[1. 数据驱动的关系网络构建方法 1](#_Toc17104)

[1.1 传统统计学的相关分析方法 1](#_Toc5572)

[1.2面向大数据的相关分析方法 4](#_Toc32210)

[1.3数据挖掘的方法 5](#_Toc25179)

[1.4基于因素与数据源关系的网络构建方法 9](#_Toc24657)

[2. 知识驱动的关系网络构建方法 12](#_Toc7850)

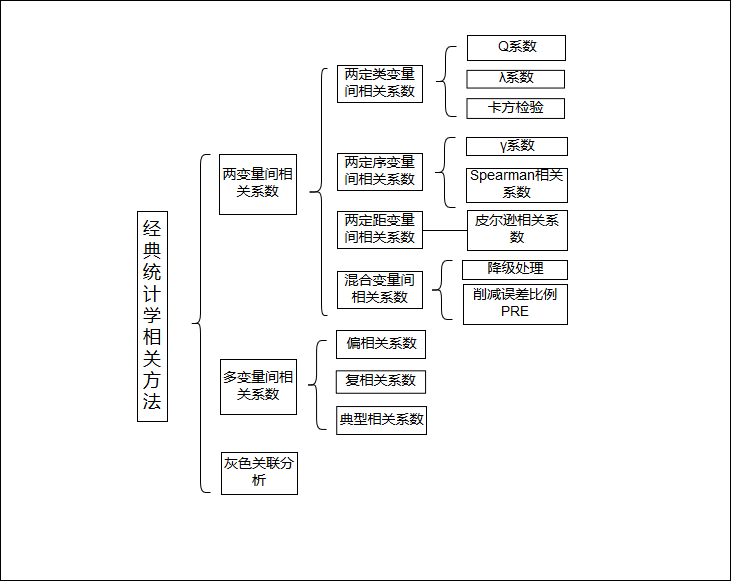
[2.1 基于知识图谱的关系网络构建方法 12](#_Toc18978)

## 数据驱动的关系网络构建方法

数据驱动的关系网络构建方法从数据的视角挖掘资源能源安全风险因素之间的关系。通过因素变量的历史数据来自动挖掘出可能蕴含在数据中的规律，揭示因素间的关系和变化原因。数据驱动的方法可以分为传统的统计学方法，面向大数据的相关分析方法，数据挖掘的方法，以及基于因素与数据源间关系的推测方法。

### 传统统计学的相关分析方法

相关分析在经典的统计研究中已经被广泛讨论，针对不同类型的变量和随机向量形成了种类繁多的相关系数。统计学的相关分析方法针对变量的数值序列进行分析，通过考虑其服从的统计模型的方法来衡量变量间的相关性。具体而言，它通过定义相关系数的计算方法给出相关性的定量度量，是一种典型的数据驱动的分析方法。统计学相关分析方法可以分为两变量间相关系数、多变量间相关系数，同时，灰色关联分析作为一种使用较为广泛的模型是对经典的方法的很好补充。



1. 两变量间相关系数

* 两定类变量间相关系数

定类变量又称为名义变量，这种变量的值代表一个类别，没有大小、次序之分，比如性别这个变量的取值为男、女这样的类别。计算两定类变量间相关系数通常需要使用列联表，也称交互作用表。其典型的相关系数为Q系数（只能计算2\*2的列联表），λ系数（可以计算任意两个定类变量将相关性），以及卡方检验等。其中最常用的是λ系数法，其优点是计算简单，前提宽松，且具有削减误差比例意义。

* 两定序变量间相关系数

定序变量也称等级变量，其取值有序的意义，即其取值区分等级和序列。例如高校教师职称分为助教、讲师、副教授、教授四个等级。两定序变量间相关性定义常用的概念包括同序对、异序对。常用系数有γ系数、Spearman系数（统计上可看为皮尔逊相关系数的特例），这两个系数本身或其变形具有削减误差比例意义，因此较为常用。其他针对定序变量的相关系数还有dy系数，肯德尔系数等。

