

文章编号:1673-0062(2012)01-0058-05

管网材料腐蚀信息语义检索系统的设计与实现

欧阳纯萍, 阳小华, 余 颖, 李晓昀, 马淑萍

(南华大学 计算机科学与技术学院, 湖南 衡阳 421001)

摘要:传统的基于关键词的信息检索不能理解用户的需要,仅仅对关键词进行简单的匹配,其结果往往包含大量与用户实际需要毫不相干的信息,同时却丢失用户实际需要的信息,使得检索的效率很低.基于本体的语义检索技术的出现,弥补了基于关键词检索的不足,成为目前构建信息检索系统的应用热点.本文主要针对燃气管网的材料腐蚀信息,设计一个基于 GIS(Geography Information System)的管网材料腐蚀信息语义检索系统,使用户检索管网空间数据和腐蚀数据时为其提供相关数据的语义信息,同时也使得检索结果更加符合用户需求.

关键词:本体;地理信息系统(GIS);语义检索;腐蚀信息

中图分类号:TP31 **文献标识码:**B

Design and Realization of Semantic Retrieval System for the Corrosion Information of Piping Material

OUYANG Chun-ping, YANG Xiao-hua, YU Ying, LI Xiao-yun, MA Shu-ping

(School of Computer Science and Technology, University of South China, Hengyang, Hunan 421001, China)

Abstract: Traditional keywords retrieval didn't understand the requirement of users, which is a process of simple matching. The results of keywords retrieval usually include much needless information, meanwhile, some information the user needs are neglected. To make up for the deficiency of keywords retrieval, semantic retrieval is proposed, which is a hotspot in the application of building information retrieval system. Taking the materials corrosion information of Gas Network as example, this paper introduces a semantic retrieval system based on GIS to search the material corrosion information of pipeline network. The aim of this system is to provide more semantic information about spatial data and corrosion data of pipeline network. Moreover, the results of semantic retrieval could meet the extensible needs of users.

key words: ontology; geography information system; semantic retrieval; corrosion information

收稿日期:2012-01-11

基金项目:湖南省自然科学基金资助项目(10JJ6097);湖南省教育厅科学研究基金资助项目(07C637)

作者简介:欧阳纯萍(1979-),女,湖南衡阳人,南华大学计算机科学与技术学院讲师,博士.主要研究方向:语义数据集成,语义信息检索系统及本体建模.

0 引言

传统的关键词检索存在很多弊端,语义技术在信息中加入表示其隐含意义的内容,为信息系统提供了基于知识和语义的检索途径,为实现智能信息检索提供了坚实基础.解决传统的基于关键词的信息检索问题的关键在于将信息检索从传统的基于语法的简单匹配提升到基于语义的知识层面.目前,GIS 技术已被广泛应用在各类管网信息系统中,而这类系统的数据查询大部分仍是基于关键词的,本文以燃气管网 GIS 系统中查询材料腐蚀信息为例,设计并实现了一个管网材料腐蚀信息语义检索系统.该系统不仅能查询一些通用的管网材料信息,还能通过管网材料的语义查询,获取更多与材料相关的腐蚀关联信息.

1 本体在信息检索中的应用

1.1 本体

本体是概念的明确的规范说明,因此,本体可以看做是一种概念模型 (cerptualization),即是通过将客观世界中一些对象的特性或本质揭示出来而得到的模型,表示为: $C = \langle D, W, Rc \rangle$,其中 C 表示概念化对象, D 表示一个域, W 是该领域中相关事物状态的集合, Rc 是域空间 $\langle D, W \rangle$ 上的概念关系的集合.其表示的含义独立于具体的环境状态.本体的特点有:

1) 明确 (explicit): 所使用的概念及使用这些概念的约束都有明确的定义.

2) 形式化 (formal): 精确的数学描述,使计算机可读.

3) 共享 (share): 本体中表达的知识是其使用者共同认可的,是相关领域中公认的^[1].

1.2 本体在信息检索中的作用

本体在信息检索中的作用主要体现在以下 3 个方面^[2-3]:

1) 建立好的本体确定了标注中所使用的词汇、术语以及描述被标注资源之间相互关系的词汇.

2) 建立好的本体确定了检索所使用的词汇,应该说,所有可供检索的字段都来自这个本体中的词汇.

3) 本体是推理的依据:当需要使用推理工具进行推理时,所有资源之间的关系以及对属性的约束等条件均出自本体.

由本体在信息检索中的作用可以看出,本体的建立是进行基于语义的信息检索的前提条件.

1.3 基于本体的语义检索

由于本体具有良好的概念层次结构和对逻辑推理的支持,因此在信息检索,特别是在基于知识的检索中得到了广泛的应用.

