知识图谱在问答系统中的应用综述

刘俪婷，师文轩，程书芝
（南开大学软件学院，天津 300350）

摘要：从基于关键词匹配、信息抽取并基于浅层语义分析的 IR-based QA 到依赖于网民贡献的问答系统，再到基于知识库的问答系统，甚至问答系统的研究转向基于自由文本方向的发展，都是由于知识库规模这个难以突破的瓶颈问题。近两年来，随着互联网文档万维网向数据万维网转型的时机，各大搜索引擎公司纷纷构建知识图谱。因此，知识图谱成为问答系统的新突破。本文介绍了问答系统的起源发展、知识图谱概论、知识图谱构建技术、知识图谱应用问答系统以及总结与展望。其中，构建技术从四个方面的信息抽取技术进行了详细的描述，包括：概念层次关系学习、命名实体识别与连接、事实知识学习和事件知识学习。研究表明，知识图谱的构建有望有效的扩大知识库的规模。

关键词：问答系统；信息抽取；知识图谱；深度学习
中图分类号：TP391.1

The application of mapping knowledge domain in question answering system

LIU Liting, SHI Wenxuan, CHENG Shuzhi
(College of Software, Nankai University, Tianjin 300350)

Abstract: From IR-based QA,Community QA to KB-based QA,even to free text based QA.All are due to the scale of the knowledge base.In recent years,with the transition opportunity of the Internet from the world wide document web to the world wide data web ,many of major search engine companies turn to build the mapping knowledge domain.Therefore, the mapping knowledge domain has become a new breakthrough in the question answering system.This paper introduces the origin and development of the question answering system, the introduction of mapping knowledge domain, the construction technology of mapping knowledge domain, The application of mapping knowledge domain in question answering system,the summary and prospect.Among them, the construction of technology from four aspects of the information extraction technology to carry out a detailed description, including: the concept of hierarchical relationship learning, Named Entity Recognition and connection, factual knowledge learning and event knowledge learning.The research shows that the construction of mapping knowledge domain is expected to expand the scale of the knowledge base effectively.

Key words: the question answering system;Information extraction;mapping knowledge domain;deep learning

基金项目：高等学校博士学科点专项科研基金（20130031120042）
作者简介：刘俪婷（1993-），女，南开大学软件工程系硕士研究生，主要研究方向：数据挖掘，机器学习等
通信联系人：师文轩（1977-），男，南开大学副教授，主要研究方向为：数据挖掘，移动计算，软件工程等. E-mail: shiwx@nankai.edu.cn

0 引言

在 2011 年 8 月，英国《nature》杂志中 Prof. Oren Etzioni 就说过：“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”。

第一个问答系统[1]诞生于 20 世纪 60 年代，一般认为是 Jaseph Weizenbaum 设计实现的“Eliza”：一款聊天机器人，通过反问的方式引导说话，用于对精神病人的心理治疗。其采用模式匹配的方式在知识库中寻找答案。

20 世纪 70 年代出现的一些专家系统标志着问答系统的新的里程碑的跨越，如：伍兹（W. Woods）设计的 LUNAR 系统，是一个专门回答有关阿波罗登月返回的月球岩石样本的地质分析问题自然语言信息检索系统。维诺格拉德（T.Winograd）建立的一个用自然语言指挥机器人动作的人机对话系统等等。由于专家系统专业范围的严格限制，一方面使得特定领域发展成果的惊人，另一方面成为通用领域发展的桎梏。

20 世纪 80 年代，Doug Lenat 发起的 CYC 研究项目，将问答系统带入“知识原则”时代，在知识库的基础上进行自然语言理解、学习以及问答的研究，于 90 年代，国内的陆汝铃、曹存根等人也开始知识库的建设研究工作。

20 世界 90 年代，由 MIT 人工智能实验室开发一个优秀的 Start 问答系统，也是包含两个数据库给出快速准确的回答，同时包含一个搜索引擎检索数据库外的知识。

2014 年以来，互联网开始了新的转型，由于 Linking Open Data 等项目的开展，其悄然从仅包含网页之间超链接的文档万维网转变为包含大量描述实体间丰富关系的数据万维网。如今的几大主流的搜索引擎公司：Google、百度和搜狗等借着互联网转型东风，开展了知识图谱的构建研究。因此，对于问答系统的发展，本文更看好知识图谱作为知识库新的突破应用到问答系统中。

科学知识图谱[2]，将复杂的科学知识领域通过数据挖掘、信息处理、知识计量和图形绘制,以可视化的方式显示科学知识的发展进程与结构关系,揭示科学知识及其活动规律,展现知识结构关系与演进规律。构建知识图谱，突破知识库发展瓶颈成为这场变革的重中之重。本文便从构建知识图谱的技术开始综述，进而展望其在问答系统中的应用。

1 知识图谱概览

知识图谱的概念于 2012 年 5 月由 Google 率先提出，目的是将搜索结果进行知识系统化，让每一个关键字都拥有一个完整的知识体系，从而提高搜索的质量。Google 的辛格博士曾说“The world is not made of strings,but is made of things”，知识图谱旨在描绘真实世界中的各种实体或概念，由此，也使得检索真正实现了基于内容的查询。

知识图谱的概念模型是构建知识图谱的基础。(1)概念 concepts(classes/types) 具有相同属性的一组对象。（2）实例 instances(entities,objects) 属于某个概念中的一个对象。(3)关系 ISA（hyponymy/hierarchy/subclassof）分类知识。(4)属性 properties(attributes) 对象拥有的特征，属于事实知识共同构成知识图谱的概念模型。有观点认为知识图谱表示的概念模型和逻辑基础可以看作本体，即哲学中的本体。通过语义网络（当前给出的定义好的网页信息的扩展）的形式表达出来，从而使得人类和电脑能更好的合作工作。综上明晰知识图谱的原理，利用数据万维网提供的知识共享和重用的知识平台，借鉴互联网上信息内容的语义描述的基础、人工智能的知识表示(例如符号主义)，实现对客观世界从字符串描述到结构化语义描述，进而对客观世界的知识映射。如图 1-1 是知识图谱的结构示意图：

