文章编号:1001 - 9081(2010)02 - 0532 - 05

海相油气地质的概念本体知识系统设计与实现

李国和[1],杨新颖[1],叶 婷[1],孙红军[2],唐先明[2],韩宝东[2]
(1. 中国石油大学(北京) 计算机科学与技术系,北京 102249;
2. 中国石油化工股份有限公司 石油勘探开发研究院,北京 100083)
(guohe_li@sina.com)

摘 要:为了有效利用和普及海相油气地质知识,采用本体的知识表示,通过各类文档转换为标准 TXT 文档后,以专业词条和关联关系词条为基础,实现文档句子的词条分割和概念本体的提取。基于 B/S 结构,实现了海相油气地质知识系统,完成了海相油气地质知识从各类文档中进行知识的自动获取。通过该系统的词条维护功能,确保提高知识获取的精度。最后在海相油气地质知识系统上实现知识检索和共享。实践证明,该知识系统可靠性强,性能良好,已达到实用水平。

关键词:本体;知识表示;知识获取;知识系统;海相油气
中图分类号: TP182; TP311.12; TP399 **文献标志码**:A

Design and implementation of ontology-based knowledge base system for marine hydrocarbon geology

LI Guo-he[1], YANG Xin-ying[1], YE Ting[1], SUN Hong-jun[2], TANG Xian-ming[2], HAN Bao-dong[2]
(1. *Department of Computer Science and Technology, China University of Petroleum, Beijing* 102249, *China*;
2. *Research Institute of Petroleum Exploration and Development, China Petroleum and Chemical Corporation, Beijing* 100083, *China*)

Abstract: In order to efficiently apply and effectively spread the knowledge of marine hydrocarbon, in terms of the ontology for knowledge representation, a variety of documents were translated into standard TXT documents in which all the clauses were segmented into a set of words by marine-facies and relevant ones to extract ontology concepts from marine-facies geology. On the basis of B/S infrastructure, the knowledge system of marine hydrocarbon was successfully implemented, on which the marine-facies knowledge could be automatically acquired from a variety of documents. The maintainable function of the knowledge system for majority words and relevant words guarantees the accuracy of knowledge acquisition. The knowledge of marine hydrocarbon was retrieved quickly and shared widely. The knowledge system is proved to be reliable, efficient and practical.

Key words: ontology; knowledge representation; knowledge acquisition; knowledge system; marine hydrocarbon

0 引言

在智能科学领域,采用语义网络构建领域知识库已得到快速发展,但由于语义关系的复杂性,尤其由领域专家按所需知识类型构建的语义网在结构、涵盖的知识点以及对问题求解的适应性限制了知识获取和利用[1]。本体论的出现进一步完善了知识网络化表示[2-3],它通过对自然语言概念体系的总体表述,建立概念联想脉络,并由概念和语义关系组成的有向无环图来表达知识和描述语义关系,也就是把一切概念转变为“是”和“所是”,使得“是”成为统摄、包容一切“所是”的最普遍概念,由关联关系词和概念组成语义体[4]。概念本体知识可表示为由节点和弧线或链式线组成,并带有标记符的有向无环图,其节点表示实体、概念和情况等,弧线表示节点间的关联关系。通过概念本体提取、关系分析,把知识体系中的名词术语抽象为一组概念与概念间关系[5]。基于概念本体理念,揭示领域有关的语义实体以及语义实体之间的数量、时间、因果、方式、状态等的语义关系构成的概念知识系统。

在中国海相油气知识体系中[6],概念之间关联关系复杂,主要包括分类关系、所属关系、因果关系等。首次采用本体论的思想理念,表征海相油气概念的本质内涵,建立海相油气概念本体及其关联关系,表示该领域知识的基本理论体系。

为了从海相油气地质文本文档中自动高效提取概念本体知识,本文在分析已有知识网络构建技术的基础上,给出了一种基于语义理解的汉语文献概念本体知识库构建方法。该方法以中文文本为处理对象,实现海相油气地质知识的自动获取,并建立中国海相油气领域的概念本体知识库及其海相油气地质知识的关联查询,达到海相油气地质知识的高效利用和推广普及[7]。

收稿日期:2009 - 08 - 28;**修回日期**:2009 - 10 - 20。 **基金项目**:国家自然科学基金资助项目(60473125);国家重大专项子课题(G5800-08-ZS-WX);中国石油(CNPC)石油科技中青年创新基金资助项目(05E7013)。

