

· 理论与探索 ·

doi: 10.15908/j.cnki.cist.2015.01.006

面向军事信息系统结构化数据的信息汇聚方法

严红 黄颖 应励志

(中国电子科技集团公司第二十八研究所 南京 210007)

摘要: 面向军事信息系统大量结构化数据,结合语义本体、数据可视化、元搜索、垂直搜索和信息聚合等技术,提出一个具有语义智能的多手段综合信息汇聚框架,并分析了基于网络本体语言(OWL)的关系数据模型的本体抽取、基于本体的思维导图生成、基于联合搜索的垂直搜索和基于模板的信息汇聚等关键技术。

关键词: 信息搜索; 信息目录; 信息聚合; 本体; 思维导图; 垂直搜索

中图分类号: G230.7 **文献标识码:** A **文章编号:** 1674-909X(2015)01-0029-06

Information Aggregation Method for Structured Data of Military Information System

Yan Hong Huang Ying Ying Lizhi

(The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China)

Abstract: Aimed at structured data of the military information system, an integrated architecture of information aggregation with semantic characteristics is proposed based on technologies of semantic ontology, data visualization, meta-search, vertical search, and information aggregation, etc. Several key technologies are analyzed, including ontology extraction from structured data model based on ontology Web language(OWL), mind map creation based on ontology, vertical searching based on meta-search, and information aggregation based on template.

Key words: information search; information directory; information aggregation; ontology; mind map; vertical search

0 引言

经过多年的研制建设,我军在数据工程方面积累了不少成果,其中结构化数据是非常重要的部分,是各类概念、关系和模型的浓缩和精华,覆盖了军事信息系统各方面。但目前业界的各类语义本体、数据挖掘和信息汇集等技术,因数据表示不一致问题,一直难以在军事信息系统中推广应用,导致对这些结构化数据信息的再使用、知识提炼和深度挖掘的欠缺。在未来网络化和服务化体系结构下,如何利用新技术实现对各类军事结构化数据的深加工是关键技术之一。

1 信息汇聚系统典型框架

针对各类结构化军事信息数据库的信息汇聚系统构架如图1所示,分为以下3层:

- 1) 应用数据层:存储各类结构化军事数据和非结构化军事应用数据,以及相关元数据;
- 2) 模型层:主要包括支持搜索的索引文件及通用/专用词库,支持语义的军事信息本体库和网络本体语言(OWL)文件,支持汇聚的汇聚模板库及结果库,以及支持挖掘的挖掘模型库及数据库等;
- 3) 汇聚处理层:主要包括各类信息汇聚所需的应

用软件及工具,支持搜索的索引建立工具、联合搜索框架和专用搜索引擎等,支持本体的语义生成编辑工具和知识地图工具,支持数据挖掘的模型库管理工具,以及支持信息汇聚的模板管理和信息汇聚工具等。

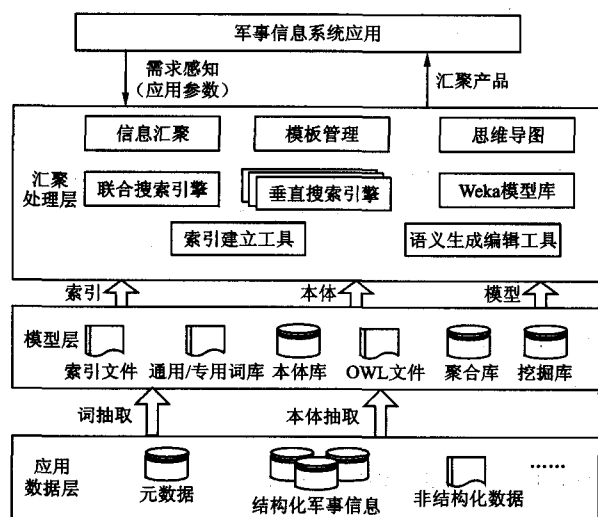


图1 信息汇聚系统框架

2 关键技术

2.1 基于OWL的关系数据模型本体抽取及存储

目前,军事信息系统已建立大量数据模型,但这些数据模型都以关系数据库形式进行存储和管理,适用于简单和基于规则的推理和计算。如何将其中的数据按照本体概念进行抽取,支撑各类计算模型和推理,是提升军事数据应用的关键。

