DOI:10.13530/j.cnki.jlis.174008

开放世界视角:面向多源词表的知识融合框架MtFFO研究

严承希　房小可

摘　要　为了解决多源词表的异构性和知识扩展的局限性问题,本文基于知识论中波普尔世界理论论证和辨析知识融合范式的概念及其有效性,并基于开放世界假设提出了面向多源词表融合的框架体系“MtFFO”,重点对外部环境信息输入框架——不同数据单元的多级化调整和交换系统,内部核心系统的知识模式匹配方式、冲突冗余识别方式,知识映射与合并策略,质量控制和知识扩展方法等逐步进行阐述和分析。MtFFO框架不仅是对知识融合方法体系的合理补充,而且为开放环境中多源词表构建和融合提供了一定的理论基础和技术参考。图5。表1。参考文献76。

关键词　开放世界假设　词表　元数据　知识融合

分类号　G250.7

“MtFFO”:Research of A Framework of Knowledge Fusion on Multi-source Thesauri from Open World Perspectives

YAN Chengxi & FANG Xiaoke

ABSTRACT

As a kind of normative and structured knowledge form, thesauri plays a significant role in the fields of information retrieval, enterprise knowledge management and service, as well as automated intelligent decision-making. Traditional closed-world thesauri have more and more severe drawbacks, such as high cost of expert-knowledge construction, low accuracy and narrow coverage of domain knowledge, obsolete knowledge-updating mechanism and weak capacity of knowledge expansion. It has become an important research domain that how to construct a proper, holistic and scientific system based on thesauri, as a powerful aid towards various systems and applications of intelligent information and knowledge management, which aims at knowledge ordering and sharing under the condition of the open interconnection. Related research has discussed in depth the structure, data sources, methods and language of thesauri integration, but lacks comprehensive induction and combing of multi-source fusion technology, theory and methods on thesauri, especially about how to embark on knowledge fusion in the open and interconnected environment.

Introducing knowledge fusion, this paper, first of all, defines the conception and connotation of the paradigm of “knowledge fusion” in the light of Popper's Three World Theory, and discerns its difference from “data integration” and “information integration”, as to clarify the category and scope of “knowledge fusion” fundamentally; Secondly, from the perspectives of OWA(Open World Assumption), this paper

通信作者:房小可,Email:xiaoke@buu.edu.cn,ORCID:0000-0001-7357-1558(Correspondence should be addressed to FANG Xiaoke,Email:xiaoke@buu.edu.cn,ORCID:0000-0001-7357-1558)

presents a more comprehensive and systematic theoretical framework of knowledge fusion called "MtFFO" in combination with KRAFT framework and other related existing models. The framework mainly contains the external environment, internal core, input unit, key factors and methods. Through feature comparison of different data source (database, metadata, and the table), external Environment Sub-frame of Information Input(EESII) is expounded as a multi-level architecture of knowledge adjustment subsystem. Next the knowledge flow dealt with EESII will receive the pattern matching and semantic recognition as the control and guidance of fusion strategies and quality to achieve optimization of knowledge transformation and composition; the respectively corresponding actual operation process is knowledge mapping and merging, and the former can be divided into two methods, knowledge aggregation based on homogeneity and logical conversion system on the meta-thesaurus, and the latter consists of simply merging, reductively merging, fully merging, or synonymy & hyponymy merging. The best way we recommend is taking target requirement, detailed context, cost and quality of knowledge unit and schema in thesauri as well as feasibility into consideration, combined with the different technology of matching, mapping and merging, and importing expert knowledge and reasonable control, so as to improve the efficiency, accuracy and robustness of the fusion system. Another innovation of the "MtFFO" framework is, to deal with the difficulty of thesauri expansion and enrichment, its introduction of metadata, general knowledge base on the basis of graph model and machine learning algorithm to mix all kinds of domain taxonomy, vocabulary and database together for data mining and knowledge discovery of high quality knowledge unit.

Although "MtFFO" is merely a theoretical framework without any validation by systems in practice, it can not only improve and develop the current theoretical system of knowledge fusion to a certain extent but also scientifically sort out and integrate a variety of relevant methods and crucial technology, providing thus theoretical basis and technical reference for solving the problems of thesauri interoperability, semantic understanding and self-enriching mechanism. 5 figs. 1 tab. 76 refs.

KEY WORDS

Open World Assumption.　Thesauri.　Metadata.　Knowledge fusion.

0　引言

词表(Thesauri)是一种由结构化规范术语组成的知识组织体系。它具有清晰的概念划分方法,如主题法、列举分类法、分面分类法等,以及明确的概念和语义关系(等价、等级和相关关系),如"用(Y)、代(D)、属(S)、分(F)、族(Z)、参(C)"等。Gilchrist[1]对不同类型的词表特征进行标注和区分,包括作为规范性通用知识宝库——叙词表(Thesaurus)、满足特定信息需求的概念或词的有效组织集合——知识分类法(Taxonomies),以及用于更精确知识表征和知识推理能力的共享概念模型的抽象形式化说明——本体(Ontologies)。从广义上来说,词表应该是所有具有知识性概念的语义化集合。词表系统的构建不仅在一定程度上解决了词汇控制、同义近义关系等文本标引的问题,也为不同情境下知识组织、知识服务和知识决策提供了规范性标准。目前大量国内外顶级通用词表和领域词表,如 Wordnet、《中图法》、《汉语主题词》表[2]、AAT(人文艺术词表)[3]、生物科学叙词表 ThesauForm[4]和医学诊断词表 NANDA taxonomy[5]都被成功应用于信息标引与检索[6]、企业知识地图和流程控制调度应用[7-8],以及医疗和农业信息资源[9]共享和辅助决策自动化等不同领域。

构建“开放、互联、有序”的网络生态环境已经成为当今企业、组织乃至国家信息化战略的重要课题之一,这也对传统的基于词表的信息组织和知识管理系统的应用提出了更高的要求,具体来说表现在两个方面。

(1)现阶段词表主要用于封闭的信息环境,良好的规范模式和可控术语可起到知识指导和概念控制的作用。但不同词表之间概念、类别、模式结构差异很大,而单一词表的构建不仅消耗极大的人力和智力成本,更难以全面覆盖到所有的知识领域范畴,也就无法满足用户的动态主体性知识需求和开放共享的要求。

(2)随着大数据时代的来临,词表陈旧的知识更新机制和薄弱的扩展能力问题日显突出。虽然面向特定领域的词表系统和集成数据库技术可以解决信息的整序问题,如企业运营中基于 Taxonomy 的组织服务和决策应用[10-11]等,但受制于封闭环境的假设和单一的业务情景,词表不能提供通用的知识模式,也无法从语义层解决异构、歧义问题,更难以解决知识价值增益与创新的问题。

因此,如何构建科学和完善的词表融合机制,使之更智能、更有效地服务于情报知识管理和科学决策,实现开放互联环境下知识有序和共享,是目前亟待解决的问题。

知识融合是从信息整合和知识科学视角下的知识库整合技术逐步衍化而形成的知识管理方法[12]。一方面知识融合需要在信息整合的基础上对不同类型的信息源进行深入的挖掘和建模,使内部知识价值和概念体系显性化,实现知识存储和服务型应用;另一方面也需要将已有事实库、知识库等规范性架构系统进行改造和融合,实现开放互联环境下的全知识共享和知识复用。本文提出开放环境下面向多源词表的知识融合框架,结合不同异构数据源如元数据和数据库系统进行知识语义扩充,能在一定程度上丰富和完善现有知识融合理论体系,科学梳理和整合不同知识融合的关键技术和方法,同时对于解决词表互操作、语义理解、自动丰富机制问题提供了理论性指导和建议。

