跨领域多来源主题词表集成与服务研究*

朱礼军[1] 赵新力[1,2] 乔晓东[1] 孙钦山[1]

[1](中国科学技术信息研究所 北京 100038) [2](中国科学技术交流中心 北京 100045)

【摘要】 在调研现有领域主题词表情况的基础上,结合 Web Service 技术和知识组织技术,针对国家科技热点监测的需要,提出了跨领域、多来源主题词表的集成服务框架。集成服务框架的核心是知识组织工具体系,它包括:主题词表转换适配器、主题词表 RDFS 表示、概念融合工具,主题词表应用程序访问接口等,并对语义集成过程及关系构词进行了重点分析。

【关键词】 主题词表集成 知识组织系统 主题词映射 知识服务 【分类号】 TP302.1

Research of Multidisciplinary and Multi – Sources Thesauri Integration and Service Architecture

Zhu Lijun[1] **Zhao Xinli**[1,2] **Qiao Xiaodong**[1] **Sun Qinshan**[1]

[1](*Institute of Scientific & Technology Information of China, Beijing 100038, China*)

[2](*China Science and Technology Exchange Center, Beijing 100045, China*)

【**Abstract**】 On the wide research of existing thesauri, Web service and knowledge organization technologies, according to the need of scientific monitor requirements, a multidisciplinary and multi – sources thesauri integration service framework is proposed. The core of the framework is knowledge organization tools system, which includes: thesauri transformation adaptor, thesauri RDFS formalism, concept merging tool and thesaurus based API. The semantic integration process and relation primitives are discussed in detail.

【**Keywords**】 Thesaurus integration KOS Thesaurus mapping Knowledge service

1 背 景

传统主题词表(Thesauri),又称叙词表,是信息资源管理中重要的检索和导航工具。随着基于内容的信息处理需求增长,主题词表作为一个知识体系已经成为概念之间可视化分析和演变分析的重要支撑工具。我国目前开发的主题词表大致可以分为行业主题词表和综合主题词表。行业主题词表如《林业汉英主题词表》、《海洋科学主题词表》、《大气科学主题词表》等。综合主题词表如《汉语主题词表》、《电子政务综合主题词表》等。

主题词表的小型化专业化发展为领域信息资源的管理带来了方便,同时也带来了一些问题。例如,交叉学科通常是研究活跃的领域,在一篇前沿论文关键词中,往往涉及到跨多个领域的术语。一篇题为《基于本体的鱼病知识获取与诊断推理集成系统研究》的博士论文关键词包括:“领域本体、知识获取、遗传算法、基于案例推理、鱼病诊断系统集成系统”,这些关键词涉及到人工智能、农业、计算机等诸多领域,一部某领域的词表往往很难全面覆盖这些词汇。另外,各个行业所编制的主题词表计算机化表示方式并不统一,为构建信息处理系统带来诸多不便。主题词表编制者(领域专家)根据自身所处的领域编制主题词表,而主题词表用户(信息技术专家)则希望能够通过统一的软件接口去访问不同领域的各种主题词表,用户要求主题词表的计算机化表示形式对用户而言是透明的。

经过了小型化、专业化的发展阶段之后,跨领域、多来源的兼容化、集成化将是主题词表研究和发展的重要方向。有必要研究主题词表的统一计算机化表示形式、规范和技术接口,从而集成其它各种行业性主题词表、兼容现有多种格式的主题词表。通过跨领域多来源的主题词表集成服务体系,使以前开发的各种主题词表能够在信息智能处理过程中充分发挥作用,使网络上的计算机