* 两定距变量间相关系数

定距变量即数值变量，其取值有数值上的差别，可以进行加减乘除运算。度量其相关性的常见方法是皮尔逊相关系数法，其主要思想是将两向量中心化后计算向量夹角的余弦值。该方法在几何、代数、三角等角度上都有良好的数学解释，且其平方具有削减误差比例意义，因此被广泛使用。也有学者针对其容易受到大样本、异常点影响的缺点进行改进。

* 混合变量间相关系数

主要采用两种方法。其一是直接采用削减误差比例法，其主要思想是如果已知X对预测Y有很大的帮助，则X与Y相关性较大。通过计算已知X预测Y产生的误差和不知道X预测Y产生的误差来进行判断。其二是采用降级处理，其主要思路是把其中一方的变量类型降级，如计算定序和定类变量相关性可将定序变量降为定类变量处理，其缺点为信息损失。

1. 多变量间相关系数

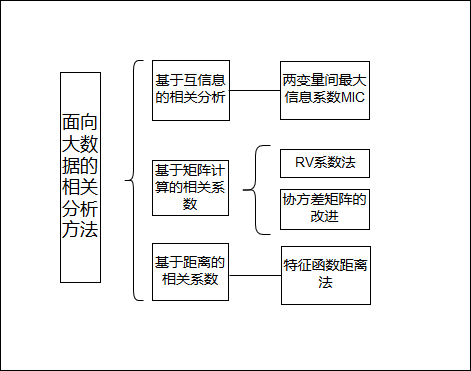
常见的多变量间相关系数有偏相关系数、复相关系数和典型相关系数，且都是两定距变量间线性相关系数（皮尔逊系数）的推广。偏相关系数衡量的是多个变量中固定其他变量时任意两个变量的相关性。复相关系数衡量一个变量对多个变量的相关性。典型关系数衡量多个变量对多个变量的相关性，是一种较为通用的多元统计方法。典型相关分析的主要思路是寻找两个投影方向，使得中心化后的两个变量在投影后皮尔逊相关系数最大。即寻找两个线性变换，将原始数据压缩到一个低维子空间，使得压缩后的数据在子空间线性相关性最大。典型相关分析在电子通信、生物信息、经济管理等领域都有广泛应用。

1. 灰色关联分析

灰色系统理论以部分信息已知部分信息未知的“小样本”“贫信息”不确定系统为研究对象，主要通过对部分知信息的生成开发提取有价值的信息，实现对系统运行行为、演化规律的正确描述和有效监控。灰色关联分析是灰色系统理论的重要组成部分之一，而且是灰色系统分析、建模、预测、决策的基石。其基本思想 是根据事物或因素的序列曲线的相似程度来判断其关联程度的。若两条曲线的形 状彼此相似，则关联度大；反之，关联度就小。灰色关联分析在经济、社会、农业、矿业、交通等各个领域都有广泛应用，如小麦生育期气候因素与产量的关联分析[39]、中国碳排放强度与产业结构的关联分析（李健）、 区域经济系统的分析[41]等等。（孙玉刚）

### 1.2面向大数据的相关分析方法

针对经典统计学不适用于高维、非线性大数据的缺陷，许多面向大数据特征的相关分析思路被提出。其中应用较多的包括基于互信息的方法、基于矩阵计算的方法、基于距离的方法。



基于互信息的方法采用信息论的思想，使用信息熵来度量系统的不确定性。通过计算已知X后Y的信息量变化程度，如果变化程度越小则说明X与Y的相关性越弱。一个经典的该方法使用案例是两变量间最大信息系数的定义（MIC），实验表明MIC比传统的皮尔逊系数更能细致刻画变量间的相关性，且对非线性的大数据相关性有很好的挖掘能力。

基于矩阵计算的方法重点考虑用矩阵论的思想进一步推广，发展经典的统计相关分析方法。其典型方法例如RV系数、协方差矩阵的改进。RV系数是一个针对两个随机向量的更具有泛化性的相关系数，可以证明多种经典统计相关分析方法都可看为RV系数广义模型的特例。协方差矩阵的改进相关研究通过简化协方差矩阵表，降低计算复杂度，更能适应大数据的高维特性。

基于距离的相关系数通过分布的特征函数与距离相关联，得到相关系数。这种方法的优点在于可以度量的相关性不限于线性关系，同时可以度量任意两个不同维度的随机向量间相关性。被广泛应用与机器学习，时间序列分析，核化学等分析领域。对于大数据的高维、非线性特性都有较好的适应性。

