

doi:10.3969/j.issn.1001-893x.2016.05.003

引用格式:俞鸿波.基于关联性的目标动向热点分析[J].电讯技术,2016,56(5):483-489.[YU Hongbo. Hot spot analysis based on correlation of target activity information[J]. Telecommunication Engineering, 2016, 56(5):483-489.]

基于关联性的目标动向热点分析*

俞鸿波**

(中国西南电子技术研究所,成都 610036)

摘要:为了从海量传感器数据中及时发现重要目标的动向,提出了一种目标动向信息表征及关联方法,即从多源异类传感器获取的信息中抽取出目标关联要素进行动向表征,利用语义决策树实现动向要素聚类,通过知识规则进行关联匹配扩展,从而发现目标动向的热点,并进一步统计分析目标活动规律与发展趋势。实验表明所提出的基于关联性的目标动向热点分析算法准确率高,具有实用价值。

关键词:信息关联;目标动向;热点分析;语义决策树

中图分类号:TN971;TP391.4 **文献标志码:**A **文章编号:**1001-893X(2016)05-0483-07

Hot Spot Analysis Based on Correlation of Target Activity Information

YU Hongbo

(Southwest China Institute of Electronic Technology, Chengdu 610036, China)

Abstract: A target activity expression and correlation method is presented in order to extract important target activity from varying and massive sensor data. Activity factors are extracted from various information to denote target activity, then, semantic decision tree is used to classify activity factors, correlative relationships between sensor data are established using knowledge and rules. By extraction of target activity information, it is possible to analyze hot spot or predict target activity trends by correspondent historical target activities. Experiment shows that the proposed target activity information correlation algorithm has high precision, which can be used in real data set.

Key words: information correlation; target activity; hot spot analysis; semantic decision tree

1 引言

随着传感器技术的发展,传感器用户每天可获得大量的数据,如何从海量信息中及时发现重要目标的动向是用户关注的重点问题。传统的信息分析方法主要是采取人工比对与综合的方式进行,如今,利用关联分析技术提取出有用信息,有助于情报人员快速分析并做出决策。

关联是目前常见的术语,主要用于关系型数据

库、语义网 Web、数据挖掘等应用领域。在关系型数据库应用领域中,关联指的是一种数据的某些特性变化会引起另一种数据对应变化的关系;在语义网中,关联数据是指用以 RDF/XML 格式描述的元数据信息,且通过链接指向其他 RDF 数据的数据,即通过网址链接形成富含元数据信息的数据关联;在数据挖掘领域,关联分析主要是指关联规则挖掘^[1],即发现数据库的项和属性之间存在的不易察

* 收稿日期:2016-01-29;修回日期:2016-03-28 Received date:2016-01-29;Revised date:2016-03-28

** 通信作者:walkeryhb@163.com Corresponding author:walkeryhb@163.com

觉的联系,而这些联系往往预先未知,且无法通过结构化数据库的简单查询操作获取。最著名的关联规则是 Apriori 算法^[2],其设计思想是通过多次迭代运算找出频繁项集,从而发现各数据项之间的隐藏关联关系,生成形如 $X=>Y$ 的规则。

在针对目标动向研究领域,国内外的研究重点主要集中在命名实体识别和突发事件的识别等。其中,命名实体识别技术的主要研究成果包括中科院自动化所的汉语词法分析系统(Institute of Computing Technology, Chinese Lexical Analysis System, ICT-CLAS)等;突发事件的识别大多面向特定的领域,如金融领域、网络故障领域、灾难性事件等,成果包括北京邮电大学陈莉萍等进行的突发事件识别的研究^[3]。文献^[4-5]针对文本信息中人物目标动向的感知与关联方面进行了相关研究,在完成文本中命名实体识别后,再运用句法分析建立人物动向语义信息,通过本体建模形成的关系模型抽取人物动向之间的关联关系。上述方法在应用中的主要限制在于仅针对文本信息中的人物动向进行分析,且前期本体建模工作量巨大。

本文主要以海空机动目标为研究对象,着重研究通过关联分析技术实现该类目标动向识别与聚类,并发现目标活动的热点、规律与趋势,其中热点包括热点平台、热点区域及相关的热点行为。