作者简介:李国和(1965 -),男,福建平和人,教授,博士生导师,博士,主要研究方向:人工智能、知识发现; 杨新颖(1984 -),男,甘肃靖远人,硕士研究生,主要研究方向:数据挖掘; 叶婷(1984 -),女,江西南昌人,硕士研究生,主要研究方向:数据挖掘; 孙红军(1968 -),男,辽宁盘锦人,高级工程师,博士,主要研究方向:石油地质; 唐先明(1970 -),男,四川平昌人,高级工程师,博士,主要研究方向:遥感地理信息系统; 韩宝东(1979 -),男,山东宁津人,工程师,主要研究方向:石油地质。

1　概念本体知识系统设计

围绕海相油气地质领域知识获取与共享(如图 1 所示),利用三元组表示法构造海相油气概念本体知识。每个陈述都分解为主语(Subject)、谓语(Predicate)和宾语(Object),并由起始节点、连接弧和终止节点三部分图示化[8]。主语为资源或资源的部分;谓语为资源的属性,包括资源的特点、性质与其他资源之间的关系等;宾语为属性的值,可以是资源、资源的部分、定语、状语或者补语(如图 2)。

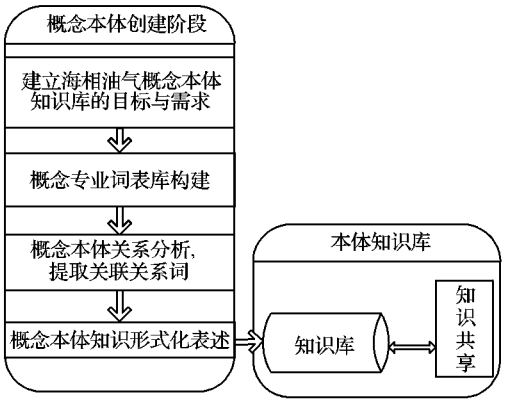

图 1　概念本体知识获取与共享

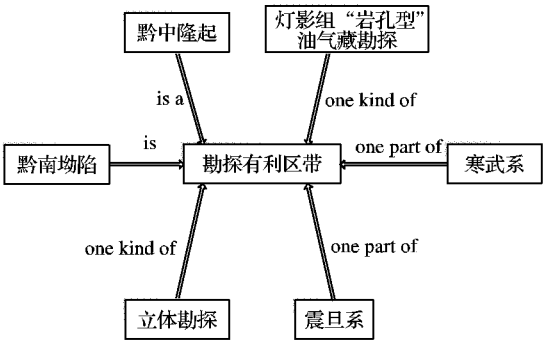

图 2　概念本体结构

概念之间的关联关系包括[9]:1)属性关系 is(attribute-of),由属性和属性值表示概念节点;2)类属关系 is a(instance-of),表示拓展节点是核心节点的一个子类;3)部分和整体的关系 is part of,把整体概念划分为多个部分的组成形式;4)继承关系 a kind of,具体类和抽象类之间的关系,具体类由一般类拓展、派生而来。

海相油气地质的概念本体知识库构建流程如图 3 所示。

1.1　文本分词词条库构建

文本分词词条库包括专业词库和属性词库,其中词条均由石油地质专家提供。专业词条涵盖了石油地质、勘探、开发等方面的专业词汇,例如沉积相、海相、河流相、大陆相等[10];属性词包含汉语言表达中修饰性质类的词条,例如主题、浅层、立体勘探等。专业词条和属性词条具有全面、专业、标准特点,确保了概念本体知识库的构建效率和正确程度。

1.2　关联关系词库构建

海相油气地质概念本体关联关系词的提取转化为单句谓语的提取。在汉语句子中,表示关联关系的词多数是动词,而文学作品或口语中常见的体词[11](指主要语法功能是充当主语和宾语的一些词类)和状态词在海相油气专业领域文档中并不多见,因此仅采用动词作关联关系词,如是、呈现、属于、占据、钻遇、组成、构成、包括等。

1.3　文本预处理

海相油气地质领域主要有 DOC、PDF、PPT、TXT、HTML 等类型文档。以 TXT 为标准文档,在进行文档的中文词条的分割、概念本体知识自动获取、基于内容的查询等操作前,将 DOC、PPT、PDF、HTML 类型转换为 TXT 类型的临时文档。该临时文档原文档建立了一一对应关系,但不长久保留。各种文档信息处理只在 TXT 标准临时文档中进行,这样简化了多种类型的文档处理,提高了概念本体知识系统的研发效率和系统的可靠性。在本项目中使用了 Lucene 文本类型转换工具包把这些不同类型的文档转换成 TXT 纯文本格式的字符串,然后统一进行分词处理[12]。

