基于个人信息领域的语义信息抽取系统①

武林仙 (华侨大学 计算机科学与技术系 福建 泉州 362021)

摘 要: 在利用本体进行信息抽取的基础上，提出了一个基于个人信息领域的语义信息抽取系统框架，将语义抽取从 WEB 领域扩展到个人信息领域；系统对个人信息领域内的网页，电子邮件，本地数据库和本地文件夹建立本体，根据本体之间的语义关联，实现个人信息领域内数据的交流。系统详细描述了语义信息抽取系统的实现过程，并以电子邮件为例重点介绍了语义信息抽取的算法。

关键词: 语义信息抽取；个人信息领域；本体；MVC 模式

Semantic Information Extraction System Based on Personal Information Domain

WU Lin-Xian

(Department of Computer Science and Technology, Huaqiao University, Quanzhou, 362021, China)

Abstract: On the basis of semantic information extraction with ontology, this paper proposes a framework of semantic information extraction system based on personal information domain and extends semantic extraction from the WEB area to the personal information domain. Building four ontologies including webpage, Emails, local database and file system, the system implements data communication in personal information domain with semantic association among their ontologies. Then the process of semantic information extraction system is described in detail. With an example of Email domain, relevant semantic information extraction algorithms are designed.

Keywords: semantic information extraction; personal information domain; ontology; MVC Model

随着个人信息的不断增长，每个人要处理的数据越来越多，数据格式越来越多样化，然而它们之间缺乏关联，往往同一主题的信息分散在计算机的各个地方，如何将分散的数据提取出主要信息，再进行语义关联，是我们急需考虑的问题。

在自然语言的处理中，信息抽取[1]是一个输入一个未知文本，产生一个确定格式数据的过程。在计算机领域中，本体是指特定领域概念共享的形式化描述[2]，本体作为各种领域内不同主体之间进行交流(对话、互操作、共享等)的一种语义基础，因此被引用到信息抽取中，称为语义信息抽取(Semantic Information Extr- action)。

1 引言

1.1 文章安排

本文先介绍目前基于语义的信息抽取系统。第 2 节提出一种基于个人信息领域的语义信息抽取系统。第 3 节以 Email 领域为例说明语义抽取的算法。第 4 节演示了系统的运行结果，并给出结论。

1.1.1 基于语义的信息抽取

目前已有的基于语义的信息抽取系统有对油画和艺术领域进行抽取的 Artequakt 系统[3]，使用传统的 IE 工具和技术来实现知识抽取；OFFEE 系统[4](全称叫 Ontology-based Fuzzy Event Extraction)是一个基于 ontology 的汉语新闻摘要的模糊事件抽取代理系统；KEUOA 系统[5]，它是一个通过使用用户自定义的知识抽取模式来从互联网页上抽取知识结构的工具。

本文设计了一个基于个人信息领域的语义信息抽取系统，系统利用本体作为语义抽取的桥梁，建立一个个人信息领域的关联本体，对个人信息常用到的网页，电子邮件，数据库文件和本地文件夹文件的信息

① 基金项目:福建省科技计划(2008I0021)
收稿时间:2009-09-19;收到修改稿时间:2009-11-05

分别进行提取和语义转化，实现它们之间的语义关联。

2　基于个人信息领域的语义信息抽取系统

2.1 框架

图 1 是系统的框架，分别对网页，电子邮件，数据库和本地文件的信息进行包装提取，将提取信息放入设计好的关联本体中，利用本体的关联规则实现信息的语义交流，最终实现语义抽取，以语义查询的结果输出。

