知识图谱研究进展 The Research Advances of Knowledge Graph
（完整 27k 字 15 页经典收藏版，文末打赏下载重点彩排 PDF）

漆桂林等作 秦陇纪 10 汇编

目录

作者简介：**漆桂林**（1977-），博士，教授，研究方向：人工智能、知识工程、语义网，email：gqi@seu.edu.cn；高桓（1984-），博士研究生，研究方向：数据挖掘，信息抽取，知识库构建；吴天星（1990-），博士研究生，研究方向：知识图谱，语义 Web，知识挖掘。通信地址：东南大学计算机科学与工程学院 南京 211189。

基金项目：本文受**国家自然科学基金**面上项目：基于图的并行 OWL 本体推理方法研究（61672153）的资助。

随着大数据时代的到来，**知识工程**受到了广泛关注，如何从海量的数据中提取有用的知识，是大数据分析的关键。知识图谱技术提供了一种从海量文本和图像中抽取**结构化知识**的手段，从而具有广阔的应用前景。本文首先简要回顾**知识图谱的历史**，探讨知识图谱研究的意义。其次，介绍知识图谱构建的关键技术，包括实体关系识别技术、知识融合技术、实体链接技术和知识推理技术等。然后，给出现有开放的知识图谱数据集的介绍。最后，给出知识图谱在情报分析中的应用案例。

关键词：人工智能，知识图谱，知识挖掘，情报分析 中图分类号：G35

With the advent of big data era, knowledge engineering has attracted wide attention, as mining knowledge from large-scale

data is critical for big data analysis. Knowledge graph techniques provide a way to extract structured knowledge from large-scale texts and images, thus have wide application prospect. In this article, we first gave a brief overview of the history of knowledge graph, and discussed the importance of knowledge graph research. We then introduced key technologies of knowledge graph, including techniques of instance relation detection, techniques of knowledge fusion, techniques of instance mapping, and techniques of knowledge reasoning. After that, we introduced some well-known open knowledge graph datasets. Finally, we presented some use cases of knowledge graph in intelligence analysis.

Keywords: Artificial intelligence, knowledge graph, knowledge mining, intelligence analysis

1 知识图谱历史回顾

知识图谱（Knowledge Graph）的概念由谷歌 2012 年正式提出，旨在实现更智能的搜索引擎，并且于 2013 年以后开始在学术界和业界普及，并在智能问答、情报分析、反欺诈等应用中发挥重要作用。知识图谱本质上是一种叫做语义网络（semantic network）的知识库，即具有有向图结构的一个知识库，其中图的结点代表实体（entity）或者概念（concept），而图的边代表实体 / 概念之间的各种语义关系，比如说两个实体之间的相似关系。语义网络[1]是 20 世纪 50 年代末 60 年代初提出，代表性人物有 M. Ross Quillian 和 Robert F. Simmons。语义网络可以看成是一种用于存储知识的数据结构，即基于图的数据结构，这里的图可以是有向图，也可以是无向图。使用语义网络，可以很方便地将自然语言的句子用图来表达和存储，用于机器翻译[2]、问答系统[3]和自然语言理解[4]。20 世纪 70 年代开始有不少工作研究语义网络跟一阶谓词逻辑之间的关系，比如说，文献[5]提供了一个算法将一个语义网络转化成谓词逻辑的形式，但是具有计算方面的优势，而文献[6]则给出了如何用语义网络来表示一阶谓词逻辑中的连接词和量词。到了 20 世纪 80 年代，人工智能研究的主流变成了**知识工程和专家系统**，特别是基于规则的专家系统开始成为研究的重点。这一时期，语义网络的理论更加完善，特别是基于语义网络的推理出现了很多工作(例如文献[7]中的工作),而且语义网络的研究开始转向具有严格逻辑语义的表示和推理。20 世纪 80 年代末到 90 年代，语义网络的工作集中在对于概念（concept）之间关系的建模，提出了术语逻辑（terminological logic）以及描述逻辑。这一时期比较有代表性的工作是 Brachman 等人提出的 CLASSIC 语言[8]和 Horrock 实现的 FaCT 推理机[9]。进入 21 世纪，语义网络有了一个新的应用场景，即语义 Web。语义 Web 是由 Web 的创始人 Berners-Lee 及其合作者提出[10]，通过 W3C①的一些标准来实现 Web 的一个扩展，从而数据可以在不同应用中共享和重用。语义 Web 跟传统 Web 的一个很大的区别是用户可以上传各种图结构的数据（采取的是 W3C 的标准 RDF），并且数据之间建立链接，从而形成链接数据[11]。链接数据项目汇集了很多高质量知识库，比如说 Freebase②、DBpedia③和 Yago④，这些知识库都是来源于人工编辑的大规模知识库－维基百科。这些高质量的知识库的发布，为谷歌知识图谱项目的成功打下了坚实的基础。

①https://www.w3.org
②https://en.wikipedia.org/wiki/Freebase
③http://wiki.dbpedia.org
④http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

谷歌知识图谱很重要的一部分是一个大规模的协同合作的知识库，叫 Freebase，即链接数据的一个数据集。Freebase 采用的数据结构是**图模型**，即可以把一个 Freebase 的知识库看成是有向图，这种数据模型相对于传统数据库的优势在于可以处理更复杂的数据以及方便数据的插入。谷歌知识图谱的**模式（Schema）**是由谷歌自己的专业团队在 Freebase 的基础上开发和设计的。谷歌知识图谱中，所有的对象都有属于它的 Type。Type 的数量不是固定的，有一个**数据结构 Collection 记录**的是计算机自动抽取出的类型，Collection 中有成千上万种类型，有些今天生成后第二天就被删除了，有些则能长期的保留在 Collection 中，如果 Collection 中的某一种类型能够长期的保留，发展到一定程度后，由专业的人员进行决策、命名，最后上升为一种 Type，作为谷歌知识图谱的一种类型保存在模式中。谷歌知识图谱的 Type 有音乐家、网球运动员等等。不过谷歌的知识图谱中的模式并没有太多去考虑类型的层次性。虽然很多文献都把知识图谱看成是一个**实体 - 关系**的有向图，但是也有一些观点认为知识图谱应该包含更抽象的概念之间的关系，比如说，谷歌和必应、雅虎一起推出了 Schema.org[2]来提供一个覆盖广泛主题（包括人物、地点、事件等）的模式（schema）。

跟早期的语义网络相比，知识图谱具有自己的特点。首先，知识图谱强调的是实体之间的关联，以及实体的属性值，虽然知识图谱中也可以有概念的层次关系，这些关系的数量相比实体之间的关系的数量要少很多，而早期的语义网络主要用于对自然语言的句子做表示；其次，知识图谱的一个重要来源是百科，特别是百科中半结构化的数据抽取得到，这跟早期语义网络主要靠人工构建不一样，通过百科获取高质量知识作为种子知识，然后通过知识挖掘技术可以快速构建大规模、高质量知识图谱；最后，知识图谱的构建强调不同来源知识的融合以及知识的清洗技术，而这些不是早期语义网络关注的重点。

⑤https://www.w3.org/TR/rdf-schema/
⑥https://www.w3.org/TR/owl2-overview/

知识图谱跟本体标准语言，比如说 RDFS⑤和 OWL⑥具有紧密的关系。一方面，知识图谱可以看成是一种知识存储的数据结构，本身并不具备形式化的语义，但是可以通过 RDFS 或者 OWL 的规则应用于知识图谱进行推理，从而赋予知识图谱形式化语义。另外一方面，并不是所有的 OWL 本体都适合转化成知识图谱，因为转化过程中会丢失语义信息（在文献[12]中，OWL EL 语言表示的本体已经被证明适合转化成知识图谱，并且可以实现高效推理机）。

下面几个小节将介绍知识图谱构建的关键技术、一些开放知识图谱以及知识图谱在情报分析的应用案例。

2 知识图谱构建技术

本节首先给出知识图谱的技术地图，然后介绍知识图谱构建的关键技术，包括关系抽取技术、知识融合技术、实体链接技术和知识推理技术。

2.1 知识图谱技术地图

构建知识图谱的主要目的是获取大量的、让计算机可读的知识。在互联网飞速发展的今天，知识大量存在于非结构化的文本数据、大量半结构化的表格和网页以及生产系统的结构化数据中。为了阐述如何构建知识图谱，本文给出了构建知识图谱的技术地图，该**技术地图**如图 1 所示。整个技术图主要分为三个部分，第一个部分是知识获取，主要阐述如何从非结构化、半结构化、以及结构化数据中获取知识。第二部是数据融合，主要阐述如何将不同数据源获取的知识进行融合构建数据之间的关联。第三部分是知识计算及应用，这一部分关注的是基于知识图谱计算功能以及基于知识图谱的应用。

