情报学报 ISSN 1000-0135
第27卷第2期163-168，2008年4月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.27 No.2，163-168 Apirl 2008

基于Web的情报知识元挖掘与语义集成地图[1)]

温有奎[1] 孙 明[1] 温 浩[2] 焦玉英[3]

（1. 西安电子科技大学经济管理学院，西安 710071；2. 西安电子科技大学通信工程学院，西安 710071；
3. 武汉大学信息管理学院，武汉 430072）

摘要 Web广泛使用的信息组织与表示语言HTML将显示方式内嵌在数据中，这使得应用程序很难将内容与显示方式分离开来。本文提出一个基于网页信息的知识元挖掘方法，通过汉语分词、词性标注预处理，用软件实现了具有三元组语义关系的知识元结构的挖掘，利用Protégé本体开发工具实现了基于知识元集成的军事情报语义网地图。试验表明，这是一种快速获取情报知识元的有效方法。

关键词 网页信息 知识元挖掘 语义网地图 军事情报

Information Knowledge Element Mining Based on Web and Semantic Integrating Map

Wen Youkui[1], Sun Ming[1], Wen Hao[2] and Jiao Yuying[3]

(1. *School of Economy and Management, Xidian Univerisity, Xi'an* 710071;
2. *School of Communiction Engineering, Xidian Univerisity, Xidian* 710071;
3. *School of Information Management of Wuhan University, Wuhan* 430072)

Abstract Information organization and expression language HTML, which is widely used in Web, embeds the display mode in the miscellaneous data, making it difficult for application program to separate the content from expression. This paper proposed a method of knowledge element mining based on the web information, through Chinese segmentation and syntactical functions notation preprocessing, we can get the mining of knowledge element structure with three element group semantics relationship realized by software. We used a Ontology development kit called Protégé to implement military Intelligence semantics network map based on the knowledge element integration. The experiment indicated this is a effective method to get Knowledge element.

Keywords Web Information, knowledge element mining, semantics network map, military Intelligence

1 引 言

随着Internet的飞速发展，Web已成为一个主要的信息来源。2006年11月2日，据美国CNN报道，Netcraft因特网监控公司的数据显示，全球WWW网站数量已经达到1亿个[1]。然而，人们发现万维网的海量信息越来越不能满足日益丰富的多样性需求。万维网上的信息虽然是机器可读的，却不是机器可理解的，由此导致了网上的信息难以被计算机自动处理。面对海量的网络信息，人工处理显然是不现实的。万维网的基石——HTML提供的链接缺乏语义，基于关键词检索的万维网搜索引擎的检索质量和效果远不令人满意[2]。当前Web的数据基本上以HTML形式存储和表示。HTML文档为显示而设计，缺乏针对内容的描述。对于文档的三要素

收稿日期：2007年1月10日

作者简介：温有奎，西安电子科技大学经济管理学院教授，主要研究方向：知识管理、知识挖掘。E-mail：wykui123@126.com。孙明，硕士研究生，研究方向：数据挖掘与知识发现。温浩，博士研究生，研究方向：模式识别与智能系统。焦玉英，武汉大学信息管理学院教授，博士研究生导师，主要研究方向：信息检索与现代咨询理论研究。

1) 国家自然科学基金资助项目(70373046)；国家自然科学基金资助项目(70473067)。

——数据、结构和显示方式，HTML将显示方式内嵌在数据中。HTML所表达的页面信息和组织方式，没有将内容形式、内在结构和表达方式相分离，也没有提供计算机可读的语义信息，因此限制了计算机在信息检索中的自动分析处理以及进一步智能化的信息处理能力。人们已经发现检索不能够只盯着字面，而应关注字面下隐藏的内容。因此，如何从Internet上智能地挖掘出有用的知识成为一个重要的研究课题。为了从海量网页汉字信息中获取有效知识，我们研究了基于网页信息的语义知识元挖掘，并将知识元集成为语义网地图，实现了快速建立军事情报知识元语义网地图的新方法。

2 Web资源的语义描述

2.1 Web信息表示

万维网上广泛采用超文本标记语言(Hyper Text Markup Language，HTML)来表达页面信息和组织方式，但HTML所表达的页面信息和组织方式主要面向用户直接阅读，没有将内容与表示分离，将显示方式内嵌在数据中，并且缺乏对数据结构的描述，这使得应用程序很难理解文档的内容，也很难抽取语义信息。

