关联图谱的研究进展及面临的挑战

尹 亮[1] 袁 飞[2,3] 谢文波[2,3] 王栋志[4] 孙崇敬[2,3]
(装甲兵工程学院 北京 100072)[1] (电子科技大学大数据研究中心 成都 611731)[2]
(电子科技大学计算机科学与工程学院 成都 611731)[3]
(西南科技大学计算机科学与技术学院 四川 绵阳 621010)[4]

摘 要 随着 Web 技术的不断发展和 Linked Open Data 等项目的相继开展,关联图谱已被广泛应用于互联网智能搜索、图书馆书目管理、医学、智能制造等领域,并取得了显著的成果。文中深刻阐述了关联图谱的定义、架构以及构建的关键技术,包括实体抽取、实体间关系抽取和知识融合等方面的研究进展,并深度分析了当前关联图谱分析与研究所面临的若干挑战问题。

关键词 关联图谱,实体抽取,关系抽取,知识融合
中图法分类号 TP391 **文献标识码** A

Research Progress and Challenges on Association Graph

YIN Liang[1] YUAN Fei[2,3] XIE Wen-bo[2,3] WANG Dong-zhi[4] SUN Chong-jing[2,3]
(Academy of Armored Force Engineering, Beijing 100072, China)[1]
(Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China)[2]
(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)[3]
(School of Computer Science and Technology, School of Computer Science and Technology, Mianyang, Sichuan 621010, China)[4]

Abstract With the development of web technology and projects such as Linked Open Data having been carried out, the association graph has made significant contributions on many areas such as Internet intelligent search, library bibliographic management, medicine and intelligent manufacturing. This paper reviewed the key topics of the association graph, including definition, framework and construction etc. The research progress on entity extraction, relationship extraction and knowledge fusion are discussed thoroughly. Furthermore, some challenges on association graph are also summarized.

Keywords Association graph, Entity extraction, Relation extraction, Knowledge fusion

1 引言

近年来,随着公开的关联数据集等项目的开展,关于关联图谱的研究也逐渐深入[1]。关联数据已成为国际互联协会(W3C)推荐的一种规范,用以发布和链接各类数据、信息和知识。到 2010 年底,基于互联网的关联数据集已经有 100 多种,覆盖了生物、地理、文化、智能制造等多个领域。

在此背景下,关联图谱的研究得到了迅速发展。在搜索领域[2],以 Google[3]、百度[4]、搜狐为首的搜索引擎公司以此为基础,构建出描述各种实体和实体之间丰富关系的关联图谱[5];图书管理领域以作者和图书信息的关联数据为基础,建立起图书书目的关联图谱[6],达到信息重用和数据开放互联的目的[7];在医学领域,构建了大量的疾病、基因以及蛋白质之间形成的关联图谱[8]等。

本文从关联图谱构建核心技术的研究进展的角度,全面而深刻地介绍关联图谱,包括关联图谱的定义、逻辑架构和体系架构。同时,从实体抽取、实体间关系抽取以及知识融合等方面对关系图谱中的关键技术进行了全面的综述。此外,还对关联图谱研究中的热点和难点问题进行了详细的分析和总结,帮助读者全面而清晰地理解关联图谱,从而更加客观地做出评价和使用。

2 关联图谱的定义与架构

2.1 关联图谱的定义

"图谱"[9-10]一词,源于数学中的图概念。在数学中,一个图(G)可以定义为一个二元组(V,E),记为:

$$G=(V,E)$$

其中,V 表示顶点的非空有限集合,E 是 V 中顶点与顶点的关系集合。

关联图谱即复杂的关联图[11],通常展现了两个可变量之

本文受国家自然科学基金(61433014,61673085),中央高校基本科研业务费专项资金(ZYGX2014Z002)资助。

尹 亮(1982-),男,硕士,工程师,主要研究方向为军事装备信息化、大数据挖掘与分析,E-mail:1804359156@qq.com;**袁 飞**(1992-),女,硕士生,主要研究方向为数据挖掘、深度学习,E-mail:boerhesi@qq.com(通信作者);**谢文波**(1990-),男,博士生,主要研究方向为大数据分析与挖掘、推荐系统;**王栋志**(1993-),男,硕士生,主要研究方向为数据挖掘;**孙崇敬**(1986-),男,博士,助理研究员,主要研究方向为机器学习、社会网络分析、隐私保护,E-mail:sunchongjing@uestc.edu.cn(通信作者)。

间关系的图表，其中，每个变量沿着互相垂直的一对轴中的另一个变量来度量。三元组是关联图谱的普遍表示形式，记为：

$$G=(V,E,S)$$

其中，V 和 E 与图谱定义中的含义相同，S 表示实体与实体之间的关联关系集合。

实体是关联图谱的基本构成元素，关联图谱的实体可以是客观世界中的一切事物。不同的实体具有不一样的属性和属性值。属性指的是实体可能具有的特征，属性值指的是特征的参数。不同的实体之间存在着不同的关系，在关联图谱中使用关系来连接两个实体，以刻画实体之间的关系，从而构成复杂的关联图谱。

目前，关联图谱已被广泛应用于各种大规模的实体关系中，它的具体存在形式有很多种，例如 Google 的知识图谱[12-13]、Facebook 的社交图谱等。关联图谱是广义上的知识图谱，旨在将各种知识通过相互之间的关系连接在一起，如基因图谱、视频图谱等。狭义的知识图谱指的是 Google 建立的知识图谱。

2.2 关联图谱的架构

关联图谱是由自身的逻辑结构以及体系架构两个部分组成。

2.2.1 关联图谱的逻辑架构

关联图谱的典型应用依赖于自身的逻辑结构，如图 1 所示。其中，第一层即为关联图谱的应用层，第二层为关联图谱自身的核心模块。

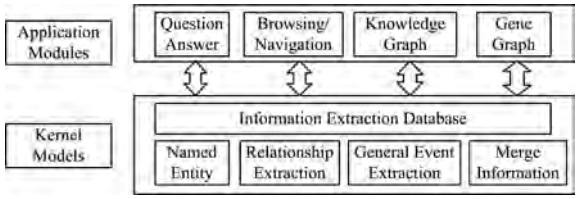

图 1 关联图谱的逻辑架构

关联图谱的第一层应用层（Application Modules）包括行业关联图谱和通用关联图谱。行业关联图谱往往需要依赖特定行业的专业知识来构建，形成具有行业特色的关联图谱，如基因图谱（Gene Graph）。通用关联图谱的信息来源广泛，可以融合更多的实体，描述实体的精确度相对于行业关联图谱而言略低，但是覆盖的范围大，应用广泛，例如问答系统（Question Answer）以及搜索引擎中的知识图谱（Knowledge Graph）等。

关联图谱从逻辑上可以分为两大层：模式层和数据层。其中，数据层是由一系列以事实为单位的知识组成。一般采用三元组（事实，属性，事实）描述知识，也可以利用四元组（事实，属性，事实，关系）进行描述；而模式层构建于数据层之上，利用已有的实体约束和规范数据层中抽象出来的知识。在关系图谱中，本体是结构化的概念模板。

关联图谱模式层的构建方式[12]一般采取自顶向下（Top-Down）或自底向上（Bottom-Up）的方式。自顶向下的方式是指依赖已有的本体编辑器（Ontology Editor）预先构建本体，当关联图谱中不存在本体时，需要依赖一些百科类或其他结构化的数据，从结构稳定的高质量数据中提炼出实体的模式信息，完成从无到有的过渡。自底向上的方式则是通过实体识别、关系识别等相关技术[14]从各种数据源（如搜索日志）中发现实体、属性以及实体和实体之间的关系，并将提取结果中置信度高的结果并入已有的关联图谱中[15]，利用实体对齐等算法进行合并。未能匹配的模式可以作为新模式并入关联图谱，并适当加入人工干预。

自顶向下的方式在保证抽取质量的前提下，有利于从信息中抽取新实体；而自底向上的方式有利于发现新的模式。因此，在实际的研究分析过程中，经常将二者结合使用，如 Google 公司的知识图谱模式层的构建。

2.2.2 关联图谱的体系架构

关联图谱的体系结构意为构建关联图谱的模式结构，如图 2 所示。构建关联图谱的数据来源丰富，可以利用结构化的数据，这类数据的结构明确、数据质量高且更新速度慢。同时，构建关联图谱也可以利用各种半结构化或非结构化的数据，抽取实体的属性键值对，以丰富对已有实体的描述。

