**基于图数据库的电影知识图谱应用研究**

**Research on the Application of Movie Knowledge Graph Based on Graph Database**

**陆晓华1，张宇2，钱进3 （图形图像）**

**一、科学问题**

**1.1 本文所涉及科学问题**

知识图谱的构建主要包括知识单元的构建，知识单元间关系的构建和知识的可视化三个部分组成。由于知识图谱的主要应用局限于搜索引擎和问答系统等方面，本论文对IMBD电影网站数据进行抓取并抽取命实体和实体关系，通过Neo4j图数据库建立一个电影知识图谱。

**1.2 同行专家如何解决**

Google在2012年就将搜索结果进行知识系统化，不同于基于关键词搜索的传统搜索引擎，知识图谱可以更好的查询复杂的关联信息。以往的知识图谱研究多基于文献来进行研究，将关键词、摘要等结构化信息作为直接的知识单元进行构建。

**1.3 本文所解决的问题**

（1）通常现实生活中的实体和关系普遍都比较复杂，而且常常查询涉及到1度以上的关联查询，如果使用关系型数据库存储知识图谱会形成性能瓶颈。对于复杂的关系网络，基于图数据库存储的优势非常明显，在关联查询的效率上会比传统的存储方式具有较大的提高，当涉及到2~3度的关联查询时，这种存储方式会比传统的存储方式效率高出几千倍甚至是几万倍。

（2）可以提高代码的灵活性，降低维护成本。如果需要对数据进行更改，只需要在图谱上插入即可解决。与此相反，关系型存储方式灵活性较差，它所有的Schema都是提前定义好的，因此要修改起来就非常麻烦。

**1.4 本文解决方案效果**

本文将知识图谱的常用算法运用到电影数据关系中，基于Neo4j图数据构建的电影知识图谱系统拥有更加全面的电影关系网，用户查询起来非常方便。

**二、研究内容**

**2.1 理论与方法介绍**

本文使用的数据来自于IMBD网站，利用爬虫技术从互联网空间抓取的文本包含HTML标签等杂讯，需要对抓取的数据进行清洗，数据准备完成之后，通过机器学习算法提取文本中的命名实体，继而通过特殊的正则模式匹配找出实体之间的关系，并将其持久化为csv文件，最后将所有的命名实体及实体关系导入Neo4j图数据库中，以供数据可视化以及知识图谱内部联系的查询。

在本文中，命名实体提取使用的是NLTK库中的最大熵算法，NLTK的命名实体识别使用的是MaxEnt分类器，此分类器工作时有两个原则：①总是试图保持均匀分布（既最大化熵）；②保持其统计概率与经验数据一致。

在本文中，实体关系抽取采用的是基于模式匹配的关系抽取，所谓基于模式匹配的关系抽取通常是运用语言学知识，在执行抽取任务之前，构造出若干基于语词、基于词性或基于语义的的模式集合并存储起来。当进行关系抽取时蒋经国预处理的语句片段与模式集合中的模式进行匹配。一旦匹配成功，就可以认为该语句片段具有对应模式的关系属性。寻找制定类型的命名实体之间的关系，首先建立所有（X, α，Y）形式的三元组，其中X,Y是制定类型的命名实体，α表示X和Y之间的字符串。提供了特殊的正则匹配方式，可以方便对词性、命名实体类别等进行模式匹配，提取我们兴趣的远组。

**2.2 验证分析与实验效果**

本文的系统主要由Python实现，主要分为三个模块：数据采集模块、数据提取模块以及Neo4j图数据库导入模块。从构建的电影知识图谱中，可以非常容易的分析电影节点以及电影人节点，以及他们之间的关系，推理出演员之间是否相识或者间接相识，从而推断出是否有合作的可能。通过Cypher查询语言，可以很容易的推演出两个演员节点之间的最短路径。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

本文虽然提出了基于图数据库的电影知识图谱的研究发现，然而系统中还存在着许多不足的地方可以提高。首先文章知识提供了有限类型的边节点和点节点。此外，文章仅仅提出了Spyher的查询功能，并没有提出相关的图算法。

