领域本体模型，为之后构建知识图谱奠定知识基础。

（2）研究文物领域的知识图谱构建与展示方案。提出了基于特征词集的文物属性知识抽取方法将web非结构化文本中具有的属性知识抽取出来，完成本体的实例化。并将本体知识库存到Neo4j图数据库中，可实现文物知识的可视化和查询检索工作。

（3）研究并设计了基于文物领域知识图谱的智能问答模型。提出了基于BERT模型的问句意图分类，使用基于Bi-LSTM模型和字典匹配模型的结合完成了槽位提取功能，最后将问句的意图和属性信息进行转化为SPARQL语句，将其送到基于Apache Fuseki搭建的SPARQL查询检索API中，完成对问句的解析返回最终答案。

（4）在上述的智能问答模型的基础上，本文构建了一个文物助手系统，还增加了语音问答、每日文物知识和博物馆资讯查询的功能，以微信公众号的展示形式为用户提供服务。

1.4 本文组织结构

第一章 绪论。本章主要介绍本文的研究背景及研究目的，并阐述了知识图谱相关技术研究现状，整理了本文的论述结构。

第二章 文物领域的本体模型构建。本章详细阐述了本体相关理论及构建知识图谱流程，介绍了知识图谱的存储与可视化技术，最后详细描述了智能问答相关技术和目前研究现状，然后基于文物领域的特点以及实际应用场景的考量，结合本体构建方法七步法和循环获取法，并详细介绍了本体构建每一步骤，利用protégé建模工具，用OWL本体描述语言对本体进行描述，最后完成了本体模型的构建，为之后知识图谱的构建提供基础。

第三章 文物领域的知识图谱构建。本章详细阐述了文物知识图谱的构建流程。首先介绍了本章知识图谱构建过程中所用到的相关技术和理论知识，详细阐述了实体抽取、属性抽取的相关方法和综述研究，并介绍了知识图谱的存储和可视化技术。然后获取文物领域相关数据并对数据进行整理分析，针对文物文本的属性抽取提出了基于特征集自动属性抽取方法来抽取web文本中属性信息，最终完成了本体的实例化工作，然后将整理好的本体知识库存储到Neo4j图数据库中，并实现文物知识的可视化。

第四章 基于文物知识图谱的智能问答模型构建。本章设计了基于文物知识图谱的智能问答模型框架，将整体框架分为三个部分：问句理解模块、问句转换模块和答案检索模块。其中，提出了基于BERT的意图识别和基于Bi-LSTM与字典规则结合的属性抽取模型，完成了问句理解，并送入到问句转化模块中得到SPARQL语句，然后在知识图谱中进行查询检索。

第五章 文物助手系统的设计与实现。该系统基于上章构建的智能问答模型提供了文物问答服务，并且根据用户需求，还增加了语音问答、每日文物知识、博物馆资讯查询等服务。

第六章 总结与展望。总结本文工作，并对本文实现过程中的难点与不足之处进行复盘分析，提出改进方法，并对知识图谱应用趋势进行展望。

第二章 文物领域本体构建

2.1 引言

中华文化历史悠久，蕴含着上下五千年的文化内涵，从古至今遗留下的文物是我们历史长河中的见证，文物的保护与传承具有重大的意义，对历史，文化，艺术都有巨大的研究价值，也是展现我们文化自信的证据。文物资源丰富，知识体系结构复杂，涉及领域范围广泛，包含人物，历史，文化，艺术，自然等学科。目前，文物领域的数据已逐步实现了数字化整理，但这些数据多以文物本体的描述信息进行整理，缺乏文物蕴含的艺术文化内涵表示，数据的规范性不够严谨，尚未形成整体的知识体系结构。本体是一个知识共享的概念体系模型，可对本体进行明确的概念描述说明及定义和约束本体之间的联系，对领域内的知识进行整合和规范化表达，如何设计一个合理的文物领域内的本体结构模型是本章研究重点。

本章主要研究文物领域下的本体构建，展现该领域下概念体系。本章首先分析了文物领域的知识特征，根据其特征结构，选择合适的本体描述语言和建模工具；然后，介绍了文物领域下的本体模型构建方法及构建的技术，详细阐述了构建本体的每一步骤，最后完成文物领域的本体模型。

2.2 本体的定义

本体（Ontology）是指一种概念化的规范[35]。”本体”这个词的使用在哲学领域内有较长的历史，是指对存在的系统描述。但在有知识共享的语境限制的上下文中，与哲学领域的含义是不同的。本体这一术语来是对概念化的明确说明。也就是说，本体是物理世界及物理世界可能存在的概念及关系的描述。本体作为一种规范机制，最重要的用途是实现对知识的共享与重用，能够在领域内对理论知识进行协定与描述概念。

本体的构成要素主要有：

（1）类：也称概念，是指类对象的基本类型。

（2）属性：对象所具有的特征或参数。

（3）关系：类与个体之间所具有的联系，关联的方式。

（4）函数术语：在声明语句当中，可用来代替具体术语的特定关系所构成的复杂结构。

（5）约束：采取形式化方式所声明的，关于接受某项断言作为输入而必须成立的情况的描述。

（6）规则：用于描述可以依据特定形式的某项断言所能够得出的逻辑推论的，if-then（前因－后果）式语句形式的声明。

（7）公理：代表永真断言，是指不用证明的基本事实。

（8）实例：也称作个体，是指某个类或者是概念对应的具体对象

2.3 本体构建方法

2.3.1 以往构建方法介绍

由于本体专业领域性强的特点，使得本体的构建成为一项复杂的系统工程，构建过程中，耗费周期长，需要有对领域知识有一定研究的专业性人才及采用智能协同的工具完成，严重依赖人力劳动，自动化难度较高。因此目前本体构建大部分基于特定领域构建且大多是手工构建。

本体构建尚未制定统一的标准和原则，目前大部分是基于 Gurber 学者于 1995 年提出的五项基本原则：

（1）明确性：对本体的描述及涉及到的概念应当尽量明确、清晰、完整及规范。

（2）一致性：制定的本体逻辑推理规则应当是严谨的，对本体进行逻辑推理的结果应当是正确的。

（3）可扩展性：本体的设计应当是随着领域的发展的变化而可扩充或删减修改的。

（4）最小的编码偏差：本体的设计应当是由于的符号的编码的不同而改变的。

（5）最小的本体承诺：本体构建的覆盖范围应当是合理的，范围过大，使得本体的概念指意不明，容易出现歧义。构建的本体应当是在特定领域满足知识共享的原则。

在不同领域内，由于应用场景的不同，在实际构建的过程中侧重点也不同，可根据特定领域的结构进行调整。随着语义网的快速发展，本体的研究也渐渐成为热点，如何构建比较权威的本体模型成为国内外研究的重点。目前，已有很多学者推出了自己的研究成果，比较成熟及应用广泛的本体构建方法有：如 METHONTOLOGY、骨架法、KACTUS 工程法、SENSUS 法、IDEF- 5 方法、七步法等。

领域本体的构建是一个比较复杂的过程，在上一节我们已经对文物知识的特点进行了分析，本节主要介绍本体构建的使用的描述语言、构建的工具和构建的方法。

本体采用本体语言来进行描述的，本体语言是一种来对本体进行编制的形式化语言。目前本体语言发展出了多种类型，如Cyc、Gellish、IDEF5、知识交换格式（Knowledge Interchange Format，KIF）、RDF/RDFS、OWL等。本文调研了不同本体建模语言，比较了他们之间的建模能力，如下表所示。从比较可以看出OWL建模能力最强，同时OWL语言基于W3C标准制定的语言，应用场景更为广泛，兼容性更强。因此本文选OWL语言来对本体进行构建，保证文物知识的可重用性。

在本体构建工具方面，我们使用了斯坦福大学开发的protégé5.5.0本体编辑工具来进行构建，在2.1.2节我们介绍了protégé的基础功能和优点。在本文采用protégé主要是由于它的图形化的编辑界面，极大方便和简单的进行操作，且支持关系数据库的连接存储，功能齐全。再者protégé 的应用广泛，有丰富的相关学习技术资料可供参考，在protégé的支持下可以很好的完成文物领域的本体构建工作。

自动构建的本体噪声较大，本文采用人工构建。本体构建的方法比较通用常见的有斯坦福医学院的七步法、骨架法、循环获取法、TOVE企业建模法和METHONLOGY法等。七步法即本体的构建过程是由七个步骤组成的，该方法整体构建的思路清晰，逻辑清楚。具体步骤如下：确定本体的专业领域及范围；调研相关领域是否有本体构建，考虑是否可复用本体；列出该领域本体重要的专业术语；定义类与类之间的层次关系，构建等级体系；定义类的属性；定义属性的约束或类型；创建类的实例。但七步法也存在如下缺陷：在构建过程中没有考虑评估本体的质量；在文物领域术语也可能来源于实例，在列出领域内的术语前，应当对领域内的信息进行收集；缺少了对本体的更新和迭代过程。循环获取法则是使用了环形结构来获取本体的方法，包括数据获取、概念学习、领域聚焦、关系学习和评价等步骤，但是该方法没有详细说明步骤的具体过程，比较空泛，没有对本体进行设计着一环节，且没有考虑对本体的重用。这两种本体构建方法都缺乏对知识图谱应用场景的考虑，基于七步法和循环获取法都各存在利弊，本文将针对文物领域知识图谱场景提出构建方法。

2.3.2 基于知识图谱场景应用的文物领域本体构建方法

针对上述本体构建方法存在的弊端，本文结合了七步法和循环获取法，并在此基础上对文物领域进行了改进，提出了基于知识图谱应用场景的文物领域本体构建方法。具体构建流程如下图。本文提出了文物领域的改进方法，首先我们根据文物领域确定了本体的范围，然后对文物领域方向进行文献阅读及调研，针对本文基于智能问答的实际应用场景抽象出该领域的专业术语和概念，采用自顶向下的方法，从应用的场景及领域的数据来这两个方面考虑对本体进行修正和设计，这样能更好的完善本体且覆盖应用需求。接下来，定义本体的概念和进行精简，再考虑是否可进行本体复用，设计本体层次结构，然后对本体进行实例化操作，最后进行本体的迭代优化。

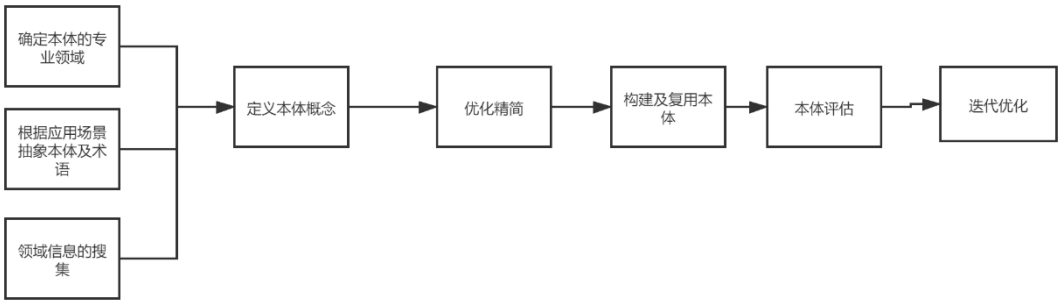


图4 本体的构建流程

下面我们针对上述的文物领域本体构建步骤进行详细说明：

1）确定本体的领域及范围，对本体进行初步抽象

明确本体构建的领域范围，如果范围过大，导致包含的知识过于繁杂，整理和构建难度大。范围过小，限制了知识或数据的获取，构建的本体不丰富和全面，无法覆盖该领域的知识。因此在构建本体之前，需确定本体的领域的范围，考虑之后的应用场景，明确构建的侧重点。

2）对本体领域进行充分调研及信息搜集

为了了解本体领域的知识体系，构建之前需对该领域进行充分的调研和广泛的搜集可供参考的领域信息。对于文物领域，信息的搜集来源可着重关注各地的各大博物馆的网站，博物馆就是对文物进行收集及展览，各地的博物馆网站有着相当丰富的文物数据，同时对文物数据进行了细致及结构化的整理。除此之外，可在百度百科页面对文物信息进行相关的补充。

3）根据实际应用场景，定义本体概念及术语

传统的本体构建方法没有考虑到最终的使用场景，缺乏针对性，最终构建的本体往往需要再次迭代更新。本文采用自顶向下及底层数据两个方面相结合，既

针对最终的应用场景进行本体优化，也根据底层的数据来源来进行本体设计，这样设计的本体模型可以更加契合且实际操作过程比较省时省力。在根据实际的应用场景基础下和前面充分的信息收集，对提炼出本体的术语和概念。这步主要是对抽象出的本体概念进行明确定义，且枚举出领域范围内的专业术语，对术语及概念进行枚举能进一步加深我们对本体的理解，同时选择合适的本体描述语言对本体进行描述。另外，这一步在枚举时应当尽可能对领域进行做到全覆盖，对领域内的知识做到尽可能的丰富。

4）精简优化本体概念，定义本体的属性

这步主要是将上一步中定义或者丰富的本体概念完成精简和优化工作，在上步中，我们目标是尽可能覆盖领域内的知识，本体的定义或者概念比较冗余和繁杂，这样直接定义的本体概念不能直接应用于最终的构建过程中，我们需要结合实际的应用场景对本体进行进一步的精简和优化，使本体的定义更加明确精炼。最后定义本体的属性。

5）构建本体及考虑对本体的复用

选择合适的本体构建工具对本体模型进行建模。同时，这步需要考虑是否有相关领域内的本体可以进行复用，并采用合适的知识融合方法对将异源知识进行合并。

6）评估本体

此步骤旨在检验通过上述步骤构建的本体是否准确及有效。传统的本体评估方法主要从本体的丰富情况、本体的结构、逻辑是否正确等几个方面来进行评估，这种方法仅关注了知识层面的评估量度，没有验证最终的实际应用场景的有效性。这步同时结合了本文构建基于知识图谱的问答的应用场景来进行本体的评估，使本体构建更加准确有效。

7）对本体进行迭代优化

本体的构建是一项复杂的工程，构建完成后还需要从多个方面来对本体进行观察，根据实际的反馈来对本体进行调整更新和迭代优化。

2.4 本体构建工具

工欲善其事，必先利其器。本体的构建是一项复杂繁琐的工作，需要大量人力来完成，为了能够更加智能化的完成构建工程，一个好的构建工具是十分必要的。目前国内外对于本体的自动构建也倾注了大量心血研究。本体工具主要包括本体编辑管理工具、本体解析工具和推理机。目前比较典型的构建工具有：protégé、

KAON、Ontolingua、Apollo 等。其中因 protégé免费开源且易扩展易操作的特点，应用最广泛。Protégé是由斯坦福医学院开发，主要功能由本体编辑，本体可视化等。本文也选用 protégé作为本体构建工具，将对其进行详细介绍。

Protégé是由 Java 开发的免费开源的本体编辑工具，成为最好用的本体编辑工具之一，得益于由以下特点：具有图形化界面，便于操作；支持多插件，扩展性强；推理机制完善，内置多种推理机且可进行扩展；支持多类型本体语言，如 OWL、RDF、RDP 等。

Jena 是由 HP Labs 用 Java 开发的开源免费的框架[36]，主要用于对建设语义网提供技术支持。Jena 支持 RDF 数据格式，集成 API 可对数据进行存储，添加、删除、推理及 SPARQL 查询， 方便用户的使用。Jena 框架架构图如图所示，主要分为 4 个部分。1.RDF 模块。为核心模块，对 RDF 数据进行处理解析工作。2.SPARQL 查询模块。提供接口，使用 SPARQL 语句对数据进行查询和管理。3 推理模块。使用推理机对数据进行逻辑推理。4.TDB 数据存储模块。提供连接数据库如 MYSQL 的接口，对数据进行高效存储。Jena 可在官网下载（https://jena.apache.org/download/index.cgi），需同时下载 apache-jena 和 apache-jena-fuseki 两个包。

2.5 本体构建

2.5.1 知识分析与提取

文物是在历史进程中由人类创造的遗存物，具有历史、艺术、科学等价值的文化产物。文物在不同的历史环境下人类社会生产的物品，以不同形式保留至今。文物信息具有复杂性，表现在：产生时代不同，所属文化不同，且种类繁杂，质地不一，功能各异。以质地特性而言，就有陶瓷器、青铜器、玉石器，金银锡器等。虽然文物信息复杂，但是，是具有可分性的，可根据文物以下的维度对文物进行分类：产生年代、产生地点、文物的材质、文物的用途、文物的形状等。

通过分析文物的定义及调研文物相关的文献，我们将对从文物本身的属性信息和文物蕴含的信息这两个方面来对一件文物进行描述定义。

文物本身的属性信息，指文物的名称，产生时期，出土地点，体积大小等基本信息。这些信息是文物本身所赋予的基础数据，也是对文物研究的基本入口。对一件文物进行知识分析，首要关注的是基础信息，继而对文物蕴含的其他信息进行挖掘和研究。

文物蕴含的信息，是指通过从文化、历史、艺术等不同维度对文物研究主观进行获取的信息。对于一件文物来说，我们将从不同维度蕴含的信息映射到一件

文物本身上，主要概括了以下了几个描述要素：文物造型要素、制作工艺要素、历史文化要素、功能要素、象征意义要素等，具体解释如下表：

表1 文物描述要素

文物蕴含信息	
文物描述要素信息	举例
造型要素	文物的形状，纹饰和色彩
制作工艺要素	文物制作采用工艺手段，如彩绘，釉下彩
历史文化要素	文物背后的历史故事
功能要素	文物的用途，如酒器、乐器
象征意义要素	文物蕴含的抽象象征意义，如吉祥、富贵等

基于上述的分析，本文文物本体的构建将从这两个方面展开进行，能够比较充分的对文物本体进行描述。本文构建将以故宫博物馆公开数据为主，同时参考各个博物馆整理的信息进行整理。

2.5.1 本体构建过程

本小节主要是根据2.3.1节中所提出的基于知识图谱应用场景的文物领域本体构建方法来完成文物领域的本体构建，基于OWL本体描述语言，采用protégé工具来进行构建工作，主要的构建过程如下所示。

1. 明确本体的专业领域和范围，对本体进行初步抽象。本文研究的是文物领域的本体构建，所以主要的范围应该限制在文物领域范围内，本体的核心是文物本体。主要目的是构建基于知识图谱的文物领域的问答系统，有效的整合文物资源，建立合理的知识架构，为系统更好的服务。

2. 对本体领域进行充分调研及信息搜集。本文主要基于故宫博物院的数据，选取了陶瓷，青铜器，金银器，玻璃器，珐琅器，玉石器这六大类的文物，初步获取到共2496例文物实例，其中陶瓷类1368例，青铜器类264例，金银器类192例，玻璃器36例，珐琅器144例，玉石器492例。

3. 根据实际应用场景，定义本体概念及枚举重要术语。根据搜集到的文物领域的相关数据，并在相关文献的指导下，我们对文物信息进行了分析和整理，列出了相关的领域专业术语，定义本体概念之间的相互联系及整理本体之间的层级结构。我们将文物按照故宫博物院的分类体系，分成陶瓷类、金银器类、玻璃器类等，同时增加文物的基础信息概念，如朝代、体积大小、出土时间、出土地点等，另外对于文物蕴含的信息的概念也进行扩展，如纹饰、工艺、象征意义等。同时列举出尽可能覆盖全的专业术语，具体的情况可见表2。

表2 文物部分相关术语举例

术语类别	部分术语举例
文物类别	金器、银器、玉器、玻璃器、陶瓷器、漆器、珐琅类、雕塑类、钟表类、文房类、织绣类、首饰类、青铜器 •••
朝代	旧石器时代、新石器时代、夏、商、周、西周、春秋、战国、秦、汉、三国、晋、南北朝、隋、唐、五代、宋、辽、元、明、清、民国 •••
纹饰	弦纹、鱼纹、瓦纹、雪花纹、勾莲纹、垂环纹、背纹、松梅纹、漩涡纹 •••
器型	碗、盆、枕、杯、瓶、盒、砚台、琴、刀、剑、桌、椅、灯、炉、壶、尊、斧、矛 •••
工艺	刻花、剔花、划花、拍花、印花、贴花、起线、画花、补花、戳纱、拍纱、钉珠、起绒、烂花 •••
• • • • • •	• • • • • •

