

OntoThesaurus Web Service API 及其应用研究^{*}

林伟明 曾新红

深圳大学图书馆 深圳 518060

[摘要] 描述中文叙词表本体共建共享系统 (OTCSS) 所提供的 Web Service API (OntoThesaurus-API)。OntoThesaurus-API 现有 7 类 16 个通用的 API 用来满足 M2M 形式的术语学服务, 并可根据新的需求进行扩展。以 OntoThesaurus-API 在 OPAC 中的应用为例, 详细介绍其使用方法。最后对 OntoThesaurus-API 在其他领域的应用做进一步的展望。

[关键词] 叙词表 本体 中文叙词表本体 OntoThesaurus-API Web Service

[分类号] G254 TP393.09

OntoThesaurus Web Service API and Its Applications

Lin Weiming Zeng Xinhong

Library of Shenzhen University Shenzhen 518060

[Abstract] This paper presents the Web Service API (OntoThesaurus-API) of OTCSS (OntoThesaurus Co-construction and Sharing System). OntoThesaurus-API consists of sixteen common interfaces in seven groups which are used for terminology services from machine to machine. They can be expanded according to new requirements. Taking the application of OntoThesaurus-API to OPAC as an example, the paper also expounds the usage of OntoThesaurus-API. Finally, it makes a prospect of applying the OntoThesaurus-API to other fields.

[Keywords] thesaurus ontology OntoThesaurus OntoThesaurus-API Web Service

1 前言

国家社会科学基金课题 (编号“05CTQ001”) 课题组构建了中文叙词表本体 (OntoThesaurus), 即利用本体技术对中文叙词表进行形式化描述和扩展, 使中文叙词表成为机器真正可理解的本体知识库, 增强了中文叙词表的共享性, 并运用 Jena、SPARQL 等技术, 实现了中文叙词表本体共建共享系统 OTCSS。该系统解决了中文叙词表本体的网络共建共享和动态完善问题, 为 OntoThesaurus 提供了检索、获取、一致性检测、修订意见发送、知识提取、词表管理/发布等功能^[1-4]。为了进一步提供面向应用系统的术语学服务, 我们基于 Apache 的 Axis2 开发了一系列 Web Service API (OntoThesaurus-API)。

2 OntoThesaurus-API

Web Service 是一种可以通过网络进行发布、发

现、调用的新应用。它使用 WSDL 语言来描述应用操作的接口, 通过 XML 消息传递的机制, 让其他应用程序通过网络来访问这些操作。其目的是为了不同的程序之间可以进行交互, 实现应用程序间的通信^[5-6]。因此通过 Web Service 可以有效地解决 M2M (Machine to Machine) 的互操作问题。如图书馆信息管理系统、数据库检索系统、搜索引擎、标签系统等, 不管这些系统基于什么平台开发、采用何种开发语言, 都可以使用 OntoThesaurus-API 提供的服务, 来实现基于中文叙词表本体的智能检索、知识链接和知识服务。

随着 Web Service 的提出与发展, 各大程序语言平台 (Java Net 等) 都对 Web Service 提供了支持, 并涌现了一批开源的 Web Service 开发框架, 如 XFire、Axis2。Axis2 是 Apache 下的第三代 Web 服务引擎, 它比前一个版本 Axis 更高效、更模块化、更面向 XML^[7]。Axis2 既可以作为独立的 Web 服务平台, 也可以内嵌到具体的 Web 应用程序中。因此, 我们选用了 Axis2 作为 OTCSS 系统的 Web Service 开发平台。

在研究了参考文献 [8] 第 4.3.2 节中的术语学服

^{*} 本文系国家社会科学基金项目“基于本体和知识集成实现中文叙词表的升级、共享和动态完善” (项目编号: 05CTQ001) 研究成果之一。

收稿日期: 2009-07-17

修回日期: 2009-09-18

本文起止页码: 119-122, 139

本文责任编辑: 易飞

务描述以及 SKOS API^[9]的服务内容之后,针对中文叙词表本体的术语学服务具体需求,我们开发了以下 7 类共 16 个通用的 Web Service API(另有若干个 API 用于特定的叙词表,并可以根据新的需求进行扩展)。

- 获取叙词。

① public Concept[] getConcepts (String key_word
int matching_mode int search_field)