**3.2 后续研究重点**

文章后续将会结合Spark Graphx，运用图算法进行大规模的知识图谱分析，进而可以方便的实现社区发现、用户影响力、人群划分等功能。

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：基于深度神经网络的法语命名实体识别模型

**（2）作者介绍**：严红，陈兴蜀，王文贤，王海舟，殷明勇

**（3）摘要**: 现有法语命名实体识别研究中，机器学习模型多使用词的字符形态特征，多语言通用命名实体模型使用字词嵌入代表的语义特征，都没有综合考虑语义、字符形态和语法特征。针对上述不足，本文设计了一种基于深度神经网络的法语命名实体识别模型 CGC-fr。首先从文本中提取单词的词嵌入、字符嵌入和语法特征向量，再由卷积神经网络(Convolutional Neural Networks, CNN)从单词的字符嵌入序列中提取单词的字符特征，然后通过双向门控循环神经网络(Bi-directional Gated Recurrent Unit Neural Network, BiGRU)和条件随机场分类器(Conditional Random Field, CRF)根据词嵌入、字符特征和语法特向量识别出法语文本中的命名实体。实验中，CGC-fr 在测试集的 F1 值能够达到82.16%，相比于机器学习模型 NERC-fr、多语言通用的神经网络模型 LSTM-CRF和 Char attention 模型，分别提升了 5.67%、1.79%和 1.06%。实验结果表明，融合三种特征的 CGC-fr 模型更具有优势。

**4.2 相关论文二**

**（1）题目**：深度学习实体关系抽取研究综述

**（2）作者介绍**：颚海红，张文静，肖思琪，程瑞，胡莺夕

**（3）摘要**：实体关系抽取作为信息抽取、自然语言理解、信息检索等领域的核心任务和重要环节,能够从文本中抽取实体对间的语义关系.近年来,深度学习在联合学习、远程监督等方面上的应用,使关系抽取任务取得了较为丰富的研究成果.目前,基于深度学习的实体关系抽取技术,在特征提取的深度和模型的精确度上已经逐渐超过了传统基于特征和核函数的方法.本文围绕有监督和远程监督两个领域,系统总结了近几年来中外学者基于深度学习的实体关系抽取研究进展,并对未来可能的研究方向进行了探讨和展望。

**4.3 相关论文三**

**（1）题目**：基于神经网络的片段级中文命名实体识别

**（2）作者介绍**：王蕾，谢云，周俊生，顾颜慧，曲维光

**（3）摘要**：命名实体识别是自然语言处理的一个重要基础任务。传统基于统计学习模型的命名实体识别方法严重依赖特征工程，特征设计需要大量人工参与和专家知识，而且已有的方法通常大多将中文命名实体识别任务看作一个字符序列标注问题，需要依赖局部字符标记区分实体边界。为了减弱系统对人工特征设计的依赖，避免字符序列化标注方法的不足，该文对基于神经网络的片段级中文命名实体识别方法进行探索研究。通过采用深度学习片段神经网络结构，实现特征的自动学习，并通过获取片段信息对片段整体分配标记，同时完成实体边界识别和分类。基于神经网络的片段级中文命名实体识别方法在MSRA数据集上对人名、地名和机构名识别的总体F1值达到90.44%。

**基于深度神经网络的中文命名实体识别**

**Chinese Named Entity Recognition Based on Deep Neural Network**

**张海楠，伍大勇，刘悦，程学旗（中文信息学报）**

**一、科学问题**

**1.1 本文所涉及科学问题**

本文涉及了深度神经网络的结构和训练方法，详细的介绍了字向量、词向量和字词结合向量三中输入特征的表示。对比不同窗口大小和隐藏节点个数情况下，在字向量、词向量、字词结合向量作为输入特征时，DNN在中文NER上的实验结果。

**1.2 同行专家如何解决**

近年来，研究者们把NER任务规约为一种序列标注任务，对于每一个输入的字，判断其标签类别，根据最终的类别标签判定命名实体的边界和类型。陈钰枫提出使用双语对齐信息来提高NER性能和双语对齐结果，可以有效提高中文NER的自适应性。邱泉清提出使用CRF模型对微博数据进行命名实体识别，利用知识库和合适的特征模板，取得了良好的效果。

**1.3 本文所解决的问题**

由于中文词语缺乏明确的边界和大小写特征，本文利用用词向量的特点，提出了一种用于深度学习框架的字词联合方法，将字特征和词特征统一地结合起来。在词特征中加入词性信息，进一步提高系统的性能。

**1.4 本文解决方案效果**

本文介绍了用于中文命名实体识别的深度神经网络，提出了字词结合的方法，有效弥补了单字识别和单词识别的不足，加入词性特征后的识别系统性能更加完好。实验对比了字向量、词向量和字词结合向量在中文NER上的结果，字词结合方法对中文NER上的结果有了较大的提升。

