计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20150874
53(2):231-246,2016

大数据融合研究:问题与挑战

孟小峰　　杜治娟
(中国人民大学信息学院　北京　100872)
(xfmeng@ruc.edu.cn)

Research on the Big Data Fusion: Issues and Challenges

Meng Xiaofeng and Du Zhijuan
(*School of Information, Renmin University of China, Beijing* 100872)

Abstract　Data characteristics and realistic demands have changed because of the large-scale data's links and crossover. The data, which has main features of large scale, multi-source heterogeneous, cross domain, cross media, cross language, dynamic evolution and generalization, is playing an important role. And the corresponding data storage, analysis and understanding are also facing a major challenge. The immediate problem to be solved is how to use the data association, cross and integration to achieve the maximization of the value of big data. Our paper believes that the key to solve this problem lies in the integration of data, so we put forward the concept of large data fusion. We use Web data, scientific data and business data fusion as a case to analyze the demand and necessity of data fusion, and propose a new task of large data fusion, but also summarize and analyze the existing fusion technologies. Finally, we analyze the challenges that may be faced in the process of large data fusion and the problems caused by large data fusion.

Key words　big data; data integration; data fusion; knowledge fusion; data management

摘　要　随着大规模数据的关联和交叉,数据特征和现实需求都发生了变化.以大规模、多源异构、跨领域、跨媒体、跨语言、动态演化、普适化为主要特征的数据发挥着更重要的作用,相应的数据存储、分析和理解也面临着重大挑战.当下亟待解决的问题是如何利用数据的关联、交叉和融合实现大数据的价值最大化.认为解决这一问题的关键在于数据的融合,所以提出了大数据融合的概念.首先以 Web 数据、科学数据和商业数据的融合作为案例分析了大数据融合的需求和必要性,并提出了大数据融合的新任务;然后,总结分析了现有融合技术;最后针对大数据融合问题可能面临的挑战和大数据融合带来的问题进行了分析.

关键词　大数据;数据集成;数据融合;知识融合;数据管理

中图法分类号　TP391

近20年里,数据产生的方式不断在扩展,数据之间的关系变得千丝万缕,呈现出大规模数据关联、交叉和融合的局面[1-2],数据出现了如下新的特征:

1) 多元性.当下数据不仅是类型多样,更重要

收稿日期:2015-09-25;**修回日期**:2016-01-12

基金项目:国家自然科学基金项目(61532010,61379050,91224008);国家"八六三"高技术研究发展计划基金项目(2013AA013204);教育部高等学校博士学科点专项科研基金项目(20130004130001);中国人民大学科学研究基金项目(11XNL010)

This work was supported by the National Natural Science Foundationof China(61532010,61379050,91224008), the National High Technology Research and Development Program of China (863 Program) (2013AA013204), the Research Fund for the Doctoral Program of Higher Education of China (20130004130001), and the Research Funds of Renmin University of China (11XNL010).

的是数据内容的“维度”多样和知识范畴的“粒度”多样,呈现出一种多元性.它体现了数据与知识之间的立体关系,而非单纯数据类型多样,与演化性成为当下大数据的精髓,是区别于大规模数据、海量数据、或早期“大数据”(量大)的最显著特征.

2) 演化性.是指数据随时间或解释的变化而变化的特性,体现了数据的动态性和知识的演化性.比如,实体的某些属性在不同时间点可能产生变化.这就要求合理建模演化行为,保证数据一致性.它与高速性共同构成了知识的动态演化性,更加贴切地体现出现实数据的本原性,而非单纯地强调速度.

3) 真实性.主要由实体的同名异义表示和异名同义表示以及关系的变化引起.这种现象普遍存在,它们增加了理解的不确定性.真实性由演化性引起,反过来又为演化性提供了印证,只有知识得到印证才能使演化更新和融合更有意义.

4) 普适性.是指在认知范围内可以达成共识关系的特征,比如,“老师”和“蜡烛”在神经元连接上具有普适性.这种普适性发现源于知识之间隐性关联的发现,它也比信息本身的增长更有价值.这是将大数据定位到知识层面的一个独特特征.

这导致大数据集成的对象已经不单是数据,而是数据和知识的复合体,可以称之为“数据湖”(data lake),其内涵到底是什么呢?偶读了68年前费孝通《乡土中国》[3],略有所悟.费老分析总结了中国乡土社会结构,指出中国社会呈现出所谓的“差序格局”,而西方社会呈现的是“团体格局”.传统数据库结构关系单一,呈现状态犹如“团体格局”,即以单个实体为本位,实体之间的关系好比一捆柴,几根成一把,几把成一扎,条理清楚,有共同的模式可循.而当下大数据来源广泛,关系复杂,远近亲疏各不同,这种关系就好比“差序格局”,以语义主题为本位,每类实体都以自我为中心按照与其他实体的语义关系为主线结成网络,这个网络按照关系的语义紧密亲疏呈现“差序”状态,就如同湖面丢下的石子形成的水波纹依中心扩散开去.这种状态随着实体间关系的变化而动态演化,并且每个网络的大小不同,体现的语义关系也不同,蕴含的价值也不同.

数据库的“团体格局”本质上是先有模式后有数据,因此数据集成可以采用中介模式的方法(全局视图(global-as-view, GAV),局部视图(local-as-view, LAV))以自顶向下的方式实现集成.数据湖的“差序格局”是先有数据后有模式,因此需要一种自底向上的方式以一种大数据融合的方法实现集成.大数据融合即建立数据间、信息间、知识片段间多维度、多粒度的关联关系,实现更多层面的知识交互,从而聚敛出数据湖中一个个维系我们社会的“水波纹”(即语义关联的紧密程度).

本文首先分析了大数据融合的现实需求并提出大数据融合的问题,探讨了现有融合技术的发展现状,并给出大数据融合的理解,指出了大数据融合面临的挑战.

1 大数据融合的案例分析与问题

大数据融合是最大程度发挥大数据价值的一种手段,它的实现可以使人类对世界的探索和认识向新的深度和广度拓展.它不同于传统的数据集成或知识库技术,需要大跨度、深层次和综合性的研究方法.下面我们通过几个不同领域的案例分析来具体探讨这一问题的本质.

1.1 公共安全领域大数据融合案例分析

公共安全领域的数据包括结构化数据和非结构数据.其中结构化数据包括人员信息(比如人员户籍库、重点人员库等)、人员行为轨迹数据(比如飞机、火车出行数据等)、车辆信息(比如车辆购买信息、违章信息等)、电信数据(比如话单)等;非结构化数据包括网页、卡口图片、重点区域的视频监控录像等.公共安全领域数据的主要应用场景是公安办案提供线索.这种数据比较复杂,规模也较大,如中国某省会城市一小部分数据构建成图,其顶点的个数和边的个数分别达到了十亿和百亿的规模.

1.1.1 实现原理

目前采取的方案是基于超大规模复杂关联数据的管理理论建立超大规模的实体-关联图.图上的每个顶点代表自然界的一个客观对象,比如人员、物品、住所等;图上的边表示实体之间的关系.如图1所示.

这种方案总体上可以分为4步实现:1)数据治理.需要把物理上相互隔离的多源异构数据通过数据治理整合到统一的数据平台,该过程是后面3步的前提和基础.目前在实际的工程实践中采用以人工为主的操作模式.2)关系构建.这个过程需要自动地构建实体之间的显式关系和隐式关系,并存储在图数据库.隐式关系的构建借助规则或机器学习.3)可视化交互分析.系统提供强大的可视化交互分析工具,帮助用户在超大规模图上做各种分析和关系推演和比对.4)基于以上3步构建各警种的具体应用.

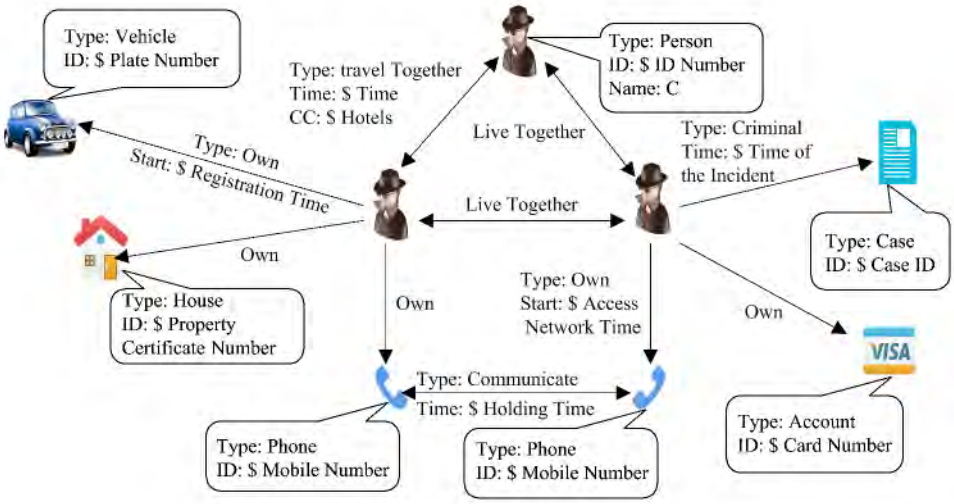

Fig. 1 An example of data fusion in the public safety field.