该接口的作用是通过各种检索关键字检索并获取叙词。其中,参数 key_word 为检索关键字(检索值);参数 matching_mode 为匹配方式(0:精确匹配 1:前方一致 2:任意一致 其中任意一致检索只能用于叙词(入口词)项以及英译名项);参数 search_field 表示检索项(0:叙词(入口词) 1:自定义分类号 2:《中国图书馆分类法》分类号(中图法分类号) 3:英译名)。如果查找结果为空则返回 null 否则返回 Concept(系统中定义的类,其属性有叙词名、叙词类型)数组。

- 获取叙词的数据类型属性(DatatypeProperty)。

② public String getConceptAndCLCCCode (String concept String separator)

通过指定的叙词获取相应的叙词及其《中图法》分类号。其中,参数 concept 为指定的叙词;参数 separator 为指定的分隔符号。该接口返回的结果是以 separator 作为分隔的叙词及其分类号字符串。

③ public String getConceptAndEngCounterpart (String concept String separator)

通过指定的叙词获取相应叙词及其英译名。其中,参数 concept separator 的说明同②。该接口返回的结果是以 separator 作为分隔的叙词及其英译名字符串。

④ public String getEngCounterpartForConcept (String concept String separator)

通过指定的叙词获取相应叙词的英译名。其中,参数 concept separator 的说明同②。该接口返回的结果是指定叙词的以 separator 作为分隔的所有英译名字符串。

- 获取叙词的等同关系词(入口词)。

⑤ public String getConceptAndNTerm (String concept String separator)

通过指定的叙词获取相应叙词及其入口词(同义词)。其中,参数 concept separator 的说明同②。该接口返回的结果是以 separator 作为分隔的叙词及其入口词字符串。

⑥ public String getNTermForConcept (String concept

String separator)

通过指定的叙词获取相应叙词的入口词。其中,参数 concept separator 的说明同②。该接口返回的结果是指定叙词的以 separator 作为分隔的所有入口词字符串。

- 获取叙词的等级关系词(以及等级关系的子关系词)。

⑦ public String getConceptAndBroader (String concept String separator)

通过指定的叙词获取相应叙词及其直接上位词。其中,参数 concept separator 的说明同②。该接口返回的结果是以 separator 作为分隔的叙词及其直接上位词字符串。

⑧ public String getBroaderByLevel (String concept String separator int level)

通过指定的叙词和指定的级数获取相应叙词的指定级数以内的所有上位词。其中,参数 concept 为指定的叙词;参数 separator 为分隔符号;参数 level 为所获取上位词的级数,从 1 开始(直接上位词级数为 1)。该接口返回的结果是指定叙词的指定级数以内的所有上位词,并以 separator 作为分隔的字符串。

⑨ public String getBroaderByPath (String concept String separator String subproperty int level)

通过指定的叙词和指定的上位词子关系、级数获取相应叙词的指定级数以内的指定子关系上位词。其中,参数 concept 为指定的叙词;参数 separator 为分隔符号;参数 subproperty 为指定需要获取的上位词子关系属性(类属/实例/整体_部分);level 为获取子关系上位词的级数,从 1 开始(直接子关系上位词级数为 1)。该接口返回的结果是指定叙词、指定级数以内的所有指定子关系上位词,并以 separator 作为分隔的字符串。

⑩ public String getConceptAndNarrower (String concept String separator)

通过指定的叙词获取相应叙词及其直接下位词。其中,参数 concept separator 的说明同②。该接口返回的结果是以 separator 作为分隔的叙词及其直接下位词字符串。

⑪ public String getNarrowerByLevel (String concept String separator int level)

通过指定的叙词和指定的级数获取相应叙词的指定级数以内的所有下位词。其中,参数 concept 为指定的叙词;参数 separator 为分隔符号;参数 level 为所获

取下位词的级数,从 1 开始(直接下位词级数为 1)。该接口返回的结果是指定叙词的指定级数以内的所有下位词,并以 separator 作为分隔的字符串。

⑫ public String getNarrowerByPath (String concept String separator String subproperty int level)

通过指定的叙词和指定的下位词子关系、级数获取相应叙词的指定级数以内的指定子关系下位词。其中,参数 concept 为指定的叙词;参数 separator 为分隔符号;参数 subproperty 为指定需要获取的下位词子关系属性(类属/实例/整体_部分);level 为获取子关系下位词的级数,从 1 开始(直接子关系下位词级数为 1)。该接口返回的结果是指定叙词、指定级数以内的所有指定子关系下位词,并以 separator 作为分隔的字符串。

