中国图书馆学报(双月刊)2006 年第 3 期
ZHONGGUO TUSHUGUANXUEBAO

●王曰芬　颜端武　邱均平

基于 Ontology 的内容分析法*

——内容分析系统架构与运行机理研究

摘　要　基于 Ontology 的内容分析系统至少应包括 Ontology 管理模块、内容资源管理模块和内容分析模块。探索了基于 Ontology 的内容分析系统中建立 Ontology,内容样本的处理和存储,内容语义分析的推理机制。图 5。参考文献 6。

关键词　Ontology　内容分析法　内容分析系统　运行机理
分类号　G254

ABSTRACT　The authors think that an ontology-based content analysis method should include an ontology management module, a content resource management module and a content analysis module. They also analyze some mechanisms in an ontology-based content analysis system. 5 figs. 6 refs.

KEY WORDS　Ontology.　Content analysis method.　Content analysis system.　Running system.
CLASS NUMBER　G254

内容分析法在网络环境下要发挥其重要作用,必须借助现代技术和先进工具创新研究思路。本文将在理论研究的基础上,进一步探索基于 Ontology 的内容分析法的分析系统架构与运行机理。

1　基于 Ontology 内容分析系统的功能与层次

1.1　基于 Ontology 内容分析系统功能

(1)能够实现对概念的分析。Ontology 是对概念的一个共享的显性描述,对概念有其明确的定义。内容样本中不管以任何形式出现的概念都能够被识别进而能够对其进行分析。如关于疾病,基于 Ontology 的内容分析法可以将各种疾病(流感、禽流感、非典型性肺炎等)作为疾病的概念来进行分析。

(2)Ontology 同时还描述了概念之间的关系,进行内容分析时,可以对内容样本中出现的各种概念之间的关系进行分析。

(3)Ontology 还描述了概念的各种规则,进行内容分析时,可以根据这些规则对内容样本进行深层次的信息挖掘。

1.2　基于 Ontology 内容分析系统的层次

在实际应用时,Guarino 提出了从详细程度与领域依赖度两个方面对 Ontology 进行划分。其中,依照领域依赖程度,Ontology 被细分为顶层 Ontology、领域 Ontology、任务 Ontology 和应用 Ontology 四类。我们在进行内容分析系统层次设计时选择参照领域 Ontology 的层次结构。

基于 Ontology 的内容分析系统的应用,首先需要有一个内容元数据库。在此基础上还需要有 Ontology 库。这两个库所处的层次可以被认为是数据层,而具体的内容分析应该是在数据层之上的特定的应用。因此,按照领域 Ontology 层次结构划分的模式,我们将内容分析放置在应用层上。应用层特定的内容分析必须通过中间层次完成相应的 RDF/S 的推理和语义分析查询等功能。由此,我们设计了基于 Ontology 的内容分析法的层次结构,如图 1 所示。其中,Ontology 库中存储的是顶层 Ontology、领域 Ontology、任务 Ontology,而内容元数据库存储的则是已经处理好的内容样本数据。

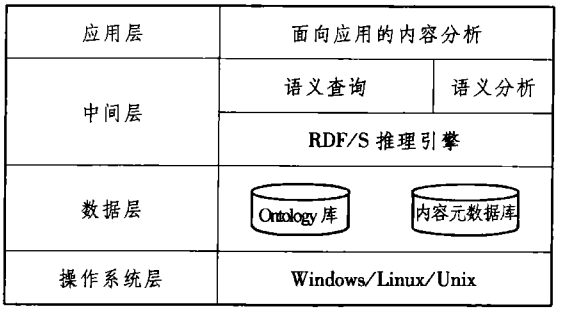

图 1　基于 Ontology 的内容分析系统层次结构

* 本文系教育部人文社会科学重点研究基地重大项目"文献计量与内容分析的综合研究"(项目编号 02JAZD870003)的研究成果之一。

2　基于 Ontology 内容分析系统的总体架构研究

Ontology 是某领域内概念的显示说明,并且描述了概念之间的内在关系。在 Ontology 的基础上进行内容分析,能够在概念或语义层次对内容对象进行相对较为全面分析。另外,Ontology 在领域范围内规定了资源的统一表述,任何自然语言或编码描述的资源都对应于 Ontology 中的某一个类或实例。内容分析在应用时需要对资源概念等进行统计分析。由于在统计某个概念时研究者可能会使用自然语言或者是事先针对内容库设计的编码表中的词条,而用这些自然语言或词条进行统计,查准率和查全率效果不是很理想,尤其是在使用自然语言进行统计某个概念时。而如果使用 Ontology 作为中介,将自然语言或者词条转换为某个规范的概念,再在内容库中进行匹配检索,就能够更加准确和全面地统计分析概念在内容资源中出现的频次等。