1 研究综述

目前,国外词表融合领域的相关研究主要集中在结构、数据源、方法和语种四个方面。

(1)词表级信息整合:对已有词典、主题词表和叙词表等传统规范性词表集合进行模式匹配和逻辑建模。如美国国家癌症生物信息研究中心(National Cancer Institute Center for Bioinformatics,NCISB)基于旨在提供基因技术和诊所咨询服务的生物医学集成词表所开发的 EVS 集成化知识管理系统[13];美国国家癌症研究所(National Cancer Institute,NCI)开发的 CDE 系统整合了 NCI 词表概念和部分术语模式,其中集成化的知识词典可以作为癌症研究数据描述的受控词表[14];Bakken 等[15]将北美护理诊断协会词表 NANDA 与 SNOMED 概念词典进行语义相似性映射,并实证评估了其在欧洲标准化委员会(European Committee for Standardization,ECS)分类结构模型和国际标准化组织(International Organization for Standardization,ISO)的参照术语模型(Reference Terminology Model,RTM)的有效性;在产品和服务集成领域,Park 等[16]构建出集成词表 IPS taxonomy,该词表集成和涵盖了企业产品服务流程和规范所需的概念和关系。

(2)词表、数据库系统以及元数据的集成:将数据库中所存储的数据单元和电子元数据描述记录进行知识抽取和转换,形成与词表聚合的辅助来源。Shah 等[17]将斯坦福医学组织微阵列数据库 TMAD 中人体组织样本的文本数据进行语义标注,然后基于 NCI theasurus 进行知识映射和处理,实现语义化查询。Shironoshita 等[18]使用癌症生物医学信息网格 caBIG 的概念和元数据进行语义化建模,实现知识表示和基于 SPAQRL 的语义化查询。

(3)基于本体工程的知识融合:本体工程的建模方法可以提供客观完整的集成化流程,包括适应性本体描述语言、标注规则、本体构建工具以及本

体映射与合并的技术支持等，典型应用如语义标注的多分面产品族本体（Multi-Faceted Product Family Ontology，MFPFO）的建模[19]。Lacoste 等[20]则首次提出云计算环境下本体即服务的框架（Ontology as a Service，OaaS），并将 UMLS 用于超大词表 Metathesaurus 的本体驱动化集成和构建。

（4）多语种词表的知识融合：基于多语种文本处理技术（Multilingual Text Processing）对不同语种的术语进行概念标准化、同义词替换识别和文档匹配，并通过本体对齐和模式识别方法实现跨语种之间语义互联和知识共享，如基于 EuroWordNet thesaurus 和 depedia 的跨语种本体融合案例[21]。

相比之下，国内在词表融合领域的研究较少，主要集中在领域主题词表集成[22]，分类主题法改造[23]，基于本体、词表和元数据的混合集成，以及词表自动丰富等方面[24-25]。由此可见，目前国内外相关研究没有对词表的知识融合体系架构、方法和技术进行全面整合和理论构建，更缺少对开放环境下如何进行知识融合问题的探索，这也是本文论述的重点和力图解决的主要问题。

2　知识融合的范式研究

目前数据集成、数据整合、信息整合、信息集成等术语被大量混用在不同的研究和分析中。数据、信息和知识之间的关系并非处于静态和完全对立的状态，而是共存于不断的相互转化和动态演变中。为了明晰不同术语和理论之间的区别，凸显出知识融合方法理论的独立性、完整性和科学性，需要对知识融合的范式进行研究和辨析。

在情报理论研究中，知识论学派将波普尔世界 3 理论看作知识管理和组织的基础，即世界 1 是物质世界，世界 2 是精神世界，世界 3 是知识世界，其中知识世界是客观知识理论、问题和论据所存在的具有自主性和永恒性的世界[26]。如果从知识论角度出发，知识融合应该属于世界 3 所考虑的范畴，显然数据集成和信息整合是属于世界 1 或者世界 2 的知识范围内。Gulledge 提到"信息整合"是以不同形式的接口信息传递的方式使那些从来不会一起工作的多个应用协同式工作和操作起来的过程[27]，也就是说整合或者集成并不是一个自然形成的过程，其隐含的意义是"需要借助外部力量"，且更侧重于将不一致的信号和系统进行协同式管理和组织。基于此，本文认为数据集成和信息整合应该是物理层面的协调性归并，没有涉及语义化和知识关联性的考虑，即不满足波普尔理论提到的自主性要求，因此"融合"显然不适用于以数据和信息作为对象。相反，知识的共享和叠加是理论价值层面的考量，是含有逻辑论据推理而自然形成的高阶领域，相比于物理性需要外力进行合成，知识概念之间演变更加柔和和自然，是潜移默化的交互和改变，从这一点来说知识融合也是情报研究和分析的最终目标。

为了进一步弄清楚数据集成、信息整合以及知识融合之间的关系，本文通过图 1 来展示知识论范式下知识融合的演变模式，具体包含了数据层、介质层和语义层，分别对应于数据集成、信息整合和知识融合。

数据层主要对异构冗余数据进行集成化处理，包括对不一致的数据格式、缺省数据、重复数据的清洗和规约等，通过数据建模和合成方法对数据作序化处理，从而获得更准确和更有意义的信息。介质层则是信息整合的范畴，不同于数据集成，信息整合更侧重于信息的交换和共享，即如何使不同系统和平台的信息可以进行协作式工作，从而实现信息资源科学服务和辅助决策。相关研究则主要围绕中间件、网格技术和分布式存储架构进行开拓和分析，如中间件 CoBase 项目[28]，信息排序算法处理[29]以及传感器之间信息交互[30]等。知识融合属于语义层面范畴，即集中解决如何实现语义化查询和推理，有效消除概念歧义，建立合理的知识质量评价体系，并深入挖掘显性知识集合，实现隐性知识扩展与知识创新。

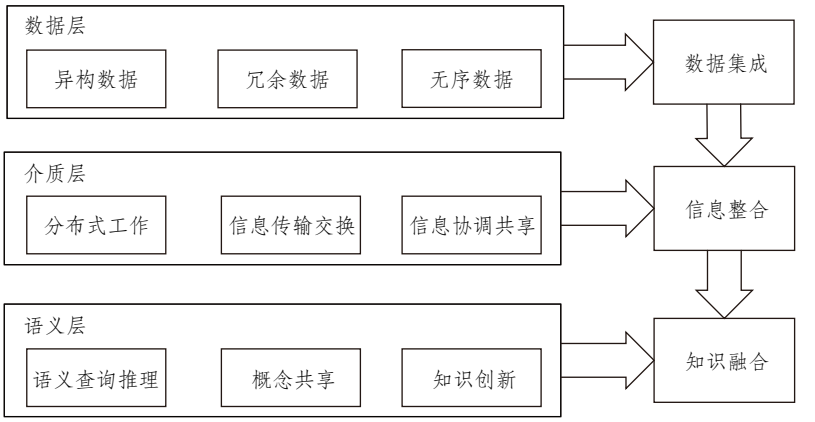

图 1　知识论范式下知识融合的演变模式

3　开放世界视角下面向多源词表的融合框架 MtFFO

传统意义上的封闭世界假设(Closed World Assumption,CWA)具有丰富的知识来源,而开放世界假设(Open World Assumption,OWA)是一种带有后实证主义色彩的不确定信息环境的逻辑解释范式,即假设尚不清楚是真是假的事实规则可能是真实的,或者将信息的缺失解释为未知信息,而不是带有负面性的信息(Negative Information)[31]。虽然开放世界的学说视角可能会影响单个知识源或知识库的建设和推理性能,但是对语义逻辑推理的准确性和知识价值的淬炼具有显著的提升效果。因此,对整个开放互联网络而言,OWA 无疑是更佳的选择。进一步来说,基于开放世界假设的词表知识融合将摒弃封闭环境中仅在组织内部或者组织联盟之间知识流传递和共享的局限,将融合的对象范围扩展到互联环境下的全领域网络(包括企业与企业、企业与政府以及企业与个体、个体与个体等)。显然 OWA 视角下多源词表融合的框架是更加科学和全面的理论框架。