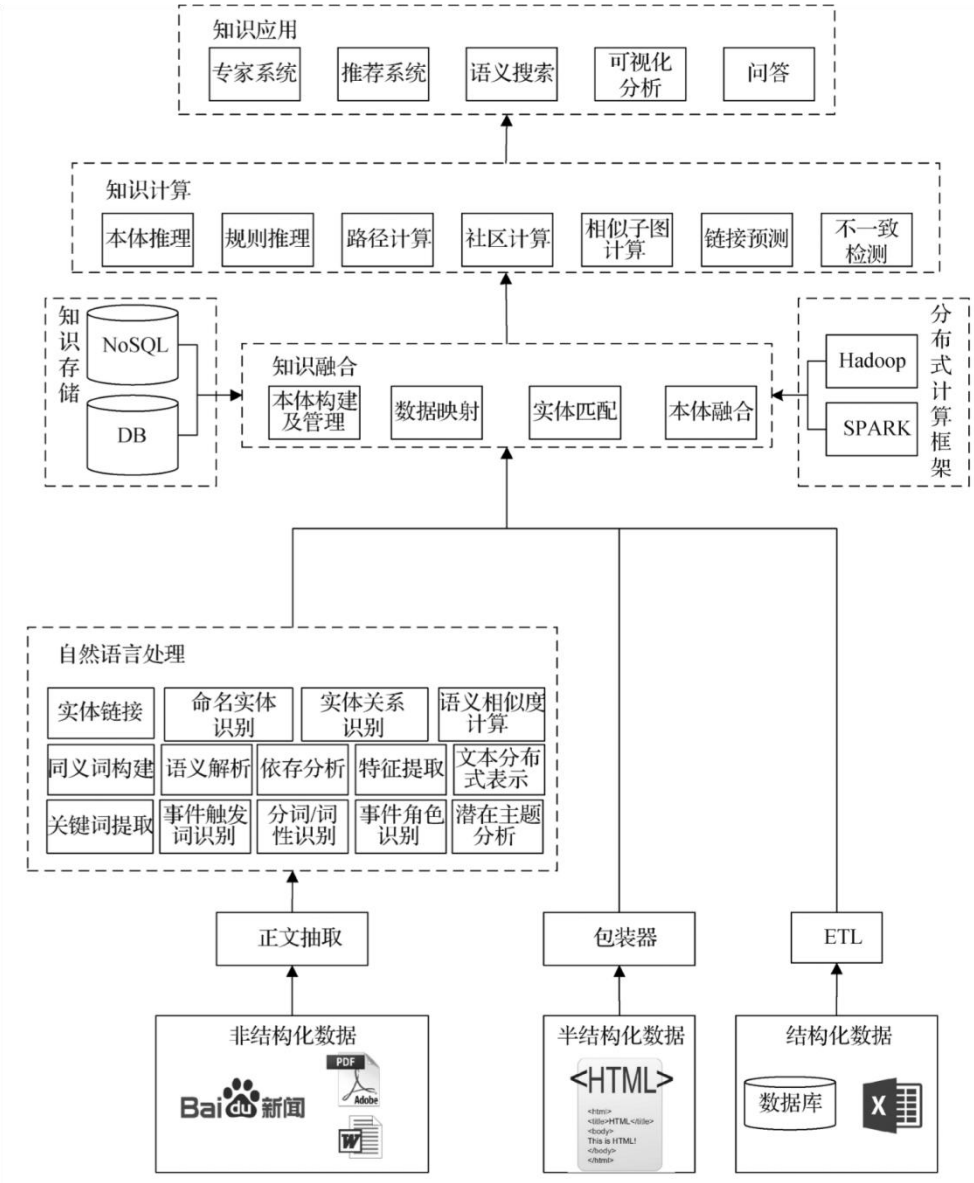

图 1 知识图谱技术

2.1.1 知识获取

在处理非结构化数据方面，首先要对用户的非结构化数据提取**正文**。目前的互联网数据存在着大量的广告，**正文提取技术**希望有效的过滤广告而只保留用户关注的文本内容。当得到正文文本后，需要通过自然语言技术识别文章中的**实体**，实体识别通常有两种方法，一种是用户本身有一个知识库则可以使用实体链接将文章中可能的候选实体链接到用户的知识库上。另一种是当用户没有知识库则需要使用命名实体识别技术识别文章中的实体。若文章中存在实体的别名或者简称还需要构建实体间的**同义词表**，这样可以使不同实体具有相同的描述。在识别实体的过程中可能会用到**分词、词性标注**，以及深度学习模型中需要用到分布式表达如词向量。同时为了得到不同粒度的知识还可能需要提取文中的关键词，获取文章的潜在主题等。当用户获得实体后，则需要关注实体间的关系，我们称为**实体关系识别**，有些实体关系识别的方法会利用句法结构来帮助确定两个实体间的关系，因此在有些算法中会利用**依存分析**或者**语义解析**。如果用户不仅仅想获取实体间的关系，还想获取一个事件的详细内容，那么则需要确定事

件的触发词并获取事件相应描述的句子，同时识别事件描述句子中实体对应事件的角色。在处理**半结构化数据**方面，主要的工作是通过包装器学习半结构化数据的**抽取规则**。由于半结构化数据具有大量的重复性的结构，因此对数据进行少量的标注，可以让机器学出一定的规则进而在整个站点下使用规则对同类型或者符合某种关系的数据进行抽取。最后当用户的数据存储在生产系统的数据库中时，需要通过 ETL 工具对用户生产系统下的数据进行重新组织、清洗、检测最后得到符合用户使用目的数据。

2.1.2 知识融合

当知识从各个数据源下获取时需要提供统一的术语将各个数据源获取的知识融合成一个庞大的**知识库**。提供统一术语的结构或者数据被称为本体，本体不仅提供了统一的**术语字典**，还构建了各个术语间的关系以及限制。本体可以让用户非常方便和灵活的根据自己的业务建立或者修改数据模型。通过数据映射技术建立本体中术语和不同数据源抽取知识中词汇的映射关系，进而将不同数据源的数据融合在一起。同时不同源的实体可能会指向现实世界的同一个**客体**，这时需要使用实体匹配将不同数据源相同客体的数据进行融合。不同本体间也会存在某些术语描述同一类数据，那么对这些本体间则需要本体融合技术把不同的本体融合。最后融合而成的知识库需要一个存储、管理的解决方案。知识存储和管理的解决方案会根据用户查询场景的不同采用不同的存储架构如 NoSQL 或者关系数据库。同时大规模的知识库也符合大数据的特征，因此需要传统的大数据平台如 Spark 或者 Hadoop 提供高性能计算能力，支持快速运算。

2.1.3 知识计算及应用

知识计算主要是根据图谱提供的信息得到更多隐含的知识，如通过本体或者规则推理技术可以获取数据中存在的隐含知识；而**链接预测**则可预测实体间隐含的关系；同时使用社会计算的不同算法在知识网络上计算获取知识图谱上存在的社区，提供知识间关联的路径；通过不一致检测技术发现数据中的噪声和缺陷。通过知识计算知识图谱可以产生大量的智能应用如可以提供精确的用户画像为精准营销系统提供潜在的客户；提供领域知识给专家系统提供决策数据，给律师、医生、公司 CEO 等提供辅助决策的意见；提供更智能的检索方式，使用户可以通过自然语言进行搜索；当然知识图谱也是**问答**必不可少的重要组建。

从图 1 可以看出，知识图谱涉及的技术非常多，每一项技术都需要专门去研究，而且已经有很多研究成果。由于篇幅的限制，本文重点介绍知识图谱构建和知识计算的几个核心技术。

2.2 实体关系识别技术

最初实体关系识别任务在 1998 年 **MUC （Message Understanding Conference）** 中 以 MUC-7 任务被引入，目的是通过填充关系模板槽的方式抽去文本中特定的关系。1998 年后，在 **ACE（Automatic Content Extraction）**中被定义为关系检测和识别的任务；2009 年 ACE 并入 **TAC(Text Analysis Conference)**，关系抽取被并入到 **KBP（Knowledge Base Population）**领域的槽填充任务。从关系任务定义上，分为**限定领域（Close Domain）**和**开放领域（Open IE）**；从方法上看，实体关系识别了从流水线识别方法逐渐过渡到端到端的识别方法。

基于统计学的方法将从文本中识别实体间关系的问题转化为**分类问题**。基于统计学的方法在实体关系识别时需要加入实体关系上下文信息确定实体间的关系，然而基于监督的方法依赖大量的标注数据，因此半监督或者无监督的方法受到了更多关注。

（1）**监督学习**：Zhou[13] 在 Kambhatla 的基础上加入了基本词组**块信息和 WordNet**，使用 **SVM** 作为**分类器**，在实体关系识别的准确率达到了 55.5%，实验表明实体类别信息的特征有助于提高关系抽取性能；Zelenko[14] 等人使用**浅层句法分析树**上最小公共子树来表达关系实例，计算两颗子树之间的核函数，通过训练例如 SVM 模型的分类器来对实例进行分。但基于核函数的方法的问题是召回率普遍较低，这是由于相似度计算过程匹配约束比较严格，因此在后续研究对基于核函数改进中，大部分是围绕改进召回率。但随着时间的推移，语料的增多、深度学习在图像和语音领域获得成功，信息抽取逐渐转向了基于神经模型的研究，相关的语料被提出作为测试标准，如 SemEval-2010 task 8[15]。基于神经网络方法的研究有，Hashimoto[16] 等人利用 Word Embedding 方法从标注语料中学习特定的**名词对**的**上下文特征**，然后将该特征加入到**神经网络分类器**中，在 SemEval-2010 task 8 上取得了 F1 值 82.8% 的效果。基于神经网络模型显著的特点是不需要加入太多的特征，一般可用的特征有词向量、位置等，因此有人提出利用基于联合抽取模型，这种模型可以同时抽取实体和其之间的关系。联合抽取模型的优点是可以避免流水线模型存在的**错误累积** [17-22]。其中比较有代表性的工作是 [20]，该方法通过提出全新的全局特征作为算法的软约束，进而同时提高关系抽取和实体抽取的准确率，该方法在 ACE 语料上比传统的流水线方法 F1 提高了 1.5%。另一项工作是 [22]，利用**双层的 LSTM-RNN 模型**训练分类模型，第一层 LSTM 输入的是词向量、位置特征和词性来识别实体的类型。训练得到的 LSTM 中隐藏层的分布式表达和实体的分类标签信息作为第二层 RNN 模型的输入，第二层的输入实体之间的依存路径，第二层训练对关系的分类，通过神经网络同时优化 LSTM 和 RNN 的模型参数，实验与另一个采用神经网络的联合抽取模型 [21] 相比在关系分类上有一定的提升。但无论是流水线方法还是联合抽取方法，都属于有监督学习，因此需要大量的训练语料，尤其是对基于神经网络的方法，需要大量的语料进行模型训练，因此这些方法都不适用于构建大规模的 Knowledge Base。