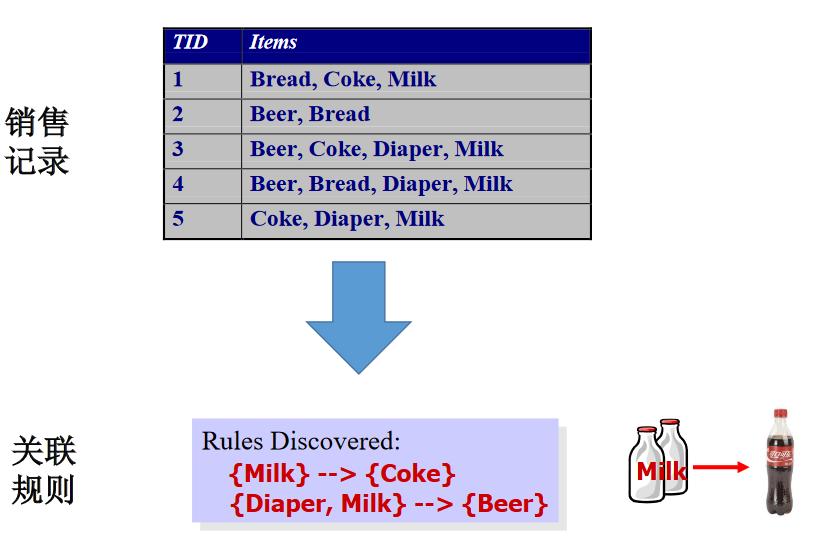
### 1.3数据挖掘的方法

数据挖掘技术是针对海量数据的处理而发展起来的一种数据处理技术 ，它通过对历史数据和实时数据进行分析 ，最后总结出预见性结论。

关联规则挖掘是一种在大型数据库通过识别数据库中的频繁模式发现变量之间关系的方法，是数据挖掘最先研究的问题之一，也是数据挖掘的一个主要研究方向。通过将资源能源安全相关因素的监测历史数据建立数据库，就可以采用关联规则挖掘的方式挖掘出不同因素间的关联关系。

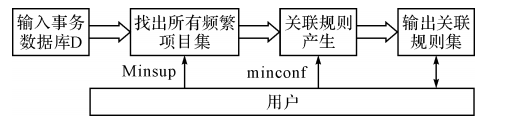
1）关联规则挖掘简介

关联规则挖掘最早是为了发现超市交易数据库中不同的商品之间的关系。关联规则可以反映出隐藏在数据库中有价值的知识。关联规则是形如X→Y的蕴涵式，存在支持度和信任度。X和Y分别称为关联规则的先导和后继。简言之，关联规则就是描述一个事物中物品甲的出现对物品乙的出现的影响有多大的一种规则。



关联规则挖掘的主要思想是通过用户指定最小支持度和最小置信度，挖掘出强关联规则。代表的方法有Apriori算法、FP-树频集算法、多层次关联规则算法、多值属性关联规则算法等等。其基本实现是一个两步的过程 ：1） 根据最小支持度找出事务数据库 D 中所有的频繁项目集 ；２） 由频繁项目集和最小置信度产生强关联规则。

经典的 Apriori方法的核心思想是利用1-频繁项目集生成２-频繁项目集 ，然后根据２-频繁项目集生成３-频繁项目集 ，以此类推 ，直至生成所有的频繁项目集 ，然后从频繁项目集中找出符合条件的关联规则。技术实现上使用层次顺序搜索的循环方法（又称做逐层搜索的迭代方法）产生频繁项集。 Apriori算法在众多领域都有广泛应用，也有许多针对不同场景和需求的改进算法。下图展示了关联规则挖掘的基本过程。