4. 精简优化本体概念，定义本体的类及类的层次结构。在对文物领域的术语进行枚举时，为了尽可能覆盖整个领域知识，定义了相关本体的子类。如陶类、瓷类为陶瓷类的子类，清乾隆、清康熙、清嘉庆等为清朝时代的子类。我们根据实际应用的考察，和用户在实际应用中对文物的搜索情况，用户对文物的细分类别并不太关注，所以，这步我们将文物类中的部分子类取消，融合成一个大类，对本体进行精简优化。

接下来，我们将精简优化好的本体或本体领域相关的术语，整理成本体类相关的词语，并按照树的结构定义本体之间的层级关系，如子类关系、并列关系等。类层次定义可根据自顶向下的方法，先定义好顶级的父类，然后细化相关的子类。也可先从底层数据进行概括，抽象出父类。在实际应用过程中，这两种方法可进行混合使用，可先定义比较明确清晰的层次，然后向上或向下进行扩展。

5.定义类的属性及属性约束。根据第4个步骤中整理出的本体类，还只是概念，描述一个领域还尚有不足，我们还需要定义类的属性来对本体进行全面的描述。本体中属性可定义为数据属性和对象属性，数据属性为本体的固有的特征，这类属性具有传递性，父类及子类均可对该类属性进行继承，如“文物”类的“名称”这个属性，其子类“陶瓷”也具有“名称”这个属性。对象属性则是描述了类与类之间的关系，例如“象征意义”这个属性，用于关联“文物”类与“寓意”类，描述某个文物可以蕴含着什么象征意义。属性编辑操作页面如下所示。

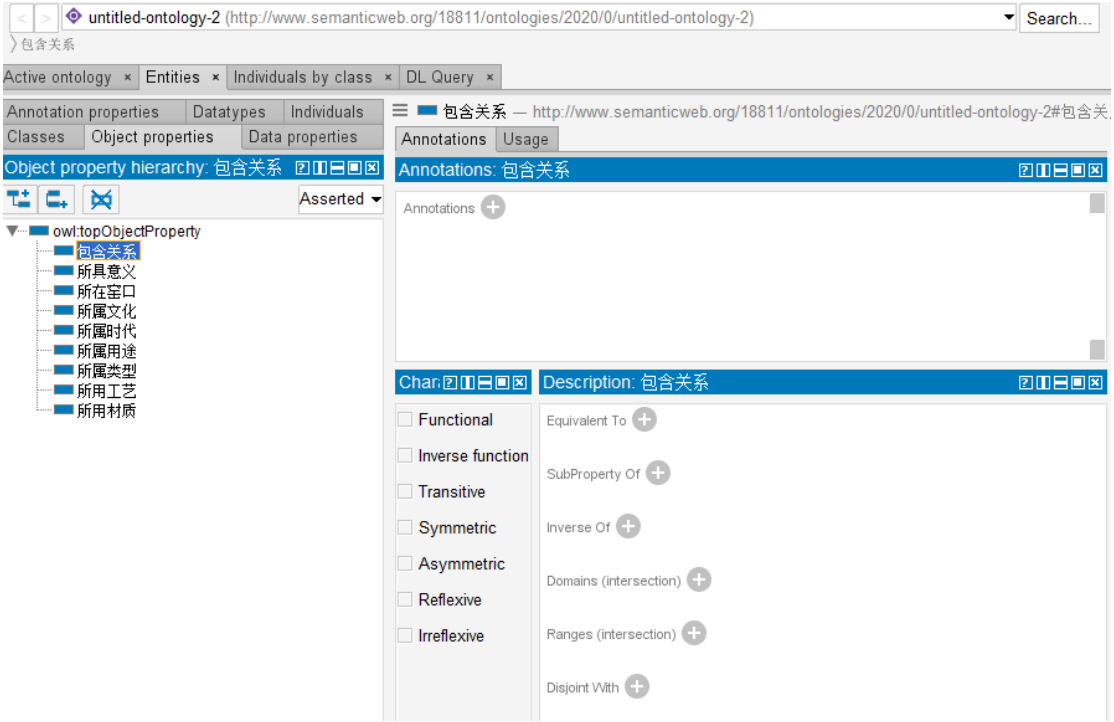


图5 poretege属性编辑页面

6. 评估本体。本文评估结合两种方法，其一采用了Jena推理机，通过对Jena中定义的描述逻辑，来判断构建文物中是否存在逻辑错误，然后进行本体实例化，判断是否符合实际应用需求。其二方法就是在国家博物馆采样1000条文物样本信息，从中抽象出关键词，然后与上述步骤构建的本体进行比对，判断设计的本体概念及定义的属性是否包括在采用的语料中，如果可以的话，说明本体的设计是可以覆盖文物领域的。本文经过比对，采样的数据在90%的场景可以被知识图谱所涵盖，本文建立的本体模型基本符合玉器效果。

7. 本体的迭代优化。经过上述的步骤，本体的构建基本构成，但是文物的内容不会一成不变，通过不断的更新获取新的文物知识，我们也相对应地对本体进行迭代更新。

通过上述七个步骤，至此我们完成了基于知识图谱的本体模型构建，最后文件以OWL文件格式保存，本文将每个步骤都以详细的进行了描述，对之后的研究具有较好的借鉴意义。

2.6 本章小结

本章详细阐述了文物本体的构建流程。首先是对本章所用到的相关技术详细说明，然后结合了课题背景对文物领域的知识进行了了解和分析，从文物本身的属性信息和文物蕴含的信息这两方面解析，为后续的构建本体奠定理论基础。然后提出了一种结合了七步法和循环获取法的文物领域本体构建方法，加入了迭代

优化的思想，考虑了本体构建的复杂性，将七步法细化成每一步具体的实现方法，通过该方法构建的本体，可以提升本体的完整性，通过protégé工具构建的半自动方法也提升了本体构建的效率。最后，我们通过结合七步法和循环获取法的本体构建方法完成了文物领域本体模型，为之后构建知识图谱奠定知识基础。

第三章 文物领域知识图谱构建

3.1 引言

知识抽取主要包括实体识别和属性、关系抽取。知识图谱的构建流程如下图所示。在第二章中我们已完成了本体模型的构建，本章主要集中于知识抽取和图谱生成这两个步骤。此外，构建一个完整的知识图谱，还应考虑到知识图谱的存储于可视化步骤，在构建文物领域的知识图谱过程中，很多通用领域的知识图谱构建方法也可借鉴过来，下面本节将构建过程中所用到的关键技术进行简单介绍，这部分内容是后续工作的研究基础。

3.2 本章相关技术

知识图谱构建主要是由本体模型构建、知识抽取、图谱生成这三个部分构成，其中知识抽取主要包括实体识别和属性、关系抽取。知识图谱的构建流程如下图所示。在第二章中我们已完成了本体模型的构建，本章主要集中于知识抽取和图谱生成这两个步骤。此外，构建一个完整的知识图谱，还应考虑到知识图谱的存储于可视化步骤，在构建文物领域的知识图谱过程中，很多通用领域的知识图谱构建方法也可借鉴过来，下面本节将构建过程中所用到的关键技术进行简单介绍，这部分内容是后续工作的研究基础。

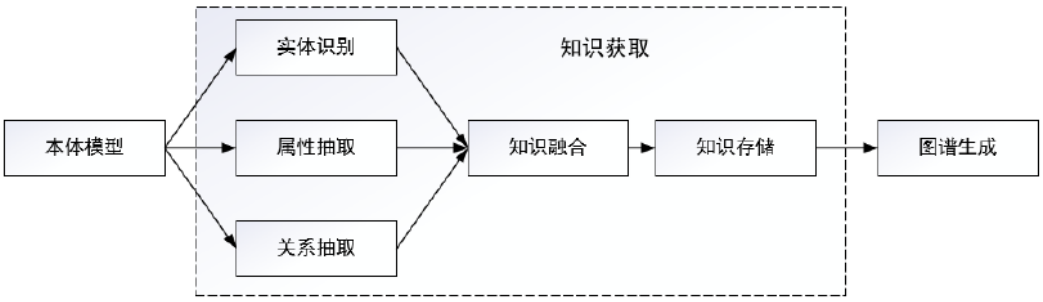


图 6 知识图谱构建流程

3.2.1 实体抽取技术

实体抽取也叫作命名实体识别技术，是指从数据源中自动识别或抽取命名实体出来。比较常见的实体抽取算法有：基于规则和词典的实体抽取方法、基于机器学习实体抽取方法以及基于深度学习实体抽取方法。

近年来，X.Wang 等人[37]面向细菌实体的命名实体识别提出了基于微生物

词典和命名规则的抽取方法，识别出文本中的细菌实体。W.Sun 等人[38]基于改进的条件随机场（Conditional Random Fields, 简称 CRF）方法识别出明信片中的地址实体。L.Zhang 等人[39]对中文微博上的语料进行了命名实体识别，使用了卷积神经网络的方法识别出微博语料中的人名、地名、机构名等相关命名实体，这种基于深度学习方法的实体抽取比规则或者基于统计的方法人工干预较少，实验效果也很好，从此基于深度学习方法的命名实体识别成为热潮。X.Wang 等人[40]也同样使用了基于深度学习和神经网络面向军事领域提出了实体识别方法，他们使用了双向长短时记忆神经网络（Bi-LSTM）和条件随机场（CRF）相结合的模型进行了实体抽取工作，取得了十分不错的效果。

在本文的文物领域，文物领域的专业性较强，有大量的专有名词且命名规则复杂，如文物名“粉彩勾莲纹天球瓶”，文物名在文本中出现时，基于机器学习等方法抽取实体识别往往会比较难以识别，为了保证实体抽取的准确度，目前文物领域相关研究大都都采用基于规则和词典的匹配方法进行对文物实体的抽取，在本文中由于选取的数据源为故宫的藏品页面，该页面数据是除了对文物描述的文本信息是非结构化的，其他信息页面都已整理成结构化的数据，所以在构建知识图谱中，本文直接在爬取语料时对文物名进行抽取，然后再抽取属性之后一一对应后进行存储。

3.2.2 实体属性抽取技术

在知识图谱中的知识是以三元组的形式进行存储的，即“实体-属性-属性值”或者“实体-关系-实体”，属性是指数据属性，属性值为文本类型，这里实体之间的关系是指对象属性，其属性值为另一实体，如“出土朝代”是文物的一个对象属性，也是文物和朝代实体之间的关系描述，而文物的“高”则为文物的数据属性。在本文中为了能够同一描述对数据属性和对象属性的抽取统称为属性抽取。

传统的属性抽取方法，按照数据源分类的话分为两种：面向结构化或者半结构化的属性抽取方法、面向非结构化数据的属性抽取方法。基于结构化或者是半结构化的数据相对来说组织结构是比较规范的，可以很容易对其进行抽取。基于非结构化的文本数据没有固定的组织结构，一般都为自然语言组成的文本格式，由于中文的语法和句法结构都较为复杂，需要对句子语义进行理解剖析，所以这种结构的数据较为难抽取。一般常见的抽取方法有：基于规则的模式匹配方法、基于统计机器学习的方法和基于深度学习的方法。

K.Yu 等人[41]面向生物领域采用了基于模式匹配的方法抽取了蛋白质之间的实体关系，以生物学文章作为数据源从而抽取出不同蛋白质之间的实体关系来构建知识库。S.Boonpa 等人[42]采用了语法和句法分析抽取了泰国童谣中人物于

故事情境之间的关系，可以根据图谱来生成童谣。Y.Huang 等人[43]提出了基于远监督的人物关系抽取，他们先使用了远程监督的方法生成具有弱标记的数据集，然后再使用监督模型来训练人物关系分类器，从而能抽取句子中人物之间的关系。L.Xue 等人[44]使用了基于管道方法（pipeline）的关系抽取，，使用了深度神经网络来自动构建了知识库。I.N.Dewi 等人[45]面向医药领域利用了卷积神经网络来抽取药物之间的相互关系，并通过实验该方法可以有效提升药物之间的实体关系抽取结果。

实体属性或关系抽取在各个领域都有相关的研究，但在文物领域却没有很多可供参考的理论研究性的论文，主要是由于文物领域的数据结构性不强，且领域的专业性强，导致数据的标注难度比较大等，因此为了属性抽取的准确率，本文采用了基于规则和词典的方法结合模式匹配的思想，对文物文本数据完成属性（关系）抽取的研究工作。

3.2.2 知识图谱存储与可视化技术

3.2.2.1 知识图谱存储技术

知识图谱中的知识结构是以图的方式进行组织的，其中图的节点和节点之间关联各种关系，如果节点与节点之间有边进行了连接说明节点代表的实体是蕴含着某种关系的。在之前存储知识时通常是采用传统的数据库技术，如果两个数据之间具有某种连接的话需要定义外键来对他们进行联系起来，但是这种方式具有缺点就是当数据越来越庞大时，在数据库查询是需要耗费比较长的时间，导致查询性能低。因此，目前大规模的知识图谱一般基于 NoSQL 数据库进行存储。

NoSQL 数据库简介

随着互联网的蓬勃发展，数据也出现了潮水般的增长趋势。利用传统数据库将知识进行存储的存储方式逐渐无法支撑庞大数据体量和大量用户的查询量，查询性能低，其基于表结构的知识组织结构也无法满足解析数据的语义需求，表达能力弱，且在数据库中进行查询检索变得难以操作。另外，在如今互联信息网中存在越来越多的复杂信息，对这些复杂结构的大数据进行存储成为传统数据库的痛点。因此 NoSQL 数据库的产生逐渐替代了传统关系型数据库的地位，在各大场景都有应用，如社交网络图谱、搜索引擎、金融领域反欺诈、游戏领域等应用场景。NoSQL 是“Not Only SQL”的缩写，意思是在选择数据的存储时考虑更加合适的存储方式，适用于关系型数据库的数据就选择关系型的，不适用的情况没必要非选择关系型的，额外之意就是 NoSQL 数据库的产生就是为了弥补不适用关系型数据库的情况，所以 NoSQL 数据库是指一种非关系型数据库，基于动态结

构进行了存储，容易适应各种数据类型和结构产生的变化，具有灵活的架构。在关系型数据库中，存储方式是基于表结构的，NoSQL 则采用了大块模块进行组合的结构，如文档、键值对、图结构等。常见的基于图结构进行存储的 NoSQL 数据库有 Neo4j、GraphDB、OrientDB 等。

NoSQL 在数据处理性能方面对比于关系型数据库具有以下几个特性：1.易于数据的分散，在用关系型数据时各个数据存在关联，从而数据不得不存储于同一个服务器内，不适用分布式的场景应用，在大量数据进行写入时，性能会十分低效。而 NoSQL 中，数据模块是独立设计的，可以将数据分散到各个服务器中，使每个服务器的数据量减少，在进行大量数据的写入操作时就变得比较容易了。2.服务器的规模增加也使在计算时的性能有有大大的提示。3. NoSQL 数据库能够对多种数据类型进行高速处理，也能进行及时保存。

Neo4j 图数据库简介

Neo4j图数据库是目前最为流行的图数据之一，它采用基于图这种数据结构对数据进行建模，是具有高性能的图引擎，能够高效计算大量的多种类型的数据，其效率大大超过了传统的关系型数据库。Neo4j数据库不同于传统的数据库只能查询两度关系以类的短程关系，还可对远距离，远范围的关系进行查询，而且查询速度非常快，同时也可过推理模块，挖掘实体之间隐藏的关系。Neo4j使用了自己设计的cypher语言对数据进行查询，cypher语言跟关系型数据库所使用的SQL语言也十分相似，在使用起来不会有十分高的门槛，比较简单易学。

Neo4j图数据库在对数据进行存储时，是基于动态结构的，相比于关系型的数据库对数据建模时，需要对数据先进行结构定义，也就是说需要对数据的内容和形式进行描述，这种预定义的结构是比较可靠和稳定的，但是当我们要对数据进行修改时，就会使操作变得复杂和困难。基于图数据结构存储的Neo4j数据库就不需要预先对数据结构相关的定义，其操作十分灵活，从而使数据库具有高扩展性，用户可以很方便对知识进行扩充。举个例子来具体说明，比如当我们要描述人与人之间的关系时，若我们使用传统的关系型数据库进行存储，需要先构建三张表结构，然后对表中的列的内容和数据结构进行定义。如果使用Neo4j来进行存储的话，只要构建两个节点和几条边就可以实现，节点代表人这个实体，

边代表的就是人与人之间具有的关系，如果后续需要扩充的话，只需增加相应的节点和边就可以了。如图7所示。

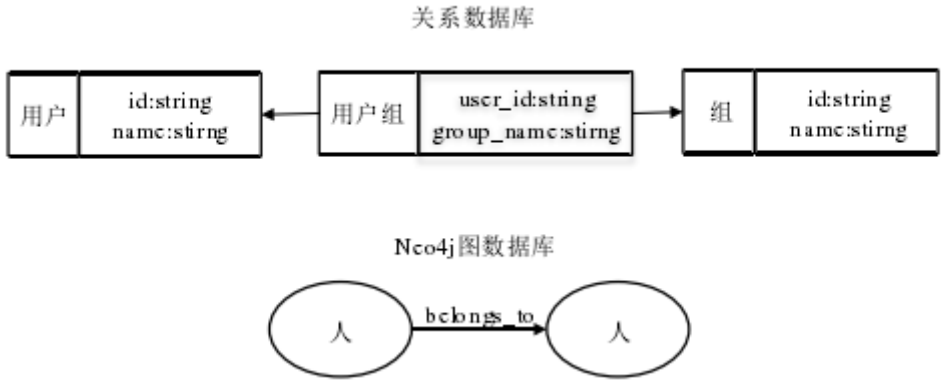


图7 数据模型对比图

基于图数据库的存储方案有多种模型可供选择，如北大自主研发的gstore、百度开发的HugeGraph、RDF4j等，但是这种类型的数据库大都是基于Java框架对RDF数据进行处理的，存在只能通过调用API的方式对RDF格式的数据进行存储，且只局限于RDF格式的数据，但是在实际应用中，可能存在如三元组形式等多种类型的数据。Neo4j的特点是基于原生的图数据库，具有高性能的查询算法，且支持多种形式的数据存储，具有可视化页面方便进行操作，而且支持对存储数据的可视化展示。

3.2.2.2 知识图谱可视化技术

数据可视化是提炼数据的关键信息，然后借助于计算机的图形软件对数据的一种图形化表示方案，它可以通过图形、颜色、符号等可视化元素将数据进行多元化展现出来，挖掘数据之间难以显示的数据特征，可以清晰有效的为用户传达与沟通信息，可视化技术可分为以下三类，并总结了各自的应用领域与任务。

表3 可视化技术的分类

可视化技术	数据特征	应用领域	应用任务
科学可视化	计算模拟数据、三维空间测量数据和医学影像数据	主要可应用于气象、物理、化学和航天等学科	寻找数据中潜在的特点、关系和模式
信息可视化	非结构且抽象的数据	主要是应用于金融、生物医学或地理信息等领域	帮助人们对数据的理解和挖掘其中的信息，发现新信息
可视分析	计算模拟数据、三维空间测量数据和医学影像数据	可应用各个领域	对数据进行整体分析并且推理新信息

