CN43-1258/TP
ISSN 1007-130X

计算机工程与科学
COMPUTER ENGINEERING & SCIENCE

2012 年第 34 卷第 9 期
Vol. 34, No. 9, 2012

文章编号:1007-130X(2012)09-0077-06

恐怖袭击事件不确定性度量及可视分析*

Uncertainty Measurement and Visual Analysis on Terroristic Attacks Data

贺怀清,王　赫

HE Huai-qing, WANG He

(中国民航大学计算机科学与技术学院,天津 300300)

(College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China)

摘　要:近年来,全球范围内恐怖主义活动愈发频繁,已经严重影响了地区稳定和世界和平。随着信息技术的发展,研究者们得以从多个方面获取恐怖袭击事件信息。然而,随着数据集规模的不断扩大,如何从大量数据中发掘隐含的信息、分析其中包含的不确定性,成为恐怖袭击事件分析过程中的重要问题。针对全球恐怖主义数据库,基于可视分析和不确定度量理论,提出了数据记录和属性不确定性的度量及可视分析方法。通过将不确定性度量结果与平行坐标、柱状图、面积图和交互式方法相结合,在不影响数据源表达的同时清晰地展示了其中包含的不确定性,为下一步基于不确定性理论的态势评估提供了信息基础。

Abstract: In recent years, terroristic activities occur more frequently and have seriously affected the regional stability and the world peace. With the development of information technology, the researchers are able to obtain information of terroristic attacks from many aspects. However, with the constant enlargement of the scale of data sets, how to explore the underlying information and analyze the uncertainty from a large number of data has become an important issue in the analysis process of terroristic attacks. On Global Terrorism Database, based on visual analysis and uncertainty measurement theory, we propose the measurement and visual analysis methods on data records and uncertainty of attributes. By integrating results of uncertainty measurement with parallel coordinates, histogram, area chart and interactive methods, the data uncertainty is clearly displayed without influence on its representation and provides information base for situation assessment based on uncertainty theory the next step.

关键词:恐怖袭击事件;不确定性度量;可视分析;平行坐标;柱状图;面积图

Key words: terroristic attacks data; uncertainty measurement; visual analysis; parallel coordinates; histogram; area chart

doi:10.3969/j.issn.1007-130X.2012.09.014

中图分类号:TP391　　**文献标识码**:A

1　引言

自 20 世纪 90 年代以来,世界恐怖活动日益频繁,以 9·11 为代表的一系列恐怖袭击事件的发生对全球各国的社会稳定、人民的生命财产安全以及社会经济的发展造成了巨大的影响和冲击[1]。目前,恐怖主义已成为影响世界稳定和地区安全的首

* **收稿日期**:2011-08-31;**修订日期**:2011-12-09
基金项目:中央高校基本科研业务费中国民航大学专项(ZXH2009C001);天津市应用基础及前沿技术研究计划资助项目(10JCYBJC00900);国家自然科学基金资助项目(60879003)
通讯地址:300300 天津市中国民航大学计算机科学与技术学院
Address: College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, P. R. China

要威胁,引起世界各国的高度重视。然而,随着恐怖袭击事件数据记录规模的日益庞大,如何有效地理解并发掘事件中包含的隐藏信息成为需要认真考虑的问题。由认知心理学可知,人类获取的知识中83.3%来源视觉,因此利用以计算机图形学为基础的信息可视化分析方法受到越来越多的关注。

Joonghoon提出的全球恐怖主义数据库(Global Terrorism Database,简称GTD) Explore根据恐怖袭击发生的国家、地区、袭击类型等分类属性,使用面积图对1970年~2008年发生的恐怖袭击数量进行了分类可视分析,并提供了良好的交互方法,使得用户可根据所关心的内容分析恐怖袭击随时间的变化趋势[2]。Yang等利用社会网络可视化和数据挖掘方法,引入相关地理信息,以恐怖袭击事件各要素节点之间的关系作为基本分析单位,对恐怖组织之间的活动模式和发展特点等内在规律进行挖掘与解释[3]。文献[4]以5W为基本思想提出一种可视分析系统,通过该系统,用户可根据时间信息(When)、地理空间特征(Where)、恐怖组织间的相互关系(Who)和不同的袭击方式(What)高效地检索恐怖袭击事件信息并发掘袭击事件的原因(Why)。Xiao等提出改进的Parallel Sets方法用于国际恐怖主义数据分析,该方法采用带降势的启发式分类值布局算法,能够自动优化分类值布局顺序,减轻视图中的可视混乱,并且减少了参与计算的分类值数目,适合于分析数据量大、分类值较多的数据集[5]。