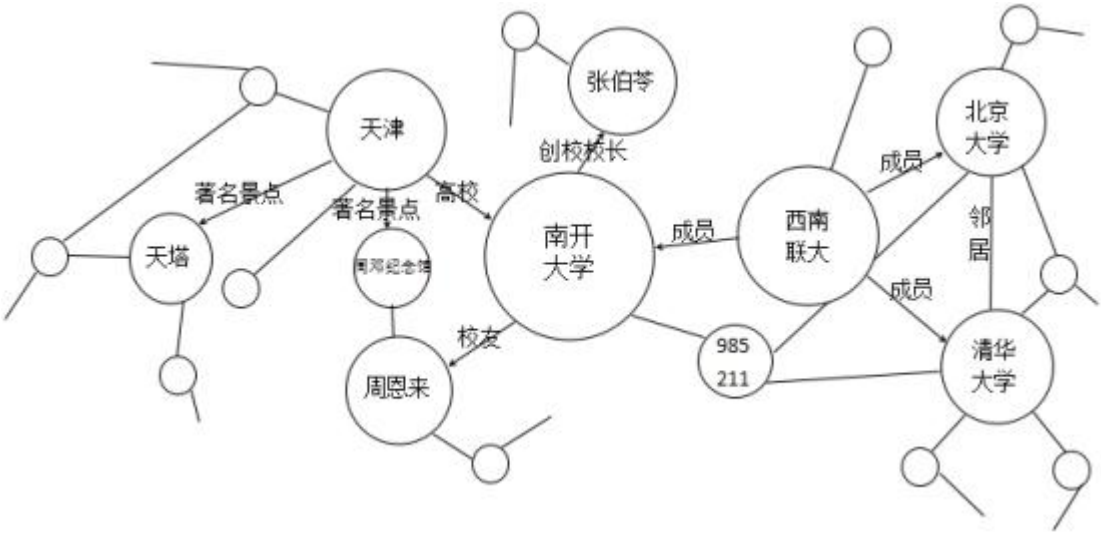

图 1-1 知识图谱结构示意图

2 知识图谱构建技术

构建知识图谱最重要的两点是数据和信息抽取。

为了满足对话搜索和复杂问答的新要求，知识图谱中不仅需要包括大量高质量常识性知识，还需要及时的补充更新。通常会选择

1》现在公认的最大的在线百科全书--维基百科（Wikipedia）：其文本为半结构化形式，在百科类站点中结构化程度最为规整，采用协同编辑实时更新与修正。

2》Freebase。类似 Wikipedia 的允许创作共享的百科类数据源，其完全采用结构化形式，规模远大于 Wikipedia。在 Google、百度、搜狗自建的知识图谱中也担任着举足轻重的地位。

3》YAGO。通用语义数据集。一个大型的语义数据集，其整理、规范和结合了来自维基百科、WordNet 和 GeoNames 的数据资源。

4》搜索日志、点评网站、社交网站、一些特定领域知识库、众包反馈等等。

知识图谱的构造与信息抽取密不可分。信息抽取就是从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术，目的在于使得信息变得更加机器可读化。简单说，是把文本中的信息进行结构化的处理，以统一的形式把信息集成起来形成巨大的数据库。

信息抽取的研究历史很长：

20 世纪 80 年代末，被几何级增长的在线与离线文本数量和从 1987-1997 年由美国国防高级研究计划委员会 DARPA 资助的 MUC（七届“消息理解研讨会”，该会议特点在于信息抽取的评测）推动着蓬勃开展起来。

到 1999-2008 年由美国国家标准与技术研究所 NIST 主办的 ACE（自动内容抽取评测会议），旨在开发支持三种不同来源（普通文本、由自动语音识别 ASR 得到的文本、由光学字符识别 OCR 得到的文本）语言文本的自动内容抽取技术。由于 ACE 评测不针对具体领域场景，对系统跨文档处理能力进行评测，采用基于漏报、误报为基础的评价体系，进而把信

息抽取技术研究引入新的高度。

以及 2009 沿用至今的 TAC-KBP 子任务，其评测会议一直推动着相关领域的研究与发展。

本文从信息抽取的四个角度分别对上述数据进行经典算法的综述。然而真正构建完备得知识图谱仍需要各个方法的综合使用、深入研究以及新方法的开辟。

2.1 概念层次关系学习

本章主要抽取思路是确定类与子类(subclassof)的关系，值得注意的这里并不是研究一个类一棵树，而是多维的概念。本章的方法以基于启发式规则（模板）和基于分布式为例，前者可以获得较高准确率的候选对概念，但是覆盖率有限，使用维基百科和大规模网页库也只是起到部分改善的作用；后者不需要预知具有可能上下位关系的词对，通过利用概念的分布特征计算上下位关系，但是准确率低。两种方式各有利弊。

2.1.1 基于启发式规则的方法

Heuristic(启发式技术=启发式扫描+启发式监控)，重点在于特征值识别技术上的更新、解决单一特征码比对的缺陷。该技术的核心在于反编译:计算机软件反向工程(Reverse engineering)，推导出他人的软件产品所使用的思路、原理、结构、算法、处理过程、运行方法甚至是源代码。破解自然中或者已经存在的，学习并产生启发。在知识图谱的构建思路为：

1>根据上下位概念的陈述模式从大规模资源中找出可能具有上下位关系的概念对。

2>层次树(一种语义描述)归纳。

例如：a.在 Wikipedia 中学习，思路[3]是：

1>把 Wikipedia 中的 category 当作概念。

2>利用 Wikipedia 中条目下所属的分类生成初始的上下位关系。

其优点：大规模概念和概念层次关系生成自动化。缺点：概念覆盖不全、存在噪声等，通常采用过滤、迭代、引用外部资源的方式进行改善。

b.在非 Wikipedia 资源网页中学习，如：YAGO 利用 Wordnet 中的良好上下位关系定义概念层次结构，联合使用 Wikipedia 概念作为叶子，进行分类挂载。

c.大规模概念结构学习，概念结构这类数据库也存在不少，常被应用的如 Freebase、Probase。其中，作为 Bing 的法宝 Probase 有自己独特的优势。其特色：

1>自动学习规模大：16 亿网页。

2>概念的层次结构规模可达上百万（200 万+）。

3>概率概念结构（Probase 中的数据，就像人脑中的思想，不是简单的非黑即白，会将不确定量化）。

4>没有自子节点的概念认为是实例。

建立一系列符合实际问题解决的启发式规则也就形成了启发式算法，40 年代就已经被提出，50 年代发展鼎盛，“贪婪算法和局部搜索”受到关注，60 年代发现不足：速度很快，但是解得质量不能保证，而且对大规模的问题收敛速度慢。在构建小型的知识图谱时仍有很多选择基于启发式规则[4]建立的实例，然而随着数据惊人的增长速度，该方法显得力不从心。不过启发式算法也并没有停止研究的步伐：模拟退火算法（Simulated Annealing Algorithm），人工神经网络（Artificial Neural Network），禁忌搜索（Tabu Search），演化算法（Evolutionary Algorithm），蚁群算法（Ant Algorithms）等等。

