# 基于知识图谱在故障诊断、预测、分析的文献综述

**摘要** 随着科学技术的发展，手机及软件越来越普及，手机软件应用层软件成为测评手机质量的一个重要手段。随之而来的是各种软件故障问题。然而随着互联网、人工智能和大数据的发展，以及人们对智能的需求的不断提升，深度挖掘知识的内容，将大量的非结构化的文本数据以有序及有关系的形式显示出来，知识图谱技术应运而生。知识图谱可以将软件故障问题的关系，以实体-关系-实体的三元组的形式展现出来，简单明了。本文首先介绍了基于知识图谱在软件故障诊断、分析、预测的应用概况，其次概括了基于知识图谱案例的研究现状与进展，指出了在构造知识图谱中所遇到的问题，即文本实体及实体关系的抽取，归纳了实体及实体关系抽取的方法与工具的研究现状和进展，在此基础上对文本实体及实体关系的研究前景进行了展望。

**关键字** 软件故障 知识图谱 实体关系抽取 研究现状和进展

## 0 引言

近些年来，随着科技的发展，用户需求的提升，手机产品性能不断提高，手机软件层应用也越来越多，越来越智能。但是随着移动上网，数据业务的增多，各种新功能的普及，给用户带来方便的同时，带来的故障问题也越来越多。除了手机的硬件问题以外，软件故障也有很多问题[[1]](#endnote-1)。比如用户的误操作、进水等等都会造成手机的不良状况，如死机、黑屏、没有声音、发烫等等一系列的问题。基于知识图谱的故障诊断、分析及预测主要是针对爬取的网站的有关软件故障现象的一些非结构化的数据，通过自然语言处理，即文本及文本的实体及关系抽取，找出软件故障的各种现象及问题之间的关系，将抽取出来的三元组数据存在图数据库里，进而生成知识图谱，从而清晰明了的看软件故障问题之间的内在联系，进而能很好地分析其故障现象，直到软件故障的问题在哪，从而达到预测的效果，减少软件故障问题的出现。但是中文知识图谱构建的开放链接相对缺乏，导致目前国内尚没有成熟的故障诊断、预测、分析等知识图谱的问题。知识图谱的建立涉及信息获取、信息标示、信息存储及信息可视化等环节，专业性强、数据量大、过程复杂。

## 1 基于知识图谱在故障诊断、预测、分析的应用

知识图谱有两种存储方式：一种是基于RDF的存储，一种是基于图数据库的存储。RDF是对象-属性-值的存储，而基于图数据库的存储是实体-关系-实体。要构造一个在故障诊断、预测、分析的知识图谱，首先要收集故障数据，将数据预处理，知识图谱的设计，把数据存入知识图谱，上层应用的开发及系统的评估。即先把语料分词，然后命名实体识别，抽取关系，然后知识推理，形成实体1-关系-实体2这样的三元组，存储检索，最后形成知识图谱。但是这里我们要分析的是故障现象和故障现象之间的内在联系，这样就可诊断故障，得知原因，实现预测及预防。那么这里要存储的就是故障现象--关系—故障现象，所以我们要将大量的非结构化的文本数据将故障现象提取出来，再抽取故障现象之间的关系，进而存在数据库之间，形成知识图谱。难点就在于如何将故障现象从每个句子中提取出来，并提取出其关系，可以是从一个句子或者任意两个句子中提取。这里把实体替换成了简单的句子，即故障现象。那么我们可以把故障现象看成是一个三元组，即实体--关系---实体，或者是对象—属性—值，那么两个故障现象就是两组三元组之间的关系，即从一个句子或多个句子中提取两个三元组及三元组的关系，然后存入图数据库，生成知识图谱。

## 2 基于知识图谱案例的研究现状与进展

现在越来越多领域都在构建知识图谱，因为处于这个大数据的时代，知识图谱能将多维性、非结构性、参差不齐的数据整合起来，能更清晰理解真实世界中的实体和实体之间的关系，而且能更高效地管理数据信息[[2]](#endnote-2)。现如今，知识图谱已经涉足多个领域，如医学、金融、商务研究、化学品、地理教育、铟产业、农业、电信业客服、媒体发布等等。而知识图谱最终的目的就是让知识被用户访问到，可被查询，可被支持行动。我国目前在HER方面知识图谱的研究，利用大量的HER数据构造好了知识图谱，那么当新的患者来临时，知识图谱就可以查询扩展、临床决策支持和预防疾病的作用[[3]](#endnote-3)。而知识图谱的核心技术主要有知识抽取、知识表示、知识融合和知识推理四大方面。又如在能源领域的知识图谱，根据改进的TextRank迭代算法得到实体的权重，利用共现得到实体的关系，最后利用Pajek软件绘制出知识图谱[[4]](#endnote-4)，该文的数据源为能源学科的文献，通过人工和自动结合的方式，改善分词效果，再实体抽取。知识图谱在电网全业务统一数据中心也有应用，它首先通过人为定义和语义描述算法对全业务统一数据中心的电网数据进行知识抽取，将实体之间的冲突消除，再使用ProjE 知识融合算法计算知识实体间详细的关系，构建三元组保存至知识图谱库中，从而实现数据搜索的智能化，其中对于大规模的数据的快速计算，使用了Spark计算引擎和Mapreduce分布式计算模型[[5]](#endnote-5)。