（2）**半（弱）监督学习**：半监督学习主要是利用少量的标注信息进行学习，这方面的工作主要是基于 Bootstrap 的方法。基于 Bootstrap 的方法主要是利用少量的实例作为**初始种子的集合**，然后利用 pattern 学习方法进行学习，通过不断的迭代，从非结构化数据中抽取实例，然后从新学到的实例中学习新的 pattern 并扩种 pattern 集合。Brin[23] 等人通过少量的实例学习**种子模板**，从网络上大量非结构化文本中抽取新的实例，同时学习新的抽取模板，其主要贡献是构建了 DIPRE 系统；Agichtein[24] 在 Brin 的基础上对新抽取的实例进行可信度的评分和完善关

系描述的模式，设计实现了 Snowball 抽取系统；此后的一些系统都沿着 Bootstrap 的方法，但会加入更合理的对 pattern 描述、更加合理的限制条件和评分策略，或者基于先前系统抽取结果上构建大规模 pattern；如 NELL（Never-Ending Language Learner）系统[25-26]，NELL 初始化一个本体和种子 pattern，从大规模的 Web 文本中学习，通过对学习到的内容进行打分来提高准确率，目前已经获得了 280 万个事实。

（3）**无监督学习**：Bollegala[27] 从搜索引擎摘要中获取和聚合**抽取模板**，将模板聚类后发现由实体对代表的隐含语义关系；Bollegala[28] 使用联合聚类（Co-clustering）算法，利用关系实例和关系模板的对偶性，提高了关系模板聚类效果，同时使用 L1 正则化 Logistics 回归模型，在关系模板聚类结果中筛选出代表性的抽取模板，使得关系抽取在准确率和召回率上都有所提高。

无监督学习一般利用语料中存在的大量冗余信息做聚类，在聚类结果的基础上给定关系，但由于聚类方法本身就存在难以描述关系和低频实例召回率低的问题，因此无监督学习一般难以得很好的抽取效果。

2.3 知识融合技术

知识融合（knowledge fusion）指的是将多个数据源抽取的知识进行融合。与传统**数据融合（data fusion）**[29] 任务的主要不同是，知识融合可能使用多个知识抽取工具为每个数据项从每个数据源中抽取相应的值，而数据融合未考虑多个抽取工具[30]。由此，知识融合除了应对抽取出来的事实本身可能存在的噪音外，还比数据融合多引入了一个噪音，就是不同抽取工具通过实体链接和本体匹配可能产生不同的结果。另外，知识融合还需要考虑本体的融合和实例的融合。

文献［30］首先从已有的数据融合方法中挑选出易于产生有意义概率的、便于使用基于 MapReduce 框架的、有前途的最新方法，然后对这些挑选出的方法做出以下改进以用于知识融合：将每个抽取工具同每个信息源配对，每对作为数据融合任务中的一个数据源，这样就变成了传统的数据融合任务；改进已有数据融合方法使其输出概率，代替原来的真假二值；根据知识融合中的数据特征修改基于 MapReduce 的框架。文献［31］提出一个将通过不同搜索引擎得到的知识卡片（即结构化的总结）融合起来的方法。针对一个实体查询，不同搜索引擎可能返回不同的知识卡片，即便同一个搜索引擎也可能返回多个知识卡片。将这些知识卡片融合起来时，同文献［30］中提出的方法类似，将知识融合中的三维问题将为二维问题，再应用传统的数据融合技术。不过，文献［31］提出了一个新的概率打分算法，用于挑选一个知识卡片最有可能指向的实体，并设计了一个基于学习的方法来做属性匹配。

在知识融合技术中，本体匹配扮演着非常重要的角色，提供了概念或者实体之间的对应关系。截止目前，人们已经提出了各种各样的本体匹配算法，一般可以分为模式匹配（schema matching）和实例匹配（instance matching），也有少量的同时考虑模式和实例的匹配[32-34]。从技术层面来讲，本体匹配可分为启发式方法、概率方法、基于图的方法、基于学习的方法和基于推理的方法。下面围绕模式匹配和实例匹配，具体介绍各自分类中几个具有代表性的匹配方法。

模式匹配主要寻找本体中属性和概念之间的对应关系，文献［35］和［36］给出比较详尽的综述。文献［37］提出一个自动的语义匹配方法，该方法首先利用像 WordNet 之类的词典以及本体的结构等信息进行模式匹配，然后将结果根据加权平均的方法整合起来，再利用一些模式（patterns）进行一致性检查，去除那些导致不一致的对应关系。该过程可循环的，直到不再找到新的对应关系为止。文献［38］也是考虑多种匹配算法的结合，利用基于术语的一些**相似度计算**算法，例如 n-gram 和编辑距离，这里算法计算的结果根据加权求和进行合并，还考虑了概念的层次关系和一些背景知识，最后通过用户定义的权重进行合并。为了应对大规模的本体，文献［39］提出一个使用**锚（anchor）**的系统，该系统以一对来自两个本体的相似概念为起点，根据这些概念的父概念和子概念等邻居信息逐渐地构建小片段，从中找出匹配的概念。新找出的匹配的概念对又可作为新的锚，然后再根据邻居信息构建新的片段。该过程不断地重复，直到未找到新的匹配概念对时停止。文献［40］则以分而治之的思想处理大规模本体，该方法先根据本体的结构对其进行划分获得组块，然后从不同本体获得的组块进行基于锚的匹配，这里的锚是指事先匹配好的实体对，最后再从匹配的组块中找出对应的概念和属性。现有的匹配方法通常是将多个匹配算法相结合，采用加权平均或加权求和的方式进行合并。但是，由于**本体结构**的不对称性等特征，这种固定的加权方法显出不足。文献［41］基于贝叶斯决策的风险最小化提出一个动态的合并方法，该方法可以根据本体的特征，在计算每个实体对的相似度时动态地选择使用哪几个匹配算法，如何合并这些算法，其灵活性带来了很好的匹配结果。

实例匹配是评估异构知识源之间实例对的相似度，用来判断这些实例是否指向给定领域的相同实体。最近几年，随着 Web 2.0 和语义 Web 技术的不断发展，越来越多的语义数据往往具有丰富实例和薄弱模式的特点，促使本体匹配的研究工作慢慢的从模式层转移到实例层[42]。文献［43］提出一个自训练的方法进行实例匹配，该方法首先根据 owl:sameAs、函数型属性（functional properties）和基数（cardinalities）构建一个核（kernel），再根据区别比较明显的属性值对递归的对该核进行扩展。文献［44］利用现有的局部敏感哈希（locality-sensitive hashing）技术来大幅提高实例匹配的可扩展性，该方法首先需要定义用于实例相似性分析的粒度，然后使用分割好的字符串技术实例相似度。文献［45］首先使用向量空间模型表示实例的描述性信息，再基于规则采用倒排索引（inverted indexes）获取最初的匹配候选，在使用用户定义的属性值对候选进行过滤，最后计算出的匹配候选相似度用来作为整合的向量距离，由此抽取出匹配结果。虽然已有方法中已有不少用于处理大规模本体的实例匹配问题，但是同时保证高效和高精度仍然是个很大的挑战。文献［46］提出了一个迭代的框架，充分利用特征明显的已有匹配方法来提高效率，同时基于相似度传播的方法利用一个加权指数函数来确保实例匹配的高精度。

2.4 实体链接技术

歧义性和多样性是自然语言的固有属性，也是实体链接的根本难点。如何挖掘更多、更加有效的消歧证据，设

计更高性能的消歧算法依然是实体链接系统的核心研究问题，值得进一步研究。下面按照不同的实体消歧方法进行分类。

基于概率生成模型方法：韩先培和孙乐 [47] 提出了一种生成概率模型，将候选实体 e 出现在某页面中的概率、特定实体 e 被表示为实体指称项的概率以及实体 e 出现在特定上下文中的概率三者相乘，得到候选实体同实体指称项之间的相似度评分值。Blanco 和 Ottaviano 等人 [48] 提出了用于搜索查询实体链接的概率模型，该方法采用了散列技术与上下文知识，有效地提高了实体链接的效率。

基于主题模型的方法：Zhang 等人 [49] 通过模型自动对文本中的实体指称进行**标注**，生成训练数据集用于训练 LDA 主题模型，然后计算实体指称和候选实体的上下文语义相似度从而消歧得到目标实体。王建勇等人 [50] 提出了对用户的兴趣主题建模的方法，首先构建关系图，图中包含了不同命名实体间的相互依赖关系，然后利用局部信息对关系图中每个命名实体赋予初始兴趣值，最后利用传播算法对不同命名实体的兴趣值进行传播得到最终兴趣值，选择具有最高兴趣值的候选实体。

基于图的方法：Han 等人 [51] 构造了一种基于图的模型，其中图节点为所有实体指称和所有候选实体；图的边分为两类，一类是实体指称和其对应的候选实体之间的边，权重为实体指称和候选实体之间的局部文本相似度，采用词袋模型和余弦距离计算得出。另一类是候选实体之间的边，权重为候选实体之间的语义相关度，采用谷歌距离计算。算法首先采集不同实体的初始置信度，然后通过图中的边对置信度进行传播和增强。Gentile 和 Zhang[52] 等人提出了基于图和语义关系的命名实体消歧方法，该方法在维基百科上建立基于图的模型，然后在该模型上计算各个命名实体的得分从而确定了目标实体，该方法在新闻数据上取得了较高的准确率。Alhelbawy 等人 [53] 也采用基于图的方法，图中的节点为所有的候选实体，边采用两种方式构建，一种是实体之间的维基百科链接，另一种是使用实体在维基百科文章中句子的共现。图中的候选实体节点通过和实体指称的相似度值被赋予初始值，采用 PageRank 选择目标实体。Hoffart 等人 [54] 使用实体的先验概率，实体指称和候选实体的上下文相似度，以及候选实体之间的内聚性构成一个加权图，从中选择出一个候选实体的密集子图作为最可能的目标实体分配给实体指称。