2.1.2 基于分布的方法

该方法通常是在一个给定的实体集合和大规模文档集合中，利用词或实体的分布相似性，自主建立实体集合的概念结构。基本思路[5][6]为：

1>实体的各种分布特征，计算其间关系。

2>定义概念层次结构的特征度量方法。

3>构造分类（Taxonomy）结构。

例如：a.基于概念结构度量的分类学习，其特点为：

1>使用大量对找出概念上下位关系有帮助的特征（如：上下文特征、术语共现、语法从属特征、重复定义等），对每个概念对定义信息函数。

2>采用增量层次聚类。增量式是针对数据变化的部分增量地更新挖掘结果，这样若能充分的利用前一次挖掘的结果，挖掘效率就会提高。层次聚类不再是一个单一聚类，而产生了一个聚类层次。层次聚类算法就是产生了一个嵌套聚类的层次，在执行步骤 t 时，将 t-1 步的聚类情况下产生一个新聚类。

3>概念结构学习转化为多标准优化问题。

b.基于隐变量的分类学习，其特点：

1>从公众分类法(folksonomy：随着互联网发展出现的一种新的分类方法，典型的代表网站是Del.icio.us，由公众为信息贴加标签 tagging)中建立实体集合的概念结构。

2>话题模型计算标签间关系。

3>四种计算分类的标签距离度量方法（Tag divergence、Hypernym-divergence、Merging-divergence、Keep-divergence）。

c.基于信念传播的分类结构学习[7]，由 Bill Freeman 与 Fredo Durand 提出的 BP 算法其特点为：通过学习局部间的联系，得到全局最优解，使其既减少了运算时间，又提高了分类质量。其优点为：

1>不需要用 pattern 找上下位概念对。

2>同时考虑上下位关系，兄弟结点关系和整个树结构。

3>非增量的分类归纳的概念模型。

基于特征分布进行分类的研究很多，该方法对于非均衡语料、特别是稀有类别的判断存在显著的优势。例如：Zheng 等提出的显示近似最优组合正/负特征[8]、徐燕等人提出的在较少类别中出现的词条具有较好的区分类别能力[9]等。但是该方法应用在构建知识图谱计算上下位关系上，准确率较低效果不佳。

2.2 命名实体识别与连接

本章从命名实体识别和实体连接的角度构建知识图谱。

2.2.1 命名实体识别

首先明确命名实体识别的任务：待处理文本中识别出七类（人名、机构名、地名、时间、日期、货币和百分比）命名实体。其中包含两个子任务：实体边界识别和确定实体类别。根据任务需求与识别特点形成识别方法：

1>命名实体的内部构成和外部语言环境具有的特征。

2>尽可能充分发现和利用实体所在上下文特征和实体内部特征。

3>相应类别实体识别模型。

4>计算特征权重的序列标注工具（工具如：MEMM、HMM、CRF）。

采用以上方法的识别结果于 2006 年在国际会议上得到了评定，英文的最好结果（准确率、召回率）为（95%、92%），该结果还算理想，已经具备了相当程度的大规模文本处理能力。而汉语的最好结果分别对于人名、地名、机构名为(66%，92%)，(89%，91%)，(89%，88%)。

第一代的实体识别，汉语的识别效果并不理想，在实际应用中，效果性能甚至还会大打折扣。回顾第一代，训练语料规模和类别数限定了识别的性能：训练语料所用的网页信息不规范，存在很多噪音，使得命名实体识别模型所依赖的上下文特征发生了明显变化，性能剧烈下降。类别数的限定，使得不能满足实际的应用。

开放域实体抽取的提出使得实体识别有了新的发展方向，实体类型更多、更细，甚至是随着时间演化的。开放域实体识别，不限定实体类别，不限定目标文本。给定某一类别的实体实例，从网页中抽取同一类别其他实体实例。从几个实例分析总结其主要方法。

a.利用查询日志(Query Log)进行开放域实体抽取：

1>在查询日志中的上下文分析种子实例学得模板，用模板找到同类别实例。

2>构造候选向量、种子上下文向量，计算相似度。

b.利用网页文档(Web Page):若处理列表型网页，种子与目标实体有相同的网页结构。以此为推动，总结网页信息的开放域实体抽取三个模块:

1>爬取模块(Fetcher)：把种子送到搜索引擎，抓取前 100 个网页作为语料。

2>抽取模块(Extractor)：针对单个网页学习模板，再使用模板抽取候选实例。

3>排序模块(Ranker)：利用种子、网页、模板、候选构造一个图，综合考虑网页(一个网页包含的高质量模板越多，该网页质量越高，反之亦然)和模板(一个模板抽取出的正确实例越多，该模板的质量越高，反之亦然)的质量，使用随机游走（Randow Walk）算法为候选打分并排序。

c.多数据源融合进行实体抽取：针对不同数据源，选取不同特征分别进行实例扩展，对结果进行融合，不同数据选取不同的模板和特征，同时计算特征候选的置信度也不同。实体抽取的数据源倾向于多样化，多源化发展。找到种子词语目标词之间的联系(相同或者类似的上下文)，从种子词生成模板，从模板找到更多同类实体。

开放域实体抽取的想法不再单纯的召回，而是进行实例扩展。针对不同数据源的特点设计方法，针对性、灵活性更强；通常包括模板抽取和计算实例候选置信度两个模块，两部分迭代进行、相互依赖。以无监督的方法为主，但缺少公认的数据集和相关评测。随着研究深入，新的问题也凸显出来：命名实体的歧义。几种基于聚类的实体消歧：

1>把所有实体指称项按其指向的目标实体进行聚类。

2>每一个实体指称项对应到一个单独的类别。

a.词袋模型：

1>利用待消歧实体周边的词来构造向量。

2>利用向量空间模型来计算两个实体指称项的相似度，进行聚类。

b.社会化网络：

1>每个人都有自己的社会关系，各自不同。

2>不同人的社会化关联信息所表现出来的网页链接特征，对网页进行聚类，实现人名聚类消歧。

c.维基百科[10]：

1>维基百科的相关实体间有连接关系。

2>连接关系反映条目间的语义相关度。

3>用实体上下文的维基条目对实体进行向量表示。

4>利用维基条目间的相关度计算指称项间的相似度(解决数据稀疏问题)。

d.多源异构知识：挖掘和集成多源异构知识，解决单一知识源覆盖度有限的问题，提高消歧性能。多源：

1>Wikipedia:捕捉概念间的语义关联。

2>WordNet:捕捉词语间的语言学关联。

3>Web 网页库：捕捉命名实体间的社会化关联。

异构[11]：语义图表示框架。同时捕捉显示语义知识(语义图的边：建模所有知识源中直接抽取出的概念之间的显示语义关联)和结构化语义知识（语义图的结构：建模概念之间的隐藏语义关联）。计算原则：若一个概念的邻居概念与另一个概念存在语义关联，则这个概念也与另一个概念存在语义关联。