- 获取叙词的相关关系词(以及相关关系的子关系词)。

⑬ public String getRelated (String concept String separator)

通过指定的叙词获取相应叙词的相关词。其中,参数 concept separator 的说明同②。该接口返回的结果是指定叙词的以 separator 作为分隔的所有相关词字符串。

⑭ public String getRelatedByPath (String concept String separator string subproperty)

通过指定的叙词和指定的相关关系子关系获取相应叙词的指定子关系相关词。其中,参数 concept 为指定的叙词;参数 separator 为分隔符号;参数 subproperty 为指定需要获取的相关关系的子关系。该接口返回的结果是指定叙词的所有指定子关系相关词,并以 separator 作为分隔的字符串。

- 获取叙词的完整款目信息。

⑮ public ConceptProperty [] getConceptProperties (String concept)

获取指定叙词的所有属性—属性值(即叙词款目信息)。其中,参数 concept 为指定的叙词。该接口返回的结果是:如果叙词不存在或该叙词不存在任何属性—属性值,返回 null 否则返回 ConceptProperty(系统中定义的类,其属性有 Concept 数组、Property、String 数组,其中 Property 表示叙词款目具体属性,如分类号、上位词;而该属性所对应的多个属性值则根据属性值类型为叙词(入口词)或字符串值分别写入 Concept 数组或 String 数组中)数组。

- 获取 OntoThesaurus 中支持的所有属性的信息。

⑯ public Property [] getProperties()

获取 OntoThesaurus 中支持的所有属性名称及其 comment label 等信息。Property 为系统中定义的类,其属性有 name、label、comment 分别用于表示属性名、属性标签、属性注释。

3 OntoThesaurus—API 的使用方法

3.1 OntoThesaurus—API 的应用示例

Web Service 的使用比较简单。目前一般采用以下三种调用方法(程序员无需了解具体的 SOAP 协议):① 使用程序语言平台本身提供的对 Web Service 调用的支持;② 通过其他公司或机构提供的组件进行访问,如微软的 htc 组件、Axis2 开发包等;③ 直接通过 URL 地址访问。

下面我们以图书馆信息管理系统中的公共联机书目查询系统 OPAC 为例,详细介绍 OntoThesaurus—API 的使用方法。其他应用系统可参照此过程开发自己的具体应用。

在 OPAC 中,可以利用 OTCSS 提供的 API 进行扩展性的智能检索应用。比如规范化检索:对于用户输入的检索词,系统运用 OntoThesaurus—API 进行检测,如果检索词为某一个主题词的入口词,系统将会使用该主题词代替此检索词进行检索,否则就使用原检索词进行检索。下面以此为例来说明如何调用 OntoThesaurus—API 来增强 OPAC 的服务。在本例中,采用直接访问 URL 的方式调用 CCT_OTCSS(根据《中国分类主题词表》一版局部建立)的 Web Service 具体的流程如图 1 所示:

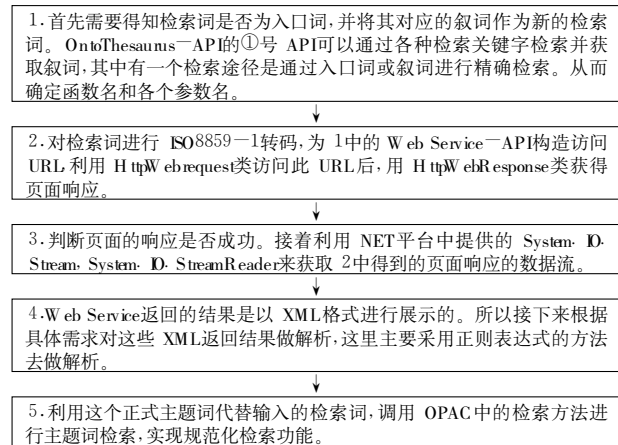


图 1 运用 OntoThesaurus—API 实现规范化检索流程

图 1 中的步骤 2 的 URL 地址为: <http://210.39.15.167:8080/ThesaurusProjectForCCT/services/Thesau>

musService/getConcepts? key_word={0}& matching_mode=0&search_field=0

其中 {0} 处填写检索词。

步骤 4 中所使用的正则表达式为:

<ax21:label>(. * ?)</ax21:label>

在本例中,如果能找到检索词的正式主题词,或者该检索词就是一个正式主题词,还可以运用 OntoThesaurus-API 的 ⑮ 号 API: public ConceptProperty[] getConceptProperties (String concept),把这个主题词的所有款目信息输出到检索结果页面,供用户浏览,用户还可以利用这个主题词的上下位词等款目信息继续进行检索。调用的步骤与上述步骤几乎一致,不同之处在于调用的 URL 地址以及对返回结果的解析过程。

例如,检索“公断”这个词,系统利用 OntoThesaurus-API 找出“公断”的正式主题词为“仲裁”,并使用“仲裁”进行主题词检索,显示检索结果如图 2 所示:



图 2 使用“公断”进行规范化检索后的结果

读者可点击“完整叙词款目”查看“仲裁”的叙词款目信息,如图 3 所示:



图 3 点击图 2 的“完整叙词款目”后显示的叙词款目

3.2 OntoThesaurus-API 在其他方面的应用

利用 OntoThesaurus-API 可以在 OPAC 中实现规范化检索,分类号、英译名扩展检索,上位词、下位词、相关词 (或它们的子关系词) 的扩展检索等,为读者提供更准确、更强大的检索服务。

在图书馆信息管理系统编目子系统中, OntoThesaurus-API 可以为标引员提供内嵌式的辅助标引功能,帮助标引员使用更为准确的主题词进行更快捷的标引。

此外, OntoThesaurus-API 还可以应用在其他领域。如在 CNKI 等数据库检索系统中,在主题词、关键词的检索功能上,可以使用它来辅助完成规范化检索、上下位扩展检索等功能;在搜索引擎中,可以利用此服务获取更多的检索词形态;在众多支持 Tag 的应用程序中,可以使用 OntoThesaurus-API 为用户输入的标签进行规范化的提示,使用户输入的标签更为规范化;在面向主题词的数据挖掘中,可运用 OntoThesaurus-API 基于主题词及其款目信息对数据进行挖掘、统计;在机器学习中,通过 OntoThesaurus-API 可以提供同义词辨析机制和主题词分析机制,进行与主题词相关的信息抽取等。

上述所有的应用都可以通过 Web Service 调用的标准方法,在不同的应用平台上进行实现。

4 结语

基于 Web Service 技术开发的 OntoThesaurus-API 可以非常广泛地应用到其他应用系统中,促进新信息环境下知识服务的发展。本文通过 OPAC 应用实例详细阐述了这些 OntoThesaurus Web Service API 的使用方法。相关领域应用系统的开发者可以参考这些方法,根据具体的 OTCSS 系统提供的 Web Service API 开发自己的具体应用。OntoThesaurus-API 今后还可以针对新的需求,扩展提供其他服务,如支持返回 JSON 数据格式,支持叙词修订信息采集 Web Service 等。

参考文献:

- [1] 曾新红, 明仲, 蒋颖, 等. 中文叙词表本体共建共享系统研究. 情报学报, 2008, 27(3): 386—394.
- [2] 曾新红. 中文叙词表本体——叙词表与本体的融合. 现代图书情报技术, 2009(1): 34—43.
- [3] 曾新红, 林伟明, 明仲. 中文叙词表本体的检索实现及其术语服务研究. 现代图书情报技术, 2008(2): 8—13.
- [4] 曾新红, 林伟明, 明仲. 中文叙词表本体一致性检测机制研究与实现. 现代图书情报技术, 2008(5): 1—9.

(下转第 139 页)

从结论①来看,首先,未参与过网络调查的应答者比参与过网络调查的应答者要重视奖励一些。网络调查者如果想吸引更多的未参与过网络调查的网民参与到网络调查中来,应该要加强对奖励因子的研究,要针对目标调查群体的特点来提供内容合适和价值合理的奖励。网络调查的取样经常会存在偏差,采用适合的奖励方式来吸引未参与过网络调查的应答者参与到网络调查中来,可能会有助于在一定程度上解决这个问题。

其次,调查内容是影响网络用户是否参与网络调查的重要因子,这表明并不是所有的调查都适合采用网络调查的方式。由于网络调查无环境压力,网民在选择是否参与某项网络调查时受外界因素的干扰更小,对没有兴趣和不熟悉的调查选择参与意愿不高。因此,调查的实施者应该根据调查内容来慎重考虑是否要采用网络调查方式,对网络用户不感兴趣和不熟悉的调查项目尽量不要使用这种方式。

