基于油田领域本体的信息抽取技术研究

文必龙,李云静
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:文中主要针对当前油田搜索引擎本身不能直接、自动、高效地从油田文本中抽取出精确信息,语义信息不清晰,而且模式不明确的现状进行分析、研究,将信息抽取技术引入到油田信息搜索引擎中,从而构建一种适合于油田领域的信息抽取系统。构建油田领域本体,在GATE框架下,对油田信息进行语法分析并生成相应的抽取规则,最后对文档进行信息抽取,展示抽取结果。该研究会为以后油田自动报表生成、知识推理、自动问答等提供依据,具有很大的实用价值。

关键词:信息抽取;GATE;本体;领域本体
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2015)07-0226-04
doi:10.3969/j.issn.1673-629X.2015.07.051

Research on Information Extraction Technology Based on Domain Ontology in Oil Field

WEN Bi-long, LI Yun-jing
(College of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Mainly aiming at the situation that current oil field search engine itself cannot be direct, automatic, efficient extraction of accurate information from the field of text, the semantic information is not clear, and the mode is uncertainty, analysis and research is made. Information extraction technology is introduced into the field of information search engine, in order to build an information extraction system suitable for oil field. Constructing the oil field ontology, under the framework of GATE, conduct syntax analysis for oil field information and generate extraction rules corresponding. Finally, to extract information from the document, display the results of extraction. For automatic report generation, knowledge reasoning, automatic question answering, provide the basis, this research has great practical value.

Key words: information extraction; GATE; ontology; domain ontology

0 引 言

当今时代是一个信息飞速发展的时代,知识更新越来越快。随着互联网的高速发展,几乎所有的信息都以电子数据的形式存储,这就使人们越来越关注如何从这些海量的信息中找到与自己相关的信息[1]。目前网络上的信息大部分是以HTML来表现,以无规则的形式来显示的。这样不仅语义信息不清晰,而且模式也不明确,导致了很多信息无法直接利用,给人们带来了许多不便。人们希望在较短时间内能够通过搜索引擎获取更为精确的信息,这是普通搜索引擎做不到的。在这样的背景下,信息抽取(Information Extraction,IE)应运而生[2]。

从数字油田到智能油田转变的过程中,油田信息越来越广泛,当前油田搜索引擎本身不能直接、自动、高效地抽取出精确信息,语义信息不清晰。针对这一现状,文中进行分析、研究,利用GATE框架,将本体和信息抽取技术引入到油田信息搜索引擎中,从而构建一种适合于油田领域的信息抽取系统。

文中收集了大量油田领域中的相关信息,构建了油田语料库,构建了油田领域本体,对自然语言处理框架GATE进行深入研究和整理,提出了适合油田领域的信息抽取系统的整体思路,在油田领域操作和使用方面有很好的参考实用价值,也为以后信息抽取的研究奠定了基础。

收稿日期:2014-09-11 **修回日期**:2014-12-16 **网络出版时间**:2015-06-23
基金项目:国家科技重大专项(2011ZX05023-005-012)
作者简介:文必龙(1967-),男,博士,教授,研究方向为软件工程与集成技术;李云静(1986-),女,硕士研究生,研究方向为软件设计开发与集成。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150623.1051.047.html

1　相关概念介绍

1.1　信息抽取

信息抽取是把文本所包含的内容进行结构化处理,形成类似表格的形式[3]。把原始文本输入信息抽取系统,结果会以固定的格式输出。从各式文档集中抽取出所需的信息点后以统一形式集成在一起[4]。

1.2　GATE 有关介绍

目前在自然语言处理领域可供参考的开源资源很少,GATE(General Architecture for Text Engineering)作为一个自然语言处理框架并且是一个开源项目,所以现阶段发展前景很好[5]。它是一个自然语言处理开放型基础架构,在各个领域应用非常广泛,提供易学的开发环境界面,所以现阶段很多信息抽取系统采用此框架[6]。

GATE 共有三个基本的组织模块:GATE 文档管理器(GDM)、GATE 图形用户接口(GGI)和语言工程可重用组件(CREOLE)[7]。这三个组织模块共同管理数据和可重用语言处理组件,达到很好的效果。组织模块如图 1 所示。