对于在网页中获取 Deep web 的数据是很有挑战性的，同时又是很有价值的。其中，Deep Web 是指网页中可访问的在线数据库，也就是网页页面通过查询接口连接后台查询数据库，动态地显示于页面的内容。而对于 DeepWeb 的实体识别有大量的研究，例如文献《基于探测查询的 Deep Web 实体识别》[12]、《基于 BP 神经网络的 Deep Web 实体识别方法》[13]。

2.2.2 实体链接

实体连接[14]的任务是给定实体指称项和它所在的文本，将其链接到给定知识库中的相应实体中。其主要步骤为两步，候选实体的发现和候选实体的链接。实体连接结构示意图如图 2-1 所示：

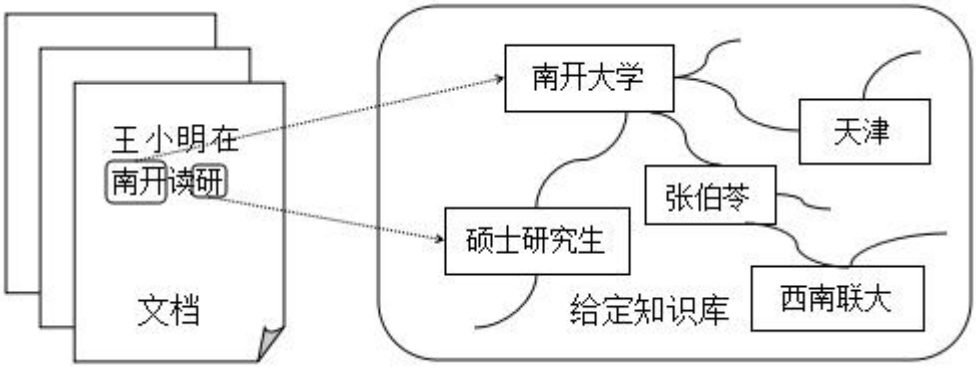

图 2-1 实体连接结构示意图

步骤一是通过给定实体指称项，连接系统根据知识、规则等信息找到实体指称项的候选实体。Wikipedia 的信息和上下文信息可作为数据源。

维基百科中：

1>维基百科中锚文本的超链接关系；

2>维基中的消歧页面[15]；

3>维基中的重定向页面来获取候选实体。

b.上下文中：主要解决缩略语的相关问题，根据

1>缩略语指称项具有很强的歧义性，而全称往往没有歧义。

2>在实体指称项文本中，缩略语的全称出现过。利用人工规则抽取实体候。

步骤二则是系统根据指称项和候选实体之间的相似度等特征，解决实体消歧问题，选择实体指称项的目标实体[16]。

其发展经历了单一实体链接向协同实体链接发展的的过程，其目的是更有效挖掘实体指称项信息，更准确计算实体指称项和实体概念之间的相似度。

单一实体链接：

1>基于词袋子模型计算相似度：用词袋子向量形式表示上下文文本（实体指称项、候选实体），通过计算向量间的夹角确定指称项与候选实体相似度，系统选择相似度最大的候选实体进行链接。

2>加入候选实体类别特征：为改善候选实体文本内容太短小导致的相似度计算不准确问题。其实现方法：训练 SVM 分类器对候选实体进行选择；训练数据由维基中的超链接获得；采用文本相似度、指称项文本中词与候选实体类别的共现信息等特征。

3>加入候选实体的流行度等特征：由于传统方法仅计算实体指称项与候选实体的相似度，忽略候选实体的背景知识和先验信息：如：实体 e 在知识库中的概率 P(e)，指称项 s 指向实体 e 的概率 P(s|e)，实体 e 出现在特定上下文环境 c 的概率 P(c|e)。

协同实体链接既要计算实体指称项与目标实体的语义相似度，又要计算目标实体间的语义相似度，采用 Pairwise（排序学习算法）优化策略。以下介绍几种协同链接的具体方法：

基于图的协同链接：由于 Pairwise 策略只考虑两两实体关系，结果不是全局最优的。采用图方法，全局考虑目标实体间的语义关联，构成两种关系指称项图(Referent Graph)：

1>指称项与实体间关系：指称项文本与实体文本的相似度，由传统的 VSM 模型(向量空间模型)得到。

2>实体间的语义关系：利用目标实体间的连接关系计算实体间的语义相关度。

例如：Han 等构建包含指称-实体、实体-实体关系的图，提出有效的协同推理算法[17]。Hoffart 等在图的基础上，利用边加权计算稠密子图，得到映射效果[18]。

b.基于深度学习的方法：传统方法中，由于没有考虑各个概念间的内在联系，计算待消歧实体上下文和目标实体语义相似度的方法可扩展性差。协同协作的方法中，提出利用深度学习的方法自动联合学习实体和文档，进而完成实体链接任务。

c.跨语言实体链接：由于传统方法要先翻译成目标语言，过程产生错误传递风险，需要大量句子级平行的双语训练预料。双语隐含主题模型将实体指称项与候选实体映射到同一个主题空间：给定一种语言的实体指称项和其所在的上下文，将其连接到另外一种语言的知识库中。例如：Wang 等提出的链接因子图模型挖掘百度百科与英文维基百科实体关联，效果良好[19]；

d.结构化数据中的实体链接，结构化数据没有上下文，任务与传统任务不同，消歧利用实体的流行度和实体共现；

e.社交数据中的实体链接，社交媒体(Twitter)作为重要信息来源。由于社交媒体上下文短，又不规范，因此综合考虑用户信息和交互信息。

2.3 事实知识学习

本章节的研究思路源于 Alexander Schutz 等人对关系抽取的定义：自动识别由一对概念和联系这对概念的关系构成的相关三元组。有观点认为三元组是实体、属性和属性值，也有观点认为是主体、谓词和客体。本文认为本质上是一致的都是抽取与本体相关的事实作为事实知识，不做过多的区别。考虑根据实体间的共享事实知识形成彼此之间的联系，从而构建

知识图谱。事实知识[20]如表 2-1 所示：

表 2-1 事实知识示例表

实体名	莱昂纳多·迪卡普里奥		
属性	属性值	属性	属性值
中文名	莱昂纳多·威尔海姆·迪卡普里奥	外文名	Leonardo Wilelm DiCaprio
国籍	美国	出生地	加利福尼亚州洛杉矶
出生日期	1974年11月11日	职业	演员、制片人
主要成就	柏林国际电影节最佳男主角奖、美国电影金球奖最佳男主角奖、第77届奥斯卡金像奖男主角奖提名、第79届奥斯卡金像奖男主角奖提名、第88届奥斯卡金像奖最佳男主角	代表作品	《泰坦尼克号》、《逍遥法外》、《血钻》、《盗梦空间》