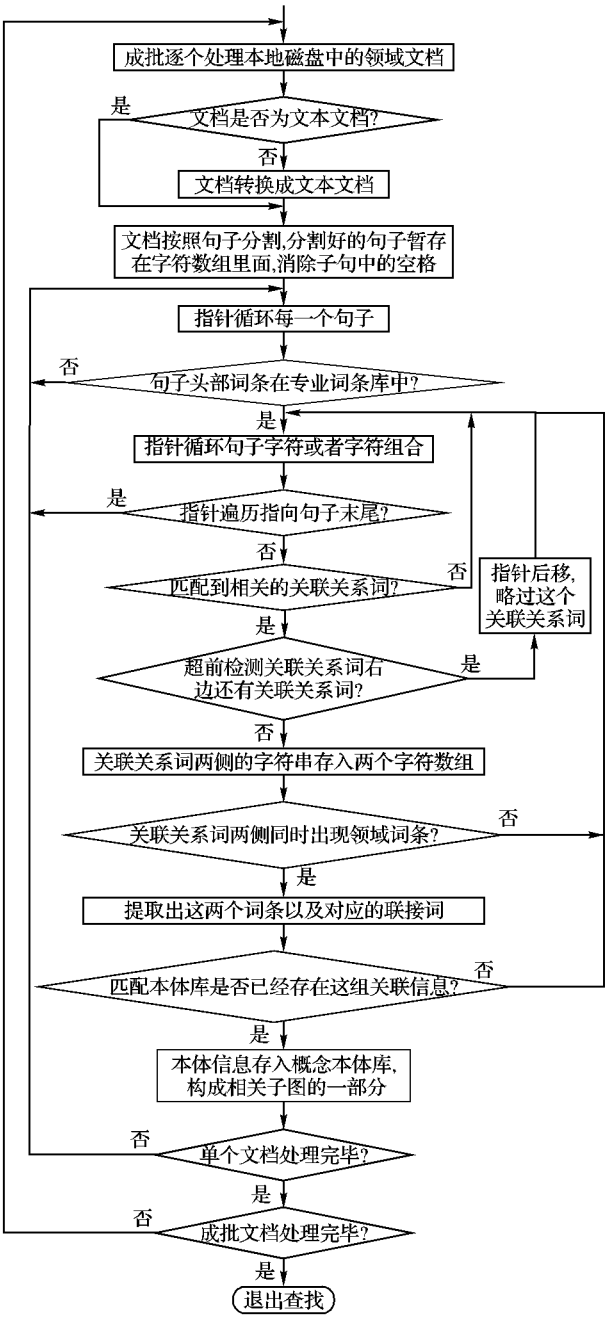

图 3　概念本体知识库构建流程

1.4　文本分词

所有文档转换成标准 TXT 类型后,首先将标准文档以句子为单位进行分割,之后每个句子通过专业词条和关联关系词进行二层过滤处理。

1)第一层过滤:遍历每个子句,从开始往后顺序遍历字符串匹配关联关系词。如果匹配成功,则进入第二层过滤,否则继续遍历句子。

2)第二层过滤:匹配关联关系词两边的专业词条,按照“最贴近关联关系词”以及“最长词条”的原则进行匹配。

所谓最贴近关联关系词就是当关联关系词匹配成功后,在匹配关联关系词两端的专业词条时,专业词条要尽量贴近关联关系词。如果专业词条与关联关系词之间的距离超过 2 个字符,那么这两个专业词条不能经过关联关系词构成有语义关系的知识结构。

所谓最长词条就是当关联关系词匹配成功后,在匹配关联关系词两端的专业词条时,关联关系词两端的专业词条可能会重叠出现在分词词条库中,例如“构造,滑脱构造”,可以选择“构造”或选择“滑脱构造”。为了消除歧义和达到最优,选择最长的专业词条作为最终匹配结果。