以上文献分别从不同角度对恐怖袭击数据进行了有效分析并提供了较好的交互方法。但是,我们注意到恐怖袭击作为一种突发事件,最基本的特征在于其内容的不确定性,而不确定性程度的高低在很大程度上决定了今后事件态势的发展变化。因此如何表达数据本身及其包含的不确定性,成为恐怖袭击数据分析中的重要问题。本文以多维数据不确定可视化方法为基础,对恐怖袭击突发事件数据及其不确定性进行了表示和分析,为今后突发事件态势可视评估的研究打下一定基础。

2 多维数据不确定性度量方法

不同于一元和二元数据,恐怖袭击数据信息通常包含若干个属性。因此,为了充分描述恐怖袭击中包含的不确定信息,必须从构成恐怖袭击数据记录的基本要素入手分析、度量其中包含的不确定性[6],主要包括:构成数据记录的单个数据项的不确定性(Data Uncertainty)、数据记录的不确定性(Record Uncertainty)和属性的不确定性(Dimension Uncertainty)。相应地,我们采用距离误差来度量数据项和数据记录的不确定性;对于属性不确定性,引入信息论中熵-知识粒度的方法。下面对几种不确定性度量方法做简要介绍。

2.1 距离误差方法

(1)数据项不确定性计算:在某个属性下数据项A_{ij}的不确定性V_{ij}可由A_{ij}与该属性数据平均值M_j之间差值的绝对值来表示。计算方法如公式(1)所示:

$$M_j = \sum_{i=1}^{n} A_{ij}/n,\quad V_{ij} = |A_{ij} - M_j| \tag{1}$$

其中,n是数据表的行数,i是数据项的行下标,j是数据项的列下标。

(2)记录不确定性的计算:由数据项不确定性V_{ij}可得到记录的不确定性R_i。计算方法如公式(2)所示:

$$R_i = \sum_{j=1}^{n} V_{ij}/m \tag{2}$$

其中,m是数据表的列数,i是数据项的行下标,j是数据项的列下标。

2.2 熵和知识粒度

熵概念[7]是热力学中描述热过程方向的物理量,用来描述热量的退化或进化。熵是系统无序性的度量,熵越大,无序性越高。Shannon将物理学中的熵概念引入到了信息论中,用来度量系统结构的不确定性,认为信息源提供的信息量越大,其不确定性也就越大。由文献[8]知道,不确定性的大小直接依赖于信息量的大小;而Shannon熵、信息熵和知识粒度则分别从不同的角度描述了信息源的信息量,因此也就反映了信息源的不确定性大小。相关概念如下:

(1)Shannon熵。

设U是论域,$X=\{X_1,X_2,\cdots,X_n\}$是U的一个划分,其上有概率分布$p_i = p(X_i)$,则称:

$$H(X) = -\sum_{i=1}^{n} p_i \mathrm{lb} p_i$$

为信息源X的信息熵。当某个p_i为零时,理解为$0\cdot\mathrm{lb}0=0$且$0\leqslant H(X)\leqslant \mathrm{lb}n$。

(2)信息熵(Information Entropy)。

设$S=\{U,A\}$是一个信息系统,U是论域,A是属性集合,$U/A=\{R_1,R_2,\cdots,R_m\}$。利用信息函数具有补的性质,提出信息系统的一种新的信息

熵定义为:

$$E(A)=\sum_{i=1}^{m}\frac{|R_i|}{|U|}\frac{|R_i^c|}{|U|}=\sum_{i=1}^{m}\frac{|R_i|}{|U|}\left(1-\frac{|R_i|}{|U|}\right)$$

其中,R_i^c 表示 R_i 的补集,即 $R_i^c=U-R_i$,$|R_i|/|U|$ 表示 R_i 在论域 U 中的概率,$|R_i^c|/|U|$ 表示 R_i 的补集在 U 中的概率。信息熵 E 随着信息粒度的变小(通过更细的划分)单调增加。当为最小划分时,E 达到最大 $1-1/|U|$;当为最大划分时,E 达到最小值 0。