由于内容分析法是针对某一系列特定内容集合的,需要对内容集合进行管理并且利用结构化语言对内容资源进行处理。因此,我们认为基于 Ontology 的内容分析系统至少应该包含 3 个模块:Ontology 管理模块、内容资源管理模块、内容分析模块。此外,内容分析与内容管理模块和 Ontology 管理模块之间还必须有一个概念匹配器进行分析查询匹配,也就是语义分析查询模块。总体架构如图 2 所示。

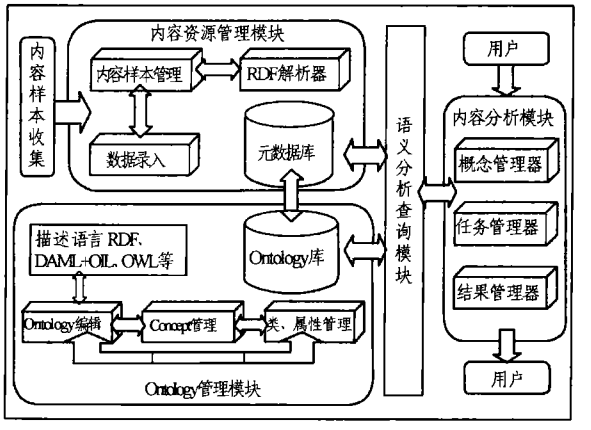

图 2　基于 Ontology 的内容分析系统总体架构

2.1　Ontology 管理模块

Ontology 管理模块的功能是针对内容分析法的应用目的,选择或构建某一领域 Ontology,同时,建立的 Ontology 需要有一定的弹性和延伸性,不仅能够使用已经存在的 Ontology,而且能够在此基础上进行扩展。

Ontology 管理模块中建立的 Ontology 主要分为两个层次:抽象层和实例层。抽象层包含了 Ontology 三层结构中的顶层 Ontology 和领域 Ontology,这个层次的 Ontology 主要是在领域专家的帮助下建立的;应用层 Ontology 则是针对不同任务由进行内容分析的专家学者通过使用 Ontology 编辑器进行设计和建立的,包含 Ontology 层次结构中的任务层 Ontology 和应用层 Ontology。

2.2　内容资源管理模块

该模块主要功能是对内容分析的对象进行采集、审核、标记和存储等。内容资源不仅包括文本内容资源,还包括 Web 内容资源,视频、音频、图像内容资源和各种数据资源。可以从网络上采集,也可以从报刊和电视等媒介中采集。对采集来的内容样本需要进行审核等一系列的操作,最后将内容中的数据,根据 Ontology 存入元数据库。

2.3　内容分析模块

内容分析模块是完成常规的内容分析功能。和其他的计算机辅助内容分析软件相比,不同之处表现在内容分析过程中对概念的查询上。基于 Ontology 的内容分析法在进行查询时主要是针对概念的语义查询,通过语义分析查询模块,并借助 Ontology 而完成的。

内容分析模块由概念管理器、任务管理器、结果分析器构成。概念管理器中管理的是进行内容分析时需要分析统计的概念或概念集;任务管理器是根据某个具体的内容分析任务,进行相应的管理操作;结果分析器对分析的结果进行存储或作进一步分析等。

2.4 语义分析查询模块

这个模块完成的功能就相当于语义检索。主要是在Jena、VicSoft. RDF等推理机制的帮助下,对内容分析所涉及到的概念及术语进行相关的语义分析查询,然后将结果反馈给内容分析模块,完成相关的内容分析任务。

