文章编号：0253-374X(2012)05-0707-04 DOI:10.3969/j.issn.0253-374x.2012.05.010

交通事故案例检索优化方法

董宪元，方守恩，王俊骅
(同济大学 道路与交通工程教育部重点实验室，上海 201804)

摘要：为提高交通事故条件下实时决策管理的准确性，提出交通事故案例检索优化方法.基于交通事故案例数值型和枚举型数据构成特征，采用信息熵方法评价数据离散程度，客观确定交通事故案例属性权重，在数据归类化处理的基础上，应用二阶聚类算法建立案例检索库，开发交通事故案例检索系统并进行案例检索试验，从案例最高相似度及案例集匹配度两方面评价案例检索精度，验证了该案例检索优化方法的有效性.

关键词：交通事故；案例属性权重；案例检索库；案例集匹配度；案例检索
中图分类号：U491.31 **文献标识码**：A

Case Retrieval Optimization Method for Traffic Accidents

DONG Xianyuan, *FANG Shouen*, *WANG Junhua*
(Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract: An optimization method for traffic accident case retrieval is proposed to improve the accuracy of real-time decision management under traffic accident conditions. Based on traffic accident data constitution of Numeric and enumeration type, attribute weights are determined objectively through information entropy method evaluating the data dispersion degree. On the basis of data classification, case retrieval base is established through second-order clustering algorithm. And a traffic accident case retrieval system is developed to have a series of tests. Then, case retrieval tests are conducted to verify effectiveness of this optimization method by analyzing both highest similarity and case set matching degree of the retrieval conclusions.

Key words: traffic accident; case attribute weight; case retrieval base; case set matching degree; case retrieval

交通事故决策管理涉及交通管制、交通救援、道路清障等不同作业需求，涉及部门广，影响因素复杂，实际仍处于半结构化管理阶段.案例推理(case-based reasoning, CBR)是一种人们利用积累的知识和经验解决当前问题的类比学习方法，适用于解决非结构化或半结构化问题，具有良好的可扩充性、可移植性以及自学习能力[1].案例检索是案例推理过程的核心，当前研究较多的案例检索算法有K-NN(k nearest neighbor)算法、决策树法、神经网络算法、支持向量机等[2].其中，作为最广泛使用的案例检索算法，K-NN算法应用简便灵活，且随着计算机技术的提高，响应迅速，保证了检索效率，尤其适用于基于属性表示的交通事故案例检索.美国佛蒙特大学Sadek等[3]将案例推理方法引入到交通诱导领域，案例各属性权重值均界定为1，采用K-NN算法检索交通流路径分流方案；荷兰代尔夫特科技大学Hoogendoorn等[4]采用案例属性数据平均模糊隶属度作为权重值计算案例相似度，实现实时路网交通安全辅助管理；加拿大不列颠哥伦比亚大学Lin等[5]提出了基于案例推理的道路安全改善方法效益评估系统.国内，华南理工大学翁小雄等[6]采用基于案例推理的动态参数匹配和协调冲突检验方法，寻求建立交叉口信号控制优化方案；西南交通大学戢晓峰[7]采用层次分析法确定案例属性权重，应用K-NN算法检索相似的决策案例进行交通拥挤管理；东南大学杨顺新[8]根据专家赋值法确定案例属性权重，应用K-NN案例检索算法查询制定高速公路交通事件应急管理措施.但是，应用K-NN算法进行交通事故案例检索仍存在以下问题：① 交通事故案例数据由数值型和枚举型数据组成，案例属性权重一般采用主观方法定性确定，检索精度及稳定性较差；② 以往仅仅以案例最高相似度为指标评价检索精度的方法过于片面.

收稿日期：2011-07-06
基金项目："十一五"国家科技支撑计划(2009BAG13A06)
第一作者：董宪元(1984—)，男，博士生，主要研究方向为道路交通安全. E-mail:dxygnn@163.com
通讯作者：方守恩(1961—)，男，教授，工学博士，博士生导师.主要研究方向为道路交通安全. E-mail:fangsek@tongji.edu.cn

本文基于道路交通事故由数值型和枚举型数据构成的特点,分析道路交通事故数据频率分布特征,采用信息熵方法评价数据离散程度,客观确定交通事故案例属性权重,并应用二阶聚类算法建立案例检索库,提高案例检索精度;以 Matlab R2008a 为仿真工具开发交通事故案例检索系统,以沪杭高速公路交通事故数据为例,采用 K-NN 案例检索算法进行案例检索试验,结合案例最高相似度和案例集匹配度综合评价案例检索精度,验证该案例检索优化方法的有效性.

1 案例属性权重设定

1.1 基于数据离散程度的案例属性权重计算

案例属性权重准确设置是实现案例检索目标的关键.本文基于交通事故案例属性数据频率分布特征,通过数据信息熵分析案例各属性数据离散程度,客观确定案例属性权重值.