图 1 公共安全领域数据融合实例

1.1.2 现实需求

当下,在工程实际中公共安全领域数据融合系统还需要很大改观. 1)所需的数据割裂地分布在多个数据源中,且数据种类多样,需要把这些割裂的多源数据自动整合为一个统一的系统. 2)目前公共安全领域的系统绝大多数提供的服务属于事后研判型,但是有些重大案件的破坏性非常大,事后再研判损失太大,急需能够做到事前预警的大数据技术和系统. 3)嫌犯可能会在作案后更改姓名、手机号码、常驻地等,这样会造成数据的演化,需要识别这种演化,这对于破案极为重要. 4)所需数据规模超大,比如为了找到涉恐人员的蛛丝马迹,需要对整个互联网和电信网路进行监控和分析处理,这里需要处理的数据目前工业界无法承受,需要控制融合的规模.

1.2 科学大数据融合案例分析

在科研领域,不仅需要数据本身,更需要与该数据有潜在密切关系的各种数据,并能够方便地分析这些数据. 例如,在查看一个基因数据时,还能循着它去看基因组、蛋白质等相关的其他数据. 为了实现这种融合,中国科学院提出了数据融合管理与服务系统,目前包括 36 个不同数据源的生物学数据,累计汇聚数据超过 40 TB;并在此基础上选取了 8 个数据源的数据进行数据解析、转换和数据关联处理,转换得到的约 830 万个数据之间建立起了约 1.4 亿个关联关系.

1.2.1 实现原理

该系统的实现原理如图 2 所示.

图 2 中,集中的数据存储库基于分布式文件系统

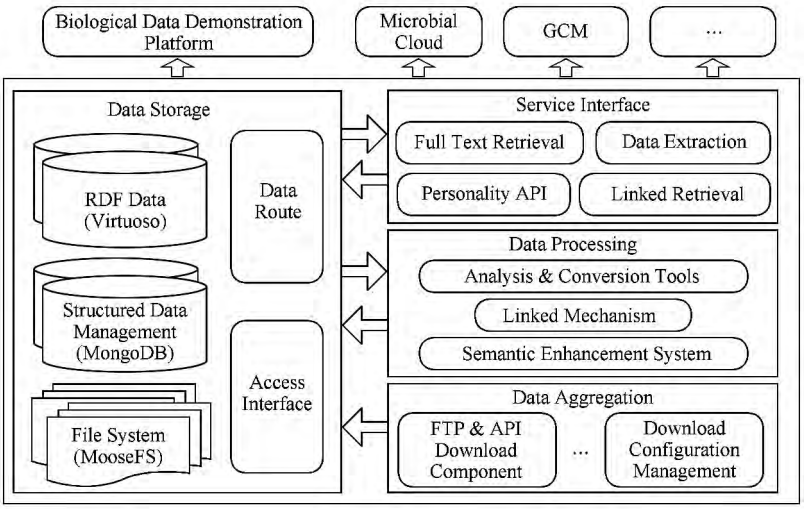

Fig. 2 Scientific data fusion management and service architecture.

图 2 一种科学数据融合管理与服务系统架构

MooseFS、MongoDB 数据库集群和 Virtuoso 数据库来构建，分别用于存储从各数据源下载的非结构化数据、从各数据源下载的结构化数据和解析原始数据后得到的数据、转换得到的 RDF 数据和通过语义增强系统新建立的关联关系数据. 这些数据通过数据汇聚模块的各种下载组件和管理系统进行汇聚. 对于下载的数据，定制化开发数据解析工具，完成数据从原始形态到“属性-值”结构化形态的转变；开发基于配置的数据转换工具，抽取对发现数据之间的关联有价值的“属性-值”并经必要的数据合并、拆分、等价变换等处理将抽取的这些数据转变为一致化的 RDF 格式；确定数据之间的关联机制并在此基础上运用相似度计算、推理、本体映射等关联发现方法来增加数据之间的语义关联关系. 最后通过服务接口模块对外提供服务.

1.2.2 现实需求

从上述介绍可以看出该系统更加注重数据的获取、存储、表达格式的统一和提供服务接口，数据关联与深度融合相对较弱. 在实现过程中发现该系统的服务稳定性和响应效率较高、对数据分析类应用的支持能力较好. 同时也发现了一些问题：1)科学数据的融合仅依靠软件工程师和计算机科学家很难完成，需要吸引各领域科研人员的广泛参与和紧密合作. 2)随着所汇聚和加工处理的数据量的增大，现有的数据存储方案会面临考验，特别是在 RDF 数据存储方面，业界缺乏具有百亿量级 RDF 数据存储管理和高服务能力的数据库系统或较成熟的解决方案. 3)目前所设计的数据关联发现方案还比较粗糙，主要应用了相似度计算方法，还有待深入研究.

1.3 Web 数据融合实例分析

在科研领域，经常需要查询学术信息，比如发表论文、承担项目、参与学术活动等. 这些信息分散在众多 Web 数据源中，对高效检索挑战很大，亟需一个跨领域、多学科的学术信息集成系统，为学术信息检索、分析提供方便. 为此，ScholarSpace[4] 应用而生，它包括 25 个学科领域的学术信息. 目前包含实体 1140 余万、三元组 1.8 亿，实体关系 66 种，支持学者、研究领域、研究课题等多条件的学术信息检索，并基于文本挖掘和社会网络分析建立学术关系网络，支持学者谱系、评审推荐等应用功能.

1.3.1 实现原理

ScholarSpace 的实现原理如图 3 所示，它的数据源于领域数据库、现存的知识库或者 Web 中的开放信息. 首先利用 Web 数据抽取技术从这些数据源中自动抽取学者、论文、科研项目、专利等实体和关系信息. 然后，在此基础上识别关联实体和发现实体之间的复杂关系，进而实现数据的融合和关联. 这一过程涉及 5 种核心技术——海联关联数据的存储技术、实体识别技术、复杂关系发现技术、实体和关系的演化处理技术和跨语言融合技术. 其中，海联关联数据的存储库用于存储自动收集的学术信息学和推演得到的知识；实体识别技术和复杂关系发现技术用于从数据源抽取并识别实体和关系，复杂关系发现还负责从已经得到的数据中推理间接关系；实体和关系的演化处理技术主要是为了应对实体和关系

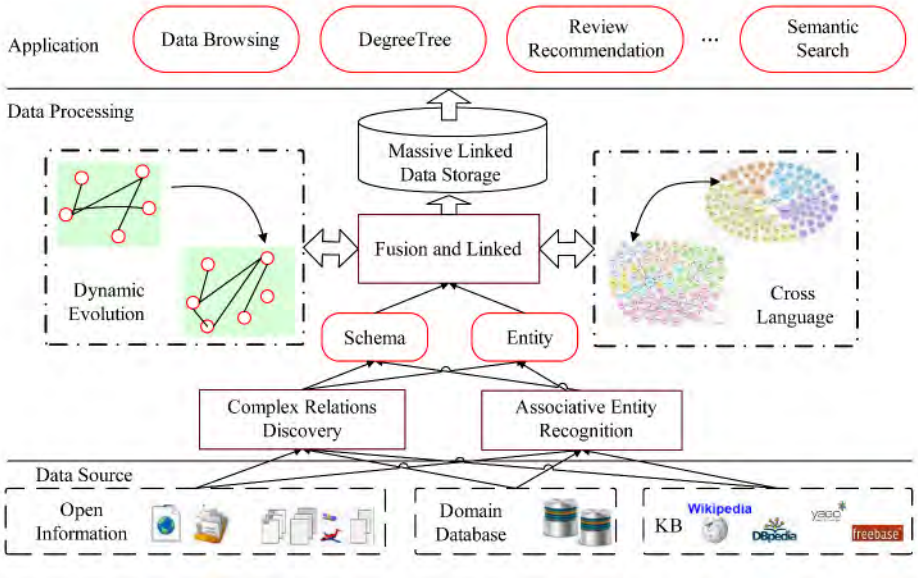

Fig. 3 The system architecture of ScholarSpace.

图 3 ScholarSpace 系统架构

随时间变化的情况,比如学者的所在单位发生了变化,或者学者因为升学发表论文的合作者发生了变化等;跨语言融合主要是因为学术无国界,同一学者的信息可以以任何语种出现,这种融合对于完成呈现学者学术信息是必要的.

1.3.2 现实需求

目前,该系统还需要不断完善.首先,数据源不断有新的出现、旧的消失,各种数据快速增长,规模越来越大,需要自适应更新策略和增量融合的方法.其次,随着大规模数据的交叉、关联和融合实体和关系的演化现象越来越明显,最难辨别的演化是看似不相似的记录表示同一实体,或者原本表示同一实体的记录因某些属性的改变而变得不太像同一实体的情况,这就需要对实体和关系的演化做细粒度分析建模.第三,学术信息可能是多语言的,例如发表的论文有中英文之分,需要做跨语言的融合.最后,学术信息中也蕴含的许多隐含关系,发现这些隐含关系意义重大,例如,"合作者"关系中可能包含"导师-学生",而"导师-学生"关系对专家推荐、学者谱系的构建等具有重要的帮助.

1.4 大数据融合的独特性与问题

通过上述分析可知,大数据时代数据的极大丰富为人们提供了更大的利用价值,但是数据的海量产生和新的特征也使人们面临的问题空前复杂化.

1)割裂的多源异构数据.目前需要处理的数据可能来自领域数据库、知识库或者 Web 页面的开放信息,从来源角度看是多源异构的.而且,这些数据被物理地存放在不同的系统中.这些割裂的多源异构数据造成了各种数据孤岛,给大数据分析处理带来非常大的挑战,需要把这些割裂的数据整合到统一的系统中.这种情况在 3 个案例中均有体现.

