**目录**

**摘要**

**1 绪论**

* 1. **研究背景及意义**

在2016年，美国Google公司率先推出的人工智能阿尔法狗战胜世界围棋冠军李世石后，人工智能引起了人们的广泛关注。随着人工智能的发展，智能问答系统技术也迅速发展起来。纵观问答系统的发展过程，近些年，问答系统取得了许多丰硕的成果。苹果公司在iPhone4s中就嵌入了其公司的人工智能产品Siri，Siri能够通过自然语言和用户进行沟通交流。Siri作为一款智能问答产品，其不仅可以通过语言和用户直接对话，还能完成用户的特定需求，例如，给某个人打电话、定个闹钟等功能。除此之外，美国微软公司也在Windows电脑中嵌入了其智能问答产品—— Cortana个人助手。同谷歌的Siri一样，Cortana一样可以满足用户这些特定的需求。近年来，国内各大互联网公司也相继推出了自己的智能问答产品，譬如阿里巴巴旗下的天猫精灵，能够和人们通过自然语言进行对话。小米公司也相继推出了小爱智能音箱，能够满足用户语音点歌，讲故事等需求。科大讯飞等公司也相继推出了智能问答个人助手相关产品。随着这些优秀产品的相继推出，人们的生产生活更加便利，人类的双手也进一步得到解放。

随着互联网的飞速发展，信息产生与传播速呈几何速度增加，如文字，声音以及视频数据。在这些海量数据中充斥着许多“垃圾”数据。因为传统的搜索引擎大多是基于关键字进行搜索的，这种信息检索方式不能很好的理解用户需求，也就不可避免的导致检索效率低下问题。因此，如何才能够提高人们获取信息的效率成为等下人们面临的主要问题。人们希望尽可能的提高信息检索的效率。与此同时，携高效性与便捷性的智能问答系统的出现有效的解决了这个问题。

基于知识图谱的智能问答系统，是一种新型的信息服务方式。不同于现有的搜素引擎，问答系统是以精确的自然语言形式返回答案，而不再是搜素引擎中返回的基于关键词匹配的相似文档排序。华盛顿大学图灵中心**主任Etzioni教授**曾明确指出：“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”[1]。因此，智能问答系统技术被认为是未来信息服务领域的颠覆性技术之一。

身体健康是人们最关注的问题，但是人们对于各种疾病的症状、用药、预防措施、治疗方法及费用等信息却知之甚少，传统信息化服务有着信息杂乱多样，获取方式低效等弊端。面对上述问题，更好的理解用户意图、更针地对用户提供疾病健康知识是关键。目前，越来越多的科研人员投入这一领域，研究如何利用人工智能技术，为人类提供更智能的服务。

* 1. **国内外研究现状**

最早知道的问答系统有BASEBALL[2]和LUNAR[3]。BASEBALL可以回答关于日期、地点、和美国棒球比赛时间等问题。LUNAR是最早的科学问答系系统。它的设计初衷是支持阿波罗任务对岩石进行地址分析。在对其评测中发现，它能正确回答人类提出的近90%的问题。BASEBALL和LUNAR的共同点是都使用了由领域专家手动编写的领域知识库。早期的问答系统限定于特定领域，可扩展性差。同时，需要领域专家对大量非监督数据进行手动收集和标注，耗费大量的时间和精力，因此很难进行较大范围的推广。

Google公司在2012年提出了知识图谱的这一概念[7]。近年来，随着知识图谱技术的不断发展，研究者开始尝试将知识图谱技术应用于问答系统。希望通过信息抽取、实体连接、实体融合等方法，将文本知识转换为计算机易于理解和表示的结构化知识，利用实体及实体间语义关系对文本数据进行更深层次的表示，从而实现对数据背后的信息进行深度挖掘与理解。之后，随着大规模知识库的出现，如YAGO[4]、Freebase[5]、DBpedia[6]等，进一步推动了基于知识图谱问答的发展。

随着知识图谱的迅速发展，智能问答技术也得到了长足的发展，传统的知识库问答技术大体上可以分为基于语义解析、基于信息抽取、基于向量建模三类。

（1）语义解析方法[8][9][10]的主要思想是把自然语言转化为一系列形式化的逻辑形式，通过自底向上地对这些逻辑形式的解析，得到一种可以表达自然语言问题的逻辑形式，进而转换为查询语句（如Cypher、Sparql等），然后利用这些语句查询知识库，从而得到问题的答案。Berant J等人通过该方法构建的知识库问答系统，获得了当时的最好成绩，其主要思路如图2.1所示。