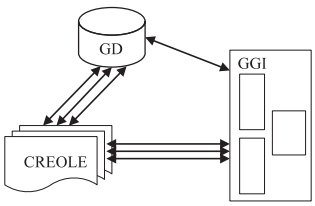

图 1　GATE 的三个基本组织模块

GATE 系统中有三组件,分别为语言组件(Language Resources,LR)、处理组件(Processing Resources,PR)和可视化组件(Visual Resources,VR),这些组件以 Java Beans 形式表现,在信息抽取系统中有很好的扩展性[8]。利用 GATE 作为基础开发框架,开发自己需要的功能组件,在此基础上进行加载,从而满足系统的需求,完善相应功能。

1.3　本体及领域本体的相关概念

本体(ontology),指事物的本身,引申为根本的。本体的概念起源于哲学领域,即“对世界上客观事物所进行的系统描述”,是研究存在的本质的哲学问题[9]。

本体描述了一个或多个领域内概念与概念之间的关系,是规范化和形式化的描述,可以实现共享,为异构系统之间的交流提供统一的语言[10]。本体论在知识获取、自然语言处理、数据库框架集成等研究领域扮演着越来越重要的作用。

不同的领域对本体的研究重点也不同:领域本体是对特定学科的研究,顶级本体对象是具有普遍意义的客观世界常识,顶级本体也称为上层本体或通用本体[11]。

目前研究热点之一就是考虑如何抽取和描述不同领域内的知识并且构建出适合该领域的本体。

文中基于本体的信息抽取系统框架图如图 2 所示。

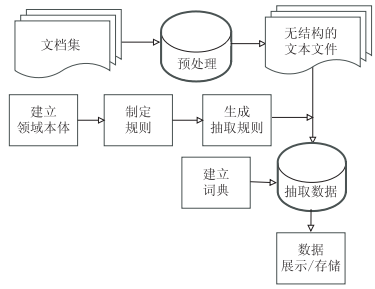

图 2　基于本体的信息抽取系统框架图

2　信息抽取关键技术介绍

2.1　抽取原理

文中通过对大量油田文章的阅读和分析,提出了一种基于油田领域的信息抽取系统。系统主要是将非结构化的信息结构化处理,再按某种语义关系关联起来存放入知识库中,进而对知识库进行自动更新。

2.2　抽取流程

基于本体的中文信息抽取系统包括 4 部分:数据转换模块、中文分词模块、领域本体构建模块以及信息抽取模块。系统流程图如图 3 所示。

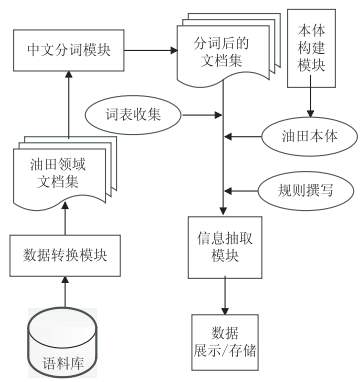

图 3　基于油田领域本体的信息抽取系统示意图

(1)数据转换处理模块。

本系统抽取用到的语料库是从 Internet 上搜集的有关油田方面的各种格式的科研文章、油田网站新闻信息和有关数据库。通过数据转换模块,将所有的信息转换为文本文档的形式,文档内容包含文章标题名称以及相应内容[12]。

(2)中文分词模块。

由于本系统中抽取模块的需要,所有要进行抽取的文本文档必须要先对其进行分词处理操作,分词软件选择中科院的 ICTCLAS,并在此基础上进行整合开发[13],分词结果达到想要的效果。

(3)本体构建模块。

文中本体构建工具选择 Protege4. 3,Protege 软件主要是对本体进行编辑处理,对知识进行获取,亦称为本体开发工具[3]。Protege 在本系统中的功能主要是构建本体和概念之间的关系,它是很重要的开发工具。

(4)信息抽取模块。

· 抽取工具。

文中采用 GATE 作为抽取工具,GATE 是开源软件,方便系统开发使用。

· 词表收集。

由于本系统是针对油田领域进行信息抽取的,所以需要收集大量的油田方面的词表,而 GATE 是国外开发的软件,处理的是英文文档集合,不适用中文,其自带的 Chinese 插件中的词表也仅供测试使用。针对这种状况,需要收集适合的词表和撰写相应的抽取规则。文中根据需要,收集了与油田相关的词表 70 多张,基本包含了油田文档中出现的相关词,例如 qukuai. lst、kaifaqu. lst 等,并写了大量的规则。

