油田环保安全领域标准智能问答关键技术研究

鲁小辉[1] 王凯月[2]

（1.中国石油化工股份有限公司安全监管部；2.中国石油化工股份有限公司胜利油田分公司技术检测中心 ）

摘 要：油田环保安全领域标准对于规范和引导油田行业安全生产、绿色发展和效率提升具有重要意义。油田环保安全领域标准知识复杂程度较高，难以形成对标准数字知识的双向理解路径，为有效解决上述问题，本论文进行油田环保安全领域标准智能问答关键技术研究。首先，进行FAQ引擎设计，包括研究基于语义相似度的问题快速匹配技术、基于深度学习的相似度重排技术，对用户行为进行评分；其次，进行KGQA引擎设计，包括研究语义库设计模型和基于Graph的搜索匹配模型；最后，设计多引擎加权打分机制，能够实现油田环保安全领域标准智能问答。

关键词：油田环保安全领域，标准，智能问答

DOI编码：10.3969/j.issn.1674-5698.2024.04.009

Research on Intelligent Q&A Technology for Oilfield Environmental Protection and Safety Standards

LU Xiao-hui[1] WANG Kai-yue[2]

（1. Safety Supervision Department of China Petroleum and Chemical Corporation

2.Technology Inspection Center of Shengli Oilfield, SINOPEC, Dongying, Shandong, China）

Abstract: The standards in the field of oilfield environmental protection and safety are of great significance for regulating and guiding the safety production, green development, and efficiency improvement of the industry. The standard knowledge in this field is relatively complex, which is difficult to be understood. To effectively solve the above problems, this paper conducts research on key technologies for intelligent Q&A of standards in the field of oilfield environmental protection and safety. Firstly, it designs the FAQ engine, including the research on the fast matching techniques based on semantic similarity and similarity rearrangement techniques based on deep learning, to rate user behavior; Secondly, it designs the KGQA engine, including the research on semantic library design models and Graph based search matching models; Finally, it designs a multi engine weighted scoring mechanism that can achieve intelligent Q&A in the field of oilfield environmental protection and safety standards.

Keywords: oilfield environmental protection and safety field, standards, intelligent Q&A

作者简介：鲁小辉，硕士，高级工程师，研究方向为石油安全管理。

王凯月，本科，工程师，研究方向为石油标准化、信息化研究。

0 引言

随着大数据、云计算、人工智能等新一代信息技术的发展和应用逐渐走向成熟，日益渗透到经济社会的各个领域。在油田环保安全领域，我国标准数字化应用水平总体处于纸质标准电子化、结构化、语义化的初级数据建设阶段，缺乏可交互标准数字化应用和探索。油田环保安全领域标准知识复杂程度较高，难以形成对标准数字知识的双向理解路径，一方面标准间的数据关联关系及数据结构复杂，可能导致机器无法准确理解人类多轮提问需求；另一方面机器生成内容可能不符合人的阅读和理解逻辑，导致人类无法高效全面理解机器生产的内容，造成人员提出的问题与机器给出的答案不匹配等问题。

针对上述不足，本文旨在深入探讨油田环保安全领域标准智能问答的关键技术。首先聚焦于FAQ（Frequently Asked Questions）引擎的设计。这一阶段主要包括基于语义相似度的问题快速匹配技术的研究，以及基于深度学习的相似度重排技术。这些技术的目的是为了更准确、迅速地识别和匹配用户提出的问题。此外，研究还涉及对用户行为进行评分，这有助于了解用户需求，从而提高问答系统的效率和准确性。其次，文章着眼于KGQA（Knowledge Graph Question Answering）引擎的设计。在这一部分，研究集中于语义库设计模型和基于图（Graph）的搜索匹配模型的开发。这些模型旨在利用知识图谱，以更加复杂和高级的方式处理和回答问题，使得答案更为精确和全面。最后，研究提出了一种多引擎加权打分机制的设计。这种机制能够综合FAQ引擎和KGQA引擎的优势，通过加权打分来确定最优的回答方案。通过这种方法，可以实现更为高效和准确的油田环保安全领域标准智能问答，从而大幅提升信息检索和问题解决的质量。整体而言，本研究在智能问答技术领域为油田环保安全标准的应用提供了一种新的视角和方法论。