收稿日期:2006 – 10 – 09

* 本文系国家自然科学基金项目“基于政务本体的信息资源类目自动映射方法研究”(项目编号:70573103)的研究成果之一。

能够通过集成服务体系所构建的术语之间的语义关系来理解信息资源的内容。

2 相关研究

在主题词表集成的理论研究方面, Hafedh Mili (1988) 提出了主题词表融合的基本原则并针对标引和检索两个应用做了融合效果评价[1]。 Marios Sintichakis (1997) 采用集合理论, 对单语种主题词表的融合过程做了形式化描述[2]。 Dachelet(1997) 提出翻译词表(Translated thesauri) 、相关词表(Correlated thesauri) 以及中间词表(Interlingua) 三种类型的主题词表集成[3]。 王军 (2006) 则从标题元数据要素抽取术语来自动构建主题词表的角度讨论了词表的扩充问题[4]。

在主题词表集成架构方面, Ralf Nikolai (1998) 等提出了基于 C/S 结构的联邦式主题词表框架体系(Thesaurus Federations)[5]。 Ralf Nikolai 等人的研究成果沿袭了联邦数据库的思想, 但是词表集成的整体层次划分和词表的服务模式并不清晰。

近年来, 随着 XML、RDF(Resource Description Framework) 、RDFS(RDF Schema) 、OWL(Web Ontology Language) 等数据要素知识描述语言的不断涌现, 主题词表的计算机表示研究和开发也非常活跃, 出现了 LIMBER (2001) 、ILRT(2001) 、 CERES(2000) 、 GEM(2001) 、DRC(2002) 、ETB(2001) 等多种基于 XML 和 RDF 的词表表示语言[6]。 W3C 组织在这些表示语言研究的基础上, 提出了 SKOS (Simple Knowledge Organization System, 2005) 作为主题词表计算机化表示的标准[7], 为主题词表集成在语言表示层次上提供了一个非常好的基础。

比较有影响力的词表集成工程有: 统一医学语言系统(Unified Medical Language System, UMLS) , 由美国国家医学图书馆(National Library Medicine, NLM) 编制, 集成了 70 多部医疗领域的词典。 通用多语言环境主题词表 (General Multilingual Environmental Thesaurus, GEMET) , 由欧洲环境署推动建设, 至今已包括 22 种语言的词汇版本与 4 种语言的定义版本。《汉语主题词表》则是由中国科学技术信息研究所牵头编写的一部大型综合性科技词表, 收词范围包括自然科学、医学、农业、工程技术等各学科领域的主要名词术语, 共收录主题词 81 198 条。

从目前的理论、方法、技术、应用等方面来看, 国内外在这个领域展开了一些基于多词表集成的算法与工具[8]、大规模本体测试环境[9]及基于多词表标引[10]等的研究, 但是尚未提出一个清晰的从语法、语义多个层面完整的跨领域、多来源主题词集成服务框架。

3 集成服务框架

Web Service 是将服务表现成单个实体发布到网上并提供 API 以供其它程序使用的一种分布式计算方式。 Web Service 核心技术包括: Web Service 描述语言 WSDL (Web Service Description Language) , 用于进行服务的统一描述、发现和集成规范; UDDI(Universal Description, Discovery and Integration) , 用于服务的发布和集成; 简单对象访问协议 SOAP(Simple Object Access Protocol) , 用于服务调用。 结合目前知识组织研究和 Web Service 技术的发展情况, 本文提出图 1 所示的主题词词表集成服务体系结构和运作机制。

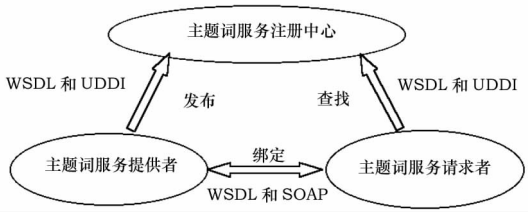

图 1 基于 Web Service 主题词集成服务体系结构与运作

基于 Web Service 的主题词表集成服务体系包含三个角色(主题词服务提供者、主题词服务请求者和服务注册中心) 以及三个操作(发布、查找、绑定) 。 主题词服务可以被其它应用系统通过网络协议来访问。 服务请求方只要遵照 Web Service 接口的定义就可以发送和接收消息。 基于 Web Service 的诸多优点, 本文所提的主题词集成服务给信息处理、信息资源的深度开发利用提供了新的综合集成方案, 降低了信息分析、处理软件和应用系统的设计、开发的复杂程度和成本。