· 规则撰写。

在信息抽取系统中,很重要的一部分就是规则的撰写。由于油田领域不同于其他领域,抽取的信息各式各样,非常复杂。因此,编写适合油田领域的规则库就显得尤为重要。文中采用 JAPE(a Java Annotation Patterns Engine)作为建立规则库的语法工具组件,JAPE 是基于正则表达式来匹配和标注信息的。JAPE 文件由 Phase 组成,Phase 又由 rule 组成。rule 包含左右两部分,左侧包含标注模式,右侧包含标注集操作描述,右侧还可以自己编写 Java 代码,来实现相应复杂的功能[14]。

规则模式匹配中常见的三种形式:

①对油田领域文档中可能出现的字符串进行匹配。例如:{Token. string = = "开发区"}、{Token. string = = "油田"}、{Token. string = = "油水井"};

②油田领域的词典和注释器等模块可能出现的词语也是匹配的对象。例如:{Token. majorType = = "dizhiyaosu"};

③从标注的对象的属性或者值进行匹配。例如:{Token. kind = = location}或者{Token. type = = "m"}。

例如识别类似"杏 5-4-平丙 083 井"、"杏 5-3-丙 060 井"的井号规则如下:

```
Phrase:KF_Basic
Input:Token Lookup
Options:control=applet
Rule:Well_Rule
//"+"代表一次或多次
({Lookup. majorType == well}
(({Lookup. majorType == number})+
{Token. string == "-"}
{Lookup. majorType == number}
{Token. string == "-"})
({Lookup. majorType == well})+
({Lookup. majorType == number})+
{Token. string == "井"}
):well
-->
:well. Well={kind="number",rule="Well_Rule"}
```

以上仅是识别井号的一个规则,文中根据要识别的内容并且添加相应标注编写规则,例如:Qukuai. jape 是识别区块的规则,并生成 qukuai 标注;Kaifaqu. jape 是识别开发区的规则,并生成 kaifaqu 标注;ontology. jape 规则是把相应标注和本体绑定,等等。例如识别开发区的规则 Qukuai. jape 如下:

```
Phase: Qukuai
Input: Lookup Token
Options: control = appelt debug = false
Rule: Qukuai
({Lookup. majorType==qukuai}
):tag
-->
{gate. FeatureMap features = Factory. newFeatureMap();
// create an annotation set consisting of all the annotations for tag
gate. AnnotationSet locSet = (gate. AnnotationSet)bindings. get("tag1");
// create an annotation set consisting of the annotation matching Lookup
gate. AnnotationSet loc = (gate. AnnotationSet)locSet. get("Lookup");
// if the annotation type Lookup doesn't exist, do nothing
if(loc ! = null && loc. size()>0)
{// if it does exist, take the first element in the set
gate. Annotation locAnn = (gate. Annotation)loc. iterator(). next();
//propagate minorType feature (and value) from loc
//features. put("locType", locAnn. getFeatures(). get("minorType"));
features. put("majorType", locAnn. getFeatures(). get("majorType"));
features. put("minorType", locAnn. getFeatures(). get("minorType"));
}
// create some new features
features. put("rule", "Qukuai");
// create a GazLoc annotation and add the features we've created
outputAS. add(locSet. firstNode(), locSet. lastNode(), "Qukuai",
features);
```

```
outputAS. removeAll( loc) ;
}
```

标注和本体绑定的规则 ontology. jape 如下:

```
Phase:KF_Basic1
Input:Qukuai
Options:control = appelt
Rule: Location_Rule1
({Qukuai}):mention
-->
:mention{
// create the ontology and class features
FeatureMap feature1 = Factory. newFeatureMap( ) ;
feature1. put( "ontology" , ontology. getURL( ) ) ;
feature1. put( "class" , "区块" ) ;
// create the new annotation
try {
annotations. add( mentionAnnots. firstNode( ). getOffset( ) ,
mentionAnnots. lastNode ( ). getOffset ( ) , " Mention" , features) ;
}
catch( InvalidOffsetException e) {
throw new JapeException( e) ;
}}
```