目前,国内外从关系数据库到本体的映射方法研究较多^[1-4],关系数据库模式和本体间存在许多近似的对应关系。例如关系数据库模式中的表(table)可以对应到本体中的类(class),因此实现关系数据库和本体之间数据的互操作性可以通过构建关系数据库模式和本体间映射的途径解决。基于关系数据模型的本体抽取软件结构如图2所示。

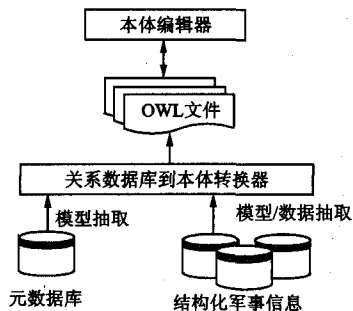


图2 基于关系数据模型的本体抽取软件结构

2.1.1 模式转换

关系数据的模式可通过元数据库进行抽取,或通过关系数据库管理系统中的元数据进行抽取,主要包括将军事数据库的表及其字段转换为类及其属性,表的每条数据记录转换为类实例对象,再根据表间参照关系设定类间关系(如同类和子类等)、属性间关系(如子、对称、传递和函数等属性)和属性的约束(如值域和基数约束等)。具体模式转换方法包括^[3]:

1) 关系数据库中不存在表 T ,将其转换为OWL模式中的同名类,表的描述信息(如Oracle数据库中表的comments)转换为类说明,即:

$$\forall T \in \text{RDB} \rightarrow \text{Class}(\text{ID}(T), \text{Cmt}(T))$$

2) 若关系数据库中的2个表 T 和 T_{sub} 之间存在父子关系(如Oracle数据库中 T 的主键(Primary key)是 T_{sub} 的参照外键(Foreign key)),则将表 T_{sub} 对应的OWL类声明为子类,表 T 对应的类声明为父类,即:

$$\forall T, T_{\text{sub}} \in \text{RDB} \wedge \text{Sub}(T_{\text{sub}}, T) \rightarrow \text{SubClassOf}(\text{ID}(T_{\text{sub}}), \text{ID}(T))$$

3) 关系数据库中的表 T 存在非外键字段 F ,将其转换为OWL模式中的同名数据类型属性。属性定义域(domain)为表 T 对应的OWL类,值域(range)为字段 F 的数据类型,字段描述(Oracle中的字段comments)转换为属性说明,即:

$$\forall T \in \text{RDB} \wedge \forall F \in \text{Field}(T) \wedge \neg \text{IsFKKey}(F, T) \rightarrow \text{DatatypeProperty}(\text{ID}(F), \text{domain}(\text{ID}(T)), \text{range}(\text{datatype}(F)), \text{Cmt}(F))$$

4) 关系数据库中不存在表 T_i 和 T_p , T_i 通过外键字段 F 与表 T_p 相关联,则将字段 F 转换为OWL模式中的同名对象属性,属性定义域为表 T_i 对应的OWL类,值域为 T_p 对应的OWL类,字段描述转换为属性说明,即:

$$\forall T_i, T_p \in \text{RDB} \wedge \forall F \in \text{Field}(T_i) \wedge \text{IsFKKey}(F, T_i) \wedge \text{Relation}(F, T_i, T_p) \rightarrow \text{ObjectProperty}(\text{ID}(F), \text{domain}(\text{ID}(T_i)), \text{range}(\text{ID}(T_p)), \text{Cmt}(F))$$