2 目标动向概念模型及动向关联模型

目标是指具备行为能力的实体或者物体,目标动向是指目标的一切行为活动。根据所使用传感器的类型,能够获取到目标动向信息的载体包括文本、图像、视频、格式化数据等。经过对相关资料的整理,目标动向可描述如下:

目标动向:={特征<固有特征、动态特征、感知特征、状态特征>、要素<发生时间、发生地点、目标行为、人物、关联人物、平台、关联平台、关联组织机构>、模式<规律、异常>}。

其中,目标动向模型由目标自身蕴含的特征和目标动向行为等来表征,如图 1 所示。目标特征包括固有特征、动态特征、感知特征、状态特征,固有特征是目标独特唯一具有的不变特征,例如目标机动能力、外形特征等;动态特征是目标的行为规律;感知特征是目标被传感器捕获后表现出的特征,如目标在声、光、电上表现出的特征;状态特征指目标当前具备的状态,如巡逻、警戒等。目标动向由动向要素和动向模式组成,其中动向要素是组成动向的实体,包括时间、地点、行为等,是组成目标动向的主

体;动向模式包括规律和异常,是目标动向在时空层面表现出的统计特性。

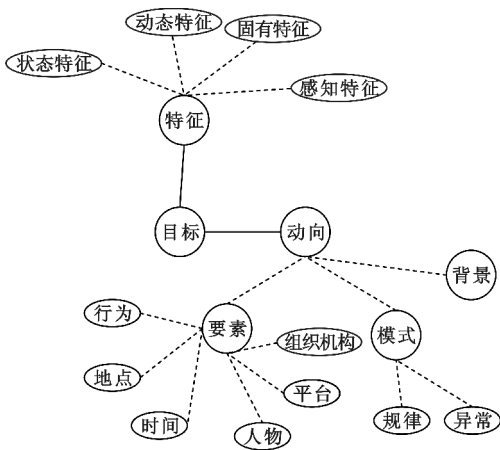


图 1 目标动向概念模型
Fig. 1 Target activity conceptual model

在应用中,目标动向分析通常以某一事件或目标作为中心或起点进行研究,其中经常从时空和类别层面对相关本体或事件进行关联组织,便于更进一步的挖掘深层次信息。虽然当前对目标动向关联的研究需求迫切,并有一些相关概念的研究,但在宏观格局上的认识并未形成体系,对目标动向间联系的类型划分也缺乏专门的探讨,导致现有概念之间缺乏好的继承和整合。因此,本文基于目标动向模型演化出目标动向关联模型,试图对概念进行整合和宏观把握。

按照目标动向的关联要素,可将其划分为目标主体型关联、背景主题型关联和行为模式型关联。其中,目标主体型关联主要关注目标动向中的相关主体特征及要素,包括时空主体、平台主体、人物主体、特征主体等;目标动向由于显见或潜在突出或不突出的主体因素也能产生关联,如某舰队行为和舰队指挥官的行为之间存在潜在关联性;行为模式型关联主要关注目标间活动模式之间的关联,比如具有相同活动模式与规律的海监巡航编队;主题型关联是指具有相似或者相同背景主题而产生的广义目标或动向事件的关联,譬如以东海巡航为讨论主题,关联目标动向包括南海维权事件和建设东海防空识别区等。基于上述研究范畴,围绕目标动向的主体事件出现多方面的关联线索与方向,形成如图 2 所示的目标动向关联模型。由于主题间具有隐式关联关系,主题间的关联关系可依据动向要素的核心性共现词语进行关联。