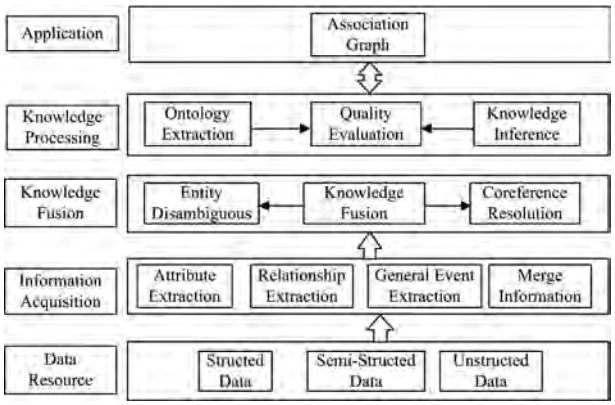

图 2 关联图谱的体系架构

从各种类型的数据源中识别出构建关联图谱需要的各种元素，即实体、实体的属性、实体和实体之间的关系等，从而形成多个相互孤立的抽取图谱（Extraction Graphs）。将抽取图谱中的实体进行归并，使其具有全局唯一标识，通过标识可以获得更多信息的实体对象，并将这些实体对象添加至关联图谱中。对所有抽取图谱中的实体进行实体对齐，融合相关知识，最终将抽取图谱融合成关联图谱。

模式是对知识的高度提炼，经过知识融合完成数据层的构建，对知识进行提炼，形成知识标准，然后在此基础上制定关联图谱的模式层标准。从本质上说，模式层的构建相当于在关联图谱中建立本体。关联图谱中的本体元素包括：概念、概念层次、属性、属性值类型、关系、关系定义域概念集以及关系值域概念集。

3 关联图谱的关键技术、定义与方法

本文主要从实体抽取、关系抽取和知识融合 3 个方面分析关联图谱构建的关键技术与方法。

3.1 实体抽取

一个固定的原始数据集由大量的实体构成，每个实体可用一个元组集合 $U=(u_1,u_1,\cdots,u_n)$ 表示，其中 u_i 是第 i 个实体对应的元组，它描述了对应实体的某个特征。通常，实体抽取方法通过在元组集合 U 上进行一个划分 $R=(R_1,R_1,\cdots,R_m)(m\leqslant n)$，将代表同一实体的多个元组组织在一个簇中。$R_j(j=1,2,\cdots,m)\subseteq U$ 是元组 U 的子集，包含 R_j 中元组的实体可以指代为同一实体，包含 R_j 和 R_k 中元组集的实体被判定为不同实体。

现有的实体提取方法主要有基于规则的方法、基于距离函数的方法和基于机器学习的方法等。

3.1.1　基于规则的实体抽取

基于规则的实体提取方法[16]针对已有实体的形式特征制定相应的规则，使得系统可以依赖这些信息自动生成精确抽取实体生成的位置和实体属性等信息。文献[15]提出了从XML存储形式的数据[17]中抽取现实生活中有具体意义的物理实体的方法。该方法基于规则的实体抽取技术，首先利用松弛技术自动获得大量包含实体抽取的候选查询集合，再利用相似度计算的方法剔除候选集合中不适用于实体抽取的查询。

通常地，一组属性相似的实体可以表示为 $e=(q_L, qa_1, \cdots, qa_k)$，$q_L$ 为所抽取实体在XML中的节点位置，所抽取的实体通过 k 个属性特征 qa_i 来描述。由此，实体的抽取过程可被看作寻找到 q_L 和 $\{qa_i\}$ 的过程。在此基础上，可将基于规则的实体抽取的基本思路归纳为：

1）根据用户的行为兴趣规则量化实体特征 $\{qa_i\}$，并利用给定规则蕴含的语义信息自动抽取实体；

2）给定规则 φ，自动生成初始位置查询 q^{φ} 和属性查询集合 A^{φ}，并使用松弛技术扩充候选集 C，以提高查询能力；

3）从 C 中选择所要提取的目标实体位置 Q，并对 Q 中的每个位置抽取对应的属性集合。

3.1.2　基于距离函数的实体抽取

文献[18]中提出了一种基于欧氏距离的实体抽取方法。该方法考虑到字符串中的不同词语具有不同的重要性，且不同词语之间存在着相关性[19]，提出了基于词特征的距离度量方式，以达到识别实体的目的。在基于距离度量的实体抽取方法中，为生成输入的每个数据集中的词序列的特征向量，利用分类器对词向量矩阵进行分类处理，其中在样本上使用的距离函数，即是从样本中抽取实体。

在基于欧氏距离度量的实体抽取方法中，首先对实体的元组信息进行再提取，形成元组的词特征向量、标签以及描述实体之间的距离测度。给定数据集用 n 个元组表示，即 $D=(r_1, r_2, \cdots, r_n)$。其中，第 i 个元组 r_i 的词特征向量记为 $\vec{x}_i=(a_1, a_2, \cdots, a_m)$，$r_i$ 所指代的实体为 $e(r_i)$。词特征向量的分量定义如下：

$$a_k=\begin{cases}0, & t_k \notin r_i \\ 1, & t_k \in r_i\end{cases}, 1 \leqslant k \leqslant m$$

那么，$\forall r_i \in D$，存在相应的词特征向量 x_i，则称 $X_D=(\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n)$ 为数据集 D 上的词特征矩阵。基于词特征矩阵构建元组的距离度量，元组的距离度量采用传统的欧氏距离来定义。因此，对于数据集 D 中的两个元组 r_i 与 r_j，其词特征向量分别为 $\vec{x}_i$ 和 $\vec{x}_j$，则两个元组之间的距离度量 $dist(r_i, r_j)$ 等于两个词特征向量差值的2-范数：

$$dist(r_i, r_j)=\| \vec{x}_i-\vec{x}_j \|_2^2$$

结合 D 的词特征矩阵 X_D 和 D 在每个元组的标签，利用上述距离度量函数 $dist(r_i, r_j)$ 学习得到一个距离度量矩阵 L，对词特征矩阵进行线性变换 $L \times X_D$，利用分类器对变换后的词特征矩阵进行分类，即可得到实体的划分结果。

3.1.3　关联图谱的体系架构

基于机器学习的方法也被广泛应用于实体抽取中，例如条件随机场方法可以达到在文本数据中进行实体识别的目的[20]。

通常有监督的机器学习算法需要大量标注训练数据，同时训练数据的规模会直接影响识别效果。因此，一些基于规则的策略被提出，这在一定程度上突破了这一制约。例如，利用实体库（字典等）可以大大提高抽取的准确率和召回率。

针对需要从少量的实体中抽取实体识别的模式，从而达到在海量文本中表现出极好的效果的实际需求，有研究者提出了迭代扩展的策略，通过利用少量的实体来建立实体抽取的特征模型，将其广泛地应用于新的数据集，从而得到新的实体。根据新实体，再次更新特征模型。基于朴素贝叶斯模型的实体抽取算法[21]就是其中最经典的一种方法，其主要借助WordNet识别文档中的实体。

WordNet是由普林斯顿大学认识科学实验室建立和维护的英文字典，它根据单词的意义对单词进行分组，每个分组则代表一个概念，称为一个语义单元，每个概念之间由各种错综复杂的关系链接。利用概念在关系链接网络中的距离和层级，可以定义单词之间的距离和深度。借助WordNet提供的查询接口量化单词之间的距离 $L(w_1, w_2)$ 和深度 $D(w_1, w_2)$，单词之间的距离越小，深度越深，单词间的语义一致性程度就越高。综合单词之间的距离和深度，挖掘不同实体内部单词对之间的相同点和不同点，然后将共同特征占总特征的比重作为衡量不同实体单词对之间的相似度。其相似度建模框架如图3所示。

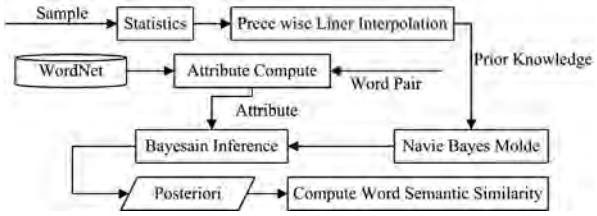

图3　单词语义相似度建模[21]

将单词 w_i 和 w_j 的共同特征量记为 V_1，不同特征量记为 V_2，则单词之间的语义相似度计算如下：

$$Sim(w_i, w_j)=\frac{\alpha V_1}{\alpha V_1+\beta V_2}$$

$$\alpha=\sum_i LW(i)\sum_j DW(j)$$

$$\beta=\sum_i (1-LW(i))\sum_j (1-DW(j))$$

其中，α 和 β 为调整因子。

在利用朴素贝叶斯模型度量单词语义相似性度的过程中，首先依据训练集数据生成不同实体的距离及深度的均值函数 $LW(i)$ 和 $DW(j)$，如图4所示，并利用均值函数计算实体在不同类别（即不同的词语序列）C 下的条件概率分布 $P(L(w_i, w_j)|C)$ 和 $P(D(w_i, w_j)|C)$，同时计算调整因子 α 和 β，最后针对特定的单词对 w_i 和 w_j，计算在不同类别 C（即不同的词语序列）下 $L(w_i, w_j)$ 与 $D(w_i, w_j)$ 分别取得的条件概率，再根据条件概率量化单词对 w_i 和 w_j 的语义相似性。其实体抽取过程与基于距离的实体方法类似，利用分类器来获得最优的划分结果。

