文章编号:1672-3961(2012)05-0035-06 DOI:10.6040/j.issn.1672-3961.2012.05.007

基于事件本体的 Web 不良信息挖掘

刘东慧[1,2],姜薇[1*]
(1. 中国矿业大学计算机科学与技术学院,江苏 徐州 221000;
2. 连云港师范高等专科学校计算机系,江苏 连云港 222000)

摘要:为挖掘互联网上的不良信息,本研究借鉴了事件语义分析技术。研究了基于事件本体的 Web 不良信息挖掘方法,重点是事件本体的构建、文本特征重构。为了验证所提方法的有效性,以信息聚类为例实现了一个基于事件本体的 Web 不良信息挖掘的原型系统。实验结果表明:基于事件本体和 k 均值的信息聚类方法的平均准确率为 72.1%,较之传统的基于 k 均值的信息聚类方法的平均准确率提高了 5.3%。

关键词:Web 挖掘;不良信息;聚类;事件本体;信息检索
中图分类号:TP393 **文献标志码**:A

Research on Web negative information mining based on event ontology

LIU Dong-hui[1,2], JIANG Wei[1*]
(1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221000, China;
2. Department of Computers, Lianyungang Teachers' College, Lianyungang 222000, China)

Abstract: In order to mine the negative information on the internet, the event-based semantic analysis technology was applied. The methods of event ontology-based Web negative information mining, especially event ontology construction and text feature reconstitution were studied. Information clustering was used as an example to validate the methods proposed. A prototype system based on event ontology was implemented. Experimental results showed that the average accuracy obtained by the event ontology-based and k-means method was 72.1%, which increased by 5.3% compared with the traditional k-means clustering method.

Key words: Web mining; negative information; clustering; event ontology; information retrieval

0 引言

Web 挖掘是针对互联网上大量的、非结构化的信息提出的一个新的研究方向。开发新的 Web 挖掘技术,以及对 Web 文档进行预处理以得到关于文档的特征表示,是 Web 挖掘研究的重点[1-4]。BROWN D E 认为互联网上的不良信息多是以事件的形式表现的[5],分析者关心的是事件发生的地点和时间等信息。NATH S V 提出了一种发现不良信息模式的方法[6],一条不良纪录包括时间、地点、概况和武器等信息。文献[7-9]提出基于 Web 信息挖掘的网络舆情分析技术,包括网络舆情发现、网络舆情溯源、网络舆情传播扩散模式以及评估舆情影响效果等。BRUIN J S 介绍了一种根据人的个性化内容[10],使用聚类的方法发现犯罪分子的异同。CHEN Hsinchun 借助计算机研究了犯罪的案例[11],主要涉及实体抽取、聚类、分类、字符串比较、社会网络分析等技术。袁占亭提出了一种基于概念的 Web 信息检索系统模型[12],给出了它的理论模型和工作机制,其核心技术是自然语言处理技术。

大量的研究已经表明,对 Web 上不良信息的特

收稿日期:2012-05-10
基金项目:江苏省高校自然科学研究项目(10KJD520008)
作者简介:刘东慧(1980 -),女,江苏连云港人,硕士研究生,主要从事概念格,信息检索等方面的研究. E-mail:denghui-liu@126.com
*通讯作者:姜薇(1966 -),女,江苏徐州人,副教授,硕士生导师,主要从事数据挖掘等方面的研究. E-mail:wjiang@cumt.edu.cn

征分析可以从事件的角度入手,借鉴事件的相关研究成果解决 Web 不良信息的挖掘问题。事件是由触发词标识,关联了对象、时间、地点等要素,比概念粒度更大的语义单元,事件之间有着内在的本质的联系。借助事件的语义分析技术挖掘 Web 犯罪信息是一种崭新的尝试。本研究借鉴了事件语义分析技术,研究了基于事件本体的 Web 不良信息挖掘方法,并以信息聚类为例实现了一个基于事件本体的 Web 不良信息挖掘的原型系统。