在知识图谱领域，可视化技术的表现形式目前主要包括环形图、雷达图、力导向图和标签云等。S.Hasani等人提到了一种叫TableView的可视化工具，这个工具可以利用知识图谱的底层结构去生成许多的实体的力导向图和预览表，通过图形的可视化展示，可以很清晰直观的了解知识图谱的结构。一般来说，常见的知识图谱通常是采用力导向图的可视化形式，比如复旦大学的通用百科知识图谱（CN-DBPedia）也是采用了力导向图对数据查询的结果进行可视化展示，搜狗引擎提出的“知立方”图谱同样也采用了力导向图来展示人物之间的关系。下图为CN-DBPedia搜索“周杰伦”得到的关系图，这就是使用力导向图的可视化形式来展示知识之间的联系，可以很清晰直观展示人物所具有的关系，并且这种可视化形式的展示也符合了知识图谱在图数据库中三元组的存储方式。

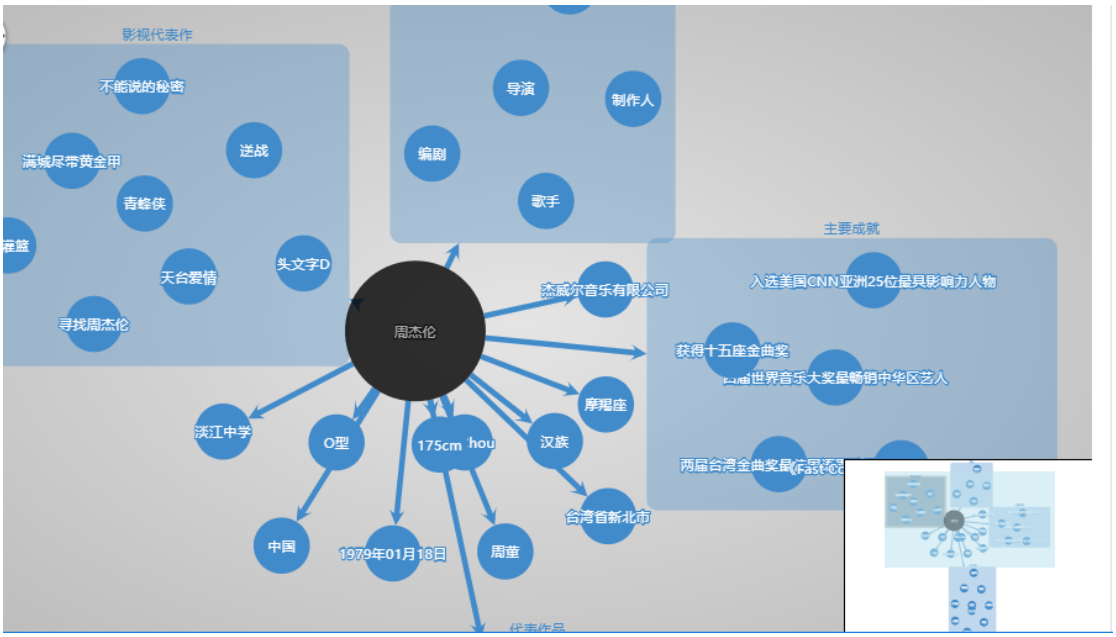


图8 在CN-DBPedia中搜索“周杰伦”得到的关系图谱

3.3 文物数据获取

3.3.1 文物数据抓取

本文的数据主要从故宫博物院进行爬虫获取，爬虫是一种自动化的脚本或程序，根据代码逻辑去自动获取网站的特定信息，从根本来说，爬虫就是通过模拟浏览器去对网站进行操作，从而获取网页中的部分信息，爬虫的基本流程如下图。

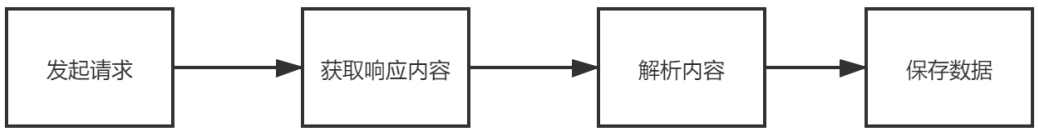


图9　爬虫基本流程图

首先通过URL向目标网站或服务器发送一个请求request，可以包含额外的请求头，即请求类型、cookie和浏览器的类型。然后等待服务器的响应，如果能正常响应，即HTTP状态码为200，服务器返回一个response，也就是我们请求的网页内容。网页内容包含多种形式数据，如html、json或者二进制格式的数据。获取到的内容后，根据不同格式的数据对页面进行解析，HTML可以使用页面解析库对内容进行拆解分析，json格式先转化为json对象进行解析，二进制数据保存为文件再进行下一步处理。然后将解析后的数据进行存储，根据实际的需求可以将数据保存文件到本地，也可以存储到如mysql、redis、mongodb等数据库。

本文的爬虫基于python实现，页面请求采用Requests库，页面解析采用Beautiful Soup库，存储数据直接存为json格式的文件至本地，为后续再进行方便操作。

本文构建的文物知识图谱主要以文物为核心，主要考虑了对文物信息的提取。由于故宫博物院文物信息全面，数据权威且比较规范化，故本文的数据源选取了故宫博物院的藏品目录下的页面来进行爬取，首先我们通过访问对应的藏品页面并解析获得该页面的文物列表和对应URL列表，在故宫博物院网站的藏品页面下将文物划分了陶瓷类、绘画类、书法类等共23类的子页面，本文选取了数据量较多的前六类作为文物数据源。获取到这六类文物的URL后，依次进行访问，通过解析页面得到该类下的所有文物列表和对应的URL，然后依次进行访问抓取该文物的详细信息。对爬虫下来的数据进行了分析，共2469条，其中部分文物在

故宫博物院上页面刊登较少的信息，之后也通过了百度百科的页面对文物信息进行了补充。我们将最后的数据以JSON格式进行了存储。

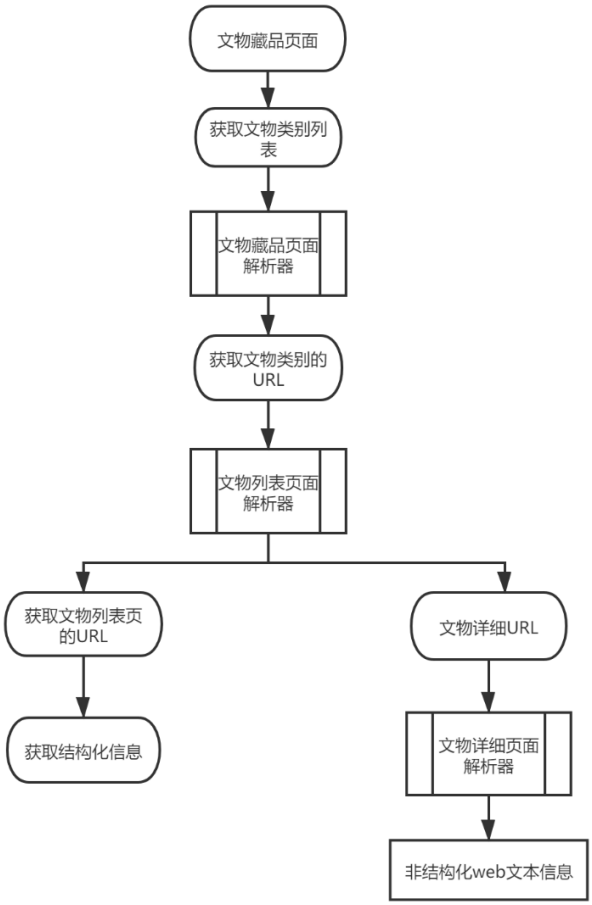


图10 文物数据爬虫基本过程

3.3.2 数据分类与信息

本文将上述爬取到的数据作为知识图谱构建的数据源，按照数据的结构方式，可将上述爬取到的数据分为两类：

（1）半结构化数据

本文获取到的半结构化数据包含两个来源：文物藏品页的文物分类数据和分类文物列表页的表格数据，分别如下所示。

文物藏品页的文物分类数据，这部分数据包含了故宫藏品所有文物分类，将文物藏品按照文物的材质进行了分类，对每个类别进行了链接，具有一定的结构性，在对文物的材质属性进行抽取时可直接将这部分的半结构化数据进行处理保存。

图11 文物藏品分类

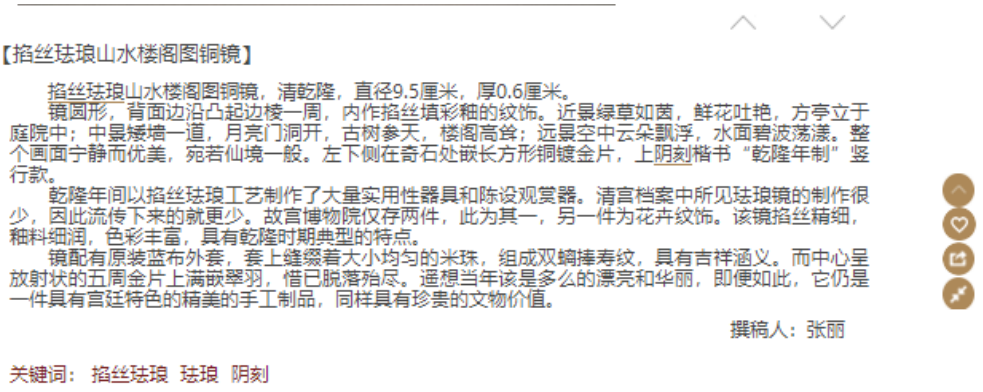
分类文物列表的数据，将文物分类链接点击进去进入该分类下所有文物列表，该文物列表在网络上以表格形式进行了展示，主要包含文物名称、文物时代、所属分类以及窑口等属性信息，表格类的数据具有规范性，在进行抽取时比较容易能够抽取出来，下图展示了文物列表页的部分数据。

文物名称	时代	分类	窑口
宜兴窑紫砂金漆云螭砚			宜兴窑
宜兴窑紫砂御题澄泥套砚			宜兴窑
宜兴窑石楳款紫砂胎镶玉槟榔木壶			金银锡器
宜兴窑紫砂双螭福寿水丞			宜兴窑
宜兴窑“时大彬”款紫砂胎剔红山水人物图执壶		漆器	
磁山文化红陶盂	新石器时代	红陶	磁山文化
磁山文化红陶平底碗	新石器时代	红陶	磁山文化
龙山文化黑陶双系壶	新石器时代	黑陶	龙山文化
齐家文化红陶盉	新石器时代	红陶	齐家文化
龙山文化黑陶双系罐	新石器时代	黑陶	龙山文化
龙山文化红陶鬶	新石器时代	红陶	龙山文化
大汶口文化白陶双系壶	新石器时代	白陶	大汶口文化

图12 文物列表页的部分数据

（2）非结构化数据

非结构化数据主要是文物详细页中的文物描述信息，以网络文本的数据形式展现在网络上，该文本形式以自然语言的构成，对每个文物的详细描述内容具有一定的相似特征，其组成结构较为相似，一般都先介绍该文物名称，再介绍文物的体积，后再描述具体文物所具有的属性。但由于中文语法结构和句法结构较为复杂，对这种自然语言文本形式的属性抽取页存在一定难度。


【掐丝珐琅山水楼阁图铜镜】

掐丝珐琅山水楼阁图铜镜，清乾隆，直径9.5厘米，厚0.6厘米。

镜圆形，背面边沿凸起边棱一周，内作掐丝填彩釉的纹饰。近景绿草如茵，鲜花吐艳，方亭立于庭院中；中景矮墙一道，月亮门洞开，古树参天，楼阁高耸；远景空中云朵飘浮，水面碧波荡漾。整个画面宁静而优美，宛若仙境一般。左下侧在奇石处嵌长方形铜镀金片，上阴刻楷书“乾隆年制”竖行款。

乾隆年间以掐丝珐琅工艺制作了大量实用性器具和陈设观赏器。清宫档案中所见珐琅镜的制作很少，因此流传下来的就更少。故宫博物院仅存两件，此为其一，另一件为花卉纹饰。该镜掐丝精细，釉料细润，色彩丰富，具有乾隆时期典型的特点。

镜配有原装蓝布外套，套上缝缀着大小均匀的米珠，组成双螭捧寿纹，具有吉祥涵义。而中心呈放射状的五周金片上满嵌翠羽，惜已脱落殆尽。遥想当年该是多么的漂亮和华丽，即便如此，它仍是一件具有宫廷特色的精美的手工制品，同样具有珍贵的文物价值。

撰稿人：张丽

关键词： 掐丝珐琅 珐琅 阴刻

图13 文物详细描述信息

3.3.3 文物属性知识抽取

经过上述爬取到的数据，我们得到了该文物的主要描述信息和部分半结构化的信息，如朝代、类别和所属文化信息。如图所示。结构化信息包括文物名、朝代、材质、工艺、类型、窑口这六类属性值，对于这些结构化信息我们直接保存对文物实例化，对非结构化的web文本信息需要处理成结构化数据，主要抽取纹饰、寓意、用途、形状及体积类这五类属性信息。首先要对web文本进行预处理，包括分词操作和词性标注，另外为了防止文物领域的专业词汇被切开，需要扩充自定义用户词典。本文使用的是jieba工具对文本进行操作，其总体的web文本信息处理过程如下图

通过对文物文本进行分析，发现文物文本的描述信息和属性知识点是具有一定的规则的。如（1）文物体积属性，一般的表示形式为高XX厘米，口径XX厘米，具有明显的规则，比较能够容易抽取出来。（2）文物的纹饰属性规则性表示不是很强，我们从几个方面进行了分析，一种是有些文物名称会直接带有纹饰属性，如白套红玻璃云龙纹瓶中带有“龙纹”这个纹饰属性，可以直接从文物名称进行抽取；第二种从描述信息进行抽取，纹饰信息一般的表示形式为饰某某纹或刻某某纹等；第三种从关键词信息进行抽取，有部分的描述信息的关键词会标明纹饰信息，可直接从关键词信息进行抽取。（3）文物寓意、材质及工艺等属性信息也可根据上下文具有的特征词进行抽取，如寓意一般的表示形式为具有某某寓意，取某某之寓意等。根据以上对文物知识的规则性，本文提出了基于特征词集的半自动抽取方法，具体操作为，首先根据文物知识具有的规则抽取出特征词集和生成规则库，然后利用规则库进行匹配，抽取出文本描述信息的属性知识点。下面我们将进行详细介绍抽取过程。

3.3.3.1 基于特征词集的文物属性知识抽取规则构建

本文采用了基于半自动化的文物属性知识抽取算法，对比人工抽取的方法，该算法极大的提高了工作效率，减少了大量人力。该算法基于文物知识的特征词集来进行抽取，本节首先将对构成特征词集的步骤及要素进行详细说明，然后提出基于特征词集的文物属性知识抽取规则，最后通过实验进行验证。

文物特征词集的构成

文物知识属性抽取规则的构建基于以下几个方面进行考虑：（1）文物描述信息的上下文知识点。（2）文物属性语法结构，主要是在描述文物属性时的特定的句法句式。（3）文物名称信息蕴含的属性信息，文物名的构词特征。（4）文物详细页面关键词栏的属性知识。（5）文物信息中直接具备固定某属性值的词语。特征词集就是指明文物属性信息在包含这三部分的特点的词语集合。

根据上述的分析，从以上五个方面总结了文物特征词集，主要包括以下特征词语：

（1）描述文物属性的具有上下文关系的特征词：主要是指出现在上下文中描述文物属性的指示词，如高为、口径、直径等词经常出现在体积类的上文中，用于、用作、用来等词经常出现在用途属性的上下文信息中，具体上下文特征表如下表所示。

表4 具有上下文关系的特征词

属性字段	特征词
纹饰	纹饰为
用途	用于、（可）用作、用以、（可以/主要）用来、
形状	形状为、形如
寓意	寓意是、的寓意、意寓、喻示、取意、寓意为、寓指、寓示
体积类	高、长、宽、足径、直径

（2）描述文物属性句式结构的特征词：通过对文物文本的分析，发现某些文物属性的描述是具有特定的句式结构，当文中出现这种句式结构很大概率是具有该文物属性。如纹饰属性信息通常以饰以…纹、描绘…纹、以…为纹饰等句式结构进行描述，具有（象征）…含义（寓意/涵义）通常用来描述寓意属性信息。体积类的属性基本是以上下文的形式来进行描述的，无其他特殊句式结构。具体的具有特点句式结构的特征词如下表所示。

表5 具有句式结构的特征词

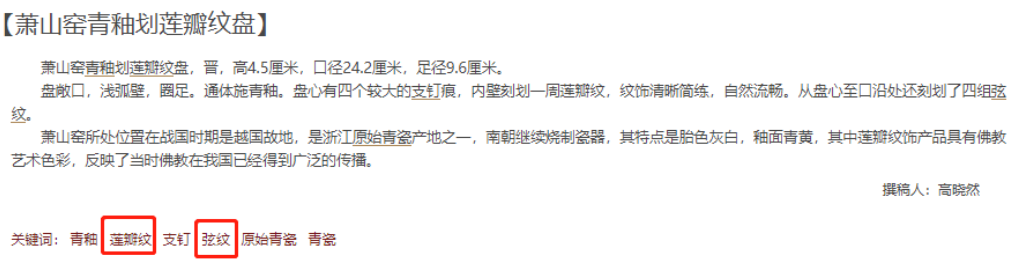
属性字段	特征句式
纹饰	饰以…纹、描绘（绘）…纹、以…为纹饰，刻划…纹
用途	是一种…用具、用作…具、是一种…具（器/皿）
形状	呈…形（形状）、…式样、作…状
寓意	具有（象征）…含义（寓意/含义/涵义）
体积类	无

（3）描述文物名称的构词特点的特征词：通过分析，文物名称通常是以叠加多种属性为结构的方法来命名的，如“萧山窑青釉划莲瓣纹盘”这个文物名称，是以窑口类型-萧山窑加上工艺-青釉然后纹饰-莲瓣纹最后为文物类型-盘这样的结构进行命名的。我们可根据文物名称的命名规则进行属性抽取。

表6 文物名称的构词特点的特征词

属性字段	构词特征词
纹饰	划…纹、…纹
用途	空
形状	…形、方（形）、
寓意	空
类型	某壶、某瓶、某杯、某罐
体积类	空

（4）存在于关键词信息栏的特征词：该网站在文物详细页中整理了该文物描述信息中的关键词，该栏的关键词是需对文物的专有词进行进一步说明，我们可以通过该关键词列表中进行抽取文物属性知识点，如通常关键词中一般会列出该文物具有的纹饰属性，如下图所示。


【萧山窑青釉划莲瓣纹盘】

萧山窑青釉划莲瓣纹盘，晋，高4.5厘米，口径24.2厘米，足径9.6厘米。

盘敞口，浅弧壁，圈足。通体施青釉。盘心有四个较大的支钉痕，内壁刻划一周莲瓣纹，纹饰清晰简练，自然流畅。从盘心至口沿处还刻划了四组弦纹。

萧山窑所处位置在战国时期是越国故地，是浙江原始青瓷产地之一，南朝继续烧制瓷器，其特点是胎色灰白，釉面青黄，其中莲瓣纹饰产品具有佛教艺术色彩，反映了当时佛教在我国已经得到广泛的传播。

撰稿人：高晓然

关键词：青釉 莲瓣纹 支钉 弦纹 原始青瓷 青瓷

图14 文物页面关键词栏的属性知识展示

（5）具备固定某属性值的特征词：经过对文物的描述信息分析，发现若文中包含此类特征词，则某属性的值就为特征词所指代的词语。如若文中出现了“陈设用”或是“具装饰性”此类的词，则该文物的用途属性则为陈设用具

3.3.3.2 基于特征词集的文物属性知识抽取算法

本文提出的基于特征词集的文物属性抽取的规则算法主要思想为：首先将爬虫获取到文物信息进行处理，将结构化信息中对应的文物属性直接进行抽取，对于非结构的文本信息先按标点进行分割成小句，然后根据匹配规则去遍历周围的词语，匹配特征词集，如果匹配成功，则将待抽取信息抽取出来，对知识进行进一步整理成本体概念的文物属性知识点。若匹配不成功，继续向前遍历，直到匹配成功或至文本末尾。