主题词表服务的核心在于知识组织工具体系, 它包括: 主题词表的电子化表示规范、多种格式主题词表向规范化主题词表转换的适配器、跨词表的语义分析工具、规范化主题词表的应用程序访问接口等, 如图 2 所示。

跨领域、多来源的主题词表, 由于其计算机化的表示和存储格式多样, 针对各种文件格式(如 EXCELL、RDB、XML、TXT) 等分别开发特定格式适配器, 将其转化成为统一的表示形式, 然后通过对词表解析, 完成不同词表的词、概念和关系的分析, 并消除词表间的冲突, 形成一致的综合词表, 存入词表库并通过本地 API 或者 WSDL 描述的 Web Service 接口, 对外提供词表服务。

4 基于 RDF Schema 的形式化表示

主题词表虽然早已经有了编制规则标准, 但是却缺乏一个统一的电子表示和交换格式。 随着 Semantic Web

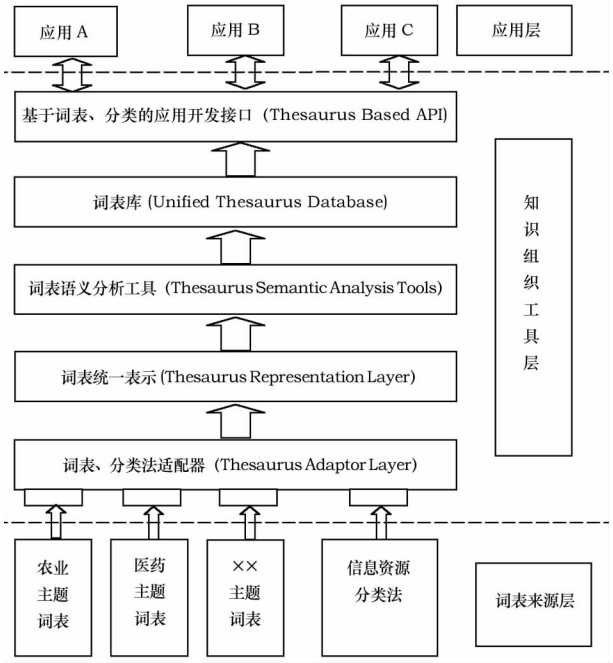

图 2 主题词表服务的核心 - 知识组织工具体系

和知识组织技术的发展, 涌现出了一些可以实现不同系统间知识交换的统一格式, 这无疑为主题词表统一形式化表示、集成和共享交换提供了坚实的技术基础。近年来, 随着语义 Web 技术研究而兴起的主题词表 RDF Schema 表示多种多样, 如 LIMBER 、ILRT、CERES、GEM、DRC、FAO、ETB、SKOS 等。随着 XML、RDF/RDFS 相关的配套解析、推理和存储工具日益完善, RDFS 非常适合作为主题词表集成框架体系中主题词表的统一中间表示格式。采用 RDFS 来表示主题词表的关键在于利用 RDFS 正确实现主题词表中的概念、关系等的构词。RDFS 规范用 RDF 定义了一些建模原语, 其中有关核心类、特性和约束的建模原语如表 1 所示。