基于深度神经网络的方法：周明和王厚峰等人 [55] 提出了一种用于实体消歧的实体表示训练方法。该方法对文章内容进行自编码，利用深度神经网络模型以有监督的方式训练实体表示，依据语义表示相似度对候选实体进行排序，但该方法是一种局部性方法，没有考虑同一文本中共同出现的实体间相关性。黄洪钊和季姮等人 [56] 基于深度神经网络和语义知识图谱，提出了一种基于图的半监督实体消歧义方法，将深度神经网络模型得到的实体间语义关联度作为图中的边权值。从实验结果得出：基于语义知识图谱的 NGD 和 VSM [57] 方法比起 Wikipedia anchor links 无论在关联性测试上还是在消歧性能上都具有更好的测试结果。相比 NGD 和 VSM ，基于 DNN[58] 的深度语义关联方法在关联性测试上还是在消歧性能上都具有更好的关联性和更高的准确性。但该方法存在两点不足，一方面在构建深度语义关联模型时采用词袋子方法，没有考虑上下文词之间位置关系，另外一方面在消歧的过程中，构建的图模型没有充分利用已消歧实体，边权值和顶点得分随着未消歧实体增加保持不变，并没有为后续的歧义实体增加信息量。

2.5 知识推理技术

知识库推理可以粗略地分为基于符号的推理和基于统计的推理。在人工智能的研究中，基于符号的推理一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。基于符号的推理可以从一个已有的知识图谱，利用规则，推理出新的实体间关系，还可以对知识图谱进行逻辑的冲突检测。基于统计的方法一般指关系机器学习方法，通过统计规律从知识图谱中学习到新的实体间关系。

2.5.1 基于符号逻辑的推理方法

为了使得语义网络同时具备形式化语义和高效推理，一些研究人员提出了易处理（tractable）概念语言，并且开发了一些商用化的语义网络系统。这些系统的提出，使得针对概念描述的一系列逻辑语言，统称**描述逻辑（description logic）**，得到了学术界和业界广泛关注。但是这些系统的推理效率难以满足日益增长的数据的需求，最终没能得到广泛应用。这一困局被利物浦大学的 Ian Horrocks 教授打破，他开发的 FaCT 系统可以处理一个比较大的医疗术语本体 GALEN，而且性能比其他类似的推理机要好得多。描述逻辑最终成为了 W3C 推荐的 Web 本体语言 OWL 的逻辑基础。

虽然描述逻辑推理机的优化取得了很大的进展，但是还是跟不上数据增长的速度，特别是当数据规模大到目前的基于内存的服务器无法处理的情况下。为了应对这一挑战，最近几年，研究人员开始考虑将描述逻辑和 RDFS 的推理**并行**来提升推理的效率和可扩展性，并且取得了很多成果。并行推理工作所借助的并行技术分为以下两类：1）单机环境下的多核、多处理器技术，比如多线程，GPU 技术等；2）多机环境下基于网络通信的分布式技术，比如 MapReduce 计算框架、Peer-To-Peer 网络框架等。很多工作尝试利用这些技术实现高效的并行推理。

单机环境下的并行技术以共享内存模型为特点，侧重于提升本体推理的时间效率。对于实时性要求较高的应用场景，这种方法成为首选。对于表达能力较低的语言，比如 RDFS、 OWL EL，单机环境下的并行技术将显著地提升本体推理效率。Goodman 等人在文献 [59] 中利用高性能计算平台 Cray XMT 实现了大规模的 RDFS 本体推理，利用平台计算资源的优势限制所有推理任务在内存完成。然而对于计算资源有限的平台，内存使用率的优化成为了不可避免的问题。Motik 等人在文献 [60] 工作中将 RDFS，以及表达能力更高的 OWL RL 等价地转换为 Datalog 程序，然后利用 Datalog 中的并行优化技术来解决内存的使用率问题。在文献 [61] 中，作者尝试利用并行与串行的混合方法来提升 OWL RL 的推理效率。Kazakov 等人在文献 [62] 中提出了利用多线程技术实现 OWL EL 分类（classification）的方法，并实现推理机 ELK。

尽管单机环境的推理技术可以满足高推理性能的需求，但是由于计算资源有限（比如内存，存储容量），推理方

法的可伸缩性（scalability）受到不同程度的限制。因此，很多工作利用分布式技术突破大规模数据的处理界限。这种方法利用多机搭建集群来实现本体推理。

Mavin[63] 是首个尝试利用 Peer-To-Peer 的分布式框架实现 **RDF 数据推理**的工作。实验结果表明，利用分布式技术可以完成很多在单机环境下无法完成的大数据量推理任务。很多工作基于 MapReduce 的开源实现（如 Hadoop，Spark 等）设计提出了大规模本体的推理方法。其中较为成功的一个尝试是 Urbani 等人在 2010 年公布的推理系统 WebPIE[64]。实验结果证实其在大集群上可以完成上百亿的 RDF 三元组的推理。他们又在这个基础上研究提出了基于 MapReduce 的 OWL RL 查询算法 [65]。利用 MapReduce 来实现 OWL EL 本体的推理算法在文献 [66] 中提出，实验证明 MapReduce 技术同样可以解决大规模的 OWL EL 本体推理。在文献 [67] 的工作中，进一步扩展 OWL EL 的推理技术，使得推理可以在多个并行计算平台完成。

2.5.2 基于统计的推理方法

知识图谱中基于统计的推理方法一般指**关系机器学习**方法。下面介绍一些典型的方法。

1. 实体关系学习方法

实体关系学习的目的是学习知识图谱中实例和实例之间的关系。这方面工作非常多，也是最近几年知识图谱的一个比较热的研究方向。按照文献 [68] 的分类，可以分为**潜在特征模型**和**图特征模型**两种。潜在特征模型通过实例的潜在特征来解释三元组。比如说，莫言获得诺贝尔文学奖的一个可能解释是他是一个有名的作家。Nickel 等人在文献 [69] 中给出了一个关系潜在特征模型，称为**双线性（bilinear）模型**，该模型考虑了潜在特征的两两交互来学习潜在的实体关系。Drumond 等人在文献 [70] 中应用两两交互的张量分解模型来学习知识图谱中的潜在关系。

翻译（translation）模型 [71] 将实体与关系统一映射至低维向量空间中，且认为关系向量中承载了头实体翻译至尾实体的潜在特征。因此，通过发掘、对比向量空间中存在类似潜在特征的实体向量对，我们可以得到知识图谱中潜在的三元组关系。全息嵌入（Holographic Embedding，HolE）模型 [72] 分别利用圆周相关计算三元组的组合表示及利用圆周卷积从组合表示中恢复出实体及关系的表示。与张量分解模型类似，HolE 可以获得大量的实体交互来学习潜在关系，而且有效减少了训练参数，提高了训练效率。

基于图特征模型的方法从知识图谱中观察到的三元组的边的特征来预测一条可能的边的存在。典型的方法有基于基于归纳逻辑程序（ILP）的方法 [73]，基于关联规则挖掘（ARM）的方法 [74] 和路径排序（path ranking）的方法 [75]。基于 ILP 的方法和基于 ARM 的方法的共同之处在于通过挖掘的方法从知识图谱中抽取一些规则，然后把这些规则应用到知识图谱上，推出新的关系。而路径排序方法则是根据两个实体间连通路径作为特征来判断两个实体是否属于某个关系。

2. 类型推理（type inference）方法

知识图谱上的类型推理目的是学习知识图谱中的实例和概念之间的属于关系。SDType[76] 利用三元组主语或谓语所连接属性的统计分布以预测实例的类型。该方法可以用在任意单数据源的知识图谱，但是无法做到跨数据集的类型推理。Tipalo[77] 与 LHD[78] 均使用 DBpedia 中特有的 abstract 数据，利用特定模式进行实例类型的抽取。此类方法依赖于特定结构的文本数据，无法扩展到其他知识库。

3. 模式归纳（schema induction）方法

模式归纳方法学习概念之间的关系，主要有基于 ILP 的方法和基于 ARM 的方法。ILP 结合了机器学习和逻辑编程技术，使得人们可以从实例和背景知识中获得逻辑结论。Lehmann 等在文献 [79] 中提出用向下精化算子学习描述逻辑的概念定义公理的方法，即从最一般的概念（即顶概念）开始，采用启发式搜索方法使该概念不断特殊化，最终得到概念的定义。为了处理像 DBpedia 这样大规模的语义数据，该方法在文献 [80] 中得到进一步扩展。这些方法都在 DL-Learner[81] 中得以实现。Völker 等人在文献 [82] 中介绍了从知识图谱中生成概念关系的统计方法，该方法通过 SPARQL 查询来获取信息，用以构建事务表。然后使用 ARM 技术从事务表中挖掘出一些相关联的概念关系。在他们后续工作中，使用负关联规则挖掘技术学习不交概念关系 [83]，并在文献 [84] 中给出了丰富的试验结果。

3 开放知识图谱

本节首先介绍当前世界范围内知名的高质量大规模**开放知识图谱**，包括 DBpedia[85][86]、Yago[87][88]、Wikidata[89]、BabelNet[90][91]、ConceptNet[92][93] 以及 Microsoft Concept Graph[94][95] 等。然后介绍中文开放知识图谱平台 OpenKG。