5) 关系数据库中不存在表 T ,该表主键仅由2个外键字段 F_a 与 F_b 组成,这2个外键分别与表 T_a 与 T_b 的主键相关联,该表除主键外不包含其他字段,则将字段 F_a 与 F_b 分别转换为同名对象属性,且两属性存在InverseOf关系,即两属性在定义域与值域上的取值互逆,表达式如下:

$\forall T, T_a, T_b \in \text{RDB} \wedge \forall F_a, F_b \in \text{Field}(T) \wedge \text{IsPKey}((F_a, F_b), T) \wedge \text{IsFKKey}(F_a, T) \wedge \text{IsFKKey}(F_b, T) \wedge \text{Relation}(F_a, T, T_a) \wedge \text{Relation}(F_b, T, T_b) \rightarrow \text{InverseOf}(\text{ObjectProperty}(\text{ID}(F_a), \text{domain}(\text{ID}(T_a)), \text{range}(\text{ID}(T_b)), \text{Cmt}(F_a)), \text{ObjectProperty}(\text{ID}(F_b), \text{domain}(\text{ID}(T_b)), \text{range}(\text{ID}(T_a)), \text{Cmt}(F_b)))$

6) 关系数据库表 T 中的非外键字段 F 取值不允许为空,则其对应的数据类型属性的基数限制 Cardinality 取值为 1,即:

$\forall T \in \text{RDB} \wedge \forall F \in \text{Field}(T) \wedge \neg \text{IsFKKey}(F, T) \wedge \neg \text{IsNull}(F) \rightarrow \text{Restriction}(\text{Class}(\text{ID}(T)), \text{Cardinality}(\text{DatatypeProperty}(F)), 1)$

7) 关系数据库表 T 中的非外键字段 F 取值允许为空,则其对应的数据类型属性的基数限制 MaxCardinality 取值为 1,即:

$\forall T \in \text{RDB} \wedge \forall F \in \text{Field}(T) \wedge \neg \text{IsFKKey}(F, T) \wedge \text{IsNull}(F) \rightarrow \text{Restriction}(\text{Class}(\text{ID}(T)), \text{MaxCardinality}(\text{DatatypeProperty}(F)), 1)$

8) 关系数据库表 T 中的外键字段 F 取值不允许为空,则其对应的对象属性的基数限制 MinCardinality 取值为 1,即:

$\forall T \in \text{RDB} \wedge \forall F \in \text{Field}(T) \wedge \text{IsFKKey}(F, T) \wedge \neg \text{IsNull}(F) \rightarrow \text{Restriction}(\text{Class}(\text{ID}(T)), \text{MinCardinality}(\text{ObjectProperty}(F)), 1)$

9) 关系数据库表 T 中的外键字段 F 取值允许为空,则其对应的对象属性的基数限制 MinCardinality 取值为 0,即:

$\forall T \in \text{RDB} \wedge \forall F \in \text{Field}(T) \wedge \text{IsFKKey}(F, T) \wedge \text{IsNull}(F) \rightarrow \text{Restriction}(\text{Class}(\text{ID}(T)), \text{MinCardinality}(\text{ObjectProperty}(F)), 0)$

2.1.2 数据转换

关系数据库数据转换为类的实例,主要是用关系数据库表里的数据为本体实例的各属性赋值而不涉及复杂的映射过程。数据转换主要经过以下 3 个步骤^[3]:

1) 将关系数据库表中的一个元组转换为 OWL 本体的一个实例,为每个实例分配唯一的标识符,考虑到每个元组的主键名唯一,因此本文将元组主键值作为实例的标识符;

2) 将关系数据库表中元组的非外键属性值转换成本体实例相应的 DatatypeProperty 的值;