基于本体的语义检索的基本流程可归纳如图 1 所示^[4].

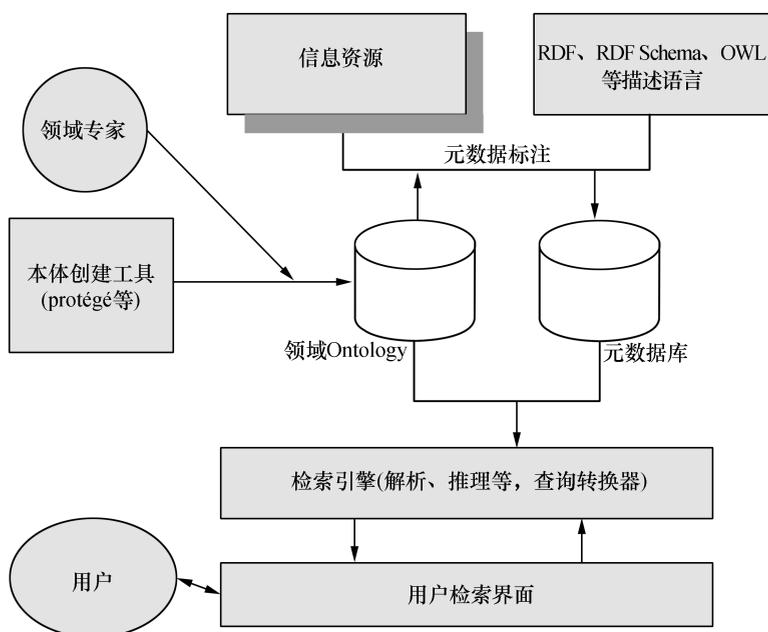


图 1 基于本体的语义检索流程

Fig. 1 Semantic retrieval process based on ontology

1) 在领域专家的帮助下,建立相关领域的本体。

2) 收集信息源中的数据,并参照已建立的本体,把收集来的数据按规定的格式存储在元数据库(关系数据库、知识库等)中。

3) 对用户检索界面获取的查询请求,查询转换器按照本体将查询请求转换成规定的格式,在本体的帮助下从元数据库中匹配出符合条件的数据集。

4) 检索的结果经定制处理后,返回给用户。

2 基于 GIS 的燃气管网腐蚀信息语义检索系统的设计

2.1 系统模型的设计

既然本体是语义检索的前提,所以在设计基于 GIS 的管网材料腐蚀信息语义检索系统模型时,也是以领域本体为基础的,如图 2 所示。

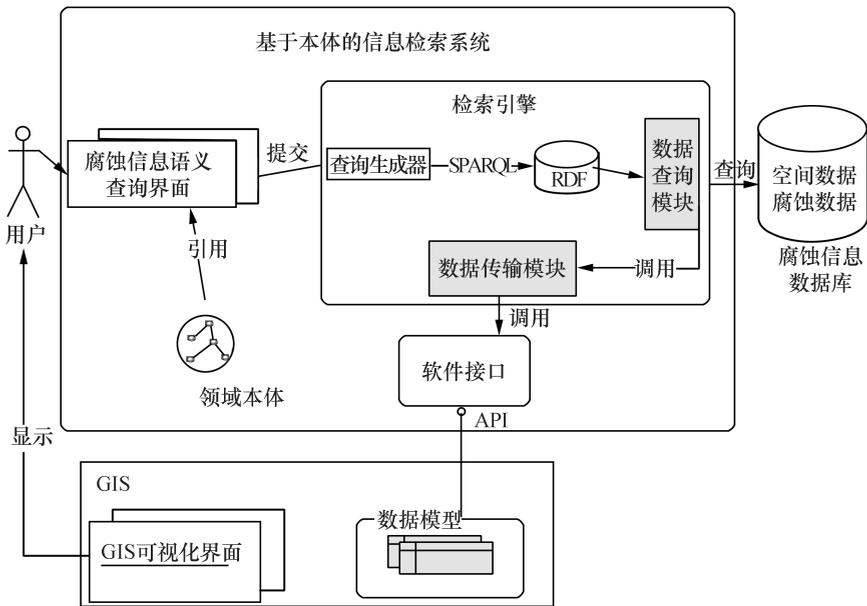


图 2 基于 GIS 的语义检索系统模型设计

Fig. 2 Design of semantic retrieval system model based on GIS

上述设计方案主要包含以下流程:1) 用户在基于本体的语义查询界面输入用户想要查询的内容,然后提交查询;2) 提交查询后,由查询生成器即查询分析器,分析查询内容,构造 SPARQL 查询语句;3) 查询燃气管网土壤腐蚀信息本体 RDF 库,得到结果后,将查询结果与管网腐蚀数据库中的属性数据库进行匹配;4) 查出管线在 GIS 数据库中对应的唯一编号(ObjectID),然后以该编号通过 ArcGIS 数据中间件 ArcSDE 查询管网腐蚀数据库中的空间数据;5) 查询空间数据后,将空间数据与属性数据进行匹配、结合,之后通过 GIS 软件接口,将该数据在 GIS 可视化界面中进行回显。