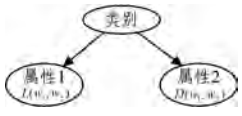

图4　基于单词语义的实体分析方法[21]

3.2　关系抽取

3.2.1　基于有监督学习的关系抽取

基于监督学习的关系抽取方法主要包括：基于特征的方法、基于核函数的方法和基于知识库的方法。

1）基于特征的方法

首先抽取与样本相关的特征并形成特征向量，然后运用

机器学习的方法训练关系抽取模型。基于 SVM 的实体关系抽取[22]是一种典型的基于特征向量的机器学习算法。基于 SVM 的实体关系抽取的关键在于对实体关系的准确描述，不仅要对实体关系的类别有准确的识别能力，还要在输入文本中准确确定实体的边界。其关系抽取流程如图 5 所示。

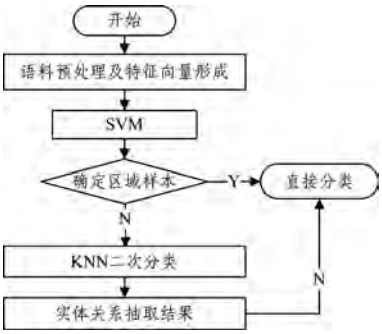

图 5 基于 SVM-KNN 的实体关系抽取流程[22]

目前，该算法只考虑一个句子中不同实体之间的关系，而不考虑多个句子之间的实体关系。根据对训练数据的分析，实体的类型对确定实体关系的类型有较大的作用；另外，两个实体在上下文中出现的顺序（前后、包含/被包含）也是实体关系识别的重要因素；最后，位于描述实体关系周围的 w 个词也属于较好的特征。在给定数据集的任意实例中，两者之间的关系与处于实体关系描述词周边的第 k 个位置的词和词性有着密切的关系。将由每个句子实例的属性特征构成的特征向量作为 SVM 支持向量机的输入，将实体关系的类别作为预测输出，完成对句子中实体关系的抽取。

2）基于核函数的方法

首先，获得两实体所在句子的结构特征，形成结构树并计算它们之间的相似度；然后，训练支持核函数的分类器进行关系抽取。

基于树核函数的关系提取[23]是基于核函数中的典型方法，其中需要解决的核心问题是将实体之间的关系结构化。为了提高实体关系抽取的性能，在实体关系的抽取过程中考虑实体相关的语义特征，生成一棵包含实体大类 TP、子类 ST 以及引用类型 MT 等信息的语义信息扩展树（Semantic Extended PT，SEPT），如图 6 所示。

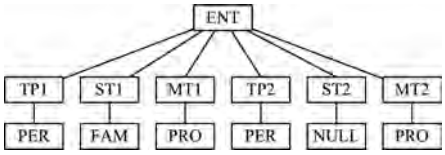

图 6 实体语义信息扩展树 SEPT

同时，去除输入文本信息中冗余的结构信息，如修饰语、并列结构等，生成去除修饰语冗余树（Modification Removed PT，MRPT），提高实体关系识别的准确率，如图 7 所示。

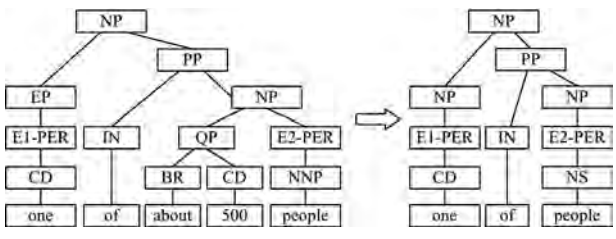

图 7 去除修饰语冗余树

最后，将生成关系时忽略的所有格后面的中心语信息补充完整，生成所有格扩充树（Possessive Extended PT，PEPT），如图 8 所示。

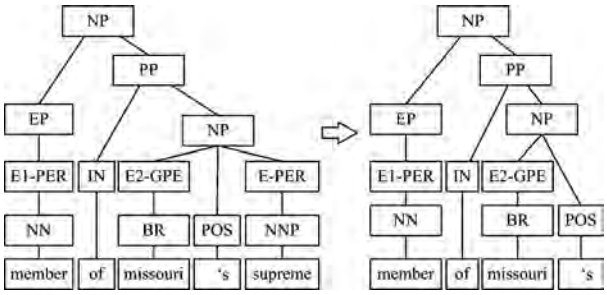

图 8 所有格扩充树 PEPT

3）基于知识库的方法

通过外部的知识库（如维基百科等）进行实体关系的抽取。

关系抽取是构建知识库的重要组成部分。知识库建成后，利用已有知识库中蕴含的实体关系，在大量未标注的语料上抽取实体之间的关系，从而达到更新已有知识库的目的。Mintz 等[24]充分使用 Freebase 中已经提取出来的实体关系，训练关系抽取的分类器，从维基百科文章中抽取出新的实体及实体之间的关系。

在基于知识库的方法中，关系抽取的结果对知识库的完备性提出了较高要求，且知识库在处理跨领域问题时效果欠佳。

4）基于开放域的关系提取

基于开放域的关系提取主要是利用高可扩展性的方式，跨领域地从提供的信息中抽取实体与实体之间的关系。

目前第一代开放域的关系提取方法已经发展到第二代。第一代的关系提取方法，如 POS（Part-of-Speech）[25]模型，利用非文本向量化的特征建立线性链式模型，从而对实体关系进行提取。但是，第一代开放域关系的提取方式无法突破交叉领域（如新闻、博客、百科等）在大规模数据量上对关系提取的需要。第二代的是基于开放域关系提取模型，如 TextRunner[26]，StatSnowBall[27]，WOE[28]，Re Verb[29]等。大多基于开放域的模型都包括以下 3 个主要步骤：1）中等程度的分析；2）学习模型的建立；3）可视化的展示[30]。

第一代和第二代信息提取方式主要存在的区别如表 1 所列[30]。

表 1 第一代和第二代信息提取方式主要存在的区别

	第一代提取方式	第二代提取方式
输入	句子和被标记的关系	句子
关系	提取指定关系	随机发现
提取	指定的关系	相互独立的关系

3.2.2 基于半监督学习的关系抽取

典型的基于半监督学习的关系抽取方法是种子模式的自扩展。先人工标注一定数量的实体，再将这些实体实例作为初始种子，使用自动训练的方法学习已有实例中的模型，并对一定量的语料进行学习；当种子集合达到一定规模时，即可利用从种子集合中抽取出来的种子集合的模式进行关系提取[31]。

Sergey[32]利用种子模式的自扩展，将其进行扩展，在书籍和作者的关系提取中表现出了较高的准确性。后来，Agichtein[33]对关系提取的实体进行标注，通过限定实体的类型达到提高关系提取的准确性和灵活性的目的。

1）基于弱监督学习的实体关系抽取[33]。不同领域的知识库给基于弱监督学习的实体关系抽取提供了基础，从知识库中抽取固定结构的关系三元组[34]，即〈主体，关系，客体〉，同时抽取出具有实体关系的上下文文本信息用于建立训练集。在用训练集学习分类器的过程中，需要对抽取的实体关系进行特征定义，采用 n-gram 特征可以有效捕捉局部范围内部词语之间的关系。令 $n=1,2,3$：

1-gram：1 个词语＋词性（$word_i/pos_i$）；

2-gram：2 个连续词语＋词性（$word_i/pos_i$，$word_{i+1}/pos_{i+1}$）；

3-gram：3 个连续词语＋词性（$word_i/pos_i$，$word_{i+1}/pos_{i+1}$，$word_{i+2}/pos_{i+2}$）。

定义实体关系特征后，采用基于朴素贝叶斯的句子分类器[35]来解决由于利用固定的实体关系从知识库中抽出的训练语料数量不能够满足弱监督学习算法的学习要求从而导致特征提取不够充分的问题，能从未标注的数据中获得更多的训练语料。

2）基于改进信息增益的关系抽取算法[36]。黄卫春等人基于弱监督的方法，根据人物关系自身的特点，提出了一种基于信息增益的中文人物二元关系的抽取算法。

根据关系描述词，即关系特征计，算每个实体对候选集在上下文信息的信息增益。在有多个候选实体对时，则取信息增益值最大的实体对与其余所有候选实体对进行模糊匹配，找到每两个实体和其各自属性共同出现的实体的信息增益值进行匹配，当所有的匹配结束时则完成了实体对关系的抽取，如图 9 所示。