例如，用HTML来描述《知识元挖掘》这本书，其代码如下：

```
〈html〉
  〈head〉
    〈title〉知识元挖掘〈/title〉
  〈/head〉
  〈body〉
    〈h1〉书名:《知识元挖掘》〈/h1〉
    〈h2〉作者:温有奎　徐国华　赖伯年　温浩〈/h2〉
    〈h2〉 Email:wykui123@126.com 〈/h2〉
  〈/body〉
〈/html〉
```

在浏览器上，〈h1〉显示的字体比〈h2〉大，但从内容上我们并不能知道〈h1〉与〈/h1〉标记(Tage)之间的内容描述的是一本书名，同样也不知道〈h2〉与〈/h2〉标记之间的内容分别表示该书的作者和作者的Email，更不知道〈h1〉〈/h1〉与〈h2〉〈/h2〉之间是否存在关系，存在何种关系。HTML语言缺乏语义的先天不足，成为智能信息检索的一大瓶颈。

20世纪90年代中期，出现了更关注信息资源内容的结构、将数据的内容与布局分开来的XML(eXtensible Markup Language)可扩展标记语言，为语义丰富、更自然的网上内容表达打开了新的局面。同样上例的内容，XML的代码如下：

```
〈? XML VERSION = "1.0" ENCODING = "GB2312" standalone = "no" ?〉
〈! DOCTYPE Book SYSTEM "http://db.xdu.edu.cn/Book.dtd"〉
〈书〉
〈标题〉《知识元挖掘》〈/标题〉
〈作者〉温有奎 徐国华 赖伯年 温浩〈/作者〉
〈Email〉 wykui123@126.com〈/ Email 〉
〈/书〉
```

XML中的〈标题〉、〈作者〉、〈Email〉等标记表达了更多的内容和结构信息。

2.2 Web语义描述

W3C在XML的基础上推荐了一种标准，用于描述万维网上的信息资源的语义。RDF(资源描述框架)就是专门用于表达Web资源的元数据，如资源的标题、作者、版权、主题等信息，用来描述Web资源的特性及资源与资源之间的关系。RDF解决了XML缺乏语义的缺点。

RDF资源描述框架定义了由资源(Resource)、性质(Property)和语句(Statement)3种对象组成的基本模型。所有能用RDF表达式来表述的事物都可称为“资源”。资源可以是整个网页，例如HTML文档；也可以是网页中的一部分，例如文档中的某个HTML或者XML元素；资源还可以是整个网页集合，例如整个网站。性质被定义为用来描述资源的某个特定方面，如特征、属性或关系。每个性质都有特定的含义，规定了它的取值范围、所描述的资源的类型以及与其他性质之间的关系。语句由一个特定的资源、一个指定的性质和这个性质的取值组成。语句的这3部分分别称为主体(Subject)、谓词(Predicate)、客体(Object)。这3部分构成了一个资源语义描述的三元组$\{P,S,O\}$。语句的客体可以是另一个资源，也可以是一个常量。RDF语句可看作一个有向标记图：每个资源和文本都是一个节点，一个三元组$\{P,S,O\}$为一个从S指向O的标记为P的箭头。资源以椭圆形节点表示，文本以矩形节点表示，而指向的性质以箭头表示。RDF提供了比XML更多的语义信息。如《知识元挖掘》这本书的

作者是温有奎等、作者的 E-mail 是 wykui123@126.com 的匿名资源的扩展三元组如图 1 所示。

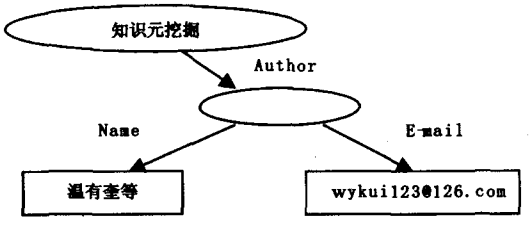

图 1 带有匿名资源的扩展三元组

其 RDF 三元组语句如下：

〈rdf: RDF〉〈rdf: Description about = "知识元挖掘"〉
〈Author〉温有奎等〈/Author〉
〈Email〉wykui123@126.com〈/Email〉
〈/rdf:Description〉〈/rdf:RDF〉