1 不良信息事件本体的构建

本体最初是一个哲学上的概念,十多年前被引入计算机领域中作为知识表示的方法并被广泛使用。STUDER R 给出了本体最流行的定义"共享概念模型的明确的形式化规范说明"[13]。这意味着本体是某些应用领域的概念以及概念间关系的预先定义的形式化表示。SANCHEZ D 提出了一种从 Web 文档中学习分类和非分类关系构建领域本体的方法[14]。在非分类关系的学习过程中,作者虽然没有明确提出"事件",但实际上已经使用了事件三元组模型,通过事件模型辅助本体的构造和学习。ZARRI G P 在论述语义 Web 时曾建议在传统的概念的本体上增补事件以更接近语义 Web 的目标[15]。HAN Y 提出了一种基于事件的人物本体模型,但是还仅限于人物本体[16],只是抽取人物涉及到的一些简单事件。刘宗田在分析了概念本体在表示事件信息不足的基础上提出了一种面向事件的本体模型[17-20],该本体模型以事件类及其关系为本体的语义单元。

通过对互联网不良信息类型的深入分析,并对每类不良信息涉及到的相关概念及其联系进行合理地设计和整理,运用 Protégé 体工具开发实现了不良信息事件本体知识库,此知识库可以为 Web 信息挖掘等应用提供辅助。

1.1 不良信息概念的内涵

不良信息类似于事件,通常包含参与人员、场所、时间、工具、类型 5 个部分,这 5 个部分即为描述一个不良信息最基本的元素。其中参与人员又可分为公安、受害人、不良信息发布者 3 类(见图 1),场所选择了最为常见的住宅、出租屋、发廊、宾馆、网吧等地点(见图 2),工具选择了最为常见的网络设备、专有网站、电话、计算机等,时间是仅有案发时间和破案时间两个基本属性的一个类,不良信息类型有色情信息、虚假信息、危害国家安全信息等共 15 个子类。

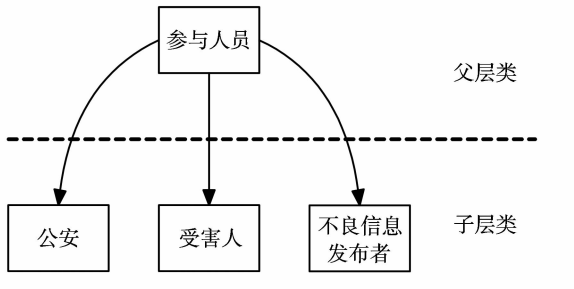

图 1 参与人员及其子类

Fig. 1 Participants and their subclasses

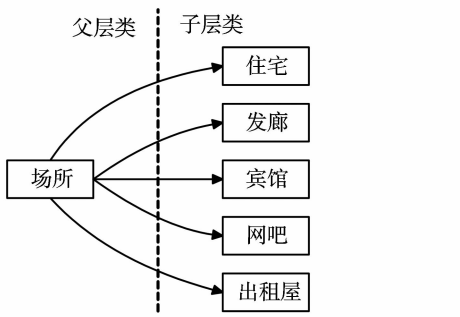

图 2 场所及其子类

Fig. 2 Place and its subclasses

1.2 不良信息的类型

根据调研情况,结合网络不良信息的现状,可将网络不良信息分为名誉损毁、敲诈勒索、煽动犯罪、版权侵犯、病毒破坏、传销、攻击、盗窃、色情、诈骗、赌博、迷信、证件货币、隐私机密侵犯、非法交易共计 15 种类型,其子孙关系见图 3。