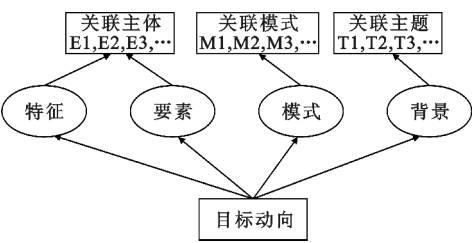


图 2 目标动向关联模型
Fig. 2 Correlation model of target activity

3 目标动向要素关联分析方法

由前文所述,目标动向信息中既包含了目标的基本特征属性,又包括了目标动向的基本要素描述,还包含了目标行为的深层次语义特征和相关知识。其中动向要素形成了目标动向的底层描述,可以从传感器数据中通过统计、匹配、识别等方式获取。目标动向要素建立了传感器数据和目标动向模式的纽带,基于目标动向要素的关联分析可支撑目标动向模式的深度挖掘与分析。因此,本文中目标动向信息分析的主要思路就是从传感器数据中抽取目标动向要素,并建立不同源传感器数据中基于目标动向的关联关系,在目标动向相关性的基础上进一步发现目标活动的热点、规律与趋势。

如图 3 所示,目标动向要素关联分析方法处理流程分为如下步骤:

- (1)从多种类、多格式、多媒介的海量信息中抽取目标动向要素;
- (2)构建目标动向要素特征向量,利用语义决策树对目标动向要素进行关联聚类;
- (3)对目标动向要素进行知识扩展,进一步关联出潜在的动向主体;
- (4)结合动向背景,对动向主体进行统计、相关排序、预测,发现动向热点。

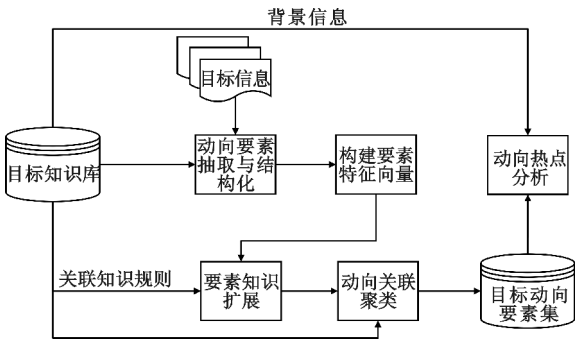


图 3 目标动向要素关联分析方法处理流程
Fig. 3 Flowchart of target activity factor correlation

3.1 目标动向要素抽取与结构化

3.1.1 时空结构化

传感器原始数据种类虽多,但数据的基本属性一般都包含空间属性和时间属性。其中,空间属性通过经纬高或者地址名称等方式表达了目标活动的相关区域;时间属性则是通过将目标的历史活动轨迹投影到时间轴的方式,来表达目标动向的时间点或者时间段。在实现关联前,必须对各类数据进行时空化分类处理以统一格式和标准,如图 4 所示。

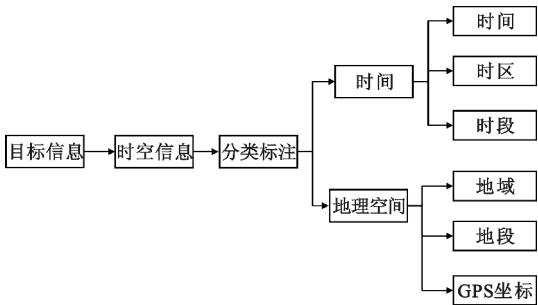


图 4 时空化数据分类处理
Fig. 4 Class of temporal and spacial data

本文使用 ICTCLAS 自动分词系统对原始数据进行分词,在得到分词序列的基础上进行自动词性标注,然后进行命名实体识别,提取时空信息。其中,基于 863 词性标注集,利用条件随机域 CRF 词性标注器进行词性标注^[6],使用 Viterbi 动态优化方法,即可求出最佳词性标注序列。

(1)时间信息抽取

词性标注后的分词序列仍然是信息量较少、语义不完整的单词,需要通过命名实体识别,识别出有语义的短语和词组。有语义的短语和词组就是要识别的命名实体,通常通过模板规则来表征实体要素的表达规则。依据中文命名实体(NE)识别方法,即可建立时间表达的实体规则。规则包括:数字+年|月|日,例如,2012 年,09 年 12 月;数字+“年代”或“世纪”,例如,20 世纪,90 年代;数字+时|分|秒|点,例如,上午 09hh:08mm:17ss。

(2)空间信息抽取与同化

地理空间的实体之间有包含和非包含的关系,包含关系的地名实体直接组合成一个实体,非包含关系的实体合并成集合形式。由于命名规则极不规范,应用时为了快速准确定位,必须进行规范化操作和地理同化。