(3)知识粒度(Knowledge Granularity)。

设 $S=\{U,A\}$ 是一个信息系统,U 是论域,A 是属性集合,$U/A=\{R_1,R_2,\cdots,R_m\}$,则 A 的知识粒度为:

$$GK(A)=\frac{1}{|U|^2}\sum_{i=1}^{m}|R_i|^2$$

其中,$\sum_{i=1}^{m}|R_i|^2$ 是由 $\bigcup_{i=1}^{m}(R_i\times R_i)$ 决定的等价关系中元素的个数。当为最小划分时,A 的粒度达到最小值 $1/|U|$;当为最大划分时,A 的粒度达到最大值 1。知识粒度 $GK(A)$ 和信息熵 $E(A)$ 满足关系:$E(A)+GK(A)=1$。

3 恐怖袭击事件数据集介绍

现有的恐怖袭击事件数据集的形式大多不统一,没有形成一个良好的数据共享环境,为分析恐怖袭击的规律、特征、发展趋势及不确定性带来了困难。美国马里兰大学提供的全球恐怖主义数据库详细记录了 1970~2008 年间近 9 万起恐怖袭击事件,涵盖了丰富的事件信息,主要包括:事件记录号(GTD ID)、发生时间(Incident Date)、事件发生的地理位置(Incident Location)、事件特征信息(Incident Information)、袭击信息(Attack Information)、目标信息(Target Information)、恐怖组织信息(Perpetrator Information)、恐怖组织信息统计(Perpetrator Statistics)、是否有恐怖组织声称对袭击负责(Perpetrator Claim of Responsibility)、武器信息(Weapon Information)、伤亡(受害者)信息(Casualty Information)、财产损失情况(Consequences)、恐怖袭击中的人质/绑匪信息(Hostage / Kidnapping Additional Information)、附加信息(Additional Information)和事件详细信息注释(Source Information)共 15 个方面的统计信息,包含了 123 个具体属性。因此,本文采用 GTD 作为实验源数据集,从中提取重点关注的各个属性数据。

本文主要从恐怖袭击发生的时间、地域性、袭击方式、针对目标、危害程度共 5 个方面进行分析。因此,我们取这五个方面所包含的 10 个主要属性,分别是年(Year)、月(Month)、日(Day)、国家(Country)、地区(Region)、袭击类型(Attack)、袭击目标(Targ)、袭击目标国籍(Natlty)、武器类型(Weap)和伤亡人数(Nkill)。其中,除了年、月、日属性直接使用整型值(int)表示外,其余 7 个属性值采用 int 型和相应文本描述两种形式表示。

4 恐怖袭击数据属性不确定性的度量方法

对于 2.2 节(1)中的概率分布 p_i,若其中某个 p_i 为 1,则每次结果 X_i 一定出现,也就是说结果是完全确定的。这样,信息源 X 并没有提供任何信息量,这种情况下信息源的不确定性量应该为 0。若 $p_1=p_2=\cdots=p_n=1/n$,则哪个结果 X_i 会出现是最不确定的,这时信息源能提供最大的信息量,信息源的不确定性量应该是最大的。

根据以上理解,本文首先使用 Shannon 熵作为恐怖袭击数据源的不确定性度量。同时,为了利用信息函数的补的性质,利用信息熵和知识粒度进一步说明某属性下取值的分布情况,通过属性值分布的集中与否表达该属性的不确定性。这里仅以 2008 年数据为例,结果如表 1 所示,根据表中的结果,下面的各节中我们将使用白色、浅灰色和深灰色的柱状图分别表达 Shannon 熵、信息熵和知识粒度。

5 恐怖袭击数据不确定性的可视分析

5.1 2D 平行坐标表示法

平行坐标是多维数据的一种 2D 表示方法,该

表 1 2008 年恐怖袭击数据属性不确定度量

	Year	Month	Day	Country	Region	Attack	Targ	Natlty	Weap	Nkill
Shannon 熵	0.000 0	3.516 1	4.945 6	4.213 7	2.430 9	1.824 5	3.372 5	4.303 9	1.712 7	2.285 5
信息熵	0.000 0	0.909 5	0.967 1	0.899 4	0.751 0	0.629 5	0.852 1	0.909 6	0.597 5	0.646 4
知识粒度	1.000 0	0.090 5	0.032 9	0.100 6	0.249 0	0.370 5	0.147 9	0.090 4	0.402 5	0.353 9