2)数据规模与数据价值的矛盾.当下,数据越来越丰富,提供了更多有价值的信息,但数据的规模也越来越大,对已有的数据存储和处理方法提出了挑战,需要对融合的规模进行控制.就像公共安全领域,如果办案时有越多的相关数据就越有可能快速破案.但是,目前需要处理的数据规模已经让工业界无法承受,只能对部分数据进行计算和处理得出结论.

3)跨媒体、跨语言的关联.需要处理的数据有结构化数据、半结构化数据和非结构数据,这对数据关联的发现提出了挑战,尤其是图片、视频、音频数据与文本数据的关联.这种情况在公共安全领域极为常见,如何自动识别它们之间的关联是工程实际中亟需的.并且数据可能源于多语种,如学术领域提到的同一作者可以发表中文、英文论文.

4)实体和关系的动态演化.数据是动态变化的,实体和关系也是随时间不断演化的,这就增加实体和关系的判别难度,容易造成数据不一致.比如,公共安全领域涉及的嫌犯在作案后更改姓名、学术领域中作者更换了所在单位等都属于此类情况.因此,需要合理建模演化行为,保证数据一致性.

5)知识的隐含性.从案例中我们也可以发现,除了显示知识还有隐式知识,隐式关系比显示知识更重要.比如生物领域中,鱼类中的掠食者在食物富集时运动轨迹呈布朗运动,或者学术案例中出现的"合作者"关系可能暗含"师生"关系.这种隐含的关系对知识的理解和数据的融合都有很大帮助.

2 现有数据融合技术分析

为了实现大数据的融合,各领域出现了一些融合方法,可以认为普遍采用 3V(海量、高速、类型多样)特征下的集成方式,如图 4 所示:

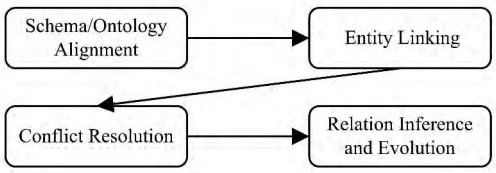

Fig. 4 Universal mode of big data fusion.

图 4 数据融合的普遍方式

当下这种融合方式普遍认为大数据融合的难点在于大数据的 3V 特征,它所需要的关键支撑技术有模式(本体)对齐技术、实体链接技术、冲突解决技术和关系推演.其中,1)模式和本体对齐技术用于应对本体的异构性和数据源的异构性[5-6];2)实体链接包含命名实体识别/记录链接和实体关联 2 部分,是大数据融合的基础;3)冲突解决是数据融合的必经之路[7-8],有时也叫做实体消岐;4)关系推演用于发现隐含知识,也可用于知识库的扩充和补全.

2.1 模式/本体对齐

模式对齐解决 2 个模式元素之间的一致性问题,主要是利用属性名称、类型、值的相似性以及属性之间的邻接关系寻找源模式与中介模式的对应关系[9-12].为了应对大数据新特征出现了演化模型[13]、概率模型[14-15]和深度匹配[16]方法.演化模型主要是检测模式映射的演化,采用尽力而为、模糊回答的方式,在一定程度上解决了数据多样性和高速性带来的问题;概率模型将中介模式按语义表示成源属性

的聚类,由此源模式会出现与其有不同程度对应关系的多个候选中介模式,然后根据查询请求为每个候选中介模式分配一个备选概率来确定最佳映射;深度匹配方法面向概念级,基于潜在的语义匹配,而不仅仅依赖于可见属性.

本体是针对特定领域中的概念而言的,用来弥合词汇异构性和语义歧义的间隙,本体对齐主要解决本体不一致问题,需要识别本体演化.本体演化分为原子变化、混合变化和复杂变化[17].原子变化反映单个本体的变化,混合变化修改本体实体的邻居,复杂变化是前两者的复合体.有时原子变化也叫基本变化,混合变化和复杂变化统称为复杂变化[18],这些变化通过日志[19]和本体版本差异[18,20]获得.一般在概念级[21]和实例级检测[22],采用图论方法表示本体变化[19]、引入 SetPi 运算来建模本体演化过程[23]、采用一致性约束跟踪本体的全局演化过程实现可溯源[17]、Pellet 推理检测不一致性[24].采用多重相似度度量与本体树结合实现多策略的本体匹配[25].

2.2 实体链接

实体链接的关键是实体识别,主要是识别相似实体和消除实体歧义,相似指多个命名实体表象可对应到一个真实实体(或称概念),歧义指一个实体表象可对应到多个真实实体.根据数据类型的不同,实体识别方法分为面向非结构化文本的命名实体识别与消歧、面向结构化数据的记录链接和 2 种数据类型之间的复杂数据实体关联方法.

2.2.1 命名实体识别

对于命名实体识别,先后出现了针对单查询、文档、短文档及社会媒体 3 种类型的识别方法.

命名实体识别最早针对单查询,且局限于维基百科和新闻文章,利用维基百科文章与围绕提及的上下文的相似性消除专有名词的歧义[26],或用统计识别方法识别命名实体并用多种提及联合消歧[27],如采用实体分类作为上下文相似度向量的一部分.

随着 Web 技术的发展,需要从普通文档中识别实体,最早采用类似文献[27]的方法,但给出了未知实体的显式模型来识别未知实体[28].接着采用识别"Wikipedia-like"链接方法,在一小部分已存在的维基百科链接上建立分类训练器,并用最频繁感知基准来度量相似度[29],或采用短语的歧义度作为其被提及的度量指标,并捕捉歧义候选对象之间的语义关联关系来识别实体[30].在此基础上出现了采用监督学习先验相似性,并采取近似算法求解联合概率分布的最大后验估计的联合推理方法[31].对于文档中提及的全局一致性,一般采用提及的迭代消歧方式解决,考虑实体间的语义相关性,但对领域和语料变化敏感[32].

随着社会媒体的发展,目前命名实体识别更倾向于社交网络、短文本,特别是微博平台[33-35],一般采用字典和启发式词组联合识别[33],或者加入微博的特殊句法(如 #,@ 等)做过滤器[34]联合推理.也有将各种相似度度量方法综合学习方法[36],当实体不在知识库中时采用从已知实体的特征(关键字句)随机抽样得到的未知实体表示的方法识别实体[37].

2.2.2 记录链接

记录链接是从数据集中识别和聚合表示现实世界中同一实体的记录(也称实体表象),即对相似度达到一定阈值的记录做聚类操作(也称共指识别).相似性一般根据领域知识设定匹配规则度量、也可用机器学习训练分类器的方法实现[38]、或利用编辑距离或欧氏距离计算[39].作出表象局部相似性判断后,接下来的工作是对实体进行邻接性聚类、相关性聚类[40-41]或密度聚类[42-44].其中后 2 类聚类采用奖励高内聚、惩罚高相关性保证歧义最小的方法与大数据的"差序"关系相辅相成.但这些方法是非增量式聚类,难以应对大数据的海量性,考虑到大数据的相互关联对实体匹配的局部决策和全局一致性的影响,以及数据更新可以及时弥补聚类过程中的错误聚类,出现了增量记录链接方法,主要从匹配规则的演化[45]和数据的演化[46-47]2 个方面为依据探讨记录链接的增量问题.

由于大数据的海量性,在相似性计算之前先根据实体的一个或多个属性值将输入记录划分为多个块,进行块内比较,提高链接效率[8].分块技术按分块函数数量分为单分块技术[48]和多分块技术[49],按冗余程度又可分为冗余消极、中立和积极[50-51]3 种.通常采用多分块技术与冗余积极相结合的方法,因为大数据富含冗余信息、实体的属性多样,且单分块技术存在假负现象、仅适用于高质量先验模式属性的情况、冗余消极和冗余中立在创建块时需要先验知识.但多分块与冗余积极相结合的方法引发了重复比较 、多余比较和不匹配比较,为此,出现了借助 MapReduce 并行分块[52]和引入 Meta-blocking 直接优化分块[53].Meta-blocking 技术首先将信息封装在块分配集并构建块图,然后将问题转化为度量图中边的权重和图修剪问题.这种做法独立于底层的分块技术,与模式无关,具有通用性.但是,它没有本质上代替原有的分块技术,它依赖于底层块集合

的冗余程度,并且分块图的构建是通过调整块构建方法中的相应参数得到的,因此目前需要一种调参少、不依赖于底层块集合的冗余程度低且模式无关的分块方法.

2.2.3 复杂数据实体关联

结构化数据与非结构化数据也存在关联关系,我们将这两者的关联称为复杂数据实体关联,它的核心任务是表象消歧.早期研究集中于从文档中识别与数据库实体对应的表象,代表性系统是 SCORE[54] 和 EROCS[55],分别采用关键字匹配和词共现原理寻求两者的对应关系.后来,针对评论信息中的实体与数据库对象的相关性,提出一种不需要识别评论中实体就能完成匹配的产生式语言模型[56-57].接着针对在线供应商的无结构产品信息与结构化的商品清单信息做链接,提出基于语义理解的有监督的学习方法[58].这些方法都无法处理实体演化的情况.当下研究热点转为寻找 Web 文本中命名实体提及与知识库中命名实体的关联关系,这种对应关系分为可链接和不可链接 2 种,不可链接是指知识库中不存在对应实体的情况,否则为可链接[59].可链接关系的核心是在知识库中寻找最优匹配实体,通过产生候选对象并对其排序得到.候选链接的产生可以通过图论的方法[60],借助语义知识[61]、概率模型[62],如果是面向社会媒体,则可以利用用户兴趣等建模链接关系[63].候选链接的排序按影响因素可以分为与实体的上下文信息无关[64]和与实体的上下文信息有关[65-66] 2 种.不可链接采用设定阈值的方法判定,并最终聚类不可链接提及[63,65].