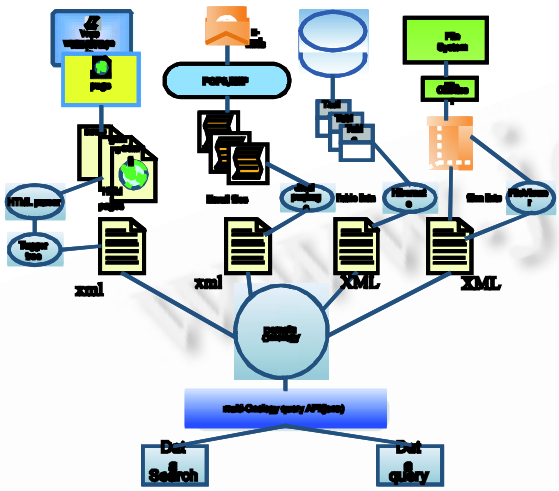


图 1　基于个人信息领域的语义信息抽取框架

2.2 抽取模式

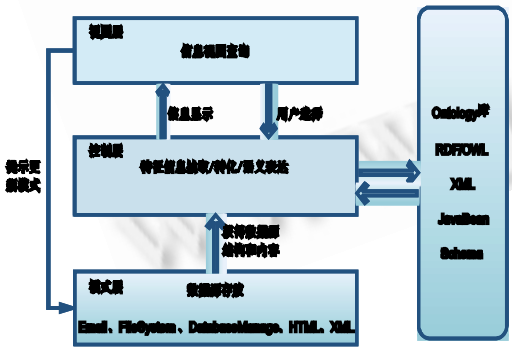


图 2　语义抽取模式

本系统的抽取模式采用的是JAVA的MVC模式实现的语义抽取模式。模式层存放各种数据源的各种描述模式，如 Email 的描述模式为(主题，发送者，接收者，时间，内容)，FileSystem 的描述模式为(文件名，位置，创建时间，修改时间，格式，内容)，Database-Manage 的描述模式为(数据库名，表名，列名，主键)，网页的描述模式为(地址，标题，内容)。控制层里有一个信息包装器和一个语义转化器，控制信息的转化和语义表达，并将结果返回给视图层。视图层里对控制层提供的信息进行语义查询，利用已有的语义查询语言接口，如图 2 所示。

2.3 抽取过程

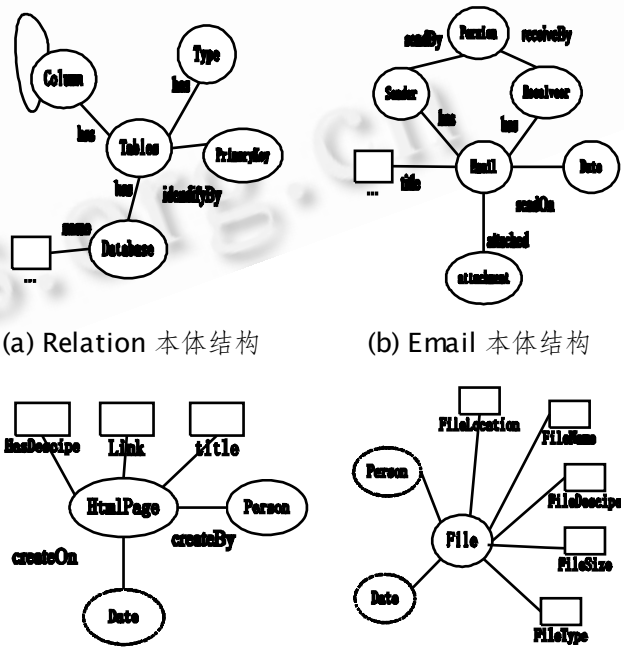
2.3.1 异构数据源的输入

系统集成了四种用户常用的应用程序，方便用户输入数据。电子邮件的客户端，网页浏览器，基于 Hibernate 的数据库管理界面和本地文件系统。当用户在正常使用这些应用程序的情况下，系统对用户正在浏览的数据进行存储并对它们进行语义自动提取。

2.3.2 领域本体的建立

系统分别在 EMAIL 数据领域，WEB 数据领域，关系数据库领域，和文件系统领域建立本体，本体的描述语言采用 OWL；并且描述了这四个领域本体的语义关联，建立了一个顶级本体。

下面是四种常用数据源的本体的结构图，方框代表属性客体，圆框代表类主体，连接表示关系：



(a) Relation 本体结构　(b) Email 本体结构

(c) Websites 本体结构　(d) File 本体结构

图 3　常用四种数据源的本体结构图

2.3.3 语义映射和信息抽取

系统通过设计一个语义包装器和一个语义转化器

来实现语义信息的抽取，采用两种不同的抽取策略实现信息的自动抽取。(1)将用户输入的数据进行包装，转化成 XML 文件保存起来，对于一些在本体没有描述到的新数据，根据 XML 元素创建 RDF 模型，再由 XML 信息生成新的 RDF 有语义的文件，即建立该数据的领域本体。(2)将每一个新数据都看成本体的实例，XML 信息中加入本体 OWL 文件中生成新的本体实例，本体 OWL 文件事先已建好，XML 元素与本体的抽象类进行语义映射，实现数据抽取。这两种方法前者灵活，可以对领域本体进行修改，但需手动修改本体文件。后者虽然只能在原有本体文件中添加实例，不能扩展领域本体，但能实现自动映射。

2.3.4 语义关联输出

系统利用 jena 包集成的 SPARQL 查询语言进行查询输出。SPARQL(Simple Protocol and RDF Query Language)是一面向 RDF 数据模型的查询语言与数据访问协议，是基于以前的 RDF 语言(RDFDB,RDQL,SERQL)发展而来，通过图形模式匹配实现查询功能；个人信息领域内的各个信息关联可以通过本体推理机来实现，如 Pellet，Race 等，都是基于逻辑描述的推理机制，也可手动建立关联，图 4 是以人和时间作为主要关联点，建立的一个语义关联图。