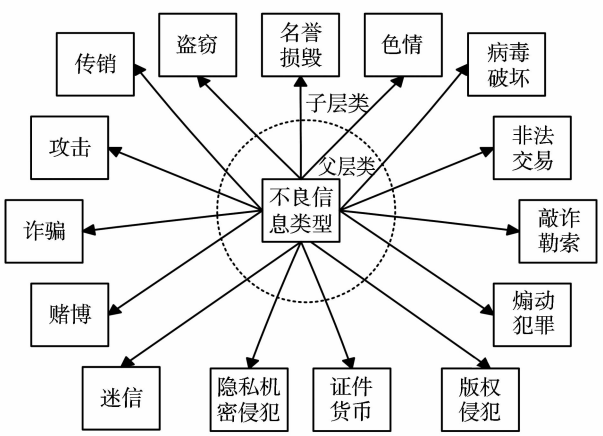

图 3 不良信息的类型

Fig. 3 Types of negative information

按照上述分类,可逐步完成知识库的创建,现仅以色情为例,列出用 Protégé_3.3.1 体工具创建知识库后生成的 OWL 部分代码如下:

```
〈owl:Ontology rdf:about = " "/〉
  〈owl:Class rdf:ID = "色情服务"〉
    〈rdfs:subClassOf〉
```

```
〈owl:Class rdf:ID = "色情"/〉
〈/rdfs:subClassOf〉
〈/owl:Class〉
〈owl:Class rdf:ID = "名誉损毁"〉
〈rdfs:subClassOf〉
〈owl:Class rdf:ID = "犯罪类型"/〉
〈/rdfs:subClassOf〉
〈/owl:Class〉
〈owl:Class rdf:ID = "盗窃"〉
〈rdfs:subClassOf rdf:resource = "#犯罪类型"/〉
〈/owl:Class〉
〈owl:Class rdf:ID = "淫秽光盘"〉
〈rdfs:subClassOf〉
〈owl:Class rdf:ID = "色情电影"/〉
〈/rdfs:subClassOf〉
〈/owl:Class〉
〈owl:Class rdf:ID = "色情小说"〉
〈rdfs:subClassOf〉
〈owl:Class rdf:ID = "色情文字"/〉
〈/rdfs:subClassOf〉
〈/owl:Class〉
〈owl:Class rdf:ID = "淫秽录像带"〉
〈rdfs:subClassOf〉
〈owl:Class rdf:about = "#色情电影"/〉
〈/rdfs:subClassOf〉
〈/owl:Class〉
…
```

1.3　不良信息概念的属性

不良信息概念的属性包括 DataTypeProperty 和 ObjectProperty 两种属性,DataTypeProperty 属性就是类(class)的基本属性。例如:警察的 DataTypeProperty 属性包括姓名、年龄、工号、服务地点等。而警察和犯罪嫌疑人之间存在着一定的关系,警察“抓”犯罪分子,犯罪分子与警察的关系是“被抓”,而且抓和被抓之间存在着逆反属性,这种属性称为 ObjectProperty 属性。

1.3.1　DataTypeProperty 设计　不良信息本体知识库的每一个类都有相关的特有属性,当发现有某几个相关的类有共同的属性时,为了便于知识库查询和推理,统一将公有属性作为父类属性,这样,各个子类便自然继承了父类的属性。

1.3.2　ObjectProperty 设计　ObjectProperty 指两个类之间的一种关联关系。不良信息事件本体中的每个类都不是孤立存在的,它必定与其他某个或者某些类有关联。在 ObjectProperty 设计的时候,必须将某个类与有联系的类进行关联,这样在做查询的时候,通过联想才能获取相关的概念。不良信息色情的 ObjectProperty 设计见图 4。