2.3 Web 语义推理

RDFS(Resource Description Framework Schema)是对 RDF 的一种扩充。RDFS 定义了类和性质(二元关系)，domain(定义域，或域)和 range(值域，或范围)约束，以及子类和子属性关系。这些类和性质可以用来描述其他类和性质，从而增强了 RDF 对资源的描述能力。然而 RDFS 没有定义推理机制，推理能力差。

本体(Ontology)是解决语义层次上万维网信息共享和交换的基础。作为一种能在语义和知识层次上描述信息的概念模型的建模工具，在计算机的许多领域得到了广泛的应用，如知识工程、数字图书馆、软件复用、信息检索和 Web 上异构信息的处理、语义 Web 等。

Berners-Lee 的语义网(Semantic Web)的提出激发了许多对标记语言的研究。W3C 总结了以上几种语言的开发经验，于 2004 年 2 月正式推出 OWL(Ontology Web Language)。OWL 是语义网发展中的一个重要的里程碑，它经过广泛的讨论并得到比较一致的认可。

OWL 从概念和属性两个方面对客观世界进行描述，描述的手段是面向对象域(Object Domain)的方式和面向数据类型域(Datatype Domain)的方式。面向对象域的方式采用 RDFS 和 OWL 自身的语法进行，用于描述概念间的分类化、层次化的继承关系，以及相互间的关联关系；用面向数据类型域方法描述时，OWL 支持 XMLS 的所有数据类型进行概念属性的定义与表达。OWL 具有描述复杂语义网和语义推理的能力。

3 智能网页文本知识元挖掘(ATKEM)系统

我们开发的智能网页文本知识元挖掘(ATKEM)系统主要包括网页格式转换、文本分词、词性标注、知识元自动抽取和知识元语义集成。

3.1 网页格式转换与分词

首先将网页行转换为 txt 文件。网页格式的转换可以采用软件方式，也可采用手工方式。转换后的 txt 文件用 ICTCLAS(中国科学院计算所汉语词法分析系统)进行分词、词性标注等预处理。对经过预处理的 txt 文件，使用我们开发的智能文本知识元挖掘(ATMEM)软件进行分析。该软件可自动挖掘出具有 6 个属性的知识元结构，其中具有语义网关系的三元组(对象名称、对象属性、对象属性值)结构的挖掘是智能文本知识元挖掘(ATMEM)的关键。

3.2 知识元自动挖掘

知识元自动挖掘算法的基本思路是：采用一个工作集合 $Fni(T)$ 来存放文本，将文本分解为段群，挖掘段群中的数值信息特征和时间信息特征，建立有效句，将有效句存放在工作集合 $Fn1(T)$；对 $Fn1(T)$ 中的句子按句号、分号和逗号进行句群语法分析，将满足知识元信息的句子存放到 $Fn2(T)$；对 $Fn2(T)$ 中的句子进行语义分析，分离出知识元的二元组语义结构，并将具有二元组语义结构的句子存放到 $Fn3(T)$；对 $Fn3(T)$ 二元组语义结构的句子分离出知识元三元组语义结构，则获得一个有效三元组知识元并存放到 $Fn4(T)$。否则转入下一个语句，如此反复直至 $Fni(T)$ 为空。将 $Fn4(T)$ 中的三元组元素进行语义合成，得到文本形式的知识元，并存放到 $Fn5(T)$ 形成知识元库。

(1)知识元有效句抽取算法

step1. 对每个文本进行段群分块处理。

step2. 提取段内时间特征值。

step3. 采用数值特征值对段内句子进行选择过滤。

step4. 知识元有效句采用知识元特征值决策树

判断。

step5. 判断值 = 1 为有效句,判断值 = 0 为无效句。

step6. 设文本分为 i 段;每个段内有 j 个句子;每个句子由"。""；""，""、"语法组成。

step7. for $i = 1$; $i = i + 1$; $i <= 1000000$

for $j = 1$:$j = j + 1$:$j <= 1000$

时间特征值提取,数值特征值定位,语法"。""；""，""、"特征符号定位分析,特征值送入决策树判断,记录此时句子序号 j;

if 判断值 = 1;

then 该句属于有效句,转到 step7,识别下一个句子;

else 转到 step7 中第一个 for 循环;

end

end

(2)三元组特征提取

通过句子的意群分析,我们将有效句分解为两部分,即对象名和对象数值。对对象名中的动词进行处理,由对象名中获取对象属性的信息,从而达到将一个有效句子分解成三元组(S,P,O)的目的,获得一条知识元[3]。如图2和图3所示。本文挖掘用的数据来源于文献[4]。