早期学者 Nikolai 等[32]提出了"多词表系统架构"(Multi-Thesaurus Systems),通过阐述词表环境(Multi Thesaurus Environments)、词表转换系统(Thesaurus Switching Systems)和词表复合物(Thesaurus Compounds),奠定了早期词表集成的理论体系。针对异构知识库的可扩展性和知识集成问题,成熟框架系统 KRAFT 项目提出了一种知识融合架构体系,该系统包括外部资源和内部组件两大部分[33],其中外部资源主要是用户代理和服务资源,而内部组件则是系统的核心融合部分,按照由内至外的顺序主要包含调解器、触发器和包装器。这些理论和方法对开放环境中面向多源词表框架的构建提供了理论基础和指导依据。鉴于此,本文从输入单元、系统结构、因素与方法等层面提出了开放世界视角下面向多源词表的知识融合框架体系"MtFFO"(Multiple thesauri - oriented Fusion Framework in Open-world Assumption),具体架构如图 2 所示。

类似 KRAFT 项目,MtFFO 系统结构也可以分为外部环境和内部核心两个构件,其中外部环境主要是三类数据源:多源词表、元数据和数据库系统,一并经过处理作为框架的输入单元。多源词表是不同词表之间的聚合体,是对规范的概念术语进行有序排列和整合。元数据和数据库系统是异构型输入信号,主要是非规范性的信息集合。内部核心则包括两个部分,即转换复合过程和输出单元。转换与复合过程是对不同知识词表和数据源进行词表类目、实体和

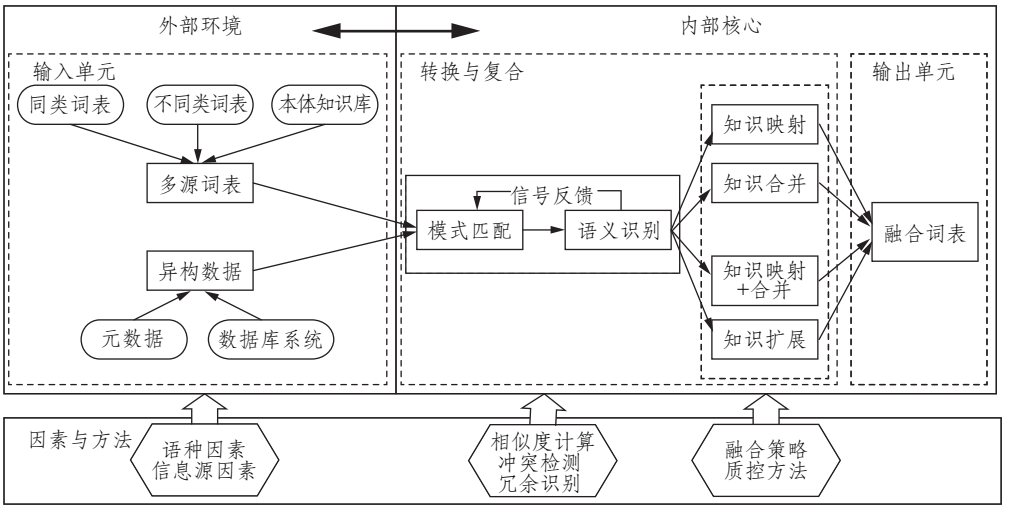

图 2　开放世界视角下面向多源词表融合的框架体系"MtFFO"

关系之间模式匹配和语义识别，其中语义识别需要对模式匹配进行实时控制和信号反馈，并通过相似度计算和冲突冗余检测机制确保不同词表之间融合的唯一性、一致性和完整性，并为词表知识复合提供指导性和质量评估标准，从而根据特定的融合策略对不同情境中的词表和数据源进行结合，如知识映射、知识合并以及知识扩展等，可以说转换复合是整个知识融合框架的关键步骤。输出单元则是不同词表和数据源经过转换与复合所获得的统一知识结构模式。

3.1　外部环境

（1）输入单元的数据源

外部环境是输入单元及其所在情境内容的总和。外部环境的对象是不同词表、元数据和数据库系统。其中多源词表既包括类型和结构的不同，各种标准通用词表如分类词表（Classification）、主题词表（Subject Headings）、叙词表（Thesaurus）和满足企业组织服务流程和个性化体系的知识词表 Taxonomy 等，也包括语种的不同，如"中图法"和欧盟多语种叙词表[34]。本体（Ontology）是一种比 Thesaurus 和 Taxonomy 在概念、约束和术语关系表述方面更精确的"词表"[35]，特别是对"相关关系"模糊松散的定义进行了明确的形式化说明，可以看作是对传统词表的丰富。尽管目前大量多源词表是基于本体语义网技术和描述逻辑理论所构建的，大有被本体取代之势，但仍有学者表示传统词表的叙词概念和知识体系仍具有本体工程所不具有的优势[1]。本文也认为本体和词表之间存在一定的差异性，并将本体知识库作为与传统词表不同的数据来源进行专门注释。狭义的元数据来源于识别、描述和定位电子资源的数据记录，广义上可以将元数据扩展为散布在开放网络中的各种对象描述记号，即数据符号单元。相比格式互异的元数据集合，虽然数据库系统具有更统一的信息管理机制和结构性表达，但由于面向的服务需求不同，技术接口也不尽一致，数据库的集成化管理已成为企业级信息管理和共享最为关注的问题。

相比词表，这些异构数据源虽然不具有规范的概念和明确的实体关系，但是数量巨大并且集中体现了最广泛的用户需求，因此完善多源词表知识融合体系必须将其纳入考虑的范围，同时也作为知识融合的高阶形式"知识创新"和"知识扩展"的重要来源。输入单元不同类型数据源的特征对比如表 1 所示。

表 1　输入单元不同类型数据源的特征比较

类型	词表	元数据	数据库
格式	词表内部一致性较高,不同词表之间具有类似的体系	元数据之间表现为异构性,一致性较低	数据库内部数据格式较一致,但不同数据库之间差异较大
单元	具有规范概念术语及其语义匹配	非规范数据,不同元数据之间差异极大	非规范信息,具有类似的体系,但不具有明确的概念成分
环境	私有组织或者公共机构所构建,但一般用户难以接触	可为任何人员创造,大部分公开,易获取	私有组织和企业内部建立,少部分提供开放接口
结构	结构化数据,具有明确的知识单元	非结构化或者结构化数据,以纯数据形式分布居多	结构化数据,信息的排列和整合具有一定的规则
数量	稀少	非常多	较少

(2)外部环境信息输入框架

Hepp 等[36]在比较工业分类系统 Taxonomy 的转换和集成研究中指出“输入标准原始语义的表达和解释是决定模型生成本体效能好坏的重要因素,特别是对于 Taxonomy 的关系”。本文根据不同数据源的特征和属性,对外部环境的数据输入单元交换和预处理流程进行总结和归纳,提出一种多层级化外部环境信息输入框架,如图 3 所示。

