Software Engineering and Applications 软件工程与应用, 2024, 13(2), 190-198
Published Online April 2024 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2024.132020

Hans汉斯

基于知识图谱的档案领域问答系统研究与应用

王建林，陈萌萌，冶存花，魏天楠

西北民族大学数学与计算机科学学院，甘肃 兰州

收稿日期：2024年3月13日；录用日期：2024年4月19日；发布日期：2024年4月29日

摘　要

在信息化时代的迅速发展下，每天都会产生大量的文书档案数据。然而，当前这些数据的利用率并不高，用户的检索效率也较低。为了改善这一状况，提出了一种基于知识图谱的自动问答系统。首先，利用自然语义处理技术(Stanford NLP)对责任者进行实体识别和关系抽取，以丰富档案知识图谱。通过这项技术，能够从文档中识别特定的实体，如人名、地点、组织机构等，并了解它们之间的关系，从而构建一个丰富的知识图谱。为了存储和管理这些信息，选择了Neo4j作为数据库。Neo4j是一个图数据库，非常适合存储和查询具有复杂关系的数据，这与知识图谱非常契合。其次，设计并实现了一个基于知识图谱的问答系统，其核心功能在于通过模板问答方法进行信息检索。通过结合自然语义处理技术和知识图谱，问答系统能够理解用户提出的问题，并在知识图谱中进行关联查询，以找到准确的答案。为了让用户更容易操作，还使用Flask框架开发了一个轻量级且易于使用的Web界面。通过整合知识图谱，问答系统可以利用自然语言处理技术为用户提供准确、快速且个性化的答案和服务。这不仅提高了文书档案数据的利用率，也提升了用户的检索效率。

关键词

知识图谱，档案，Neo4j，Flask

Research and Application of Archive Domain Question Answering System Based on Knowledge Graph

Janlin Wang, Mengmeng Chen, Cunhua Ye, Tiannan Wei

School of Mathematics and Computer Science, Northwest Minzu University, Lanzhou Gansu

Received: Mar. 13th, 2024; accepted: Apr. 19th, 2024; published: Apr. 29th, 2024

文章引用: 王建林, 陈萌萌, 冶存花, 魏天楠. 基于知识图谱的档案领域问答系统研究与应用[J]. 软件工程与应用, 2024, 13(2): 190-198. DOI: 10.12677/sea.2024.132020

Abstract

In the rapidly developing era of information technology, a large amount of documentary archive data is generated every day. However, the utilization rate of this data is currently not high, and the efficiency of user retrieval is also low. In order to improve this situation, a knowledge graph-based automatic question answering system is proposed. Firstly, natural language processing technology (Stanford NLP) is used to identify entities and extract relationships among responsible parties, enriching our archive knowledge graph. This technology helps identify specific entities such as names, locations, and organizations from documents and understand the relationships between them, thereby building a rich knowledge graph.Neo4j was chosen as the database to store and manage this information. Neo4j is a graph database that is well-suited for storing and querying data with complex relationships, making it a perfect fit for our knowledge graph. Secondly, a question answering system based on the knowledge graph is designed and implemented, with its core function being information retrieval through template-based question answering. By combining natural language processing technology and the knowledge graph, our question answering system can understand user queries and perform related queries in the knowledge graph to find accurate answers. To make it easier for users to operate, a lightweight and user-friendly web interface was developed using the Flask framework. By integrating the knowledge graph, our question answering system can provide users with accurate, fast, and personalized answers and services using natural language processing technology. This not only improves the utilization rate of documentary archive data but also enhances user retrieval efficiency.