2）关联规则挖掘的应用

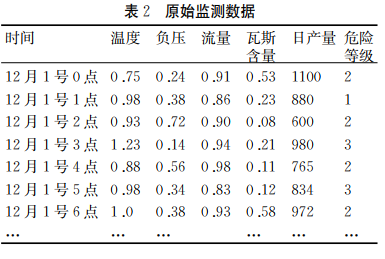
关联规则数据挖掘方法应用十分广泛。在网络安全领域，入侵检测系统 (IDS)运用关联规则生成入侵检测标准规则库，通过判断网络数据与规则库的匹配度产生告警数据流。隐私保护中，隐私保护关联规则算法基于数据干扰和分布重构思想对原始数据进行保护，避免商业敏感数据、个人隐私数据泄露。利用关联规则还可进行蛋白质结构预测，根据蛋白质序列的特点可以对这些序列数据进行量化和处理，再将关联规则算法用于序列数据集上寻找蛋白质序列中的关联关系。在遥感导航定位等科学领域，关联规则算法也被用于分析事件出现的原因，解释其与指标之间的关系。以下给出了关联规则算法用于社会经济相关问题一些的案例，能更具体阐明其用途。

（张鑫，陈燕）使用Apriori算法对经济指标关联进行了分析。选取1999～ 2012年大连市的GDP、CPI、失业率、港口货物吞吐 量、港口旅客吞吐量、航空旅客吞吐量增量六大指标来进行关联挖掘分析。得出的关联规则例如在最小支持度为0.5和最小可信度为0.7的水平 下，GDP增量范围在[0，0．17)且航空旅客吞吐量增量范围在[0，0．19)时，CPI增量％范围在[一2， 1．7)。得出了港口货物吞吐量、航空旅 客吞吐量、港口旅客吞吐量等与经济发展三大指标存在潜在关联，为利用吞吐量来预测经济发展趋势提供理论依据。

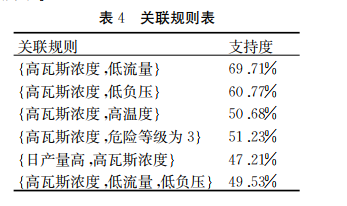
（王晨绫，刘伟）基于Apriori算法对微信热点事件的关联进行了分析与研究。使用微信平台萃取出来的大量热点事件，剔除无效数据，将Apriori算法应用在获取的大量微信热点事件上，分析出影响微信事件成为热点事件的原因。得出结论为微信热点事件的形成与事件主体是否知名，事件发布平台是否知名，事件是否有知名平 台转发，事件阅读数，事件转发数等因素成正相关。

（贾慧娟, 卫晨）使用Apriori算法对煤矿安全数据进行了分析，为了挖掘的方便及有效性 ，首先对数据进行预处理，把连续的数值转化为离散的等级值，再输入到顶尖数据挖掘平台（TipDM ）中的Apriori算法模型，分析了影响煤矿安全预警的因素（如瓦斯浓度 、矿压 、ＣＯ 浓度 、温度 、流量 、水 、火等）与矿井安全之间的内在关联信息。得出结论为影响煤矿安全隐患的主要因素 是瓦斯含量和日产量。同时，数据库管理人员可以利用挖掘结果对可能出现由瓦斯 、ＣＯ 浓度超限等工况信息引起的安全隐患直接监控指挥 、实时控制并及时进行处理。

以下是该研究中使用的原始数据、处理后数据以及输出结果的部分示例。



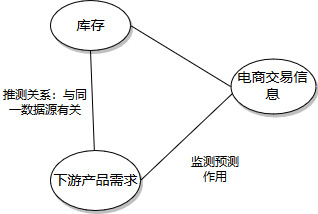




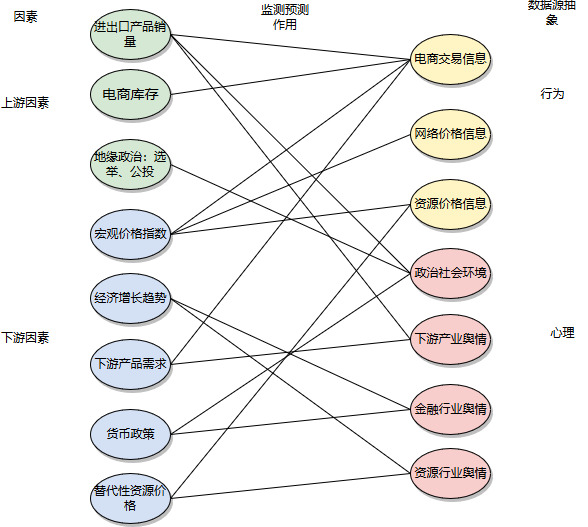
可以从中得出的关联规则如瓦斯浓度高 → 低流量的置信度约为 ７０％ ，瓦斯浓度高 → 低负压的置信度约为 ６１％等，可见通过Apriori算法建立了因素之间，因素与危险等级之间的关系。该研究对于进行资源能源安全因素关系挖掘和风险评定都有很好的借鉴意义。