基于特征词集的文物属性抽取规则算法具体实现如下：

（1）文物数据的预处理。对爬取的结构化信息直接映射到文物本体属性，进行存储。对非结构化的文物描述信息进行进行预处理操作，将文物领域的专有

词汇定义成用户自定义词典导入到jieba工具中。对文物描述信息按标点符号切割成小句，对文物关键词栏整理成字典格式。

（2）对文物的描述信息进行知识抽取。将切割成小句的文物描述信息，循环制定的匹配规则去遍历周围的词语，如果匹配成功则进行下一步处理，若匹配失败则继续向下一个小句进行匹配，直到匹配成功或至文本末尾。将匹配成功的文物知识抽取出来，与文物本体设计的属性进行映射。经过对数据的分析，部分文物属性知识是符合上文构建的文物本体设计的属性，则可以直接进行映射，将抽取的属性知识作为文物本体属性值。但是我们发现部分抽取出来的属性知识并不能直接作为文物本体的属性，需要进行进一步转换处理， 如在抽取用途属性时，按照规则会抽出“是一种盛汤浆或饭食的器皿”，其中“盛汤浆或饭食”需要转换为定义的本体属性生活用具，然后将生活用具映射到该本体上。所以这里需要对抽取的属性知识进行进一步转换。

（3）对文物详细页面下的关键词栏进行属性抽取。关键词栏是对文物领域的专有名词进行进一步的说明，通过点击该关键词在其上方会出现该词的描述信息框，如下图所示。我们将此关键词信息整理成字典形式，key为关键词文本，value为关键词对应的描述信息。遍历关键词字典，对关键词字典的key按照对应的抽取规则进行检测，value信息按标点符号切成小句，同样按照第二步中的文本描述信息的规则进行抽取。

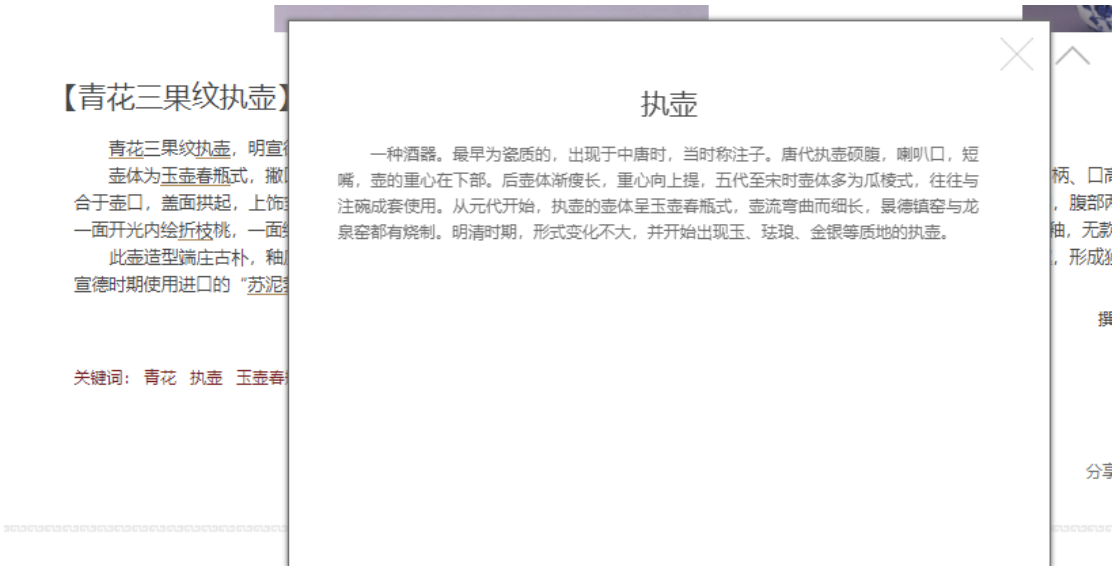


图15 关键词页面相关知识

（4）完成文物属性的抽取。将抽取出的文物属性知识映射到文物本体，将最后的数据整理成如图所示的json格式。

图16 文物属性抽取数据示例

本文主要构建了纹饰、用途、形状、寓意、类型、体积类这六类的抽取规则，我们从数据中抽取了100条的文物信息进行人工比对，经过初步分析，文物寓意属性知识点的抽取规则较不全面，通常寓意属性知识在文物描述信息较为隐晦，需要对文本整体语义进行理解，比较难抽取；类型、形状等其他属性具有较强的匹配规则，预计抽取的准确率较高。总体来说，文物的描述信息的描述语句大多比较规范，构建的抽取规则基本能够覆盖大部分的属性特征，较估计抽取的准确率相对比较高。

通过上述对文物信息的属性抽取，共包含2469条文物实例，其中包含2034条朝代属性知识，2469条文物类型属性知识点，1012条纹饰类属性知识点，寓意类知识点567条，形状属性知识点892条，用途属性知识点1134条，工艺属性知识点896条，材质属性知识点934条，窑口属性知识点1029条。部分文物属性的知识展示如表5所示。

表7 部分文物属性的知识展示

文物名称	朝代	类型	用途	纹饰	形状	寓意
嵌松石长剑	战国后期	剑	兵器	无	弧曲状	无
斗彩瓜蝶纹瓶	清嘉庆	瓶	陈设器	瓜蝶纹	圆形	吉祥
古铜彩双耳炉	清乾隆	炉	生活用具	卷草纹	扁圆形	富贵
斗彩鸳鸯卧莲碗	清乾隆	碗	生活用具	鸳鸯纹	圆形	美好
钧窑鼓式三足洗	宋	洗	陈设用具	无	无	如意
珐琅彩双环瓶	清乾隆	瓶	陈设用具	弦纹	壶形	富贵
越窑青釉执壶	清康熙	壶	酒器	细碎片纹	长圆形	无

本节主要完成了对文物属性知识的抽取，采用了基于特征词集的抽取规则算法，分别从五个方面进行了综合考虑：（1）文物描述信息的上下文知识点。（2）文物属性语法结构，主要是在描述文物属性时的特定的句法句式。（3）文物名称信息蕴含的属性信息，文物名的构词特征。（4）文物详细页面关键词栏的属性知识。（5）文物信息中直接具备固定某属性值的词语。将这五类的特征词构成特征词集按照正则的匹配规则依次遍历去匹配，将匹配到的属性知识点映射到文物本体的属性值上，最终完成对文物属性知识的抽取，经过抽取100条的样本信息进行验证，构建的抽取规则基本能够覆盖大部分的属性特征，较估计抽取的准确率相对比较高。这一步骤实际也是对本体的实例化操作，得到了本体实例，并抽取的本体知识整理成三元组格式，如下图所示，我们下一节将本体数据进行存储和可视化操作。

```
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/中文名> "黄釉暗云龙莲瓣盘" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/朝代> "清顺治" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/材质> "陶瓷" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/窑口> "官窑" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/高> "4.4厘米" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/足径> "5.8厘米" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/口径> "24.8厘米" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/纹饰> "龙纹" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/类型> "盘" .
<http://kg.BLCU.edu/entity/黄釉暗云龙莲瓣盘> <http://kg.BLCU.edu/ontology/property/工艺> "釉下彩" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/中文名> "青花人物诸葛碗长春万古款" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/材质> "陶瓷" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/窑口> "景德镇窑" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/高> "7.9厘米" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/足径> "7.4厘米" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/口径> "14.2厘米" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/类型> "碗" .
<http://kg.BLCU.edu/entity/青花人物诸葛碗长春万古款> <http://kg.BLCU.edu/ontology/property/工艺> "釉下彩" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/中文名> "白釉描金云龙纹罐" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/朝代> "明嘉靖" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/材质> "陶瓷" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/窑口> "官窑" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/高> "14.5厘米" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/足径> "7.7厘米" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/类型> "罐" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/工艺> "青花五彩" .
<http://kg.BLCU.edu/entity/白釉描金云龙纹罐> <http://kg.BLCU.edu/ontology/property/纹饰> "龙纹" .
```

图17 本体三元组数据示例

3.4 文物知识图谱的存储

在上一节我们文物知识进行了属性抽取，同时也完成了文物的本体实例化工作，本节完成对文物知识图谱的存储工作的展示。

文物知识图谱通常以图的结构进行存储，本文同时采用了Neo4j图数据库存储然后对本体进行可视化。

Neo4j图数据库是目前最为流行的图数据之一，它采用基于图这种数据结构对数据进行建模，是具有高性能的图引擎，能够高效计算大量的多种类型的数据，其效率大大超过了传统的关系型数据库。Neo4j数据库不同于传统的数据库只能

查询两度关系以类的短程关系，还可对远距离，远范围的关系进行查询，而且查询速度非常快，同时也可过推理模块，挖掘实体之间隐藏的关系。Neo4j使用了自己设计的cypher语言对数据进行查询，cypher语言跟关系型数据库所使用的SQL语言也十分相似，在使用起来不会有十分高的门槛，比较简单易学。

知识图谱中的知识存储方式是以图为结构的，其属性关系图是由节点、属性和其之间联系的关系组成的。由于互联网知识的急剧增加，使知识图谱中的三元组数据也呈现了巨大的增长，节点之间的关系也随之变得更加复杂。直接将RDF格式的数据存储到文本中进行数据的操作使查询效率也变得十分低效，因此探索一种新的基于三元组格式数据的存储方式显得十分必要。

由于数据之间存在连接关系复杂且具有多变性，为了使知识图谱能够具有扩展性和稳定的维护功能，将RDF格式的三元组数据和从自然语言中抽取出的三元组放到Neo4j图数据库中进行保存。与传统的关系型数据库相比，Neo4j的查询效率是遥遥领先的，所以能够满足大规模三元组数据的查询需求的。因此，本文将构建好的文物知识库存储到Neo4j图数据中，具体流程如下图。

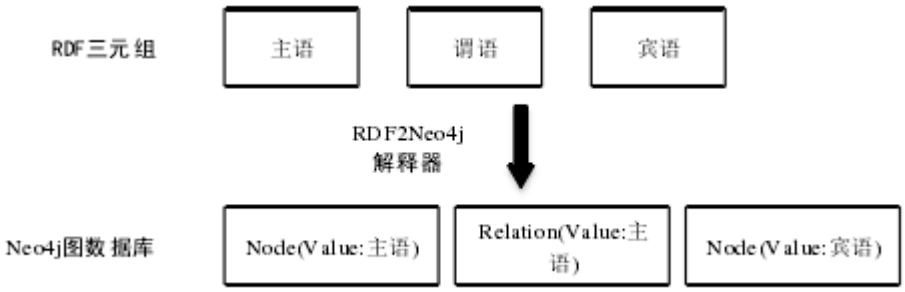


图18 RDF三元组数据转换到图数据库

经过上文的数据处理操作获得三元组数据，将三元组存储到Neo4j图数据库中，具体操作如下：

（1）RDF三元组数据具有官方的数据解析API：Jena API可以获取到RDF数据集中的每个三元组的主语、宾语和谓语信息，进而封装为Java中的对象。

（2）编写RDF2Neo4j解析器，将通过API解析出的对象数据，提取对象属性等信息，构建Cypher语句，将主语和宾语映射为数据库中的Node类节点数据，谓语映射为Relation类的关系数据。

（3）JAVA程序中给定Neo4j的相关配置等，通过Neo4j API可以实现将三元组数据导入图数据库中，导入的数据会在图数据库中构成图结构并保留原有的关系信息，部分数据展示如下。

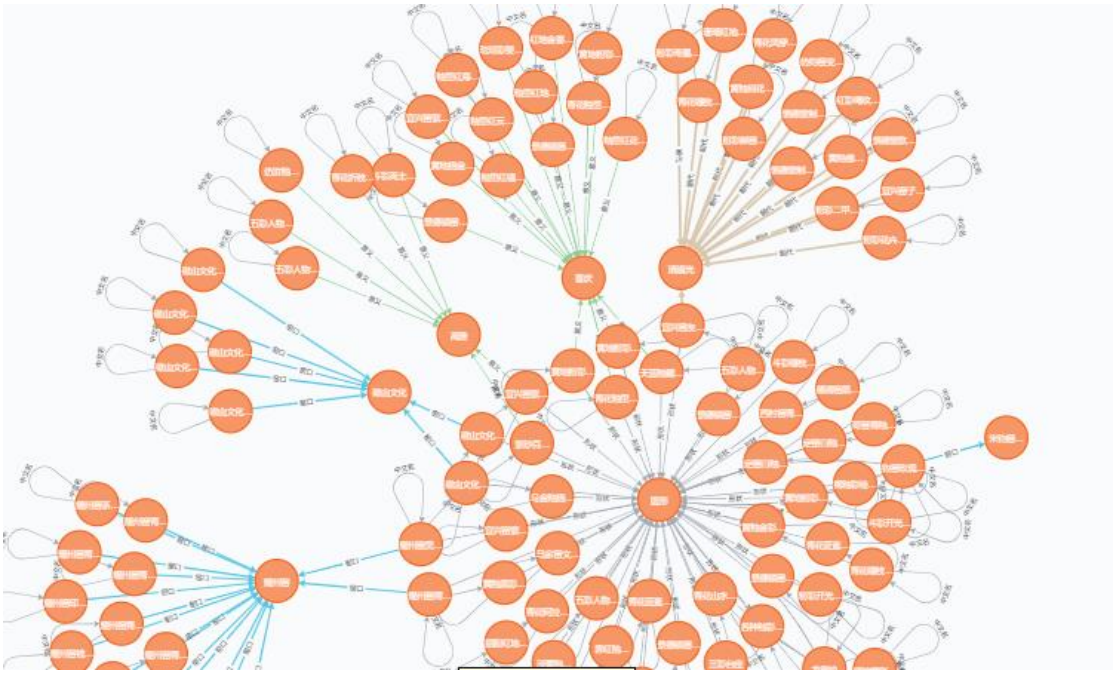


图19 三元组数据在Neo4j进行持久化存储

3.5 本章小结

本章详细阐述了文物知识图谱的构建流程。首先介绍了本章知识图谱构建过程中所用到的相关技术和理论知识，详细阐述了实体抽取、属性抽取的相关方法和综述研究，并介绍了知识图谱的存储和可视化技术。然后获取文物领域相关数据并对数据进行整理分析，针对文物文本的属性抽取提出了基于特征集自动属性抽取方法来抽取web文本中属性信息，最终完成了本体的实例化工作，然后将整理好的本体知识库存储到Neo4j图数据库中，并实现文物知识的可视化。

第四章 基于文物知识图谱的智能问答系统

4.1 引言

第三章我们构建了文物领域的知识图谱，这章我们将考虑如何对知识图谱的信息进行有效合理的利用，在知识图谱的实际应用有很多，如推荐、搜索、智能问答等场景。在日常生活中，人们获取自己想要的信息，通常是以问答的形式来解决自己的疑惑，用户使用自然语言进行询问是基于一种最熟悉最习惯的方式，使计算机与用户的交互性增强，用户可以很快速便捷的获取到自己想要的信息。因此本文以智能问答为文物知识图谱的实际应用，构建了基于文物知识图谱文物领域的智能问答系统。

4.2 相关研究

4.2.1 基于知识图谱的自动问答

基于知识图谱的问答主要核心技术是需要理解用户输入的自然问句的意图并对其进行解析理解。在机器处理过程中，存在的难点就是如何将自然语言问句转化为机器可理解的语言，因其自然语言问句的表述方式与在知识图谱中的存储方式有着很大的区别。如输入“请问周董演过什么电影？”，在知识图谱中以三元组的格式进行存储的，如以下三元组（“周杰伦”,“电影”,“《不能说的秘密》”）以及与其相关联的知识三元组（“周杰伦”,“代表作”,“《不能说的秘密》”）、（“周杰伦”,“别称”,“周董”）等。问答系统需要做的就是如何在用户输入的自然语言问句中找出与三元组对应的实体与实体之间的关系，如“周董”与“周杰伦”，“电影”与“代表作”如何把这些实体对进行相互对应起来。目前基于知识图谱的问答主要有两类研究思路：一类是基于语义解析的方法，另一类是基于信息抽取的方法。语义解析（Semantic Parsing-Based，SP-based）[46]这种方法通常是构建一个语义解析器，然后通过语义解析器将自然语言问句转化为知识图谱可理解的结构化的表达。之后把转化好的结构化表达输入到知识图谱检索模块中查询相关的三元组信息，提取答案。基于信息抽取（Information －Retrieve-Based，IR-based）[50]的方法主要思路是：先从问句中抽取实体，然后通过知识图谱查询，得到以该实体节点为中心的知识图谱子图，子图中的每一节点或者

边可能都可以作为问题答案的候选项，然后根据观察可以制定相关规则或者模板进行信息抽取，得到问题的关键信息后对候选答案进行筛选得到最终答案。

随着深度学习的快速发展和在各种自然语言的任务中表现出优异的效果，是当前最火的研究趋势之一，研究者们也开始采用深度学习的方法利用在知识图谱问答中，基于深度学习的知识图谱问答（Neural Network-Based,NN-based）已成为目前的主流方式。深度学习较传统的神经网络有更强大的表征能力，可以对自然语言句进行深层次的语义挖掘。深度学习在知识图谱主要两个应用方面：一种是基于传统语义解析的模型的基础上进行的改进，将语义解析使用的传统实体识别、关系抽取的方法替代为基于深度学习的方法。第二种是基于向量语义相似度计算的端到端方法，主要思路是：将知识图谱的实体以及实体之间的关系和自然语言问句都通过神经网络转化为向量表征形式，然后通过计算问句向量和知识图谱实体以及关系的知识向量语义相似度来判断是否是正确答案，这种方法的主要技术要点在于知识图谱中的知识表述形式问题，这个方向的研究目前还不是很成熟，本文主要采用的是基于语义解析的深度学习问答方法，重点介绍基于深度学习方法的技术方法。

4.2.2 基于深度学习的自动问答

深度学习的重要一个优势是可以对大规模的数据进行训练，从大规模的数据中解决文本语义解析的问题，能够有效地挖掘问句中隐藏地语义信息。

在问答领域的评测任务上深度学习方法表现大放异彩，有着十分优异的表现，榜单上几乎是完全超越传统方法，由于基于深度学习的神经网络端到端的特性，训练时十分方便，只需很少的人工参与，且实验效果也非常好，研究者一般都是利用深度学习方法在对自然语言处理任务进行实验。目前在基于知识图谱的问答方面，有很多研究者相继了发表了自己的研究成果。Borders 等[54]利用向量表征的语言模型来解决知识图谱问答，主要方法是将自然语言问句和知识库中三元组都用 word2vec 模型映射为低维词向量的表征形式，然后通过计算它们之间的语义相似度来找出问题的答案。Weston[55]在基于记忆网络（Memory Network）的基础上，提出了 MemNNs 模型，该模型可在大规模的问答集上进行训练，也使之有复杂度更高的推理功能；Yih 等人[56]将知识问答分割成两个步骤：命名实体识别和关系分类，并分别使用了卷积神经网络模型（CNN）来实现这两个功能，最后完成问题答案的抽取。Yang 等人[57]在 Yih 的基础上进行了改进，提出基于端到端的方法，将实体识别和关系映射看作是一个整体过程进行处理。Yih 等人[58]将提出了 Multi-Column CNN（MCCNNs）模型，并利用了不同角度的信息如上下文、答案路径和类型等进行了分析，并基于卷积神经网络模型（CNN）模型进行了抽取，最后通过累计

分数对候选答案进行排序。Yang 等[59]在 MCCNNs 模型的基础上加入了注意力机制，使模型更能关注有用的信息，同时还引入知识库的信息作为模型的额外信息，使答案向量能补充到更多的知识信息，从另一方面看还解决了语料信息在解析过程中语义信息不足的问题。Zhang 等人[60]也关注到注意力机制，为了研究注意力机制的有效性，利用不同方式的注意力表示形式来获取不同的问句向量表征。