3) 根据关系数据库元组的外键创建本体实例

之间的关系,将元组中外键值映射到一个本体实例,用该本体实例作为由外键生成 ObjectProperty 的值。

2.2 基于本体的思维导图生成

思维导图(Mind map)是一种常用可视分析方法^[5-6],又称心智图或脑图,源于以等级层次形式存储的图式解释,是有效表达发散性思维的可视化工具。思维导图的核心是将形象图画与抽象思维结合起来,用图画和线条表示思维的发散结构。它运用图文将各级主题的关系用相互隶属与相关的层级关系表现出来,将关键词与图像和颜色等建立链接,并利用记忆、阅读和思维的规律,找出逻辑与想象之间的结合点,从而记录思维过程并启发人类的大脑潜能。

本体中蕴含了大量概念/实体之间的关系,如知网(HowNet)包括上下位关系、事件必要角色框架、事件关系与角色转换以及同义、反义、整体部分、宿主-属性、实体-属性-属性值、实体-相应事件、制成品-材料和各种动态角色等关系^[7]。

在信息汇聚中,根据用户需求搜索关键词,可在本体库中搜索到相关概念/实体,并依据本体中相关关系,以图形化方式显示,形成思维导图或知识地图。图 3 给出了基于本体的思维导图软件框架结构。

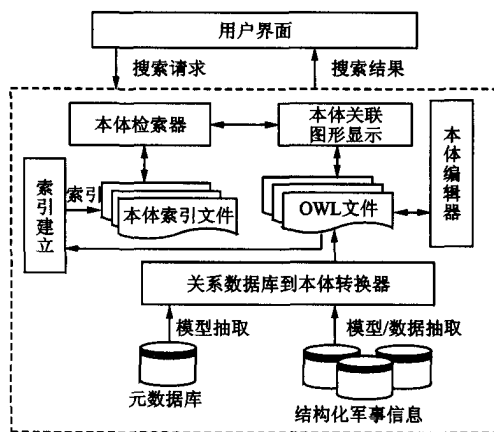


图3 基于本体的思维导图软件框架结构

本体编辑工具选用的 Protégé 是美国斯坦福大学的医学院研制开发的一个开放源码的本体编辑器,也是基于知识的编辑器。Protégé 是一种基于 Java 语言的开源软件,具有强大的集成环境及良好的扩展性,支持本体整合插件、本体可视化插件和语

言转换等应用插件。TGvizTab 是一种本体可视化插件,以可视化方式显示本体库中概念以及概念间相互关系。

2.3 基于元搜索的垂直搜索

搜索是支持信息汇聚的重要基础技术,在信息汇聚的过程中需用大量多种搜索技术和手段获取与用户需求相关的数据,通过基于元搜索的联合搜索框架进行各类搜索的集成和调用,并研制数据库、垂直和语义等搜索引擎。

元搜索引擎^[8-9]没有自己的数据,而是将用户的查询请求同时向多个搜索引擎递交,将返回的结果进行重复排除和重新排序等处理后,作为自己的结果返回给用户。其优点是返回结果的信息量更大、更全,缺点是不能充分利用使用搜索引擎的功能,用户需做更多筛选。

垂直搜索引擎^[8]是针对某个行业的专业搜索引擎,是搜索引擎的细分和延伸,是对网页库中某类专门的信息进行一次整合,定向分字段抽取需要的数据进行处理后再以某种形式返回给用户。因此,应建立面向军事应用的垂直搜索引擎,以解决实际军事应用问题。

元搜索引擎由检索请求提交、检索接口代理和检索结果显示机制 3 部分组成,其结构示意图如图 4 所示。

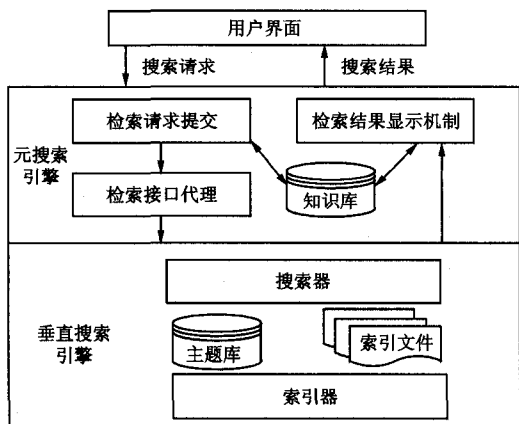


图4 元搜索引擎结构示意图

1) 检索请求提交:负责实现用户个性化的检索设置要求,包括调用的搜索引擎、检索时间和结果数量等限制。一般的元搜索引擎设定了调用的独立搜索引擎,如 widewaysearch;有些元搜索引擎让用户自己选择所用的搜索引擎,或通过分析用户兴趣和网络实际情况选择,这有利于提高用户的查询准确