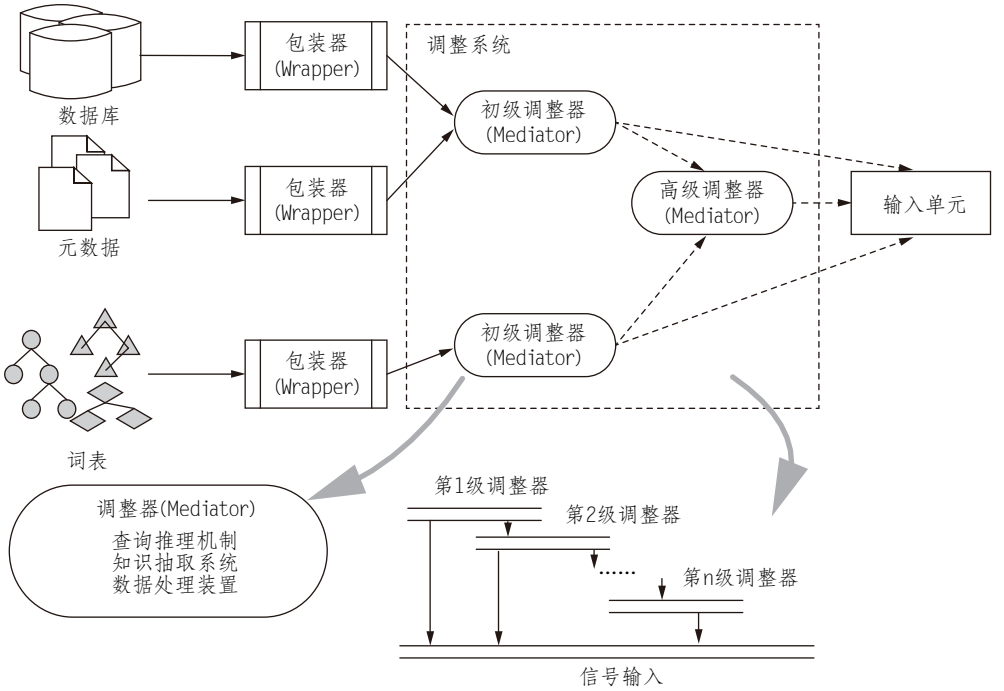

图 3　外部环境中输入单元的 MtFFO 子框架

不同数据源进入调整系统需依赖不同的交　互接口,交互接口层可以用包装器(Wrapper)与

调整系统进行互联，但互联操作并不进行具体的知识抽取和挖掘。调整系统本质上是一个不同级别调整器进行知识交换和处理的容器。调整器（Mediator）是早期信息集成系统中智能代理技术的应用形式，如构建在对象定义语言（Objective Definition Language，ODL）和协调器基础上的 MOMIS 分布式信息集成系统（Mediator Environment for Multiple Information Sources，MOMIS）[37]。如果将调整器看成独立的知识调整单元，那么它不仅应该具有数据基本处理和概念传递的功能，还应该具备知识抽取、查询和消息回馈打包能力以及完备的推理机制。

与此同时，引入多级化调整系统架构可以解决不同数据来源所夹带的知识异构性和噪声冗余问题。按照自下而上的原则，从最初的输入信号开始，初级调整器将知识交换的结果提交给上一级调整器进行智能判断和再加工，同时高层调整器需要及时回馈处理结果，形成知识闭环。具体而言，对于词表来源的知识流，调整器仅需对简单概念、属性和关系进行交换和处理，所需的调整器数目较少、层次较低；而异构数据源或者信息流是只经过包装器过滤的一般数据信号，并非知识单元，需要进行知识抽取、知识编码以及知识挖掘等步骤才能获取初步具有规范概念的知识单元，当然异构数据源所需的调整器数目和层次相应较高，开销较大。所以多层调整器需要逐级设置完善调控机制，其中重点问题是如何设计可靠的挖掘算法和优化的控制模型，以平衡好计算开销和知识抽取精度两个重要因素。值得一提的是，功能更强大的多级化调整器（Mediator）机制需要适应跨语种词表的互操作和语义推理[38]，类似的技术和系统如 Jurwordnet、领域专家分析系统 THALEN[39]等，与业务情景和需求紧密相关。

3.2　内部核心

（1）模式匹配与语义识别

模式匹配与语义识别是 MtFFO 内部核心的管道和先验条件，是对不同知识源进行异构知识表征和知识元定位的识别技术手段。通过定位和标注不同或相同的知识体，根据知识需求确定有效的知识关联，并进行再规范、简化和扩充，如 Po 等[40]将词汇知识抽取（Lexical Knowledge Extraction）过程分为图式语义化和词汇标注规范化两部分，这种语义标注可能是对 Taxonomy 的术语再分类和词典模式再设计等。按照概念、关系和属性等知识元素，模式匹配可以分为模式层次匹配、实例层次匹配、结构层次匹配和基于语言或限制规则的匹配等[41]。典型的匹配技术是相似度计算和模糊集合算法。不同实例和类别之间具有相似的语义环境，包括相同的上位类、下位类、同义关系和属性交叉等[42]，通过计算不同词表的实体和类别特征之间的相似性测度，可以为知识映射和合并提供指导依据，典型的实例如 Batet 利用语义相似度计算本体、叙词表和生物医药领域语料库的实证研究[43]，以及 Castano 根据叙词表 Artemis 的语言附属度开发的计算方法 H-Match[44]。模糊匹配则是一种基于模糊集合和关系推理的非精确式匹配方式，典型的案例如智能个人信息分层 iPHI，基于软计算理论中的模糊匹配算法，将 Taxonomy 的概念与个人信息的元数据进行匹配，实现优选术语的个性化查询和检索[45]。

模式匹配是为解决知识映射和合并策略问题，而冲突检测与冗余识别则属于质量控制的范畴。冲突检测是基于一致性原则的知识处理技术，对不同词表和知识流中的概念、实例以及属性关系等进行知识单元的矛盾消解和知识流的语义消歧，是在进行策略性知识融合前所必须进行的知识异常识别操作，如从认知特征的角度来测量不同实体类别语义匹配距离的相似度计算方法——"匹配距离相似度测量法"（Matching-Distance Similarity Measure，MDSM），可在一定程度上消除传统语义测量所忽略的不对称性和情景依赖性因素问题[46]。由于开放性世界环境中真实性对不确定性的容忍，不是简单的"是或非"问题，因而冲突检测技术还需要包括真实性甄别，即如何识别和确定不同词表

知识单元的真值和真值空缺问题。常用的检测方法是专家规则和统计学习模型,如启发规则[47-48]、准确性识别算法(TRUTHFlNDER)[49]、贝叶斯概率模型[50]以及分布式信任修正技术(Distributed Belief Revision)[51]等。另外,知识流之间不仅存在各种冲突性问题,同时也不可避免地遇到不同程度的概念和属性冗余问题,特别是描述逻辑规则中互斥关系的冗余性错误问题[52]。已有方法,如对分 HS—树(Hitting Set Tree)模型[53]以及实例冗余冲突检测算法[54]都是基于机器学习算法进行冗余识别,也有学者[55]引入时间属性概率因素对实体演化模型做出了一定修正。

匹配和识别的过程之间并不孤立,二者之间需反复进行正确信号的反馈和传递,从而实现知识转化和复合的最优化。

(2)知识映射与知识合并

知识映射与知识合并是模式识别后对不同知识流和词表中的概念、关系和属性所进行的知识关联化操作。知识映射是在不改变原始词表的基础上,对概念知识流和词表以及词表之间进行有效语义化连接的方式。概念映射需要对互斥关系概念、重叠概念以及属从概念进行更准确的定位和互连,而属性和关系映射则主要包括对不同词表的等价关系、偏序关系和相关关系的语义化处理以及对属性约束集合的对应关系构建等。多源词表的知识映射思想可以分为两类情况:基于词表同质性的知识聚合和元词表逻辑转换系统的构建,具体如图 4 所示。