2.3 冲突解决

冲突分为模式冲突、标示符冲突和数据冲突,其中模式冲突由数据源的模式异构引起,比如属性名、语义等不同的情况,标识冲突主要是指异名同义现象;数据冲突主要是指同一属性具有多种不同的值.冲突的解决一般是在实体或属性级别,采用识别函数.目前主要集中在实体级别的真假甄别和演化问题.

真假甄别问题也称事实(fact)甄别问题,即从所有冲突的值中甄别正确的值(真值),真值可以不止一个,但多个真值间语义上相同[39,67-68].影响真值度量的因素可分为数据源的复制关系[69]和依赖关系[70]以及值的新鲜度和相似度.对于真值的度量一般采用投票的策略[71],并在此基础上进行独立性衰减[72],然后根据值的置信度(源的可信性的函数)[73]、值的贝叶斯后验概率(用数据源的精度和复制概率表示)[72],或者以源的独立性、组的可靠性、值的真实性为参数设定真值的完全分布函数,并将对数似然值最大化求下界推理得到真值结果.这些方法侧重于有效地检测假数据的正相关性,但不适用于真实数据的正相关或负相关,并且模型依赖于单一真值的假设,然而,某些事实可以有多个真值,如某人可能有多个职业.所以,真值度量因素中又增加了多真值[74],源的正、负相关性[75],值演化性[76]及数据抽取维度[77]等不确定因素.其中,值的演化性用一定时间区间内的数据项的值的一组状态变迁序列度量.

实体演化是指实体随时间演化会出现看似不相似的记录表示同一实体的现象,但已有的方法大多是判别相似记录是否表示同一实体[67,78],不适合这种情况[79],这是因为实体的属性值可能随时间变化.所以,对于随时间变化的实体,需要细粒度分析变化.最早实体演化建模采用时间衰减模型捕获实体属性值在时间跨度范围内改变的可能性,其中,采用歧义衰减度量相同属性值变得不一样的概率,一致性衰减度量不同属性值变得一样的概率[80].但只捕获了属性值变或不变的概率,为此出现了采用突变模型来学习随时间推移属性值再次出现的概率加以改进的方法[81],这种方法考虑了属性值来回变化的情况和实体内/间的演化,依据全部历史时间点做决策.

2.4 关系推演

我们希望自动地找到关联数据中的路径模式和自然语言中的关系词汇之间的对应关系.这种对应关系对于理解复杂数据非常重要.这就涉及到关系推演问题.关系推演包括 3 种情况:已知一个实体和一条关系,推断另一实体,或者已知 2 个实体预测它们之间的关系;实体间间接关系的推理;关系的演化度量.

对于前 2 种情况,大多数采用嵌入表示[82-83]和图特征模型[84-86]进行关系的推理与预测.嵌入表示即将实体和关系都表示为低维(如 d 维)向量 $\boldsymbol{h}$ 或 $\boldsymbol{t}$,并且定义一个评分函数 $f_r(\boldsymbol{h},\boldsymbol{t})$ 来确定元组的合理性,主要模型有双线性模型、多层感知模型和潜在距离模型,这些模型的对比如表 1 所示.其中 RESCAL 模型是典型的双线性模型,E-MLP,ER-MLP 和 NTN 属于多层感知模型,其余的是潜在距离模型.多层感知模型参数复杂,后 4 种可以处理复杂关系,KG2E 模型将实体和关系表示为高斯分布,其他模型将实体和关系映射为超平面中的点.

Table 1 Embedding Models Comparison

表 1 嵌入模型比较

Model	Score Function $f_r(\boldsymbol{h},\boldsymbol{t})$	Memory Complexity
RESCAL[87]	$\boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}_r\boldsymbol{t}, \boldsymbol{W}_r\in\mathrm{R}^{k\times k}$	$O(n_e d_e+n_r d_r^2)$
E-MLP[88]	$\boldsymbol{W}_k^{\mathrm{T}} f(\boldsymbol{A}_k^{\mathrm{T}}\boldsymbol{\phi}_{ij}^{\text{E-MLP}}), \boldsymbol{A}_k=[A_k^s;A_k^o], \boldsymbol{\phi}_{ij}^{\text{E-MLP}}=[e_i;e_j]$	$O(n_r(d_a+d_a\times 2d_e)+n_r d_e)$
ER-MLP[89]	$\boldsymbol{W}^{\mathrm{T}} f(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{\phi}_{ij}^{\text{ER-MLP}}), \boldsymbol{\phi}_{ij}^{\text{ER-MLP}}=[e_i;e_j;e_k]$	$O(d_c+d_c(d_r\times 2d_e)+n_r d_r+n_e d_e)$
NTN[88]	$\boldsymbol{u}_r^{\mathrm{T}} f(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{W}_r\boldsymbol{t}+\boldsymbol{W}_{rh}\boldsymbol{h}+\boldsymbol{W}_{rt}\boldsymbol{t}+\boldsymbol{b}_r), \boldsymbol{u}_r,\boldsymbol{b}_r\in\mathrm{R}^s, \boldsymbol{W}_r\in\mathrm{R}^{d\times d\times s}, \boldsymbol{W}_{rh},\boldsymbol{W}_{rt}\in\mathrm{R}^{s\times d}$	$O(n_e d_e+n_r(sd_e^2\times 2sd_e+2s))$
SE[90]	$\lVert\boldsymbol{W}_{rh}\boldsymbol{h}-\boldsymbol{W}_{rt}\boldsymbol{t}\rVert_{l1/2}, \boldsymbol{W}_{rh},\boldsymbol{W}_{rt}\in\mathrm{R}^{d_r\times d_e}$	$O(n_e d_e+2d_r n_r)$
TransE[91]	$\lVert\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\rVert_{l1/2}, \boldsymbol{r}\in\mathrm{R}^{d_r}$	$O(n_e d_e+n_r d_r)$
TransR[92]	$\lVert\boldsymbol{h}\boldsymbol{M}_r+\boldsymbol{r}-\boldsymbol{t}\boldsymbol{M}_r\rVert_{l1/2}, \boldsymbol{r}\in\mathrm{R}^{d_r}, \boldsymbol{M}_r\in\mathrm{R}^{k\times k}$	$O(n_e d_e+(d_r+d_r^2)n_r)$
TransH[93]	$\lVert(\boldsymbol{h}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r)+d_r-(\boldsymbol{t}-\boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{t}\boldsymbol{w}_r)\rVert_2^2, \boldsymbol{w}_r,d_r\in\mathrm{R}^{d_r}$	$O(n_e d_2+2n_r d_r)$
KG2E_EL[94]	$\frac{1}{2}\left\{\boldsymbol{\mu}^{\mathrm{T}}\sum{}^{-1}\boldsymbol{\mu}-\log\det(\sum_h+\sum_t+\sum_r)\right\}$	$O(2n_e d_e+2n_r d_r)$
KG2E_KL[94]	$\frac{1}{2}\left\{tr(\sum_r^{-1}(\sum_h+\sum_t))+\boldsymbol{\mu}^{\mathrm{T}}\sum_r^{-1}\boldsymbol{\mu}-\log\frac{\det((\sum_h+\sum_t))}{\det(\sum_r)}\right\}, \boldsymbol{\mu}=\boldsymbol{\mu}_h-\boldsymbol{\mu}_t-\boldsymbol{\mu}_r$	$O(2n_e d_e+2n_r d_r)$

图特征模型被广泛用于链接预测,它认为相似的实体很可能相关,相邻的节点或者有路径相连的节点很可能相似.衡量实体相似性的方法分为局部相似、全局相似和准局部相似[95].局部相似性计算只依赖于所涉及实体的直接附近实体,不能模拟大范围的依赖关系[96];全局相似性考虑了所有路径上的实体,预测性能比局部相似方法好,但计算更昂贵[97];准局部相似方法通过路径实体的相似度和有限长度的随机游走平衡了预测精度和计算复杂度[98].此外,还有其他方法,例如:采用路径排序算法延伸有限长度的随机游走来预测多关系知识图的链接[99];从大型知识图中提取逻辑规则,可以处理知识图的开放世界假设[100].

嵌入表示和图特征模型互补[101],前者擅长通过新引入潜在变量建模全局关系模式,并且当元组可以用少量的隐变量解释时计算效率很高;后者擅长建模局部和准局部图模式,并且当元组可以由邻居实体或与其有较短路径的实体解释时计算效率很高.因此出现了很多两者结合的研究方法[101-102].

关系推演的另一个方面是实体关系的演化,它表现为聚类随时间的变化,这类方法认为首先应该为记录创建软聚类,即在作出每个记录应该属于哪个聚类的决定之前,一个记录可以同时属于多个聚类,然后收集证据在软聚类的基础上迭代细化聚类[80].但是现实中缺乏演化证据,为此出现了 2 阶段聚类[103]方法.第 1 阶段假设记录静态并基于属性值相似度分组做实体静态匹配,为第 2 阶段的演化决策收集证据;第 2 阶段考虑时间维度,从初始分组合并聚类,合并的条件是一个实体从一个聚类中的某个状态演化到另一个聚类中的某个状态.这种方法记录匹配阶段不考虑演化情况,聚类决策时采用演化决策,既节省了时间,又不损害匹配精度.