注意，网络文本的信息结构有三种：结构化数据(置信度高、规模小、缺乏个性化的属性信息)，半结构化数据(置信度较高、规模较大、个性化的信息、形式多样、含有噪音)和纯文本(置信度低、复杂多样、规模大)。通常结构化和半结构化的数据可以用网页结构进行信息块的识别、模板的学习以及属性值的抽取从而实现关系抽取。非结构化文本的实体关系抽取成为了研究重点，以下从传统关系抽取、开放关系抽取和关系发现分别进行具体方法的介绍：

2.3.1 传统关系抽取

目前传统关系抽取主要采用统计机器学习的方法，将关系实例转换成高维空间中的特征向量或直接用离散结构来表示，在标注语料库上训练生成分类模型，然后再识别实体间关系。

a. 基于特征向量方法：获取各种有效的词法、句法、语义等特征，并有效地集成起来，从而产生描述实体语义关系的各种局部特征和简单的全局特征。其中，特征选取从自由文本及其句法结构中抽取出各种表面特征以及结构化特征。优点是：实用，计算速度快。缺点是挖掘有效性有限，再提高不易。

b. 基于核函数方法：有效挖掘反映语义关系的结构化信息及有效计算结构化信息之间的相似度。其优点是机构化信息特征挖掘有效，缺点是计算速度缓慢，实用性差。其中，核

代表卷积树核，用两个句法树之间的公共子树的数目来衡量相似度：

1>标准的卷积树核函数(CTK)，在计算两子树相似度时，只考虑子树本身，不考虑子树上下文信息。

2>上下文相关卷积树核函数(CS-CTK)，在计算子树相似度量，同时考虑子树的祖先信息，并对不同祖先的子树相似度加权平均。

c.基于神经网络的方法，设计合理的网络结构，捕捉更多信息，关系抽取准确度提升。该方法属于一种深度的监督学习下的机器学习模型(深度学习)，是通过训练关系分类器，来判别句子中实体之间的语义关系。可规避传统特征抽取需要的 NLP 预处理和人工设计特征产生的错误。不同的网络结构捕捉不同的信息：

1>递归神经网络(RNN)，构建过程更多考虑句子的句法结构，需要依赖复杂的句法分析工具。

2>卷积神经网络(CNN)，通过卷积操作完成句子级信息的捕获，不需要复杂的 NLP（Natural Language Processing)工具。

2.3.2 开放域关系抽取

开放域关系抽取的特点是不限定关系类别和目标文本，也因此产生相应的难题，并且获取训练语料也是一个难题。

a.针对网页，采用一种 TextRunner(传统信息抽取系统专注于在小而均匀的语料中抽取少量精确狭窄预先要求的信息。TextRunner 采用全新的开放信息抽取的方式可实现不行人工输入的情况下，抽取大量相关元组集合)的方式：

1>用户输入特定的谓词和论元。

2>搜索引擎返回与用户输入相关的句子。

3>用 TextRunner 抽取出谓词论元三元组。

b.针对在 Wikipedia 文本中抽取关系信息，采用一种机阅读(Machine Reading)[21]的方式。其具体方法：

1>Infoboxes(维基的信息框)抽取关系信息。

2>维基条目文本中进行回标，产生训练语料。

c.从 NYT(纽约时报)中抽取 Freebase 的关系类别：比较多个数据库发现，Freebase 的实体类别和关系类别最为丰富；

1>人工标注训练集不可行，需要寻求无监督或弱监督的关系抽取方法；

2>将弱监督关系抽取看作是多示例问题；

3>利用分段卷积网络自动学习特性：设计分段最大池化层，更好保留句子结构化信息。

d.基于细粒度实体类型特征发现的弱监督关系抽取，强化粗粒度实体类型特征对特定关系类别的指示作用。其具体方法：

1>细粒度实体类型，将实体类型分类看成文本分类问题，利用搜索引擎扩展实体指称，再用预先训练的类别分类器分类扩展后的实体指称文本。

2>三种融合方法：替换方法、扩展方法和选择方法。

2.3.3 开放域关系发现

关系发现在知识图谱中也就是链接预测(知识库补全)：

1>利用知识图谱中现有的知识推断未知知识(未知的连接)。

2>知识图谱本身蕴含着推理模式。

从知识表示角度，可以分为符号逻辑表示和分布式表示。符号逻辑表示表达能力强、人类可理解，可提供精确的结果。其方法介绍：基于符号逻辑的关系发现，关键技术：事实的抽取、规则的学习和事实在规则上的推理。其中，逻辑推理规则学习：

1>统计关系路径的共现，学习霍恩子句表示的推理规则。

2>利用实体间在图中的链接特征学习关系分类器，得到路径与关系的推理规则。

但是随着知识库的规模越来越大，逻辑表示很难高效的扩展到大规模知识库上；而分布式表示学习可用线性参数来表示指数数量的区域，把知识库中的实体和关系表示为低维空间的对象(向量)及其操作(空间转换)，能够蕴含其在知识库中的性质。其特点:

1>每个实体、关系的表示是通过优化整个知识库的目标函数编码得到的，凝练了整个知识库的信息，能够进行全面的知识推理。拥有一个丰富的相似度空间。

2>是一个统计学习过程，事物间的隐式规律蕴含在对象的表示中，反映在推理过程中。

根据分布式知识表示方式的特点，相应的关系发现方法的介绍：知识库表示学习的方法：

a.张量分解法[22]：张量分解方法将知识图谱编码为一个三维邻接张量，然后分解为一个核心张量和因子矩阵的乘积。分解后的结果作为对应三元组存在与否的概率。

b.基于翻译的模型[23]：将关系看做头部实体到尾部实体的翻译。

c.基于神经网络的能量函数，为三元组打分，通过惩罚错误的三元组完成学习过程。

以上方法中，所有关系的转换矩阵都有相同的自由度(自由变量数目)，并且头部和尾部实体共享同一个转换矩阵。然而，知识库中关系存在：

1>异质性。关系链接很多实体对(复杂关系)；关系链接很少的实体对(简单关系)。

2>不平衡性。关系链接很多头部(尾部)实体和很少的尾部(头部)实体。

不同复杂度的关系应当使用不同表达能力的模型学习，头部尾部实体应分开学习。可使用不同稀疏度的矩阵学习不同复杂性的关系。

d.基于稀疏矩阵的知识库补全方法：在公式 TransR：$f(h,r,t)=\|M_r h-r+M_r t\|^2_2$ 中，每种关系赋予了一个转换 M_r，用稀疏矩阵替换转换矩阵。在使用稀疏矩阵的前提下，可根据不同关系的复杂度调节转换矩阵自由度。这种做法对数据有很强的适应性。

e.基于高斯分布表示学习的补全方法，针对不确定性问题。符号不确定性：实体可用粒度大小表示，关系可用指示性强弱表示。实体不确定性：包含事实越多，含义不确定性越小；关系不确定性：关系越复杂，含义不确定性越大。用多维高斯分布表示：均值向量表示该符号的位置(含义)，协方差矩阵表示该符号的多样性(不确定性)。