方式一:自动匹配切词技术。将地址按照数据库中地址/地名信息自动匹配切词,该方式的优点是具有唯一性和高效性。

方式二:GPS 定位。对 GPS 类定位数据,提取数据的经纬度位置、高度信息,并转换到统一的椭球坐标系下。同时,根据数据的产生以及传感器特点(包括传感器的类型特点、精度等),计算数据定位的误差范围,进行定位匹配处理。

方式三:地址库匹配。即将地址数据与在知识库中存储的地理数据信息进行相关性匹配,从而获取相应的地址信息。

3.1.2 主体信息结构处理

动向要素中主体信息包括人物、关联人物、平台、关联平台、目标行为等要素,对该类信息的抽取根据传感器数据文件类型的不同而有所区别。对于图像类数据,利用图像解译算法对数据自动或交互式添加语义标签。例如,利用目标检测与识别等算法对数据自动添加语义标签,并将目标名称及属性对应的语义标签添加到平台要素中。对于视频类数据,则首先进行视频文件的结构化处理,结构化过程首先通过边界检测方法将视频分割为多个镜头,然后提取每个镜头的关键帧来代表整个视频的内容。在完成了镜头边界检测和关键帧提取后,利用前述的图像解译算法及语音识别算法实现对视频内主体信息的提取。对于文字类信息,通过建立专业领域词典,提取各要素,例如:目标行为要素包括巡逻、侦察、航渡等。具体过程本文不再赘述。

3.2 动向要素特征向量的语义决策树关联

根据提取出的描述目标动向事件的基本要素构建一个描述动向的特征向量,假设描述第 i 项传感器中包含向量 $\mathbf{X}^{(i)}$:

$$\mathbf{X}^{(i)} = (\mathbf{A}_1^{(i)}, \mathbf{A}_2^{(i)}, \mathbf{A}_3^{(i)}, \mathbf{A}_4^{(i)}, \cdots, \mathbf{A}_n^{(i)})^T. \quad (1)$$

式中: $\mathbf{A}_n^{(i)}$ 代表第 n 维的动向要素。

多源传感器获取目标信息使用基于要素知识模板的向量空间表示后,成为规则的动向要素描述数据。然后利用语义决策树的方法,依据要素关系进行关联。选择语义决策树的主要原因一是特征向量由非数值型的属性组成,决策树算法在聚类过程中可比对属性的标称值,适合于标称属性向量的非数值计算;二是多源目标信息关联是一个知识发现的过程,对于能挖掘出多少知识信息不能确定,很大程

度上与数据隐藏了多少潜在的知识有关,因此,不能预计关联聚类结果的数目,决策树能够依据属性取值进行自由分裂和合并,从而适应聚类数目不确定的情况。

在决策树算法应用中,通过从根结点到叶子结点的排列来分类样本实例。对于目标动向所属的文件来说,决策树的结点代表了文件中目标动向的所属分类,而每个叶结点则对应动向所包含特征向量中要素的对象取值。通过选择最有助于分类目标动向实例的属性作为建立决策树的上层结点,所以根结点一般是特征向量中最优的属性。

使用信息增益来统计并衡量属性区分给定训练样例的能力^[7]。为了精确定义信息增益,使用熵刻画任意样例集的纯度。给定包含 c 个聚类的样本集合 S ,那么 S 的熵定义为

$$Entropy(S) = \sum_{i=1}^c -p_i \lg p_i. \quad (2)$$

式中: p_i 为样本集合 S 中属于类别 i 的概率。

此时,属性 A 相对于样本集合 S 的信息增益 $Gain(S, A)$ 被定义为

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v). \quad (3)$$

式中: $Value(A)$ 是属性 A 的取值范围集合; S_v 是样本集合中属性值为 v 的子集合,即 $S_v = \{s \in S | A(s) = v\}$ 。通常具有最高信息增益的属性是最好的属性,使用最好的属性作为建立决策树的根节点。本文使用 ID3 算法训练, ID3 算法增长树的每一步使用属性的信息增益作为选取最佳属性的度量标准。

由训练集 S 构造最佳决策树的算法描述如下^[7]:

