语义 Mashup 技术研究*

李 峰
(中国科学院国家科学图书馆 北京 100190)
(中国科学院研究生院 北京 100049)

【摘要】介绍语义集成融汇概念,语义网技术在集成融汇中的作用;总结语义集成融汇关键技术,包括语义化数据描述技术、基于语义的协议规范、基于本体的融汇推理技术三个方面;分析国外主要研究项目,包括 KC3 Browser、Bio2RDF、SBWS 和 Semantic REST 等;最后指出如何推动其发展。

【关键词】集成融汇 语义网 RDF 本体
【分类号】G250.7

Study on Technologies of Semantic Mashup

Li Feng
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
(Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【**Abstract**】This paper first introduces the concept of Semantic Mashup and the role of Semantic Web in the Mashup process. Then it summarizes the crucial technologies of Semantic Mashup, such as semantic data describing technology, semantic protocol specifications, Ontology-based Mashup reasoning technology. In addition, the paper analyzes some oversea on-going research projects, specifically for KC3 Browser, Bio2RDF, SBWS and Semantic REST. And finally, it gives some advice on how to promote the development of Semantic Mashup.

【**Keywords**】Mashup Semantic Web RDF Ontology

1 引 言

当前,集成融汇(Mashup)的主要方法包括:数据类型融汇法(Datatype-based)、模式融汇法(Model-based)、大纲融汇法(Schema-based)、实例融汇法(Instance-based)等,这些方法虽然支持异构资源的融汇,但在可用性、灵活性、兼容性、表现性以及支持机器自动融汇方面存在一定的局限性,而语义集成融汇法(Semantics-based)则代表未来的发展方向[1]。语义集成融汇(Semantic Mashup)是将语义网技术与 Mashup 技术相结合,通过对集成融汇的资源对象、服务对象进行语义描述,并建立对象间的语义关系,支持不同来源、不同类型资源间的语义映射、推理,实现复杂语义层面的信息对象的融合、汇集。

语义网技术在 Mashup 中可以起到特殊作用[2]:

(1)通过在网上发布 RDF 格式数据,不但可以克服异构数据集成的障碍,而且可以支持数据集之间的关联整合,支持利用语义网浏览器(比如 Piggy Bank)浏览关联数据项。RDF 链接可以被语义网搜索引擎所采集和标引,例如 Sindice.com 支持采集数据的复杂查询,由于查询结果是结构化的数据,这些结果可以很方便地被其他应用集成融汇。

收稿日期:2009-10-16
收修改稿日期:2009-10-27
*本文系中国科学院 2009 年新增能力项目"知识服务集成融汇关键技术研究"的研究成果之一。

(2)利用语义网的链接、推理技术,可以创建本体和数据实体之间的映射,进而支持不同领域数据的语义融汇。同时,采用语义网技术对资源进行标注,使用基于内容与注释的分离方式对服务描述和服务请求中的领域概念进行扩展,有助于服务发现及工作流重组。

2　语义集成融汇的关键技术

2.1　语义化数据描述技术

(1)基于 RDF、URI 的数据组织

RDF 是用于描述网络资源的标记语言,通过 RDF 用户可以使用自定义的词汇表描述任何资源,由于使用的是结构化的 XML 数据,搜索引擎可以理解元数据的精确含义,使得搜索变得更加智能、准确。同时,网络上任何可用的资源都可以通过通用资源标识符 URI 进行定位。基于 RDF、URI 等技术,在组织、发布 Web 数据的同时可以建立不同数据源之间的数据链接,进而支持对数据、信息、知识进行语义揭示、共享和链接操作。目前,很多组织机构都在积极推动语义化数据组织建设,例如 W3C 设立了 Linked Open Data 项目[3],其目标就是采用 RDF、URI 等技术对开放网络资源进行重新组织,发布语义化、链接化的数据集,支持语义层面的开放调用与集成融汇。该项目已经发布了 100 多个语义化数据集,包括 DBpedia[4],WordNet[5],RDF Book Mashup[6]等。其中,DBpedia 是从 Wikipedia 抽取结构化信息,利用 RDF 进行描述并建立对象间 URI 链接。每个三元组由主、谓、宾三部分组成,每个部分可以是文字或 URI,每个 URI 都意味着一个指向新链接资源的描述。这种描述通常包含新的指向其他 URI 的 RDF 链接。RDF 链接使得利用语义网浏览器可以浏览一个数据源中的数据项与其他数据源中相关的数据。RDF 链接可以被语义网搜索引擎采集到,在这些采集的数据上做更复杂的检索,而且检索结果也是结构化的数据,可以被其他的应用程序所集成。