(3)去掉词性标记获得知识元

去掉词性标记,建立由时间、地区、领域、对象名称、对象属性、对象值等属性集成的知识元,并自动存入知识元库。最后可以对挖掘到的知识元进行汇总,存入到总表中,以便于以后的查阅和关联推理使用。结果如图3所示。

(4)知识元自动存储

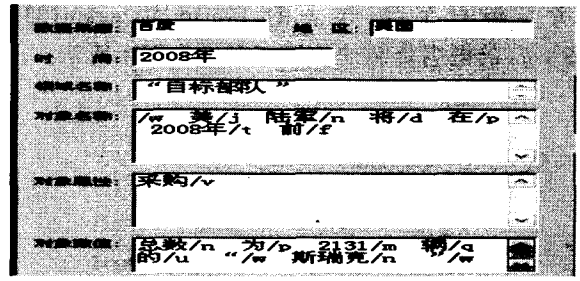

图2 获取对象属性

可以对挖掘到的知识元进行模糊和精确查询,并将查询结果反馈到用户界面,还可以将挖掘的信息生成简要文本输出。结果如图4所示。

4 知识元军事情报语义网集成地图

利用 Protégé 工具将挖掘出的知识元用本体语言 OWL 开发,实现了领域知识元集成的语义网地图。

4.1 Protégé 介绍

Protégé 是由斯坦福大学医学信息化研究小组开发的本体编辑和知识获取软件,是一个基于 Java 环境的开放式架构的开发知识建模工具。其扩展的 OWL 插件是目前最为强大的 OWL 本体构建工具。Protégé 不仅具有良好的可扩展性和简单灵活的用户定制界面,还具有如下一些特性:支持图形化本体编辑模式、支持数据库存储模式、基于 OWL 数据库的多人开发模式和支持逻辑检测功能等。最新版本的 Protégé 还增加了对资源多语言描述的支持。更为重要的是,Protégé 还拥有超过 60000 人的注册用户和