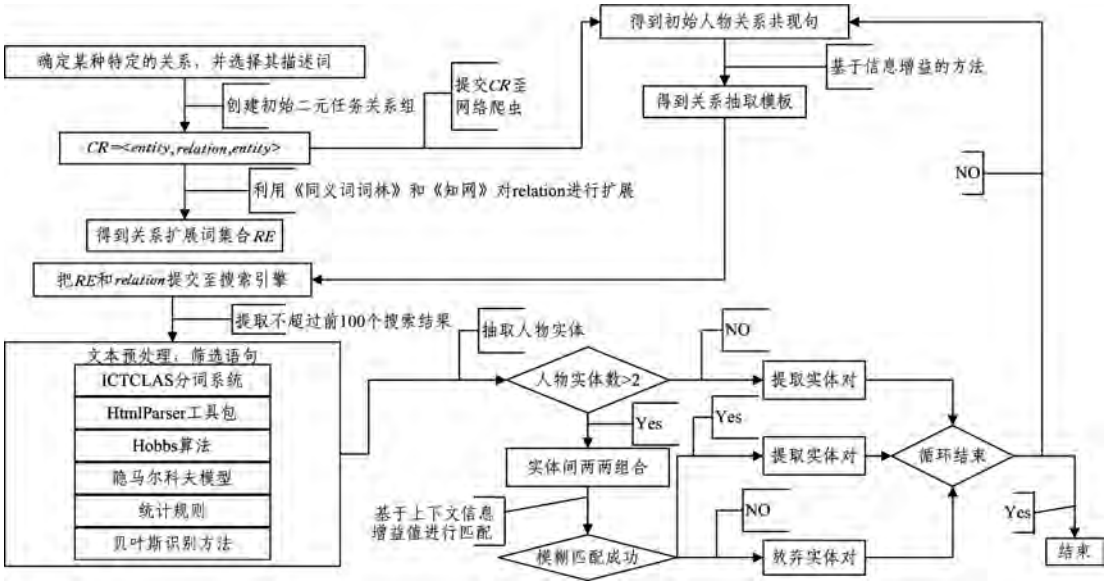

图 9　基于信息增益的人物关系提取方法[36]

3.2.3　基于无监督学习的关系抽取

2004 年，Hasegawa 等[37]首次提出了一种基于无监督学习的实体关系抽取方法。首先将实体及其上下文收集起来，然后将实体的上下文信息作为实体的特征进行全连通聚类，最后连接具有相同关系的实体，再根据实体所具有的特征明确实体之间的关系类型。但是，这种方法会受到关系出现次数的直接影响，出现次数较少的关系往往会被忽略。为了改进无监督的关系提取方法，Hasegawa 等[38]提出了 StatSnowBall 模型，该模型不同于传统的面向开发域的信息提取框架，可自动产生抽取器的模板。在 StatSnowBall 模型的基础上，又有研究者将实体识别加入关系提取中，有效地提高了关系提取的准确性和召回率。

当前，无论是半监督式或强监督式的实体关系抽取方法，都需要建立关系类型体系作为抽取模型的标签进行学习。为了避免建立复杂的关系类型体系，使用两个实体的上下文的关键信息作为实体关系[39]。同样地，使用关系三元组来描述实体关系，即〈$entiry_1$，$relation$，$entity_2$〉。为了挖掘三元实体关系侯选组以便进一步筛选，使用如下两个指标。

1）距离。距离用来衡量两个实体之间的词距。随着词距的增加，候选关系三元组的数量也会增多，需要给定两个实体的关系出现的最大词距 $maxDistance$。同时，两个实体之间还会存在其他实体会干扰主要目标实体关系抽取，因此还需给定两个实体之间的最大其余实体个数 $maxEntityDistance$。

2）关系指示词。将出现在两个实体的上下文的名词和动词作为两个目标实体的候选指示词，为下一阶段的筛选奠定基础。

由于挑选的候选关系指示词存在较大的噪声，因此需要对候选关系指示词进行再抽取。从信息论的角度来说，信息熵能够衡量某个关系特征在三元实体对中的信息量，故利用信息增益计算每个关系指示词对实体关系类型的区分度，区分度越高的词语越有可能是两个实体间的实体关系类别。

3.2.4　基于实体相关性的关系抽取

关联数据在搜索引擎中有着重要的意义，这主要是因为网络搜索引擎可以吸引有兴趣的用户参与。给用户一个机会探索实体的信息，探索的结果与他们最初的意图有关，例如谷歌、雅虎。

现有的知识图（如 DBpedia 和 Freebase）是由几个实体通过一个给定的实体关系显式地连接在一起。然而，实体可以连接许多不同的实体，由于没有任何量化的连接强度产生的知识基础，因此形成了一个庞大的连接网络，导致其很难跟踪和遍历。

Nitish 等[40]提出了实体关联图（ENRG），其为一个给定的实体提供了相关实体的排名，同时利用 DBpedia 类型对齐进行聚类。此外，ENRG 可以发现两个实体之间的内在关系

的出处并探索没有明确地包含在可用的知识库中的弱关系。

实体关联图(ENRG)通过计算各实体对之间的关联度来构建实体之间的关系。利用维基百科的分布式语义实体关联性(diser)[13]可以建立一个实体语义分布的高维概念空间。diser以每一篇维基百科文章为载体,生成一个高维向量尺寸,从而确定目标实体。简单地计算了它们对应的两个实体之间的余弦值[41]得分diser向量来作为两个实体之间的语义关联性,从而生成检索相关的维基百科概念的列表,并根据给定的实体的关联性得分对它们进行排名。

此外,还有研究者提出了一种基于知网和术语相关度的关系抽取方法[42]。实体关系的抽取流程是利用句法规则提取实体的上下文信息,并利用自然语言处理技术和互信息的方式来计算不同术语之间的相关度,使用提取的关键特征作为关键词,在知网语义关系框架中定位关系所处的位置,最后为实体关系指明具体的标签。

两个术语的互信息表示两个术语在概念上的相似性和联系的紧密程度,用如下公式表示:

$$mi(x,y)=\frac{p(x,y)}{p(x)p(y)}$$

其中,线 x 和 y 表示两个不同的术语概念,$p(x,y)$是两个术语在某语境中的共现频率,$p(x)$和 $p(y)$则代表两个术语在语境中独立出现的频率。

在自然语言处理领域的上下文假设,即若两个概念相近的术语,则上下文特征相似,对于术语的上下文,通过句法分析抽取对不同的上下文特征模式进行抽取,都会抽取出一系列的词组集,用该词组集定义该术语的上下文特征向量,则两个术语之间的相似度用余弦相似度来计算。

$$\cos(x,y)=\frac{xy}{|x||y|}$$

综合两个术语的互信息与上下文特征,计算两个术语的相关度:

$$correlation(x,y)=\begin{cases}\cos(x,y), & \text{if } mi(x,y)>\alpha\\ 0, & \text{otherwise}\end{cases}$$

其中,α 是预先设定的阈值,只有当计算出的相关度大于该阈值时才说明两个术语之间存在一定程度的相关性,这有效避免了互信息在两个术语处于低频时的突变缺陷。

3.3 知识融合

关联图谱中的信息来源广泛,实体覆盖范围广,存在着信息重复、实体信息不全、实体之间关联关系不明确等问题。知识融合是在关联图谱中输入信息,进行更高层次的知识组织,使得异构数据得到清洗、加工、整合等清洗和提取处理,并将得到的信息与人的思想相融合,从而形成高质量的关联图谱。

3.3.1 基于贝叶斯的知识融合

完全依靠模型自学习的方式构建Bayes网络,会因为数据学习算法的搜索空间过于庞大而导致搜索效率很低[43]。因此,需要收集领域专家知识[44],根据这些领域专家知识得到被探测的实体的先验知识,同时结合相关理论,确定数据之间的关联关系。由于融合背景和融合目的的不同,可以通过不同的数学模型计算从知识来源处获得输出数据的条件概率[45]。

知识融合的本质是一个多维的决策问题,利用Bayes准则,可以利用各个输入知识的先验概率和后验概率来确定决策的结果。Bayes融合的核心是最大后验概率(Maximum a Posteriori,MAP)的利用。

Bayes方法的核心是构建Bayes网络,而Bayes网络通常用一个二元组 $B=(G,\Theta)$来描述。其中,知识融合框架中的定性知识部分用有向无环图 $G=(X,E)$来描述,$X=(X_1,X_2,\cdots,X_n)$表示所描述领域的变量集合,而 A 为网络节点之间的有向弧集合,有向弧 a_{ij} 描述了两个领域变量之间的依赖关系,为因果推理提供基础;$\Theta=(\theta_1,\theta_2,\cdots,\theta_n)$表示网络中的条件概率的参数集合,是对知识部分的定量描述,$\theta_i=P(X_i|\pi(X_i))$为给定领域知识 X_i 的父节点的状态集合下领域知识 X_i 的条件概率分布。因此,Bayes网络能够用图和条件概率的参数集合来表述不同领域的联合概率分布:

$$P(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P(X_i|\pi(X_i))$$

知识融合就是学习给定数据集 $C=\{X_1,X_2,\cdots,X_n\}$时的最优Bayes网络结构[46]。设 S^h 表示对Bayes网络结构的假设,即找到使 $p(S^h|C)$最大化的网络结构。通过Bayes定理,可以推导出:

$$p(S^h|C)=\frac{p(S^h,C)}{p(C)}=\frac{p(C,S^h)}{p(C)}$$

当前,基于独立性的检验方法与基于“评分+搜索”的算法是Bayes网络结构的主流学习算法。基于独立性的检验方法的计算量小,收敛速度快,但是网络结构的准确性较差。而基于“评分+搜索”的算法的准确率高,但随着网络规模的增加,搜索空间呈指数级增长,因此增加专家领域经验来确定部分网络结构能够大大缩小搜索空间,从而除去不必要的网络结构。为了避免不同专家领域的主观性,采用DEMPSTER合成法则对领域知识进行融合,具体融合思路如下:

1)基于潜在的链接关系,建立有向完全图,代表不同领域变量之间存在相互依赖的关系;

2)选择 m 个专家对潜在的完全有向关系图中的任何一组变量 X_i 与 X_j 的因果关系进行可信度分配;

3)采用DEMPSTER合成法则对领域知识进行融合,同时决定不同变量之间的因果关系,对不必要的网络结构拓扑进行裁剪,使之能够近似最优网络结构;

4)搜索状态空间,计算每个节点的当前评分,当所有节点的评分不再发生变化时,最优Bayes网络的拓扑结构构建完成。

基于专家知识融合的Bayes网络结构学习方法在一定程度上可以解决由于数据驱动学习的Bayes网络搜索状态空间过大造成的效率过低的问题,利用专家的领域知识能够识别网络中原有的因果关系,通过不同领域专家的意见,一方面避免了主观性,另一方面能够缩小搜索空间,同时利用证据理论,综合多个领域专家的意见,提高了知识融合的效率。

3.3.2 基于证据组的知识融合

根据知识源提供的数据和先验知识,利用数据挖掘、实验测试等手段获取识别框架上的基础概率分配函数或者信任函数,为后续的证据组合[47]处理提供不确定性的数学描述。证据组合按照基础概率分配的基本组合规则,导出其他组合规则,实现知识融合[48]。

根据融合得到的基础概率分配,选择最大支持度的假设作为最优判断。根据基础概率分配计算信任函数和合理性函数,利用这两个函数可以描述证据对于可能性的支持度,并将它们作为上下界而构成不确定性区间。

研究者提出了一种基于可靠度的直觉模糊数排序方法的模糊多属性决策的证据可靠性评估方法。利用该方法对时域信息序列中相邻的时间节点的证据可靠性进行评估，再基于证据折扣运算和 Dempster 证据组合规则提出一种基于复合可靠度的时域证据组合方法[49]。

基于 D-S 证据理论的知识融合[50]算法主要包含以下 3 个过程[51]。

1）确定知识融合问题的识别框架，同时给定不同知识在识别框架上的基础概率分配。令 F 表示成功，$\bar{F}$ 表示失败，则知识识别框架为：

$\Theta=\{F,\bar{F}\}$

在该知识框架下的基础概率分配为：

$m_j(F)=p_F(1-m_j(\Theta))$

$m_j(\bar{F})=(1-p_F)(1-m_j(\Theta))$

利用信息熵计算知识的不确定性：

$m_j(\Theta)=-k[p_F\ln p_F+(1-p_F)\ln(1-p_F)]$

其中，$k\in(0,1)$为调节因子。

2）知识融合过程即计算证据组合后知识识别框架下的基础概率重分配问题。设有 M 个证据在 Θ 下的基础概率分配分别为 $m_1,m_2,\cdots,m_M$，则其计算公式如下：

$$m(F)=\frac{\sum\limits_{\cap X_{ji}=F}\prod\limits_{j=1}^{M}m_j(X_{ji})}{1-\sum\limits_{\cap X_{ji}=F}\prod\limits_{j=1}^{M}m_j(X_{ji})}$$

$$m(\bar{F})=\frac{\sum\limits_{\cap X_{ji}=\bar{F}}\prod\limits_{j=1}^{M}m_j(X_{ji})}{1-\sum\limits_{\cap X_{ji}=\bar{F}}\prod\limits_{j=1}^{M}m_j(X_{ji})}$$

3）根据知识融合的结果，通过选择合适的判别准则识别方案的支持度，并作出相应决策。

3.3.3　基于模糊集理论的知识融合

模糊集理论放宽了概率论公理中的限制条件，具有更广泛的应用空间。基于模糊集理论的模糊逻辑和模糊积分，很好地解决了知识融合的问题。

在求得各种知识源组的模糊测量值后，将此值作为知识源组合的可靠性。利用模糊积分计算出不同知识源提供的最大一致度。模糊积分中的隶属函数根据实际情况而定，实现知识源测量值的模糊化。根据融合后得到的结果，判断最优的假设，选择最大值或者最小值作为评价标准[52]。

基于模糊积分的知识融合的实现思路如下：首先，计算模糊测量 g_λ，模糊测量是为了度量不同知识库进行组合的可靠性，其值通过专家先验知识得出；然后，根据如下公式计算不同知识库下的模糊测量值，其中参数 λ 根据具体问题附加的特定规则选择不同的值。

$$g_\lambda(I)=\frac{1}{\lambda}\left(\prod_{i\in I}(1+\lambda g_i)-1\right)$$

1）清晰化。清晰化过程是将融合的信息转换成人们能够理解的概念。例如，将一个已经经过知识融合后的处于[0，1]间的数值映射到它本身的特征空间，即恢复出数据原有的知识特征。

2）基于改进型混合粒子群优化算法的模糊知识融合。此融合算法过程包括 3 个部分：构造合适的模糊知识库编码方式、模糊知识融合算法、融合知识库的评估。目前很多学者将多模糊知识库的融合问题转化为多目标优化问题[53]。改进型的混合粒子群可有效避免传统粒子群优化算法容易陷入局部最优的缺陷，提高了算法的效率和准确性。

模糊知识融合算法是知识融合的核心，算法过程包含两个步骤：模糊知识库编码与基于粒子群算法的模糊知识融合。在编码阶段，粒子由 3 部分组成：第一部分对模糊规则的输入采用二进制编码，第二部分对模糊函数采用实数编码，第三部分对模糊规则的输出采用实数编码。

对模糊加权融合模型进行研究，将隶属度和融合权值进行比较，确定利用模糊贴近度方法求得的融合权值具有高实用性、可靠性和精确性。

4　关联图谱的垂直应用

关联图谱将大量、异构、动态变化的数据有效地组织在一起，提供了一套快速且高效的数据管理、分析与应用的结构，因而被广泛地应用在搜索引擎、基因序列分析、图书管理、智能制造等领域，成为了这些领域发展的动力。

4.1　语义搜索

语义搜索的概念最早是由 Tim Berners-Lee 于 2001 年在 *Scientific American* 上发表的一篇文章中提出来的。

语义搜索引擎中的核心是知识图谱。维基百科中对知识图谱的定义如下：知识图谱是 Google 用于增强其搜索引擎功能的知识库[54]。本质上，知识图谱是一种解释了实体之间关系的语义网络[55]。

然而，仅依靠语义搜索让搜索结果更加符合用户的真实需求是不行的。对于 Google 而言，知识图谱还考虑了实体在 Web 上的声誉、实体的权威性以及其他人对该实体的信任程度。为此，还需要其他相关图谱的协助，包括以下类型。

1）社交图谱[56]（Social Graph）。社交图谱中描述了一个实体和其他实体在 Web 上可以找到的所有联系。这些联系主要包括一方到另一方的直接链接、个人的联系、共享的群组动态以及从内容分享角度而言更加直接的互动。

2）链接图谱[57]（Link Graph）。链接图谱包含了 Google 通过 Page Rank 所积累下来的所有信息，主要包括一个网页以何种方式链接到其他网页，以此作为两个网页实体之间关系的重要性和一个网页对另一个网页重视程度的决定因素。

3）交流图谱[58]（Engagement Graph）。交流图谱描述了实体与实体之间交流的级别，包括评论、交流、+1、赞和踩等。社交图谱和交流图谱的区别在于互动程度的不同。

社交图谱、链接图谱、交流图谱和知识图谱这 4 个图谱作为一个整体，每天协助 Google 解决大约十亿个用户问题。

4.2　图书馆书目图谱

图书馆的书目索引本质上满足关联数据的特性。关联数据[59]是国际互联协会推荐的一种规范，用来发布和链接各类数据、信息和知识。2006 年，Tim 基于语义网首次提出了关联数据的概念，并总结了对实体进行语义描述及其相互关系需要满足的特性：