度和对用户的响应速度。

2) 检索接口代理:负责将用户的检索请求转化成满足不同成员搜索引擎本地化要求的格式。由于不同搜索引擎所支持的查询方式不同,如有些搜索引擎支持 stemming 方式,即便是同一种方式,也有不同的表达方法,因此需将元搜索引擎中的查询请求映射到对应的搜索引擎中,使得语义信息不丢失。

3) 检索结果显示机制:负责将所有成员搜索引擎检索结果去重、合并、排序处理并按一定的格式显示。元搜索引擎的结果一般有网页标题、内容摘要、所指向网页地址、相关度、信息返回时间以及所采用的引擎标志等,这些搜索结果是多个独立搜索引擎的并集。元搜索引擎的结果应该具有多种排序方式以满足不同用户的需求。

搜索结果的聚类浏览技术用于信息检索结果的可视化输出以方便用户浏览。聚类算法和类别标签是聚类浏览技术的两个重要组成部分。聚类算法决定了搜索结果的组织结构和运行效率,而类别标签则是帮助用户迅速确认生成的文档类目相关与否的重要信息,是提高搜索结果可用性的基本体现。

2.4 基于模板并具有上下文感知的信息汇集生成

信息聚合指将来自于多个分布的、异构的信息资源中的内容整合在一起。在信息汇聚中,如何分解用户的需求是关键。参考任务类知识需求模型^[10],建立一个作战需求模型,将作战过程分解为具有逻辑约束和依赖关系的一系列任务单元。用户充当某些角色,作为任务的执行者,在工作中承担一定任务。在过程建模中将任务指派给某些角色,而用户为完成某些任务,在执行任务时对信息提出不同需求。

先建立任务类信息需求模板,可表示为 $R_{km} = \{A_c, M_{ac}\}$ 。其中 A_c 为任务类; $M_{ac} = \langle T, W, V \rangle$, T 为任务所需信息维度的集合; W 为获取信息的方法或过滤条件集合,过滤条件可组合为新的过滤条件; V 为任务需求的集合, $v_i = (t_i, w_i)$, $t_i \in T, w_i \in W$ 。

将作战过程定义为: $P_k = \{A, B, F, R, U\}$ 。其中 A 为由作战过程分解而得到的任务集合, $A = \{a_1, a_2, \dots, a_m\}$; $B = \{a_i, A_{ci}\}$, 为第 i 个任务所属任务类; $F = \{a_i, a_j, C\}$; C 为 a_i 和 a_j 之间的约束关系; R 为完成任务所需角色的集合, $R = \{R_1, R_2, \dots, R_m\}$, 其中 $R_i = \{r_1, r_2, \dots, r_n\}$; U 为用户集合。

结合指挥信息系统应用,即可感知作战阶段、用户角色和作战背景等,可自动匹配某一模板,实现自动的信息汇聚。

3 应用示例

在某项目中实现了一个通用的信息聚合软件,

汇聚模板设置样例界面如图 5 所示,图中为制定的钓鱼岛需求模板,设置了相关数据的维度和可搜索的关键词。根据该模板进行信息的汇聚,汇聚结果可按用户要求进一步整编,整编后信息汇聚结果示例如图 6 所示^[11]。同时提供一个知识地图生产和编辑界面,如图 7 所示。