Keywords

Knowledge Graph, Archives, Neo4j, Flask

1. 引言

随着互联网和信息数字化的快速发展，档案管理正经历数字化转型，以便于管理、保存和提升档案信息的检索及共享效率。档案，作为重要的文化遗产和信息资源，在传统管理中面临信息碎片化和检索困难等问题，迫切需要利用新技术提升管理效率和质量，推动档案工作向智能化、数字化方向发展。

中国档案馆藏档案量的迅速增加以及向公众开放的档案数量的持续上升，显示了档案数量和可访问性的显著增长，这为档案的信息化建设和共享创造了更广阔的空间。2019 年末，中国档案馆的档案总量达到 8.29 亿卷，年增长率为 10.39%，公开档案量达到 1.32 亿卷，年增长率为 17.38%。在此背景下，全国档案局长馆长会议提出加快推进档案信息化战略转型，强调利用互联网、云计算、大数据、人工智能等新兴技术改善档案管理和推动档案信息资源共享服务平台的建设。

档案领域受大数据和云计算等理念的深刻影响，面临新的挑战和机遇。传统档案服务方式已难以满足日益增长的多样化档案信息资源的利用需求。档案部门需紧跟信息化技术发展，创新管理方式，提高档案查询和共享效率，以实现档案信息资源的更有效利用。这一切凸显了档案行业对信息化技术的迫切需求和其在当前社会中的重要性。

2. 国内外研究现状

2.1. 档案信息数字化

随着多媒体技术的不断进步，档案信息的数据类型变得日益丰富，从传统的纸质档案逐渐转变为来自多个来源、形式各异的数字档案。数字档案馆利用计算机技术将各类档案文件电子化，构建了综合的档案信息管理平台。档案数字化可分为两个层次：浅层次是对档案标题和目录信息进行数字化，而深层次则是对档案全文信息进行综合利用。

当前，大多数数字档案馆仍处于档案信息数字化的初级阶段。随着计算机技术的不断发展，许多机构开始尝试将知识图谱技术应用到数字档案馆中，以实现档案的智能检索和全文信息化。例如，欧洲委员会信息社会技术(IST)的 Good Practice Guidelines 项目为数字档案馆的建设提供了指导[1]。而世界电子图书馆项目在互联网上提供了来自全球各地的多语种档案资料，为档案领域的数字化转型提供了示范和参考。

人工智能技术在多媒体信息提取方面取得了显著进展，对档案知识的抽取尤为关键。例如，美国卡内基梅隆大学开发的 Mormedia 项目是最早的内容分析系统之一，利用语音引擎分析音频信号，能够从音频材料中提取关键信息。此外，香港中文大学的 IViEW 系统提出了一种全新的基于内容特征提取的数字视频管理方案。国内，浙江省公安厅的数字档案室在文字识别技术方面取得了显著成效，并被评为 2018 年“全国示范数字档案室”。

总的来说，数字档案馆的进步和智能检索技术的运用为档案信息的管理和应用提供了新的可能性和工具。利用知识图谱的构建技术，我们能够深度挖掘档案之间的关联性，为智能检索档案提供了一种创新的方式。

2.2. 知识图谱技术

知识图谱作为描绘实体间关系的语义网络，在知识工程领域扮演着重要角色。全球范围内，已经开发出多个知识库资源，包括国际上的 Freebase、Wikidata、DBpedia、YAGO，以及国内的复旦大学发布的中文概念图谱 CNProbase 等[2]。本研究提出的知识图谱架构专注于档案领域，与泛化的通用知识图谱不同，它利用特定领域的知识来快速构建知识库，比如医疗、地理、军事、农业知识图谱等[3]。研究工作主要聚焦于基于现有元数据(如 EAD、Dublin Core 等)探索元数据语义互操作性及其映射关系[4]。在实践中，有学者通过引入语义本体概念在企业档案数据应用中进行语义分析，并在此基础上建立了联通电子档案知识图谱系统[5]；另有学者利用 Protégé、OWL 等技术构建了科研档案的知识图谱语义模型[6]；也有学者针对数字人文发展，提出了档案时空本体模型和数据抽取框架，构建了档案关联数据知识图谱，实现可视化展示[7]。此外，还有学者通过语义分析文档，并建立了云服务推荐系统[8]。