2.5 现存技术的局限性

经过前 4 小节的技术梳理发现面对大数据融合,现存融合技术还存在如下局限性:

2.5.1 实体链接技术存在的局限性

首先,现有的实体链接基本是实体识别、冲突解决、共指识别串行化执行,不感知彼此的相互影响.但是,这样做有 3 方面的弊端:1)实体识别过程中产生的错误会依次向后续过程传播,这种错误不可恢复;2)共指识别和冲突解决的结果不能向前反馈;3)实体识别过程和冲突解决过程可能会产生不一致的输出.但实际中这三者相互影响,前者为后两者提供更多的特征,后两者为前者提供已消歧的链接信息辅助聚类.所以有人提出交叉迭代[104]的方法,这种联合链接方法也是目前的一大研究热点.

其次,共指识别还面临的一大挑战是实体关系的演化,已有方法[80]没有考虑可靠性和更新程度、局部决策对与之关联表象的影响,并且直接面向动态数据,演化模型依赖于训练数据集和演化证据的质量,匹配精度高,但时间代价不是大数据能够承受的.

最后,复杂实体关联方法在适用范围、准确率等方面都存在一定的不足,主要挑战性在于:1)非结构化数据中一般不显式包含属性名,其实体属性也不一定都完全出现在结构化数据中,反之亦然.并且,2 类实体之间是需要做近似匹配还是精确匹配也需要区别;2)新实体的发现也是目前的一大难点,关键在于相似性判定阈值的确定没有有效的解决办法;

3)大数据融合向跨语言融合迈进,所以需要相关实体跨语言、跨文档的关联,目前研究成果不多[105-107].其中,未知链接的处理对于跨语言、跨文档的链接更加复杂;实体链接中存在隐喻情况、一个实体在多个文档中出现的情况、提及的边界重叠的情况、嵌套提及、嵌套链接的情况,以及实体的相关性,这些情况都是目前亟待解决的问题.

2.5.2 冲突解决技术的局限性

目前,冲突解决的侧重点在于知识的真假甄别,但是对于大数据融合还不够,还存在以下2个问题:

首先,消歧方法依赖于实际参照数据的可用性(如数据标注),参照数据一般源于维基百科,缺乏领域性和针对性,这使得实用性变窄.对于其他领域,如新闻,仅有一小部分标注样本可用,所以必须采取超越维基百科的消歧策略.

其次,引发冲突的一个关键因素是信息的质量[7,67],如数据本身的新鲜度、对特定需求的价值量等,并且对于新鲜度和价值量不同的多真值问题,如何设计质量评估函数是一个挑战性问题.此外在真假甄别过程中有2个假设:假值服从均匀分布;不匹配即为完全不同,但这个假设在对于现实过于绝对,以至于已有方法不能很好地处理错误产生的不确定性.此外,所有冲突解决技术都有一个假定前提,即假定模式对齐和实体识别已完成,并且数据也已经对齐.但这个假设在大数据环境下过于理想化.

2.5.3 关系推演技术存在的局限性

关系推演主要集中在关系的推理和关系的演化建模.关系推理方面目前只考虑了直接关系和多路径关系的推理,缺乏对关系之间复杂模式的考虑,如自动通过元组(人,离不开,空气)推断出元组(鱼,离不开,水)这种类比关系.并且关系推演借助于知识表示,目前有嵌入表示和RDF图2种表示.嵌入表示方法存在复杂关系表示与系统可扩展性不能兼顾的问题[93];采用RDF图表示时,传统的图相似性计算只是考虑到图结构的相似性,典型的如图的编辑距离和最小公共子图等,显然这种量度不能很好地反映语义上的相似性.有时实体间图结构的编辑距离比较大,但是它们的语义却等价,所以采用RDF图表示时要重点考虑语义关系.无论采用哪种表示形式,都需要考虑推理关系的可信性,自动过滤无意义的推理关系.

此外,演化建模对冲突识别与解决影响很大,虽然现有方法捕获了实体属性值的改变,但未考虑属性值变化的复杂模式,如用属性的再现概率建模实体演化[81],当一个属性值在后续时间内不再出现,则所有情况下记录表示同一实体的可能性相同,但这个说法与实际相悖.如一个讲师在2年后成为副教授是可能的,但1年后变为助教的可能性是不存在的,明显前一种表示同一实体的可能性远大于后一种,而文献[81]认为这种概率是相同.这说明,建模变化需要考虑属性本身的变化模式,如语义相关度等.

3 大数据融合的理解

由上述分析可知,大数据价值链是一个"离散数据→集成化数据→知识理解→普适机理凝练→解释客观现象、回归自然"这样一条阶梯式循环过程,每一个链条是对大数据的一次价值提升.为了实现这一价值,本文提出大数据融合的概念,即它是一种处理

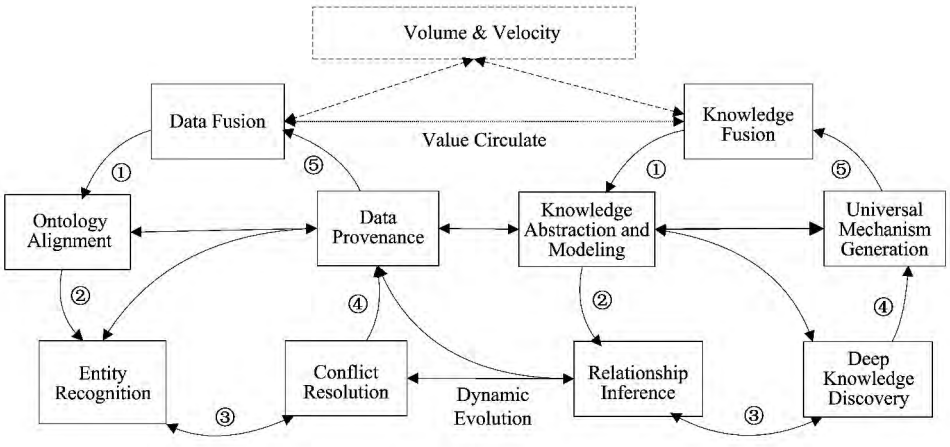

Fig. 5 The big data fusion architecture.

图5 大数据融合框架

大数据的手段,用于从大数据中发现知识,并按照知识的语义逻辑关联融合形成更接近人类思维的知识,包括数据融合和知识融合2个步骤,如图5所示.

数据融合负责将多源数据动态提取、整合并且转化为知识资源,为知识融合奠定基础.而知识融合负责对知识和知识间的关系进行不同粒度的理解,使知识具有不同层次的可理解性和可领悟性,进而方便解释客观现象.数据融合和知识融合不是孤立存在的,知识融合中获取的知识可以作为数据融合的参考因素辅助数据融合;而数据融合也不仅是为知识融合提供集成化数据,其中的一些方法同样对知识融合有借鉴作用.此外,还有2个贯穿整个大数据融合过程的操作,即数据溯源和动态演化,它们保证了大数据融合的与时俱进和可理解性.这种融合方式的优势在于通过双环互动、启发动态演化地逐步探索大数据融合问题,并且融合过程的每个步骤都是大数据价值的一次提升过程.

3.1 数据融合

数据融合需要用动态的方式统一不同的数据源,将数据转化为知识资源.这个过程对用户透明,缺乏可解释性和可操作性,并且大数据的海量性和动态演化加大了错误恢复的难度,传统融合方法没有考虑这一点.因此,必须建立大数据融合的可溯源机制.

另外,大数据的关联性使得融合步骤之间相互影响,传统的流水线式融合不再满足现有融合需求.面对新的融合需求,反馈迭代机制显得极为重要.

为此,我们给出数据融合的实现步骤:对齐本体、模式,加速融合效率;识别相同实体、链接关联实体;甄别真伪、合并冲突数据,并将处理结果反馈给实体识别阶段,提高识别效率;对数据起源,实体识别和冲突解决过程溯源、跟踪数据的演化.

1) 模式/本体对齐.模式/本体对齐是大数据融合的辅助步骤,用于提高融合效率,重点对齐演化引起的不一致性.大数据的高通量性和演化性导致事后补救难度大,所以需要采用"以防为主,防治结合"的策略.此外,还可以变相思维,利用模板[108-109]在捕捉经验方面的优势为频繁错配的本体建立对齐模板以便重复使用.所以,我们认为本体演化对齐应该分3步完成,即本体的演化管理、不一致性的预防和补救、对齐模板的挖掘.

2) 实体识别.实体识别是数据融合的基础,大数据环境中实体识别有别于传统实体识别的方面在于:①实体之间的语义关联性较强,且存在演化性.②实体的属性特征以及所在的语境信息、冲突实体的解决结果和共指识别结果都可能对实体识别产生影响.所以,识别实体应该是实体识别、冲突解决、共指识别三者迭代优化、逐步求精的过程.③推演出的新知识、发现的深度知识,以及得到的普适机理都有可能对实体识别起到启发作用,所以,反馈结果极为重要.

3) 冲突解决.冲突解决是大数据融合的必要条件,它的第一要务是消歧.大数据的真实性和演化性是引发冲突的导火索,如数据本身的新鲜度和贡献给特定查询的价值量等,这就引发了新鲜度和价值量不同的多真值问题,需要评估信息质量,合并不确定性信息.此外,知识融合中推演出的关系也可能对其起到启发作用,所以要将这种新知识动态地引入冲突解决过程,并保持这种知识的演化.所以,本文给出冲突解决的步骤,即真假甄别、不确定性合并和动态演化.