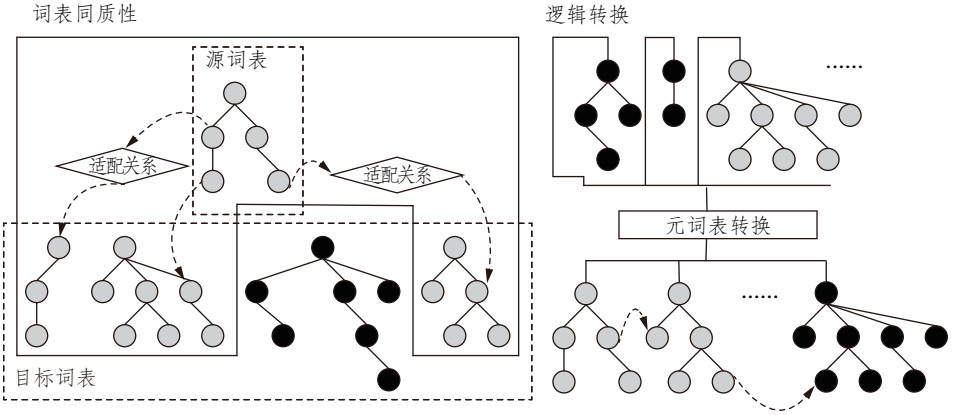

图 4 多源词表知识映射方法

元词表逻辑转换系统的构建是基于专家知识映射和控制方法的统称。传统专家知识映射方法是基于人工数据匹配和知识工程技术对知识单元进行映射。这种方法较早用于大型词表的构建,如北美医用词表 NCI Theasurus[56]和 FAO 维护的多语种词表 AGROVOC[57]。为了解决多对象和多语种问题,Kramer 等[58]提出一种面向非专家模式的跨语言检索工具——叙词表联合的元词表映射系统(Meta-thesaurus of Thesaurus Federation),并已用于知名的国际语词表——通用欧洲多语环境叙词表(General European Multilingual Environment Thesaurus,GEMET)系统中。这种方法虽然可以进行人工修正和反馈控制,属于可控词表映射装置,但并没有针对映射优先顺序进行策略判断和自动化归类,融合词表构建开销较大。

词表同质性聚合是基于数据挖掘技术和统计模型对大量输入词表和知识单元进行自动化建模和匹配,实现同质性词表之间的识别和高度聚合,进而有效提高知识映射的效率。目前机器学习算法是针对不同数据集合进行有效聚合和归类的主流自动化技术。监督分类学习的

思想是将源词表中的术语看作特征变量进行训练，同时把需要融合的目标词表的术语看作不同的类别，进而转化为一个样本多标签分类问题，如 Hardiker 等[59]对国际护理实践分类法框架和北美护理诊断知识分类表术语（North American Nursing Diagnosis Association，NANDA taxonomy）之间的词表融合就采用了自动分类式映射，并利用组合术语语言 GRAIL 对融合知识进行统一表征。按照模型训练集的不同，一般可以把融合扩展算法归为两类。①显性映射方法：指若干待映射的对象根据已有分类特征进行归类决策，整个过程源词表数据一直是训练集，目标词表中术语集合不参与训练阶段，典型的算法如最大熵模型[60]。②潜在映射方法：采用二步训练法，第一阶段使用半监督分类学习训练源词表，第二步学习阶段则用与源词表存在潜在映射关系的目标集合进行半监督分类。由于潜在映射方法的数据训练集合显著增多，且不同词表术语概念范畴存在较大语义距离和不同特征维度，为了保证分类的准确性和概念关系的精准度，往往会采用集成学习算法进行建模，如 Bootstrapping 迭代学习[61]。

相比监督分类算法需要大量的标准训练集和专家知识的判定，Madhavan[62]提出效率更高的叙词表聚类模型 Cupid，即根据不同词表之间具有的相似结构和计算语言特征进行匹配（Linguistic Matching 算法），对特征相近的词表优先聚合，有效提高词表知识融合的计算开销和精确度。其他一些更优化的算法则可结合多种语义维度，如 HMatch 2.0 结合网络分析理论将结构角色作用考虑到本体演化中，结合特征维度（类别、属性）和 RDF 三元组个数等进行加权计算作为映射函数的输入，模型不仅考虑了原始本体内在的关系，更纳入了新连接本体的影响作用[63]。

冲突和冗余检测的目的是对异常知识单元的识别，知识合并则是在此基础上采用一定策略对异常点进行排除和优化的融合模型修正方法。如果将知识映射看作静态的知识融合技术，知识合并则更加强调对质量的动态控制，即对词表和知识单元的概念、关系以及属性的增删和修改操作等。Yang 等[64]将分类树集成技术（Category-tree Integration Technique）用于词表分类之间的语义化操作（分裂、插入）等，获得比映射更佳的融合效果。按照合并重构方式，本文可以将词表知识合并分为简单合并、约简合并和完全合并三种，如图 5 所示。

①简单合并：需要先对知识融合需求和各个源词表的结构、模式和关系进行全面的理解，然后基于匹配和度量对不同词表中同层概念进行知识聚合，通过添加更上位的概念结点或者创建不同词表概念之间的关系连接，实现简单的知识合并。由于合并策略强调质量控制，因此合并过程中需要适当地对冗余概念、关系或者需求外的知识单元进行舍弃，当然这也意味着源词表的结构和知识单元可能被破坏。形式概念分析（Formal Concept Analysis，FCA）方法是一种基于语境背景知识合并和集成的构造方法，概念格是 FCA 的核心数据结构，属于简单合并的一种典型方法。Sahoo 等[65]基于 FCA 抽取和集成多个癫痫分类系统数据，如国际抗癫痫联合会的推荐术语集、美国国家神经障碍与中风研究所的通用数据和神经电磁本体（Neural Electro Magnetic Ontologies），从而构建出癫痫及其发作主题的 Taxonomy EpSO。Stumme 等[66]则提出支持合并过程保持全局结构描述、自下而上的本体合并方法 FCA-Merge。该方法将从源本体集合中抽取的指定领域实体按照概念和关系进行数学建模，通过本体专家进行翻译和合并，然而其缺点是可能造成本体质量下降。

②约简合并：在简单合并基础上的再修整和精炼，这一阶段的冲突性测试和概念质量需要被重点考虑。一种可行的方式是局部概念关系的剪枝和合并，如基于"ATOM"原则的概念和关系冗余消解的合并方法[67]，具体是根据"二步法"思想：首先对源 Taxonomy 和目标集合进行关系匹配形成概念图，第二步是对概念图的合并和精炼。另外一种则是全局视角的合并方法。