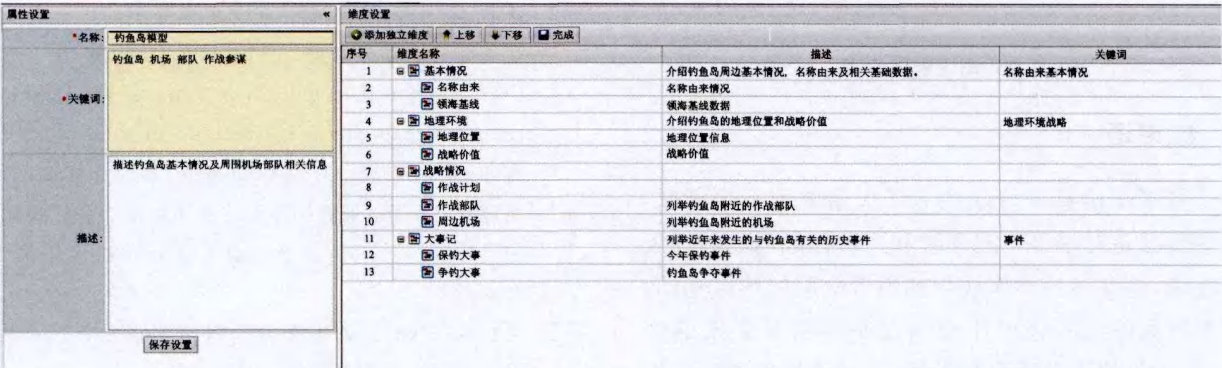


图 5 汇聚模板设置样例界面

钓鱼岛专题

目录

1 基本情况

1.1 名称由来

1.2 领海基线

2 地理环境

2.1 地理位置

2.2 战略价值

3 战略情况

3.1 战略部队

3.2 作战计划

3.3 周边机场

4 大事记

4.1 保钓大事

4.2 争钓大事

1.基本情况

介绍钓鱼岛周边基本情况,名称由来及相关基础数据。

1.1 名称由来

中国有关钓鱼岛的最早文献出自明朝永乐元年(1403年)的《顺风相送》,称该岛为“钓鱼屿”。其后文献及官方舆图亦采用“钓鱼屿”名称,见诸如明朝嘉靖十三年(1534年)第十一次册封使陈侃所著《使琉球录》、嘉靖四十一年(1562年)浙江提督胡宗宪编纂之《筹海图编》、清乾隆三十二年(1767年)乾隆皇帝钦命绘制之《坤舆全图》(《坤舆全图》使用闽南语发音,称为“好鱼须”[Hao-yu-su],即“钓鱼屿”)。台湾沿用“钓鱼台”名称至今。大陆现代则称该岛为“钓鱼岛”,有时也用“钓鱼台”的名称。

明朝郑若曾编著的《筹海图编》卷二《福建使往日本针路》记载,出使琉球使船须经小琉球、鸡笼山、梅花瓶、彭嘉山、钓鱼屿、黄麻屿、赤屿后才到达姑米山,因此钓鱼岛在明朝版图内,不属琉球国管辖,证明钓鱼岛在明朝归中国政府管辖。

1.2 领海基线

钓鱼岛、黄尾屿、南小岛、北小岛、南屿、北屿、飞濤屿的领海基线为下列各相邻基点之间的直线连线:

1.钓鱼岛1北纬25°44.1'东经123°27.5'

2.钓鱼岛2北纬25°44.2'东经123°27.4'

3.钓鱼岛3北纬25°44.4'东经123°27.4'