(2)基于 RDFa 的语义标注

RDFa 是直接在 XHTML 内包含 RDF 数据的一种机制。RDFa 使用 XHTML 元素和属性的集合,使网页能够包含任意数量和任意复杂性的机器可读的语义数据,并显示标准的 XHTML 内容[7]。如例 1 所示,通过在 HTML 标签中添加有关的描述属性,指明该标签的语义。本例中通过在 < a > 标签中增加属性:rel = "license",表明该链接的语义是版权协议;在 < h2 > 标签中增加属性:property = "dc:title",表明该内容为标题;在 < h3 > 标签中增加属性:property = "dc:creator",表明该标签内容是作者。利用这些语义标记,第三方系统就可以快捷、准确地从网页中抽取相关信息项,进而完成与其他资源的语义融汇。

例 1:

```
< a rel = "license" href = http://creativecommons. org/licenses/by/3.0 >
  A Creative Commons License </a >
< div xmlns:dc = http://purl. org/dc/elements/1. 1/ >
  < h2 property = "dc:title" > The trouble with Bob </h2 >
  < h3 property = "dc:creator" > Alice </h3 >
  ……
</div >
```

(3)基于 Microformats 的语义标注

与 RDFa 类似,Microformats 是在 HTML 元素的类、属性中嵌入轻量级的语义标签,实现语义标注目标。Microformats 包括基本微格式(Elemental Microformat)和复合微格式(Compound Microformat)[8]。基本微格式是解决单一问题的最小解决方案,采用 XHTML 支持的 Rel、Rev、Class 等属性定义具有语义的属性集,嵌入到网页文件中直接使用,或者作为复合微格式的基本组成要素。复合微格式由基本微格式和标准的 XHTML 元素组成,解决描述复合数据类型现存标准方案与 XHTML 之间的准确转换问题。基于微格式标注的网页数据不但直接支持语义融汇,通过利用 GRDDL(Gleaning Resource Descriptions from Dialects of Languages)、XSLT 等技术可以将其转换为 RDF 格式数据,支持基于语义网标准的融汇操作。例如英国爱丁堡大学 Harry Halpin 专门研究 Microformats 转换,试图为每种 Microformats 给出一种同类型的 RDF 模型(见表 1)[9]。当前,利用 Microformats 和 GRDDL 开发社交网络的研究成为语义 Mashup 在社交网络中的一种典型应用。

表 1　Microformats 与 RDF 模型的映射

微格式名称	含义	RDF 模型
DC	文档、内容元数据	Dublin Core
XFN	社会网络	FOAF
hCard	人	FOAF
hCalendar	日程表和事件	RDF Calendar
hReview	观点、评定等级、评论	RDF Review
Relicense	版权	Creative Commons
relTag	标签、关键词、分类	Tag Ontology

(4)直接从 HTML 网页中抽取语义内容

直接从 HTML 网页中抽取语义内容是语义融汇技术研究的重要方面。MIT 开发了 Firefox 插件 Piggy Bank[10],目标是支持用户从网页抽取信息实现语义融汇。当网页中不包含 RDF 数据时,Piggy Bank 可以调用 Screen Scrapers 把网页信息重新构造成语义网格式,以 RDF 格式进行保存,供以后检索和与其他人共享。MIT 还开发了 Solvent[11]工具以帮助编写 Scraper。

2.2 支持语义的协议规范

基于 Web Service 的开放 API 是支持集成融汇的关键技术。为实现从 Web Service 到 Semantic Service 的转变,W3C 提出基于 SAWSDL(Semantic Annotations for WSDL and XML Schema)的 WSDL 语义注释方案,支持 SOAP 的语义化信息交换[12];利用 RDFa 和 GRDDL 支持 REST 语义标注,实现基于 REST 的语义融汇[13]。