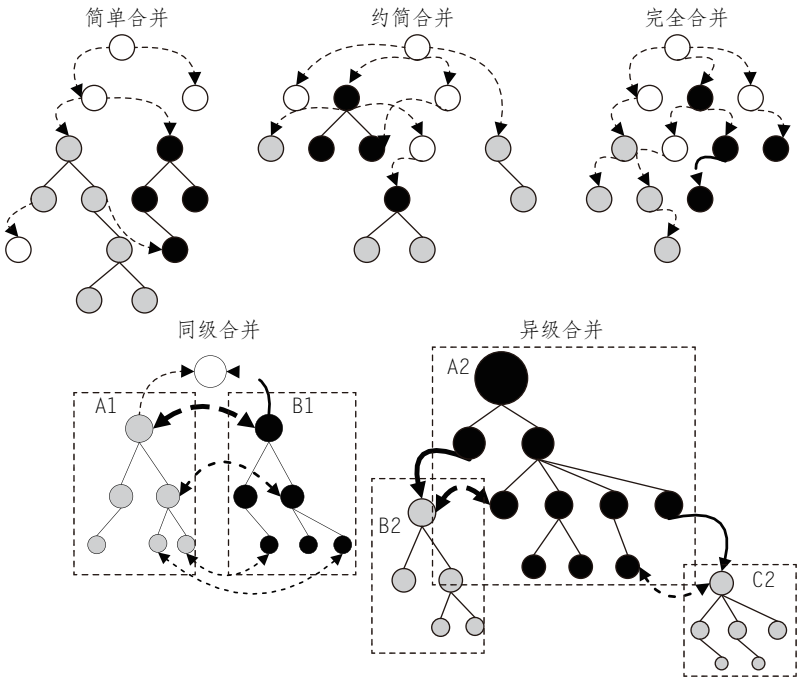

图 5　多源词表知识合并方式与策略

随着本体资源的复杂性增加,局部合并修正可能无法将新本体结构展示出来,Rouane-Hacene等[68]提出基于关系概念分析(Relational Concept Analysis,RCA)的合并方法可以在清除源本体的原始结构情况下进行全局调整和重构(Reshuffling & Refactoring),从而降低缺省实例的错误率。这种处理方式效率适中,适合中小型业务场景下的词表融合。简单合并和约简合并属于不完全知识融合的范畴,即从源词表集合出发,在逐步合并和修整中加入知识需求的限制,因此合并后生成的词表缺乏整体性的模式和概念体系,可能存在认知误差。

③完全合并:从本体集成编辑和设计工程思维出发,以知识融合需求为导向,事先构建出合理的融合词表概念体系和概念集合,然后进行关系和属性函数的拼接,可以看成是一种新超大词表的重构过程。Bellamy 等[69]提到将逻辑模式的关联作为开发集成词表的基本原则,进而实现不同系统之间词表的合并,但重构时耗成本比较高。

根据词表重构策略方法的不同,知识合并也可以分为同级合并和异级合并。图 5 中对同级合并和异级合并的模式进行抽象性描述,即两个同级别(词族首之间)是等价关系,或者可以组成一个新的上位概念的同级下位互斥概念对(见 A1 与 B1),叙词表联并(Thesaurus Unions)就是典型的基于同级合并策略的词表[70];另外一种是异级合并策略,考虑到不同词表的词族首可能存在上下位关系,上位词表一般是高阶的通用型词表,具有更宽泛的知识范围和抽象概念,如 A2,而下位词表则一般是领域词表和行业分类表(Taxonomy)等,含有更具体和精细的概念和关系,如 B2 和 C2。Roulin 提出词表耦合法(Coupling of Micro-thesauri)就是针对不同粒度的词表进行结合的策略形式[71]。由于领域词表的上位概念和通用词表的下位概念在一定意义上可能存在同质性,或者出现概念重叠现象。因而词表耦合还提供新词表同层次的概念合并策略,如只保留最规范的知识单元,从而有效消除知识的重叠性问题(如图 5 中双向箭头虚线连接就是知识重叠区域)。

Raunich 概括了当前本体融合集成技术的问题[72]，如单一策略合并和匹配的局限，半自动模式集成的不足，匹配和合并的划分问题，以及缺乏灵活性和高成本的专家知识集成辅助设计问题等。由此可见，不同知识映射与知识合并的方式具有各自的优势。因此，在进行词表知识融合的时候，需要充分考量目标需求、词表知识单元和体系的质量，以及融合成本和可行性的平衡，同时可以将不同匹配、映射和合并技术进行有效的结合，并合理引入专家知识控制，从而提高获取知识融合系统的准确度和鲁棒性。

(3)知识扩展

由于目前规范性词表较少，且词表概念术语的扩展和丰富陷入瓶颈。大量面向多源词表的集成技术主要以私有词表和领域本体为对象，大大限制了知识融合的知识来源范围，不利于开放互连环境下词表的丰富和扩展，更无法对隐性知识作进一步挖掘。Nigam 等[73]考虑将知识库和大量的元数据进行结合，即将文本语义标注的元数据(如基因和组织相关的微阵列)与 UMLS 概念知识库结合构建出原型系统，实现了本体驱动下的生物医学数据的语义标注和查询，其中概念识别工具是知识融合的关键点，该系统采用来自密歇根大学开发的字典标注工具 Mgrep[74]。关联数据技术是语义网环境中 Web 元数据进行语义化链接的最佳实践方案，用户可以通过 HTTP/URL 协议进行数字化资源对象(Thing)的访问。如 Fernandez 等[75]将关联数据技术用于在线教育信息互联，并基于 FOAF 和 W3C Ontology 对 Youtube 等网站进行视频资源类别的语义化标注，然后基于 ODP 的 TextWise 主题分类工具实现了文本归类和融合。

除此以外，随着关联数据的发展，开放环境下本体工程所构建的大型通用词表系统逐渐成为知识扩展领域的另外一大知识来源，如大众开放式分类目录网站 DMOZ、维基百科知识库 DBpedia 等。DBpedia 是德国莱比锡大学以维基百科为蓝本进行人工编辑的大型开放百科型词表，具有 97 种语言和近 3 500 万个对象描述；德国马克斯·普朗克研究所研制的 Yago 系统对 Wordnet 和 DBpedia 进行实体抽取，创造出包含近 1 000 万个实体和 1.2 亿条关系的跨领域知识库；Freebase 则被 Google 所收购，并用于知识图谱和搜索引擎的优化。

本文认为利用映射合并的方法将上述传统词表(组织私有词表)、元数据、各种数据库和通用知识库进行有效融合，实现概念的自动丰富和关系、属性的扩展，将催生出更新、质量更高的知识单元。同时从知识创新的角度来看，高层的知识融合应该是知识发现，即不仅仅是词表之间的显性知识聚合，更需要对隐性知识的挖掘和扩展。基于机器学习和统计模型的知识扩展和优化是一种多源词表知识融合的补充手段，例如 Rhodes 等[76]将多目标分类法(Multi-target Classification)引入知识扩展，具体是基于关联增量学习策略的神经网络算法对共现数据进行分类，有效识别词汇中元素之间的关系和隐知识结构，实现了术语概念的预测，并证明了其性能优于朴素贝叶斯模型。

4　总结与展望

尽管作为规范化知识组织工具的各种语义词表和词典已经被各类行业领域和组织机构所使用和认同，但是受制于知识异构性、领域局限性以及专家知识驱动的词表静态更新策略，词表的发展和普及应用受到了一定的限制。与此同时，随着大数据和信息技术的普及，开放世界领域中大量元数据和数据库系统资源造成严重的信息过载，进而知识融合技术应运而生，为知识有效聚合和解决异构性互操作问题提供了全新的视角和手段。结合相关研究理论，本文首先探讨了知识论视角的知识融合范式，结合波普尔世界理论对知识融合概念的模糊性进行辨析，然后基于开放世界假设提出了面向多源词表融合的框架体系"MtFFO"，并对框架的主要构件"外部环境"和"内部核心系统"进行阐述。一方面描述了外部输入单元的多级化知识交换、调整和过滤的处理模式，另一方面则集中在

知识流的模式匹配、识别、映射、合并理论和方法的阐述,重点讨论了基于不同融合策略和质量控制方式的映射和合并方法,并补充了知识库和元数据的融合方法和技术,将词表的自动丰富和知识扩展机制整合到框架中。本文提出的MtFFO框架仅属于理论定性的研究方式,并没有通过实验进行验证。未来的工作将考虑如何从技术角度对此框架进行原型系统的研发,解决多源词表知识融合的实践问题,从而证明MtFFO框架的有效性和可行性。

参考文献

[1] Gilchrist A. Thesauri, taxonomies and ontologies-an etymological note[J]. Journal of Documentation, 2003, 59(1):7-18.