创建结点 N 。

如果 S 为空,则返回节点,标记失败。

If S 中的元组都在同类 C 中, Then 返回 N 作为叶子结点,用类 C 标记结点 N 。

If 属性列表为空, Then 返回 N 作为叶结点,用 S 中的多数类标记结点 N 。

For splitting_criterion 每个输出 j

设 S_j 是 S 中满足输出 j 的数据元组的集合;

If S_j 为空

增加叶结点 N , 标记为训练集中的多数类;

Else

增加由 Generate_decision_tree(S_j , Attribute_list) 返回的

结点到 N 。
End For
返回根节点

说明:Generate_decision_tree 函数通过对决策树各分支的子集递归调用的方式建立决策树结点的分支。

根据动向特征向量中要素构成的假设决策树理论上有很多,但很多决策结果在实际应用中比较少见,本文关注的决策树主要是目标在地点的活动行为,使用 Tree(目标,地点,行为,时间)表示。语义决策树的关联聚类见图 5。

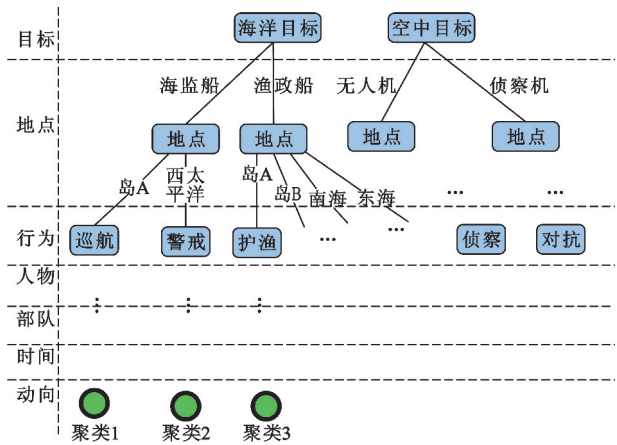


图 5 语义决策树关联聚类处理

Fig. 5 Correlative classification of decision tree

语义决策树的聚类结果代表目标动向要素属性值合取的析取式,每一条路径都代表了一类目标动向的聚类结果。例如,<目标=中国海监编队∧地点=某海域∧时间=2012 年∧行为=巡逻>代表了在“某海域”主题下搜索到的信息中,2012 年中国海监船编队在某海域活动的行为等情况。

3.3 扩展关联查询

基于要素的聚类关联能够发现某个主题下海空机动目标的活动情况,根据关联知识规则进行关联人物和关联组织机构的扩展,能够发现关键人物的活动信息,并发现潜在目标关联信息,如从某舰船编队的动向扩展关联出编队指挥官的相关行为和言论。关联要素的扩展查询需要建立基于目标要素的知识规则库。知识规则库是基于要素词典建立的知识扩展映射表格,是基于领域知识对要素项的进一步联想与扩展。

根据决策树聚类 and 扩展查询的结果,把传感器

数据聚类成不同的分组,保存数据之间的分组关系,作为已知信息用于后续热点分析。

3.4 动向热点分析

动向热点定义为指定时间段内频繁发生的目标动向,对海空机动目标来说,就是某类目标频繁的行为活动。下面介绍动向热点的分析方法的具体步骤。

(1)对某类目标的活跃度进行评估。在上述关联聚类结果中,对目标名称或者类属进行相似性检测,经过统计后即可发现目标活动的频率、目标热点活动区域、目标活动的热点时间段,并对目标活跃度进行演化预测,若演化预测出的目标活跃度提升达到预警值,则提示相关部门采取应对措施。

本文中,目标动向的趋向性演化预测采用马尔科夫预测模型^[8],具体描述如下:

第 t 个时期的状态概率向量可表示为 $S(t) = (S_1^{(t)}, S_2^{(t)}, \dots, S_n^{(t)})$,其中 $S_i^{(t)}$ 表示第 t 个时刻预测系统处于状态 i 的概率,所以有 $S_i^{(t)} \geq 0 (i = 1, 2, \dots, n)$ 和 $\sum_{i=1}^n S_i^{(t)} = 1$ 。特别地,初始化状态概率为 $S_1^{(0)}, S_2^{(0)}, \dots, S_n^{(0)}$,相应的初始状态概率向量为 $S(0)$ 。在进行目标动向趋向性分析时,马尔科夫预测模型的核心是状态转移方程,表述如下:

$$S(t) = S(t-1) * P = S(t) * P'$$
 (4)