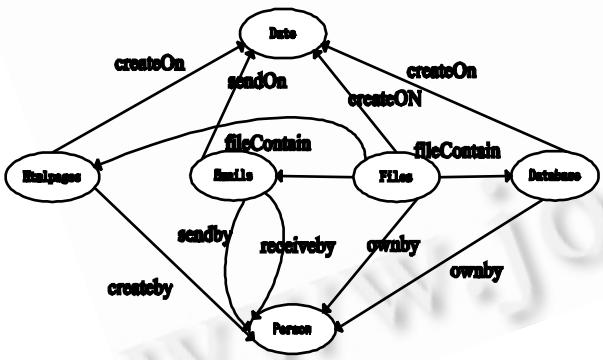


图 4 语义关联图

3 语义抽取算法

下面以电子邮件的语义抽取为例，介绍语义包装器的算法。用户在电子邮件客户端输入登录信息后，系统通过 POP3 协议和 IMAP 协议的下载得到所有 Message 信息，并将 Message 信息的基本属性(如发送者，接收者，主题等)，参考属性项文件(XSD)添加到 JDOM 树的结点上，并保存到 XML 文件中；之后通过 SAX 解析 Email 领域本体的关系结构，以三元组的集合表示，将 XML 所提取的信息以实例添加到领域本体文件中，添加相应本体标签后就能获得带实例的本体文件，实现邮件的语义抽取。

算法如下：

```
EmialToOnto(Message mes_file, String emailOnto)
//包装器部分
    createDomTree();//首先建立邮件结点树
    For allMessageProperty of allmp do
    addContent(mes_file.getProperty()); //将邮件的每个属性信息添加到树结点上
    End For
XMLOutputter(new File(xml_file)); //以 XML 文件输出
//语义转化器部分
onModel=ModelFactory.createOntologyModel(OntModelSpec.OWL); //建立 Email 本体
onModel.read(emailOnto); //加载并读取 Email 本体
Statements stmlAll=listStatements(null, prop,res);
    For all statement stmt of stmlAll do
    ontModel.createIndividual(n,c);//添加新实例
    Property arg1=ontModel.getProperty (pro);
    ontModel.add(arg0, arg1, "this is an email new content");//添加实例新属性
    Do For
ontModel.write(instanceEmailOnto,"RDF/XML-ABBREV"); //输出带实例的 Email 本体
```

4 演示与结论

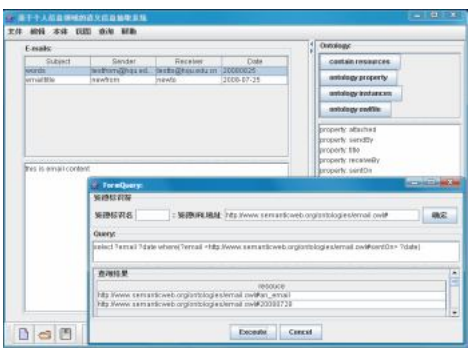


图 5 系统演示结果

(下转第 48 页)

(上接第 11 页)

图 5 是系统的演示结果。文件菜单用来打开本体文件，本体菜单里有四个本体显示窗口(分别对应网页，电子邮件，数据库和本地文件的本体)，视图菜单可以以树的形式显示本体的 OWL 层次，查询菜单实现结构化查询和关键字查询。

目前基于语义的信息抽取很多是对 WEB 网页的信息抽取[6]，本文将信息抽取的领域扩展到个人信息领域，利用本体的语义关联实现了个人信息领域内网页，电子邮件，本地数据库和本地文件夹语义关联，提出了一个基于个人信息领域语义信息抽取系统，并加以实现，不仅扩展了信息抽取的领域，而且实现了各领域之间的数据共享和语义交流。

参考文献

1 Ellen R. Information extraction as a stepping stone toward story understanding. Montreal: MIpress,1999.

2 Fensel D. The semantic web and its languages. IEEE Computer Society 15, 6 (November/December), 67－73.

3 Alani H, Kim S, Millard DE, Weal MJ, Hall W, Lewis PH, Shadbolt NR. (2003) Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, 18 (1):14－21.

4 Lee CS, Jian ZW, Huang LK. A Fuzzy Ontology and Its Application to News Summarization. IEEE, 2005.35,5.

5 Vargas-vera M, Motta E, Domingue J, Shum S B, Lanzoni M. Knowledge Extraction by using an Ontology-based Annotation Tool. K-CAP 2001 workshop on Knowledge Markup and Semantic Annotation.

6 周明建,高济,李飞.基于本体论的 Web 信息抽取. 计算机辅助设计与图形学学报, 2004,16(4):535－541.