3. 系统架构

3.1. 系统结构

相比于传统检索方法，自动问答系统在处理简单、模板化的问题上展现了更高的速度、针对性和准确性，同时其返回的结果也更易于用户理解。该系统主要由四个模块组成：数据收集、图谱构建、问题理解、以及用户界面。系统的整体架构见图 1。

本系统的数据主要来源包括甘肃省档案局提供的低密级档案文件、政府网站发布的公务文件以及部分文书档案目录数据。在知识图谱构建阶段，首先对收集的文书档案目录数据进行处理和整合。然后，利用 Stanford NLP 库对责任者和题名中的实体及其关系进行抽取。考虑到档案数据中包含大量的专有名词，部分未能识别的内容需要手动进行划分。接着，使用 Python 语言构建知识图谱，并将这些数据存储

在 Neo4j 数据库中。问题理解模块与用户界面紧密相关，因为问题理解部分需要处理用户在界面中输入的查询。为了方便用户操作，我们利用 Flask 框架构建了一个轻量级的 Web 网站，用户可以在该网站上输入问题并获取查询结果。

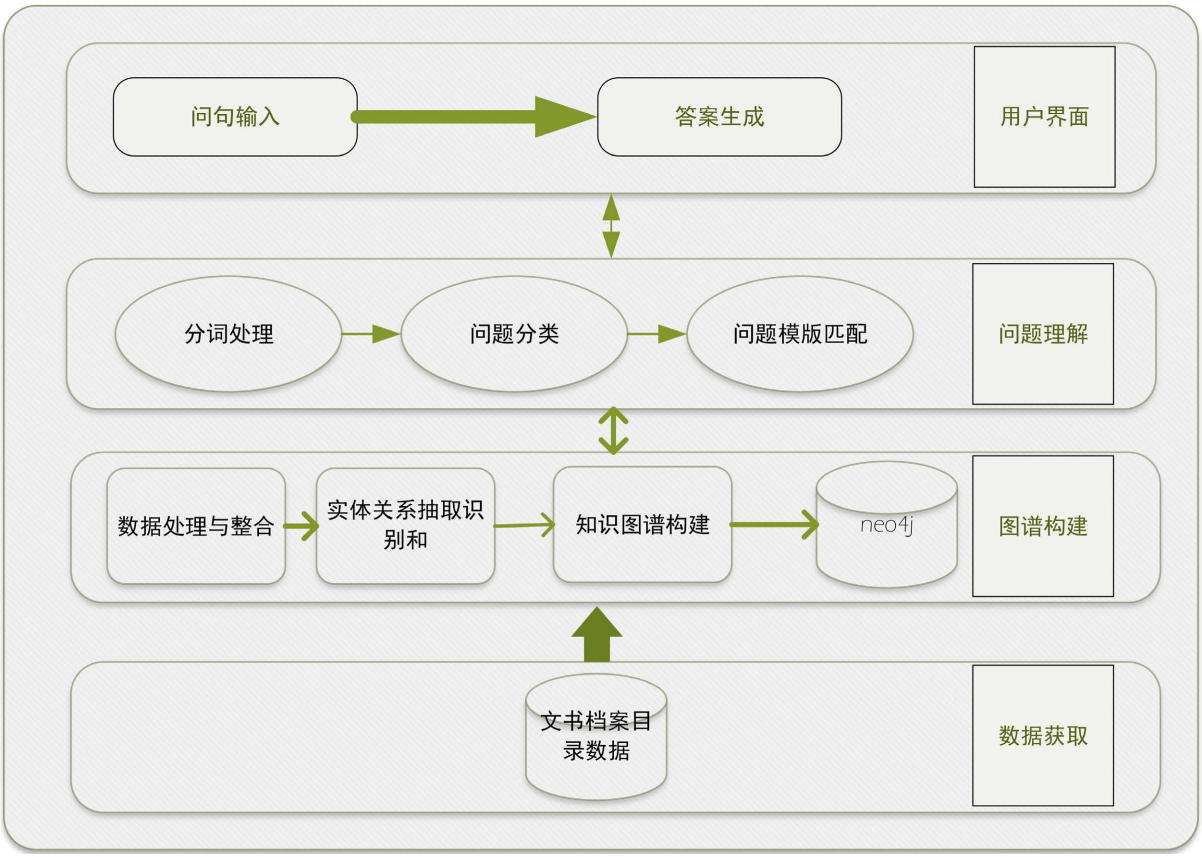

Figure 1. Architecture of archive directory-based question answering system
图 1. 档案领域问答系统架构