1）可以作为任何事物的标志、名称；

2）全局唯一的名称；

3）访问标志名称时，可以获得有用的信息；

4）访问标志名称时，有途径可以获得更多的相关信息。

2008 年,瑞典基于关联数据的形式,开始共享链接数据,发布了 LIBRIS 国家书目(libris. kb. se),并利用 LIBRIS 和 DBPedia 之间的关系将两个数据库的书目进行关联,形成了多个图书馆书目关联的数据集[60-61]。使用的链接数据信息包括英国国家书目(bnb. data. bl. uk)和 Open Library。这两个数据库均使用了链接数据的相关格式(RDF,JSON 和 Turtle)来记录书籍和作者的信息,极大地方便了信息的重用[62]。

2010 年被 Antoine 称为图书馆关联数据元年,源于多种基于关联数据的词表和 KOS 涌现,积累了大量有质量的内容实体以及实体和实体之间的关系。此后涌现出了大量相关研究[63],大量的图书馆开始考虑如何将已有的 MARC 数据转化为链接数据。如图 10 所示,截止到 2014 年 4 月,基于链接数据原则将图书馆中以文档为中心的信息转换原则转换为以数据为中心的对象,从而创建了一个新的图书信息世界。

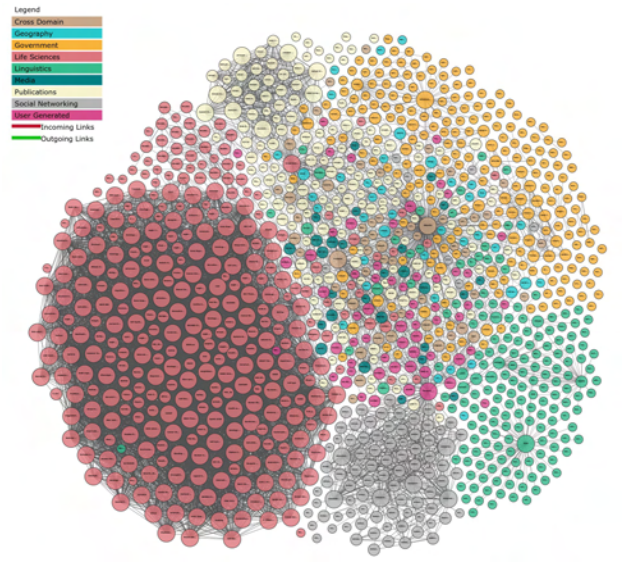

图 10 截止到 2017 年 4 月开放出版物的链接数据集[64]

4.3 医学关联图谱

控制生物信号转导、正常代谢和基因调控的生物网络在许多生命体的基本机制中发挥着关键的作用,如生长、发育和环境反应。在这些生命网络中包含着成千上万的基因、蛋白质、化合物和 RNA 等,它们共同进行着复杂的相互作用。

元素之间的相互作用即构成了元素之间的相互关系。基于生物学功能,将这些关系分为:

1)蛋白质-蛋白质相互作用(PPI);

2)化合物-蛋白质相互作用(CPI);

3)转录基因(TF)或小的非编码 RNA(sRNA)对于游靶基因的调节作用;

4)酶对于其底物/产物化合物的化学反应;

5)转运蛋白与其底物之间的关系。

这些元素是医学关联图谱中的基本构成元素,元素与元素之间复杂的相互作用即为关联图谱中的边,从而建立起了生物世界的关联图谱——生命网络的基本信息。

大量已发表的论文、高通量数据集和第三方数据库都基于生命元素以及生命元素之间的相互关系来构建关联图谱,并取得了突破性进展[65]。

在蛋白质之间的相互作用方面,已有研究利用拟南芥蛋白质相互作用构建关联图谱(AtPIN)[66]。通过在拟南芥蛋白质关联图谱中进行实验验证或者计算预测,得到了关于拟南芥 PPI 的信息。

在蛋白质与基因之间相互作用方面,利用关联图谱进行基于计算的预测技术,同时为理解关联图谱中元素的交互作用提供了有价值的线索。利用运输者分类数据库(TCDB)[67]承载每个转运蛋白家族的标准序列,能够通过结构域连接将转运蛋白与相应的底物连接,或基于同源的预测。微 RNA(miRNA)靶预测工具以高可信度连接 miRNA 和潜在靶基因[68]。

基于基因和基因组的关联图谱,有利于对基因所控制的相关性状以及基因分组进行研究。在公共存储库中提供以太字节和 PB 级转录数据,例如 ArrayExpress,NCBI(国家生物技术信息中心)[69]以及基因表达综合征数据库等的建立,使研究人员能够通过表达分析对具有相似表达模式的基因进行分组。

因此,综合基于生物功能的这些异质生物相互作用的关联图谱,无论是基于网络路径还是网络水平,都将是研究基因功能的有价值的资源。

5 关联图谱研究面临的挑战

5.1 实体抽取方面面临的挑战

实体抽取的挑战主要源于两个方面:自然语言的复杂性和实体抽取任务的开放性。

自然语言的复杂性主要体现在 3 个方面:

1)自然语言的歧义性,即在不同的语言环境中,相同的词语会有不同的含义;

2)自然语言的多样性,即相同的意思,可以用不同的方式表达;

3)自然语言中特有的语法结构。

实体抽取任务的开发性主要由两个部分组成:

1)实体抽取任务面向的对象是多种多样的,即实体信息需要从不同数据中抽取得到;

2)使用实体抽取的网页开发,不同的网站使用的信息抽取技术往往是不相同的。

目前,普遍采用弱监督或者半监督的方式,但是这些处理方式仍无法解决由上述两种问题带来的挑战,例如:自然语言在不同的应用场景面对的难点差别很大;不同文化环境下的自然语言存在着较大的文化差异;抽取不同语言中的相同实体对象具有较大的难度。因此,解决实体抽取仍是一项长期而艰巨的任务。

5.2 知识融合方面面临的挑战

关联图谱的发展与更新需要不断融合新知识。在过去的 10 年中,数据的产生方式不断扩充和发展。虽然关联数据提供了统一的数据格式,但是如果在应用中直接使用关联数据,会出现大量由数据质量导致的问题。

1)数据不完整;

2)利用关联数据建立的索引无法检索到相关数据;

3)相同操作获得的数据是不同的;

4)不同数据源的数据不一致;

而且在大数据的情况下,数据呈现出一些新的特征。

1)多元性:数据的类别多样、数据内容维度多样以及数据和知识之间的“立体关系”。

2）演化性：数据随着时间的变化而变化，因此知识融合体现出动态演化的特性。

3）真实性：数据源于真实的生活环境，直接原因是自然语言的多样性。真实性增加了知识融合的复杂性，但也正是真实知识的佐证为演化和融合提供了支撑。

大数据下的知识融合，可以有效地从海量数据中提取出高品质的信息，最大程度地发挥大数据的作用。但是，大数据下的知识融合是一个跨领域的研究问题，已有的融合方法已经无法适应，这需要各领域的研究人员广泛参与，并提出更深更广的研究思路和方法，从而实现更深层次的融合。

5.3　应用层面临的挑战

关联图谱已被广泛运用于搜索、推荐、智能问答等多个领域，但是这些领域中建立的关联图谱不够完善，仍存在着亟需解决的问题。利用关联图谱进行智能语义搜索，需要解决结果展示以及优化搜索结果的问题；利用关联图谱进行音乐推荐，需要构建和完善曲库；利用关联图谱进行智能客服，需要解决自动客服帮助有限的问题；利用关联图谱进行智能问答操作，需要更多的尝试和完备的学科知识库等。

同时，在很多重要的领域中，并没有建立相应的关联图谱，尤其是在现代军事领域中，将装备研制作为军用装备产品寿命的起始，结合军用装备的特点实施标准化要求，在不同的军用装备产品的生产周期中，需要进行的论证、研制、实验和使用等步骤应遵循的标准是不同的；每一个技术活动是否存在相应的标准作为依据；每一个标准中的条目可以解决的决策问题等。这些问题的解决需要建立基于标准的关联图谱，这些关联图谱被广泛应用在具体的装备或者工作上，为其找到应贯彻的标准以及标准中具体对应的条目。关联图谱的建立对实现武器装备的先进性、实用性和经济性有着巨大的作用。

关联图谱具有应用价值和经济价值，我们应最大可能地应用关联图谱的价值，但是，关联图谱对于输入数据的利用和理解还很有限，利用三元组的方式描述关联图谱的表现力不足，关联图谱应用“深”比“广”更加困难的问题还需解决。

结束语　本文对关联图谱的定义、架构、关键技术和垂直应用等方面做了全面而深刻的综述，深入地阐述了关联图谱构建中的实体抽取、关系抽取和知识融合方面的核心技术；概括地介绍了关联谱图在搜索引擎、图书馆书目管理以及医学方面发挥的巨大作用；总结了目前研究关联图谱所面临的挑战，以及未来应研究的一些问题。关联图谱的重要性不仅在于它拥有强大的数据组织、知识管理和知识融合的能力，还在于它是一把利剑，能为相关科学领域拓展出一条新的有效的研究道路。在未来的数年内，关联图谱的相关研究仍是各个科研领域中的前沿研究问题。

参 考 文 献

[1] SINGHAL A. Official Google Blog: Introducing the Knowledge Graph: things, not strings[EB/OL]. http://www.mendeley.com/catalog/official-google-blog-introducing-knowledge-graph-things-not-strings.