如果关联关系词两边的专业词条匹配成功,则以概念本体知识形式存入知识库中;如果关联关系词一端或者两端专业词条匹配不成功,则回到第一层过滤继续遍历句子。

1.5 概念本体知识存储

关系数据模型只能表达“属性/值”二元关系。通过对数据语义的提升,实现在关系数据库中语义的表达。知识的表示形式是:N 元关系 $R'(A_1,A_2,\cdots,A_n)$,A_i 的取值为 $P_i(i \in [1 \,..\, n])$,R' 是等价的一组 n 个二元关系:$A_i(H, P_i)$,其中 P_i 为语义节点,A_i 为节点 H 和节点 P_i 之间的弧[13]。概念本体知识结构形式:$co = \langle Concept_A, link, Concept_B \rangle$,其中 $Concept_A$ 和 $Concept_B$ 为概念,$link$ 为 $Concept_A$ 和 $Concept_B$ 间的关联关系词。概念本体知识库可抽象为集合 $KB = \{\langle Concept_A, link, Concept_B \rangle\}$。输入概念本体 co,概念本体知识库构建如下:

```
For each k ∈ KB
  If ∃ k, k = co then Exit For
    If ∃ k, k ∩ co ≠ ∅ then
      Append co. x to KB for co. x ∉ k
    Else
      Append co to KB
    Endif
End For
```

如果概念本体的部分 $co.x$ 在知识库,则通过增加标记的形式建立概念本体子类图关系;如果概念本体的部分 $co.x$ 不在知识库,则增加 $co.x$ 外,还增加标记。在程序实现上,海相油气地质的概念本体知识网状数据结构定义了如下:

```
struct knowledgenets
{public int ID;                          //唯一性标记
 public string link;                     //关联关系词
 public String conceptA;                 //概念 A
 public String conceptB;                 //概念 B
 public int map_sign;                    //子图标记
}
```

其中:ID 为本条记录的唯一性约束,表示概念本体在整个网络中都是唯一的;link 为关联关系词,表示概念本体的关联关系;概念 A 为核心词,在子图中主要描述的对象;概念 B 为资源或者属性,描述主语;Map_sign 为子图标记,标志概念本体知识属于哪一个子图以及在子图中的位置关系。

可以看出,知识库的构建可以避免概念本体的冗余,而且增强知识系统的知识完备性和一致性,保证知识系统的正确应用。

1.6 归一化处理

归一化处理主要包括:重复关联关系词消除和关联关系词规则化。

1)重复关联关系词消除。

当语义相同的两个关联关系词连在一起时,对其做了一个标记检测。只采用一个关联关系词,另外一个关联关系词在文本分词时不予处理。如“关键有利区带与构造是:通南巴山挠曲与多向构造复合”,其中“是”和“:”是两个关联关系词紧靠在一起,文本分词时先处理与“是”相关联的概念 A(关键有利区带与构造)和概念 B(通南巴山挠曲),处理完之后存入知识库,并对“是”做一个标识。当处理“:”时,系统会提示指针跳过“:”检测下一个关联关系词。

2)关联关系词规则化。

对于文档中出现的一些标点符号,赋予一定语义。对于不同的写作风格,在进行文本分词后,根据汉语言标点符号使用规则,将不规则的标点转换成规则的字符关联关系词或标点符号后再存入知识库。如“关键有利区带与构造:通南巴山挠曲与多向构造复合;镇巴正交迭加构造与滑脱构造;南大巴山推覆中带构造。”中的“:”统一转化成“、”之后再存入概念本体知识库。转换目的:1)词条检索后能以规范写作方法显示给用户;2)便于概念本体知识库中检测知识结构和知识结构扩展。