3.2. 问题处理流程

Aho-Corasick 算法是一种用于字符串搜索的高效算法，由 Alfred V. Aho 和 Margaret J. Corasick 在 1975 年共同提出。该算法能够在一个主文本字符串中同时查找多个模式串(即“子串”或“关键词”)，并且不论搜索多少个模式串，时间复杂度都保持为线性，使其在许多应用场景中非常有效，特别是在处理大量数据时。本文进行问题分类时使用了 Aho-Corasick 算法，其作用主要是进行了特征词提取。问答系统的应用的具体流程见图 2。

1) 问题输入：输入档案领域相关的问题与档案查询问题，例如，蔡向海档案在哪？

2) 问题过滤：主要过滤敏感词和非法词。

3) 待处理：待处理包括异常空格处理、分词、关键词提取、异常修正和分类。在进行关键词提取时使用的是 Aho-Corasick 算法。Aho-Corasick 算法的核心思想是构建一个有限状态机(也称为“trie 树”或“关键词树”)，其中包含了所有模式串的信息。这个有限状态机不仅包含了模式串的所有字符，还预计算了失败指针(failure pointer)，这些指针在搜索过程中当前字符不匹配时提供了回退的路径。因此，即使在面对大量不匹配的情况下，算法也能快速跳过那些不需要的字符，直接转移到下一个潜在的匹配位置。