· GATE 抽取结果。

经过收集词表、构建本体、撰写规则等操作以后,把这些组件组装成类似管道的形式,按顺序加载到 GATE 中,最后就可以利用 GATE 对油田文档集进行信息抽取。

3　实验结果评估

由于系统所需的语料必须是较新的文档,所以以采油厂的最新文本文档为例,抽取井号、开发区、区块、地质要素和断层等内容。从表 1 可以看出,本系统能较为准确地抽取出油田领域中人们关心的信息,实用性比较强。

用准确率和召回率对实验结果进行检测,其中:

准确率 = 系统抽取到的正确数量/系统抽取到的总数

召回率 = 系统抽取到的正确数量/语料库中的总数

表 1　抽取结果

抽取内容	井号	开发区	区块	地质要素	断层
实际数	410	420	361	287	374
抽取数	388	380	330	252	279
正确数	370	316	298	212	192
准确率	0.95	0.83	0.90	0.84	0.69
召回率	0.90	0.75	0.83	0.73	0.51

从实验数据可以看出,基于油田领域本体的信息抽取系统取得了比较满意的准确率和召回率,说明系统性能良好,也证明了系统设计的合理性和可行性。但同时,存在一些因素影响系统的准确率和召回率:没有专业词典支持,ICTCLAS 的分词不能保证 100% 正确,存在错误分词的情况,而且某些分词错误会影响下一处分词。

4　结束语

文中基于油田领域本体的信息抽取系统,利用 GATE 框架对油田信息中的井号、开发区、区块、地质要素、断层等进行识别,取得了较好的效果。然而油田领域信息丰富并且由于中文语言本身的特点,导致中文表达方式多样、句式复杂,今后还有许多后续工作,如对同义词进行识别,复杂关系识别,句式分析等,因此抽取结果有待提高。

参考文献:

[1] 李　飒. 基于 GATE 的中文信息抽取系统的开发和实现[D]. 北京:中国科学院研究生院(文献情报中心),2006.

[2] Soderland S. Learning information extraction rules for semi-structured andfree text[J]. Machine Learning, 1999, 34(1-3):233-272.

[3] 李　毅,保鹏飞,薛万国. 中文电子病历的信息抽取研究[J]. 生物医学工程学杂志,2010,27(4):757-762.

[4] 刘金亮. 汽车行业垂直搜索系统原型的设计与关键模块的实现[D]. 北京:北京邮电大学,2008.

[5] Applet D E, Israel D J. Introduction toinformation extraction technology[C]//Proc of IJCAI-99. [s. l.]:[s. n.],1999.

[6] Kenter T, Maynard D. Using gate as an annotation tool[EB/OL]. 2005. http://gate. ac. uk/.

[7] Tablan V, Maynard D, Bontcheva K, et al. GATE-an application developer's guide[EB/OL]. 2004. http://gate. ac. uk/.

[8] 张　伟. 基于视觉特征的 Web 信息抽取技术的研究与实现[D]. 上海:华东师范大学,2008.

[9] 常平梅. 一种多本体支持的语义标注模型的研究[D]. 大连:大连海事大学,2010.

[10] 奉国和. 基于 GATE 的数字图书馆信息抽取技术概述[J]. 情报杂志,2009,28(5):171-174.

[11] 姜彩红,乔晓东,朱礼军. 基于本体的专利摘要知识抽取[J]. 现代图书情报技术,2009(2):23-28.

[12] 吴　芳,郑　君,刘金亮,等. 基于 GATE 框架的中文信息抽取技术的研究[J]. 电脑知识与技术,2009,5(24):6857-6858.

[13] 杜洪伟. 软件安全领域垂直搜索引擎的优化设计与实现[D]. 天津:天津大学,2010.

[14] 袁萌伽. 异构本体间映射方法研究[D]. 哈尔滨:哈尔滨工程大学,2008.

基于油田领域本体的信息抽取技术研究

作者： 文必龙， 李云静， WEN Bi-long， LI Yun-jing
作者单位： 东北石油大学 计算机与信息技术学院,黑龙江 大庆,163318
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(7)