[2] BRACHMAN R J. What IS-A is and isn't: An analysis of taxonomic links in semantic networks[J]. United States Journal of Computer, 1983, 16(10): 30-36.

[3] STEINER T, VERBORGH R, TRONCY R, et al. Adding realtime coverage to the google knowledge graph[C]// International Conference on Posters & Demonstrations Track-Volume 914. CEUR-WS. org, 2012: 65-68.

[4] WANG Z C, WANG Z G, LI J Z, et al. Knowledge extraction from Chinese wiki encyclopedias[J]. Frontiers of Information Technology & Electronic Engineering, 2012, 13(4): 268-280.

[5] ZENG Y, WANG H, HAO H, et al. Statistical and structural analysis of web-based collaborative knowledge bases generated from Wiki Encyclopedia[C]// 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01. IEEE Computer Society, 2012: 553-557.

[6] HUANG Z, CHUANG W, ONG T H, et al. A graph-based recommender system for digital library[C]// 2nd ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, 2002: 65-73.

[7] SIEK J G, LEE L Q, LUMSDAINE A. The Boost Graph Library: User Guide and Reference Manual, Portable Documents[M]. Canada: Pearson Education, 2001: 17.

[8] DAI X, LI J, LIU T, et al. HRGRN: A Graph Search-Empowered Integrative Database of Arabidopsis Signaling Transduction, Metabolism and Gene Regulation Networks[J]. Plant and Cell Physiology, 2016, 57(1): 12.

[9] DIESTEL R, KR? L D, SEYMOUR P. Graph theory[J]. Oberwolfach Reports, 2016, 13(1): 51-86.

[10] WALKINGSHAW A D, ALEKANDROVSKY B L, VAN-HOFF A A, et al. Generating an Implied Object Graph Based on User Behavior: U. S. Patent Application 14/691, 370[OL]. https://patents.google.com/patent/US20150227563.

[11] NARAYANAN S, NANDAGOPAL V, SUN E. Automatically generating nodes and edges in an integrated social graph: U. S. Patent 9, 002, 898[P]. 2015-4-7.

[12] 李涓子. 知识图谱：大数据语义链接的基石[EB/OL]. http://www.cipsc.org.cn/ kg2/. LI Juan-zi. Knowledge graph: the foundation for big data semantic link[EB/OL]. (2015-02-20). http://www.cipsc.org.cn/kg2.

[13] 刘峤，李杨，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.

[14] 徐增林，盛泳潘，贺丽荣，王雅芳. 知识图谱技术综述[J]. 电子科技大学学报，2016，45(4)：589-606.

[15] 耿霞，张继军，李蔚妍. 知识图谱构建技术综述[J]. 计算机科学，2014，41(7)：148-152.

[16] 刘显敏，李建中. 基于建规则的 XML 实体抽取方法[J]. 计算机研究与发展，2014，51(1)：64-75.

[17] Wikimedia Foundation Inc. simple API for XML [EB/OL]. [2013_05_07]. http:// en. wikipedia. org/ wiki/simpl_API_for_XML.

[18] 黎玲利，高宏. 基于距离度量的实体识别算法[J]. 智能计算机与应用，2014，4(6)：61-63.

[19] 刘雪莉，王宏志，等. 基于实体的相似性连接算法[J]. 软件学报，2015，26(6)：1421-1437.

[20] 贾真，何大可，杨燕，等. 基于弱监督学习的中文网络百科关系抽取[J]. 智能系统学报，2015，10(1)：113-119.

[21] 王俊华，左万利，闫昭. 基于朴素贝叶斯模型的单词语义相似度度量[J]. 计算机研究与发展，2015，52(7)：1499-1509.

[22] 刘绍毓，周杰，李弼程，等. 基于多分类 SVM-KNN 的实体关系抽取方法[J]. 数据采集与处理，2015，30(1)：202-210.

[23] 刘晓勇. 一种基于树核函数的半监督关系抽取方法研究[J]. 山

东大学学报(工学版),2015,45(2):22-26.
[24] MINTZ M,BILLS S,SNOW R,et al. Distant Supervision for Relation Extraction Without Labeled Data[C]// Joint Conference of the Meeting of the Acl & the International Joint Conference on Natural Language Processing of the Afnlp: Volume. Association for Computational Linguistics,2009:1003-1011.
[25] CHENG A,XIA F,GAO J. A comparison of unsupervised methods for Part-of-Speech Tagging in Chinese[C]// 23rd International Conference on Computational Linguistics:Posters. Association for Computational Linguistics,2010:135-143.
[26] BANKO M,CAFARELLA M J,SODERLAND S,et al. Open Information Extraction from the Web[C]// International Joint Conference on Artifical Intelligence. 2007:2670-2676.
[27] ZHU J,NIE Z,LIU X,et al. StatSnowball:a statistical approach to extracting entity relationships[C]// Proceedings of the 18th international conference on World wide web. ACM,2009:101-110.
[28] WU F,WELD D S. Open information extraction using Wikipedia [C]// Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,2010:118-127.
[29] FADER A,SODERLAND S,ETZIONI O. Identifying relations for open information extraction[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics,2011:1535-1545.
[30] ETZIONI O,CAFARELLA M ,BANKO M. Open information extraction[OL]. https://doi. org/10. 1142/S2425038416300032
[31] BATISTA D S,MARTINS B,SILVA M J. Semi-supervised bootstrapping of relationship extractors with distributional semantics [C]// 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon,Portugal,2015:499-504.
[32] BRIN S. Extracting patterns and relations from world wide web [C]// WebDB Workshop at 6th International Conference on Extending Database Technology (WebDB'98). 1998:172-183.
[33] AGICHTEIN E,GRAVANO L. Snowball:Extracting relations from large plain-text collections[C]// fifth ACM conference on Digital libraries. ACM,2000:85-94.
[34] 罗甫林,黄鸿,刘嘉敏,等. 基于半监督稀疏流形嵌入的高光谱影像特征提取[J]. 电子与信息学报,2016,38(9):2321-2329.
[35] MADAAN A,MITTAL A,RAMAKRISHNAN G,et al. Numerical relation extraction with minimal supervision[C]// Thirtieth AAAI Conference on Artificial Intelligence. 2016.
[36] NICKEL M,MURPHY K,TRESP V,et al. A Review of Relational Machine Learning for Knowledge Graphs[J]. Proceedings of the IEEE,2015,104(1):11-33.
[37] 黄卫春,徐力,熊李艳,等. 基于信息增益的 Web 人物关系抽取[J]. 计算机应用研究,2016,33(8):2286-2289.
[38] HASEGAWA T,SEKINE S,GRISHMAN R. Discovering relations among named entities from large corpora[C]// Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004:415.
[39] ZHU J,NIE Z,LIU X,et al. StatSnowball:a statistical approach to extracting entity relationships[C]// 18th International Conference on World Wide Web. ACM,2009:101-110.
[40] 秦兵,刘安安,刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展,2015,52(5):1029-1035.
[41] AGGARWAL N,BUITELAAR P. Wikipedia-based distributional semantics for entity relatedness[C]// 2014 AAAI Fall Symposium Series. 2014.
[42] ZOBEL J,MOFFAT A. Exploring the similarity space[J]. Acm Sigir Forum,1998,32(1):18-34.
[43] TOUTANOVA K,CHEN D,PANTEL P,et al. Representing Text for Joint Embedding of Text and Knowledge Bases[C]// EMNLP. 2015:1499-1509.
[44] 刘明辉,王磊,党林阁,等. 非确定先验信息的贝叶斯网结构学习方法[J]. 计算机工程,2010,36(5):165-167.
[45] SHAW C A,CAMPBELL I M. Variant interpretation through Bayesian fusion of frequency and genomic knowledge[J]. Genome medicine,2015,7(1):4.
[46] COUSSEMENT K,BENOIT D F,ANTIOCO M. A Bayesian approach for incorporating expert opinions into decision support systems:A case study of online consumer-satisfaction detection [J]. Decision Support Systems,2015,79(C):24-32.
[47] 张振海,王晓明,党建武,等. 基于专家知识融合的贝叶斯网络结构学习方法[J]. 计算机工程与应用,2014,50(2):1-4.
[48] MARCHEGGIANI D,TITOV I. Discrete-state variational auto encoders for joint discovery and factorization of relations[J]. Transactions of the Association for Computational Linguistics, 2016(4):231-244.
[49] 韩立岩,周芳. 基于 DS 证据理论的知识融合及其应用[J]. 北京航空航天大学学报,2006,32(1):65-68.
[50] 宋亚飞,王晓丹,雷蕾 . 基于直觉模糊集的时域证据组合方法研究[J]. 自动化学报,2016,42(9):1322-1338.
[51] SHAFER G. A mathermatical theory of evidence [M]. Princeton,NJ:Princeton University Press,1976.
[52] 郭强,关欣,潘丽娜,等. 一种基于条件证据网络的多源异类知识融合识别方法[J]. 控制与决策,2015,30(12):2153-2160.
[53] 屈强,刘中晅,陈波. 基于修正倒数型距离贴近度的传感器数据模糊加权融合法[J]. 计算机工程,2016,42(5):313-316.
[54] XIE N,WANG W,MA B. et al. Research on an Agricultural Knowledge Fusion Method for Big Data[OL]. https://www.researchgate. net/publication/277962505_Research_on_an_Agricultural_Knowledge_Fusion_Method_for_Big_Data.
[55] 陈云翔,蔡忠义,张诤敏,等. 基于证据理论和直觉模糊集的群决策信息集结方法[J]. 系统工程与电子技术,2015,37(3):594-598.
[56] Wikipedia. Knowledge graph[EB/OL]. [2016-05-09]. https://en. Wikipedia. org/wiki/ Knowledge _Graph.
[57] HASNAIN A,DUNNE N,DECKER S. Knowledge Processing with Big Data and Semantic Web Technologies[R]. 2015.
[58] WALKINGSHAW A D,ALEKSANDROVSKY B L,VAN-HOFF A A,et al. Generating an Implied Object Graph Based on User Behavior:U. S. Patent Application 14/691,370[P]. 2015-4-20.
[59] LI Y,MARTINEZ O,CHEN X,et al. In a World That Counts: Clustering and Detecting Fake Social Engagement at Scale[C]// 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016: 111-120.