方法最早由 Inserberg 等人提出。高维空间点集中的一个点被表示为一条连接 N 条竖直、平行且等距坐标轴的折线，折线与竖直坐标轴的交点表示该折线在该属性上的取值。平行坐标有良好的数学基础，是信息可视化的一种重要方法，克服了传统笛卡尔坐标系空间容易耗尽、难以表达三维以上数据的问题。由表 1 可知，对于 2008 年数据其 Year 属性对应的熵为 0、知识粒度为 1，相应其属性不确定性为 0，因此在以后的可视分析中，将不再显示 Year 属性的相关信息，重点关注其余 9 个属性。图 1 是 2008 年恐怖袭击数据不确定性的平行坐标表示。

如前所述，在 2D 平行坐标表示中，除了利用平行坐标显示源数据外，分别采用颜色和柱状图表示数据记录和属性的不确定性。对于数据记录的不确定性[6]，根据距离误差度量方法的结果，首先对其进行归一化处理，使得记录的不确定性取值在[0,1]区间。这里，将不确定划分为 5 个区间，■■■■■五种颜色分别表示不确定性区间(0～0.2,0.2～0.4,0.4～0.6,0.6～0.8,0.8～1.0)。可以看出，数据记录在最右边的 un_record 轴(最右边深色坐标轴)上分布较均匀，反映了数据源整体的不确定程度较大。

属性不确定性首先通过 Shannon 熵(黄色柱状图)进行表示，Shannon 熵反映了信息源中某属性取各个属性值的可能性，其值越大则取各属性值的概率分布越接近，属性的不确定性就越大。例如，从图中可以看到 Day 属性(左数第 2 个轴)对应的 Shannon 熵最大，则意味着恐怖袭击可能发生在一个月任何一天，不确定性较大；而 Country 属性(左数第 3 个轴)的 Shannon 熵相对较小，表示了恐怖袭击可能集中发生在某个或某些国家，不确定性相对较小。而通过信息熵(浅灰柱状图)和知识粒度(深灰色柱状图)同样可以清晰地反映属性的不确定性，例如，对于 AttackType 属性(左数第 5 个轴)，其信息熵较小而知识粒度较大，说明了大多恐怖袭击可能都采取某种比较固定袭击方式，如炸弹袭击。

5.2 交互方法

图 1 所使用的 2D 平行坐标及柱状图虽然能够有效表达大规模多维数据集及其记录和属性不确定性，但是随着数据量的增加，即使采用了不同颜色表达数据记录的不确定性，仍然会造成不同不确定性区间内的数据记录在相邻属性轴之间的遮盖现象。为了更加清晰地表示和观察恐怖袭击数据，引入隐藏(Hiding)交互方法，只显示用户选择的特定不确定性区间内的数据记录，使显示结果更加直观。如图 2 所示，显示不确定范围在 0.2～0.4区间的数据记录。

5.3 3D 平行坐标表示法

在 2D 平行坐标表示法中，对应每个属性坐标轴，我们可以观察到数据记录在各属性轴上的取值情况、数据记录的不确定性和属性不确定性；也可观察到任意两个相邻属性轴间的关系。但同时，我们也可以看到 2D 平行坐标本身无法表达任意一个属性与其他属性之间的关系。为此，本文引入 3D 平行坐标[9]的方法，如图 3 所示。

在 3D 平行坐标表示法中，以 Month 属性为中心轴，其余 8 个属性以 Month 轴为中心环绕排布。对应每一个属性，依次由 month 轴(中心轴)、属性轴(黄色轴)、记录不确定轴(绿色轴)组成一个平行坐标平面，且在不确定轴的旁边显示基于熵-知识粒度的属性不确定性度量。图 3 中，除了用■■■■■表示记录不确定性外，对源数据也采用不同颜色进

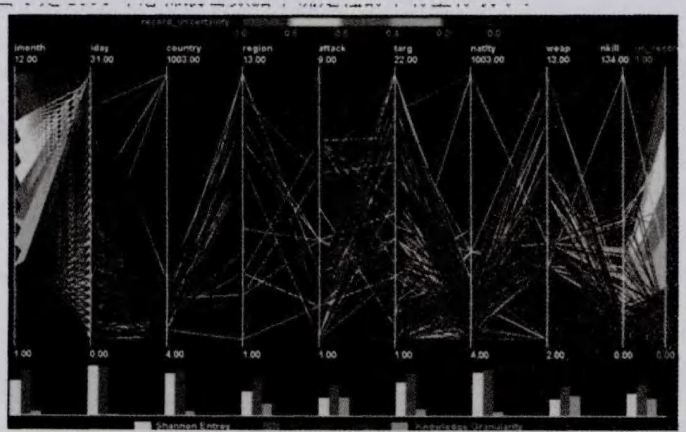

图 1 恐怖袭击及其不确定性的平行坐标表示

行表示。用户可以通过四个方向键在水平和垂直方向旋转查看 3D 平行坐标图的各个坐标平面,进一步增加方法的交互性。通过扩展的 3D 平行坐标表示方法,我们可以查看任意一个属性与其他属性之间的关系。

