# 第一章 知识图谱介绍

### 1.1 什么是知识图谱？

如果你此前没有接触过，估计你看到“知识图谱”这个概念的时候就像丈二的和尚——摸不着头脑，我第一次听到“知识图谱”时，也是这样的感受。

听起来很高大上啊，但具体是个什么东西，完全不知道。如果你是爱学习的人，估计接下来会使用百度、必应、谷歌等搜索引擎去搜索知识图谱的概念。但是搜索引擎给出的答案并不能满足我们的需求。现在网络中关于知识图谱的信息还很少，现有的博客文章也大部分是抄袭已发表的论文或者是对论文的解读。

知识图谱是由谷歌公司提出的概念。谷歌于2012年5月17日（美国时间：5月16日）发布了它的知识图谱项目，并宣布以此为基础构建下一代智能化搜索引擎。因为公开披露的资料还很少，知识图谱就显得不那么平易近人，让人们一时难以看清这一技术的内涵和价值。

虽然知识图谱是一个很新的概念，但实际上它并不是全新的研究领域，它是从现有的概念演变过来的，大致经历了如下6个过程，图1.1按时间给出了相关概念的演变过程。

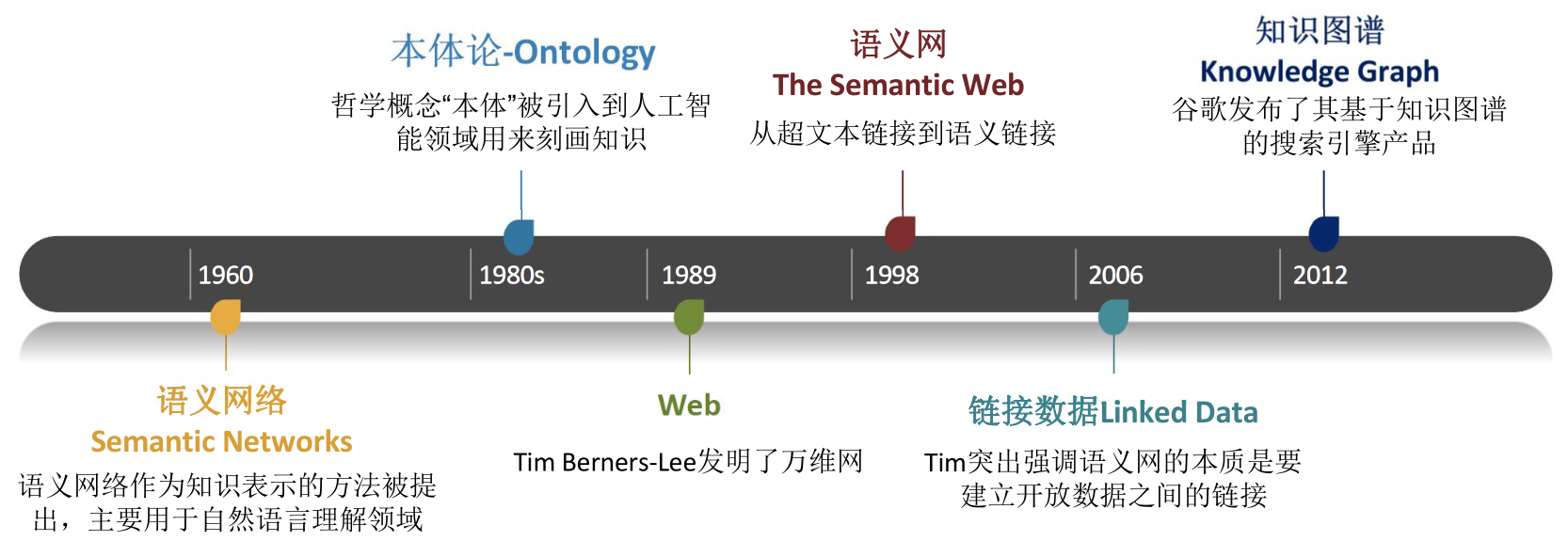


图1.1 相关概念演变

Google知识图谱（英语：Google Knowledge Graph，也称Google知识图）是[Google](https://zh.wikipedia.org/wiki/Google)的一个[知识库](https://zh.wikipedia.org/wiki/%E7%9F%A5%E8%AF%86%E5%BA%93)，其使用[语义检索](https://zh.wikipedia.org/wiki/%E8%AF%AD%E4%B9%89%E6%A3%80%E7%B4%A2)从多种来源收集信息，以提高Google搜索的质量。知识图谱2012年加入Google搜索，2012年5月16日正式发布[[1]](https://zh.wikipedia.org/wiki/Google%E7%9F%A5%E8%AF%86%E5%9B%BE%E8%B0%B1#cite_note-announce-1)，首先可在[美国](https://zh.wikipedia.org/wiki/%E7%BE%8E%E5%9B%BD)使用。知识图谱除了显示其他网站的链接列表，还提供结构化及详细的关于主题的信息。其目标是，用户将能够使用此功能提供的信息来解决他们查询的问题，而不必导航到其他网站并自己汇总信息。[1]

上一段是维基百科所给的“Google知识图谱”的定义，从上面我们可以解读出以下几点重要信息。

（1）知识图谱是一个知识库，这个知识库中的知识来源渠道很多。

（2）知识图谱可以提高谷歌搜索引擎的质量，即可以改善用户体验。

（3）使用知识图谱技术，用户搜索到的知识不再是单一的网页链接列表，还有更详细的与搜索主题直接相关的信息。

我们直接使用谷歌搜索引擎体验一下，图1.2是搜索词为“达芬奇”的搜索结果，可以看到结果页面不仅在左侧给出了相关网页的链接，最重要的是在右侧还给出了达芬奇的出生年月、父母、作品等信息，而且在右侧底部一栏，还推荐了像达芬奇一样的其他艺术家。