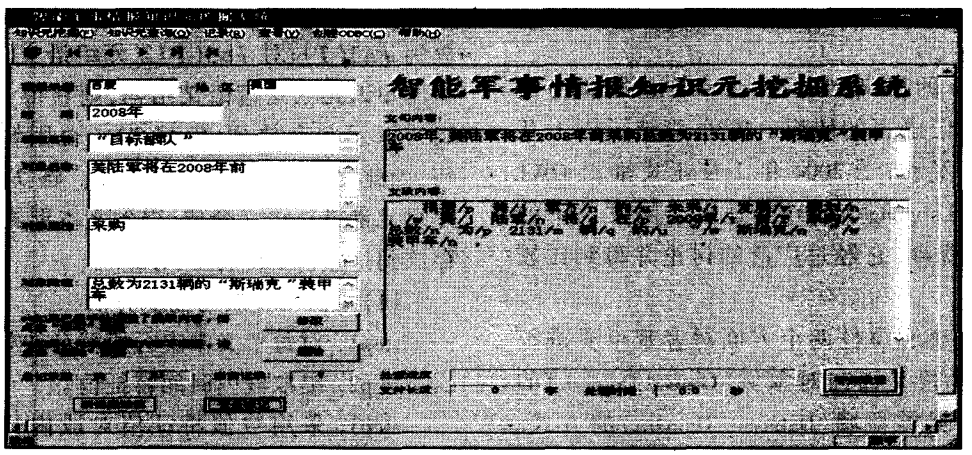

图3 去掉词性标记后获得知识元

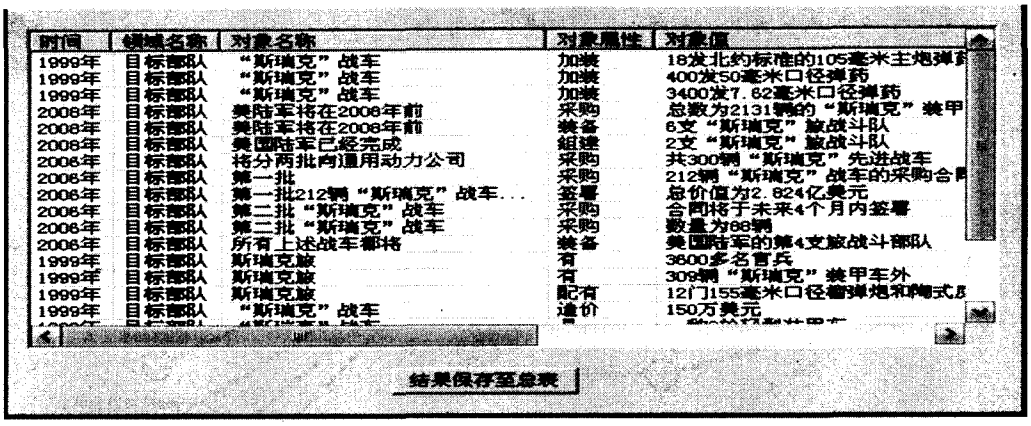

图 4　知识元库

邮件列表用户[5]，高效的技术服务支持以及丰富的技术资料和本体资源。

4.2　利用 Protégé 实现语义网地图

我们使用的本体建模工具是 Protégé 3.2 版本。开发过程有如下 4 个步骤：

(1)定义类和子类

打开 Protégé 的主页面会出现 OWL Classes(OWL 类)，Properties(属性)，Forms(表单)，Individuals(个体)，Metedata(元类)几个标签。选择 OWL Classes 进行编辑。在 Asserted Hierarchy(添加层)中会有所有类的超类 owl:Thing。点击 Asserted Hierarchy 旁边的 Create subclass，或者在 owl: Thing 点击右键选择 Create subclass，会出现 Protégé 自动定义名为 Class-1 的类。选中 CLASS EDITOR(类编辑器)的 Name 项，输入"对象值"来替换自动定义的名字。然后再分别建立"对象名称"、"领域名称"3 个子类(这 3 个子类是兄弟关系)。然后再分别在 3 个子类中继续添加要加入的子类(subclass)。

(2)建立互相排斥的属性

在"对象名称"中，"官兵"、"榴弹炮"、"装甲战车"、"陶式反坦克导弹"等属于同一类中的不同个体，它们互相具有排他性(owl:disjointWith)。这类对象要定义为互相排斥的属性。在选中"对象名称"的状态下，点击右下角的 Disjoints 的第 3 个按钮，在出现的 Add sibling to disjoints(将互为兄弟节点的类设为排他)对话框中，选择 Mutually between all siblings。这样"官兵"、"榴弹炮"、"装甲战车"、"陶式反坦克导弹"就有互相排斥的属性了。结果如图 5 所示。

(3)建立 ObjectProperty

新建一个 ObjectProperty(注意不是

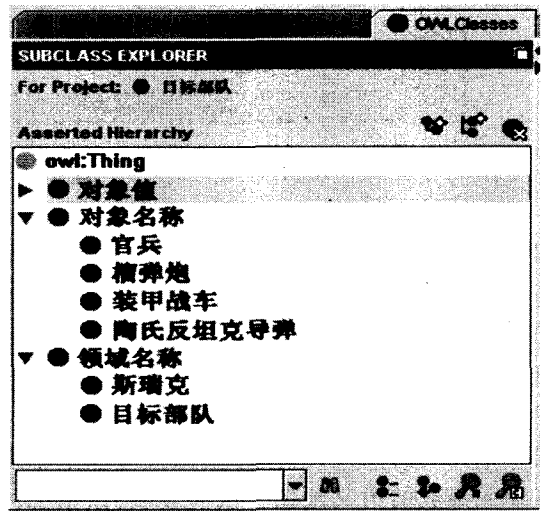

图 5　互相排斥的属性

DataProperty)，选择 Properties 标签，将 Name 改为"最高时速"，在 Domain(定义域)中把该属性的主体的类定义为"斯瑞克"，在 Range(范围)中把该对象主体所对应的作用范围定义为"每小时 100 公里"。以此类推，再将其他对象属性及其所对应的定义域和范围一一加入进去。