(1)SAWSDL

SAWSDL 试图通过从 WSDL 和 XSD 元素中引用语义模型的方法来填补 Web 服务和语义网的鸿沟,这些语义模型主要指 RDF 和 OWL,目的是将本体对象的 URI 作为属性嵌入在 WSDL 中。SAWSDL 采用两种标注方法:

① 利用扩展属性 modelReference,指定 WSDL 或 XML Schema 组件与某些语义模型中概念之间的关联,标注 WSDL 接口、操作和故障,以及 XML Schema 类型定义、元素声明和属性声明。

② 利用扩展属性 liftingSchemaMapping 和 loweringSchemaMapping,指定语义数据和 XML 之间的映射,嵌入到 XML Schema 元素声明和类型定义中。

modelReference 用于直接引用语义模型中的概念,如果一个组件或元素不能被直接引用,就可使用 liftingSchemaMapping 和 loweringSchemaMapping 来指明数据映射转换,前者用于 XML 到语义数据的转换,后者用于语义数据到 XML 的转换。2007 年 8 月 28 日,SAWSDL 成为 W3C 的推荐标准。

(2)SA – REST

借鉴 WSDL – S 和 SAWSDL 中的相关思想,SA – REST 通过使用 RDFa 和 GRDDL 添加和提取注释,实现为 REST 服务增加语义信息的目标。与 SAWSDL 不同,SA – REST 注释必须添加到由 HTML 组成的网页中的服务。SA – REST 采用 RDFa 将 RDF 三元组嵌入到 XML、HTML 或者 XHTML 文档中,支持 URI 和命名空间。采用 GRDDL 实现 XML、XHTML 格式到语义网数据的转换,它可以利用类似 XSLT 的程序从 XHTML 和 XML 的内容中提取声明,也可以从 HTML 相关的微格式中抽取 RDF/XML。作为轻量级融汇协议,语义 REST 代表未来发展方向。

2.3 基于本体的融汇推理技术

语义化访问协议、数据描述规范为数据的获取、融合提供了基本支持,但要支持计算机自动化数据分析、过滤、组合等操作,还需要建立相应的语义推理机制。当前研究主要采用本体对数据源进行描述,并利用本体之间的推理机制实现不同来源数据的语义融合。例如,在美国国立卫生研究院和美国国立医学图书馆项目 SMGP(Semantic Mashup of Gene and Pathway)[14]中,研究人员采用本体映射方法实现尼古丁依赖性研究相关资源的语义融汇。尼古丁依赖性研究主要涉及两种基因资源(Entrez Gene 和 Homolo Gene)和三种 Pathway 资源(KEGG、Reactome 和 BioCyc)。如图 1 所示,项目组采用 OWL 为基因资源创建了知识模型 EKoM(Entrez Knowledge Model),然后把该模型与现有的关于 Pathway 资源的 BioPAX 本体集成。

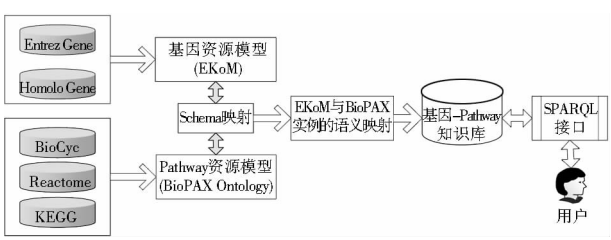

图 1 SMGP 本体映射融汇框架

在 EKoM 和 BioPAX 集成过程中,研究人员发现三个潜在的相似概念:Pathway、Protein、Interaction。通过选择在 EKoM 中重用概念“Pathway”和“Protein”实现两类资源之间的语义关联。经过以上本体集成,系统支持用户通过 RDF 查询语言 SPARQL 对相关资源的语义查询,并利用两个本体之间的映射关系发现相关的基因信息和 Pathway 信息。与 SMGP 类似的项目还有加拿大卡尔顿大学的 yOWL[15]。

3 语义集成融汇主要研究项目

3.1 KC3 Browser

KC3 Browser[16]是由日本的情报通信研究机构(National Institute of Information and Communications