1 研究现状综述

在这个信息爆炸与科技高速发展的时代，如何从繁杂的海量数据中提取准确所需信息，成为了研究的重点。全问答（question answering, QA）作为解决这一问题的关键手段之一，通过对数据信息进行检索、分析，最终向用户提供问题的答案或相关信息。问答系统在语言建模的核心挑战在于如何更好地理解和处理自然语言。其中，词嵌入（word embedding）方法扮演了重要角色。词嵌入是一种将词语转换为向量的技术，可以使计算机更好地理解词语之间的关系和含义。传统的词嵌入方法主要包括Mikolov T等人于2013年提出的Word2Vec[1]和Pennington J等人于2014年提出的GloVe[2]模型。Word2Vec模型是通过将词汇量化为向量，实现了对词语之间关系的定量度量，主要包括连续词袋模型（CBOW）和跳跃模型（skip-gram）两种模型。 在CBOW模型中，通过上下文来预测目标单词；而在skip-gram模型中，则是通过目标单词来预测上下文。这两种模型都能有效地捕捉词与词之间的联系。GloVe模型则与Word2Vec有所不同。GloVe更加关注单词同时出现的概率比率，而非简单地关注共现概率分布。它的特点在于不需要计算共现次数为零的单词对，从而减少了计算量和数据存储空间。GloVe通过分析整个数据集的词汇共现信息，从而更好地捕捉到单词间的全局关系。这两种方法都在问答系统的语言理解能力提升上起到了关键作用。通过这些先进的词嵌入技术，问答系统能够更有效地处理复杂的语言信息，更准确地理解用户的问题，并提供相关的答案。然而，这些传统的词嵌入方法在处理词义多样性、上下文灵活变化等方面还存在局限性，这也推动了后续如BERT[3]等更先进的语言表示模型的发展。

BERT是一个具有里程碑意义的自然语言处理（NLP）模型，由Google在2018年推出。它的核心是采用了生成式的掩码语言模型（Masked Language Model, MLM）和双向Transformer结构。BERT的训练分为两个阶段，首先是预训练阶段，BERT在大规模的数据集（如：BooksCorpus和英语维基百科[4]）上进行无监督学习，这一阶段的目的是让模型学习语言的基本规律和模式；其次是微

调阶段，针对特定的NLP任务（如：情感分析、问答系统、命名实体识别等），将任务相关的数据输入到预训练好的BERT模型中，并进行参数的微调，使模型适应特定的任务。

问答系统想要满足用户需求，主要需处理3个问题：问题分析、信息检索和答案生成。根据问答系统信息源的数据类型的不同，可将问答系统分为：（1）数据来源于结构化知识图谱的问答系统；（2）数据来源于对话、问答对的基于问答对的问答系统；（3）数据来源于自由文本的基于机器阅读理解的问答系统。其中，关于基于知识图谱问答系统的应用，大多集中在特定领域，如：医疗领域、金融投资领域、电商领域、聊天机器人领域等。基于问答对的问答系统使用较为普遍，早期美国在为用户提供航班信息时开发的SLS项目，欧盟开发的列车时刻信息系统和保险合约查询电话呼叫中心等[5]。目前，各大IT公司也开发了各种聊天机器人，如：苹果的Siri、微软小冰、小米、小爱等。基于机器阅读理解的问答系统是由计算机自动根据给定的语料资料来回答用户所提出的问题，目前受到了越来越广泛的关注，与FAQ、KBQA等优势互补，形成更完备、更智能的问答系统。

油田环保安全领域标准智能问答技术研究面向特定领域，相较于一般的智能问答系统具有更高的专业性和准确性。因此，油田环保安全领域标准问答系统在构建过程中，重点采用基于知识图谱、基于机器阅读理解的方法来开展智能问题系统研究。

2 油田环保安全领域标准智能问答引擎设计

2.1 FAQ引擎

（1）工作原理

FAQ引擎是基于常见问题的相似比对引擎，主要涉及收集并组织常见问题及其答案，然后通过用户界面使用户能够搜索或浏览这些问题。当用户提出查询时，引擎通过关键词匹配或使用自然语言处理技术来理解用户的查询意图，从而检索出最相关的答案。这些答案随后以易于理解的格式呈现给用户。此外，FAQ系统通常会收集用户反馈，以不断优化答案的准确性和相关性，有时还会集成聊天机器人来提供更实时的互动。

（2）基于语义相似度的问题快速匹配技术

基于语义相似度的问题快速匹配技术是一种智能技术，用于计算用户输入问题与数据库中存储问题之间的语义相似度。它通过分析和理解问题的意义而不仅仅是关键词，能够识别出语义上最为接近的问题。这种技术运用自然语言处理（NLP）和机器学习算法，以确保提供的候选答案在语义上与用户的查询尽可能匹配，从而提高问题解答的准确性和效率。

（3）基于深度学习的相似度重排技术

应用深度学习技术，将问题与候选答案进行深度语义比对，计算用户输入与候选答案之间的相似度，根据计算结果，对候选答案集进行重新排序。

（4）用户行为评分

对于推荐的相似问题，如果用户点击后，系统会记录该事件，对当前提问问题与点击的相似问题建立一个关系。相似问题之间的关系可在系统投票选举环节，进行加权评分，提高推选答案的排名。