4) 问题分析：主要就是将提取的关键词和问题的类别，进行 CQL 查询语句的翻译。

5) 答案输出：将查询的结果进行组织输出到 Web 页面供用户查看。

3.3. 问题分类器

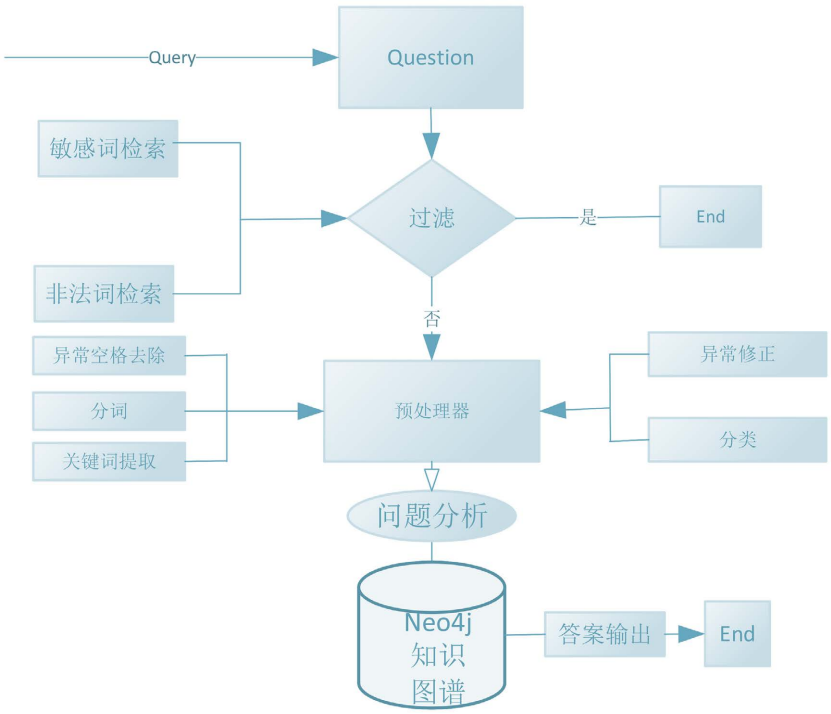

Figure 2. Question processing workflow
图 2. 问题处理流程

将知识图谱中的实体概念和属性等词加入领域词库，同时初始化分词器，完成领域分词器的构建。针对用户对档案查询的需求，共定义 6 类问题类别，见表 1 所示。

Table 1. Question categories
表 1. 问题类别

问题标识	问题类别	例子
archive_search	查档案	与市交通局相关的档案在哪里？
responsibleperson_secarch	查责任者	文件编号的档案资料的责任者有哪些
archive_name_search	查档案馆	蔡向海档案在哪?
creat_data_search	查创建时间	某档案什么时候创建的
directory_structure_search	查目录结构	某档案的目录结构
regulations_inquiry	询问规章制度	查阅档案要遵守什么规定
service_time_inquiry	询问服务时间	什么时候可以查阅档案

当问题输入时，首先进行疑问词识别(‘谁’、‘在哪’、‘什么时候’等词)构造出问句类别向量，问题分类大多是从统计学的角度进行分类。由于本文初步问题分类别少，特征突出，所以本文基于 LibSVM [9]进行多分类器的训练。