最终,在可视化界面中可以定位和查询某条燃气管线周围土壤的腐蚀信息,如某条管线某月的各腐蚀因素的数值,该数值可以图表的形式进行展示,同时用户可根据土壤腐蚀标准对照各腐蚀因素的数值,进而判定出腐蚀级别,不同的腐蚀

级别可以不同的颜色在 GIS 可视化地图中进行直观标注。

2.2 语义查询的设计

“燃气管网腐蚀信息本体”是本文所设计的语义检索系统的基础,但是由于篇幅的原因,该本体的详细构建步骤就不在此文中进行介绍。本文着重于探索语义检索与 GIS 技术的结合,将查询过程进行分解,用以展示基于 GIS 的语义检索系统的过程和机理。由于采用的本体描述语言是 RDF,因此相应的我们采用针对 RDF 的语义查询语言 SPARQL。一个 RDF 图是一个三段式的结构,每一个三段式结构包含主语、谓语和宾语。这些结构有不同的来源。SPARQL 查询语言则是基于图模式匹配的。这些简单的图模式即是三段结构模式,与 RDF 三段结构很相似,但在主语、谓语或宾语部分可能用变量代替了 RDF 术语。组合这些三段结构形成一个基本图模式,这里需要对图进行

严格匹配,以形成模式。

根据图 2 中所示的管网土壤腐蚀信息语义检索系统模型的设计方案,可以看到系统使用关系数据库作为后端存储,当用户提交了查询之后,查询分析器生成 SPARQL 查询语句,SPARQL 查询语句查询 RDF 库,然后得出结果,再用该结果与关系数据库进行匹配,然后得到关键的查询结果。所以 SPARQL 语言处理涉及到查询模式到关系数据库的映射问题,当查询少量数据时,以基本三元组模式为粒度完成到关系数据库 RDB 映射的方案查询效率比较高,但是对于大量的数据查询时,由于循环次数的增加,查询效率会大大降低。

为了解决这一问题,另一种办法是将关系数据库(RDB)转化为本体(RDBtransformat to Ontology),这种转化认为只有一个关系数据存在,而本体是由关系数据库构建起来的。因此,这个转化的输入是关系数据库而结果是本体。

根据以上方法将关系数据库转为本体实例数据后,以之前输入的查询条件经查询分析模块分析后形成的 SPARQL 语句来查询该本体实例数据,得到关键的管线号,之后,我们就可以以该管线号来操作 GIS 数据库,并最终将结果显示到 ArcGIS。具体流程如图 3 所示。

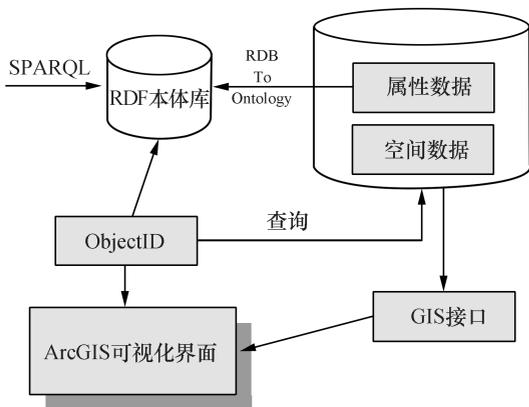


图 3 基于 GIS 和 RDF 的查询流程设计

Fig. 3 Design of query process based on GIS and RDF

2.3 GIS 数据与属性数据的集成设计

燃气管网腐蚀信息空间数据库系统中除了空间数据,还包含大量的腐蚀信息属性数据,这些属性数据有些是空间数据的相关属性,有些则是与空间数据存在某种联系。建立在对空间数据查询和腐蚀信息查询基础上的属性表,与 ArcSDE 都有一列属于相同域的属性字段(之前曾提到的