2.2 KGQA引擎

（1）工作原理

先对油田环保安全领域标准问题库进行梳理，形成知识图谱的三元组结构，在此基础上，定义基于知识图谱的问题模板，构建问题模板库。

油田行业知识图谱，是油田领域标准的结构化语义知识库，用于以符号形式描述物理世界中的概念及其相互关系，其基本组成单位是【实体，关系，实体】或者【主体，谓词，客体】三元组，以及实体及其相关属性值对，实体之间通过关系相互联结，构成网状的知识结构。

在问答系统中，把三元组定义为【主体，属性，答案】，这样在问答过程中，通过给定主体与属性两个维度查找知识图谱中的答案。

问答过程主要分为两个步骤，其一是问题理解，通过问题理解识别到该问题的具体意图，也就是问题的知识主体与知识属性；其二是答案搜索，

根据问题理解获得的知识主体与知识属性，查找知识图谱中的具体答案。例如：油田含油污泥处置后泥渣利用污染物控制限制值的知识本体如下：

【油田含油污泥处置后泥渣利用污染物As（mg/kg），控制限制值，≤30（mg/kg）】

知识主体：油田含油污泥处置后泥渣利用污染物As

知识属性：控制限制值

知识答案：≤30（mg/kg）

用户可以提问“油田含油污泥处置后泥渣利用污染物As的控制限制值？”“含油污泥处置后利用As的限制值？”“泥渣利用污染物As限制值？”，通过问题理解，识别到用户的意图是【油田含油污泥处置后泥渣利用污染物As，控制限制值，？】，再通过答案搜索，查询知识图谱中的具体答案，“≤30（mg/kg）”。

问题理解是基于语义表达式，通过关键术语来匹配用户问题，例如：油田含油污泥处置后泥渣利用污染物As的控制限制值问题，可定义表达式如下：

【油田含油污泥处置后】【泥渣利用污染物As】【控制限制值】

【油田含油污泥处置后】【控制限制值】【泥渣利用污染物As】

至此就完成了一个简单的意图，但是泥渣利用污染物有着通用元素的意图，为了扩大表达式的覆盖范围，可以将此抽象为一个实体，如下所示：

【$对象实体】【$指标实体】

其中，【$对象实体】表示一个实体，代表泥渣利用污染物。当用户提问“泥渣利用污染物As？”，匹配结果如下：

匹配表达式：【$对象实体】【$指标实体】

匹配实体：【$对象实体：泥渣利用污染物】

匹配意图：【泥渣利用污染物的As，控制限制值，？】

当用户提问“泥渣利用污染物的As的控制限制值？”。

【泥渣利用污染物】【As】≠【泥渣利用污染物】【As控制限制值】

这时需要将“As”与“As控制限制值”定义为同义词组【As控制限制值;As】，通过同义词扩展表达式的覆盖范围。

【泥渣利用污染物】【As】=【泥渣利用污染物】【As】

（2）语义库设计模型

底层语义库由词库、对象库、知识库3部分组成。

构建词库的目的主要是为了分词、构造语义表达式以及使用词本身携带的语义信息进行语义相似度计算。词库是由多个词类组成，词类由词类名和一个或多个同义或同类词构成。在“泥渣利用污染物的As”的实例中，【$对象实体】定义为词类，其属性为实体词类，其下定义的所有词均为具体实体。【As控制限制值;As】定义为同义词组，其属性为普通词而非实体词，意味着并不需要识别该词的词类。

构建对象库（语义库）的目的主要是为了实例化对象类，从而快速创建某一领域的知识点，是对知识库中的对象类实例以及属性知识点与相互关系的严格刻画。对象库由属性名、标准问题模板和一组属性语义表达式所构成。对象库中的语义表达式使用词库中的词类，由一个或多个实体对象符或关键词组成，例如：【$对象实体】【检测方法】，其中【$对象实体】为实体对象符，【检测方法】为关键词。

构建知识库的目的主要是为了根据业务需求来组织和管理知识点。实例可以是对象类的实例化，当实例为对象类实例时，该实例下所有的知识点都是属性知识点，实例语义在实例化对象的过程中替换属性语义表达式中的“对象符”，进而生成知识点的语义表达式。

（3）基于Graph的搜索匹配模型

KGQA引擎基于Graph的DFS（深度优先搜索）实现语义表达式的快速模式匹配。

深度优先搜索属于图算法的一种，英文缩写为DFS即Depth First Search。其过程简要来说是对每一个可能的分支路径深入到不能再深入为止，而且每个节点只能访问一次。

深度优先搜索的特点：每次深度优先搜索的结果必然是图的一个连通分量。深度优先搜索可以从多点发起。如果将每个节点在深度优先搜索过程中的“结束时间”排序（具体做法是创建一个list，然后在每个节点的相邻节点都已被访问的情况下，将该节点加入list结尾，然后逆转整个链表），则我们可以得到所谓的“拓扑排序”，即topological sort.