### 1.4基于因素与数据源关系的网络构建方法

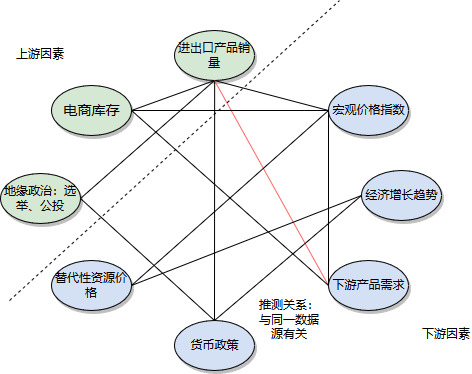
由于使用大数据进行资源能源安全风险因素的监测预测时，往往会出现不同的因素对应同样的，或是具有同样含义的数据源。例如有研究表明，产品库存的预测可以采用淘宝网电商数据（商品历史库存价格、购买量、收藏量、访问量等），同时需求的预测也可以采用电商数据，即库存和需求都可以使用电商数据进行监测预测。正因为因素与数据源之间有相关关系，数据源才能作为监测预测的信息源，而两因素与同一数据源有相关关系，可以推断出这两因素间也存在某种相关关系。可以使用这种“与同一数据源相关”的关系搭建出因素的相关性网络。如下图所示。



下图为选取了部分上下游相关因素构建的与数据源之间的关系图。展示因素与数据源抽象之间的关系：监测预测作用。



下图展示了上述因素之间的关系边的来源：两个因素具有推测关系：与同一数据源有关。边的权重：两因素共同相关的数据源个数。红色边表示权重为2，其余边权重为1。



以上例子的构建针对可监控因素（四级因素）可以容易地扩展到针对三级因素：只要其包含的四级因素中有一个与某一数据源有关，则建立三级因素到该数据源的边，边的权重为所有包含的四级因素边的权重加和。针对二级因素，一级因素也可以按照同样方式构建。

这种方法的一个明显的优势在于，可以充分利用使用大数据进行因素监测预测的模型和数据，直接在此基础上推测关系，而不需要建立额外的模型，再搜集历史数据去分析因素间的相关性。

## 知识驱动的关系网络构建方法

知识驱动的方法从知识的视角构建关系网络。其网络表示的关系是基于人类知识的语义、推理等规律性认识。其中基于经验和规则的方法是直接将领域专家的知识加入到网络中，表现出因素之间在业务逻辑上的关系。而基于知识图谱的方法则是自动挖掘出互联网文本中的因素间语义关系，即不断挖掘新的知识加入网络，更具智能性。

### 2.1 基于知识图谱的关系网络构建方法

知识图谱本质上是一种语义网络，其结点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系。资源能源安全因素关系网络可以看作一个特定领域的知识图谱。其中的点代表资源能源安全因素，边代表因素之间的语义关系，边的权值可以为因素间关系的关联强度。因而可将关系网络的构建问题转化为知识图谱的构建问题。这样既可以完成对给定的因素建立关系网络，也可以完成新因素的自动扩充，关系网络的不断更新迭代。

1）通用知识图谱

知识图谱的概念是由谷歌公司提出的。2012年 5月17日，谷歌发布知识图谱项目Knowledge Graph ，并宣布以此为基础构建下一代智能化搜索引擎。其目的是为了让用户能够更快更简单的发现新的信息和知识，将搜索结果进行知识系统化，任何一个关键词都能获得完整的知识体系。该项目始于2010 年谷歌收购Metaweb公司，并籍此获得了该公司的语义搜索核心技术，其中的关键技术包括从互联网的网页中抽取出实体及其属性信息，以及实体间的关系。