2 概念本体知识系统检索

海相油气地质概念本体知识系统的检索流程如图 4 所示。

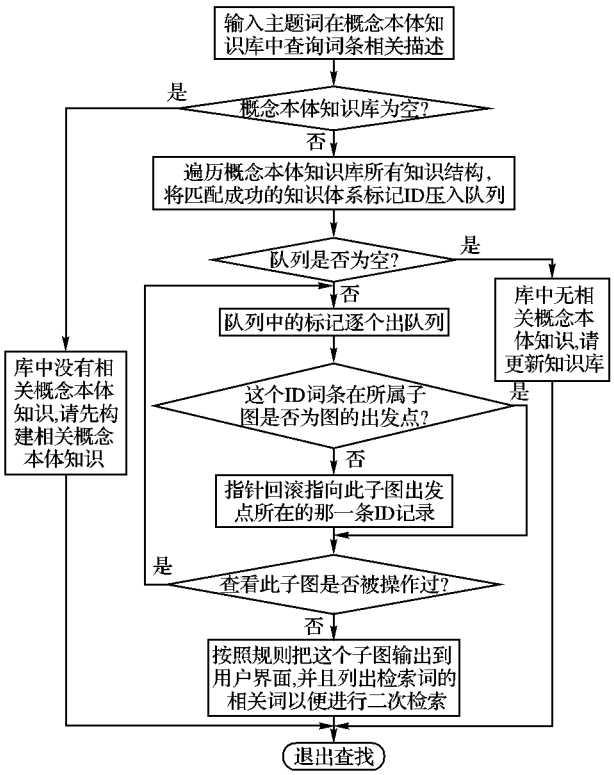

图 4 概念本体知识检索

2.1 语义检索与显示

主要采用队列结构来操作概念本体知识库,具体操作如下。

1)压队列 Enqueue 操作:输入检索词后,根据检索词遍历概念本体知识库。如果在概念本体知识库中检索到搜索词,那么把相关的概念本体子图的标记按照次序压入队列。

2)出队列 Dequeue 操作:按照“先进先出”的原则,根据每个子图标记和知识文字组成形式输出完整的概念本体知识。

概念本体知识以“关联关系词 + 谓语”的形式输出到浏览器端。例如:检索“黔南坳陷”,显示给用户“黔南坳陷……是立体勘探的有利地区”。具体算法如下:

1)输入检索词在概念本体知识库中顺序遍历匹配概念 A,匹配成功转 2);

2)将匹配成功的概念 A 对应的概念 B 压入队列,压入成功转 3);

3)概念本体知识库中的所有概念 A 部分都被匹配结束,则转 4),否则转 1)继续匹配;

4)将队列中的概念 B 部分全部输出到浏览器相关词位置显示给用户查看。

2.2　相关词检索

相关词是与搜索词语义上具有内在联系的词条,它们同处于概念本体知识库中有关联关系的节点上。通过搜索词检索到相应的概念本体知识后,同时也检索到所有的相关词(如图 5 所示)。通过相关词的列表选择相关词,进一步可以实现相关词的检索。

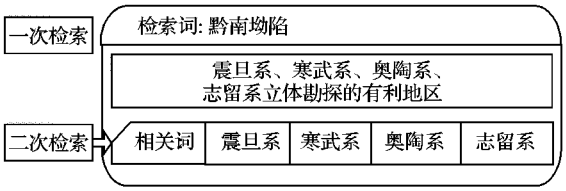

图 5　搜索词和相关词

如果用户对某个相关词感兴趣,可以直接点击这个相关词,系统会调用检索算法在知识库中检索这个词的语义描述以及与其对应的另外一组相关词,如果检索成功,则会把这个相关词的描述和相关词呈现给用户。

3　概念本体知识系统实现

1)海相油气地质知识管理系统是以海相油气地质综合知识库为基础、推理机为核心,智能处理器为关键,面向海相油气地质知识管理的集成软件系统,实现海相油气地质的智能信息处理和知识发布和共享(如图 6 所示)。整个系统在 Microsoft. NET + IIS 环境上开发,采用 B/S 三层软件体系结构,以 Oracle 为数据库服务器,Web Logic 和/或 IIS 为应用服务器,客户端为 Web 浏览器(如:IE 浏览器)。海相油气地质概念本体知识系统是海相油气地质知识管理系统的重要组成部分。海相油气地质资料文档通过“用户数据通道”进行词汇分割、专业词汇和关联词汇提取,采用概念语义挖掘算法,形成概念本体知识保留在概念本体库中。用户通过“用户知识通道”可以进行概念本体知识的检索,也可以用手工的方式进行本体知识的维护和录入。

2)中石化《中国海相油气地质信息系统建设及档案管理》项目中提供海相油气地质专题成果和研究成果数据库,包含大量的各种类型的文档材料。专业词条主要来自石油地质专业词典中的词条和海相油气地质专家提供的词条。关联关系词条主要采用海相油气地质专家总结、提炼使用频度比较高的词条,这些丰富、完备的基础词条确保文档词条的精确分割。实现了 AddWord 和 AddLinkWord 接口分别用于专业词条库和关联关系词库进行必要的增加、删除和修改的维护,也可通过该接口实现交互方式或批处理方式追加标准化、专业化的新词汇。目前,专业词和关联关系词在词条库中涵盖专业领域中的大部分基本词条,基本达到了海相油气地质概念本体知识系统的要求。