3.1 开放知识图谱

DBpedia 是一个大规模的多语言百科知识图谱，可视为是维基百科的结构化版本。

DBpedia 使用固定的模式对维基百科中的实体信息进行抽取，包括 abstract、infobox、 category 和 page link 等信息。图 2 示例了如何将维基百科中实体 “Busan” 的 infobox 信息转换成 RDF 三元组。DBpedia 目前拥有 127 种语言的超过两千八百万个实体与数亿个 RDF 三元组，并且作为链接数据的核心，与许多其他数据集均存在实体映射关系。而根据抽样评测 [96]，DBpedia 中 RDF 三元组的正确率达 88%。DBpedia 支持数据集的完全下载。

Yago 是一个整合了维基百科与 WordNet[97] 的大规模本体，它首先制定一些固定的规则对维基百科中每个实体的 infobox 进行抽取，然后利用维基百科的 category 进行实体类别推断（Type Inference）获得了大量实体与概念之间的 IsA 关系（如:“Elvis Presley” IsA “American Rock Singers”），最后将维基百科的 category 与 WordNet 中的 Synset（一个 Synset 表示一个概念）进行映射，从而利用了 WordNet 严格定义的 Taxonomy 完成大规模本体的构建。随着时间的推移，Yago 的开发人员为该本体中的 RDF 三元组增加了时间与空间信息，从而完成了 Yago2[98] 的构建，又利用相同的方法对不同语言维基百科的进行抽取，完成了 Yago3[99] 的构建。目前，Yago 拥有 10 种语言

约 459 万个实体，2400 万个 Facts，Yago 中 Facts 的正确率约为 95%。Yago 支持数据集的完全下载。

WikiText syntax

```
{{Infobox Korean settlement
|title = Busan Metropolitan City
...
|area_km2 = 763.46
|pop = 3635389
|region = [[Yeongnam]]
}}
```

RDF serialization

```
dbp:Busan dbp:title "Busan Metropolitan City"
dbp:Busan dbp:area_km2 "763.46"^xsd:float
dbp:Busan dbp:pop "3635389"^xsd:int
dbp:Busan dbp:region dbp:Yeongnam
```

图 2 RDF 三元组

Wikidata 是一个可以自由协作编辑的多语言百科知识库，它由维基媒体基金会发起，期望将维基百科、维基文库、维基导游等项目中**结构化知识**进行抽取、存储、关联。Wikidata 中的每个实体存在多个不同语言的标签，别名，描述，以及声明（statement），比如 Wikidata 会给出实体 “London” 的中文标签 “ 伦敦 ”，中文描述 “ 英国首都 ” 以及图 3 给出了一个关于 “London” 的声明的具体例子。“London” 的一个声明由一个 claim 与一个 reference 组成，claim 包括 property:“Population”、value:“8173900” 以及一些 qualifiers（备注说明）组成，而 reference 则表示一个 claim 的出处，可以为空值。目前 Wikidata 目前支持超过 350 种语言，拥有近 2500 万个实体及超过 7000 万的声明 [100]，并且目前 Freebase 正在往 Wikidata 上进行迁移以进一步支持 Google 的语义搜索。Wikidata 支持数据集的完全下载。

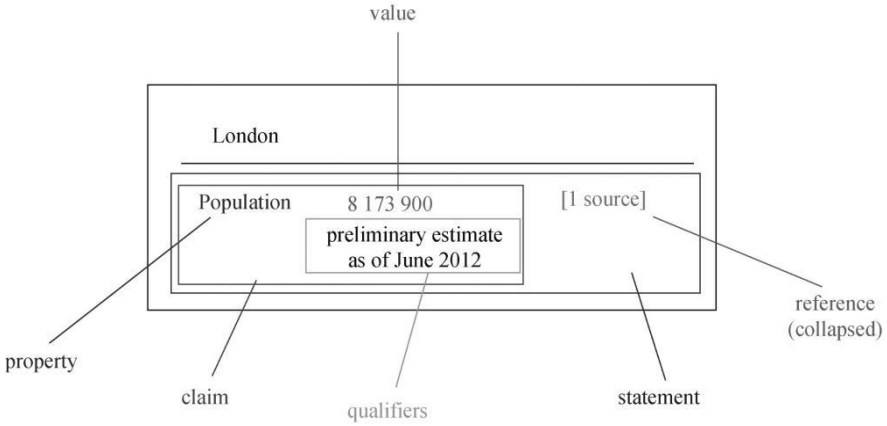

图 3 “London”声明示例

BabelNet 是目前世界范围内最大的多语言百科同义词典，它本身可被视为一个由概念、实体、关系构成的语义网络（Semantic Network）。BabelNet 目前有超过 1400 万个词目，每个词目对应一个 synset。每个 synset 包含所有表达相同含义的不同语言的**同义词**。比如:“ 中国 ”、“ 中华人民共和国 ”、“China” 以及“People’s Republic of China” 均存在于一个 synset 中。BabelNet 由 WordNet 中的英文 synsets 与维基百科页面进行映射，再利用维基百科中的跨语言页面链接以及翻译系统，从而得到 BabelNet 的初始版本。目前 BabelNet 又整合了 Wikidata、 GeoNames、OmegaWiki 等多种资源，共拥有 271 个语言版本。由于 BabelNet 中的错误来源主要在于维基百科与 WordNet 之间的映射，而映射目前的正确率大约在 91%。关于数据集的使用， BabelNet 目前支持 HTTP API 调用，而数据集的完全下载需要经过非商用的认证后才能完成。

ConceptNet 是一个大规模的多语言常识知识库，其本质为一个以自然语言的方式描述人类常识的大型语义网络。ConceptNet 起源于一个众包项目 Open Mind Common Sense，自 1999 年开始通过文本抽取、众包、融合现有知识库中的常识知识以及设计一些游戏从而不断获取常识知识。ConceptNet 中共拥有 36 种固定的关系，如 IsA、UsedFor、CapableOf 等，图 4 给出了一个具体的例子，从中可以更加清晰地了解 ConceptNet 的结构。ConceptNet 目前拥有 304 个语言的版本，共有超过 390 万个概念，2800 万个声明（statements，即语义网络中边的数量），正确率约为 81%。另外，ConceptNet 目前支持数据集的完全下载。

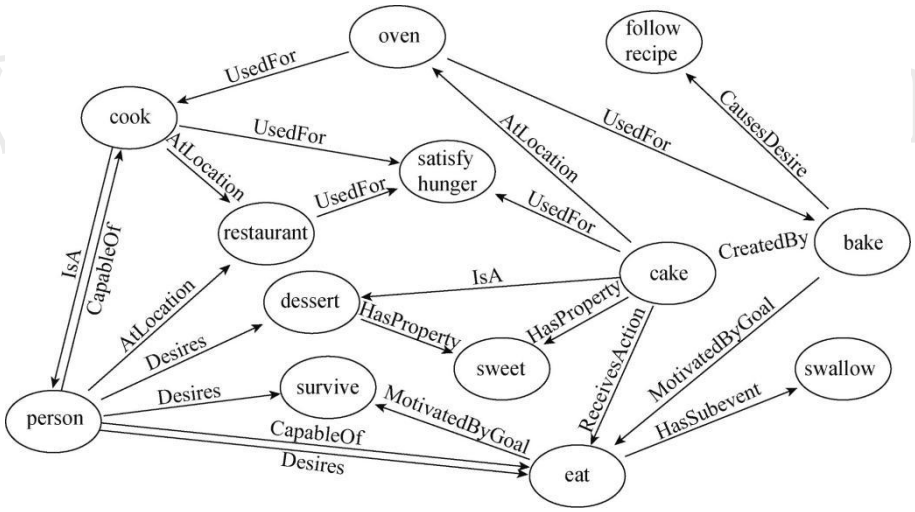

图 4 Concept Net 示例

Microsoft Concept Graph 是一个大规模的英文 Taxonomy，其中主要包含的是概念间以及实例（等同于上文中的实体）概念间的 IsA 关系，其中并不区分 instanceOf 与 subclassOf 关系。 Microsoft Concept Graph 的前身是 Probase，它过自动化地抽取自数十亿网页与搜索引擎查询记录，其中每一个 IsA 关系均附带一个概率值，即该知识库中的每个 IsA 关系不是绝对的，而是存在一个成立的概率值以支持各种应用，如短文本理解、基于 taxonomy 的关键词搜索和万维网表格理解等。目前，Microsoft Concept Graph 拥有约 530 万个概念，1250 万个实例以及 8500 万个 IsA 关系（正确率约为 92.8%）。关于数据集的使用，Microsoft Concept Graph 目前支持 HTTP API 调用，而数据集的完全下载需要经过非商用的认证后才能完成。

除了上述知识图谱外，中文目前可用大规模**开放知识图谱**有 **Zhishi.me**[101]、**Zhishi. schema**[102] 与 **XLore**[103] 等。Zhishi.me 是第一份构建中文链接数据的工作，与 DBpedia 类似， Zhishi.me 首先指定固定的抽取规则对百度百科、互动百科和中文维基百科中的实体信息进行抽取，包括 abstract、infobox、category 等信息；然后对源自不同百科的实体进行对齐，从而完成数据集的链接。目前 Zhishi.me 中拥有约 1000 万个实体与一亿两千万个 RDF 三元组，所有数据可以通过在线 SPARQL Endpoint 查询得到。Zhishi.schema 是一个大规模的中文模式（Schema）知识库，其本质是一个语义网络，其中包含三种概念间的关系，即 equal、related 与 subClassOf 关系。Zhishi.schema 抽取自社交站点的分类目录（Category Taxonomy）及标签云（Tag Cloud），目前拥有约 40 万的中文概念与 150 万 RDF 三元组，正确率约为 84%，并支持数据集的完全下载。XLore 是一个大型的中英文知识图谱，它旨在从各种不同的中英文在线百科中抽取 RDF 三元组，并建立中英文实体间的跨语言链接。目前，XLore 大约有 66 万个概念，5 万个属性，1000 万的实体，所有数据可以通过在线 SPARQL Endpoint 查询得到。