社交网站 Facebook 于 2013 年推出了 Graph Search[126]产品，其核心技术就是通过知识图谱将人、 地点、事情等联系在一起，并以直观的方式支持精 确的自然语言查询，例如输入查询式：“我朋友喜欢的餐厅”“住在纽约并且喜欢篮球和中国电影的朋友”等，知识图谱会帮助用户在庞大的社交网络中找到与自己最具相关性的人、照片、地点和兴趣等[7]。 Graph Search提供的上述服务贴近个人的生活，满足了用户发现知识以及寻找最具相关性的人的需求。（知识图谱技术综述）

微软的Bing Search[116]和Twitter[118]等大型社交服务站点达成了合作协议，在用户个性化内容的搜集、 定制化方面具有显著的优势。

中文知识图谱目前也有不少相关研究。早期的中文知识库主要采用人工编辑的方式进行构建，例如中国科学院计算机语言信息中心董振东领导的知网(HowNet) 项目。由于中文知识图谱的构建对中文信息处理和检索具有重要的研究和应用价值，近年来吸引了大量的研究。例如在业界，出现了百度知心、搜狗知立方等商业应用。其中，搜狗的知立方[119]是国内搜索引擎行业的第一款知识图谱产品，它通过整合互联网上的碎片化语义信息，对用户的搜索进行逻辑推荐与计算，并将最核心的知识反馈给用户。百度将知识图谱命名为知心[120]，主要致力于构建一个庞大的通用型知识 网络，以图文并茂的形式展现知识的方方面面[7]。图为搜狗知立方在实际搜索中的应用，扩展用户搜索结果，发现更多内容。



在学术界，清华大学建 成了第1个大规模中英文跨语言知识图谱XLore④、 中国科学院计算技术研究所基于开放知识网络 (OpenKN)建立了“人立方、事立方、知立方”原型系统。中国科学院数学与系统科学研究院陆汝钤院士 提出知件(Knowware)的概念、上海交通大学构建并发布了中文知识图谱研究平台zhishi．me④、复旦 大学GDM实验室。推出的中文知识图谱项目等[2一， 这些项目的特点是知识库规模较大，涵盖的知识领 域较广泛，并且能为用户提供一定的智能搜索及问答服务。

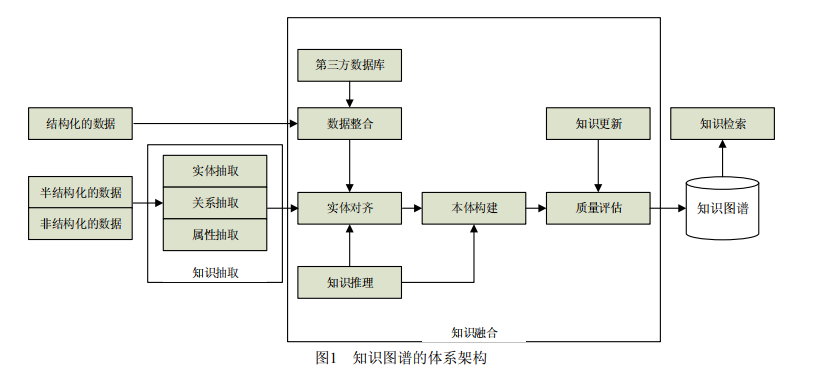
2）垂直行业知识图谱应用

行业知识图谱也称垂直知识图谱，其知识来源针对某些特定行业，如医疗行业、金融行业等。垂直知识图谱的构建和应用也有广泛的研究。

例如在金融行业中，反欺诈是一个重要的环节。它的难点在于如何将不同税务子系统中的数据整合在一起。通过知识图谱，一方面有利于组织相关的知识碎片，通过深入的语义分析与推理，可对信息内容的一致性充分验证，从而识别或提前发现欺诈行为；另一方面，知识图谱本身就是一种基于图结构的关系网络，基于这种图结构能够帮助人们更有效 地分析复杂税务关系中存在的潜在风险[127]。知识图谱还可以在精准营销方面，通过链接的多个数据源，形成对用户或用户群体的完整知识体系描述，从而更好地去认识用户的行为。