表 1 RDFS 中主要的类、特性和约束构词

核心类	rdfs: Resource: 所有用 RDF 表达式所描述的事物都被看成是 rdfs: Resource 的实例。 rdf: Property 用来刻画 rdfs: Resource 实例的所有特性的类。 rdfs: Class 用来定义 RDFS 中的概念(concepts)。
核心特性	rdf: type 关系建立了资源和类之间的 instance - of 关系的模型。 rdfs: subClassOf 关系建立了类之间的包容层次模型。 rdfs: subPropertyOf 关系建立了特性之间的包含层次关系模型。
核心约束	rdfs: ConstraintResource 定义了所有约束的类。 rdfs: ConstraintPropety 是 rdfs: ConstraintResource 和 rdf: Property 的子集, 它包括了所有用来定义约束的特性。 rdfs: range, rdfs: domain 等。

RDFS 机制提供了 RDF 模型中使用的一个基本类型系统[11]。在 RDFS 构词的基础之上, 需要定义更多的其它构词, 来描述主题词表的基本概念和关系, 以及不同主题词表之间概念的映射关系。以 SKOS 为例, 为了描述主题词表和不同主题词表之间的映射关系, 增加了以下构词[12], 如表 2 所示。

表 2 SKOS 中主题词表构词

类构词	概念构词	skos: Concept, skos: TopConcept, skos: CollectableProperty
	概念集合构词	skos: Collection member, skos: CollectableProperty, skos: OrderedCollection, skos: memberList
特性构词	标签属性构词	skos: prefLabel, skos: altLabel, skos: hiddenLabel, skos: prefSymbol, skos: altSymbol, skos: subjectIndicator
	文档属性构词	skos: note, skos: definition, skos: scopeNote, skos: example, skos: historyNote, skos: editorialNote, skos: changeNote
	语义关系构词	skos: semanticRelation, skos: broader, skos: narrower, skos: related
	概念模式构词	skos: ConceptScheme
	主题标引构词	skos: Subject, skos: isSubjectOf, skos: primarySubject, skos: isPrimary, skos: SubjectOf

利用 SKOS 可以对概念及其关系进行描述, 生成基于 SKOS 的主题词表。如下所示:

```
<rdf: RDF
  xmlns: rdf = " http: //www. w3. org/1999/02/22 - rdf - syntax - ns
    #"
  xmlns: skos = " http: //www. w3. org/2004/02/skos/core#" >
    < skos: TopConcept rdf: about = " http: /example. com/Concept/
      0010" >
    < skos: prefLabel > Materials < /skos: prefLabel >
    < skos: inScheme rdf: resource = " http: /example. com/thesau-
      rus"/ >
    < skos: narrower rdf: resource = " http: /example. com/Concept/
      0011"/ >
  </skos: TopConcept >
  < skos: Concept rdf: about = " http: /example. com/Concept/0011"
    >
    < skos: prefLabel > Marble </skos: prefLabel >
    < skos: scopeNote > A granular crystalline limestone </skos:
      scopeNote >
    < skos: inScheme rdf: resource = " http: /example. com/thesau-
      rus"/ >
    < skos: broader rdf: resource = " http: /example. com/Concept/
      0010"/ >
  </skos: Concept >
</rdf: RDF >
```

词表的集成不仅仅是同型概念的合并问题, 更重要的是通过对期刊文献数据的挖掘发现跨领域的主题词表术语之间的关联关系, 在领域专家的参与下, 完成跨领域词表词之间的概念关联。因此, 关联关系的可扩展性是选择合适的词表系统 RDFS 形式化表示的重要因素。

5 语义集成过程

多来源主题词表集成过程中, 除了在词表的计算机化表示形式上达成一致外, 还需要在集成后词表的词汇、结构等层面上进行一致性处理, 即现实语义上的集成。在语义集成过程中, 需要解决以下问题:

(1)同义词和多义词分析

跨领域多来源词表中,不可避免会遇到同义词和多义词问题,这两类问题可以划归为字形层面的融合问题。

(2)概念映射建立

两个主题词表的概念之间,可能存在多种关系,如完全相等、不完全相等(大部分相等、小部分相等)。通过概念映射建立映射文件,该文件是词库结构调整和生成融合词库的描述文件。