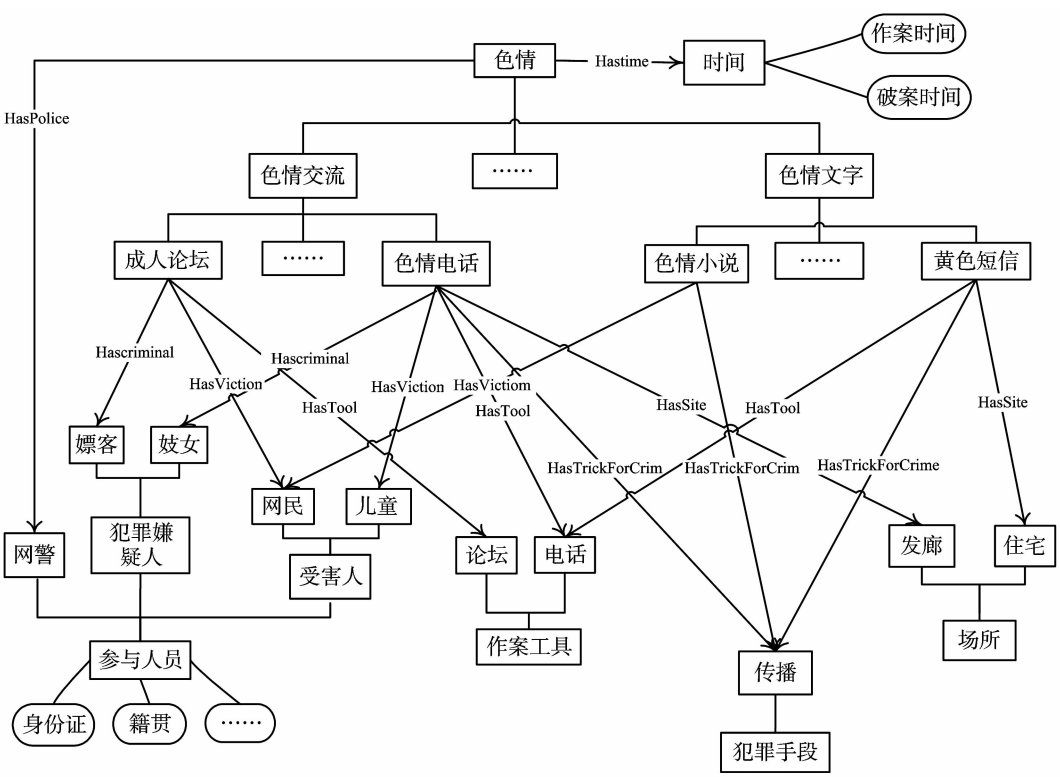

图 4　色情 ObjectProperty

Fig. 4　Pornography ObjectProperty

2 基于事件本体的 Web 不良信息挖掘

2.1 总体流程

图5给出了基于事件本体的 Web 不良信息挖掘的总体流程。

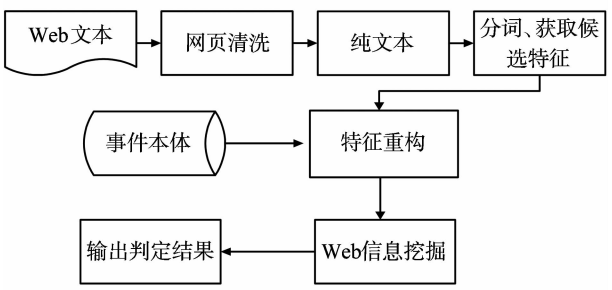

图5 基于事件本体的 Web 不良信息挖掘流程图

Fig. 5 Flow chart of Web negative information mining based on event ontology

图5所示流程主要包含以下4个模块：(1) Web 的预处理。Web 文本大多是 Html 格式的，经过清洗后，滤除了超级链接、广告等信息，留下了纯文本内容。(2) 获取候选特征。文本经分词、滤除停用词后，剩下的词作为文本的候选特征。(3) 特征重构。借助本体、同义词典等工具合并及补充某些特征。(4) Web 信息挖掘。信息挖掘包括各种技术，比如聚类、分类、检索、关联发现等，可根据需要选择使用。

本研究在实现实验系统时，模块(1)使用的是实验室已有的网页清洗工具，模块(2)使用的是中科院开源分词工具 ICTCLAS3.0。下面重点介绍文本特征的重构及使用了聚类技术的 Web 犯罪信息挖掘。

2.2 基于事件本体的 Web 文本特征重构

文本的特征是个高维、稀疏的问题。对文本特征进行降维，不仅提高了文本处理的速度，而且提高了文本处理的精度。事件本体的对象、时间、地点等要素在语言表现上都有不同的形式，但很多是同义的概念，比如事件“卖淫”的参与者：“妓女”、“小姐”、“三陪女”、“娼妓”等词，可以把其合并为概念“妓女”，既压缩了特征的维数，又增加了特征词“妓女”的权重。

有些文本的特征需要扩展后，才能更易处理，才能提高处理的精度。文本中有了某个事件，即使没有出现某些要素，借助事件本体也可以推断出文本省略的是哪些要素，根据事件与事件之间的关系，还可以联想到其他的事件。比如，信息“需要购买毒品的速和我联系”，对事件“购买毒品”，可以推断参与者有“吸毒者”、“贩毒者”等。又如，用户查询不良信息“法轮功”，根据事件本体可以联想到场所“明慧网”、参与者“李洪志”等。

2.3 基于 k 均值的 Web 不良信息聚类

信息聚类属于无指导的机器学习，它起源于信息分类，但是聚类不等于分类，聚类与分类的不同在于，聚类所要求划分的类是未知的。这适合于 Web 上不良热点信息的实时发现，因为在很多情况下，不良信息的热点都是未知的。

2.3.1 文本的表示 在本实验系统中，文本的表示采用了向量空间模型。向量空间模型是广泛使用的文本表示模型，它简便有效，且在实践中取得了很好的效果。

向量空间模型将文档映射为一个特征向量可表示为式(1)。

$$\boldsymbol{V}(d) = \{t_1, w_1(d); t_2, w_2(d); \cdots; t_m, w_m(d)\}, \tag{1}$$

其中 $t_i(i = 1, 2, \cdots, m)$ 为一列互不相同的词条项，$w_i(d)$ 为 t_i 在文档 d 中的权值，计算方法是 TF * IDF，TF 指词在文档中出现的次数，IDF 指该词在文档集合中分布情况的一种量化，常用的计算方法是 $\log(N/n_k + 0.01)$，其中 N 为文档集合中文档的数目，n_k 为出现该词的文章数。