图2.1

图2.1中，红色部分即逻辑形式，绿色部分表示来自用户的问题，蓝色部分为语义解析进行的相关操作，而形成的语义解析树的根节点则是最终的语义解析结果，可以通过查询语句直接在知识库中查询并得到最终答案。该方法效果较好，但是非常依赖相关领域专家，并且需要花费巨大的时间和精力。

（2）信息抽取方法[11]，使用自然语言处理方法提取问题中的实体，然后查询知识库获取以该实体为中心的子图，子图中的每一个节点或边都可以作为候选答案，对问题进行建模，获取问题特征，使用机器学习方法训练分类器对特征进行分类，进而得到问题的答案。



图2.2

图2.2是Yao X等人在2014年提出的通过信息抽取方法。其思想是先提取问题的主要特征词及其依赖关系（左），然后转换为特征图（右），特征图删除了不重要信息，只保留原始问题相关信息，本质上是一个信息抽取的过程。该方法相比语义解析方法，减少了对人工定义规则的依赖，但是能否对问题构建良好的特征决定了最终问答系统的好坏。

（3）向量建模方法[12][13][14]主体思想是把用户问题和答案全部向量化，通过训练数据进行训练，使得问题与答案的向量尽可能接近（通常是以向量点乘的形式）。使用模型对问题和候选答案的得分进行筛选，分数最高的答案为最终答案。



图2.3

图2.3展示了Bordes A等人在2014年提出的向量建模法用于问答系统的模型框架，该方法不但免去了人工构建规则和特征，并且从应用角度来说更加易于实现。

随着深度学习在自然语言处理领域的广泛应用，近年来，开始涌现了许多利用深度学习技术进行问答的方法。Dong L等人[15]在2015年提出用卷积神经网络对上述向量建模方法进行提升，Yih S W等人[16]在2015年提出用卷积神经网络对语义解析方法进行提升，该方法获得了当时的最好成绩。随后又出现了使用LSTM结合Attention机制进行问答的方法[17]。深度学习方法所拥有的端到端的优势有着很好的前景，伴随着AI技术的发展，知识库问答技术也将得到更好、更广的应用。。

* 1. **本文研究内容和研究目标**

本文研究的主要目标是针对限定领域内的疾病知识领域，从医药健康网站上获取相关数据，结合自然语言处理技术构建以疾病为中心的领域知识图谱；之后，使用该图谱，设计并实现一个功能可用，界面美观的疾病知识智能问答系统，从而实现用户输入自然语言问句，系统返回问答结果。

1. **相关理论与关键技术概述**

**2.1 爬虫技术**

**2.2 知识图谱技术**

* 1. **智能问答技术**

**3疾病知识图谱的构建**

**3.1 数据获取**

**3.2 知识图谱构建**

* 1. **本章小结**

**4智能问答系统算法设计与实现**

**4.1 系统环境配置**

**4.2 系统架构设计**

**4.3 系统分析**

**4.4 系统实现**

**4.5 系统演示与分析**

* 1. **本章小结**

**5总结与展望**

**5.1 总结**

**5.2 展望**

**致谢**

**参考文献**

[1] Etzioni O. Search needs a shake-up[J]. Nature, 2011, 476(7358): 25-26.

[2] Green Jr B F, Wolf A K, Chomsky C, et al. Baseball: an automatic question-answerer[C]//Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. 1961: 219-224.

[3]Woods, William A. "Progress in natural language understanding: an application to lunar geology." *Proceedings of the June 4-8, 1973, national computer conference and exposition*. 1973.

[4]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.

[5] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008: 1247-1250.

[6] Lehmann J, Isele R, Jakob M, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.

[7]刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述[J].计算机研究与发展,2016,53(03):582-600.

[8] 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.

[9] Cai, Qingqing, and Alexander Yates. "Large-scale semantic parsing via schema matching and lexicon extension." *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013.

[10] 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.

[11] 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.

[12] Bordes A, Chopra S, Weston J. Question answering with subgraph embeddings[J]. arXiv preprint arXiv:1406.3676, 2014.

[13] 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.

[14] 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.

[15] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 260-269.

[16] Yih S W, Chang M W, He X, et al. Semantic parsing via staged query graph generation: Question answering with knowledge base[J]. 2015.

[17] 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.