(3)概念的合并

概念合并的重点是概念间属分关系的发现问题,属于概念层面的融合问题。属分关系在不同的主题词表实现中,有各种不同的含义。例如,在一些词表中,BT 意味着类包含关系,(is – a 关系),而有些 BT 可能还意味着实例、部分,地理从属等各种关系。在融合的时候,需要有更加一些的关系构词来区别和描述这些关系。

(4)相关关系发现

词表的集成,不仅仅是同型概念的合并问题,更重要的是通过对期刊文献[illegible]的挖掘发现跨领域的主题词表术语之间的关联关系,在领域专家的干预下,完成跨领域词表词之间的概念关联。

(5)融合后词汇表的一致性处理

语义集成后要保证并得到词表内部结构的一致性,检查是否存在违反非自返性等 8 类错误关系检查。

基于语义分析工具所进行的语义集成处理流程如图 3 所示。

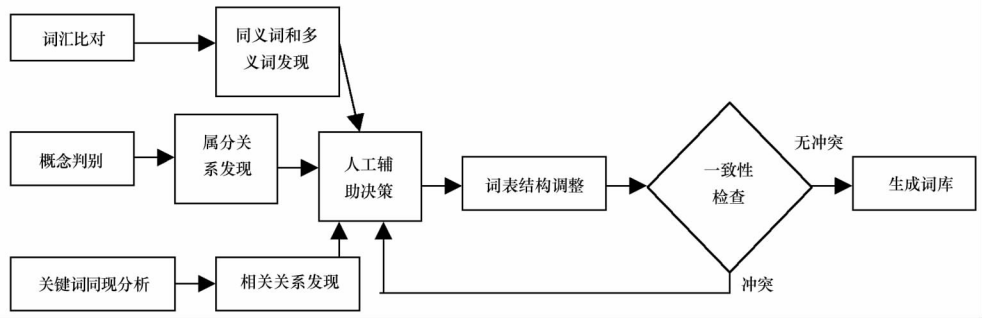

图 3 语义集成处理流程

主题词表集成在词形层面上主要是通过比较两个表中的文本上匹配的概念(KSL, Ontolingua),然后与用户进行交互,确定融合点和融合操作,调整词表结构,确定后,生成融合文件。多词表融合问题的难点和关键在于:属分关系的发现和相关关系的进一步细化。经过总结和归纳,词形和概念层次需要细化和描述的主要关系如表 3。

表 3 术语、概念关系表

术语关系	用代关系	Use UsedFor IsVersionOf HasVersion IsTranslatedFromHasTranslation
概念关系	属分关系	Broader/NarrowerTerm IsA IsPartOf HasPart HasInstantiation IsSpatialPartOf HasSpatialPart IsConceptuallyPartOf HasConceptualPart IsCollectionMemberOf HasCollectionMember
	相关关系	RelatedTerm IsReferencedBy References IsRequiredBy Requires IsBasedOn IsBasisFor IsDerivedFrom HasDerivate HasLinkTo IsLinkedFrom IsMappedTo HasMapping

语义集成过程的目标就是在多个主题词表之间,利用语义分析工具,建立一系列这些关系构词描述的多个表之间的映射文件。将这些映射文件交给应用程序后,对相应的主题词表进行处理,得到融合后的综合表。