4) 数据溯源.数据溯源是传统数据融合不具备的,用于建立大数据融合的可回溯机制,追溯融合结果的数据来源以及演化过程,及时发现和更正错误.它的关键是数据起源的表示以及数据演化的中间过程的跟踪,其中,中间过程包括实体识别和冲突解决过程.所以,需要建立实体识别溯源机制,用于跟踪融合结果是由哪些待统一实体所产生;建立冲突解决溯源机制,用于处理融合结果元组中的每个值来自于哪些记录的哪个属性值以及通过何种冲突解决方法得来.

3.2 知识融合

知识融合是将数据融合阶段获得的笼统的知识转化为可领悟知识,面向需求提供知识服务.它需要挖掘隐含知识,寻找潜在知识关联,进而实现知识的深层次理解,以便更好地解释数据.为此,我们给出知识融合的实现步骤:对知识进行抽象和建模,为后续知识融合提供方便;通过对表层知识的推理、理解,得出显式深度知识,如通过多路径关系推理得到间接知识;通过推理、归纳等方法发现隐式深度知识,如类比关系等;对知识资源、深度知识等剖析、解释、归纳出普适机理.

1) 知识抽象和建模.实体和关系可以有多种不同组合,形成的知识也多种多样.所以,需要针对实体与关系的自身特点建立知识表示空间.通常将知识建模为RDF图或者嵌入表示为低维稠密的向量空间.RDF图既不损失语义关联又能很好地表示知识,它的一个难点是需要对RDF图携带的3种信息——

描述性属性、语义关系以及两者兼顾的语义图结构进行概念描述,这一步对后续深度知识发现特别重要.采用嵌入表示的方法主要是为了缓解数据稀疏,建立统一的语义表示空间,实现知识迁移,它的挑战性在于缺乏对各语言单位统一的语义表示与分析手段.

2)关系推演.关系推演是一种显式深度知识发现,包括多路径关系的推理、新关系的预测和关系的演化建模.多路径关系推理的难点在于组合语义模型的设计和推理关系的可用性确定.新关系的预测是指根据历史知识预测2个实体之间可能存在的关联关系,或者给定一个实体和一种关系,预测与之对应的实体.这种预测的关键在于实体和关系的表示.关系的演化建模中关系可以是属性关系,也可以是语义关系,所以需要对关系变化做细粒度的分析.此外,发现的深度知识对关系推演具有参考价值,所以还需要考虑深度知识发现反馈的结果.

3)深度知识发现.深度知识发现对知识融合非常重要,尤其是隐式深度知识发现,它包含以下3种:①关系型深度知识,包含类比关系、上下位关系、因果关系、正/负相关关系、频繁/顺序共现关系、序列关系等,例如,人离不开空气与鱼离不开水这种类比关系.②数据分布型深度知识,即知识服从某些数据分布,如高斯分布、幂律分布、长尾分布等.例如,当关注数少于105时社交网络中节点的度分布服从指数为2.267的幂率分布[110].③性质型深度知识,即知识具有某种性质,如局部封闭世界、长城记忆、无标度等,常见的如知识图谱建模可假设满足局部封闭世界假设[89].

4)普适机理的剖析和归纳.目前知识融合依然缺乏对知识资源中存在的关系普适化.为此,我们首先要从理性或直觉中建立问题的模型,通过对数据呈现的现象进行概括性描述或者归纳学习得到普适模型,然后将模型与数据结合提供适当的泛化能力,比如,"谷歌大脑"可以通过深度学习无监督地辨别任何猫[111].另一方面,人的智力能透过现象看到本质,只有发现大数据所呈现出的普遍现象背后的普适原理才能对客观世界产生更大的影响.比如,社会网络中社群的消失现象,他们背后的普适原理是生物进化论.所以可以将其作为知识建模、深度知识发现和关系推演的一个参考因素,从而提高融合效率.

3.3 贯穿要素

数据融合与知识融合是一个相互启发、协调逐步融合的过程,两者受一些共同因素的影响,如动态演化性、海量性和高速性.这些因素直接影响融合技术.

1)动态演化.知识的动态演化贯穿整个大数据融合过程,它影响着数据融合、知识融合的各种技术,所以还需要结合其他方法具体考虑.但是,必须做的2个工作是:①对动态变化的跟踪和知识演化的建模,由于大数据的特殊性,需要考虑变化的复杂模式,如语义关系等,最好能从中挖掘概念模板以应对数据的高速性和海量性;②动态性多数据存储、索引带来的挑战,这也是影响大数据融合的关键因素,亟待解决.

2)海量性、高速性.对于海量性和高速性,主要是解决它们带来的负面影响,对这2个因素的处理直接关系到大数据融合的性能和效率.目前使用最多的方法是利用MapReduce解决,也有优化硬件技术的方法.我们认为,要想从根本上解决海量性和高速性带来的负面影响还需要采用软硬件同时优化的策略.

4 大数据融合面临的挑战

大数据融合是一个多学科、跨领域的研究问题,它的实现还面临着诸多挑战.

4.1 融合过程面临的挑战性分析

大数据融合具有其特殊性,所以,需要审视融合方式.此外,如何控制融合结果的规模、如何存储也是亟待解决的问题.

1)融合方式的变革.由上述分析可知,已有的融合方法关注点在于集成多源数据提供统一访问和集成化知识,它始终围绕着"大"来定义,缺乏理解、知识结构松散,没有揭示数据背后的深层意义.但是,大数据融合中知识的隐含性以及知识的理解、分析对融合大有帮助.比如,公共安全领域要想做到预警,就需要对数据进行理解、归纳数据背后的规律.所以,大数据融合需要数据集成与知识理解相互启发进行,而非单向串行,并且知识理解应该更注重揭示数据背后的深层意义,尽可能地形成机理.即需要变革融合的思维和技术,使其既能像处理传统数据那样处理大数据,又能采用碳原子合理组合成为钻石的方式获得高品质的知识.

2)融合规模的可控性.大数据融合一方面是为了提供更丰富的数据资源,所以需要尽可能融合相关数据和尽可能寻找数据之间的关联关系,如知识库的补全、知识库的扩充等;另一方面为了提供更有效

的知识服务,大数据融合考虑了知识间隐含的关系、特征,以及知识群体产生的普适机理等,这样,融合的规模会不断增加,所以大数据的融合必然要考虑融合的规模,并且对融合结果规模的控制是不容忽视的一个环节,它决定着融合结果的可用性.

3) 数据的存储和维护. 大数据的融合对数据存储挑战性更大. 首先,数据产生速度快、流通量大,所以需要考虑知识库的索引和更新问题,尤其是针对新的知识表示形式的索引方法和随时间增量更新的策略. 其次,大数据融合需要用动态的方式统一不同的数据源. 这个过程对用户透明,缺乏可解释性和可操作性,并且大数据的海量性和动态演化加大了错误恢复的难度. 因此,必须建立大数据融合的可溯源机制,且需要与数据存储配合.

4.2 融合结果带来的挑战性分析

大数据融合是为了实现大数据的大价值而提出的,然而,它的出现也不可避免地引发新的问题.

1) 大数据融合与隐私保护的矛盾. 数据融合使得数据集间的关联更紧密、关联关系更清晰,隐私泄露也越容易,这种泄露在用户发布数据时不可预知. 所以需要研究主动降低隐私泄露风险的策略和风险评估模型,用于有效地预测隐私泄露的风险程度,提供风险预警和降低风险的建议策略. 其次,数据的融合使更多的数据由于数据之间的关联性在无形中被公开化,从而无形中泄露了用户的敏感信息. 因此,当下的隐私性体现在不泄露用户敏感信息的前提下融合数据,这就需要尝试新的数据发布技术,尽量减少信息损失并且最大程度地保护用户隐私. 此外,大数据融合是一个动态性过程,数据也是与时俱进的,所以,相应的隐私保护策略也应该具有动态性. 最后,为了建立大数据融合的可回溯机制,追溯融合结果的数据来源以及演化过程,大数据融合采用了数据溯源技术,这项技术具有两面性,一方面可以作为依据向用户解释造成风险的原因,给出降低风险的建议;另一方面,数据溯源本身也可能带来隐私泄露问题,还需要有针对数据溯源技术的隐私保护技术.

2) 与实际应用对接. 大数据融合是为了更好地提供知识服务,其中数据融合提供集成化的知识,知识融合在此基础上进一步理解,获得了知识的隐性特征、规律,并对其进行验证、剖析、归纳出知识间呈现的普适性质、现象,甚至是内在机理. 那么如何将获取的深度知识、普适机理等成本低廉、直观、快速地应用到现实当中就成为问题. 有一个普遍的想法是:如果出现了类似的情境,是否可以利用已有的结论提出假设,然后在相同的环境设置下,调整一个或多个变化因素,观察事态变化以验证假设,这一过程的核心在于如何将可控模拟仿真的方法、大数据融合的理论与实际应用相结合,围绕现实中特定问题,依据大数据融合理论得到的相关历史知识、经验,包括规律、性质、机理、现象等,结合特定领域或情境下的知识,通过模拟、仿真的手段,生成相应的可执行方案. 这样做还有一个好处是充分利用领域理论,运用数学、物理等工具,进行理论建模、解析、逻辑演绎、公式推演和证明,用于得出推论,理解模型,仿真和实验的假设、过程和结果等. 所以,可控模拟仿真的方法、大数据融合的理论与实际应用相结合是目前亟待解决的一个问题.