在中文的研究方面，Xie 等人[61]在进行实体识别和属性链接过程使用了基于深度学习的方法。Yang 等人[62]则在实体识别的过程中利用了 GBDT 的机器学习方法，属性链接则结合了传统的神经网络模型和基于深度学习的卷积神经网络模型。Lai 等人[63]结合了人工构建的特征，且设计了基于层次的细粒度分词方法用于属性链接。杜泽宇等[64]构建了一个电商领域的知识图谱，并在此基础上构建了一个问答算法框架（CEQA），主要技术有基于 CRF 的实体识别和 Word2vec 的属性链接算法。王银丽等人[65]基于限定领域构建了智能问答系统，主要针对 FAQ 的问句匹配。

基于深度学习的问答算法在大规模问答语料的支持下，展现了几乎不需要人工干预、效果超过大多数传统方法等优势，因此本文智能问答系统使用的问答算法也采用了基于深度学习的问答算法，这在下一节我们将进行详细介绍。

4.3 相关模型介绍

4.3.1 LSTM模型

长短期记忆网络(Long Short-Term Memory)是基于时间递归神经网络（RNN）的改进模型，其神经网络结构由一个或多个具有记忆和可遗忘功能的单元构成。主要解决了当较长的序列输入到RNN网络中会存在远距离的信息遗失的问题，LSTM能够以更长的距离进行传递，使语义信息能够更加持久的存于网络中。LSTM将神经元设计成一种具有门结构（Gate Term）的细胞单位，LSTM网络结构具有的门单元有：输入门（Input Gate）、输出门（Output Gate）、遗忘门（Forget Gate）。LSTM结构如下图所示。

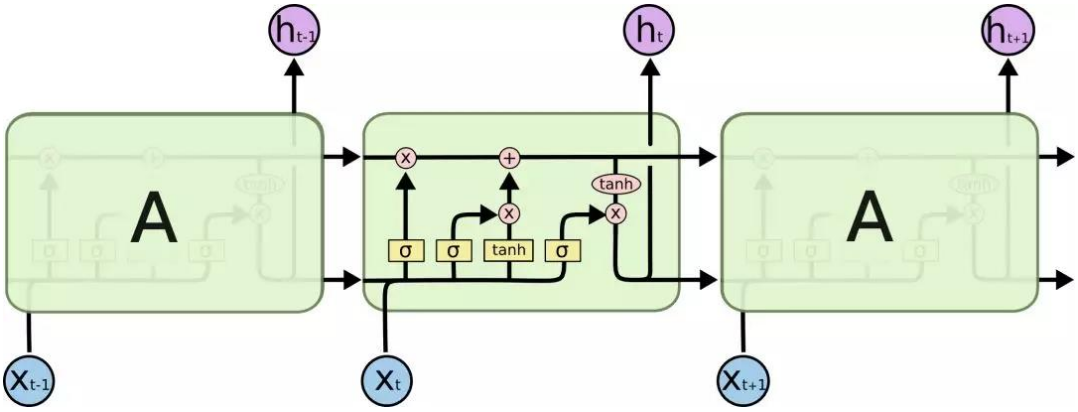


图20 LSTM模型结构

遗忘门用于判断是否丢弃细胞的信息量，使用sigmoid函数进行表示，公式如下：

$$f_t = \sigma(W_f \cdot h_{t-1}) + U_f \cdot x_t + b_f \quad (1)$$

其中x_t为当前时刻的输入，h_{t-1}代表的是上一时刻隐藏层的输出，U_f为输入的权重，W_f为遗忘门的权重，b_f是偏置项。

下一步通过输入门（Input Gate）来决定单元要更新的信息，将在单元状态中将新的信息进行存储。计算方式如下：

$$i_t = \sigma(W_i \cdot h_{t-1}) + U_i \cdot x_t + b_i \quad (2)$$

$$c_t = \tanh(W_c \cdot h_{t-1} + U_c \cdot x_t) + b_C \quad (3)$$

$$c_t = f_t c_{t-1} + i_t c_t \quad (4)$$

其中U_t、W_t、U_C、W_C均为参数矩阵，b_i、b_c为偏置项参数，c_t为当前时刻细胞单元的状态，c_{t-1}是上一时刻的细胞单元的状态。

最后决定细胞单元输出什么信息，使用输出门（Output Gate）决定LSTM模型的输出向量，计算公式如下：

$$o_t = \sigma(W_o \cdot h_{t-1} + U_o \cdot x_t) + b_o \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

其中U_o、W_o均为参数矩阵，b_o为偏置项参数，h_t为当前LSTM的输出。

4.3.2 Bi-LSTM模型

LSTM只是单向结构，没有考虑词语在句子序列中的前后顺序，无法获得上下文的语义信息，丢失了从后向前的编码信息，从而使语义判断存在不够准确的缺

点。研究者们因此提出了双向长短期记忆网络Bi-LSTM[11]，使用了前向和后向两个LSTM细胞单元对•输入的序列进行处理，最后的输出结果为两个LSTM单元输出的向量拼接。对比与LSTM模型，Bi-LSTM模型既能保留LSTM的优点，能够处理长距离依赖的问题，同时也可以更好的融合兼顾上下文的信息，从而提取更深层次的语义信息，其结构如图所示。

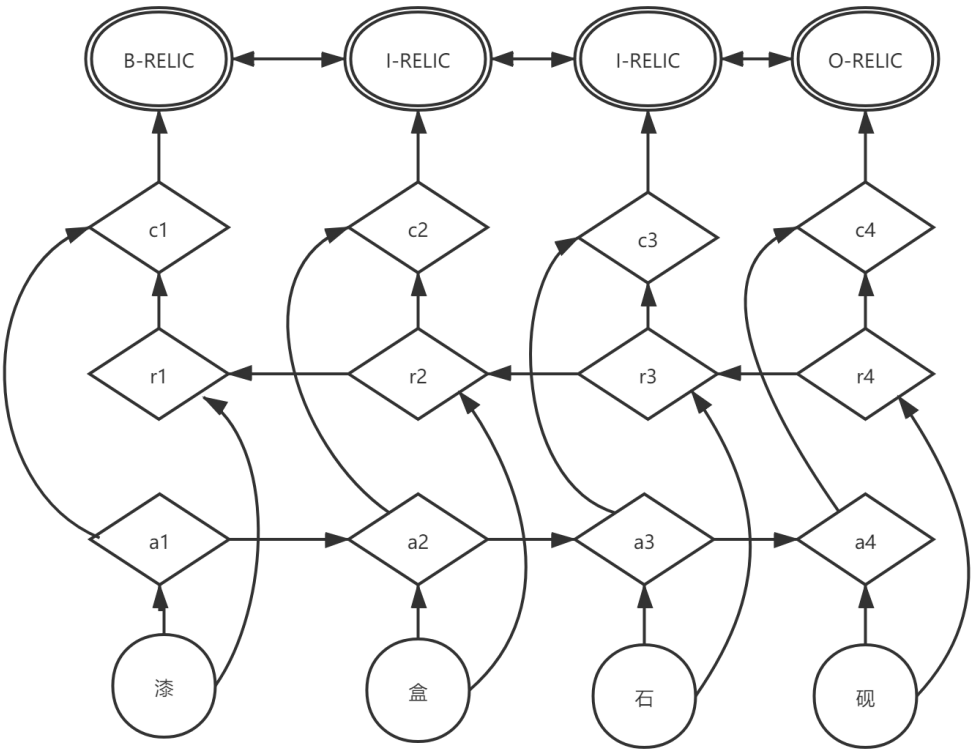


图21 Bi-LSTM+CRF模型结构

将输入向量输入到Bi-LSTM的前向和后向LSTM单元中，进行拼接得到输出向量$\mathrm{h_t}$，其处理公式如下：

$$\mathrm{h_t} = [\overleftarrow{h_t}; \overrightarrow{h_t}] \tag{7}$$

其中$\overleftarrow{h_t}$是前向单元序列的输出，$\overrightarrow{h_t}$为后项单元序列的输出，然后sigmoid层得到Bi-LSTM的输出当作自己的输入，sigmoid函数如下表示：

$$\mathrm{sigmoid(x)} = \frac{1}{1+e^{-x}} \tag{8}$$

将神经网络序列用于标注任务时，通常模型把任务转换为分类任务来解决，在使用Bi-LSTM模型时，输出当前最大概率的结果，但这样做没有考虑到基于全局的路径信息，也就是说当前的输出结果可能是不符合命名实体的规则的，因此引入CRF模型融合到Bi-LSTM模型中，CRF层可对输出结果进行约束，如句子的开头应该是B-或者O字符，不能是I-字符。

将Bi-LSTM的sigmoid层后加上CRF层，即把sigmoid层输出当作CRF层的输入，使用Bi-LSTM层提取到句子序列的深层语义关系，CRF层可以获取到句子的标记信息，将两者融合起来，这样的模型输出不再是局部概率最大的结果，而是会输出基于整体概率最大的标签序列，基于CRF的概率计算公式如下：

$$S(X,y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{9}$$

其中A为状态转移矩阵，P代表输出矩阵，X为输入特征，y代表对应的标签。

4.3.3 BERT模型

2018年Peters等人发表了基于RNN的ELMo预训练模型，引入上下文动态调整单词的embedding，从而解决了静态word embedding训练时无法区分多义词的难题。NLP领域又开始掀起了研究预训练模型的热潮，OpenAI相继推出基于Transformer的GPT模型[4]，Google在此基础上改进了Transformer，采用双向语言模型进行预训练的BERT模型，且数据规模相比GPT模型增大不少。下图回顾了预训练模型的发展过程和最新的研究成果。

预训练模型在多个任务上都有着惊人的表现，从而开创了NLP研究的新范式，即首先利用大量的无监督的语料数据对模型进行预训练，然后再使用少量的监督语料对语言模型进行微调，在预训练模型基础上还可叠加如CNN、RNN等其他模型，也可直接叠加一个输出层，进而完成文本分类、问答、阅读理解、文本语义相似度等具体的NLP任务。下图展示了基于BERT的微调可以支持的NLP任务类型，包括句对关系分类，单句的文本分类、智能问答和序列标注任务。

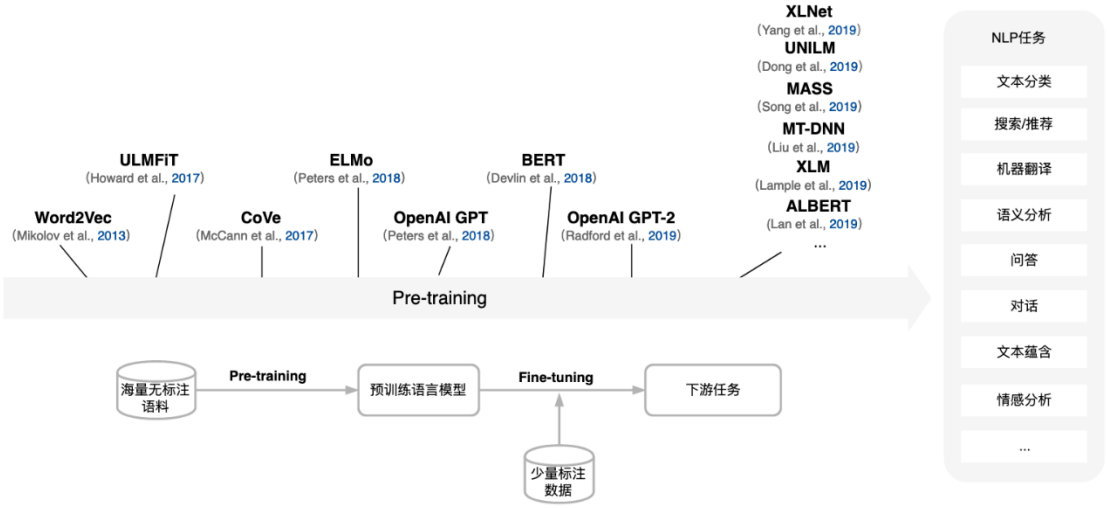


图22 预训练模型发展历程

本文采用了BERT模型对问句进行意图分类，BERT模型自出就横扫了自然语言处理的各项任务的榜首，刷新了各个任务的最佳水平，效果显著，且比其他基于RNN预训练模型训练速度快。成为很多公司实现AI技术的底层模型，如谷歌、微软、百度等机器翻译模型都使用了BERT模型。下面将对BERT模型进行详细介绍。

BERT是基于Transformer的深度双向语言模型，基本结构如下图所示。 与GPT不同的是，BERT采用的是Transformer的编码部分对特征进行提取，BERT分为预训练和下游任务微调两个阶段。

如图所示，输入BERT模型的是一个线性的序列，针对不同的任务，可以输入单句文本或者是句对文本，句首采用了[CLS]符号表示，句尾使用[SEP]符号表示，如果输入的是句对文本，句子之间需增加[SEP]符号。输入特征由Token向量、Segment向量和Position向量这三个向量组成，分别为单词表征、句子表征、位置表征。

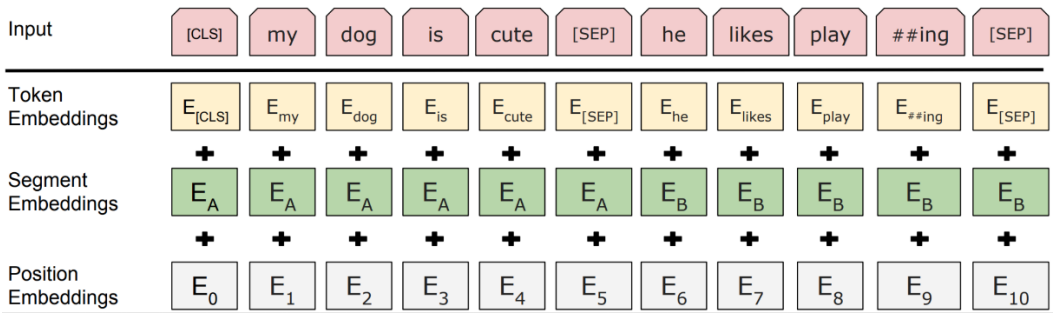


图23 BERT组成结构

BERT采用了MLM(Masked LM)和NSP(Next Sentence Prediction)两种策略对模型进行预训练。

MLM，即Masked Language Model，对输入的单词序列，通过随机掩盖15%的单词，掩盖的单词使用[MASK]符号进行标记，然后预测被掩盖的单词进行预训练。相比传统的语言模型只能通过单方向进行预测，MLM可以任意从left-to-right或right-to-left两个方向预测被掩盖的单词。但是这样的做法也存在两个缺点：（1）预训练阶段采用了[MASK]符号对随机单词进行了掩盖，但是在下游的微调任务时并没有掩盖单词的操作，会造成预训练与微调任务不一致问题。（2）预训练阶段每个batch中只有15%被掩盖的单词会被预测，而不是整个句子，因此模型的收敛速度需要花费比单向语言模型更多的时间。对于第一个缺点，可以把需要掩盖的词抽80%的单词进行替换，10%的单词随机替换其他单词，10%保持不变。由于Transformer在训练过程中不知道哪些词是被随机替换的，哪些词需要进行预测，这样就会强迫每个词的表征都会对上下文进行参考。对于第二个缺点，目前没有明确的改进方法，但是从模型训练的效果来看付出的代价是值得的。

NSP，Next Sentence Prediction。很多任务是需要理解句子之间的关系，如智能问答、语义推理等下游任务，而传统的语言模型中没有考虑到句子与句子之间关系的学习，所以为了理解句对之间的语义关系，引入预测下一句模型，增加对句子A与B之间关系的预测任务。训练语料可以语料库中随机抽取一对句对A与B，其中以50%的几率B为A的下一句，50%的时间是B随机从语料中抽取的句子。

根据参数设置的不同，Google提出了两种不同规模大小的BERT模型，如下图所示。

表8 BERT模型规格

BERT模型规格	Layers	Hidden Size	Attention Head	参数量
Base	12	768	12	110M
Large	24	1024	16	340M

4.4 问答系统整体框架设计

本节主要实现一个基于文物知识图谱的智能问答系统模型，该模型可通过文物领域的自然语言进行语义分析，充分理解用户的询问意图，然后将用户的自然语言的问句转换为知识图谱可理解的结构化查询语句，最后通过检索查询知识图谱三元组数据返回给用户答案。本节将对智能问答系统的算法进行详细介绍。

在第三章构建的文物知识图谱的基础上，本文的问答系统模型主要分为三个大模块：（1）问句理解模块：该模块是智能问答的核心模块，由两个步骤构成，

用户询问意图识别及句中关键信息抽取（即槽位识别），唯有在准确的理解用户的询问意图和识别句中关键信息的基础上，后续阶段的知识检索及答案推理模块才有意义。否则，整个最后检索的答案也是无效的。（2）问句转换模块：通过上步骤的问句理解模块输出的用户询问意图和句中关键槽位信息，将这两部分的信息结合进行转换知识图谱可理解的结构化检索语句。（3）答案检索模块：将转化好的知识图谱检索语句在知识图谱中进行查询检索，获得用户所需的答案。

本文基于文物知识图谱的问答系统架构如下图所示。可以看到，用户通过终端输入自然语言形式的问句，向问答系统进行提问，首先问句通过经过问句理解模块，在该模块进行三个操作，先识别问句的询问意图，然后对句中的关键槽位信息进行抽取，最后对槽位自然语言语句进行改写成本体设计的概念信息，将意图槽位这两部分的数据传递到到问句转化模块，然后该模块进一步将意图槽位信息转化为Sparql语句进行输出，最后答案检索模型通过知识图谱中知识库信息完成sparql语句的查询，并最终返回查询语句的答案到用户终端上。

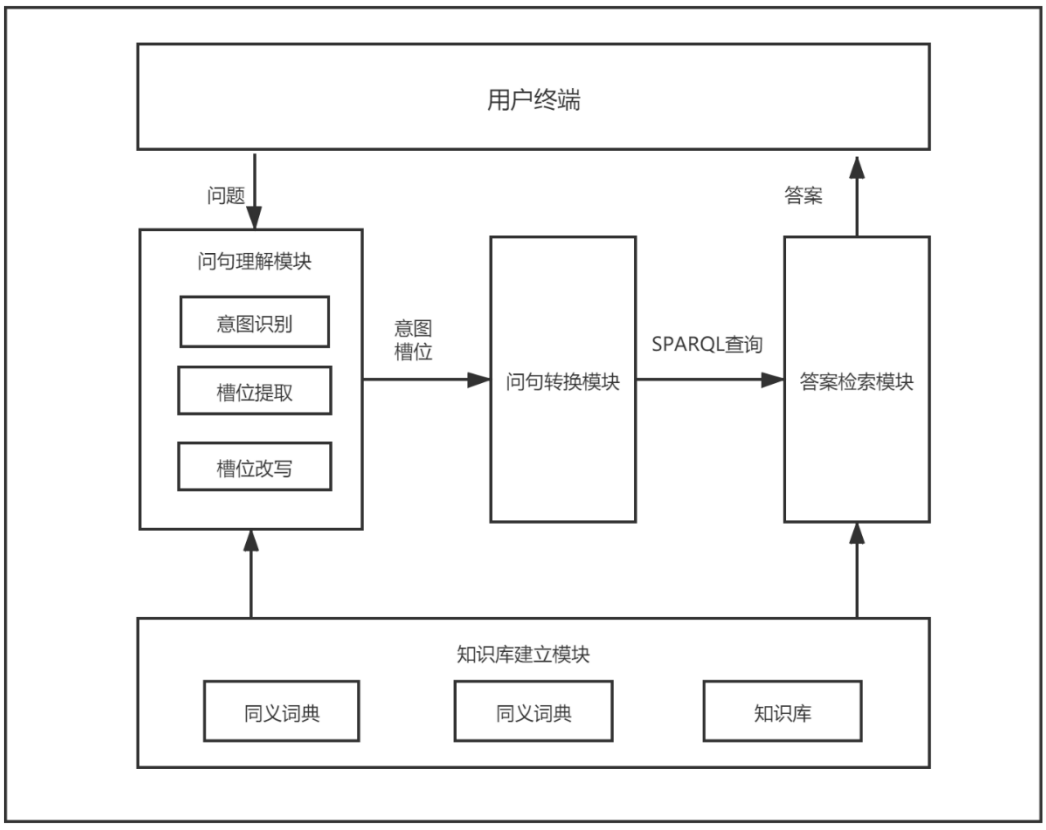


图24 基于文物知识图谱的问答系统架构

我们根据一个问句实例对整个系统模块进行更深的理解。示例流程图如下图。用户输入“文物出土于清乾隆时期且纹饰为弦纹的文物有哪些？”我们先对问句进行预处理操作，主要是分词操作，我们将文物领域的专有词包括文物名称、