## 3 实体关系抽取方法的研究现状与进展

实体关系抽取的研究最早是以美国MUC最早引入的，后来组织了自动内容抽取，即ACE[[6]](#endnote-6)。

MUC、ACE测评会议的实体关系抽取涉及的关系类型仅仅局限于命名实体之间的实体的关系，命名实体包括（人名、地名、组织、机构名等等）。近年来，将深度学习与实体关系抽取结合起来，出现了许多新方法，而且现在的信息抽取技术，主要有实体抽取、关系抽取和事件抽取。如基于CNN 和双向 LSTM 融合的实体关系抽取[[7]](#endnote-7)，是用来关系抽取，CNN由于卷积核的多样性能充分的提取特征，而RNN能考虑到远距离词之间的关系，改文巧好将二者结合，使模型有效性提高了不少，但是其间也需要加入少量人工特征，即词性标注和实体类别等。在海量法律文书实体关系抽取技术中用的也是CNN，但是不同的是在CNN的基础上并加入了改进核函数的多实体关系抽取技术，实现对多对实体关系进行挖掘的目标[[8]](#endnote-8)。我国早期对于实体和关系的抽取实现了联合抽取，即基于CRF的联合抽取[[9]](#endnote-9)。因为如果分开抽取，先抽取实体，再抽取关系，但是当命名实体识别错误时，就会导致关系的抽取错误，即关系的抽取受限于实体的识别，联合抽取可以提高命名实体和关系识别的召回率。而后来又提出基于双向GRU神经网络和双层注意力机制的中文文本中人物关系抽取研究[[10]](#endnote-10)，引入了注意力机制，将不同的预料赋予不同的权重，摒弃了噪声，提高了分类的精确度，而GRU则是在RNN的基础上改进的，也提升了抽取效果。同时还有中文开放式多元实体关系抽取[[11]](#endnote-11)，其关键技术主要是对文本进行预处理和依存关系分析，然后处理实体词，再使用经过训练的逻辑回归分类器进行过滤，提高实体关系多元组的准确性。后来提出了一种改进的实体关系抽取方法-- OptMultiR [[12]](#endnote-12)，其贡献在于利用远程监督和已有的知识库将实体和实体关系进行对齐，直接利用源文本进行训练省去了大量的人工标注，而且该文将原有的贪心算法替换成基于状态压缩的动态规划算法，使关系性能抽取得到了很好的提升。现在目前成熟的实体关系抽取的研究方法需要确定领域，要大量的标注训练预料，指明关系类型，选取特征，训练分类器等这些工作均需人工处理，所以提出了一种基于远程监督和强化学习的关系抽取，远程监督可用于提取不同知识库之间的冗余信息，可以自动进行预料标注，针对其中的噪声，使用词汇语义相似度算法，提高了提取精度[[13]](#endnote-13)。

在基于多核学习的生物实体关系提取方法中[[14]](#endnote-14)。其中，作者将基于特征的内核与扩展路径图内核结合到一个多核学习方法中，在五个预料评估上也是基于特征的核方法，并且扩展了路径图核方法和多核学习方法，实验结果表明，改进方法很好。

## 事件抽取

信息抽取分为三种，实体抽取、关系抽取和事件抽取，而一般情况下实体和关系抽取二者是结合在一起的，事件抽取会比较困那，因为一般都需要预先指定事件类型,然后或基于机器学习方法,或基于模式匹配方法识别事件类型和事件元素。

事件抽取的方法可以分为句子级的事件抽取和篇章级的事件抽取，我们这里需采用的时句子级的，但是最好先是句子级后篇章级，因为可能会有上下的联系。篇章级的事件抽取方法则包含了跨句、跨文档抽取的特征信息 句子级特征是所有事件抽取方法通用的特征，而篇章级特征则属于面向实际任务挖掘的特效特征。