2.3.2 文本特征的重构

文本经分词、过滤停用词后，其他的词都作为候选特征项。对文本进行特征压缩和扩展的具体做法如下：

(1) 事件本体的许多要素都有同义词的表达。事件本体相当于一个专用领域的同义词库，借助它可以实现文本特征的合并。

(2) 本研究只进行了短信息的特征扩展。规定凡是信息的长度小于 100 个字的就进行特征扩展。首先找到文本包含的事件类型，然后根据事件本体添加事件类所包含的诸要素到文本特征中，添加的要素的权值设为该事件在文本中的权值。

2.3.3 聚类实验

为了考查事件本体在 Web 信息聚类中的作用，用两种方法进行了实验对比：(1)基于 k 均值的聚类(记作方法 M_1)，直接将候选特征作为文本的特征；(2)基于事件本体和 k 均值的文本聚类(记作方法 M_2)，将文本的候选特征进行重构，重构后的特征作为文本的特征。

在实验中，语料分为迷信、色情、诈骗、赌博、非法交易共5个类别。针对每个类别，人工收集了30篇文本信息，共150篇文本信息。使用 k 均值聚类

时,指定 $k=5$,即聚类结果为 5 个类。聚类的时候,一篇文本只能聚到一个类别。

采用的评估指标是 F 值

$$F=\frac{P*R*2}{P+R}, \tag{2}$$

式(2)中: P 是准确率,计算方法如式(3)所示;R 是召回率,计算方法如式(4)所示。

$$\text{准确率}(P)=\frac{\text{聚类正确的文本数}}{\text{实际聚类的文本数}}, \tag{3}$$

$$\text{召回率}(R)=\frac{\text{聚类正确的文本数}}{\text{应该聚类的文本数}}。 \tag{4}$$

两种聚类方法得到的实验结果见表 1、2。

表 1　方法 M_1 得到的 F 值
Table 1　F value obtained by M_1

类别	应该聚类的文本数量	实际聚类的文本数量	聚类正确的文本数量	召回率/%
迷信	30	37	21	62.7
色情	30	29	20	67.8
诈骗	30	31	19	62.3
赌博	30	22	18	69.2
非法贩卖	30	31	22	72.1
平均 F 值	—	—	—	66.8

表 2　方法 M_2 得到的 F 值
Table 2　F value obtained by M_2

类别	应该聚类的文本数量	实际聚类的文本数量	聚类正确的文本数量	召回率/%
迷信	30	32	23	74.2
色情	30	29	21	71.2
诈骗	30	34	22	68.8
赌博	30	28	18	62.1
非法贩卖	30	27	24	84.2
平均 F 值	—	—	—	72.1

从表 1 和表 2 可见,基于事件本体和 k 均值的信息聚类方法,较之基于 k 均值的信息聚类方法,在 F 值方面有一定的提高。采用基于 k 均值的信息聚类方法,平均准确率为 66.8%;采用事件本体和 k 均值的信息聚类方法,平均准确率为 72.1%,平均准确率提高了 5.3%。但从聚类的结果来看,效果还不是很理想,这主要是与不良信息事件本体构建的规模有关。目前,整个不良信息事件本体共包含 392 个概念,规模偏小。这将直接影响到文本特征的重构,进而影响聚类的效果。

3　结语

本研究以不良信息事件本体为文本信息处理的语义资源,研究了 Web 犯罪信息的挖掘方法。以不良信息聚类为 Web 信息挖掘的实例,比较了基于事件本体和 k 均值的信息聚类方法和基于 k 均值的信息聚类方法的 F 值,实验结果表明了事件本体在 Web 犯罪信息挖掘应用中的可行性和有效性。下一步的研究重点是事件本体规模的扩大、应用领域的推广。

参考文献:

[1] 王继成,潘金贵,张福炎. Web 文本挖掘技术研究[J]. 计算机研究与发展,2000,37(5):513-520.
WANG Jicheng, PAN Jingui, ZHANG Fuyan. Research on Web text mining[J]. Journal of Computer Research & Development, 2000, 37(5):513-520.

[2] 邹涛,王继成,张福炎. 基于 WWW 的资料搜集系统的设计与实现[J]. 情报学报, 1999, 18(3):195-201.
ZOU Tao, WANG Jicheng, ZHANG Fuyan. Design and implementation of information gathering system based on WWW[J]. Journal of the China Society for Scientific and Technical Information, 1999, 18(3):195-201.