纹饰类型、工艺等属性值信息，通过jieba工具分成“文物/出土于/清乾隆/时期/且/纹饰/为/弦纹/的/文物/有/哪些？”，将分好的句子输入到意图识别模型中进行判断该问句的意图，然后该模型输出该句子的意图类别为查询文物名称，然后再将句子输入到槽位识别的模型中对句子中的关键词信息进行抽取，抽取出的信息：{朝代：“清乾隆”；纹饰：“弦纹”}；将得到的信息转化为SPARQL语句到检索模块进行查询，SPARQL语句如下图所示，再比如问句“粉彩勾莲纹天球瓶的出土年代是什么时候？”，能识别出意图类别为朝代，经过槽位提取模块可以准确提取出文物名称为“粉彩勾莲纹天球瓶”。

4.5 问句理解与解析

4.5.1 意图识别数据集及模型构建

4.5.1.1 数据集构建

在文物的限定领域内，对文物的基础信息进行提问的类型可进行罗列，所以基于领域内的问句意图识别可以转换为问句的关系分类问题。根据第三章构建的文物本体模型和用户常问问题类型的调研，本文制定了12种询问意图类型。整理如下表所示。

表9 意图类型

意图类型	示例问句
文物名称	纹饰为弦纹的文物有哪些？
朝代	粉彩勾莲纹天球瓶的出土于哪个年代？
材质	掐丝珐琅三足熏炉的材质是什么？
用途	蓝玻璃方花觚这个文物是干什么用的？
类型	青釉狗圈是一种什么文物？
工艺	斗彩瓜蝶纹瓶是怎么制作的？
形状	画珐琅开光山水人物鱼缸是什么形状？
窑口	黄玉佛手花插属于哪个窑口？
寓意	白玉羊首耳瓶具有什么寓意吗？
体积	黄釉描金双耳罐有多高呢？
文化	玉鹰攫人首佩所属哪个文化？
文物数量	清朝时期的文物数量有多少？

本文通过问题模板进行对实验数据进行构造，针对每个意图类型，通过扩展和泛化问题的表示形式，人工构造每个类型的问题模板，然后通过文物领域的词典对问题模板进行填充获得问句语料信息。

首先针对每个类型，总结了常见的句式结构，对每一种句式结构中的词基于同义词转换，词语相似度计算进行语义泛化，对模板中的规则进行详细说明如下：尖括号标注的实体类型名是后续通过抽取文物领域的词典进行填充的词语，举例<文物名称>可从文物词典中文物名称类抽取一个实例如“粉彩勾莲纹天球瓶”；括号内的内容按照斜线分割是指从中选择一个词；方括号内的内容可选择可不选择；每种句式都可添加前缀或后缀，前后缀的内容整理如下表所示。

表10 句式前缀后缀词

前缀词	1.请问、你知道、你说、我问你、问问你、请问你知道、我考你、请告诉我、回答我、我想知道、
	2.给我、向我、为我、帮我、替我 + 说说、讲讲、说下、讲下、说一下、讲一下、回答、搜下、查下、搜索、检索、查询、百度、谷歌、
后缀词	吗、呢、嘛、吧、呗、呀、喔、好吗、可以吗、行吗、行不、好不、好不好、行不行、可不可以

本文列举了朝代、纹饰、寓意类型的问题模板如下表所示。

表11 朝代类型问题模板

朝代类型-问句模板
<文物名称>(出土于/属于/归属于/是/为/归为/产生于)(什么/哪个)(朝代/年代/时代/时期/时间/阶段)
<文物名称>[的](朝代/年代/时代/时期/时间/阶段)是(什么/什么时候/何时/什么时间/几时/哪时)
<文物名称>[的](朝代/年代/时代/时期/时间/阶段)

表12 寓意类型问题模板

寓意类型-问句模板
<文物名称>的(寓意/含义/涵义/蕴意/)(是/有/包含)(什么/啥)
<文物名称>(是/有/包含)(什么/啥)(寓意/含义/涵义/蕴意/)
<文物名称>(象征/意喻/意寓/喻示/寓为/寓示/寓指/)[着](什么/啥)

表13 纹饰类型问题模板

纹饰类型-问句模板
<文物名称>的(纹饰/装饰/饰样/花纹/纹样/纹案/图案/图饰/图纹/花样/)(是/有/包含)(什么/啥)

<文物名称>(是/有/包含)(什么/啥)(纹饰/装饰/饰样/花纹/纹样/纹案/图案/图饰/图纹/花样/)
<文物名称>(装饰/饰以//喻示/寓为/寓示/寓指)[着](什么/啥)(纹饰/装饰/饰样/花纹/纹样/纹案/图案/图饰/图纹/花样/)

通过上述的模板方式使用python语言编写了自动化的程序，共生成了2万多条文物领域的问句数据集，作为问句意图分类的训练测试语料， 并以6:2:2的比例划分了训练集、验证集和测试集。部分数据集样本如下图所示。

```
你知道纹饰为弦纹的文物有哪些吗？  0
回答下仿木纹釉碗属于哪个朝代的    1
告诉我仿朱漆菊瓣式盘这个文物象征什么意义？ 7
请问你知道青花八宝纹双耳宝月瓶是采用了什么工艺制作的吗？    5
帮我查下粉彩勾莲纹天球瓶的出土于哪个年代喔？    1
请搜下掐丝珐琅三足熏炉的材质是什么？  2
替我查下蓝玻璃方花觚这个文物是干什么用的好不好？    3
我想知道青釉狗圈是一种什么文物？  4
说说斗彩瓜蝶纹瓶是怎么制作的？    5
问问你画珐琅开光山水人物鱼缸是什么形状？  5
讲讲黄玉佛手花插属于哪个窑口好吗？    6
查询下白玉羊首耳瓶具有什么寓意吗？    7
建阳窑黑釉盏有多高呢？    8
玉鹰攫人首佩所属哪个文化？    9
清朝时期的文物数量有多少？    10
```

图25 问句意图分类数据样本

data_generator.py的部分核心代码如下：

```
import random
import re
import json
import requests
import urllib.parse


# 前后缀词列表
PREFIX_ASK = ['', '', '', '请问', '你知道', '请问', '你知道', '请问', '你知道', '请问你知道', '你说', '我问你', '我问问你', '我考考你', '考考你',
              '我考一考你', '考一考你', '告诉我', '请告诉我', '请你告诉我', '回答我', '请回答我', '请你回答我', '我想知道']
PLEASE_DO = ['', '请', '请你', '', '请', '请你', '你能', '你会', '你可以', '能不能', '可不可以']
PREFIX_DO = ['', '', '', '', '', '给我', '给我', '给我', '给我', '给我们', '为我', '为我们', '帮我', '帮我们', '向我', '向我们']
DO = ['', '', '', '讲讲', '说说', '讲下', '说下', '讲一讲', '说一说', '讲一下', '说一下', '回答', '所搜', '播放', '放', '查一下', '搜一下', '查', '搜']
SUFFIX = ['', '', '', '', '', '', '', '', '', '', '', '么', '了', '呀', '啊', '吧', '啦', '呢', '吗', '嘛', '呗', '哼', '哈', '嘿', '嗯', '哦', '噢',
          '喔', '行不行', '好不好','好吗', '好吗', '好吗', '好吗', '好吗', '好吗', '好不', '行吗', '行不']
```

```
def gen_by_rule(num):
    positive = []
    negative = []
    for i in range(num):
        entity = random.choice(entities) # 从文物知识图谱中获取对应实体
        part_b = ''
        part_a = ''
        mode = random.random()
        if mode <= 0.1:
            part_b = ''
            part_a = ''
        elif mode <= 0.35: # <文物名称>(出土于/属于/归属于/是/为/归为/产生于)(什么/哪个)(朝代/年代/时代/时期/时间/阶段)
            mode2 = random.randint(0, 5)
            if mode2 == 1:
                part_b = random.choice(PREFIX_ASK)
            elif mode2 == 2:
                part_b = random.choice(DO)
            elif mode2 == 3:
                part_b = random.choice(PLEASE_DO) + random.choice(DO)
            elif mode2 == 4:
                part_b = random.choice(PREFIX_DO) + random.choice(DO)
            elif mode2 == 5:
                part_b = random.choice(PLEASE_DO) + random.choice(PREFIX_DO) + random.choice(DO)
            part_a = random.choice(['出土于','属于','归属于','是','为','归为','产生于']) + random.choice(['什么', '哪个']) +\
                     random.choice(['朝代','年代','时代','时期','时间','阶段']) + random.choice(SUFFIX)
        elif mode <= 0.6: # <文物名称>[的](朝代/年代/时代/时期/时间/阶段)是(什么/什么时候/何时/什么时间/几时/哪时)
            mode2 = random.randint(0, 5)
            if mode2 == 1:
```

4.5.1.2 基于BERT的意图分类模型

本文针对意图分类问题使用了基于BERT的算法模型进行训练，2018年谷歌公开发布了训练好的中文BERT模型BERT-Base-Chinese。BERT-Base-Chinese模型由12个transformer的encoder层，768个hidden，12个注意力heads，110M参数构成。

本文在该模型的输出结果上添加了全连接网络并使用了softmax作为分类器，然后进行BERT模型的微调。其抽象结果如图所示。

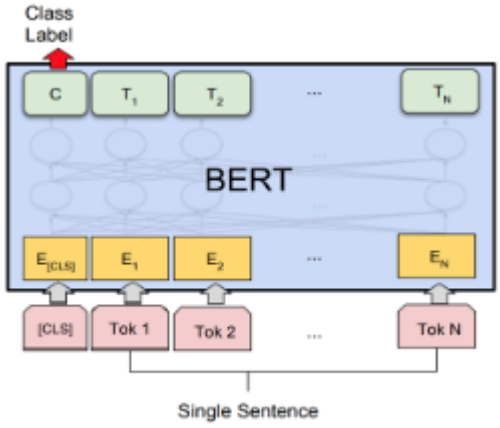


图26 BERT模型结构

图中从下往上，第一行为输入问句的序列，第二行为BERT模型输入层的输出，也就是问句序列的词嵌入，第三行为BERT模型的输出，最上面的ClassLabel指BERT模型输出中“C”对应的向量表征作为整个问句的特征进行分类，“C”指下文中“[CLS]”。具体网络结构如下图所示。

模型的训练过程如下：

1.在BERT模型的输入层阶段，首先将文本中的问题的文本信息都转换为字在字典中所对应的编号，其中字典是利用所有文本数据来构建的字粒度字典，对于字典中不存在的字，将它视为字典中的“[UNK]”，找到“[UNK]”在字典中所对应的编号，这样就得到了图中Token Embeddings。

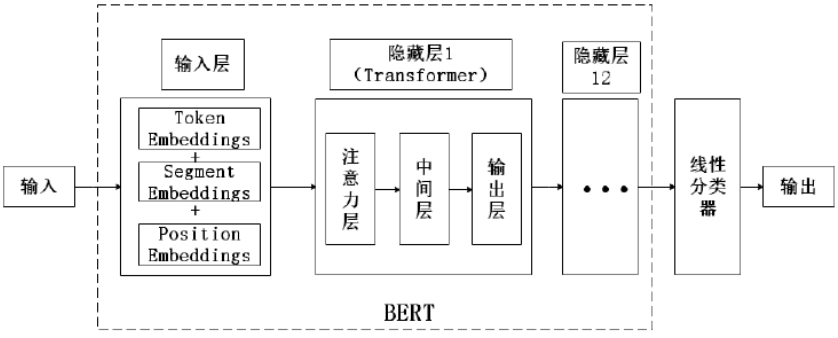


图27 BERT模型具体网络结构

然后根据Token Embeddings得到Segment Embeddings和Position Embeddings，其中Segment Embeddings由于我们这里处理单句输入，只有一种Segment Embedding，即表示问句的文本数据。Position Embeddings中的数字表示字在句子中的位置。

2.通过输入层之后，进入到BERT模型的隐藏层。每个隐藏层由Transformer构成，每个Transformer又由注意力层、中间层和输出层构成。注意力层是Transformer的核心，将输入层的向量进入到Transformer的注意力层，对每一个head，先通过定义的query、key和value的权重矩阵来求对应的query、key和value向量，再将query和key向量相乘，再经过softmax层，然后得到注意力权重矩阵，最后将value向量进行相乘得到最后分数。

3.最后将中间层得到的结果传入到输出层，首先经过全连接、Dropout层最后通过Norm层得到整个Transformer的输出。因为本文使用的是12层的隐藏层，所以循环12次上述操作，最后得到BERT模型的输出。

4.最后将对BERT模型的微调。具体是将文本数据中特殊的“[CLS]”标签的最后一层隐藏层的词向量表征为单个问句的表征，输入到简单线性分类器中，微调BERT模型的参数，完成对问句意图分类的任务。

通过上面BERT模型的微调，最终得到每个句子的词向量，如图所示。

图28 BERT模型训练的句子向量

4.5.1.3 实验分析

模型使用了5折交叉验证方式，将数据集按照4:1的比例进行划分，其中4份数据用于模型的训练，剩下的1份用于模型的验证测试。

为了对比，本文还采用了机器学习方法SVM和TextCNN模型（文本卷积神经网络模型），其中TextCNN模型的参数设定为：词向量维度为300维，序列长度为30，卷积核数量为64，卷积核大小采用多个尺寸[2,3,4,5,6,7]，dropout比例为0.3，每批训练大小为256。

实验结果如下表：

表14 KGQA意图识别实验模型结果

模型	准确率	召回率	F1值
SVM	96.45	96.32	96.38
TextCNN	99.45	99.32	99.38
BERT	99.58	99.47	99.52

从实验结果来看，BERT的模型实验效果最好，但是与TextCNN效果差别也不是很大，都达到了99%以上的准确度，这是因为数据的意图类别之间的语义关系差别比较大，比较容易区分，而且创建的数据集都较规范，规则性强，所以模型的精确度较高。

4.5.2 槽位提取数据集及模型构建

在完成了问句意图模型的分类后，接下来我们对问句中的关键信息进行提取。比如问句“文物出土于清乾隆时期且纹饰为弦纹的文物有哪些？”经过意图识别的结果为文物名称查询，

而经过槽位提取就是要把句中的关键信息抽取出来，抽取出的信息：{朝代：“清乾隆”；纹饰：“弦纹”}。

我们在构建知识图谱时，构建了文物领域的词典，包括文物名称词典、纹饰名、制作工艺方法等属性名词词典，并将自定义的词典导入了jieba工具，使这种文物专有名词不被切割，然后直接用词典对应的方法将文物名词和其他属性名词进行抽取。但是基于这种方法不能完全做到问句的语义理解，可能会把句子中任意位置匹配到的词语抽取出来判定为属性名，比如在文物名词中有“盘”，“鼎”等单字的文物名，如果用户是问“盘类的文物都有哪些”就容易把“盘”字抽取出来当作文物名称，也会把文物名中含有的纹饰属性直接抽取出来，

因此本系统采用了结合基于规则词典和基于深度学习模型的方法对槽位值提取。利用规则和词典的方法用于提取窑口、形状、类型、材质等包含限定值的属性，基于BILSTM+CRF的模型提取文物名称、寓意、朝代、体积类（高、长、宽等量度）等属性。

首先对问句数据进行基于字的标注，共标注了2469条数据，作为BILSTM+CRF模型的训练集和测试集，并且定义了不同属性的标注类型，如文物名称标注定位为B-RELIC和I-RELIC，寓意标注定义B- SYMBOL和I-SYMBOL，标注示例数据如图所示。

```
龙  B-RELIC
泉  I-RELIC
炉  O-RELIC
是  O
哪  O
个  O
朝  O
代  O
的  O
```

图29 标注数据示例

4.6.1.1基于BILSTM+CRF的槽位提取模型

本文将使用Bi-LSTM+CRF模型对槽位进行抽取，槽位提取识别的主要算法流程如图所示，首先将数据集映射为词向量，然后经过Bi-LSTM+CRF模型进行训练，获得标注问句的结果，经实验测试获得92%的准确率。

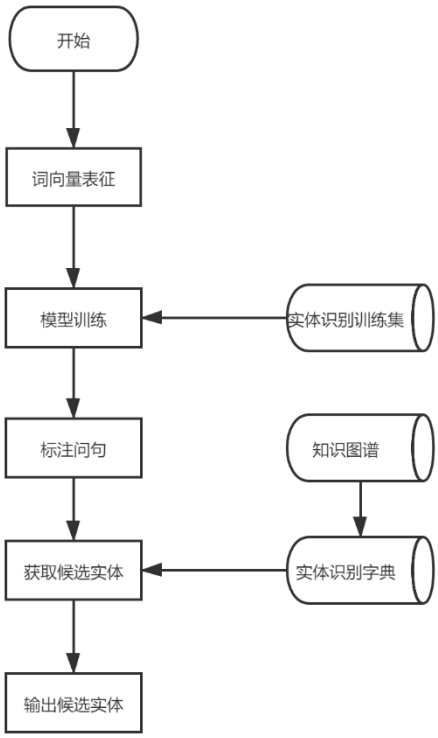


图30 槽位提取识别算法流程

模型训练算法过程如下：

表15 模型训练算法过程如下

算法：基于Bi-LSTM+CRF模型的槽位提取算法
输入：标注实体的问句序列以及实体字典 输出：问句实体的标注标签 方法： 1）准备数据集，将训练集准备输入到模型中 2）将训练集的问句通过word2vec映射为词向量表征，记为字符Q，并按批次输入至模型中 3） 初始化Bi-LSTM+CRF模型，基于梯度下降的方法来计算网络的权重参数 4）对问句Q_{user}进行分词处理 5）对问句Q_{user}中每个单词输入到隐藏层得到编码处理（上下文状态） 6）然后输入到解码层对Q_{user}进行计算 （全连接层） 7）将第6）步骤的输出送到sigmoid层，得到句子中局部最大的概率输出 8）将sigmoid的输出当作CRF的输出，基于CRF的约束，得到整体最大的概率输出 9）输出最后问句的标注结果

4.6.1.1 基于规则与词典槽位提取模型

基于规则和词典的槽位提取本系统采用了基于REFO库的规则库，REFO的意思为“对象的正则表达式”。REFO是基于python开发的一个项目库，提供与python十分相似的re模块功能，但re模块仅使用了字符串匹配，REFO可对任意序列进行匹配，并且匹配规则不仅是每个对象匹配相等，还可以匹配为任意python函数。

其使用的语法规则与re模块略有不同，与依存句法分析思想类似，类似构建正则表达式的语法树。具体匹配步骤如下所示：

1）先基于构建的文物领域词典去匹配问句中包含的关键词

2）对关键词对象进行定义，每个单词有两个对象属性：该词的文本值和词性，需要同时匹配这两个属性才算匹配成功。

3）定义将词典的词使用关键词定义进行实例化

4）定义关键词和问句的匹配函数

5）构建文物匹配规则

6）构建函数匹配问句中的关键词即属性值和对应的属性

7）将返回的属性和属性名填入槽位信息中

通过上节Bi-LSTM+CRF模型返回的结果，将提取出的槽位与使用字典提取的槽位结合，至此就完成了问句中所有槽位的提取。将提取出的槽位信息经过分析发现，部分的信息需要进行槽位值的改写，本文通过词典进行修正，主要是对朝代、形状等属性的修正。