再次,适当的调查形式、良好的技术能力和让人放心的隐私保证是网络调查实施者应该重点关注的问题。对于调查形式因子,由于被调查者关注“调查形式”,为提高回复率,网络调查可以尽可能提供更多的形式给被调查者选择;对于技术能力因子,技术能力包含了病毒、问卷太长和网页速度太慢三个变量,如果这三个问题处理不好,对网民来讲这是他们显而易见的损失,特别是在病毒问题上有过深刻教训的网民,往往遵循的原则是:不熟悉的邮件不打开,不信任的网站不上去;对于“个人信息”因子,这是一个普遍存在的问题,这就要求调查方在实施调查的时候重视对个人隐私保密政策的说明,尽可能地降低应答者的顾虑。

从结论②来看,调查者身份因子对网民参与网络调查的意愿没有重要影响。这给我们的重要启示是我们没有必要找一个很大的或者很著名的机构合作。只要问卷本身设计得好,隐私保证能够让网民放心,技术

能力到位,我们的网络调查就一定会有一个好的结果。

参考文献:

- [1] CNNIC. 中国互联网发展状况统计报告. [2009-07-15]. <http://www.cnnic.net.cn/index.htm>.
- [2] Couper M P. Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 2000, 64(4): 464-494.
- [3] 陈钢华, 黄远水. 旅游者重游决策的影响因素实证研究——基于网络调查. *旅游学刊*, 2008, 23(11): 69-74.
- [4] Manfreda K L, Bosnjak M, Berzelak J, et al. Web surveys versus other survey modes. *International Journal of Market Research*, 2008, 50(1): 79-104.
- [5] Tnelli A D. Use of Internet tools for survey research. *Information Technology, Learning and Performance Journal*, 2003, 21(1): 31-37.
- [6] Vehovar V, Batagelj Z, Lozar Manfreda K, et al. Nonresponse in web surveys. In: Groves R M. *Survey Nonresponse*. New York: John Wiley & Sons, 2002: 229-242.
- [7] Bosnjak M M, Tuten T L. Classifying response behaviors in web-based surveys. *Journal of Computer-mediated Communication*, 2001, 6(3). [2009-04-15]. <http://www.ascusc.org/jcmc/vol6/issue3/bosnjak.html>.
- [8] Manfreda K L, Vehovar V. Survey design features influencing response rates in web surveys. *ICIS*, 2002. [2009-05-01]. http://www.icis.dk/ICIS_papers/C2_4_3.pdf.
- [9] Kiesler S, Sproull L S. Response effects in the electronic survey. *Public Opinion Quarterly*, 1986, 50(3): 402-413.
- [10] Bachmann P D, Elfrink J, Vazzana G. E-mail and snailmail face off in rematch. *Marketing Research*, 2000, 11-15.
- [11] 王宝海, 王日旭. 网络调查如何提高回复率. *市场研究*, 2003(6): 49-50.
- [12] 李锐, 宋铁英. 国内网络调查研究分析. *情报科学*, 2005(6): 891-895.
- [13] 方佳明, 邵培基. 影响网络调查适用性的因素分析. *管理学报*, 2006(4): 493-497.
- [14] 方佳明, 邵培基, 粟婕, 等. 基于网络的问卷调查回复率影响因素实证研究. *管理评论*, 2006(10): 12-17.

〔作者简介〕常亚平,男,1963年生,教授,博士生导师,发表论文 80 余篇。

朱东红,女,1982年生,博士,发表论文 10 余篇。

(上接第 122 页)

- [5] 柴晓路,梁宇奇. *Web Services 技术、架构和应用*. 北京:电子工业出版社,2003: 12-18.
- [6] Banerjee A, Corer A. *C# Web 服务高级编程*. 康博,译. 北京:清华大学出版社,2002: 5-8.
- [7] Axis2. [2008-02-07]. <http://ws.apache.org/axis2/index.html>.

〔作者简介〕林伟明,男,1982年生,助理馆员,发表论文 4 篇。

曾新红,女,1968年生,研究馆员,发表论文 20 余篇,出版译著 1 部(合译)。

ml

- [8] Tudhope D, Koch T, Heery R. Terminology services and technology: JISC state of the art review. [2006-10-17]. <http://nkos.slis.kent.edu>.
- [9] SKOS API. [2008-02-07]. <http://www.w3.org/2001/sw/Eu-rope/reports/thes/skosapi.html>.