4.3 融合技术面临的挑战性分析

大数据融合是一个多学科、跨领域、跨语言的研究问题,所以面临的挑战更加复杂.

1) 跨领域、跨学科融合问题. 大数据融合的对象具有多样性,它既可以是结构化数据(如表格、列表等)、非结构化数据(如文本、图片、视频等)、半结构化的社会媒体数据(如微博、博客等复杂类型数据),也可以是知识,如规律、模型、机理等,它不仅以多种形式共存,还出现在不同领域,出现了多类型、跨领域融合的现象. 针对这种跨领域的多形式数据进行知识融合不是简单的匹配融合,需要充分考虑各种数据形式的特点,同时需要研究它们的差异所在以及如何合理地处理这些差异,这是数据融合面临的一个挑战. 在知识融合过程中上层机理是相通的,如金融市场呈现出的长期记忆性和社会网络中注意力流的长期记忆性,它们都呈现出了长期记忆现象,那么,它们在分析、处理方法上就可以相互借鉴. 此外,系统科学从全局、整体出发,研究数据的宏观现象、特征等与数据库领域的局部、微观现象的发现形成互补,可以相互借鉴. 这种跨学科寻找在知识融合中适合地处理多形式数据的方法. 这种借助多种学科的方法使得知识融合更有价值和意义.

2) 跨语言、跨媒体融合问题. 人类语言的多样性决定了实体以及实体之间语义关系会出现多个不同语种表示的情况,即出现了跨语言特征,由此人们迫切希望以自己的母语为主要语种构建知识库或表达实体及其关联关系以获得更好的知识服务. 此外,探索跨语言的数据关联有助于提高知识库的覆盖率,然而,当下缺乏这种跨语言的大规模知识库,例如,DBpedia 以英文为主,仅提供少量的德语和法语版本,其他小语种就更没有对应的知识库了. 但是,

小语种知识也非常重要.例如,伊朗发生暴动,媒体上发布的相关新闻采用非通用语种,大家很难理解时态的发展,所以,有必要构建跨语言的知识库,有必要探索跨语言的融合方法.

5 结束语

本文提出了大数据融合的问题,探讨了大数据融合的关键技术和面临的挑战.大数据融合实质是为获取高品质知识、最大程度地发挥大数据的价值而提出,它的重要性毋庸置疑.但是,作为一个多学科、跨领域的研究问题,传统的融合方法已经无法适应.面对大数据融合这一类新颖挑战,不仅需要各领域科研人员的广泛参与和紧密合作,更迫切需要将研究方法向新的深度和广度拓展,做到大跨度、深层次融合.

致谢 感谢北京明略软件系统有限公司冯是聪提供安全领域案例;感谢中国科学院计算机网络信息中心侯艳飞、韩岳岐、黎建辉提供科学研究领域案例;感谢为本论文提供修改意见的老师和同学!

参 考 文 献

[1] Suchanek F M, Weikum G. Knowledge bases in the age of big data analytics [J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1713-1714

[2] Suchanek F, Weikum G. Knowledge harvesting in the big-data era [C] //Proc of the 2013 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 933-938

[3] Fei Xiaotong. Native China [M]. Beijing: Peking University Press, 1998
(费孝通. 乡土中国[M]. 北京: 北京大学出版社, 1998)

[4] WAMDM. ScholarSpace [EB/OL]. [2015-12-12]. http://c-dblp.cn

[5] Shvaiko P, Euzenat J. Ontology matching: State of the art and future challenges [J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(1): 158-176

[6] Zhao L, Ichise R. Ontology integration for linked data [J]. Journal on Data Semantics, 2014, 3(4): 237-254

[7] Jan M. Linked data integration [D]. Prague, Czechia: Charles University in Prague, 2013

[8] Dong X L, Srivastava D. Big data integration [C] //Proc of the 29th IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2013: 1245-1248

[9] Bellahsene Z, Bonifati A, Rahm E. Schema Matching and Mapping [M]. Berlin: Springer, 2011

[10] Euzenat J, Shvaiko P. Ontology Matching [M]. Berlin: Springer, 2007

[11] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching [J]. The VLDB Journal, 2001, 10(4): 334-350

[12] Shvaiko P, Euzenat J. A survey of schema-based matching approaches [J]. Journal on Data Semantics IV, 2005: 146-171

[13] Franklin M, Halevy A, Maier D. From databases to dataspaces: A new abstraction for information management [J]. ACM Sigmod Record, 2005, 34(4): 27-33

[14] Das Sarma A, Dong X, Halevy A. Bootstrapping pay-as-you-go data integration systems [C] //Proc of the 2008 ACM SIGMOD Int Conf on Management of data. New York: ACM, 2008: 861-874

[15] Dong X, Halevy A Y, Yu C. Data integration with uncertainty [C] //Proc of the 33rd Int Conf on Very Large Data Bases. New York: ACM, 2007: 687-698

[16] Kulkarni S, Srinivasa S, Khasnabish J N, et al. Sortinghat: A framework for deep matching between classes of entities [C] //Proc of the 30th IEEE Int Conf on Data Engineering Workshops (ICDEW). Pisctaway, NJ: IEEE, 2014: 90-93

[17] Stojanovic L. Methods and tools for ontology evolution [D]. Karlsruhe: Karlsruhe University Dissertation, 2004

[18] Klein M C A. Change management for distributed ontologies [D]. Amsterdam, The Netherlands: Vrije Universiteit Amsterdam, 2004

[19] Javed M, Abgaz Y M, Pahl C. Ontology change management and identification of change patterns [J]. Journal on Data Semantics, 2013, 2(2/3): 119-143

[20] Hartung M, Groß A, Rahm E. COnto-Diff: Generation of complex evolution mappings for life science ontologies [J]. Journal of Biomedical Informatics, 2013, 46(1): 15-32

[21] Ding L, Shinavier J, Shangguan Z, et al. SameAs networks and beyond: Analyzing deployment status and implications of owl: SameAs in linked data [M] //The Semantic Web-ISWC 2010. Berlin: Springer, 2010: 145-160

[22] Luong P H, Dieng-Kuntz R. A rule-based approach for semantic annotation evolution [J]. Computational Intelligence, 2007, 23(3): 320-338

[23] Liu L, Zhang P, Fan R, et al. Modeling ontology evolution with SetPi [J]. Information Sciences, 2014, 255(1): 155-169

[24] Djedidi R, Aufaure M A. ONTO-EVO AL an ontology evolution approach guided by pattern modeling and quality evaluation [C] //Proc of the 6th Int Symp on Foundations of Information and Knowledge Systems (FoIKS 2010). Berlin: Springer, 2010: 286-305

[25] Kumar S, Singh V. Multi-strategy based matching technique for ontology integration [M] //Computational Intelligence in Data Mining-Volume 3. Berlin: Springer, 2015: 135-148

[26] Bunescu R C, Pasca M. Using encyclopedic knowledge for named entity disambiguation [C] //Proc of EACL 2006. Cambridge, MA: MIT Press, 2006: 9-16

[27] Cucerzan S. Large-scale named entity disambiguation based on Wikipedia data [C] //Proc of EMNLP-CoNLL 2007. Cambridge, MA: MIT Press, 2007: 708-716

[28] Hoffart J, Altun Y, Weikum G. Discovering emerging entities with ambiguous names [C] //Proc of the 23rd Int Conf on World Wide Web. New York: ACM, 2014: 385-396

[29] Csomai A, Mihalcea R. Linking documents to encyclopedic knowledge [J]. Intelligent Systems, 2008, 23(5): 34-41

[30] Milne D, Witten I H. Learning to link with Wikipedia [C] //Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 509-518

[31] Kulkarni S, Singh A, Ramakrishnan G, et al. Collective annotation of Wikipedia entities in Web text [C] //Proc of the 15th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2009: 457-466

[32] Ratinov L, Roth D, Downey D, et al. Local and global algorithms for disambiguation to Wikipedia [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Stroudsburg, PA: Association for Computational Linguistics, 2011: 1375-1384

[33] Ferragina P, Scaiella U. Fast and accurate annotation of short texts with Wikipedia pages [J]. IEEE Software, 2012, 29(1): 70-75

[34] Guo S, Chang M W, Kiciman E. To link or not to link? A study on end-to-end tweet entity linking [C] //Proc of HLT-NAACL 2013. Stroudsburg, PA: Association for Computational Linguistics, 2013: 1020-1030

[35] Basave A E C, Rizzo G, Varga A, et al. Making sense of microposts (# microposts2014) named entity extraction & linking challenge [C] //Proc of the 4th Workshop on Making Sense of Microposts (# Microposts2014). New York: ACM, 2014: 54-60

[36] Ceccarelli D, Lucchese C, Orlando S, et al. Learning relatedness measures for entity linking [C] //Proc of the 22nd ACM Int Conf on Information & Knowledge Management. New York: ACM, 2013: 139-148

[37] Jin Y, Kıcıman E, Wang K, et al. Entity linking at the tail: Sparse signals, unknown entities, and phrase models [C] //Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 453-462

[38] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 269-278

[39] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate record detection: A survey [J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(1): 1-16

[40] Charikar M, Guruswami V, Wirth A. Clustering with qualitative information [C] //Proc of the 44th Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2003: 524-533

[41] Bansal N, Blum A, Chawla S. Correlation clustering [J]. Machine Learning, 2004, 56(1/2/3): 89-113