表16 槽位值改写示例

用户问题中的实体	修正实体
清朝时期	清朝
清政府	
大清王朝	
大清	
清王朝	
晚清	
清廷	
乾隆年间	清乾隆
嘉庆	清嘉庆

4.6 问句转化与查询

在获得了问句意图，类别和句子的槽位信息后，接下来我们要将这两部分信息结合起来转化为知识图谱可识别的检索语句。我们根据不同的问句意图类别制定了相对应的SPARQL查询语句模板，然后将抽取并改写好的问句中的槽位信息填入对应的SPARQL查询语句模板中。

如对于问题“青花八宝纹双耳宝月瓶是采用了什么工艺制作的？”生成的查询问句如图所示。

```
PREFIX : <http://kg.BLCU.edu/ontology/property/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?o WHERE {
  ?s :中文名 "青花八宝纹双耳宝月瓶".
  ?s :工艺 ?o.
}
ORDER BY ?o
```

图31 SPARQL查询语句示例

4.7 知识库答案检索

本系统采用了Apache Fuseki作为SPARQL的服务器，配置好Fuseki server后，将之前的构建好的文物领域的RDF三元组数据上传到Fuseki服务器中，运行fuseki server，基于文物的知识图谱检索服务就搭建好了，将生成的相对应SPARQL查询语句通过问答检索模块，向Fuseki server发送HTTP请求，利用Fuseki Server提供的查询API进行知识检索，得到问句的答案，并返回给用户终端。

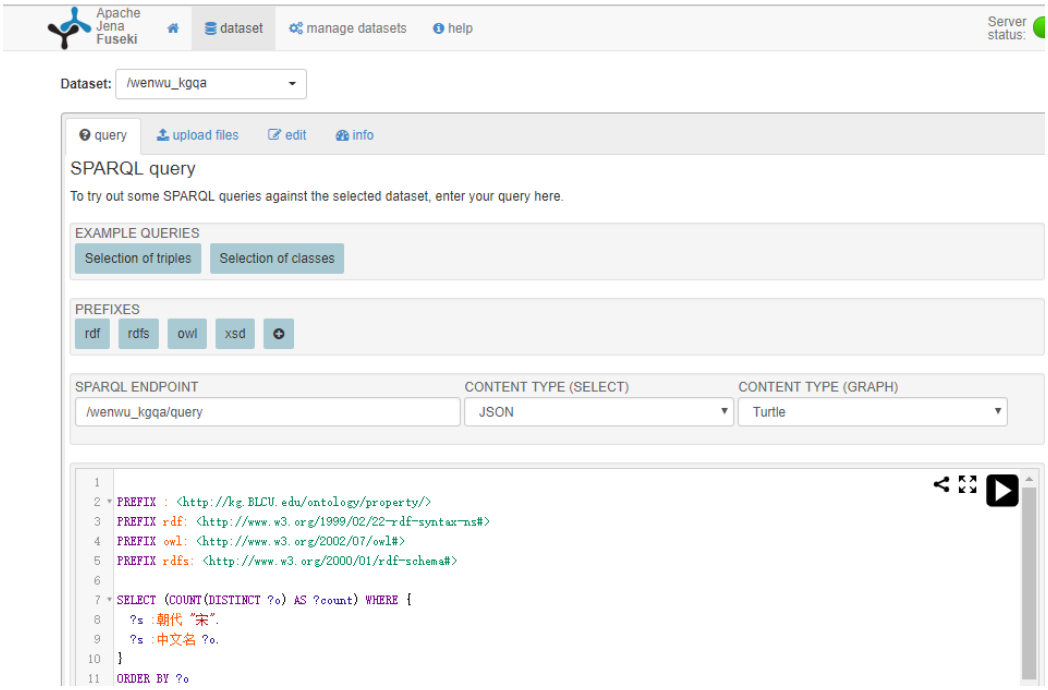


图32 fuseki server查询界面图

4.8 本章小结

本章主要详细介绍了构建智能问答系统模型步骤，本文的问答系统模型主要分为三个大模块：问句理解模块：包括用户询问意图识别及句中关键信息抽取（即槽位识别）、问句转换模块和答案检索模块。其中，详细介绍了基于BERT的意图识别的模型和结合Bi-LSTM和规则字典的槽位提取的模型算法，在意图识别模型实验数据集的构建中，本文提出了基于特征词集的自动构建问句数据集算法，共生成2万对条数据集。之后详细介绍了BERT模型的算法流程，随后在该算法的基础上进行了实验方面的设计，并给出实验结果。在槽位提取的模型中，先详细介绍了Bi-LSTM+CRF的模型原理和训练过程，并给出对应的实验结果，然后详细介绍了基于规则和词典的槽位提取方法，最后将两者算法的结果合并为最终结果。将问句理解模块的输出转化为SPARQL语句，送入到答案检索模块中，最终得到问句的答案。本章的实验结果表明，该系统设计的算法能够

有效的基于文物知识图谱提供问答服务，同时该系统框架可以很方便的根据其他领域需求加入其他规则或模型，可以有效的进行融合，兼具比较好地移植性和稳定性。

第五章 文物助手系统设计与实现

5.1 引言

基于第四章构建的文物领域智能问答模型，本章在此基础构建了一个智能文物助手系统。该系统提供了智能问答、语音问答、文物知识扩展、查博物馆信息等功能，本系统基于微信公众号开发。

5.2 系统需求分析

5.2.1 背景需求

互联网的高速发展，给人们的生活带来非常多的便捷，越来越多的产品和信息进行数字化转型。虽然我们国家的文博资源十分丰富，但是真正用户使用的频率却并不高，具研究统计，我们国家的民众平均两年才进入博物馆一次，然而在处于欧美地区的国家，当地民众走进博物馆的次数平均达到三到五次。其中原因，一方面，社会的快速发展使大众的心态也逐渐日益浮躁，很难静下心来去专注一件事，更甚，随着手机应用端的蓬勃发展，大家的业余时间都花费在手机上，走进博物馆去感受文物的魅力成为一件大众兴趣度不高的事件;另一方面，博物馆的展陈方式略显呆板陈旧，同时游览博物馆时深奥晦涩的讲解让人感觉乏味单调。为了让博物馆更加“鲜活”起来，打破之前的刻板僵化印象，打开其中的“奇妙”，让文物真正活起来，同时是为了充分挖掘，阐释及传播文物价值，文博专家和学者正借助科技的手段，让文物资源活起来，推广文物知识，复原文物历史，不再将文物简单的陈列在博物馆，而是以多种形式多方面展示文物特点，挖掘文物背后的历史，以生动的讲述形式和多样的展示方式，跟着文物感受时代的风云涌动，朝代的更替变化，探索未知的历史记忆。中央也一直高度重视文化工作的推进，致力提升民族的文化自信，号召响应，各大博物馆都在对文物资源进行数字化整理，文物与现代科技正在进行深度的合作与融合。

在这背景需求下，因此提供一个用户友好的文物知识领域的应用，能够满足用户的获取文物知识的需求，同时，综合各种文物信息到同一应用入口使用户能够方便快捷获取自己想要的信息，是十分必要且有意义的工作。

5.2.2 功能需求

本文设计的文物助手包含以下功能：

（1）文物智能问答

我们提供这样的一个智能问答接口，可供用户随时查阅，比如用户在逛博物馆时对某个文物产生兴趣，想对文物进行进一步了解，用户可使用自然语言的方式对该助手进行友好的交互，基于构建好的文物知识图谱的智能问答系统，可以对用户提供的问句进行方便快速的回答。

（2）语音问答

在日常生活中，常常存在不方便用文字的形式进行交互的情况，也存在小孩、老人或其他特殊用户对手机使用情况的不太熟悉，为了方便更多的受众用户，开放语音的接口显得十分必要，该系统提供普通话转文字的功能，用户可以很方便的通过微信的语音端口进行对文物知识的查询。

（3）每日文物知识

该功能主要目的是扩展文物知识，当用户想了解更多的文物知识，该文物助手可随机推送一个文物知识点，包括文物基本信息，文物相关的历史事件。

（4）博物馆资讯

当用户制定一个游览博物馆的计划时，一般需要对该博物馆网站去搜寻相关的信息，网站信息繁杂，找寻自己想要的信息需要花一定时间。该文物助手提供了对博物馆相关信息查询的接口，该功能主要提供对博物馆相关信息的查询，包括博物馆基本介绍、开放时间、近期展览等信息。

5.2.3 性能需求

本文构建的文物助手系统目前还在实验阶段，因此在性能方面可以满足以下的要求即可：

1）易用性：用户使用流畅，交互性良好，用户通过自然语言的方式或者语音的方式即可对服务请求；同时结果展示清晰明了。

2）实时性：所有的用户请求均在秒级别可得到回应。

3）准确性：能够针对用户不同的询问意图可以正确识别且得到相对应合适的回复。

5.3 系统关键功能设计与实现

5.3.1 系统架构

智能文物主要系统架构如图所示。该系统从上而下分别为服务层、技术层及数据层。服务层主要是面向各阶层对文物知识有需求的用户如游客用户、学生用户等，本助手系统主要基于微信公众号开发，在微信端对用户提供了文物智能问答、语音问答、每日文物知识及博物馆资讯等服务；技术层主要包含了问答系统所需的算法模型，包括意图识别、槽位提取和问句转化与查询等技术方法；数据资源层主要为存储文物领域知识图谱构建好的 RDF 数据、每日文物知识功能所需的文本资源和查询博物馆资讯的语料数据。

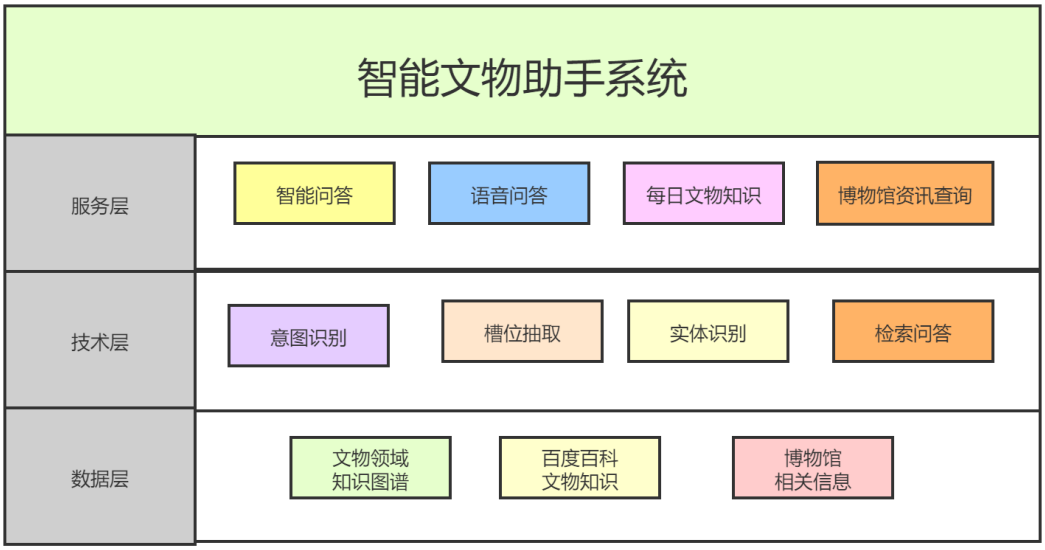


图 33 智能文物助手系统架构

5.3.2 系统实现

5.3.2.1 智能问答模块实现

智能问答模块时基于文物知识图谱的 QA 模型，该模型的具体结构在第四章进行了详细说明，本章主要是将 QA 模型迁移到微信公众号的服务端，整体的模型技术不会有改动，增加了对微信公众号的信息接收和发送到微信公众号的数据发送的相关技术。具体流程是用户输入一条语句至微信公众号，该代码进行捕获进行转化格式并发送到 QA 模型的输入端，然后通过训练好的模型，经过一些列操作，返回问句的答案，并将答案处理成微信端所需要的格式，最终将显示在屏幕上。具体流程图所示。

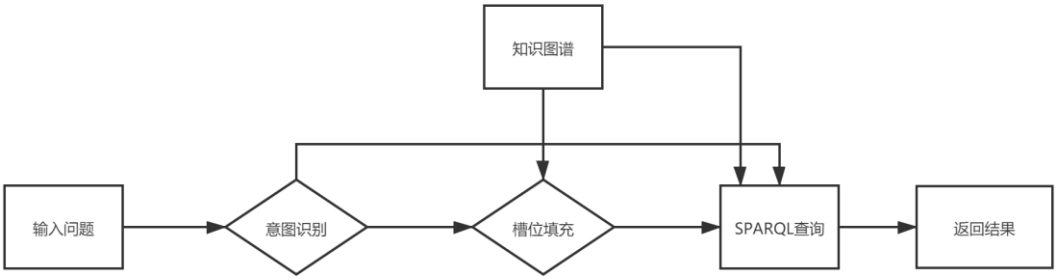


图 34 智能问答模块流程

5.3.2.2 语音问答模块实现

本系统基于微信公众号开发，微信公众号提供丰富的智能接口，其中就提供了语音识别的接口，主要实现步骤如下：

（1）首先登陆你申请的微信公众号，进入到平台中。

（2）然后进入开发目录下的接口权限菜单，并在接口权限页面下，微信公众号提供了十分多的接口服务，这里选择接收语音识别结果这项接口服务，并将接口设置为开启状态。

图35 接口权限页面

（3）开启语音识别接口后，当用户发送语音信息至公众号时，微信服务端将接收的语音信息存至XML数据包中，同时也会增加一个Recongnition字段。语音信息的XML数据包如下，其中Format代表语音格式，一般存为amr格式，Recongnition为语音识别的结果，使用UTF8编码保存。

```
<xml>
<ToUserName><![CDATA[toUser]]></ToUserName>
<FromUserName><![CDATA[fromUser]]></FromUserName>
<CreateTime>1357290913</CreateTime>
<MsgType><![CDATA[voice]]></MsgType>
<MediaId><![CDATA[media_id]]></MediaId>
<Format><![CDATA[Format]]></Format>
<Recognition><![CDATA[腾讯微信团队]]></Recognition>
<MsgId>1234567890123456</MsgId>
</xml>
```

图36 XML语音数据包

（4）对用户输入的语音数据进行语义解析，对语义接口进行封装，这一步使用智能问答的语义解析模块。

（5）微信端识别语音信息，修改接收信息的接口，增加对语音信息的接收，并对其进行回复。

```python
class Msg(object):
    def __init__(self, xmlData):
        self.ToUserName = xmlData.find('ToUserName').text
        self.FromUserName = xmlData.find('FromUserName').text
        self.CreateTime = xmlData.find('CreateTime').text
        self.MsgType = xmlData.find('MsgType').text
        self.MsgId = xmlData.find('MsgId').text
        self.Content = xmlData.find('Content').text

class TextMsg(Msg):
    def __init__(self, xmlData):
        Msg.__init__(self, xmlData)
        self.Content = xmlData.find('Content').text.encode("utf-8")


class ImageMsg(Msg):
    def __init__(self, xmlData):
        Msg.__init__(self, xmlData)
        self.PicUrl = xmlData.find('PicUrl').text
        self.MediaId = xmlData.find('MediaId').text

class VoiceMsg(Msg):
    def __init__(self, xmlData):
        Msg.__init__(self, xmlData)
        self.Recongnition = xmlData.find('Recongnition').text.encode("utf-8")
```

```python
def POST(self):
    try:
        webData = web.data()
        # 后台打印日志
        print('Handle Post webdata is ', webData)
        recMsg = receive.parse_xml(webData)
        if isinstance(recMsg, receive.Msg):
            toUser = recMsg.FromUserName
            fromUser = recMsg.ToUserName
            if recMsg.MsgType == 'text':
                # result = ""
                question = recMsg.Content
                result = query.parse(question)
                replyMsg = reply.TextMsg(toUser, fromUser, result)
                return replyMsg.send()
            if recMsg.MsgType == 'image':
                mediaId = recMsg.MsgId
                replyMsg = reply.ImageMsg(toUser, fromUser, mediaId)
                return replyMsg.send()
            if recMsg.MsgType == 'voice':
                question = recMsg.Recongnition
                result = query.parse(question)
                replyMsg = reply.VoiceMsg(toUser, fromUser, result)
                return replyMsg.send()
            else:
                return reply.Msg().send()
```

图37 修改API接口服务增加语音信息

5.3.2.3 每日文物知识及博物馆资讯查询模块实现

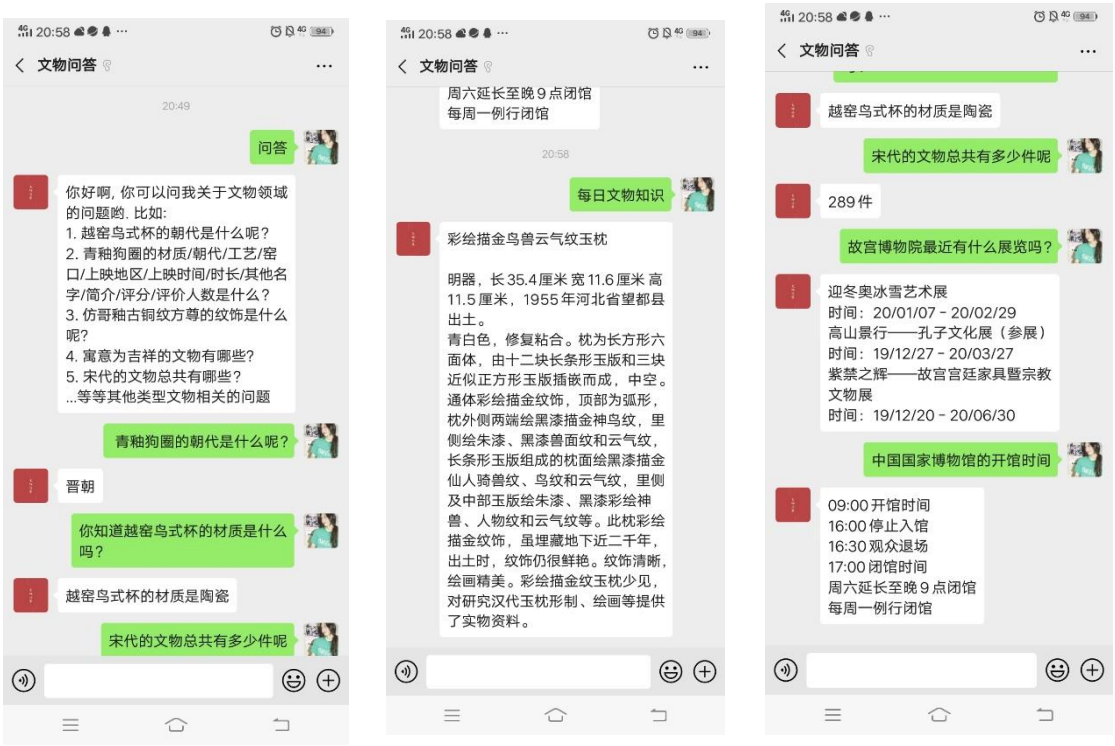
在本文中，这两个模块都主要是通过关键字进行匹配，如果用户输入每日文物知识，则系统会自动从文物知识库中随机抽取一条信息对用户进行推送。该模块除了使用知识图谱中的信息，本文还基于百度百科爬取了1322条相关信息作为补充，并进行了整理。数据整理如下图。