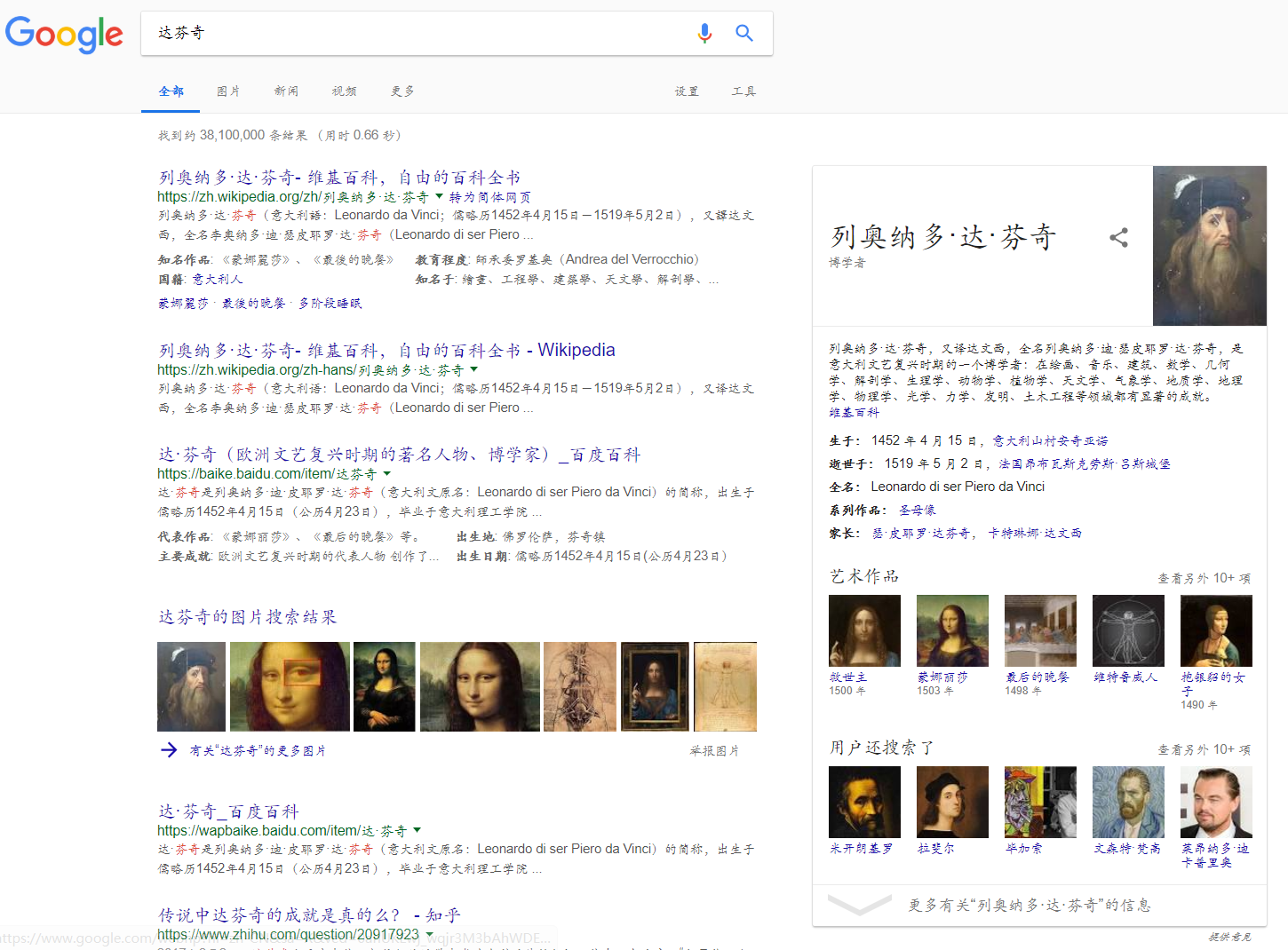


图1.2 “达芬奇”结果页面

图1.2展现了谷歌知识图谱的在实际应用中所表现的作用。人类对于图像的接收能力是最好的，我们借用一张图来简单说明搜索“达芬奇”如何能给出图1.2右侧的结果。

如图1.3所示，学过数据结构的人可能一下就能联想到“图”这个概念，图上有很多的结点，各个结点之间通过一条线（实际上是关系）联系起来，达芬奇这个结点记录了达芬奇的个人信息；而其他结点又通过连线与达芬奇建立了联系，那么使用我们熟悉的图搜索算法就能顺藤摸瓜的找到蒙娜丽莎、米开朗基罗、意大利等结点，也可以很直接的就能访问到与达芬奇相关的结点信息。我想你现在应该大致对知识图谱有一个自己的概念了。



图1.3 达芬奇知识图谱

现在给出知识图谱一个更直观一点的定义：**知识图谱是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系。其基本组成单位是“实体-关系-实体”三元组，以及实体及其相关属性-值对，实体通过关系相互联结。**[2]

我们拿数据结构中的“图”进行类比，图中的结点就是现实世界中的实体，不过这些实体不是只有一类，比如图1.3中就有地点、艺术作品、人类等多种不同的结点；实体与实体之间的语义关系就是“图”的边，所以知识图谱实际上就是通过关系链接而成的网状知识库。现在你应该对知识图谱有一个比较清晰的认识了。

### 1.2 知识图谱架构及构建方法

知识图谱架构分为逻辑架构和技术架构。一般的将知识图谱逻辑架构划分为两层，分别为数据层和模式层。

在数据层中，知识以事实为单位存储在数据库中，对于数据库的选择，可以使用关系型数据库，也可以使用图形数据库。在前文中已经描述了，知识图谱实际上就是一张很大很大的网，各个结点之间是有联系的，如果使用关系型数据库，那么在查询的时候就不可避免的会涉及到连接查询，而我们都很清楚，连接查询在关系型数据库中执行的比较慢，而且由于知识图谱的特殊性，基本上都会涉及到链接查询，所以不推荐使用关系型数据库。