**二、研究内容**

**2.1 理论与方法介绍**

本文选择深度神经网络结构进行中文NER任务，第一层是输入层，它主要负责将输入窗口的字或者词进行词向量的映射，所有出现在字典中的字或者词都有一个固定长度的低维向量。当输入窗口产生新的字词后，输入层将对这些字词进行向量映射，将其按顺序进行组合，获得该窗口下的窗口向量，窗口向量作为NER第二层的输入。第二层是一个标准的神经网络，它具有两个线性层和一个位于中间的非线性层。第三层是采用Viterbi算法实现的输出层，它主要负责对输入的句子进行最优标签序列的搜索。

**2.2 验证分析与实验效果**

本文共进行了三组实验进行人名、地名、机构名的识别，第一组实验的目的是对DNN参数进行选择，第二组实验对比字向量、词向量和字词结合向量的NER性能。第三组实验是比较加入词性特征后的系统识别性能。

实验效果：，当窗口大小在３～５时，系统的识别效果最好。字词结合系统对于大部分识别任务是窗口大小不敏感的，在小窗口下也可以达到很好的效果。对于人名的识别任务，由于词典稀疏问题，词向量方法中人名会被映射成一个特殊的向量，降低了识别效果而人名更加关注于姓氏和语句结构，因此字向量的效果更加好。

**三、论文存在问题及后续研究重点**

**3.1 论文存在问题**

本文并未将和命名实体相关的专用特征加入到系统中去，因此实验产生的F1值并不能达到一个满意的值。

**3.2 后续研究重点**

本文下一步将考虑加入姓氏集、地区特征集等相关特征，进一步的观察该方法的性能.

**四、该问题相关研究成果**

**4.1 相关论文一**

**（1）题目**：中文微博命名实体识别

**（2）作者介绍**：邱泉清，苗夺谦，张志飞

**（3）摘要**:微博这一媒体形式的迅速发展为命名实体识别提供了一个新的载体。根据微博文本的特点,提出针对中文微博的命名实体识别方法。首先,对微博文本做规范化处理,消除由于微博表达不规范造成的干扰;在建立中文人名库、常用地点库等知识库的基础上,选取适合微博的特征模板,使用条件随机场方法进行实体识别;同时,将正确的识别结果添加到知识库中以提升识别效果。在真实微博数据上的实验表明,该方法能够有效地完成中文微博的命名实体识别任务。

**4.2 相关论文二**

**（1）题目**：融合全局词语边界特征的中文命名实体识别方法

**（2）作者介绍**：刘冰洋，伍大勇，刘欣然，程学旗

**（3）摘要**：目前在中文命名实体识别的任务中经常采用有监督的字序列标注模型。我们在实际应用中发现，基于字序列标注模型的中文命名实体识别模型对于词语边界的识别错误是影响识别效果的主要因素之一，边界错误平均占错误结果中的47.5%。该文通过在平均感知机模型中引入全局的词语边界特征，使得人名、地名、机构名识别的Ｆ值平均提升了0.04并降低了边界错误占错误结果的比例。

**4.3 相关论文三**

**（1）题目**：基于词向量聚类的中文微博产品命名实体识别

**（2）作者介绍**：王洪亮

**（3）摘要**：随着微博等社交平台的兴起，如何针对微博数据进行产品命名实体识别成为了自然语言处理领域研究的热点之一，也是实现舆情监督和商业智能的基础．传统的命名实体识别技术没有考虑中文微博口语化、不规范等特点，且忽略了深层语义对命名实体识别的重要作用．因此，考虑中文微博的特殊性，提出一种融合全局上下文信息的词向量特征选择方法，分别采用主题模型和神经网络词向量聚类两种方法获取深层语义信息，并结合层叠条件随机场进行中文微博的命名实体识别．实验结果表明，基于词向量聚类的中文微博产品命名实体识别方法取得了较好的效果。

**（论文读书笔记二另起一页）**

以上为论文读书笔记的基本模板，内容尽量言简意赅，周五之前上传到服务器，我整理后发给老师，大部分人每周2-3篇论文和读书笔记（每周的读书笔记放在同一Word里），大概到**9.10**（之后我会询问具体时间）总量**20**篇请尽量全部交齐。（做成论文笔记为大家以后写论文查找有帮助，这基本就是一篇论文的思路）。