3 基于Ontology内容分析系统的运行机理研究

3.1 利用Ontology编辑器建立Ontology

可以参照Ontology的构建方法学,根据内容分析所需分析的问题及其相关领域,在领域专家的帮助下建立相应的Ontology。而建立相关Ontology时一般都是通过Ontology编辑器进行的。目前流行的Ontology编辑器主要有:OILED,Protégé等等。Protégé是斯坦福大学开发的可视化的Ontology编辑工具,它可以很方便地建立Ontology中的类、属性、属性的约束(domain和range),以及实例(instance)等。另外它又是一个开源项目,许多组织针对Protégé开发了许多插件,扩展了它的功能,其中OWL Plugin就支持W3C组织最新的Ontology描述语言OWL语言。Protégé建立的Ontology以OWL形式存在,并能够被转换为RDF文件。在此基础上可以利用Jena进行相关的推理检索。

3.2 内容样本的处理及存储

除了结构化内容,需要分别处理的内容样本类型还包括:(1)非结构化内容,例如自由文档。这类内容样本,信息量巨大,不可能对一篇文档进行完全分析,因此有必要采用一些方法抽取关键的词汇。通常有3种方法:①对题目进行抽取,因为题目对于内容而言往往能够表达整个内容的中心思想。②对关键词和小标题进行抽取。除了题目,内容的关键词或小标题对于内容也有非常重要的提示。③对内容中出现频率较高的词条进行抽取,因为这类词属于该文档内容的文眼,在一定程度上反映了内容信息。(2)半结构化内容,主要是XML表示的文档。对这类内容的处理比非结构化内容的处理要简单得多,因为我们可以使用相关语言来抽取。

内容样本的存储目前主要采用两种方式:一种是直接在Ontology中以实例的形式存储,另一种是通过相应的关系数据库(如MySQL,ORCLE等)进行存储。利用Ontology直接存储内容源并对其进行推理及内容分析比利用关系数据库进行存储内容源更加简单方便,但是利用关系数据库进行存储内容源的内容数据的存储量更大,并且能够处理更加复杂的推理。

3.3 基于Ontology内容语义分析的推理机制

Ontology是以RDF为基础的,对于Ontology的推理分析应该是对RDF的推理分析。有许多研究小组开发出了针对RDF的API接口。最有名的就是由惠普实验室开发出的Jena。Jena其实是一个JAVA包,在JAVA语言中引用这个Jena后能够实现对RDF以XML形式的读取和书写,也能够对RDF图进行查询,还能对图进行相关的并、交、差等操作。另外也有在.NET框架下的RDF推理工具——VicSoft. RdfParser,它是一个.NET的组件,且能够对RDF数据进行操作,它能够将RDF/XML文件转换成RDF图,还能对RDF中的三元组进行查询操作。

上述两者的推理原理是相同的,都是在将Ontology文件读入内存的基础上,利用自身的接口,对内存中的Ontology进行处理。其一般推理流程如下:

(1)读取Ontology,就是将RDF图、RDF/XML以及三元组形式的Ontology读入内存。其中Jena为读取和建立Ontology提供了一个Model类,该类中有Read方法,而VicSoft. RDF则提供了IrdfParser接口,有Load方法来读取指定位置的Ontology文件。图3和图4是用这两种工具读取Ontology文件的示例。

```
//利用 Jena 包读取 Ontology
import com. hp. hpl. jena. rdf. model. * ;
……
//建立空 Model
ModelIm = ModelFactory. createOntoModel;
//读取存在 C 盘的以 RDF/XML 形式存储的本体
m. read( new FileInputStream( "C:\\OntCA. rdf - xml,
                              "http://nowhere/","");
……
```

图3 利用Jena读取Ontology

```
//利用 VicSoft. RDF 包读取 Ontology
using VicSoft. Rdf;
……
//建立空 Parser
IRdfParser parser = new RdfParser( );
string rdfPath = "C:\\OntCA. rdf - xml;
//建立 URI
Uri rdfUri = new Uri( rdfPath);
//读取存在 C 盘的以 RDF/XML 形式存储的本体
parser. Load( rdfUri);
```