在本文里我们最后要得到的知识图谱的节点是故障现象及故障原因是事件，事件抽取不仅需要底层的语言学方面的知识，例如词汇、命名实体、句法知识。

事件抽取主要包括两个步骤，第一就是事件类型识别，即识别触发词，触发词是能够触动事件发生的词，另外一个就是事件元素识别，即事件的参与者。

根据学习方式不同，可以分为基于流水线模型的事件抽取方法和基于联合模型的事件抽取方法[[15]](#endnote-15)。

其中神经网络在事件抽取上也可以应用。基于神经网络的事件抽取，而作为一种有监督的多元分类任务，事件抽取包括两个步骤：特征选择和分类模型[[16]](#endnote-16)

## 4 基于文本的实体关系抽取工具的研究现状与进展

Gensim是一款开源的第三方Python工具包，用于从原始的非结构化的文本中，无监督地学习到文本隐层的主题向量表达。它支持包括TF-IDF，LSA，LDA，和word2vec在内的多种主题模型算法，支持流式训练，并提供了诸如相似度计算，信息检索等一些常用任务的API接口。

Stanford CoreNLP是基于Java实现关系抽取的，其方法如有监督、远程监督、神经网络等方法，能实现对自然语言文本的分析，包括词性还原、词性标注、命名实体标注、共指分解、句法分析以及依存分析等功能。

NLTK是Python处理自然语言的第三方模块，但处理中文文本存在一定的局限性，如缺少中文语料库，没有中文停用词，无法完成中文文本的停词过滤，而且无法完成中文文本的分割。所以总体上来说，NLTK是不太适合处理中文文本的[[17]](#endnote-17)。

FudanNLP也是一个基于Java语言的中文开源工具包，同时包含机器学习算法和数据集。可以实现信息检索：文本分类和新闻聚类；中文处理包含中文分词、词性标注、实体名识别、关键词抽取、依存句法分析、时间短语识别；另外一个功能就是结构化学习，其包含在线学习、层次分类、聚类、精确推理等等。

斯坦福大学开发了知识抽取工具（三元组抽取），即支持中文的开源的deepdive，通过弱监督学习，从非结构化的文本中抽取结构化的关系数据 。本项目修改了自然语言处理的model包，使它支持中文，并提供中文tutorial。后续将持续更新一些针对中文的优化。

OLLIE是是华盛顿大学研发的知识库KnowItAll的三元组抽取组件，支持基于语法依赖树的关系抽取，对于长线依赖效果更好，能远距离的抽取关系。Reverb也是华盛顿大学研发的开放三元抽取工具，可从英文的句子中抽取实体关系的三元组，优点是不需要提前指定关系，支持全网规模的信息抽取。

SOFIE是一个自动化本体扩展系统，由max planck institute开发。它可以解析自然语言文件，从文本中抽取基于本体的事件，将它们链接到本体上，并基于逻辑推理进行消歧。

OpenIE是开放域知识的抽取工具

中文如果不容易实现三元组的抽取，可以将中文转换成英文，再进行词性标注，去停用词实体标注，抽实体和实体之间的关系等等。

IKAnalyzer是基于Java语言的中文开源工具，支持细粒度和智能分词两种切分模式，即支持英文字母、数字也支持中文词汇等分词处理。

ICTCLAS（NLPIR）是基于多种语言的中文开源工具，如Java、C++、C、C#。有多种功能，如中文分词、命名实体识别、词性标注、自定义用户词典，并且新增微博分词、新词发现与关键词提取；可视化界面操作和API调用给用户提供了大大的方便。

OpenCCG，指开放式自然语言CCG文库，基于Java的开源自然语言处理文库，能够实现基于Mark Steedman的组合的范畴语法形式为主的文本解析，包括句法分析和依存分析。