模式层建立在数据层之上，模式层听起来很抽象，可以和关系型数据库的范式类比，模式层实际上是对图谱中的知识进行约束。通常采用本体库来管理知识图谱的模式层，你可以把模式层当作一个模具工厂，这个工厂生产出各种各样的模具，下层的知识都需要按照自己所对应的模具去规范自己，而本体就是我们这里的模具。

看到本体，估计又有丈二和尚摸不着头脑的味道，本体本来是哲学里面的概念，后来才被引入到信息科学领域，Neches给出的本体论在信息科学中的定义是：给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外延规则的定义。

上面的说法还是晦涩难懂，在此我仅讲讲我个人的理解。本体论是概念层的，它面向对象中的“类”相似，是一个通用的东西，实例是本体的具体呈现。比如在本体库里面有一个“人类”的本体，人类应该有的属性包括姓名、性别、年龄等等，而张三、李四、王二等等就是“人类”这个本体中的实例，它们都应该遵循本体的约束，不能有多余的属性等信息。更简单的理解就是类和对象的关系，本体对应类，而对象对应实例。

知识图谱的技术架构，主要指构建过程中所采用的技术架构，图1.4是知识图谱常见的整体技术架构，也是常见知识图谱的构建流程，知识图谱的构建过程是从原始粗语料库出发，通过一系列自动化、半自动化，甚至是人工的方式提取可用的事实三元组，然后通过知识融合筛选掉其中的冗余、错误知识，最后将抽取的可用知识存入数据库。可以看到这是一个更新迭代的过程，网络中的知识是每天都在更新的，我们的知识图谱也要随之更新才能适应变化，每一次迭代包括4个步骤：知识获取、知识提取、知识融合和知识处理。

构建知识图谱分为两种方式，分别是自顶向下和自底向上。自顶向下常从百科类网站等结构化的数据中提取知识并加入到知识库中；自底向上则通过一定的技术手段，从公开的非结构化数据中提取知识，选取置信度较高的知识加入知识库中。本文同时借助了百科类网站的半结构化数据，和新闻类网站的非结构化数据进行构建。下面介绍了按数据来源进行区分的四种不同构建方法。

（1）基于网络百科资源的知识图谱构建方法，这种方法通常会从互动百科、维基百科、百度百科等资源中抽取自己所需的知识。以互动百科为例，它提供了百科知识分类树，可以很方便的根据这棵分类树查找自己所需的知识，每一个词条都是一个单独与之对应的页面内，该页面首先是词条名称、然后会列出该词条的详细解释，紧接着会列出词条的其它属性信息，通过html解析技术就可以提取出其中的值-属性对信息，而且不需要添加任何规则就可以很方便的提取知识，达到了快速高效的理念。