ObjectID),通过 ObjectID 可以将空间数据与属性数据联系起来,这样就实现了空间数据和属性数据的无缝集成^[5]。

3 基于 GIS 的燃气管网腐蚀信息语义检索系统的实现

根据设计好的模型和方法,我们使用 Eclipse、ArcGIS、Jena 等开发工具完成了基于 GIS 的燃气管网腐蚀信息语义检索原型系统的开发。该系统通过用户对燃气管网腐蚀相关信息的查询,进而实现对现役的或拟建的燃气管道中采集的土壤腐蚀性数据的研究。可以分析,比对总结其特征分布和变化规律,绘制城市地下管网周围土壤环境腐蚀性等级分布和变化趋势图,并根据专家标准对燃气管道所在土壤环境的腐蚀性进行评价。

图 4 显示的是燃气管网腐蚀信息语义检索系统的查询界面,用户可以输入要查询的管线所在地区,比如“海淀区”、“丰台区”等等,或直接输入管线名称,或者输入标志性建筑的名称,比如“中央电视台”。



图 4 系统检索界面

Fig. 4 Retrieval interface of system

查询到的管线信息将在 ArcGIS 的界面上进行回显。由于采用的是语义检索机制,因此在检索到相应的管线后,在 ArcGIS 上该管线的信息显示处还会出现一些与管线腐蚀状况有关的关联信息。如可以获取到某条管线所处土壤环境某年的全部腐蚀因素走势图,如图 5 所示。

同时通过语义扩展查询,还可以得出影响土壤腐蚀因素的评价信息,如图 6 所示。

得到某条管线各土壤腐蚀因素的土壤腐蚀性

评价后,还可将结果在地图上定位并分颜色渲染显示,不同的颜色代表不同的腐蚀等级,比如红色代表某管线所处土壤环境的土壤腐蚀性等级较高。

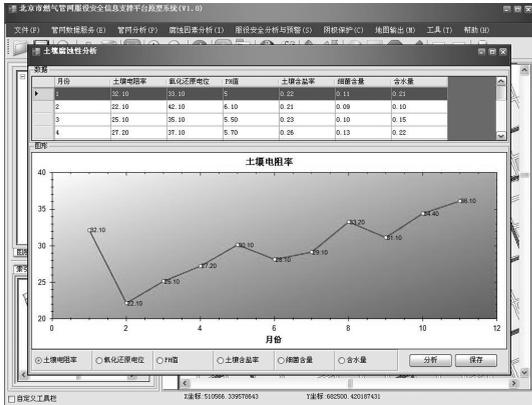


图5 土壤腐蚀因素全年曲线图

Fig.5 The annual curve of soil corrosion factor



图6 某些土壤腐蚀因素在某个月的腐蚀性评价

Fig.6 Pipeline corrosion evaluation based on some soil corrosion factor in a certain month

4 结论

本文首先分析了传统信息检索系统的不足,提出了一种基于GIS的燃气管网腐蚀信息语义检索模型,接着研究了语义Web环境下实现语义检

索的机制,及相关支撑技术,并讨论了本体在语义检索中的应用以及基于本体的语义检索的基本思想.然后着重研究了GIS和语义检索技术相结合在燃气管网腐蚀信息系统开发中的应用.并对该系统进行了设计与实现,本文取得的主要研究成果如下:

1)首先在理解了需求和理清了支撑语义检索的本体所包含的类、属性等等之间的关系基础上,构建了燃气管网腐蚀信息本体,并以此设计了燃气管网腐蚀信息语义系统模型。

2)对系统的语义搜索引擎进行了设计,重点研究了关系数据库与本体之间的映射关系,对相关规则加以研究,实现了关系数据库(RDB)到本体(Ontology)的转化。

3)对燃气管网的数据结构和数据类型进行分析,利用ArcSDE数据引擎技术把燃气管网的数据格式、接口和协议映射为统一的模型和格式,实现了空间数据和属性数据的无缝集成。

本文对语义技术与GIS技术的结合做了初步的探索,但由于个人能力的限制,在管网腐蚀信息本体的建立和查询引擎的设计上还有很多因素未加以考虑,个别问题还需要具体深入地再加以研究,系统模型的设计还可以更加完善.这些都将在下一步的工作给予完善和补充。

参考文献:

- [1] 杜小勇,李曼,王珊. 本体学习研究综述[J]. 软件学报,2006,17(9):1837-1847.
- [2] 李勇,张志刚. 基于本体语义检索技术研究[J]. 计算机工程与科学,2008,30(4):17-19.
- [3] Ouyang Chungping, Hu Changjun, Liu Zhenyu. Design and realization of materials service safety assessment system based on ontology[J]. Advance Material Research,2011, 186:94-99.
- [4] 余传明. 语义检索的原理及其实现[J]. 情报理论与实践,2007,30(2):182-184.
- [5] 王育红. 空间数据集成及冲突消解方法综述[J]. 测绘科学,2011,36(2):81-83.