(下转第 35 页)

量训练构建图像的数字化字典，对大量未标记的遥感图像数据分类。

对于包含海量信息的遥感图像，如何充分挖掘信息以更贴合遥感图像分类要求成为了研究的重点。遥感图像涵盖的纹理特征、光谱特征和空间特征都可以单独作为图像分类的依据，然而图像的这些特征信息在图像分类时尚未充分利用。对提取的这 3 类特征进行多角度充分利用，并将其共同作为分类依据，结合深度神经网络的训练模型，来提高遥感图像的分类精度，成为了该领域目前的研究热点。

结束语　本文首先介绍了遥感图像分类的相关概念，分析遥感图像分类目前存在的问题，并对深度神经网络的历史、原理进行简要介绍，探讨了几种深度神经网络的原理结构；然后分别对遥感图像分类的研究现状和深度神经网络对遥感图像分类的研究现状做出阐述；最后总结了深度神经网络对遥感图像分类的发展趋势。

参 考 文 献

[1] HINTON G E，SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science，2006，313(5786)：504-507.

[2] 庞荣. 深度神经网络算法研究及应用[D]. 成都：西南交通大学，2016.

[3] HINTON G E，OSINDERO S，TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation，2006，18(7)：1527-1554.

[4] 吴正文. 卷积神经网络在图像分类中的应用研究[D]. 成都：电子科技大学，2015.

[5] LIPPMANN R P. Pattern Classification Using Neural Networks[J]. IEEE Communications Magazine，2002，27(11)：47-64

[6] 卢柳叶，张青峰，李光录. 基于 BP 神经网络的遥感影像分类研究[J]. 测绘科学，2012，37(6)：140-143.

[7] 卜晓波，龚珍，黎华. 基于遗传算法改进 BP 神经网络的遥感影像分类研究[J]. 安徽农业科学，2013，41(33)：13056-13058，13079.

[8] 胡永森，王力，吴良才，等. 加权变异粒子群 BP 神经网络在遥感影像分类中的应用[J]. 地理空间信息，2016，14(12)：37-40.

[9] 谭秀辉. 自组织神经网络在信息处理中的应用研究[D]. 太原：中北大学，2015.

[10] 杜华强，范文义. Matlab 自组织神经网络在遥感图像分类中的应用[J]. 东北林业大学学报，2003(4)：51-53.

[11] 李石华，金宝轩. 基于 Matlab 的自组织神经网络在地形复杂区遥感图像分类中的应用研究[C]// 第二届“测绘科学前沿技术论坛”论文精选. 2010：4.

[12] 尹汪宏，李朝峰，张俊本，等. 基于混合核函数的自组织神经网络遥感图像分类[J]. 计算机工程与设计，2009，30(2)：388-391.

[13] 任军号，吉沛琦，耿跃. SOM 神经网络改进及在遥感图像分类中的应用[J]. 计算机应用研究，2011，28(3)：1170-1172，1182.

[14] 瞿继双，瞿松柏，王自杰. 基于特征的模糊神经网络遥感图像目标分类识别[J]. 遥感学报，2009，13(1)：67-74.

[15] 崔曦. 神经网络及模糊算法的遥感数据分类研究[D]. 西安：西安科技大学，2008.

[16] 张强. 基于模糊神经网络的遥感影像分类研究[D]. 昆明：昆明理工大学，2006.

[17] EINEN D，ROLFE J，FERGUS R. Understanding deep architectures using a recursive convolutional network[J]. arXiv：1312.1847v2，2014.

[18] SIMONYAN K，ZISSERMAN A. Very deep concolutional networks or lager-scale image recognition[J]. arXiv：1409. 1556，Sept. 4，2014.

[19] 曹林林，李海涛，韩颜顺，等. 卷积神经网络在高分遥感影像分类中的应用[J]. 测绘科学，2016，41(9)：170-175.

[20] 邢晨. 基于深度学习的高光谱遥感图像分类[D]. 武汉：中国地质大学，2016.

[21] 王巧玉. 基于深度学习的高光谱遥感图像分类[D]. 厦门：华侨大学，2016.

[22] 刘大伟，韩玲，韩晓勇. 基于深度学习的高分辨率遥感影像分类研究[J]. 光学学报，2016，36(4)：306-314.

[23] 黄鸿，何凯，郑新磊，等. 基于深度学习的高光谱图像空-谱联合特征提取[J]. 激光与光电子学进展，2017，54(10)：180-188.

[24] 付秀丽，黎玲萍，毛克彪，等. 基于卷积神经网络模型的遥感图像分类[J]. 高技术通讯，2017，27(3)：203-212.

[25] 张日升，张燕琴. 基于深度学习的高分辨率遥感图像识别与分类研究[J]. 信息通信，2017(1)：110-111.

[26] LOPUHIN K. Full pipeline demo：poly-＞pixels－＞ML－＞poly(Version4. 0) [EB/OL]. https：//www. kaggle. com/lopuhin/full-pipeline-demo-poly-pixels-ml-poly.

[27] ZFTurbo. 0. 51276 Public LB Solution(Version1. 0) [EB/OL]. https：//www. kaggle. com/c/dstl-satellite-imagery-feature-detection/discussion/29829

[28] ChrisCC. Object-based solution for DSTL (Version1. 0) [EB/OL]. https：//www. kaggle. com/chriscc/object-based-solution-for-dstl.

(上接第 21 页)

[34] LI G Z，YANG J Y. Feature selection for ensemble learning and its application[M]// Machine Learning in Bioinformatics. 2008：135-155.

[35] PENG Y H，WU Z Q，JIANG J M. A novel feature selection approach for biomedical data classification [J]. Journal of Biomedical Informatics，2010，43(1)：15-23.

[36] CHIN A J，MIRZAL A，et al. Supervised Unsupervised，and Semi-Supervised Feature Selection：A Review on Gene Selection [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2016，13(5)：971-989.

[37] OPITZ D W. Feature Selection for Ensembles[C]// Proceedings of National Conference on Artificial Intelligence. Orlando，FL，1999：379-384.

[38] ABEEL T，HELLEPUTTE T，VAN D P Y，et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods [J]. IEEE/ACM Transactions on computational Biology and Bioinformatics，2010，26(3)：392-398.

[39] WONG H S，ZHANG S，SHEN Y，et al. A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics，2012，9(4)：1257-1263.

[40] 张靖，胡学钢，张玉红，等. K-split Lasso：有效的肿瘤特征基因选择方法[J]. 计算机科学与探索，2012，6(12)：1136-1143.

[41] JIN L L，LIANG H. Deep Learning for Underwater Image Recognition in Small Sample Size Situations [C]// IEEE Conference on Oceans. Aberdeen UK：IEEE Press，2017.

[42] HINTON G. Reducing the Dimensionality of Data with Neural Networks [J]. Science，2016，313(5786)：504-507.

[43] 孙志远，鲁成祥，史忠植，等. 深度学习研究与进展[J]. 计算机科学，2016，43(2)：1-8.