图4 利用VicSoft. RDF读取Ontology

(2)对读取的 Ontology 进行推理,这归根到底就是对 RDF 的推理。RDF 是 Ontology 的核心,Ontology 中的任何事物,包括属性、实例、约束等在 RDF 看来都是一个资源。这个资源有属性,也有值。不同的是,这个属性可能是自己定义的,也可能是 W3C 组织提供的标准,如 rdf:type,rdf:types 等。图 5 显示的是 RDF 的表示形式,在 Ontology 中它表示一个实例,即"神经网络运动控制"是"操纵控制"的一个实例。同样,相应的约束、属性等也能够用 RDF 形式来表示。

```
<rdf:Description rdf:about = "#神经网络运动控制" >
    < rdf:type rdf:resource = "#操纵控制"/ >
</rdf:Description >
```

图 5　Ontology 的 RDF 形式

于是,对 Ontology 的推理最终就是转化为对 RDF 的推理。Jena 和 VicSoft 提供的其实就是对 RDF 的推理接口。Jena 提供了两种推理方法:一种是利用 Model 提供的方法直接进行推理,如 listSubjectsWithProperty,getProperty 等方法;另一种则是利用 RDQL 语言直接对 RDF 进行查询。而 VicSoft. RDF 则提供了针对三元组(S,P,O)的查询操作,如 getSubjects,getObjects 等方法。

4　结　语

在课题研究中,为了考察所设计的基于 Ontology 的内容分析系统的实践应用,我们选择分析样本进行了模拟系统的应用研究。按照总体架构开发了内容分析系统,利用 Protégé 构建了分析样本的 Ontology,并借助 PDF 工具的推理机制,对分析样本进行了关于某类文献、作者等的内容语义分析。

参考文献

1　Guarino N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In: Pazienza M T, eds. Information Extraction: A Muitidisciplinary Approach to an Emerging Information Technology, Springer Vedag, 1997, 139 - 170

2　Ontology 理论研究和应用建模. http://gis. pku. edu. cn/Resources/TR/ontology_Study_ application. doc. 2004 - 08 - 12

3　常春. Ontology 在农业信息管理中的构建与转化:[博士论文]. 北京:中国农业科学院科技文献信息中心,2004

4　Matthew Horridge, Holger Knublauch, Alan Rector, Robert Stevens, Chris Wroe. A Practical Guide To Building OWL Ontologies Using The Protege-OWL Plugin and CO-ODE Tools Edition1. 0. http://www. co-ode. org/resources/tutorials/ProtegeOWLTutorial. pdf

5　Peter Ph. Mohler & Cornelia Zuell, Observe! A Popperian Critique of Automatic Content Analysis. http://www. cavi. univ-paris3. fr/lexicometrica/jadt/jadt2000/pdf/10/10. pdf

6　WP4: Ontology Infrastructure, Knowledge-Assisted Content Analysis, Semantic Reasoning and Intelligent Content Retrieval for the first 18 months. http://www. acemedia. org/aceMedia/ project/work_breakdown/wp4. html

王曰芬　南京理工大学经济管理学院副教授,博士研究生。通信地址:南京。邮编 210094。

颜端武　南京理工大学经济管理学院信息管理系讲师,博士研究生。通信地址同上。

邱均平　武汉大学中国科学评价研究中心教授,博士生导师。通信地址:湖北武汉。邮编 430072。

(来稿时间:2005 - 08 - 29)

(上接第 16 页)

参考文献

1　张其仔. 新经济社会学. 北京:中国社会科学出版社,2001

2,3,5　林南著;张磊译. 社会资本—关于社会结构与行动的理论. 上海:上海世纪出版集团,上海人民出版社,2005

4　罗家德. NQ 风暴——关系管理的智慧. 北京:社会科学文献出版社,2002

6　Donald, W. King, King Research. Survey of Library and Cooperative Library Organizations: 1985 - 1986 . Washington DC: Enterprises for New Directions, 1987.

7　Sheffield University Library. http://www. shef. ac. uk/library/about(2005 - 05 - 27 查询)

高　凡　西南交通大学图书馆副馆长,副研究馆员。通信地址:四川成都。邮编 610031。

徐引篪　中国科学院文献情报中心研究员,博士生导师。通信地址:北京中关村北四环西路 33 号。邮编 100080。

(来稿时间:2006 - 01 - 05)