f.基于动态映射矩阵的补全方法[24]：通过对事实中的实体和关系构造动态映射矩阵，表示知识的多样性。主要解决“一对多、多对一、多对多”类型的关系。

分布式表示和逻辑表示融合的混合方法：融合分布表示的概率图模型。将表示学习和概率推理融合：

1>将知识图谱中的元素进行向量化的表示。

2>用马尔科夫逻辑网学习推理规则。

3>在事实推理中，将向量化中得到的候选元组的相似度值作为随机游走的状态转移概率注入到逻辑推理中，减少搜索空间，提高效率。

当然，一些具体操作的例子值得借鉴。对（半）结构化文档：如 COHEN W 等人提出的的方法 WR[25]、KOSALA R 等人提出的树自动机[26]、TENIER S 等人提出的利用最小子树[27]、ZHAI Yan-hong 等基于树匹配的部分匹配方法[28]、ALVAREZ M 等人利用聚类技术[29]以及 HONG Ming-cai 等基于机器学习两阶段语义标注的方法[30]、iASA 语义标注方法[31]等。

对纯文本自由结构的文档：MeatAnnot 系统、方法 Armadillo、KnowItAll 和 PAN-KOW、基于 Cyc 的自动语义标注系统、KIM 平台、iOkra 框架等。其中都是些自动或者半自动的知识抽取技术，并且英文抽取的技术已经较为成熟、效果良好，中文抽取技术由于中文自身的特点具有一定的挑战，仍具有很大的研究空间。

2.4 事件知识学习

事件抽取就是针对特定领域的事件进行相关信息的抽取。一个事件包括时间、地点、原因、实体行为等多个因素。与关系抽取相比，多一个事件触发词，更多的论点个数，需要预测的目标也不再是两个概念间的语义关系，而是触发词及事件类型，事件元素及其扮演的角色。

事件抽取的典型方法:

基于模式匹配的方法：

1>平面模式主要基于词袋等字符串特征。

2>结构模式更多地考虑了句子的结构信息，融入句法分析特征。

存在一定缺点：领域相关，可扩展性差；需要大量的人工标注，耗时耗力。一般过程如图 2-2 所示：

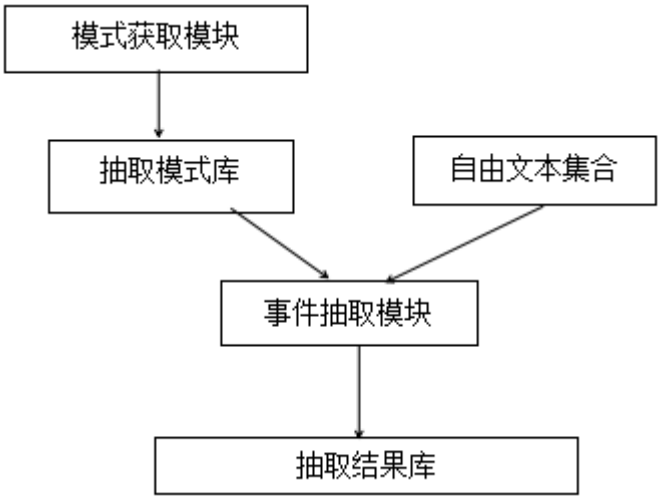

图 2-2 模式匹配过程示意图

b.基于机器学习的方法从特征和结构两个角度：

基于特征的方法：将事件抽取看成一个多分类的问题。

1.利用句子级信息，从句子中提取特征，利用最大熵、朴素贝叶斯和支持向量机等模型去完成事件抽取。代表方法：基于动态最大池化技术的卷积神经网络(DMCNN)[32]：

1>4 部分组成：嵌入词学习(word-embedding learning)、词本身特征代表(lexical-level feature representation)、句子上下文特征提取(sentence-level feature extraction)和论点分类输出(argument classifier output)。

2>传统的 CNN 利用最大池化技术，选取最大值代表每个特征映射，但事件抽取任务中一句话存在两个或者多个事件。提出动态最大池化技术来根据触发词和论点候选的位置，在不损失最大值的前提下，获取更多的有用信息。

2.利用篇章级信息[33]，在句子级信息基础上更多地考虑篇章级信息和丰富的背景知识(跨文档信息、跨语言信息、跨文本事件信息、跨实体信息)。代表方法：基于概率软逻辑推

理[34]，可利用全局信息(如，事件之间的相关性)和更精确的局部信息(隐含的局部信息):

1>局部模块基于局部，信息做出初步分类(会考虑到细粒度实体类别等精准的信息)。

2>全局模块从语料中收集全局信息，学习到事件与话题之间、事件与事件间的共现信息。

3>综合以上信息进行全局推理。

通常，按照触发器检测(Trigger Detection)、论点分类(Argyment Classification)、属性分类(Attributes Classification)和上报分类(Reportability Classification)的顺序进行多分类。然而将事件抽取分成多个步骤，会导致错误由上而下传递，没有考虑触发词和时间元素间的相互影响。

基于结构的方法：将事件抽取成一个最优结构预测：

1>看做依存树结构预测问题。

2>自定义联合结构的预测问题：触发词和事件元素联合预测。一个词存在歧义，同时考虑事件元素和事件触发词的标注会提升最终效果。

基于事件的抽取也是自然语言处理领域重要的研究方向之一，是信息抽取较高级的任务、是热点和难点，因此相关研究很多，该思路的提出是为了实现自动地将非结构化的文本结构化，方便进一步的利用信息。事件抽取对于突发事件[35]、灾难性事件[36]的研究相对较多并且更好准确，并且擅长进一步追踪分析。当然，提高新类型事件的抽取能力是新的研究挑战与目标。

3 知识图谱应用问答系统

问答系统[37]的发展经历了长期的探索与改进，现有的问答系统分为：聊天机器人、基于知识库的问答系统、问答式检索系统和基于自由文本的问答系统。其中，前三者都制约于知识库规模的大小，知识库覆盖的问答高效而准确，反之性能骤减为零。而后者因为避免了建立大规模知识库被看好代表着问答系统的发展方向：Cymfony 公司的 Textract 系统是一个成功的代表。

知识图谱的诞生使得知识库的规模有了量的提升。据不完全统计，目前 Google Graph 包含了 5 亿实体和 35 亿条事实，同时百度和搜狗推出的知识图谱规模仅略小于 Google 的。同时技术上，对于怎样做的问题，若采用深度分析逻辑化的模式，需要至少 500 条规则覆盖所有模式，而应用知识图谱则是可以开发的。所谓知识图谱就是数据的管理手段，这种技术能弥补机器的缺失：机器对语言认知和概念认知的巨大障碍，即，缺失的类似人脑中的海量而有组织的知识体系。当然，知识图谱也非万能的，对于在线的问询，也存在语调识别不出，以及机器的自我学习能力制约的问题。知识图谱的发展也是机遇与挑战并存。