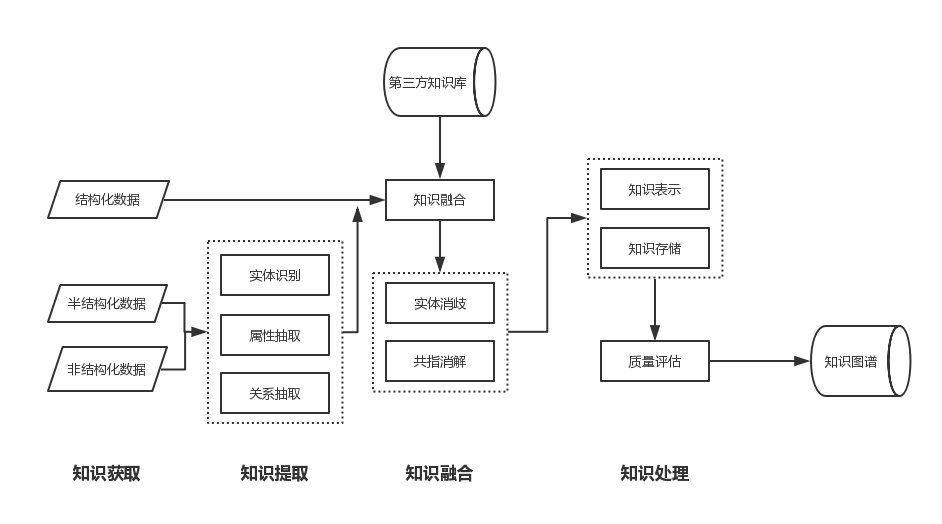


图1.4 知识图谱技术架构

（2）基于结构化数据的知识图谱构建方法，知识图谱的主要工作就是将数据进行组织，并且进行结构化存储，结构化数据是知识图谱的重要数据来源，因为结构化数据的质量高，易于处理，所以首先都会考虑将结构化知识加入知识库中。现在广泛采用的结构化数据有Yago、DrugBank等知识库。

（3）基于半结构化数据的知识图谱构建方法，我们常见的半结构化数据就是百科类网站所提供的知识，相比英文类的百科网站，中文百科类网站的结构化程度还远远赶不上维基百科，实体的很多属性都隐藏在表格中，这时就需要对这些半结构化数据进行抽取。一般我们可以通过观察具体网站来针对性的编写知识提取程序，因为同一个百科类网站的数据组织方式基本都是一样的，我们只需要找到它组织数据的方式，然后再通过定义一些规则就可以实现了知识的抽取。

（4）基于非结构化数据的知识图谱构建方法，这是目前学术前沿的研究方向，也是热门的研究方向，因为对于特定领域的知识图谱，仅仅依靠百科类数据资源无法满足需求，所以还需要从非结构化数据中提取领域知识。相比上面三种数据来源，目前从非结构化数据进行知识的抽取应用并不广泛，对非结构化数据的处理对技术的要求非常高，而且现在对非结构化数据进行处理的公开技术知识也极度欠缺，而对中文领域开放文本的处理也并不成熟，可想而知使用非结构化数据构建知识图谱难度之大。

### 1.3 知识图谱研究现状

知识图谱项目始于2010年谷歌收购Metaweb公司，籍此获得了该公司的语义搜索核心技术，该公司的语义搜索技术包括从互联网网页中抽取实体和属性信息，以及抽取实体间关系的技术。

我国对于中文领域的知识图谱研究已经起步了，以中科院计算机语言信息中心董振东领导的知网项目是一个代表，它是一个规模小、质量高、限定领域的知识库，早期主要采用人工方式进行构建。

业界也建立了百度知心、搜狗知立方等知识图谱，百度、搜狗都将其应用到了自己的搜索引擎产品中，国内日常大多都使用百度或搜狗搜索引擎。如图1.5所示，使用百度搜索“汪涵的妻子”，搜索引擎会直接给出答案“杨乐乐”，而不是给出网页链接列表，这就是通过知识图谱实现的，在已有知识图谱的基础上使用知识推理即可想前文那样给出精准的答案。中文知识图谱的构建在中文信息处理和检索有很重要的研究与应用价值。



图1.5 “汪涵的妻子”结果页面

各大高校与科研机构也都推出了自己的中文知识图谱项目，比如清华大学的中英文跨语言知识图谱XLore、上海交通大学发布的中文知识图谱研究平台zhi.me、中国科学院陆汝玲院士提出了知件的概念、复旦大学GDM实验室推出的中文知识图谱项目等。

目前对知识图谱的应用大多集中在医疗、金融、教育行业，将知识图谱与具体领域结合，实现为用户提供智能搜索和智能问答的服务，表1.1给出了现有部分知识图谱相关产品。

表1.1 现有知识图谱部分产品

|  |  |  |
| --- | --- | --- |
| Knowledge Base | Products | Data Source |
| Konwledge Graph | Google Seach Engine  Google Now | Wikipedia,Freebase,Web Open Data |
| Wolfram Alpha | Apple Siri | Mathematica |
| Sarori/Probase | Bing Seach Engine | Wikipedia,Web Open Data |
| Watson KB | IBM Watson System | Web Dictionaries,The World Book,Encyclopedia |
| DBpedia KB | DBpedia | Wikipedia |
| YAGO KB | YAGO | Wikipedia |
| NELL KB | NELL | Web Open Data |
| Facebook KB | Shopycat | Social Network Data |
| Zhilifang KB | Sougou Seach Engine | Web Open Data |
| Zhixin KB | Baidu Zhixin Platform | User Generated Content |
| Cross-Lingual KB | XLORE | Chinese/English Encyclopedia,Wikipedia |
| Zhishi.me KB | Zhishi.me | Chinese Encyclopedia |