4. 档案领域知识图谱的构建

4.1. 数据整理与预处理

数据主要来源于所整理的文书归档目录数据，原始数据有 616,124 条。见图 3。

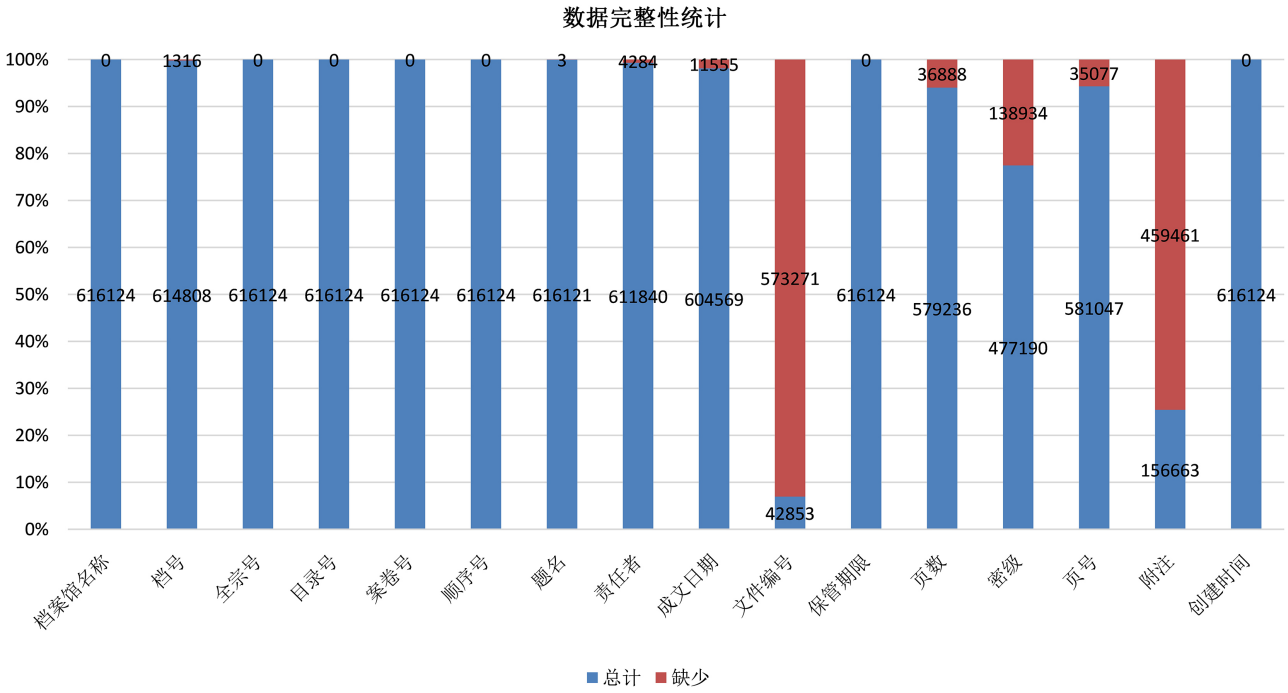

Figure 3. Data integrity statistics
图 3. 数据完整性统计

完整性统计显示，数据集中存在缺失值的列包括文件编号、附注、密级、页数、成文日期、责任者、档案号和题名。文件编号的缺失值最多，而题目的缺失值最少。因此，有必要对原始数据进行预处理。

在处理含有空值和重复数据的数据集时，特别是在面对具有复杂标识符的数据集，如临夏回族自治州档案馆的档案数据，需要采取细致的方法来确保数据的准确性和完整性。数据集中存在较多的空值，尤其是在文件编号和附注列，这可能会在数据分析和建模时引发问题。为了解决这些问题，可以采取几种策略。

首先是空值处理。针对档号列的 1316 个缺失值，可以通过全宗号、目录号、案卷号和顺序号这四个属性的组合来填充档号的缺失值。对于其他列中的空值，可以采用“无”进行填充，以确保数据的完整性。

其次是删除重复数据。通过使用 Python 的 Pandas 库进行数据查重后，发现存在重复的记录。为了解决数据冗余问题，可以选择删除重复数据的策略，保留信息最完整的一条记录，删除其余重复项。这一步骤有助于提高数据集的质量和分析的准确性。

经过空值填充和删除重复数据的处理后，数据集从原始状态经过清洗，最终剩余 501,136 条记录。这一过程不仅减少了数据的冗余，还提高了数据集的利用率和分析的可靠性。

综上所述，通过对数据集中的空值进行填充、删除重复记录等步骤，可以有效地提升数据的质量，为后续的数据分析和建模打下坚实的基础。在处理类似数据时，重要的是识别出数据中的特殊模式和问题，并采取适当的方法来解决这些问题，以确保分析结果的准确性和可靠性。

4.2. 实体和关系抽取

接下来，需要对档号和成文日期进行规范化。对于档号，按照全宗号、目录号、案卷号、顺序号的顺序排列，不可颠倒。一般有两种形式：一种是用“－”相接，多用于检索工具上；另一种是将各代码填于专栏或戳记内，多用于案卷的封面或脊背上。全宗号由档案馆指定给立档单位的代码组成，案卷目录号表示全宗内案卷所属目录的代码，案卷号则表示案卷目录内每一案卷的顺序号。为了减少冗余，可以将全宗号、目录号、案卷号、顺序号四个字段省略，因为它们已经构成了档号。

对于成文日期，需要进行规范化，不满足条件的部分用 0 进行补足。在进行时间格式检查后发现许多时间格式不正确，因此只保留了年份。

此外，责任者字段有些是无效的，直接将其去除。

另外，保管期限分为永久、定期 30 年和定期 10 年，分别用代码“Y”、“D30”、“D10”标识。因为密级和保管期限可以作为分类的条件，所以将其作为档案下面的两个实体。