从基于关键词匹配、信息抽取并基于浅层语义分析的 IR-based QA 到依赖于网民贡献的问答系统，再到知识库问答系统。问答系统一般包括三个主要组成部分:问题分析、信息检索和答案抽取。一般问答系统流程如图 3-1 所示：

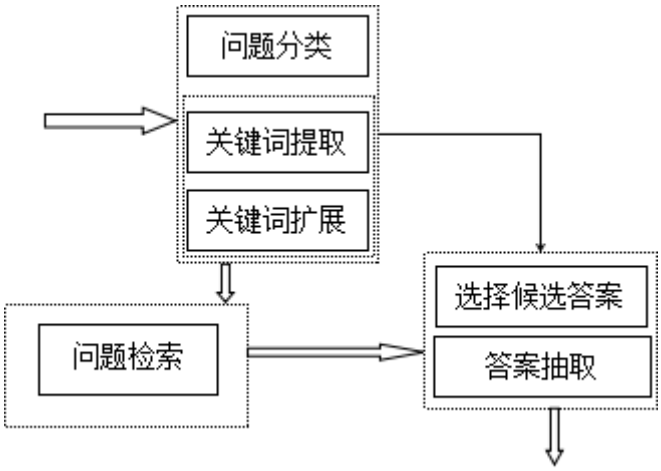

图 3-1 一般问答系统流程示意图

发展以来，知识库规模的发展成为问答系统发展的瓶颈。互联网的转型，知识图谱的诞生逐渐成为突破瓶颈的可能。知识库问答系统将会有所突破，逐渐吸引发展研究的目光：先进行问句语义解析，再进行语义表示，最后进行语义匹配查询得到推理的过程。

将深度学习的算法模型应用知识图谱的构建可以使得知识图谱信息更加的丰富，表示更多的信息。进而，将问句分析在知识图谱中进行最优子图的匹配实现问啥答啥的智能问答系统。采用知识图谱的方式查询相比传统搜索引擎简单字符串匹配，让搜索结果与查询内容更加紧实。同时知识图谱的构建源自海量数据的整合处理，也避免了跨领域问题查询偏差。也就是说，在知识图谱这张知识网编织成功的基础上，通过各种逻辑算法分析处理自然语言实现抓住最符合问题的关键匹配知识图谱，从而找出答案，实现智能问答。

问答系统实现案例中，最为成功的当属 IBM 于 2011 年研发的超级计算机“沃森”，在美国的知识竞赛节目《危险边缘 jeopardy!》的出色表现，成功战胜人类选手。其采用深度开放域技术，其逻辑架构图如图 3-2 所示：

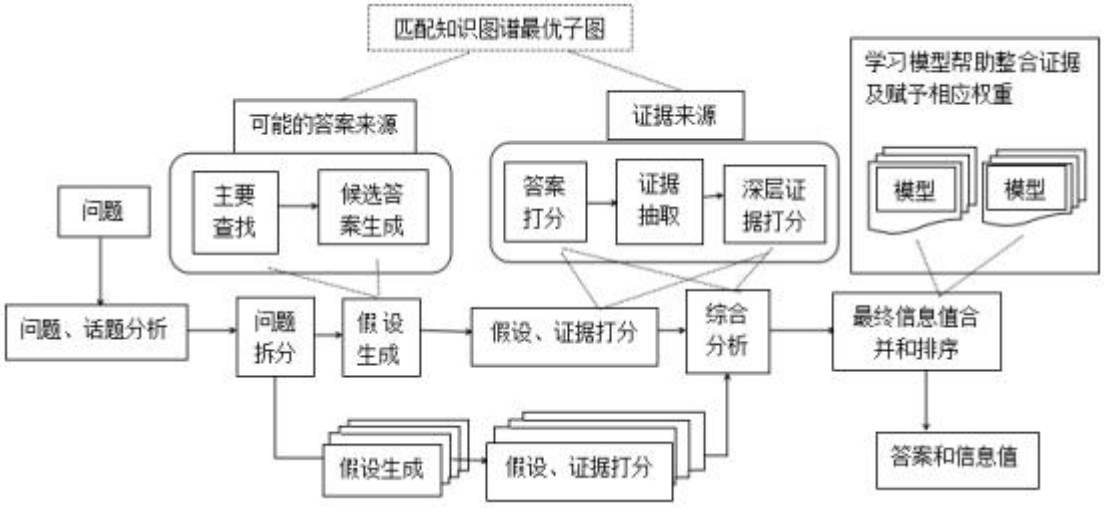

图 3-2 深度开放域“沃森”问答系统逻辑架构图

仔细分析比赛现场可发现，沃森在现场获取选手的错误答案的信息同时进行改正方面还存在不足，同时沃森有一定的排除干扰信息的能力，抓住问题的要点回答。由此可看出知识图谱固然重要，是问答系统的重要支撑，同时仅有强大而全面的知识库也是不足够的，还需要其他技术的支撑才能使得问答系统更为智能。自然语言理解(Natural language

understanding)技术使得具有更强的理解并交互的能力甚至一些双关、暗喻的知识。同时动态分析各类假设和问题、精细的个性化分析能力、在相关数据的基础上优化问题解答都是沃森成功的关键。IBM 超级电脑沃森的中国团队专家、IBM 中国研究院资深经理潘越接受采访时说到沃森系统几个特点：

1》沃森的领域知识库包括百科全书、字典、地理类娱乐类的专题数据库、新闻报道、经典著作等约 70GB。

2》依赖上下文去除歧义。如：考察上下文中提到的人、事件、地点等其它相关对象。从多个证据的角度去解释这一问题。即使在一个算法中的去歧义失败了，其它上百个算法会从其它方面给出答案是否正确的证据。这就削弱了对单个算法去歧义能力的依赖。

同时特别提到对于自然语言的处理是沃森的优势所在。

4 总结与展望

在深度学习应用于知识图谱中是近年来研究的热潮，也是研究的起步。深度学习在很多领域与技术已经有显著地成效，能够有一定的学习能力却不能完全自主学习；在自然语言的处理上已经有了飞跃，却存在不擅长的领域，没有达到完美的状态。同时深度学习在知识库问答的应用上也是重要技术，同样存在很多的挑战。知识图谱应用的问答系统中是存在两大难点的，一是问句的解析挂载查询：查询单关系类型问题存在的困难较少容易优化，而查询者实际应用中往往不是简单单关系的查询，使得难度提升。将问题分解：由于搜集并审议了各子问题的相关证据，最终答案的分值应该比未分解时较高。通过实验表明，即使不需要分解的问题，也会因分解提高整体的置信度。二是知识图谱的构建：知识图谱研究越来倾向于多源异构，这就使得需要处理上的变化。日益剧增的海量数据，一种不停更新不断智能学习的算法是需要的，才能实现一张知识面广而准确的知识图谱。知识图谱在问答系统中处于基础建筑的关键作用，也是决定上层建筑的坚实基础，对于问答系统来说，各个技术各个算法的完美结合做到统一，相互配合才是技术发展、市场应用的关键。