Technology)开发的一个语义融汇浏览器。它根据用户的环境,如兴趣、目的、计算环境等,对无缝的知识存取实现自由链接浏览,并带有各种可视化组件。KC3 Browser 是一个三层架构,最底层是 Web Service 层,它提供各种信息源,支持信息抽取、知识挖掘。中间是 Mashup 层,负责把内容和服务集成起来作为知识集,又称为 K－Workspace,由语义资源、可视化资源、个人和社会资源、内容资源组成。K－Workspace 中所有的内容和图表对象采用基于知识模型元数据标注,内部或外部的语义链接称为 K－Links,由链接生成器产生。最顶层是 Presentation 层,支持显示、发表、共享、重用 K－Workspace 等功能。K－Workspace 包含一个或多个 K－Gadgets,它提供简单的功能单元,通过协同工作支持基于用户上下文信息的语义 Mashup。目前,KC3 Browser 被应用到一个应对东亚和欧盟自然灾害的风险情报项目中,提供预警、时间轴、地图、文档、用户概要、专家信息、三维地图、三维语义空间、风险分析图表等 K－Gadgets。实验表明采用语义融汇和自由链接法超越了网络上时间、空间和领域界限,从而提高了系统的知识分析、共享和重用能力。

3.2　Bio2RDF

Bio2RDF[17] 是由 Genome Canada/Genome Quebec 资助的生物信息学知识融汇项目。

Bio2RDF 利用三元组库 Sesame、本体编辑工具 Protégé、语义浏览器 Piggy Bank、RDF 图观察器 Welkin 和 LSID 浏览器等工具,实现了生物信息学数据的集成融汇。具体过程分三步:

(1) 建立一个关于不同数据提供者的命名空间的列表,使得创建标准 URI 成为可能。

(2) 分析数据源,并用 RDF 模型表示。

(3) 建立一个 RDFizer,把 KEGG、PDB、MGI、HGNC 和 NCBI 等数据源信息转换成 RDF 格式。

经过以上三步产生的 RDF 文档被存储在三元组库中,之后就可以采用相关软件工具对三元组库进行语义分析以及基于 SeRQL 的语义查询。Bio2RDF 之所以选择 RDF,是因为其图形匹配能力可为数据语义融汇提供强有力的支持,当两个 RDF 图之间存在某个 URL 相同时,这两个图即可合并成统一的数据结构。类似的还有 Stephens 主持的一个集成项目[18],同样采用 RDF 技术整合不同生物医学数据源,以支持药物发现。

3.3　SBWS 和 Semantic REST

美国 BBN 公司对语义网与 Web Service 的结合提出了两种解决方案,分别利用 Web Service 语义桥和 Semantic REST 方式实现语义集成融汇[19],目前该项目已经完成了一个结合两种方案的应用实例,实现了对 Web2.0 资源的分布式语义查询。其中,SBWS(Semantic Bridge for Web Service)通过在传统的 SOAP 和 REST 协议上增加语义标签,支持语义层次的服务组合。它封装了一套由 WSDL 或者 WADL 文档描述的 Web Service 操作,并为这些服务创建 SPARQL 查询端,SBWS 通过分析 SPARQL 的 SELECT 或者 CONSTRUCT 决定 Web Service 操作的组合方式。SBWS 利用 WSDL 文档和 OWL－S 文档查询 SOAP 类型的 Web Service,基于 SOAP 服务由 WSDL 文档定义相关操作。对于 REST 服务没有对应的标准,SBWS 选择来自于 Sun Microsystem 的 WADL 作为 REST 服务的详细说明。WADL 被设计成 WSDL 的一种简单选择,用一种简单格式描述网络应用程序,同时定义了输入输出参数的格式,并通过对 WADL 文档自定义注释描述 REST 服务的语义。与 SBWS 不同,Semantic REST 利用 SPARQL 和 RDF 定义约束和模型,从而实现直接从 REST 服务终端查询、检索、修改、删除、增加 RDF 数据的目标。Semantic REST 操作面向两种类型的端点:类层和资源层。类层端点表示在远端服务器上的同一类资源,这些端点提供同类资源的创建和对同类资源的扩展查询。资源层端点表示某一个资源的 URI,这些资源层的端点提供 RDF 格式的资源信息。