5.4 面积图表示法

根据 5.2 节和 5.3 节的 2D 和 3D 平行坐标表示方法,我们可以从整体上观察源数据信息和属性的不确定性;通过使用 5 种不同颜色及 Hiding 方法,可以根据不同需要查看不同不确定范围内的数据记录。由信息系统的不可区分关系 $IND(P)$(P 为信息系统的属性子集)可知,当某些数据记录在 P 下有相同的属性值时,根据等价类划分思想,在平行坐标中这些记录在 P 包含的各属性轴上会相交于一点。当观察一个具体的属性时,我们无法获知记录取某个属性值的可能性的大小;而且随着数据规模的增大,折线间的互相干扰会进一步影响信息的获取。因此,引入面积图表示法[2]以更加清晰地描述某个属性下的属性值不确定性分布;同时,通过面积图也可以反映一年当中恐怖袭击发生次数随时间的变化趋势,如图 4 所示。

图 4 以 country 属性为划分依据,首先统计每个国家恐怖袭击发生的次数;然后以 Month 属性为横轴,袭击次数统计为纵轴,按照 Country 属性值由大到小依次绘制各个国家的面积图。

从累积面积图中可以看出,某些子图的面积要明显大于其余子图。如颜色■表示的面积图代表的是国家伊拉克,说明一年当中伊拉克发生恐怖袭击的次数最多(概率/可能性最大)。因此,可得到这样的结论:对于某个属性,若其取值集中于某个或某几个属性值(即这些属性值对应的概率分布明显较大;相应地,这些属性值对应的等价类划分子集的基数也明显较大),则其对应的 Shannon 熵、信息熵较小,从而知识粒度较大,信息源提供信息量较少,属性不确定性较低。通过面积图的方法,一定程度上弥补了单纯依靠平行坐标和柱状图表示数据记录和属性不确定性中存在的不足。另外,由于面积图的累积效果,面积图最上面的边界也反映了恐怖袭击发生次数随时间的变化趋势,这都是传统 2D 和 3D 平行坐标所无法提供的信息。