3.2 中文开放知识图谱联盟介绍

中文开放知识图谱联盟 OpenKG 旨在推动**中文知识图谱**的开放与互联，推动知识图谱技术在中国的普及与应用，为中国人工智能的发展以及创新创业做出贡献。联盟已经搭建有 OpenKG.CN 技术平台（图 5），目前已有 35 家机构入驻。吸引了国内最著名知识图谱资源的加入，如 Zhishi.me， CN-DBPedia, PKUBase。并已经包含了来自于常识、医疗、金融、城市、出行等 15 个类目的开放知识图谱。

图 5 OpenKG

4 知识图谱在情报分析的案例

4.1 股票投研情报分析

通过知识图谱相关技术从招股书、年报、公司公告、券商研究报告、新闻等半结构化表格和非结构化文本数据中批量自动抽取公司的股东、子公司、供应商、客户、合作伙伴、竞争对手等信息，构建出公司的知识图谱。在某个宏观经济事件或者企业相关事件发生的时候，券商分析师、交易员、基金公司基金经理等投资研究人员可以通过

此图谱做更深层次的分析和更好的投资决策，比如在美国限制向中兴通讯出口的消息发布之后，如果我们有中兴通讯的客户供应商、合作伙伴以及竞争对手的关系图谱，就能在中兴通讯停牌的情况下快速地筛选出受影响的国际国内上市公司从而挖掘投资机会或者进行投资组合风险控制（图 6）。

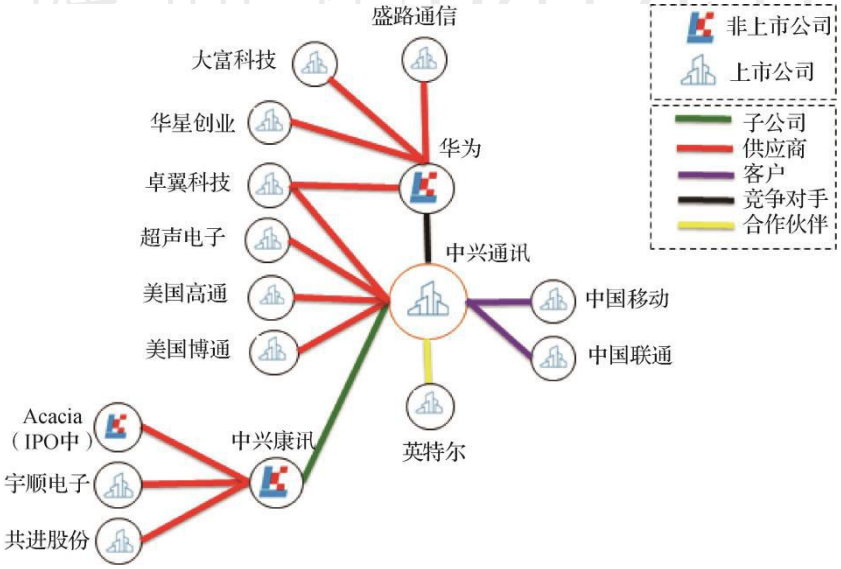

图 6　中兴通讯关系图谱

4.2　公安情报分析

通过融合企业和个人银行资金交易明细、通话、出行、住宿、工商、税务等信息构建初步的 “ 资金账户-人-公司”关联知识图谱。同时从案件描述、笔录等非结构化文本中抽取人（受害人、嫌疑人、报案人）、事、物、组织、卡号、时间、地点等信息，链接并补充到原有的知识图谱中形成一个完整的证据链。辅助公安刑侦、经侦、银行进行案件线索侦查和挖掘同伙。比如银行和公安经侦监控资金账户，当有一段时间内有大量资金流动并集中到某个账户的时候很可能是非法集资，系统触发预警（图 7）。

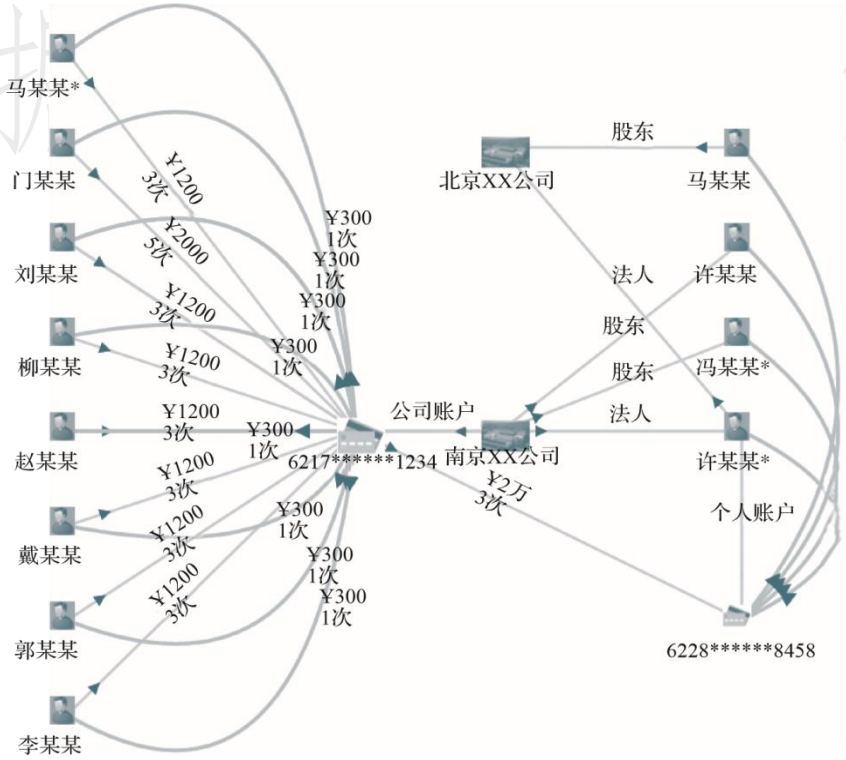

图7 公安情报分析示例

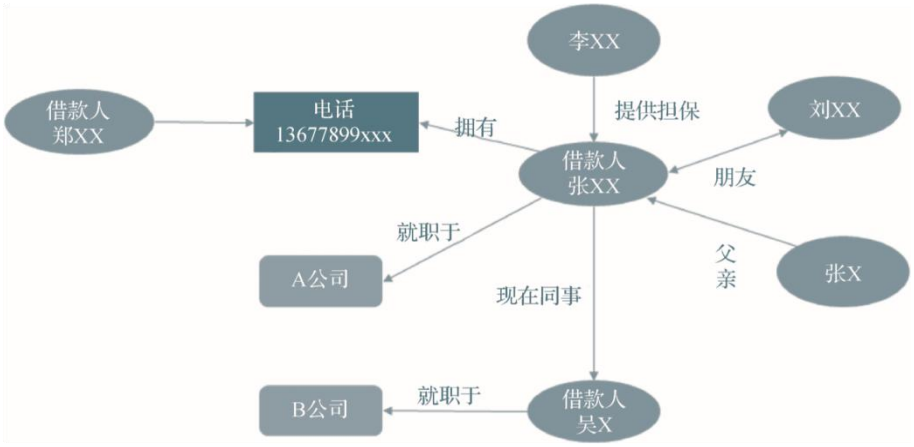

图8 反欺诈情报分析示例

4.3 反欺诈情报分析

通过融合来自不同数据源的信息构成知识图谱，同时引入领域专家建立业务专家规则。我们通过数据不一致性检测，利用绘制出的知识图谱可以识别潜在的欺诈风险。比如借款人张 xx 和借款人吴 x 填写信息为同事，但是两个人填写的公司名却不一样，以及同一个电话号码属于两个借款人，这些不一致性很可能有欺诈行为。（图 8）

5 总结

知识图谱是**知识工程**的一个分支，以知识工程中**语义网络**作为理论基础，并且结合了**机器学习**，**自然语言处理**和**知识表示和推理**的最新成果，在**大数据**的推动下受到了业界和学术界的广泛关注。

知识图谱对于解决大数据中**文本分析**和**图像理解**问题发挥重要作用。目前，知识图谱研究已经取得了很多成果，形成了一些开放的知识图谱。但是，知识图谱的发展还存在以下障碍。首先，虽然大数据时代已经产生了海量的数据，但是数据发布缺乏规范，而且数据质量不高，从这些数据中挖掘高质量的知识需要处理数据噪音问题。其次，垂直领域的知识图谱构建缺乏自然语言处理方面的资源，特别是词典的匮乏使得垂直领域知识图谱构建代价很大。最后，知识图谱构建缺乏开源的工具，目前很多研究工作都不具备实用性，而且很少有工具发布。通用的知识图谱构建平台还很难实现。

参考文献

[1]sowa J F. Principles of semantic networks: Exploration in the representation of Knowledge[J]. Frame Problem in Artificial Intelligence, 1991(2-3):135 - 157.

[2]simmons r F. technologies for Machine Translation[J]. Future Generation Computer Systems, 1986, 2(2):83-94.

[3]simmons r F. natural language Questionanswering systems: 1969[J]. communications of the ACM, 1970, 13(1):15-30.

[4-103] 余略。引文：漆桂林,高桓,吴天星.知识图谱研究进展[J].情报工程,2017,3(1):004-025。**本文原网址** http://tie.istic.ac.cn/ch/reader/view_abstract.aspx?file_no=201701002。

[x] 秦陇纪. 数据科学与大数据技术专业概论；人工智能研究现状及教育应用；文献数据共词分析的神经网络训练；大数据简化之技术体系[EB/OL]. 数据简化 DataSimp（微信公众号）,2017-08-23.