最后，为了减少冗余，添加一个时间实体对其进行分类。

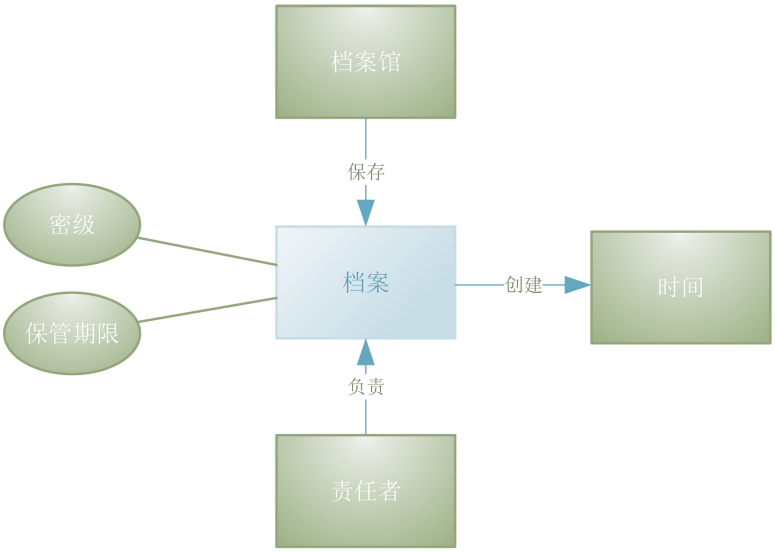

Figure 4. Basic structure of knowledge graph
图 4. 知识图谱基本结构

综上所述，共有六个实体分别是档案馆、档案、责任者、时间、密级、保管期限。见表 2。关系有五个，见表 3。知识图谱基本结构见图 4。

Table 2. Entity table
表 2. 实体表

实体	实体名	说明
档案馆	ArchiveName	例如甘肃省档案馆、兰州市档案馆
档案	Archive	包含档号、题名、附注、页号、文件编号等属性
责任者	ResponsiblePerson	如氮肥厂、甘宁青邮政管理局
时间	CreationData	年
密级	SecurityLevel	如公开、无密、保密
保管期限	RetentionPeriod	如永久、长期、30 天等

Table 3. Relationship table
表 3. 关系表

关系名	说明
BELONGS_TO	档案属于档案馆
HAS_CREATION_DATE	档案创建于哪一年
HAS_RESPONSIBLE_PERSON	档案的责任人是谁
HAS_RETENTION_PERIOD	档案的保存期限
HAS_SECURITY_LEVEL	档案的密级

4.3. 实体和关系构建

知识图谱的构建通常使用 Neo4j 数据库配合 Python 编程语言进行，这是目前比较主流的方式。Python 语言具有很好的扩展性，拥有许多方便调用的包。在进行知识图谱构建之前，需要先安装好 Neo4j 数据库、Python 以及相关的包，比如 py2neo。要注意，Neo4j 的安装需要在 Java 环境上，因此必须先安装好 Java 并查看其对应的版本，以确保兼容性。

在 Python 中配置好 Neo4j 的相关信息后，可以调用 py2neo 中的 Graph、Node 和 Relationship 进行关系构建。首先，需要进行数据收集，将处理好的数据使用 pandas 库读入 Python 中，然后调用 Node 进行节点创建，调用 Relationship 进行关系创建。当完全创建好之后，就可以在 Neo4j 中查看创建好的知识图谱了。