图 6 信息汇聚结果示例

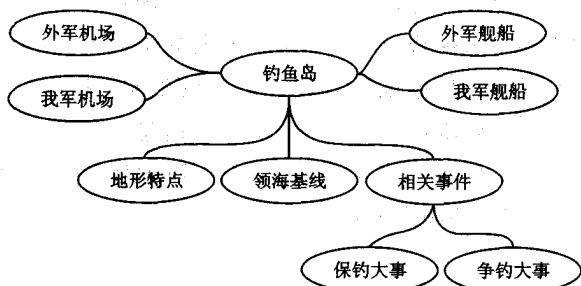


图7 知识地图界面

4 结束语

在军事信息系统中建立了大量的关系数据库,这些数据是对军事作战环境和军事作战活动的抽象和凝练,是支持各类作战指挥军事应用的数据基础。基于数据库的查询已不能满足军事信息系统的需求,智能化的语义搜索和发现、相关信息的关联和汇聚是未来发展的趋势。本文描述的基于结构化数据的语义抽取、垂直搜索和信息汇聚等方法,在军事信息系统领域初步提出,应具有广泛的应用前景,未来在分布式本体库集成和军事系统的语义推理知识库建立等方面将开展进一步研究。

参考文献(References):

- [1] 张晓明,胡长军,李华昱,等. 从关系数据库到本体映射研究综述[J]. 小型微型计算机系统,2009,30(7): 1366-1375.
Zhang Xiaoming, Hu Changjun, Li Huayu, et al. Survey on mapping from relational database to ontology [J]. Journal of Chinese Computer Systems, 2009, 30 (7): 1366-1375. (in Chinese)
- [2] 瞿裕忠,胡伟,郑东栋,等. 关系数据库模式和本体间映射的研究综述[J]. 计算机研究与发展,2008,45 (2): 300-309.
Qu Yuzhong, Hu Wei, Zheng Dongdong, et al. Mapping between relational database schemas and ontologies; the state of the art[J]. Journal of Computer Research and Development, 2008, 45 (2): 300-309. (in Chinese)
- [3] 何璐. 基于关系数据库的本体生成器的设计与实现 [D]. 武汉: 武汉科技大学计算机应用技术, 2008.

- [4] 冯波,郝文宁,宋杰,等. 基于关系数据库的军事训练本体自动构建模型[J]. 指挥信息系统与技术,2013,4 (5): 18-23.
Feng Bo, Hao Wenning, Song Jie, et al. Auto-constructed model for military training ontology based on relational database[J]. Command Information System and Technology, 2013, 4(5): 18-23. (in Chinese)
- [5] 陈欣,蓝国兴,段枫,等. 基于思维导图的仿真实验方法研究[J]. 工兵学报,2013,34(3): 346-352.
Chen Xin, Lan Guoxing, Duan Feng, et al. Design of simulation experiment based on the mind map[J]. Acta Armamentarii, 2013, 34(3): 346-352. (in Chinese)
- [6] 沈冠町. 基于扩展思维导图的协同讨论信息可视化系统设计[D]. 合肥: 合肥工业大学计算机应用技术, 2012.
- [7] 贾君枝,邵杨芳,刘艳玲,等. 汉语框架网络本体研究 [M]. 北京: 科学出版社, 2012.
- [8] 吴建强. 垂直搜索引擎爬虫系统的研究与实现[D]. 贵阳: 贵州大学计算机软件与理论, 2008.
- [9] 李红梅. 智能元搜索引擎关键技术研究[D]. 西安: 西安电子科技大学计算机系统结构, 2009.
- [10] 谢强,张磊. 基于任务类知识需求模板和用户模型的知识需求研究[J]. 武汉大学学报: 工学版, 2006, 39 (2): 36-41.
Xie Qiang, Zhang Lei. Study on knowledge need based on task class knowledge need model and user model [J]. Engineering Journal of Wuhan University, 2006, 39(2): 36-41. (in Chinese)
- [11] 钓鱼岛[EB/OL]. [2014-05-08]. <http://baike.baidu.com/view/2876.htm?sublemmaid=6475776>.

作者简介:

严红,女(1970—),研究员级高级工程师,研究方向为指挥信息系统总体和数据库技术。

黄颖,女(1975—),高级工程师,研究方向为软件总体和数据库技术。

应励志,男(1987—),工程师,研究方向为军事信息系统软件设计与开发。

(本文编辑:李素华)