6 集成词表性能评价

国内倪[illegible]等研究者在调研国外电子政务主题词表系统时,从词表遵循的标准、类目级别、词条总数、族首词[illegible]式主题词[illegible]词[illegible]率、关联比、参照度、维护部门等多个角度对主题词表系统做了评价[13]。从集成词表应用的角度来考虑,集成后的词表评价方法与词表的应用目标紧密相关,对词表集成前后的性能评价应该从词表的应用目的来分析。如果用于标引,则可以考虑融合前后标引成功率是否有显著提升来衡量,如果用于检索[illegible],则可以使用融合前后词[illegible]计算来评价。本文构建集成词表的应用目的主要有两个:一是供跨学科的科技监测和热点分析使用,二是向其它的信息分析软件和系统提供大规模的概念关系分析服务。因此,本文认为集成主题词表的评价指标应该包括:词表涵盖的学科门类 W(宽度,族首词数[illegible]表知识体系的深度 D(词表的平均层次数[illegible]度均衡性 G(所收主题词在学科上的分布)、关联关系的复杂程度 C、主题词的时[illegible]指是否收录科技文献关键词[illegible]科技文献关键词总数的比值)、主题词的时效指数[illegible]旧词的比率)等,从知识容量、知识结构、知识时效、知识变迁等多方面对集成词表的性能进行评价。

7 结 语

在调研现有领域主题词表情况的基础上,本文提出

了跨领域、多来源主题词表的集成服务框架，并已经开始了概念相似度计算方法研究和概念映射工具原型系统等的开发。后续工作将进一步完善该集成服务框架设计并着手主题词表转换适配[illegible]、概念融合工具的开发、集成以及大规模主题词表存储与访问技术的研究工作。最终，通过该框架，实现多格式、多来源的主题词表快速转换、集成、服务，并根据[illegible]评价指标动态地完善所构建重点领域主题词表、词汇集。

（**致　谢**：非常感谢中国科学技术信息研究所钱启霖、王惠临，北京邮电大学吴斌，北京大学陈文广等多位[illegible]的支持、帮助，最终得以形成此文）。

参考文献：

1 Mili H, Rada R. Merging thesauri: principles and evaluation, Pattern Analysis and Machine Intelligence. IEEE Transactions, 1988, 10(2): 204 – 220

2 Marios Sintichakis, Panos Constantopoulos. A Method for Monolingual Thesauri Merging. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM SIGIR: ACM Press, 1997. 129 – 138

3 Dachelet R. Multilingual querying and multilingual thesauri in Aquarelle. Technical Report, INRIA – Aquarelle, March, 1997

4 Jun Wang. Automatic thesaurus development: Term extraction from title metadata. Journal of the American Society for Information Science and Technology, 2006, 57(7): 907 – 920

5 Ralf Nikolai, Andreas Traupe, Ralf Kramer. Thesaurus Federations: A Framework for the Flexible Integration of Heterogeneous, Autonomous Thesauri. Proceedings of the Advances in Digital Libraries Conference. 1998. 46 – 52

6 Alistair Miles, Brian Matthews. Review of RDF Thesaurus Work. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.2.html (Accessed June. 27, 2006)

7 Alistair J. Miles, Nikki Rogers, Dave Beckett. An RDF Schema for Thesauri (SKOS – Core 1.0 Guide). http://www.w3.org/2001/sw/Europe/reports/thes/8.1/ (Accessed Jun. 27, 2006)

8 Noy N F and Musen M A. PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI – 2000)

9 McGuinness, Deborah L, Richard Fikes, et al. An Environment for Merging and Testing Large Ontologies. In Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning, 2000

10 马然，侯汉清. 基于多词表的自动标引技术研究——对《华社》新闻稿自动标引的实验. 情报学报，2002, 21(3): 273 – 277

11 Dan Brickley, R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. W3c Recommendation. http://www.w3.org/TR/rdf – schema/ (Accessed Jun. 27, 2006)

12 Alistair Miles, Brian Matthews. Inter – Thesaurus Mapping – A guide to the SKOS – Mapping RDF schema for inter – thesaurus mapping. http://www.w3.org/2001/sw/Europe/reports/thes/8.4/ (Accessed Jun. 27, 2006)

13 倪[illegible]，赵[illegible]力、钱[illegible]森垚. 电子政务主题词表编制及网络应用的调查分析. 情报学报，2003, 22(5): 565 – 571

（作者 E – mail: zhulj@istic.ac.cn）

《下　期　要　目》