在医疗行业，耶鲁大学拥有全球最大的神经科学数据库 Senselab[128]，然而，脑科学研究还需要综合从微观分子层面一直到宏观行为层面的各个层次的知识。 因此，耶鲁大学的脑计划研究人员将不同层次的与脑研究相关的数据进行检索、比较、分析、整合、 建模、仿真，绘制出了描述脑结构的神经网络图谱， 从而解决了当前神经科学所面临的海量数据问题，从微观基因到宏观行为，从多个层次上加深了人类 对大脑的理解，达到了“认识大脑、保护大脑、创 造大脑”的目标。

3）知识图谱的构建



知识图谱的构建有两种方式，其一是自顶向下法：将百科中已有的结构化数据整理，直接加入知识库。其二是自底向上法：从公开采集的数据中自动抽取知识模式加入。自顶向下法在知识图谱的早期应用中被广泛采用，而自底向上法则是目前的研究重点以及未来发展的趋势。下面主要介绍自底向上法的构建过程。

* 信息抽取

从各种类型的数据源中提取出实体(概念)、属性以及实体间的相互关系，形成本体化的知识表达。信息抽取的主要技术方法是实体抽取、关系抽取、属性抽取。

* 知识融合

在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应于多个不同的实体等。知识融合中涉及的主要技术方法是实体链接（实体消歧、共指消歧）、知识合并（外部数据库、关系数据库）等。

* 知识加工

对于经过融合的新知识，需要经过质量评估之后(部分需要人工参与甄别)，才能将合格的部分加入到知识库中，以确保知识库的质量。新增数据之后，可以进行知识推理、拓展现有知识、得到新知识。知识加工中涉及的主要技术方法有本体构建（实体并列关系相似度计算、实体上下位关系抽取以及本体的生成）、质量评估（模型评估、综合评估）等。

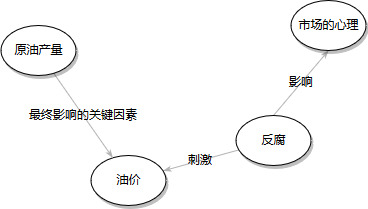
在知识加工环节中，知识推理是一个备受关注的研究领域。知识推理是指从知识库中已有的实体关系数据出发，经过计算机推理，建立实体问的新关联，从而拓展和丰富知识网络。通过知识推理，能够自动地从现有知识中发现新的知识，更加具有人类智能。如已知(乾隆，父亲，雍正)和(雍正，父亲，康熙)，可以得到(乾隆，祖父，康熙)或(康熙，孙子，乾隆）。再如（虎，科，猫科)和(猫科，目，食肉目)，可以推出(老虎，目，食肉目)。知识推理的相关方法可以分为基于逻辑的推理，包括一阶谓词逻辑、描述逻辑、基于规则的推理。以及基于图的推理，包括基于神经网络模型或Path Ranking算法等。如Socher等人将知识库中的实体表达为词向量的形式，进而采用神经张量网络模型进行关系推理。

）知识图谱在资源能源安全因素关系网络中的应用

可以用下面这句话展示基于知识图谱的关系网络构建过程。

*“不过，虽然沙特的反腐刺激了油价，但原油的产量才是最终影响油价的关键因素，随着反腐行动的平息，其对市场的心理影响也将减弱，油价短期存在冲高回落的可能。”*

从中可以抽取出实体：反腐行动、油价、原油产量、市场心理。其中存在的语义关系对包括（反腐行动、刺激、油价）、（原油产量、最终影响的关键因素、油价）、（反腐行动、影响、市场心理）。根据这句话可以画出如下的关系图谱。



总体而言，基于知识图谱的方法将资源能源安全因素关系网络看作一个特定领域的知识图谱。采用特定领域的知识图谱构建方法，技术包括文本挖掘、自然语言处理、知识推理等。采用的数据源是与资源能源安全相关的互联网、以及传统文本。主要的架构是从数据源中使用实体抽取方法挖掘新的影响因素，使用关系抽取、属性抽取方法挖掘文本中的因素间的语义关系，再对挖掘出的实体和关系进行融合、消除矛盾和歧义，进行质量评估，将达到质量要求的实体和关系加入到知识图谱中。还可以进一步进行知识推理实现语义关系即知识的自动扩展。