[2] 封庆阳. 试论分类主题一体化的现实道路——关于《中图法》的类目与《汉语主题词表》的叙词对应规范问题的探讨[J]. 北京图书馆馆刊, 1986(4):12-20.(Feng Qingyang. The practical approach of classification subject integration: discussion on the issue of standardization corresponding to category in *CLC* and thesaurus in *Chinese Thesaurus*[J].Journal of the National Library of China, 1986(4):12-20.)

[3] Soergel D. The Art and Architecture Thesaurus (AAT): a critical appraisal[J]. Visual Resources, 1995, 10(4): 369-400.

[4] Laporte M A, Mougenot I, Garnier E. ThesauForm—traits: a Web based collaborative tool to develop a thesaurus for plant functional diversity research[J]. Ecological Informatics, 2012, 11(3):34-44.

[5] Henry S B, Warren J J, Lange L, et al. A review of major nursing vocabularies and the extent to which they have the characteristics required for implementation in computer-based systems[J]. Journal of the American Medical Informatics Association Jamia, 1998, 5(4):321-328.

[6] Binding C, Tudhope D, Blocks D, et al. Query expansion via conceptual distance in thesaurus indexed collections [J]. Journal of Documentation, 2006, 62(4):509-533.

[7] Vanpoucke E, Boyer K K, Vereecke A. Supply chain information flow strategies: an empirical taxonomy[J]. International Journal of Operations & Production Management, 2009, 29(12):1213-1241.

[8] Asadi A, Mancuso V. Asurvey on opportunistic scheduling in wireless communications[J]. Communications Surveys & Tutorials IEEE, 2013, 15(4):1671-1688.

[9] Wu C H, Huang H, Nikolskaya A, et al. The iProClass integrated database for protein functional analysis[J]. Computational Biology & Chemistry, 2004, 28(1):87-96.

[10] Doll W J, Torkzadeh G. Developing a multidimensional measure of system-use in an organizational context[J]. Information & Management, 1998, 33(4):171-185.

[11] Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials [J]. International Journal of Medical Informatics, 2011, 80(6):371-388.

[12] 邱均平, 余厚强. 知识科学视角下国际知识融合研究进展与趋势[J]. 图书情报工作, 2015, 59(8):126-132.(Qiu Junping, Yu Houqiang. Research progress and trends of international knowledge fusion at the perspective of knowledge science[J]. Library and Information Service, 2015, 59(8):126-132.)

[13] Covitz P A, Hartel F W, Schaefer C F, et al. caCORE: a common infrastructure for cancer informatics[J]. Bioinformatics, 2003, 19(18):2404-2412.

[14] Nadkarni P M, Brandt C A. The common data elements for cancer research: remarks on functions and structure[J]. Methods of Information in Medicine, 2006, 45(45):594-601.

[15] Bakken S, Warren J J, Lundberg C, et al. An evaluation of the usefulness of two terminology models for integrating nursing diagnosis concepts into SNOMED clinical terms[J]. International Journal of Medical Informatics, 2003, 68(1-3):71-77.

[16] Park Y, Geum Y, Lee H. Toward integration of products and services: taxonomy and typology[J]. Journal of Engi-

neering & Technology Management,2012,29(4):528-545.

[17] Shah N H,Rubin D L,Espinosa I,et al. Annotation and query of tissue microarray data using the NCI Thesaurus [J]. BMC Bioinformatics,2007,8(1):1-9.

[18] Shironoshita E P,Jean-Mary Y R,Bradley R M,et al. SemCDI:a query formulation for semantic data integration in caBIG[J]. Journal of the American Medical Informatics Association,2008,15(4):559-68.

[19] Lim S C J,Liu Y,Lee W B. A methodology for building a semantically annotated multi-faceted ontology for product family modelling[J]. Advanced Engineering Informatics,2011,25(2):147-161.

[20] Flahive A,Taniar D,Rahayu W. Ontology as aservice (OaaS):a case for sub-ontology merging on the cloud[J]. Journal of Supercomputing,2013,65(1):185-216.

[21] Bouma G. Cross-lingual ontology alignment using EuroWordNet and Wikipedia[J]. Lrec,2010.

[22] 鲍秀林. 国家叙词库框架及其语义映射研究[D]. 北京:中国科学技术信息研究所,2015.(Bao Xiulin. Research on the framwork and semantic mapping of the National Thesauri System[D]. Beijing:Institute of Scientific and Technical Information of China,2015.)

[23] 卜书庆,贺玲勇,宗玥,等. 分类主题一体化的知识组织系统研发述要——兼论《中图法》第五版编辑维护系统[J]. 国家图书馆学刊,2011,22(4):28-34.(Bu Shuqing,He Lingyong,Zong Yue,et al. R & D stated on knowledge organization system of classification subject integration——and on the *CLC* 5 editor and maintain system[J]. Journal of the National Library of China,2011,22(4):28-34.)

[24] 张哲. 利用本体和主题词表的集成查询元数据[J]. 情报杂志,2004,23(4):16-18.(Zhang Zhe. Integration ontology and thesauri for RDF schema creation[J]. Journal of Information,2004,23(4):16-18.)

[25] 王军,程煜华. 基于传统知识组织资源的本体自动构建[J]. 情报学报,2009,28(5):651-657.(Wang Jun,Cheng Yuhua. An automatic approach to ontology building by integrating traditional knowledge organization resources[J]. Journal of the China Society for Scientific and Technical Information,2009,28(5):651-657.)

[26] 叶继元. 信息组织[M]. 北京:电子工业出版社,2015:43-44.(Ye Jiyuan. Information organization[M]. Beijing:Publishing House of Electronics Industry,2015:43-44.)

[27] Gulledge T. What is integration?[J]. Industrial Management & Data Systems,2006,106(1):5-20.

[28] Chu W W,Yang H,Chiang K,et al. CoBase:a scalable and extensible cooperative information system[J]. Journal of Intelligent Information Systems,1996,6(2):223-259.

[29] Telang A,Mishra R,Chakravarthy S. Ranking issues for information integration[C]//Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop. Washington D.C.:IEEE Computer Society,2007:257-260.

[30] Jayasimha D N,Iyengar S S,Kashyap R L. Information integration and synchronization in distributed sensor networks[J]. IEEE Transactions on Systems Man & Cybernetics,1991,21(5):1032-1043.

[31] Keet C M. Open world assumption[M]//Encyclopedia of Systems Biology. New York:Springer,2013:1567-1567.

[32] Nikolai R,Traupe A,Kramer R. Thesaurus federations:a framework for the flexible integration of heterogeneous, autonomous thesauri[C]//Proceedings of the Advances in Digital Libraries Conference. Washington D.C.:IEEE Computer Society,1998:46-55.

[33] Gray P M D,Preece A,Fiddian N J,et al. KRAFT:knowledge fusion from distributed databases and knowledge [C]//Proceedings of the 8th International Workshop on Database and Expert Systems Applications. Washington D.C:IEEE Computer Society,1997:682.

[34] Steinberger R,Pouliquen B,Hagman J. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC[C]//International Conference on Computational Linguistics and Intelligent Text Processing. Berlin:Springer-Verlag,2002:415-424.