Appx. 新闻五则(5670 字)

附 i. 早报,4 月 21 日,星期五

1、中国首艘货运飞船**天舟一号**顺利发射升空，它将与天宫二号对接并送去 6 吨重的补给 澳专家:美国都没做到；
2、河北“政法王”张越被控**受贿 1.57 亿** 哽咽认罪；
3、韩媒称 99 家乐天玛特门店中，**87 家在华乐天超市停业** 营业损失已高达 2000 亿韩元(约合人民币 12 亿元)；
4、**17 万平米污水渗坑**追责：大城副县长、环保局长被停职；
5、食药监总局:“浏阳河”酒氰化物超标 33.5 倍；
6、俄轰炸机 24 小时内两次逼近阿拉斯加 距美领土 57 公里；
7、山西破获特大跨省贩毒案 缴获冰毒海洛因 16 公斤 头目曾是国企副总 ；
8、黑龙江严查非法转基因种子:进村排查不让一粒下地；
9、北京：下月起购二手房须以本人银行卡转账付款;；
10、韩美双方已根据《驻韩美军地位协定》走完“萨德”落地所需的供地程序，美军得以正式准备部署“萨德”反导系统；
11、Facebook：增强现实（AR）技术就将取得突破 5 年内超级眼镜将取代智能手机；
12、最高法规范国家赔偿监督程序：符合条件须 7 日内立案；
【心语】做一个努力的人好处在于，人人见了你都会想帮你。但如果你自己不先做出一点努力的样子，人家想拉你一把，都不知你的手在哪里。

附 ii. 2017 年 4 月 21 日周五读报！一切美好从“勇往直前”开始！

1、骄傲！天舟一号飞船发射成功！**天舟一号**七大独门功夫：①首次执行货运飞船飞行试验，最大可运载 6.5 吨；②首次为空间站“太空加油”；③首次将测量系统“搬”到天上；④首次大规模推动核心元器件自主可控；⑤首次开展全自主快速交会对接试验；⑥首次搭载多项空间应用与技术试验载荷；⑦首次实施主动离轨受控陨落。

2、国务院：推出进一步减税措施，**增值税**税率由四档减至 17%、11%与 6%三档，预计全年将再减轻各类市场主体税负 **3800 多亿元**人民币，持续推动实体经济降成本增后劲。

3、科技部：到 2020 年，我国**研究与发展(R&D)人员**年人均研发经费由 2014 年的 37 万元/年提升到 2020 年的 **50 万元/年**，与发达国家之间的差距进一步缩小。

4、截至 2016 年 12 月，中国**网络视频用户**规模已达 5.45 亿，网络视频用户使用率为 74.5%，会员付费收入表现出强劲的增长趋势，广告收入仍占据收入主体份额。近 40%用户遇到广告时会先在网上看点其他内容。35%的网络视频用户有付费看视频经历。

5、看完你还敢喝吗？国家食药监总局通告：标称长沙市浏阳河酒业有限公司生产的“浏阳河红色经典”酒，氰化物检出值比标准规定高出 33.5 倍。

6、7 月 1 日起，商业健康保险个人所得税税前扣除试点政策推至全国。标志着此前仅在试点城市推出的个人税优健康险产品将向全国开放。

7、联合国旗下机构“Better Than Cash Alliance”(优于现金联盟)发布报告称，在支付宝和微信支付的推动下，2016 年中国社交网络支付(支付宝和微信)市场规模达到了 2.9 万亿美元，在过去 4 年中增长了 20 倍。

8、诡异的平静：沪指连续 85 日“跌不破 1%” 为 1992 年以来最长记录。分析认为，市场稳定可能和国家队有关，也可能与全球市场趋势相关。大盘平静的背后，个股表现并不太平，今年以来近七成个股下跌，近期个股“闪崩”还在延续。

9、矛盾升级！易到用车三个联合创始人周航、杨芸、汤鹏发布联合声明，正式辞去易到所有相关职务。联合声明称发布不到 1 小时后，就有消息称，易到本计划开董事会商议开除周航，只是“又被抢先了”。

10、世界军力谁家强？德国《焦点》周刊报道，“全球火力指数”借助兵力、军备等 40 多项标准对世界各国军队进行比较(核武器未计)。在接受调查的 126 个国家中，前 10 名依次为美国、俄罗斯、中国、印度、法国、英国、日本、土耳其、德国、意大利。朝鲜在装甲车方面排第 7，仅次于叙利亚。

11、美国军车开进萨德部署地，当地居民誓死阻拦并与警察发生冲突，造成 2 名居民跌倒受伤，其中 1 人被当场送往医院。僵持发生 40 多分钟后，军车在警察引导下进入星州高尔夫球场。据悉，本次运输的重型机械是兼具推土机和挖掘机复合功能的作业设备。

12、吃过的苦、抗过的压，不是毫无意义的。很多时候，正是它们的磨砺，让我们变得更加从容美好。所有你觉得永远不会过去的，最后都会过去，打不倒我们的，都必将使我们更强大。哪怕路途遥远，也请你披盔戴甲、勇往直前。

美好一天从“勇往直前”开始！

附 iii. 2017 年 4 月 21 日（丁酉鸡年三月二十五）周五 / 早读分享：

1、【李克强：要把医疗健康产业做成我国支柱产业】19 日，李总理在考察威海威高集团时，表示：我国已进入**中等收入国家**行列，群众对医疗健康的需求日益增长。但该产业目前占我国产业比重不足 5%，与发达国家存在不小差距，市场潜力巨大。你们的事业事关人民群众的健康幸福，功德无量。要把这一产业做成我国支柱产业。

2、【北京大学廉政建设会议 公布涉＂郭文贵案＂人员】中纪委网站消息，共有 1700 多人参加的北大党风廉政建设警示大会上，共通报 13 起典型案例，其中：副校长兼总务长王仰麟、副校长兼教务长高松等受到党内处分。会议还公布了**北大方正**董事长魏新、CEO 李友、总裁余丽等人的案情。

3、【中国**残疾人福利基金会**去年收入超 5 亿】该会 2016 年度总收入为 58822．55 万元，完成全年预算 141．74%，其中捐赠收入 55081．71 万元。年度总支出为 61112．41 万元，完成全年预算的 155．15%，其中慈善活动支出 60090．20 万元。

4、此条发不出！

5、【我国医保参保超过 13 亿人】人社部副部长游钧 19 日表示，我国以更加公平、更可持续为目标，深入推进社会保障制度改革，

加快实施全民参保计划，社会保障覆盖范围持续扩大，其中养老保险参保人数达到 8．9 亿人，医疗保险参保人数超过 13 亿人，基本实现了全民覆盖。

6、【中国已与 46 个国家和地区学历学位互认】其中，＂一带一路＂国家 24 个。截止目前，已经审批的中外合作办学共有 2539 个。其中，本科以上层次项目和机构 1248 个。中国高校已在境外举办了 4 个机构和 98 个办学项目，分布在 14 个国家和地区。

7、【保监会：要把业务扩张激进的异常机构作为监管重点】当前金融工作面临的国内外形势依然错综复杂。当前金融工作的不稳定不确定性因素并未减少。

8、【国际货币基金组织警告英国退欧可能破坏全球金融稳定】IMF 在周三发布的一份报告中说：＂英国退欧带来的挑战可能削弱金融稳定，其影响方式在当前难以估计或预测＂。

9、【昨日股市收评：沪深两市终结四连阴 白酒板块成涨领头羊】沪深两市早盘低开，随后震荡整理横盘至下午开盘半小时。受雄安概念集体大跌影响，沪指及创业板出现小幅跳水。临近尾声，指数在保险和券商带领下直线反弹翻红，两市股指成功收复失地，K 线四连阴终结束，涨 0．04%。

10、【先言潮声】现在过的每一天，都是余生最年轻的一天。请不要老得快快，却明白得太迟！

美好的一天从珍惜大好的时光开始！

附 iv. 2017 年 4 月 21 日（星期五）农历丁酉年三月廿五　丁酉年 甲辰月 戊寅日

每天三分钟 知晓天下事

A、【国内】

【置顶】中国首艘**货运飞船天舟一号**发射成功，运送超 6 吨货物首飞太空；

1）习近平北海考察论古今：写好丝路新篇章，“向海经济”“向海之路”，是习近平考察调研中反复提及的词汇；李克强 19 日考察山东威海港：既要扩大出口，又要加大进口；习近平在**贵州全票当选**中共十九大代表；

2）中国公布第一批增补**藏南地区公开使用地名** 6 个，如印方反应强烈，预计中方还会公布第二批、第三批命名；

3）[法制与反腐] 司法部：及时了解农民工下岗就业、上访等情况；广东环保厅原厅长李清（正厅级）涉嫌受贿被公诉；

4）新华网：加强市场监管是为了更好地保护投资者利益；

5）国产智能手机华为、OPPO、vivo 形成三足鼎立，三星苹果市场份额缩水；

6）河南首个涉毒举报奖励办法出台，最高奖励 30 万元；

7）全球 18 家顶级电信运营商拟两年内采用“未来网络”，将具备大带宽、大连接、高可靠、低时延的特征；

8）[军事]《新闻联播》画面释放重要信息：辽宁舰原舰长张峥将有重任；歼-18 中国战术轰炸机一技术取得已重大突破；海警 3062 船在广州撞货船沉没，曾在黄岩岛执法；