这种方法可以通过图形界面或者 Cypher 查询语言在 Neo4j 中查看构建好的知识图谱。

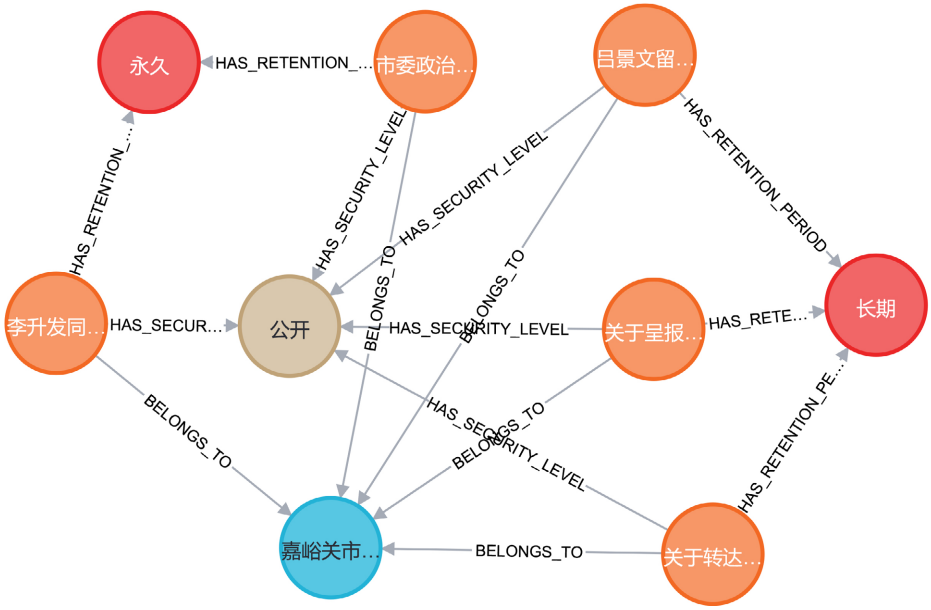

Figure 5. Example of Knowledge Graph in Archive Domain
图 5. 档案领域知识图谱示例

5. 结论

本研究主要致力于构建基于档案知识图谱的问答系统。主要任务包括构建档案知识图谱的方法、开发基于多跳知识图谱问答的模型，以及建立基于文书档案知识图谱的问答系统等。我们通过构建文书档

案本体和利用 Stanford NLP 进行命名实体识别及关系抽取，成功填补了档案领域知识图谱的空白。构建完成的部分知识图谱示例见图 5。

尽管已经取得了一些成果，但仍有改进的空间。首先，我们需要更新和扩充档案知识图谱。尽管知识图谱在处理大规模数据时具有优势，但由于资源限制，我们目前的档案知识图谱仅基于实验室现有的结构化文书档案目录数据。在未来，我们将考虑引入更多的结构化和半结构化数据，以丰富知识图谱，提高其完整性和有效性。

其次，需要提高实体关系识别的准确性。虽然像 Stanford NLP 这样的工具在处理中文文本方面表现出色，但在文书档案数据集上的表现还有待提升。在未来的工作中，我们将考虑使用迭代添加词典等方法，以提高模型在文书档案数据上的识别准确率。

总的来说，通过以上改进，期望在后续的工作中进一步优化问答系统，提升其在档案知识图谱构建、实体识别和关系抽取等方面的性能，以实现更高的研究和应用价值。

参考文献

[1] 郭雪薇, 董晶. 基于特征关联分析的档案信息关联模型[J]. 电子设计工程, 2019, 27(1): 47-52.

[2] 雷洁, 李思经, 赵瑞雪, 等. 面向科研档案管理的知识图谱构建与应用研究[J]. 数字图书馆论坛, 2020(5): 8-15.

[3] 王电化, 钱涛, 钱立新, 等. 面向档案的知识图谱构建方法研究[J]. 湖北科技学院学报, 2020, 40(1): 127-130.

[4] 周程, 戴贵奇, 周卓畅, 等. 基于知识图谱的数字档案服务模式探究[J]. 兰台内外, 2023(26): 1-3.

[5] 杨茜雅. 中国联通电子档案数据挖掘与智能利用的研究[J]. 档案学研究, 2018(6): 105-109.

[6] 雷洁, 赵瑞雪, 李思经, 等. 知识图谱驱动的科研档案大数据管理系统构建研究[J]. 数字图书馆论坛, 2020(2): 19-27.

[7] 舒忠梅. 数字人文背景下的档案知识图谱构建研究[J]. 山西档案, 2020(2): 53-60.

[8] Balaji, B.S., Karthikeyan, N.K. and Kumar, R. (2018) Fuzzy Service Conceptual Ontology System for Cloud Service Recommendation. *Computers & Electrical Engineering*, **69**, 435-446. https://doi.org/10.1016/j.compeleceng.2016.09.013

[9] 张巍, 陈俊杰. 信息熵方法及在中文问题分类中的应用[J]. 计算机工程与应用, 2013, 49(10): 129-131, 179.