(4)获得语义网地图模型

选择菜单中 Project 下的 Configure，在 Configure file 中的 Tab widgets 选择 Jambalaya Tab(在前面打勾)，在 Jambalaya 标签中就可看到我们所建立的语义网模型了。结果如图 6 所示。

5　小　结

Web 正在成为信息的主要来源，而 Web 广泛使用的信息组织与表示语言却是主要用于浏览目的的

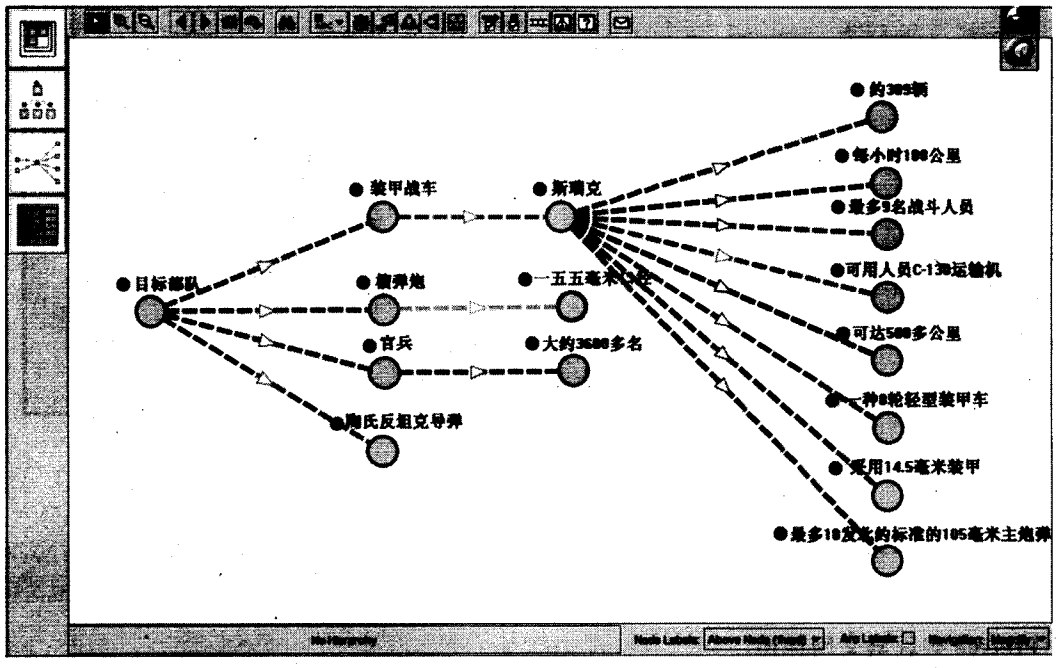

图6　“目标部队”语义网地图

HTML语言。为了有效地利用网上资源，网页挖掘成为一个重要的研究领域。我们开发了一个基于网页信息的知识元挖掘软件，成功地建立了智能军事情报知识元语义集成系统。本文利用该软件对预处理的文本信息进行汉语语义分析并挖掘出有语义关系的知识元结构，利用本体建模工具形成了军事情报知识元集成的语义地图。试验表明，该系统可以从网页上挖掘有用的知识元，并以此集成建立领域知识地图。这是一种快速获取情报知识元的有效方法。知识元库的建立以及不同文本格式的输出为用户的不同需求提供了个性化服务，更重要的是它为我们下一步实现推理和知识发现打下了数据基础。

参　考　文　献

[1] 风天. 全球网站数量今天达到1亿个里程碑[OL]. [2006-11-06]. http://news.zol.com.cn/42/422504.html.

[2] 宋炜，张铭. 语义网简明教程. 北京：高等教育出版社，2004.

[3] 温有奎，徐国华，赖伯年，温浩. 知识元挖掘. 西安：西安电子科技大学出版社，2005.

[4] 江　水.“目标部队”的利器——“斯瑞克”战车[J/OL]. 信息导刊，2004(14). http://www.people.com.cn/GB/paper2836/11781/1061973.html.

[5] Protégé 官方下载网站. http://protege.stanford.edu/.

（责任编辑　许增棋）