### 1.4 知识图谱类型及应用

现在用于商用的知识图谱主要分为以下几种类型：

（1）领域无关知识图谱，这种类型的知识图谱一般不限于特定领域，它尽可能的包含更多的知识，内容比较广泛，谷歌知识图谱就是一个典型的领域无关知识图谱，百度所构建的百度知心也是一个领域无关的知识图谱，这种知识图谱常使用本体对知识进行分类校验。

（2）特定领域知识图谱，这类知识图谱都是针对某一特定领域，其内容没有领域无关知识图谱广泛，但是在特定领域的知识会更加全面而深入，更加具有针对性，现在热门的医疗、金融、教育知识图谱都是特定领域知识图谱。

（3）跨语言知识图谱，这类知识图谱包含了多种语言的描述，比如DBPedia、Freebase都属于跨语言知识图谱。

当前对于知识图谱的应用主要集中在三个方面，分别为搜索引擎、智能问答系统和聊天机器人。

（1）对于搜索引擎我想不需要在做过多的介绍了，前文所有对知识图谱的介绍都是围绕着搜索引擎的，在网络高度发达的今天，搜索引擎是人们获取信息的重要工具，而且知识图谱最开始也是谷歌为了提高其搜索引擎的体验而提出来的概念，将知识图谱应用于搜索引擎上是重要的方向。

（2）智能问答系统是构建在知识图谱之上的应用，知识图谱为其提供数据支撑，当然问答系统的数据不一定都是以知识图谱的方式进行组织的，也可能是以其他方式组织的。第一个问答系统是由麻省理工学院开发的STRAT。

（3）聊天机器人也可以归为智能问答系统的一种，微软小冰是聊天机器人的代表，它采用网络公开聊天语料，通过数据挖掘、情感分析和意图识别来实现与用户的交流。而京东JIMI、阿里小蜜更偏向于客服功能，使用的语料主要是商品信息和订单信息。

### 1.5 本文主要内容及结构

前面介绍了很多关于知识图谱的信息，本文旨在使用网络开放语料库构建一个特定领域的知识图谱，所选择的领域为经济责任审计，对于从事计算机工作的人来说，经济责任审计可能表现的抽象，不过这完全没有关系，我也属于经济责任审计领域的门外汉，我们要学习的是技术，只需要知道要构建的是经济责任审计这一特定领域的知识图谱。

本文将从零开始，基于百科网站资料构建一个相对完善的经济责任审计领域知识图谱，使用编程语言主要是Python3和Java，所以本文不考虑零基础读者，在此之前需要掌握一定的编程基础，才能读懂本文内容。

本文旨在通过具体案例，带领读者入门知识图谱领域，完全涵盖知识获取、知识提取、知识融合、知识存储四个步骤；由于本文所构建的知识图谱不是开放领域的，所以需要对知识进行筛选，只有属于经济责任审计领域的知识才是我们所需要的，因此在上面的步骤之上需要加上一个领域词汇判定的步骤。

由于是入门类书籍，所以会在某些步骤选择性过滤掉一些技术，比如本文只使用了一个类别来约束所有的实体，所以就丢弃了本体库约束的方式；同时实体消歧、共指消解本文也不涉及，在知识融合阶段只通过技术手段筛选掉已有知识中的冗余、错误信息。

大致结构还没有写

**参考资料：**

[1] <https://zh.wikipedia.org/wiki/Google>知识图谱

[2] 刘峤, 李杨, 段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3):582-600.