[参考文献] (References)

[1] 王树西， 问答系统:核心技术、发展趋势[J]. 计算机工程与应用， 2005， 41(18):1-3.
[2] 曹树金， 吴育冰， 韦景竹，等. 知识图谱研究的脉络、流派与趋势--基于 SSCI 与 CSSCI 期刊论文的计量与可视化[J]. 中国图书馆学报， 2015(5):16-34.
[3] Shirakawa M, Nakayama K, Hara T, et al. Concept vector extraction from Wikipedia category network.[C]// International Conference on Ubiquitous Information Management and Communication, Icuimc 2009, Suwon, Korea, January. 2009:71-79.
[4] 廉成洋， 毛宇光， 黄玉明. 基于启发式规则的 Web 信息抽取技术研究[J]. 计算机技术与发展， 2009，19(8).
[5] 唐杰， 梁邦勇， 李涓子，等. 语义 Web 中的本体自动映射[J]. 计算机学报， 2006， 29(11):1956-1976.
[6] 靖红芳， 王斌， 杨雅辉. 基于类别分布的特征选择框架[C]// 全国信息检索与内容安全学术会议. 2008:1586-1593.
[7] 刘洁晶， 张建光. 信念传播算法在分类模型中的应用[J]. 福建电脑， 2012， 28(10):6-6.
[8] Zheng Z. Feature selection for text categorization on imbalanced data[J]. Acm Sigkdd Explorations Newsletter, 2004, 6(1):80-89.
[9] 徐燕， 李锦涛， 王斌，等. 不均衡数据集上文本分类的特征选择研究[J]. 计算机研究与发展， 2007，44(z2):58-62.
[10] Han X, Zhao J. Named entity disambiguation by leveraging wikipedia semantic knowledge[C]. Conference on Information and Knowledge Management, 2009.
[11] 肖芳. 异构系统中实体识别研究[J]. 自动化与信息工程， 2009， 30(3):35-37.
[12] 李石生， 刘海博， 路小英，等. 基于探测查询的 Deep Web 实体识别[C]// 全国搜索引擎和网上信息挖掘学术研讨会. 2008.
[13] 徐红艳，党晓婉， 冯勇，等. 基于 BP 神经网络的 Deep Web 实体识别方法[J]. 计算机应用， 2013，33(3):776-779.

[14] 陆伟， 武川. 实体链接研究综述[J]. 情报学报， 2015(1):105-112.
[15] Han,X.&Zhao,J.2009.Named entity disambiguation by leveraging wikipedia semantic knowledge[C]//Proceeding of the 18th ACM conference on Information and knowledge management,pp.215-224.
[16] 张涛， 刘康， 赵军. 一种基于图模型的维基概念相似度计算方法及其在实体链接系统中的应用[J]. 中文信息学报， 2015， 29(2):58-67.
[17] Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method[C]// Proceeding of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July. 2011:765--774.
[18] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]// Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011:782-792.
[19] Wang Z, Li J, Wang Z, et al. Cross-lingual knowledge linking across wiki knowledge bases[C]// International Conference on World Wide Web. ACM, 2012:459-468.
[20] 刘鹏博， 车海燕， 陈伟. 知识抽取技术综述[J]. 计算机应用研究， 2010， 27(9):3222-3226.
[21] Poon H. Markov logic for machine reading[C]// University of Washington, 2011.
[22] Nickel M, Tresp V. Tensor Factorization for Multi-relational Learning[M]. Springer Berlin Heidelberg:Machine Learning and Knowledge Discovery in Databases, 2013:617-621.
[23] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[J]. AAAI - Association for the Advancement of Artificial Intelligence, 2014.
[24] Ji G, He S, Xu L, et al. Knowledge Graph Embedding via Dynamic Mapping Matrix[C]// Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing. 2015.
[25] COHEN W，HURST M，JENSEN L． A flexible learning system for wrapping tables and lists in HTML documents［C］/ /Proc of the 11th International World Wide Web Conference． New York: ACM Press，2002: 232-241．
[26] KOSALA R，BLOCKEEL H，BRUYNOOGHE M，et al． Information extraction from structured documents using k-testable tree automaton inference［J］． Data ＆ Knowledge Engineering，2006，58(2):129-158．
[27] TENIER S，TOUSSAINT Y，NAPOLI A，et al． Instantiation of relations for semantic annotation［C］/ /Proc of IEEE/WIC/ACM International Conference on Web Intelligence． Washington DC:IEEE Computer Society，2006: 463-472．
[28] ZHAI Yan-hong，LIU Bing． Structured data extraction from the Web based on partial tree alignment［J］． IEEE Trans on Knowledge and Data Engineering，2006，18(12): 1614-1628．
[29] ALVAREZ M，PAN A，RAPOSO J，et al． Extracting lists of data records from semi-structured Web pages［J］． Data ＆ Knowledge Engineering 2008，64(2): 491-509．
[30] HONG Ming-cai，TANG Jie，LI Juan-zi． Semantic annotation using horizontal and vertical contexts［C］/ /Proc of the 1st Asian Semantic Web Conference． Berlin: Springer，2006: 58-64．
[31] TANG Jie，LI Juan-zi，LU Hong-jun，et al． iASA: learning to annotate the semantic Web［J］． Journal on Data Semantic，2005，3740(4): 110-145．
[32] Chen Y, Xu L, Liu K, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]// The, Meeting of the Association for Computational Linguistics. 2015.
[33] Liao S, Grishman R. Using Document Level Cross-Event Inference to Improve Event Extraction.[C]// ACL 2010, Proceedings of the, Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden. 2010:789-797.
[34] He S, Liu K, Zhang Y, et al. Question Answering over Linked Data Using First-order Logic[C]// Conference on Empirical Methods in Natural Language Processing. 2014.
[35] 陈湧. 基于知识元的突发事件案例信息抽取及检索[D]. 大连：大连理工大学, 2014.
[36] 钟涛, 陈群秀. 基于 Web 主题性信息检索的灾难性事件信息抽取系统[C]// 中文信息处理国际会议. 2007.
[37] 毛先领， 李晓明. 问答系统研究综述[J]. 计算机科学与探索， 2012， 06(3):193-207.