[3] 仲兆满, 刘宗田. 利用事件影响关系识别文本集合中重要事件的方法[J]. 模式识别与人工智能, 2010, 23(3):307-313.
ZHONG Zhaoman, LIU Zongtian. Identifying important events from texts using event influence relations[J]. Pattern Recognition and Artificial Intelligence, 2010, 23(3):307-313.

[4] LAWRENCE S, GILES C L. Searching the world wide Web[J]. Science, 1998, 280(3):98-100.

[5] BROWN D E. The regional crime analysis program (RECAP): a frame work for mining data to catch criminals [C]// Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. San Diego: IEEE, 1998: 2848-2853.

[6] NATH S V. Crime pattern detection using data mining [C]// Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. Hong Kong: IEEE, 2006: 41-44.

[7] 梅中玲. 基于 Web 信息挖掘的网络舆情分析技术[J]. 中国人民公安大学学报:自然科学版, 2007,54(4):85-88.
MEI Zhongling. The technology of network public opinion analysis base on Web information mining[J]. Journal of Chinese People's Public Security University: Science and Technology, 2007, 54(4):85-88.

[8] 张玉峰,何超. 基于 Web 挖掘的网络舆情智能分析研究[J]. 情报理论与实践, 2011, 34(4):64-68.
ZHANG Yufeng, HE Chao. Intelligent analysis study of network public opinion base on Web mining[J]. Information Studies: Theory & Application, 2011, 34(4):64-68.

[9] AGGARWAL C, ALGARAWI, YU P. Intelligent crawling on the world wide Web with arbitrary predicates [C]// Proceedings of the 10th International WWW Conference. Hong Kong: ACM, 2001.

[10] BRUIN J S, COCX T K, KOSTERS W A, et al. Data mining approaches to criminal career analysis[C]// Proceedings of the Sixth International Conference on Data Mining (ICDM'06). Las Vegas, Vevada: ICDM, 2006: 171-177.

[11] CHEN Hsinchun, CHUNG Wingyan, QIN Yi, et al. Crime data mining: an overview and case studies[C]// Proceedings of the National Conference for Digital Government Research. Boston, Massachusetts: IEEE, 2003: 45-48.

[12] 袁占亭,张爱民,张秋余. 基于概念的 Web 信息检索[J]. 计算机工程与应用, 2003, 39(36):173-175.
YUAN Zhanting, ZHANG Aimin, ZHANG Qiuyu. Concept-based Web information retrieval[J]. Computer Engineering and Applications, 2003, 39(36):173-175.

[13] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineer, principles and methods[J]. Data and Knowledge Engineering, 1998, 25(1-2):161-197.

[14] SANCHEZ D, MORENO A. A methodology for knowledge acquisition from the Web[J]. International Journal of Knowledge-Based and Intelligent Engineering Systems, 2006, 10:453-475.

[15] ZARRI G P. Semantic Web and knowledge representation[C]// Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02). Aix-En-Provence, France: IEEE, 2002: 1529-4188.

[16] HAN Y. Reconstruction of people information based on an event ontology[C]// Proceedings of Natural Language Processing and Knowledge Engineering, NLP-KE 2007. Beijing, China: IEEE, 2007: 446-451.

[17] 刘宗田,黄美丽,周文,等. 面向事件的本体研究. 计算机科学, 2009, 36(11):189-192.
LIU Zongtian, HUANG Meili, ZHOU Wen, et al. Research on event-oriented ontology model[J]. Computer Science, 2009, 36(11):189-192.

[18] 王海涛,曹存根,高颖. 基于领域本体的半结构化文本知识自动获取方法的设计和实现[J]. 计算机学报, 2005,28(12):2010-2018.
WANG Haitao, CAO Cungen, GAO Ying. Design and implementation of a system for ontology-mediated knowledge acquisition from semi-structured text[J]. Chinese Journal Of Computers, 2005, 28(12):2010-2018.

[19] 李曼,王大治,杜小勇,等. 基于领域本体的 Web 服务动态组合[J]. 计算机学报,2005,28(4):644-650.
LI Man, WANG Dazhi, DU Xiaoyong, et al. Dynamic composition of Web services based on domain ontology [J]. Chinese Journal of Computers, 2005, 28(4):644-650.

[20] ZHOW W, LIU Z T. Event-based knowledge acquisition for ontology learning[C]// Proceedings of the 6th IEEE International Conference on Cognitive Information (ICCI'07). Lake Tahoe, California: IEEE, 2007: 498-501.

(编辑:陈燕)