4　结　语

尽管目前已经出现了一些语义集成融汇的相关应用,但由于语义网技术的许多方面还不成熟,对于语义网与 Mashup 结合的研究也只是刚刚起步,所以为了推动语义集成融汇的发展也需要一些相应的机制,如鼓励服务提供者提供简单的 REST API,可以在 Web 浏览器中被第三方使用;鼓励服务提供者以 RDF 格式发布数据,或者建立相应的系统把遗留数据以 RDF 格式导入到网络中,并且提供一个 SPARQL 终端,使得语义网的搜索引擎可以检索到相应的数据集;鼓励社区用户使用共享的语义术语标记内容等。

参考文献：

[1] Cheung K H, Yip K Y, Townsend J P, et al. HCLS 2.0/3.0: Health Care and Life Sciences Data Mashup Using Web 2.0/3.0 [J]. *Journal of Biomedical Informatics*, 2008,41(5):694 - 705.

[2] Goble C, Stevens R. State of the Nation in Data Integration for Bioinformatics [J]. *Journal of Biomedical Informatics*, 2008, 41(5):687 - 693.

[3] Linked Data: Principles and State of the Art [R/OL]. [2009 - 05 - 14]. http://www.w3.org/2008/Talks/WWW2008 - W3CTrack - LOD.pdf.

[4] DBpedia[EB/OL]. [2009 - 05 - 30]. http://dbpedia.org/.

[5] WordNet[EB/OL]. [2009 - 05 - 30]. http://wordnet.princeton.edu/.

[6] RDF Book Mashup[EB/OL]. [2009 - 05 - 30]. http://www4.wiwiss.fu - berlin.de/bizer/bookmashup/.

[7] RDFa Primer [EB/OL]. [2009 - 05 - 30]. http://www.w3.org/TR/xhtml - rdfa - primer/.

[8] Microformats[EB/OL]. [2009 - 05 - 30]. http://microformats.org/.

[9] Social Semantic Mashups with GRDDL and Microformats [EB/OL]. [2009 - 05 - 30]. http://2006.xmlconference.org/proceedings/127/slides.html.

[10] Piggy Bank [EB/OL]. [2009 - 05 - 30]. http://simile.mit.edu/wiki/Piggy_Bank.

[11] Solvent [EB/OL]. [2009 - 05 - 30]. http://simile.mit.edu/wiki/Solvent.

[12] Semantic Annotations for WSDL and XML Schema [EB/OL]. [2009 - 05 - 30]. http://www.w3.org/2002/ws/sawsdl/spec/.

[13] Lathem J, Gomadam K, Sheth A P. SA - REST and (S) Mashups: Adding Semantics to RESTful Services[C]. In: *Proceedings of the International Conference on Semantic Computing*. 2007: 469 - 476.

[14] Sahoo S S, Bodenreider O, Rutter J L, et al. An Ontology - driven Semantic Mashup of Gene and Biological Pathway Information: Application to the Domain of Nicotine Dependence[J]. *Journal of Biomedical Informatics*, 2008, 41(5):752 - 765.

[15] Villanueva - Rosales N, Dumontier M. yOWL: An Ontology - driven Knowledge Base for Yeast Biologists [J]. *Journal of Biomedical Informatics*, 2008, 41(5):779 - 789.

[16] Michiaki Iwazume, Ken Kaneiwa, Koji Zettsu, et al. KC3 Browser: Semantic Mash - up and Link - free Browsing[EB/OL]. [2009 - 05 - 30]. http://www2008.org/papers/pdf/p1209 - iwazume.pdf.

[17] Belleau F, Nolin M A, Tourigny N, et al. Bio2RDF: Towards a Mashup to Build Bioinformatics Knowledge Systems[J]. *Journal of Biomedical Informatics*, 2008, 41(5): 706 - 719.

[18] Stephens S, LaVigna D, DiLascio M, et al. Aggregation of Bioinformatics Data Using Semantic Web Technology[J]. *Web Semantics*, 2006, 4(3):216 - 221.

[19] Battle R, Benson E. Bridging the Semantic Web and Web2.0 with Representational State Transfer (REST) [J]. *Journal of Web Semantics*, 2008, 6(1): 61 - 69.

（作者 E - mail:lifeng@mail.las.ac.cn）