KGQA引擎将定义的语义表达式，拆解为Graph节点并存放于Graph内存数据库中，例如：

【$污染物】【控制】【方法】
【$污染物】【控制】【流程】
【$污染物】【处置】【方法】
【$污染物】【处置】【依据】
【$污染物】【监测】
【$污染物】【利用】

转化为如图1所示Graph有向图结构。

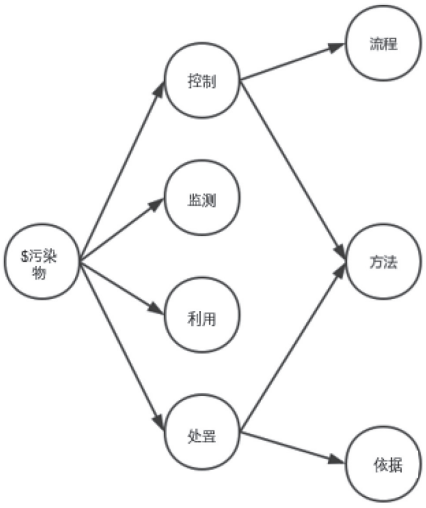

图1 Graph有向图结构

2.3 多引擎加权打分机制

油田环保安全领域标准智能问答系统采用的是一种基于柔性多引擎加权打分的问答机制，将基于模板的匹配结果（KGQA）与基于语义相似度的匹配结果（FAQ）进行基于多特征加权的结果整合，输出综合打分最高的一组结果作为候选结果。

多引擎调度采用线程池技术进行管理，处理过程中将任务添加到队列，然后在创建线程后自动启动这些任务，线程池线程都是后台线程。每个线程都使用默认的堆栈大小，以默认的优先级运行，并处于多线程单元中。如果某个线程在托管代码中空闲（如：正在等待某个事件），则线程池将插入另一个辅助线程来使所有处理器保持繁忙。如果所有线程池线程都始终保持繁忙，但队列中包含挂起的工作，则线程池将在一段时间后创建另一个辅助线程但线程的数目永远不会超过最大值。超过最大值的线程可以排队，但它们要等到其他线程完成后才启动。

3 结果与讨论

针对当前油田环保安全标准领域智能问答人机互动性较差，尚未形成人机双向理解路径，基于FAQ引擎和KGQA引擎及相关技术开展面向油田环保安全标准的双向阅读理解和智能问答的语言预训练，训练结果如图2所示。

智能问答系统虽然已经取得了一定进展，但仍存在一些问题和局限性，需要进一步改进。（1）现有系统可能在处理复杂、模糊或多层次的查询时遇到困难，尤其是涉及抽象概念或深层次语义理解的问题。（2）油田环保安全领域标准智能问答系统提供信息的准确性和可靠性有待进一步提升，特别是在处理少见的话题时。（3）油田环保安全领域

图2 智能问答结果示例

标准智能问答系统的效果很大程度上取决于其油田环保安全标准知识库的质量和时效性，需要定期更新和必要的维护。解决这些问题需要综合运用更先进的自然语言处理技术，深度学习算法、用户界面设计原则和数据保护措施。随着技术的不断进步，油田环保安全领域标准智能问答系统的性能和用户体验预期将持续提高。

参考文献

[1] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[A]. Proceedings of the 26th International Conference on Neural Information Processing Systems[C]. Lake Tahoe: ACM, 2013：3111－3119.

[2] Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation[A]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)[C]. Doha: Association for Computational Linguistics, 2014：1532－1543.

[3] Devlin J, Chang MW, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[A]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies[C]. Minneapolis: Association for Computational Linguistics, 2019：4171－4186.

[4] Zhu YK, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[A]. Proceedings of the 2015 IEEE International Conference on Computer Vision[C]. Santiago: IEEE, 2015:19–27.

[5] Den Os E, Boves L, Lamel L, et al. Overview of the ARISE project[A]. Proceedings of the 6th European Conference on Speech Communication and Technology[C]. Budapest: ISCA, 1999:1527–1530.

[6] 武鸿浩. 公安领域中知识图谱的构建与应用研究[J]. 网络安全技术与应用, 2018(8): 93–94+127.

[7] 孙利宇,钱家俊. 公安知识图谱助力智慧警务落地[J]. 数字通信世界, 2018(7):23+48.

[8] 刘峤,李杨,段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3) :582–600.

[9] 王鑫,邹磊,王朝坤,等. 知识图谱数据管理研究综述[J]. 软件学报, 2019, 30(7) :2139–2174.