{"文物名": "狩猎纹骨饰", "知识信息": "装饰品，高7.7厘米、直径4.5厘米，1956年内蒙古自治区包头市郊出土。此饰件用兽骨制成，呈圆筒形，一端平齐，另一端斜口。 匈奴以
{"文物名": "针灸画像砖", "知识信息": "墓室内装饰图像，长94.5厘米、宽91.5厘米、厚24厘米，山东省微山市两城出土。禽兽及牧畜为生。外壁用针刻画出飞鸟、奔跑的野猪、
{"文物名": "日晷", "知识信息": "计时仪器，边长27.4厘米、厚3.5厘米，1897年内蒙古自治区托克托出土。此日晷为方形，用致密泥质大理石制成。晷面中央有一直径1厘米的圆
{"文物名": "观伎画像砖", "知识信息": "墓室内装饰图像，长45.4厘米、宽40厘米、厚5.3厘米，1954年四川省成都市扬子山出土。此砖的左上方绘一男一女席地而坐，其中男者头
{"文物名": "漆盒石砚", "知识信息":"文具，长21.5厘米、宽7.4厘米，1978年山东省临沂市金雀山出土。此砚盒木胎，盖内外髹赭漆，内有长方形凹槽可扣住石板砚，有方形凹槽
{"文物名": "陶猪圈", "知识信息":"明器，高15.5厘米，湖南省长沙市出土。此陶猪圈模型由圆形围墙构成，围墙上有宽檐以保护墙壁，圈内卧猪。猪圈外高台上架筑厕所，下部与
{"文物名": "绘龙虎纹陶壶", "知识信息":"明器，通高48.5厘米，口径18.8厘米，足径18.1厘米，1953年河南洛阳烧沟汉墓出土。壶直口，长颈束腰，溜肩鼓腹，口上有扁圆，圈足
{"文物名": "彩绘描金乌兽云气纹玉枕", "知识信息":"明器，长35.4厘米 宽11.6厘米 高11.5厘米，1955年河北省望都县出土。青白色，修复粘合。枕为长方形六面体，由十二块长
{"文物名": "酿酒画像砖", "知识信息":"墓室内装饰图像，高28.4厘米、宽38.3厘米，1954年四川省彭县出土。此画像砖反映了酒肆酿酒和销售的情景。画面正中是一妇人正在大釜
{"文物名": "彩绘石骑马人", "知识信息":"明器，高78厘米，长77.2厘米，宽25厘米，1955年河北省望都县出土。此石雕系用整块石灰石雕成，骑马的僮仆头戴黑色平巾帻，身着红
{"文物名": "军司马印", "知识信息":"汉代官印，海东市蒲家墩出土。铜质、方形，桥形纽，印面阴刻篆书“军司马印”四字。军司马为部曲官职中的一级。按汉代军制，司马一职在
{"文物名": "三羊铜樽", "知识信息":"盛酒器，高21厘米，口径23.6厘米。此樽为隆盖，盖顶中间有-环钮。外有3个卧羊形钮。器身如圆筒形，直壁，腹两侧有兽衔环耳。平底，下
{"文物名": "部曲陶俑", "知识信息":"明器，高38.8厘米，四川省崖墓出土。此俑身穿短襦，交领右衽，左手执箕，右手执铲，腰佩环首刀，是东汉豪强大族私家武装部曲或家丁的
{"文物名": "铜朱雀衔环杯", "知识信息":"宽9.5厘米 通高11.2厘米，河北满城陵山二号汉墓出土。通体错金。器形为朱雀衔环矗立于两高足之间的兽背上。朱雀昂首翘尾，喙部衔
{"文物名": "鎏金银螭龙纹铜壶", "知识信息":"腹径37厘米 口径20.2厘米 通高59.5厘米，1968年于河北省满城县陵山出土。小口微侈 ，鼓腹圈足，上腹部饰一对铺首衔环。盖为
{"文物名": "错金博山炉", "知识信息":"薰炉，高26厘米，足径9.7厘米，1968年于河北省满城县陵山出土。器型似豆形，盖肖博山，通体错金。座把呈透雕三龙出水状，龙首顶托

图38 文物知识数据示例

5.4 系统展示

该文物助手系统以微信公众号为前端为用户提供查询服务。图 a 展示了文物智能问答服务，可实现对文物的基本信息问答。图 b 展示了每日文物知识推送，图 C 展示了对博物馆资讯的查询。



图（a） 图（b） 图（c）

图 39 文物助手功能展示

5.5 本章小结

本章在构建好的文物知识图谱的问答系统基础上构建了智能文物助手系统，对问答功能进行了扩展并迁移到移动端，包括智能问答，语音问答、每日文物知识和博物馆资讯这四个功能模块。本章对该系统进行了详细描述并介绍了系统架构和具体的实现路线。最终以微信公众号的前端形式为用户提供服务。

第六章 总结与展望

6.1 总结

本文完成构建了基于本体的文物知识图谱的工作，然后在此基础上构建基于文物知识图谱智能问答系统，并将该系统迁移至微信公众号，并对其功能进行了进一步扩展，增加了语音识别功能、每日文物知识和博物馆资讯查询功能，能够很方便快速对用户的输入的信息进行处理并返回相应结果，为传播文物信息有一定的价值。

在构建基于文物领域的知识图谱时，我们首先分析了基于文物领域的知识特点，然后结合实际应用场景的考量，根据了本体构建理论，对本体构建方法的七步法和循环获取法进行结合和改进，提出了基于知识图谱应用场景的文物领域本体构建方法，并利用protégé建模工具定义了本体的属性及其概念之间的语义关系，完成了文物本体的构建；接下来从故宫博物院的藏品页面爬取了部分文物数据作为知识库的数据源，并提出了基于特征词集的文物属性知识抽取方法将web非结构化文本中具有的属性知识抽取出来，完成本体的实例化。最后将本体知识库存到Neo4j图数据库中，可实现文物知识的可视化和查询检索工作。

智能问答系统可以为用户提供更加高效、直接并准确的知识获取方式，相比于传统的基于搜素引擎的知识获取方式，智能问答系统更符合人们自然语音交互的方式，给人们获取知识提供了新的便利。本文基于文物知识图谱构建了智能问答系统，为用户获取文物知识提供遍历。首先我们设计了智能问答系统的框架，本文将问答系统框架分为三个大模块：问句理解模块：包括用户询问意图识别及句中关键信息抽取（即槽位提取）、问句转换模块和答案检索模块。首先将问句进行在意图识别模块基于BERT模型完成对问句的意图分类，使用基于Bi-LSTM模型和字典匹配模型的结合完成了槽位提取功能，最后将问句的意图和属性信息进行转化为SPARQL语句，将其送到基于Apache Fuseki搭建的SPARQL查询检索API中，完成对问句的解析返回最终答案。

设计了智能问答系统框架后，接下来本文将搭建一个基于问答的文物助手服务，可以回答还文物的基础信息，另外还增加了语音问答、每日文物知识和博物馆资讯查询的功能，以微信公众号的展示形式为用户提供服务。

6.2 本文研究的局限及对相关研究的展望

本文主要介绍了基于文物领域知识图谱构建过程，并在此基础上构建了智能问答系统，最终实现了一个智能文物助手系统，为用户获取文物知识提供新的便利。但是在论文整个流程中还存在了一些不足的地方：

（1）在从web文本中根据特征词集的方式对属性进行自动抽取时，由于这种基于特征总结的方法无法覆盖全方位的信息，还存有遗漏的属性尚未抽取出来，无法达到较高的准确率，还需要进一步探索更加好的属性抽取方法。

（2）目前所构建的智能问答模型虽然可以支持基本的文物信息的查询，但是对一些较复杂的问句还无法提供解析进行回答，还需要进一步完善问句的推理工作，或者引入一些额外的信息作为知识补充融合到智能问答模型中也是一个可探索的研究方向。

（3）本文构建的文物助手系统的功能尚不完备，可在此基础上进行进一步完善，可以增加更多的模块，如对文物信息更加细粒度的分类查询服务、相似文物推荐等更多与文物相关的服务。

参考文献

[1] Berners-Lee T, Hendler J, Lassila O. The semantic web[J]. Scientific american, 2001, 284(5): 34-43.

[2] Zneika M, Lucchese C, Vodislav D, et al. Summarizing linked data RDF graphs using approximate graph pattern mining[C]. 2016.

[3] 韩道军, 甘甜, 叶曼曼, 等. 基于形式概念分析的本体构建方法研究[J]. 计算机工程, 2016 (2016 年 02): 300-306.

[4] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods[J]. Semantic web, 2017, 8(3): 489-508.

[5] 陈悦, 刘则渊, 陈劲, 等. 科学知识图谱的发展历程[D]. , 2008.

[6] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. 2007: 697-706.

[7] http://www.wikipedia.org/

[8] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A never-ending Chinese knowledge extraction system[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.

[9] Lian H, Qin Z, He T, et al. Knowledge graph construction based on judicial data with social media[C]//2017 14th Web Information Systems and Applications Conference (WISA). IEEE, 2017: 225-227.

[10] 李嘉锐, 崔运鹏, 张学福, 等. 水稻本体实例构建研究[J]. 数字图书馆论坛, 2014 (11): 43-47.

[11] 李明鑫, 王松. 近十年国内知识图谱研究脉络及主题分析[J]. 图书情报知识, 2016, 4: 93-101.

[12] 孟祥龙,樊洲,肖洋,等.中药炮制学学科研究前沿、热点及发展趋势---基于知识图谱的“中药炮制”可视化研究[J].世界科学技术K-中医药现代化,2015(7):1514-1524

[13] 余菜花,廉同辉,刘军.中国低碳研究的知识图谱分析[J].资源科学,2012(10):1959-1964

[14] 程赛琰,丁磊,魏淑娟.基于知识图谱分析的电子政务研究现状、热点与趋势[J].图书与情报,2013(1):116-123

[15] 谢靖,章鑫鑫.基于CSSCI(2000-2011)的中国文学学科知识图谱研究[J].图书与情报,2014(2):108-114

[16]李伟平,权德庆.我国体育消费研究前沿与热点---基于科学知识图谱的可视化研究[J].西安体育学院学报,2014(1):41-44
[17]辛伟.知识图谱在军事心理学研究中的应用[D].西安:第四军医大学,2014
[18]BERNERS-LEE T., Linked data[EB/OL].(2016-02-23).https://www.w3.org/Design Issues/Linked Data.html.
[19]Initiative, D.C.M., et al., Dublin core metadata element set,version 1.1[EB/OL].(2012-06-14). http://dublincore.org/documents/2012/06/14/dces/.
[20]Initiative, D.C.M., et al., Dcmi metadata terms, dcmi recommendation[EB/OL].(2012-06-14). http://dublincore.org/documents/2012/06/14/dcmi-terms/.
[21]Bechhofer, S., Miles, A., Skos simple knowledge organization system reference[R]. W3C recommendation, 2009.
[22]Martin Doerr. The CIDOC CRM-an ontological approach to semantic interoperability of metadata. AI Magazine[J],2003, 24(3):75-92.
[23]Condron, L., Tittemore, C.P.:Functional requirements for bibliographic records,2004.
[24]Antoine Isaac(ed.).Europeana data model primer[EB/OL](2017-08-22). Europeana technical document.http://pro.europeana.eu/edmdocumentation.
[25]万静，严欢春，邢立栋.浅析知识图谱在智慧博物馆中的应用前景[C].互联网时代的数字博物馆, 2017.
[26]Cherny, E., Haase, P., Mouromtsev, D., et al., Application of CIDOC-CRM for the Russian Heritage Cloud Platform[C]. Proceedings of New Trends in Databases and Information Systems:ADBIS 2015Short Papers and Workshops.September 8-11.2015:448-457.
[27]翟琼,刘宏哲. CIDOC CRM在数字博物馆中的应用[J].中国博物馆, 2015(2):26-30.
[28]Antoine I A, Jacco V O A, Guus S A. Amsterdam Museum Linked Open Data[J]. Semantic Web, 2013, 4(3):237-243.
[29]闫晓创.欧洲文化遗产资源的在线整合实践研究[J].中国档案, 2017(4):74-75.
[30]林炀平.文物知识图谱构建与检索关键技术研究与实现[D]. 浙江大学 2017
[31]邱超. 基于Web文本的文物知识图谱自动生成方法研究[D]. 西北大学 2016
[32]张娜. 文物知识图谱构建关键技术研究与应用[D].浙江大学,2019.
[33]故宫博物院.故宫博物院92周年院庆“发现•养心殿——主题数字体验展”端门数字馆开展[EB/OL] https://www.dpm.org.cn/show/246075.html 2017.10.10

[34]李婷.国内首个博物馆数字平台建成，上博参观路线最优解，董其昌朋友圈一手掌握[EB/OL] https://wenhui.whb.cn/zhuzhan/xinwen/20180504/196957.html 2018.05.04

[35]Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge acquisition, 1993, 5(2): 199-221.

[36]Carroll J J, Dickinson I, Dollin C, et al. Jena: implementing the semantic web recommendations[C]//Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. 2004: 74-83.

[37]Wang X, Jiang X, Liu M, et al. Bacterial named entity recognition based on dictionary and conditional random field[C]//2017 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2017: 439-444.

[38]Sun W. Chinese named entity recognition using modified conditional random field on postal address[C]//2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2017: 1-6.

[39]Zhang L, Zhao H. Named entity recognition for Chinese microblog with convolutional neural network[C]//2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2017: 87-92.

[40]Wang X, Zhou X, Li D, et al. Military Scenario Named Entity Recognition Method Based on Deep Neural Network[C]//2018 International Conference on Control, Automation and Information Sciences (ICCAIS). IEEE, 2018: 137-140.

[41]Yu K, Zhao T, Zhao P, et al. Extraction of protein-protein interactions using natural language processing based pattern matching[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017: 1292-1295.

[42]Boonpa S, Rimcharoen S, Charoenporn T. Relationship extraction from Thai children's tales for generating illustration[C]//2017 2nd International Conference on Information Technology (INCIT). IEEE, 2017: 1-5.

[43]Huang Y, Jia Y, Huang J, et al. Multi-language person social relation extraction model based on distant supervision[C]//2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2018: 368-374.

[44]Xue L, Qing S, Pengzhou Z. Relation Extraction Based on Deep Learning[C]//2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). IEEE, 2018: 687-691.45

[45] Dewi I N, Dong S, Hu J. Drug-drug interaction relation extraction with deep convolutional neural networks[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017: 1795-1802.

[46] Zettlemoyer L S, Collins M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[J]. arXiv preprint arXiv:1207.1420, 2012.

[47] Zettlemoyer L S, Collins M. Learning context-dependent mappings from sentences to logical form[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 976-984.

[48] Kwiatkowski T, Choi E, Artzi Y, et al. Scaling semantic parsers with on-the-fly ontology matching[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1545-1556.

[49] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 1533-1544.

[50] Berant J, Liang P. Semantic parsing via paraphrasing[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 1415-1425.

[51] Fader A, Zettlemoyer L, Etzioni O. Open question answering over curated and extracted knowledge bases[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 1156-1165.

[52] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models[C]//Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014: 165-180.

[53] Yao X, Van Durme B. Information extraction over structured data: Question answering with freebase[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 956-966.

[54] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models[C]//Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, Heidelberg, 2014: 165-180.

[55] Weston J, Chopra S, Bordes A. Memory networks[J]. arXiv preprint arXiv:1410.3916, 2014.

[56] Yih W, He X, Meek C. Semantic parsing for single-relation question answering[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014: 643-648.

[57] Yang M C, Duan N, Zhou M, et al. Joint relational embeddings for knowledge-based question answering[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 645-650.

[58] Yih W, He X, Meek C. Semantic parsing for single-relation question answering[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014: 643-648.

[59] Yang M C, Duan N, Zhou M, et al. Joint relational embeddings for knowledge-based question answering[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 645-650.

[60] Zhang Y, Liu K, He S, et al. Question answering over knowledge base with neural attention combining global knowledge information[J]. arXiv preprint arXiv:1606.00979, 2016.

[61] Xie Z, Zeng Z, Zhou G, et al. Knowledge base question answering based on deep learning models[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 300-311.

[62] Yang F, Gan L, Li A, et al. Combining deep learning with information retrieval for question answering[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 917-925.

[63] Lai Y, Lin Y, Chen J, et al. Open domain question answering system based on knowledge base[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 722-733.

[64] 杜泽宇, 杨燕, 贺暉. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件, 2017, 34(5): 153-159.

[65] 王银丽, 高光来. 限定领域内智能问答系统的研究与实现[C]//2008 年全国模式识别学术会议. 中国自动化学会, 2009: 426-431.

作者在攻读硕士学位期间的科研情况

发表论文情况

1. Deng Y, Lu M, Li H, et al. A Deep Learning Baseline for the Classification of Chinese Word Semantic Relations[C]//Workshop on Chinese Lexical Semantics. Springer, Cham, 2018: 630-642.

2. Lu M, Liu P. Denoising Distant Supervision for Relation Extraction with Entropy Weight Method[C]//China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019: 294-305.

.

致谢

只有当写到致谢时，才深深感到毕业的临近。如何整理这三年来的回忆，如何与过去的美好告别，想到过完这一两个月我就要真正离开美好的象牙塔生活，我心底开始泛起淡淡的涟漪。人们总是不太擅长离别，面对离别的我同样手足无措，可是人生就是这样，旅途很长，我们开始整理行囊，然后勇敢的告别，从这一站下车，开始迈向下一站。我将这三年来的点点滴滴一一收藏至心中，将这三年来没有说出的感谢浓缩至这短短的篇章。

首先要感谢的是我的导师—刘鹏远老师，我尤记得调剂的时候，那时我发邮件的第一个老师就是刘鹏远老师，和老师通了很多封的邮件，那些文字中无不透露出老师爆炸般的亲和力，还记得那时和老师聊的真的很开心，所以也很毅然加入老师的研究团队。在与老师相处的三年中，发现除了亲和力，更让我钦佩的是老师认真负责的工作态度，为教学科研项目熬夜也是不罕见的事件，有好几次我们组的人在为投稿会议论文熬夜撰写中，没想到老师也陪着我们一起连夜为我们指导，实在感动。在我撰写论文的过程中，也是时常督促我，由于我拖延的毛病在递交初稿时，差不多凌晨12点才交上，我原以为老师会在第二天给我进行修改，没想到老师还是连夜替我修改了，感受到了老师作为科研者的钻研认真的态度。在生活中，老师也经常给予关心与帮助，在我感到迷茫时也给了我很多宝贵的建议。能够成为刘老师的学生感到很幸运。

其次，还要感谢信科院的老师们，于东老师、邵艳秋老师、杨尔弘老师、荀恩东老师、项若曦老师、罗智勇老师等等，你们严谨认真的教学态度让我学习到很多，信科院的老师永远最棒！

然后还要感谢我SOTA的伙伴们（刘老师的研究团队），在这个团队中，我收获了很多，我们一起科研一起奋斗一起玩耍，度过了很开心的三年时光，这里，特别感谢和我一起开学一起毕业的杜成玉同学，三年来，我们相互依赖相互陪伴相互分享，在我不开心的时候总是能安慰我，遇到棘手的事情也能替我想办法解决，还记得我们一起去约伴去吃重庆小面一起去故宫游玩，这些平凡的点滴都将成为我人生中美好的回忆。然后还要感谢赵硕丰、潘月、张颖、王伟康、胡晗、毕洪亮、田永胜等师弟师妹们，你们对待生活的热情，对待学业的认真，使我也深深受到鼓舞。

接下来，还要感谢我亲爱的父母，在我写论文期间给了我莫大的支持，并且为了使我安心能够完成论文，给我创造了写论文的绝佳环境，让我只要每天吃好饭就行，家务活都不让我干，而且还为了我营养时刻担忧着，每天变着样给我做菜，各种水果也安排上。无论何时，家里永远都是最温馨的港湾，父母永远是我最大的后盾，感谢我亲爱的家人们，你们永远是我前进的最大力量，爱你们。

2020年，由于新冠病毒的爆发，使我们的生活遭受到了重大改变，在这特殊的背景下，最后将最真挚的感谢送给我们无私的医护人员及所有为之奋斗的人们，是你们的奉献与付出，才让我们中国的疫情尽快的控制住。目前，世界疫情还在继续蔓延着，希望世界疫情早点结束，早日回归到正常的日常生活中，希望生病的人早点康复，希望健康的人平平安安，希望世界和平，Love And Peace。