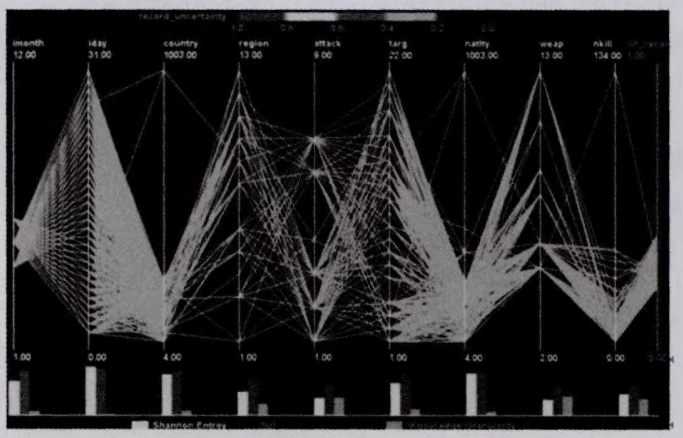

图 2 使用 Hiding 的恐怖袭击数据不确定性可视化方法

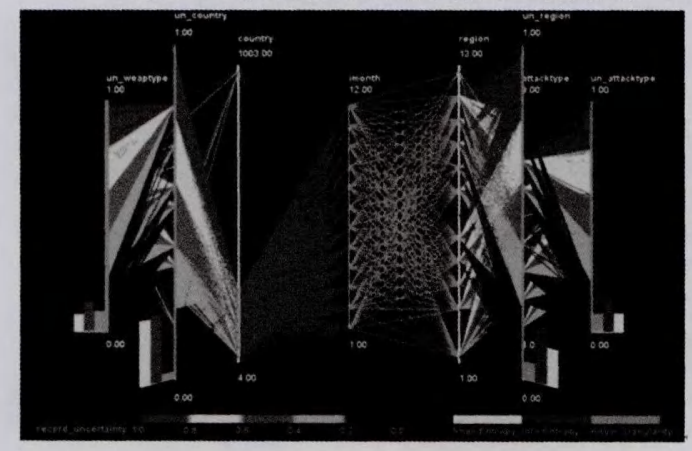

图 3 恐怖袭击数据不确定性的 3D 平行坐标表示法

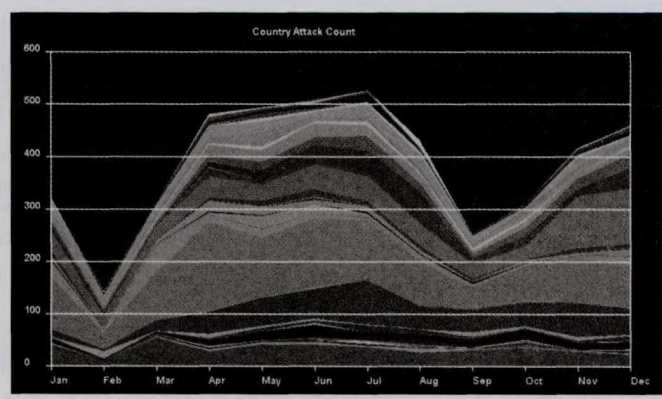

图 4 恐怖袭击发生次数的 Country 面积图表示法

6 结束语

已有的许多信息可视化文献主要关注于数据的研究(Data Exploration)和交互式操作(Interaction),并没有明确地提出信息源不确定的度量与表达。本文在已有方法的基础上,对于信息源中数据记录和属性不确定性进行度量和可视表达。通过将 2D 和 3D 平行坐标与记录不确定性的距离误差度量方法、属性不确定性的熵、知识粒度度量方法进行结合,在表达数据源本身的同时,清晰地展示了信息源的不确定性,并提供一定的交互方法;为了弥补属性值的等价类划分对平行坐标表示产生的影响,文中引用了面积图的表示方法,进一步说明了属性不确定性的意义。

下一步工作的重点是以不确定性度量和可视分析结果为基础,结合贝叶斯网络等传统态势评估方法,提出一种可视的、交互性好的恐怖袭击突发事件态势评估方法。

参考文献:

[1] Li Jian-he, Wang Cun-kui, Mei Jian-ming, et al. The Character and New Tendency of the Contemporary Terrorism[J]. Journal of Chinese People's Public Security University(Social Sciences Edition), 2008, 24(3): 1-7.

[2] Joonghoon Lee. Exploring Global Terrorism Data: A Web-based Visualization of Temporal Data[J]. Crossroads, 2008, 15(2): 7-14.

[3] Yang Yu-bin, Li Ning, Zhang Yao. Networked Data Mining Based on Social Network Visualization[J]. Journal of Software, 2008, 19(8): 1980-1994.

[4] Wang X Y, Miller E, Smarick K, et al. Investigative Visual Analysis of Global Terrorism[J]. Computer Graphics Forum, 2008, 27(3): 919-626.

[5] Xiao Wei-dong, Zhou Cheng, Sun Yang, et al. Improvement of Parallel Sets and Its Application in Analyzing Global Terrorism Database[J]. Journal of National University of Defense Technology, 2011, 33(1): 115-119.

[6] Xie Zai-xian, Huang Shi-ping, Ward M O, et al. Exploratory Visualization of Multivariate Data with Variable Quality[C]// Visual Analytics Science and Technology, 2006 IEEE Symposium on Digital Object Identifier, 2006: 183-190.

[7] 梁吉业,李德玉. 信息系统中的不确定性与知识获取[M]. 北京:科学出版社,2005.

[8] 张文修,梁怡,徐萍. 基于包含度的不确定推理[M]. 北京:清华大学出版社,2007.

[9] Forsell C, Johansson J. Task-Based Evaluation of Multi-Relational 3D and Standard 2D Parallel Coordinates[C]// Proc of SPIE-The International Society for Optical Engineering, 2007, 6495(64950C): 1-12.

贺怀清(1969 -),女,吉林白山人,博士,教授,CCF 会员(13105S),研究方向为图形图像与可视化。**E-mail**: huaiqinghe@yahoo.com.cn

HE Huai-qing, born in 1969, PhD, professor, CCF member (13105S), her research interests include graphics, image and visualization.

王赫(1986 -),男,辽宁大连人,硕士生,CCF 会员(17421G),研究方向为图形图像与可视化。**E-mail**: wanghemit@163.com

WANG He, born in 1986, MS candidate, CCF member (17421G), his research interests include graphics, image and visualization.