式中: P 为初始的概率转移矩阵。则未来第 t 个时期概率 $S(t)$ 可根据初始的概率向量和初始状态转移矩阵推测出。

(2)对于关联出的热点目标动向信息,通过对应传感器信息文件的主题进行相似度计算,将相似度高的主题提取出来,作为目标动向对应的热点主题呈现给用户。

主题相似度计算采用相似度 $SimF(c1, c2)$ 来计算:

$$SimF(c1, c2) = \begin{cases} 0, & distance(c1, c2) = 0 \\ \frac{1}{\alpha \times distance(c1, c2) + 1} - \frac{1}{5}, & distance(c1, c2) > 0 \\ \frac{1}{\beta \times |distance(c1, c2)|} - \frac{1}{5}, & distance(c1, c2) < 0 \\ 1, & distance(c1, c2) = \infty \end{cases}$$
 (5)

式中:参数 α 和 β 值的大小决定了主题 $c1$ 和 $c2$ 间语义相似度的取值。在同样的距离下, α 和 β 的值越大对应的相似度就越小,通常语义距离为 1 时, α 和 β 的取值要保证相似度在 0.5 以上。

4 仿真结果

4.1 数据描述

以某海域事件为背景构建仿真数据,仿真某海域事件主题下的3 000份文件,包括文本、图像、视频和目标数据,其中涉及到的主题包括“巡航”、“登岛维权”等共7大类13个小类。

4.2 测评指标

可由热点识别准确率和召回率对热点分析结果进行量化分析。热点分析准确率指标定义为

$$R=\frac{|E_d\cap E_b|}{E_d}。$$
 (6)

式中: E_b 表示标示出的热点类别; E_d 表示经过关联分析后检索出的热点类别。

4.3 关联及热点识别结果

利用3.1节方法,对某海域冲突事件仿真数据经过抽取后,形成包括“海监船编队”、“渔政船”、“某无人机”等平台及对应的时间要素、地理空间要素和行为要素,根据3.2节和3.3节介绍的语义决策树方法完成关联聚类后进行热点分析,共形成平台-地理-行为热点17类,经过交互类属合并后为14类。关联出目标动向后,分析出的热点主题包括巡航、领土安全、购岛、中日关系等。通过式(6)计算,热点分析准确率达到92.8%,满足了应用中对动向热点准确识别的需求。

以月度作为统计周期,统计了半年的传感器数据,对其中关注的典型目标包括海监船、海监飞机和渔政船的活跃度进行评估,形成了图6所示结果,可以看出到2014年4月形成了动向峰值,与主题聚类的结果形成了互印证。

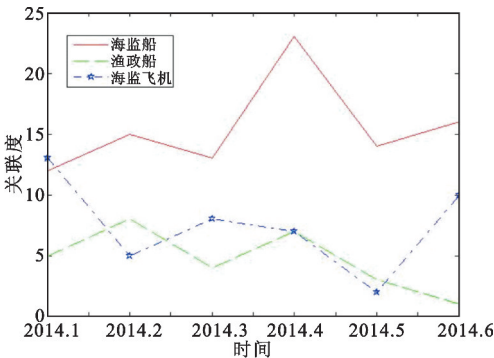


图6 典型目标动向趋向性分析

Fig.6 The activity trend analysis graph of concerned targets

5 结束语

本文主要研究了目标动向信息表征及关联方法,从多源异类传感器获取的信息中抽取出目标关联要素进行动向表征,利用语义决策树实现动向要素聚类,并通过关联知识规则进行匹配扩展,建立目标动向间的关联关系,从而发现目标动向的热点,并进一步统计分析目标活动规律与发展趋势。文中重点描述了目标动向概念模型及动向关联模型的构建及动向要素的关联聚类方法。与传统的方法相比,本文方法的主要优点在于:

首先,方法针对的数据类型包括多源信息数据,除了文本之外,还包括了图像、视频、格式化传感器数据等;

其次,方法综合使用基于规则的命名实体识别技术和模式识别技术解决目标动向的抽取与聚类,避免了采用复杂的句法分析,从目标行为宏观语义的角度实现了目标信息关联;

此外,本文的方法区别于关键词检索的不同在于,关键词检索使用户只能查询哪些信息文件中出现了用户关注目标,返回的结果集数据量巨大,因此,还需要结合上下文采用人机交互的方式,从结果集中综合判定获取目标信息的正确性;而本文方法则是在关联过程中,利用目标动向的背景知识进行主题事件的聚合与印证,关联过程不再是仅用特定词语的检索,更加具有针对性和智能化的特点,因此,关联分析的结果更加精确。

后续工作重点突破目标行为的语义理解与描述,从更细粒度分析动向特征的内涵与模式。

参考文献:

[1] 程舒通,徐从富. 关联规则挖掘技术研究进展[J]. 计算机应用研究,2009,26(9):3210-3213.
CHENG Shutong,XU Congfu. State-of-art on association rules mining technology [J]. Application Research of Computers,2009,26(9):3210-3213. (in Chinese)
[2] 张冲. Apriori 算法在物联网(IOT)数据挖掘中的应用[J]. 科技通报,2014,30(11):162-164.
ZHANG Chong. Study of IOT data excavating based on apriori algorithm[J]. Bulletin of Science and Technology,2014,30(11):162-164. (in Chinese)
[3] 陈莉萍,杜军平. 突发事件热点话题识别系统及关键问题研究[J]. 计算机工程与应用,2011,47(32):19-22.
CHEN Liping,DU Junping. Study on hot topics identification and key issues about emergency events[J]. Com-

- puter Engineering and Applications, 2011, 47 (32) : 19 – 22. (in Chinese)
- [4] 陈天莹, 苏智慧. 基于语义推理的文本信息关联关系分析技术 [J]. 电讯技术, 2014, 54 (1) : 67–73.
CHEN Tianying, SU Zhihui. Text information relationship analysis based on semantic reasoning [J]. Telecommunication Engineering, 2014, 54 (1) : 67–73. (in Chinese)
- [5] 陈天莹. 文本信息中的目标动向感知技术 [J]. 电讯技术, 2014, 54 (6) : 730–734.
CHEN Tianying. Target activity awareness on text information [J]. Telecommunication Engineering, 2014, 54 (6) : 730–734. (in Chinese)
- [6] 周晶, 吴军华, 陈佳, 等. 基于条件随机域 CRF 模型的文本信息抽取 [J]. 计算机工程与设计, 2008, 29 (23) : 6094–6097.
ZHOU Jing, WU Junhua, CHEN Jia, et al. Using conditional random fields model for text information extraction [J]. Computer Engineering and Design, 2008, 29 (23) : 6094–6097. (in Chinese)
- [7] 王小巍, 蒋玉明. 决策树 ID3 算法的分析与改进 [J]. 计算机工程与设计, 2011, 32 (9) : 3069–3076.
WANG Xiaowei, JIANG Yuming. Analysis and improvement of ID3 decision tree algorithm [J]. Computer Engineering and Design, 2011, 32 (9) : 3069–3076. (in Chinese)
- [8] 章登义, 欧阳黼霏, 吴文李. 针对时间序列多步预测的聚类隐马尔科夫模型 [J]. 电子学报, 2014, 42 (12) : 2359–2364.
ZHANG Dengyi, OUYANG Chufei, WU Wenli. Cluster-based hidden Markov model in time series multi-step prediction [J]. Acta Electronica Sinica, 2014, 42 (12) : 2359–2364. (in Chinese)

作者简介:



俞鸿波 (1977—), 男, 湖北人, 2005 年于西北工业大学获博士学位, 现为高级工程师, 主要研究方向为图像处理与信息融合。

YU Hongbo was born in Hubei Province, in 1977. He received the Ph. D. degree from Northwestern Polytechnical University in 2005.

He is now a senior engineer. His research concerns image processing and information fusion.

Email: walkeryhb@163.com