[42] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227

[43] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] //Proc of KDD'96. New York: ACM, 1996: 226-231

[44] Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496

[45] Whang S E, Garcia-Molina H. Entity resolution with evolving rules [J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1326-1337

[46] Whang S E, Garcia-Molina H. Incremental entity resolution on rules and data [J]. The VLDB Journal, 2014, 23(1): 77-102

[47] Gruenheid A, Dong X L, Srivastava D. Incremental record linkage [J]. Proceedings of the VLDB Endowment, 2014, 7(9): 697-708

[48] Bitton D, DeWitt D J. Duplicate record elimination in large data files [J]. ACM Trans on Database Systems, 1983, 8(2): 255-265

[49] Hernández M A, Stolfo S J. Real-world data is dirty: Data cleansing and the merge/purge problem [J]. Data Mining and Knowledge Discovery, 1998, 2(1): 9-37

[50] McCallum A, Nigam K, Ungar L H. Efficient clustering of high-dimensional data sets with application to reference matching [C] //Proc of the 6th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2000: 169-178

[51] Gravano L, Ipeirotis P G, Jagadish H V, et al. Approximate string joins in a database (Almost) for free [C] //Proc of the 27th Int Conf on Very Large Data Bases. New York: ACM, 2001: 491-500

[52] Kolb L, Thor A, Rahm E. Load balancing for MapReduce-based entity resolution [C] //Proc of the 28th Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2012: 618-629

[53] Papadakis G, Koutrika G, Palpanas T, et al. Meta-blocking: Taking entity resolutionto the next level [J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(8): 1946-1960

[54] Roy P, Mohania M, Bamba B, et al. Towards automatic association of relevant unstructured content with structured query results [C] //Proc of the 14th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2005: 405-412

[55] Chakaravarthy V T, Gupta H, Roy P, et al. Efficiently linking text documents with relevant structured information [C] //Proc of the 32nd Int Conf on Very Large Data Bases. New York: ACM, 2006: 667-678

[56] Dalvi N, Kumar R, Pang B, et al. Matching reviews to objects using a language model [C] //Proc of the 2009 Conf on Empirical Methods in Natural Language Processing: Volume 2. Stroudsburg, PA: Association for Computational Linguistics, 2009: 609-618

[57] Dalvi N, Kumar R, Pang B, et al. A translation model for matching reviews to objects [C] //Proc of the 18th ACM Conf on Information and Knowledge Management. New York: ACM, 2009: 167-176

[58] Kannan A, Givoni I E, Agrawal R, et al. Matching unstructured product offers to structured product specifications [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2011: 404-412

[59] Hoffart J, Altun Y, Weikum G. Discovering emerging entities with ambiguous names [C] //Proc of the 23rd Int Conf on World Wide Web. New York: ACM, 2014: 385-396

[60] Han X, Sun L, Zhao J. Collective entity linking in Web text: A graph-based method [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2011: 765-774

[61] Shen W, Wang J, Luo P, et al. Linden: Linking named entities with knowledge base via semantic knowledge [C] // Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 449-458

[62] Shen W, Han J, Wang J. A probabilistic model for linking named entities in Web text with heterogeneous information networks [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 1199-1210

[63] Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling [C] // Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 68-76

[64] Shen W, Wang J, Luo P, et al. LIEGE: Link entities in Web lists with knowledge base [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1424-1432

[65] Li Y, Wang C, Han F, et al. Mining evidences for named entity disambiguation [C] //Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1070-1078

[66] Ceccarelli D, Lucchese C, Orlando S, et al. Learning relatedness measures for entity linking [C] //Proc of the 22nd ACM Int Conf on Information & Knowledge Management. New York: ACM, 2013: 139-148

[67] Dong X L, Naumann F. Data fusion: Resolving data conflicts for integration [J]. Proceedings of the VLDB Endowment, 2009, 2(2): 1654-1655

[68] Li X, Dong X L, Lyons K, et al. Truth finding on the deep Web: Is the problem solved? [J]. Proceedings of the VLDB Endowment, 2012, 6(2): 97-108

[69] Dong X L, Srivastava D. Large-scale copy detection [C] // Proc of the 2011 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2011: 1205-1208

[70] Pasternack J, Roth D. Latent credibility analysis [C] //Proc of the 22nd Int Conf on World Wide Web. 2013: 1009-1020

[71] Dong X L, Saha B, Srivastava D. Less is more: Selecting sources wisely for integration [J]. Proceedings of the VLDB Endowment, 2012, 6(2): 37-48

[72] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 550-561

[73] Yin X, Han J, Yu P S. Truth discovery with multiple conflicting information providers on the Web [J]. IEEE Trans on Knowledge and Data Engineering, 2008, 20(6): 796-808

[74] Zhao B, Rubinstein B I P, Gemmell J, et al. A Bayesian approach to discovering truth from conflicting sources for data integration [J]. Proceedings of the VLDB Endowment, 2012, 6(5): 550-561

[75] Pochampally R, Das Sarma A, Dong X L, et al. Fusing data with correlations [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 433-444

[76] Dong X L, Berti-Equille L, Srivastava D. Truth discovery and copying detection in a dynamic world [J]. Proceedings of the VLDB Endowment, 2009, 2(1): 562-573

[77] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion [J]. Proceedings of the VLDB Endowment, 2014, 7(10): 881-892

[78] Getoor L, Machanavajjhala A. Entity resolution: Theory, practice & open challenges [J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2018-2019

[79] Köpcke H, Thor A, Rahm E. Evaluation of entity resolution approaches on real-world match problems [J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 484-493

[80] Li P, Dong X L, Maurino A, et al. Linking temporal records [J]. Frontiers of Computer Science, 2012, 6(3): 293-312

[81] Chiang Y H, Doan A H, Naughton J F. Modeling entity evolution for temporal record matching [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 1175-1186

[82] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[83] Getoor L, Diehl C P. Link mining: A survey [J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 3-12

[84] Hong L, Zou L, Lian X, et al. Subgraph matching with set similarity in a large graph database [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(9): 2507-2521

[85] Wang D, Zou L, Zhao D. Top-k queries on RDF graphs [J]. Information Sciences, 2015, 316: 201-217

[86] Zheng W, Zou L, Lian X, et al. Efficient graph similarity search over large graph databases [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(4): 964-978

[87] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data [C] //Proc of the 28th Int Conf on Machine Learning (ICML 2011). New York: ACM, 2011: 809-816

[88] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 926-934

[89] Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A Web-scale approach to probabilistic knowledge fusion [C] //Proc of the 20th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601-610

[90] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases [C] //Proc of the 25th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2011: 301-306

[91] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 2787-2795

[92] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C] //Proc of AAAI. Menlo Park, CA: AAAI Press, 2015: 2181-2187

[93] Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes [C] //Proc of the 28th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2014: 1112-1119

[94] He S, Liu K, Ji G, et al. Learning to represent knowledge graphs with Gaussian embedding [C] //Proc of the 24th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2015: 623-632

[95] Lü L, Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170

[96] Barabási A L, Albert R. Emergence of scaling in random networks [J]. Science, 1999, 286(5439): 509-512

[97] Leicht E A, Holme P, Newman M E J. Vertex similarity in networks [J]. Physical Review E, 2006, 73(2): 026120

[98] Liu W, Lv L. Link prediction based on local random walk [J]. Europhysics Letters, 2010, 89(5): 58007

[99] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks [J]. Machine Learning, 2010, 81(1): 53-67

[100] Galárraga L, Teflioudi C, Hose K, et al. Fast rule mining in ontological knowledge bases with AMIE + [J]. The VLDB Journal, 2015, 24(6): 707-730

[101] Nickel M, Jiang X, Tresp V. Reducing the rank in relational factorization models by including observable patterns [C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014: 1179-1187

[102] Rendle S. Factorization machines with LIBFM [J]. ACM Trans on Intelligent Systems and Technology, 2012, 3(3): 57

[103] Chiang Y H, Doan A H, Naughton J F. Tracking entities in the dynamic world: A fast algorithm for matching temporal records [J]. Proceedings of the VLDB Endowment, 2014, 7(6): 469-480

[104] Dutta S, Weikum G. C3EL: A joint model for cross-document co-reference resolution and entity linking [C] // Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Cambridge, MA: MIT Press, 2015: 846-856

[105] Monahan S, Lehmann J, Nyberg T, et al. Cross-lingual cross-document coreference with entity linking [C] //Proc of the Text Analysis Conf. New York: NIST, 2011: 1-10

[106] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A Nucleus for a Web of Open Data [M]. Berlin: Springer, 2007

[107] Zhang T, Liu K, Zhao J. Cross lingual entity linking with bilingual topic model [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2013: 2218-2224

[108] Clark P. Knowledge Patterns [M] //Knowledge Engineering: Practice and Patterns. Berlin: Springer, 2008: 1-3

[109] Newell A. The knowledge level [J]. Artificial Intelligence, 1982, 18(1): 87-127

[110] Weng J, Lim E P, Jiang J, et al. Twitterrank: Finding topic sensitive influential twitterers [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining (WSDM 2010). New York: ACM, 2010: 261-270

[111] Bengio Y. Learning deep architectures for AI [J]. Foundations and Trends® in Machine Learning, 2009, 2(1): 1-127

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Fellow of China Computer Federation. His main research interests include cloud data management, Web data management, flash-based databases, privacy protection etc.

Du Zhijuan, born in 1986. PhD candidate at Renmin University of China. Member of China Computer Federation. Her main research interests include Web data management and cloud data management.