CRF，条件随机场用于中文分词、词性标注及命名实体识别，是一种基于遵循马尔可夫性的概率图模型的数学算法。

综上所述，大多数用于中文自然语言处理的开源工具包均是基于Java语言实现的，而处理英文的自语言工具包多半是基于Python语言写的。

## 5 结束语

知识图谱是人工智能的重要分支，它模拟了人的思维方式，对数据进行高效的知识管理、知识获取。本文旨在通过爬虫的技术将故障的一些数据爬取出来之后，通过自然语言处理生成知识图谱，这样能更清楚地了解故障与故障之间的联系，从而实现分析及预测，预防故障的发生。自然语言处理主要是用到了实体关系抽取技术，而在前文的介绍中有许多算法和工具，算法总的来说主要有无监督方法、有监督方法、远程监督方法和Bootstrapping方法等等。实体及关系抽取有先实体抽取再关系抽取和实体和关系的联合抽取，现如今分开抽取的算法较多，但是分开抽取关系的抽取受实体抽取的限制存在缺陷，而联合抽取较好一些，但是没有特别成熟的算法。实体关系抽取具体的算法主要有基于深度卷积神经网络的实体关系抽取、循环卷积神经维纳格罗算法、基于CRF的命名实体、半监督学习的方法结合词语规则和SVM的模型及开放中文实体关系抽取。而实体关系抽取还面临的一大难题就是需要进行大量的人工标注预料。然而也可以用很多中文实体抽取关系的工具如NTLK、Sanford CoreNLP以及华盛顿大学的OLLIE、SOFIE等等。但是NTLK并不太适合中文文本的处理，缺少中文语料和停用词，大部分开发的工具都不是很适合处理中文文本，处理英文效果很好。

知识图谱流程：知识获取，知识融汇，知识存储，查询式语意理解，知识检索和可视化

测评标准是通过准确率、回召值和F值来确定的。

## 参考文献

1. 赵敏.手机应用层软件测试及故障分析[J]. 质量技术监督研究, 2009(3):57-60. [↑](#endnote-ref-1)
2. 刘宝,车礼东,黄红花,郭兵,宋振乾,李红霞,范晓明,董瑞.基于自然语言处理(NLP)技术建立化学品危险评估知识图谱的研究[J].计算机与应用化学,2018,35(07):605-610. [↑](#endnote-ref-2)
3. 何霆,吴雅婷,王华珍,熊英杰,孙偲,徐汉川.基于EHR的医疗知识图谱研究与应用综述[J].哈尔滨工业大学学报,2018,50(11):137-144. [↑](#endnote-ref-3)
4. 赵京胜,肖娜,高翔.基于自然语言处理的能源领域知识图谱[J].信息技术与信息化,2018(05):55-58. [↑](#endnote-ref-4)
5. 王渊,彭晨辉,王志强,范强,姚一杨,华召云.知识图谱在电网全业务统一数据中心的应用[J/OL].计算机工程与应用:1-7[2019-01-27].http://kns.cnki.net/kcms/detail/11.2127.TP.20190122.1702.015.html. [↑](#endnote-ref-5)
6. 刘绍毓,李弼程,郭志刚,王波,陈刚.实体关系抽取研究综述[J].信息工程大学学报,2016,17(05):541-547. [↑](#endnote-ref-6)
7. 张晓斌,陈福才,黄瑞阳.基于CNN和双向LSTM融合的实体关系抽取[J].网络与信息安全学报,2018,4(09):44-51. [↑](#endnote-ref-7)
8. 高丹,彭敦陆,刘丛.海量法律文书中基于CNN的实体关系抽取技术[J].小型微型计算机系统,2018,39(05):1021-1026. [↑](#endnote-ref-8)
9. 徐秋妍. 基于CRF的命名实体和关系的联合抽取[D].上海交通大学,2012. [↑](#endnote-ref-9)
10. 张兰霞,胡文心.基于双向GRU神经网络和双层注意力机制的中文文本中人物关系抽取研究[J].计算机应用与软件,2018,35(11):130-135+189. [↑](#endnote-ref-10)
11. 李颖,郝晓燕,王勇.中文开放式多元实体关系抽取[J].计算机科学,2017,44(S1):80-83. [↑](#endnote-ref-11)
12. 延浩然,靳小龙,贾岩涛,程学旗.一种改进的实体关系抽取算法——OptMultiR[J].中文信息学报,2018,32(09):66-74. [↑](#endnote-ref-12)
13. 张硕望. 一种基于远程监督的中文实体关系抽取方法[D].南华大学,2018. [↑](#endnote-ref-13)
14. Xu D, Pan J, Wang B, et al. Biological entity relationship extraction method based on multiple kernel learning[C]// IEEE International Conference on Bioinformatics & Biomedicine. 2017. [↑](#endnote-ref-14)
15. 秦彦霞,张民,郑德权.神经网络事件抽取技术综述[J].智能计算机与应用,2018,8(03):1-5+10. [↑](#endnote-ref-15)
16. 秦彦霞,张民,郑德权.神经网络事件抽取技术综述[J].智能计算机与应用,2018,8(03):1-5+10. [↑](#endnote-ref-16)
17. 李晨,刘卫国.基于NLTK的中文文本内容抽取方法[J].计算机系统应用,2019,28(01):275-278. [↑](#endnote-ref-17)