[35] Aitchison J,Clarke S D. The thesaurus:a historical viewpoint,with a look to the future[J]. Cataloging & Classification Quarterly,2004,37(3-4):5-21.

[36] Hepp M. Products and services ontologies:a methodology for deriving OWL ontologies from industrial categorization

standards[J]. Ssrn Electronic Journal,2006,2(2):72-99.

[37] Beneventano D,Bergamaschi S,Guerra F,et al. The MOMIS approach to information integration[C]//Proceedings of the 3rd International Conference on Enterprise Information Systems (ICEIS 2001). New York:ACM,2001:253-268.

[38] Sagri M T,Tiscornia D. Semantic lexicons for accessing legal information[C]//Electronic Government. Berlin: Springer-Verlag,2004:72-81.

[39] Francesconi E,Peruginelli G. Semantic interoperability among thesauri:a challenge in the multicultural legal domain[C]//Business Information Systems Workshops: BIS' 2010 International Workshops. Berlin: Springer-Verlag,2010:280-291.

[40] Po L,Sorrentino S. Automatic generation of probabilistic relationships for improving schema matching[J]. Information Systems,2011,36(2):192-208.

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

[42] Rodriguez M A,Egenhofer M J. Determining semantic similarity among entity classes from different ontologies[J]. Knowledge & Data Engineering IEEE Transactions on,2003,15(2):442-456.

[43] Batet M,Sánchez D,Valls A. An ontology-based measure to compute semantic similarity in biomedicine[J]. Journal of Biomedical Informatics,2011,44(44):118-25.

[44] Castano S,Ferrara A,Montanelli S. H-match:an algorithm for dynamically matching ontologies in peer-based systems[C]//Proceedings of the First International Conference on Semantic Web and Databases. Aachen:CEUR-WS.org,2003:231-250.

[45] Martin T P,Azvine B. Acquisition of soft taxonomies for intelligent personal hierarchies and the soft semantic Web[J]. BT Technology Journal,2003,21(4):113-122.

[46] Egenhofer M J. Comparing geospatial entity classes:an asymmetric and context-dependent similarity measure[J]. International Journal of Geographical Information Science,2004,18(3):229-256.

[47] Duong T H,Jo G S,Jung J J,et al. Complexity analysis of ontology integration methodologies:a comparative study[J]. Journal of Universal Computerence,2009,15(4):877-897.

[48] Hunter A,Summerton R. Fusion rules for context-dependent aggregation of structured news reports[J]. Journal of Applied Non-Classical Logics,2004,14(3):329-366.

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

[50] 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.

[51] Malheiro B. Beliefs and conflicts in a real world multiagent system[J]. International Computer Science Conventions Academic Press,1998.

[52] Noshairwan M W,Qadir M A,Fahad M. Sufficient knowledge omission error and redundant disjoint relation in ontology[C]//Advances in Soft Computing:Awic'2007. Berlin:Springer-Verlag,2007:260-265.

[53] Grimm S,Wissmann J. Elimination of redundancy in ontologies[C]//The Semantic Web:Research and Applications. Berlin:Springer-Verlag,2011:260-274.

[54] Lim E P,Chiang R H L. The integration of relationship instances from heterogeneous databases[J]. Decision Support Systems,2000,29(2):153-167.

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

[56] Sioutos N,Coronado S D,Haber M W,et al. NCI thesaurus:a semantic model integrating cancer-related clinical and molecular information[J]. Journal of Biomedical Informatics,2007,40(1):30-43.

[57] Caracciolo C,Stellato A,Morshed A,et al. The AGROVOC linked dataset[J]. Semantic Web,2013.

[58] Kramer R,Nikolai R,Habeck C. Thesaurus federations:loosely integrated thesauri for document retrieval in net-

works based on Internet technologies[J]. International Journal on Digital Libraries,1997,1(2):122-131.

[59] Hardiker N R,Rector A L. Structural validation of nursing terminologies[J]. Journal of the American Medical Informatics Association,2001,8(3):212-221.

[60] Yang C Z,Chen I X,Hung C T,et al. Improving hierarchical taxonomy integration with semantic feature expansion on category-specific terms[C]//Information Retrieval Technology. Berlin:Springer-Verlag,2008:225-236.

[61] Zhang D,Lee W S. Web taxonomy integration through co-bootstrapping[C]//Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,2004:410-417.

[62] Madhavan J,Bernstein P A,Rahm E. Generic schema matching with Cupid[C]//Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco C.A.:Morgan Kaufmann Publishers Inc. 2001:49-58.

[63] Colombi S,Chodorowski M J,Teyssier R. The HMatch 2.0 suite for ontology matchmaking[C]//Proceedings of SWAP 2007,the 4th Italian Semantic Web Workshop. Aachen:CEUR-WS.org,2007:348-370.

[64] Yang C C,Lin J,Wei C P. Retaining knowledge for document management:category-tree integration by exploiting category relationships and hierarchical structures[J]. Journal of the Association for Information Science and Technology,2010,61(7):1313-1331.

[65] Sahoo S,Dlhatoo S,Kgupta D,et al. Epilepsy and seizure ontology:towards an epilepsy informatics infrastructure for clinical research and patient care[J]. Journal of the American Medical Informatics Association,2014,21(1):82-89.

[66] Stumme G,Maedche A. FCA-MERGE:bottom-up merging of ontologies[C]//Proceedings of the 17th International Joint Conference on Artificial Intelligence. San Francisco C.A.:Morgan Kaufmann Publishers Inc,2001:225-230.

[67] Raunich S,Rahm E. Target-driven merging of taxonomies with ATOM[J]. Information Systems,2014,42(2):1-14.

[68] Rouaneh-Hacene M,Valtchev P,Nkambou R. Supporting ontology design through large-scale FCA-based ontology restructuring[C]//Proceedings of the 19th International Conference on Conceptual Structures for Discovering Knowledge. Berlin:Springer-Verlag,2011:257-269.

[69] Bellamy L J,Geyer T A W,Wilkinson J. Development of a functional model which integrates human factors,safety management systems and wider organisational issues[J]. Safety Science,2008,46(3):461-492.

[70] Rada R,Martin B K. Augmenting thesauri for information systems[J]. Acm Transactions on Information Systems,1987,5(4):378-392.

[71] Roulin C. Sub-thesauri as part of a metathesaurus[C]//Proceedings of the 5th International Study Conference on Classification Research. Amsterdam:Elsevier,1992:329-336.

[72] Raunich S,Rahm E. ATOM:automatic target-driven ontology merging[C]//Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. Washington D.C.:IEEE Computer Society,2011:1276-1279.

[73] Shah N H,Jonquet C,Chiang A P,et al. Ontology-driven indexing of public datasets for translational bioinformatics[J]. BMC Bioinformatics,2009,10 Suppl 2(2):S1-S1.

[74] Bhatia N,Shah N,Rubin D,et al. Comparing concept recognizers for ontology-based indexing:MGREP vs. MetaMap[C]//AMIA Summit on Translational Bioinformatics.New York:BioMed Central Ltd,2009:1-9(S14).

[75] Fernandez M,D'Aquin M,Motta E. Linking data across universities:an integrated video lectures dataset[C]//Proceedings of the 10th International Conference on The Semantic Web:Volume Part II. Berlin:Springer-Verlag,2011:49-64.

[76] Rhodes B J. Taxonomic knowledge structure discovery from imagery-based data using the neural associative incremental learning (NAIL) algorithm[J]. Information Fusion,2007,8(3):295-315.

严承希　北京大学信息管理系博士研究生。北京 100871。

房小可　北京联合大学应用文理学院讲师。北京 100191。

（收稿日期：2017-03-08；修回日期：2017-03-29）