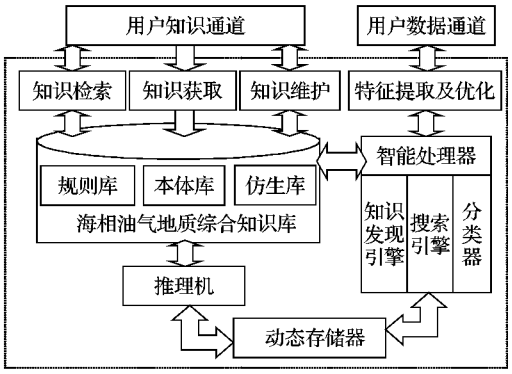

图 6　知识管理的体系结构

3)通过文档选择界面,选择 PPT、PDF、DOC、TXT 类型文档(如图 7 所示),以批处理方式,用于海相油气地质的知识获取。把符合条件的概念本体知识结构按照子图的形式存入到 Oracle 数据库(如图 8 所示)。为了方便用户了解文档内容,文档也可以转换为标准 TXT 文档后进行浏览。目前,已使用 PPT、PDF、DOC、TXT 各类文档 100 篇,总共大小20 GB。文档类型转换标准 TXT 的效率是 100 MBps, 文本词汇分割效率是 10 MBps。

图 7　选择各类文档

图 8　概念本体知识库

4)在海相油气地质概念本体知识系统中输入检索词(如

图9所示),根据概念本体知识结构和知识可视化规则,不仅显示出了搜索词的解释和描述,而且通过相关词算法在概念本体知识库中进行相关词条的检索,并显示出来,可进行相关词的检索。目前,概念本体知识结构的发现与存储的效率为微秒级别。概念本体知识只要是概念本体网络节点中存在,就能在概念本体知识库中检索。现在的词条库具有30万条基本覆盖的石油地质专业词条。词条库的完备程度直接影响到文本分词的效果和概念本体知识库的构建质量。目前,基于30万词条和根据已收集的海相油气地质相关概念的研究资料文档构建概念本体知识库之后,海相油气地质概念本体知识已涵盖该领域90%以上。

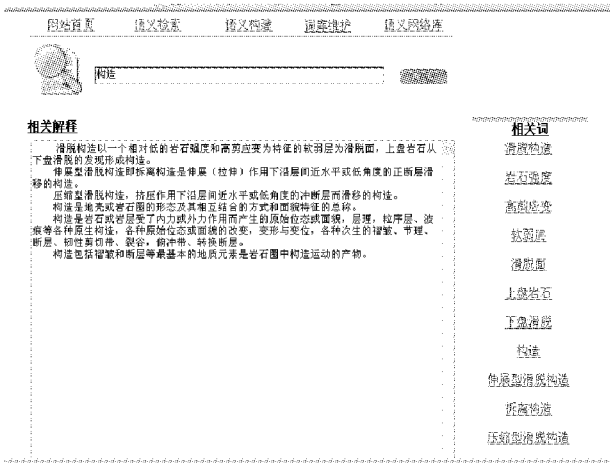

图9 语义检索

4 结语

海相油气地质概念本体知识系统是中石化《中国海相油气地质信息系统建设及档案管理》项目中"海相油气地质知识管理关键技术的研究"的组成部分。从海相油气地质知识利用的角度出发,利用本体论思想进行海相油气地质的知识表示,通过节点表示概念,有向边表示类型化的语义链,并以一定数据结构存入到关系数据库,形成海相油气地质概念本体知识原型。系统的设计目标是经过文档类型转换、文本分割、概念本体的分析与构建,解决海相油气地质知识从文档中自动获取的问题。海相油气地质的概念本体知识系统模型以海相油气地质知识库为基础,利用知识检索机制,实现了海相油气地质知识发布、共享,尤其是推广普及海相油气地质知识起到积极作用。但目前概念本体知识库还缺乏面向机器的应用,如通过本体知识的相似性计算,应用于海相油气地质的各类文档的查询等。此外,领域专业词条的质量和数量影响到概念本体知识获取的准确性,需要研究可行的方法,避免过度依赖专家,进一步完备知识系统。

参考文献:

[1] 王大亮, 孙建涛, 陆玉昌, 等. 基于HowNet构造语义场的方法[J]. 北京:清华大学学报:自然科学版,2005,45(1):78-80.
[2] NATALYA F N, DEBORAH L M. Ontology Development 101: A Guide to Creating Your First Ontology[D]. Stanford, USA: Stanford University, 2000.
[3] 邓志鸿, 唐世渭, 张铭, 等. Ontology研究综述[J]. 北京大学学报:自然科学版,2002,38(5):731-738.
[4] 李洁, 丁颖. 语义网、语义网格和语义网络[J]. 计算机与现代化,2007(7):39-41.
[5] 张德政, 庄洪波. 基于领域本体网络模型的知识获取技术[J]. 计算机工程,2007,33(7):190-200.
[6] 金顺爱. 中国海相油气地质勘探与研究———访李德生院士[J]. 海相油气地质,2005,10(2):1-8.
[7] 夏新宇, 陶士振, 戴金星. 中国海相碳酸盐岩油气田的现状和若干特征[J]. 海相油气地质,2000,5(1/2):6-11.
[8] 唐培丽, 王树明, 胡明. 基于语义的汉语文献主题词提取算法研究[J]. 吉林大学学报:信息科学版,2005,23(5):536-540.
[9] 庄洪波, 张德政, 赵秀君. 基于图的语义网络构造算法研究[J]. 计算机应用研究,2007,24(9):193-194.
[10] 马永生. 中国海相碳酸盐岩油气资源、勘探重大科技问题及对策[J]. 海相油气地质,2000,5(1/2):15.
[11] 黄伯荣,廖序东. 现代汉语(上)[M]. 北京:高等教育出版社,1997.
[12] 段寿建, 夏幼明, 甘健侯. 基于本体和Lucene的语义检索模型设计与实现[J]. 现代电子技术,2009,32(12):36-38.
[13] 张德政, 彭嘉宁, 范红霞. 中医专家系统技术综述及新系统实现研究[J]. 计算机应用研究,2007,24(12):7-9.

(上接第531页)

对感染速度较慢、危害较小的可疑蠕虫,在网络容忍范围之内,可以考虑不启动预警系统,从而避免由于启动预警而影响网络通信效率。当感染速度一旦超过这个容忍的阈值,说明有严重危害的蠕虫出现,需要马上启动预警系统,使网络处于警戒状态。

当N很大时,由于此机制是采用了C/S结构,势必会造成很大的通信量。可以采分层计算的思想,把这个N很大的网络分成很多个较为适中的网络,每一个相对独立的网络作为信息采集点,设计成多级容忍预警机制,可减少网络的通信量。

4 结语

本文综合蠕虫传播的关键因素,提出了一种基于贪婪算法的容忍预警机制。设计了估算启动预警系统的阈值($1/\sigma$)的具体方案,预测蠕虫感染速度最快的时刻,当超出网络容忍范围时才会启动预警系统,可以在一定程度上减轻网络的负载,提高了预警的效率。

参考文献:

[1] ZUO C C, GONG WEI-BO, TOWSLEY D, *et al*. Monitoring and early Detection for Internet worms[EB/OL]. [2009-08-01]. http://www-unix.ecs.umass.edu/~gong/papers/earlyDetectionJournal.pdf.
[2] 张新宇, 卿斯汉, 李琦, 等. 一种基于本地网络的蠕虫协同检测方法[J]. 软件学报,2007,18(2):412-420.
[3] 唐振江, 何慧, 云晓春. 基于多特征相似度的蠕虫检测[J]. 高技术通讯,2005(8):11-17.
[4] 姜启源, 谢金星, 叶俊. 数学模型[M]. 3版. 北京:高等教育出版社,2005.
[5] 任江涛, 孙婧. 一种基于信息增益及遗传算法的特征选择算法[J]. 计算机科学,2006,33(10):193-195,251.
[6] 李强, 康健, 向阳. 大规模蠕虫在线追踪培养皿[J]. 计算机应用,2007,27(11):2696-2698.
[7] FFRAGA J S, POWELL D. A fault and intrusion-tolerant file system[C]// Proceedings of the 3rd International Conference on Computer Security. [S.l.]: IEEE, 1985:203-218.
[8] KNG S T, CHEN P M, WANG Y-M, *et al*. SubVirt implementing malware with virtual machines [C]// Proceedings of the 2006 IEEE Symposium on Security and Priacy. Washington, DC: IEEE Computer Society,2006:314-327.