9）[港澳] 梁振英率团考察粤港澳大湾区城市群；莫言赴港畅谈文学创作；

10）[台湾] 第八届海峡两岸文创展台北开幕；近千民众聚集抗议台当局“年金改革”；台湾专业踢踏舞团献艺天津演绎中国魂。

B、【国际】

1）中国常驻联合国代表刘结一代表金砖国家在联大“可持续发展目标筹资问题高级别讨论会”上做共同发言；中国科学家姚檀栋获“地理学诺贝尔奖”瑞典国王颁奖；

2）朝鲜声称正准备再次核试，中方回应：坚决反对任何加剧对立紧张的言行；

3）英首相宣布将于 6 月 8 日提前大选，或有望扫平脱欧“绊脚石”；

4）韩国“亲信门”事件核心人物崔顺实之女拒不回国，已起诉丹麦检方；

5）美国军车开进萨德部署地，韩国居民誓死阻拦致 2 伤；

6）我国新发现一颗“紫金山彗星”，绕日周期长达 114 天；哈工大学生团队自主研制的“紫丁香一号”微纳卫星装载在“天鹅座”货运飞船进入太空；

7）印度一载 56 人大巴在北部山区西姆拉地区倾覆坠河严重损毁，至少 44 人死亡；

8）澳大利亚政府宣布提高外国人入籍“门槛”，申请者需接受更加严格的英语考试，并被要求遵循澳大利亚的价值观。

C、【财经证券】

1）**3800 亿**：李克强承诺的第一个减税大礼包来了；国务院：若失业率大幅攀升，加大财政货币政策调整力度；第二批 PPP 资产证券化项目即将推出，环保项目占半数；

2）国研中心专家：我国已出现明显的货币超发现象；

3）民生银行“30 亿”风险事件发酵，有民生银行的员工及家属也购买了这份理财；

4）上交所官员顾斌公开表示：A 股将发新股 500 只，融资 3000 亿；白糖期权昨在郑商所上市，完善产业风险管理体系；

5）昨日收盘：沪指 3172.10/+0.04% 深成指 10359.09/+0.10% 创业板 1850.39/+0.27% 恒指 24056.98/+0.97%。

D、【文教体娱】

1）**795 万大学生毕业**在即，陕西等多省出台扶持政策；北京拟为大学生村官涨薪，江苏基层就业或有考研加分；南邮党委书记刘陈：坚定不移走大信息的特色发展之路；

2）越剧表演艺术大师徐玉兰在华东医院去世，享年 96 岁，她的代表作有《北地王》《西厢记》《春香传》《红楼梦》《追鱼》《西园记》等；

3）世界俱乐部排名：恒大仍排亚洲前 4，上港排名下降 1 位位列亚洲第 20 位；

4）新一届中国女排如今已经重新集结，教练组也组建完毕，丁霞接班惠若琪成女排新队魂！

5）佛山大沥镇北村水闸附近一工地发现至少 5 枚恐龙蛋，距今约 7000 万年。

E、【生活服务】

1）北京严打炒地炒房行为，产业项目严禁擅自改居住；北京景山公园可赏百岁黑牡丹；

2）上海迪士尼进入最美赏花季，拥有种类繁多的植物；

3）广东将于 2017 年-2019 年 3 年投 20 亿建万间规范化村卫生站；

4）南沙居民身份证网上应用 APP“微警认证”发布，有效证明“我就是我”；

5）山东今年将在济南、青岛、烟台、潍坊、威海 5 个市开展既有多层住宅加装电梯试点工作。

F、【健康养生】

1）将手握紧 30 秒钟，打开后手掌变白的现象是会马上消失，还是会保持一段时间？可以测出血管好不好；动脉硬化通常在青少年时期发生，至中老年时加重、发病，在手掌上有可能表现为血管弹性不好、血液循 环不畅。

2）瑞士专家最近证实，精神压力（坏心情）可引起血管内膜收缩，加速血管老化；昼夜颠倒，心脑血管的生物钟会被打乱，让血

管收缩、血液流动缓慢、黏稠度增加。

（编辑：西安知非 自新华、中新、腾讯、凤凰、东方财富网）

附 v. 2017 年 4 月 21 日农历干支、节日、名人和事件

导语：2017 年 4 月 21 日是怒族传统节日**仙女节**，因为是当地杜鹃花盛开的季节，所以仙女节又称**鲜花节**。那么 2017 年 4 月 21 日是的节日、星座、出生名人呢？订阅秦陇纪 10 公众号，关注哦。

2017 年 4 月 21 日节日

4 月 21 日是公历一年中的第 111 天(闰年第 112 天)，离全年的结束还有 254 天。

2017 年 4 月 21 日节日：仙女节、鸡蛋会、全国企业家活动日、罗马城纪念日。

仙女节

仙女节是云南省贡山一带怒族人民的民间传统节日。当地又称鲜花节，在每年农历三月十五日举行。届时，以各村寨为单位选择有钟乳石的山洞为仙女洞，人们纷纷带上祭祀用品前去祭祀。这时候，也是当地杜鹃花盛开的季节，人们还要为"仙女"献上一束束杜鹃花。

2017 年 4 月 21 日农历

公历：2017 年 4 月 21 日 星期五

农历：二〇一七年 三月小 廿五日

回历：1438 年 7 月 24 日

干支：丁酉年 甲辰月 戊寅日

八字：丁酉 甲辰 戊寅 壬子

五行：山下火 佛灯火 城头土 桑柘木

生肖：属鸡

星座：金牛座

星宿：壁宿(壁水貐)

值神：司命(黄道日)

冲煞：虎日冲(壬申)猴 煞北

2017 年 4 月 21 日星座

白羊座(3.21-4.19)

4 月 21 日性格：4 月 21 日出生的人是高贵有品格的人，对他们来说，卓越而且完整的专业技术比什么都重要。

他们经常走在时代的前端，成为创造流行的人。不过话说回来，这天出生的人的私生活却可能常处于风暴中，婚姻纪录不止一次，外遇更是家常便饭。

4 月 21 日出生名人

1713 年——路易·德·诺阿耶，第三代诺阿耶公爵阿德里安·毛瑞斯·德·诺阿耶之子。

1729 年——凯瑟琳大帝(叶卡捷琳娜二世)，俄罗斯帝国的女沙皇。

1782 年——**福禄培尔**诞生，德国教育家。

1816 年——夏洛特·勃朗特诞生，英国小说家。代表作《简·爱》。(逝于 1855 年)。

1828 年——依波利特·阿道尔夫·丹纳，法国评论家与史学家。

1871 年——**杨昌济**，中国教育家。

1882 年——**珀西·威廉斯·布里奇曼**，美国物理学家，1946 年诺贝尔物理学奖得主。

1917 年——吴祖光，著名剧作家、导演。

1926 年——英国女王伊丽莎白二世出生。

1953 年——胡因梦，台湾女演员。

1958 年——臼井仪人，日本漫画家

1962 年——狄莺，台湾女演员。

1967 年——刘志升，台湾棒球选手。

1973 年——小西克幸，日本动画声优。

1978 年——金巧巧，中国著名女演员。

1979 年——詹姆斯·麦卡沃伊，英国男演员。

1980 年——下野紘，日本动画声优。

1980 年——星野桂，日本漫画家。

1981 年——朱亚文，内地男演员。

1981 年——BcAZA·SndikAsH 世界知名 DJ。

1985 年——叶咏捷，台湾棒球选手，效力中华职棒兄弟象队。

1991 年——弗兰克·蒂尔内(frank·dillane),少年伏地魔扮演者.

历史上今天大事件

1899 年——**美西战争爆发** 。

1930 年——五大国签署海军条约 。

1949 年——毛泽东、朱德发布《向全国进军命令》，中国人民解放军**百万雄师开始横渡长江**。

1960 年——郑州黄河大桥建成通车。

1985 年——北京万人马拉松赛举行。

1986 年——长江科学考察漂流探险队在成都成立 。

1994 年——中国政府宣布**全面禁止传销**。

2008 年——服役达 27 年的世界上第一种可正式作战的隐身战斗机美国 F—117A 隐身战斗机在执行完最后一次任务后退出现役。

2009 年——"世界数字图书馆"开放 。

2010 年——青海玉树地震全国哀悼日

历史上的今天**马克·吐温逝世**，他的作品对美国的文坛产生了深渊的影响。

Appx. 数据简化 DataSimp 社区（300 字）

附 vi. 数据简化 DataSimp 社区译文志愿者招募启事

“数据简化 DataSimp”社区翻译组、媒体组缺少志愿者，当下需要：①设计黑白静态和三彩色动态社区 LOGO 图标；②翻译美欧 IT 大数据、人工智能、编程开发技术文章的至少投一篇高质量首译美欧数据科学技术论文，正式成为数据简化 DataSimp 社区贡献者。非诚勿扰，季度无贡献者自动退出。加入**数据简化 DataSimp 社区**，请在公号后台留言，或加 QinlongGEcai（请备注：姓名-单位-职务-手机号）微信。社区筹备详情，请阅读本公号文章《科研江湖中的一眼清泉之数据简化 DataSimp 社区及学会》。

Data Simplification and Sciences Wechat and Toutiao Public Account, QinDragon2010@qq.com, 2017.04.21Fri, Xi’an, Shaanxi, China:

LIFE

Life begins at the end of your comfort zone.

-- Neale Donald Walsch

THE DAY

The strength of purpose and the clarity of your vision, along with the tenacity to pursue it, is your underlying driver of success.

-- Ragy Tomas

长按下面二维码“识别图中二维码”关注公众号：数据简化 DataSimp（搜索此名称也行）。

文末**打赏后“阅读原文”可百度网盘下载**全部重点彩标 PDF 完整版文档。

（西安秦陇纪 10 数据简化 DataSimp 综合汇编，欢迎有志于**数据简化**之传媒、技术的实力伙伴加入全球“数据简化 DataSimp”社区！欢迎转载注明出处：**秦陇纪 10 数据简化 DataSimp 公众号、头条号“数据简化 DataSimp**、科学 Sciences”汇译编，投稿邮箱 QinDragon2010@qq.com）