假设系统可能处于多种不同的状态,而每种状态出现的概率为 $P_i(i=1,2,\cdots,m)$,则系统的信息熵定义为[9]

$$H=-k\sum_{i=1}^{m}p_i\ln p_i \tag{1}$$

式中,取 $k=1/\ln m$.

交通事故案例属性 j 的信息熵为

$$H_j=-k\sum_{i=1}^{m}p_i\ln p_i \tag{2}$$

信息熵确定的值域为[0,1],可直接用于随机变量离散程度的对比分析,信息熵的值越接近于 0,离散程度越小,检索所需权重值越小;信息熵越接近于 1,离散程度越大,检索所需权重值越大.交通事故案例属性 j 的权重定义为

$$w_j=\frac{H_j}{\sum_{j=1}^{n}H_j} \tag{3}$$

1.2 沪杭高速公路交通事故案例属性权重

以沪杭高速公路 2005 年~2008 年 3542 条交通事故案例数据为分析对象,对交通事故的发生位置、发生日期、发生时段、事故形态、严重程度和天气状况数据的频率分布进行分析,通过式(2)计算数据信息熵,评价案例各属性数据离散程度.基于式(3)计算案例各属性权重值见表 1.

表 1 交通事故案例属性权重

Tab.1 Attributes weights of the traffic accident case

案例属性	事故发生位置	事故发生日期	事故发生时段	事故形态	事故严重程度	天气状况
离散度	0.809 561	0.998 445	0.826 354	0.597 972	0.179 309	0.459 111
权重值	0.209 148	0.257 946	0.213 487	0.154 485	0.046 324	0.118 610

2 交通事故案例检索库

基于 K-NN 算法进行案例检索时,因事故形态、事故严重程度及天气状况三种属性数据权重值过小,检索精度相对较低.现采用二阶聚类算法在原沪杭高速公路交通事故案例库的基础上建立案例检索库,综合提高案例各属性数据检索准确度.为兼顾检索效率与检索精度,有效进行二阶聚类,需同时对数值型事故位置数据以及权重值偏小的以上三类属性数据进行归类化处理,建立案例检索库,提高案例检索精度和效率.

2.1 数据归类化处理

基于数据频率分布特征,采用累计频率法将事故发生位置划分为事故多发位置和事故偶发位置[10],同时,以数据频率分布均衡化为目标,对交通事故形态、交通事故严重程度、天气状况数据进行归类化处理,数据归类化处理结果分别见表 2~表 5.

表 2 交通事故位置数据归类化处理

Tab.2 Data classification of the traffic accident location

交通事故属性	数据归类化	位置桩号
事故发生位置	事故多发位置	K108;K120;K116;K110;K114;K135;K105;K117;K75;K76;K138;K109;K99
	事故偶发位置	其他

表 3 交通事故形态数据归类化处理

Tab.3 Data classification of the traffic accident form

交通事故属性	事故形态							
事故形态	车辆相撞	车辆与道路设施相撞	翻车	刮擦	碾压	失火	坠车	其他
数据归类化	车辆相撞	车辆与道路设施相撞	其他					

表 4 交通事故严重程度数据归类化处理

Tab.4 Data classification of the traffic accident severity

交通事故属性	事故严重程度		
事故严重程度	财损	伤人	死亡
数据归类化	财损	死伤	

表 5　天气状况数据归类化处理

Tab.5　Data classification of the weather condition

交通事故属性	天气状况					
天气状况	晴	大风	雾	雪	阴	雨
数据归类化	晴	异常				

2.2　建立交通事故案例检索库

基于沪杭高速公路交通事故案例库，采用二阶聚类算法建立案例检索库. 现将原交通事故案例库聚为 9 类组成案例检索库，聚类效果良好，如图 1～图 2所示. 图1，图2是根据数据分析软件SPSS的分析结果.

分群	1	2	3	4	5	6	7	8	9
大小	14.0% (496)	5.4% (190)	2.3% (753)	4.0% (143)	8.9% (317)	8.9% (316)	7.7% (274)	15.0% (531)	14.7% (522)
特征	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 多发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)	发生位置事故 偶发点 (100.0%)
	时段 高峰 (100.0%)	时段 高峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)	时段 低峰 (100.0%)
	事故形态车辆 相撞 (53.4%)	事故形态车辆 相撞 (60.5%)	事故形态车辆 相撞 (60.8%)	事故形态车辆 相撞 (72.0%)	事故形态车辆 与道路设施 相撞 (55.2%)	事故形态车辆 相撞 (55.2%)	事故形态车辆 其它 (100.0%)	事故形态车辆 相撞 (100.0%)	事故形态车辆 相撞 (55.9%)
	天气 晴 (100.0%)	天气 异常 (100.0%)	天气 晴 (74.4%)	天气 晴 (82.5%)	天气 晴 (100.0%)	天气 晴 (100.0%)	天气 晴 (100.0%)	天气 晴 (100.0%)	天气 异常 (100.0%)
	严重程度 财损 (100.0%)	严重程度 财损 (100.0%)	严重程度 财损 (100.0%)	严重程度 死伤 (100.0%)	严重程度 财损 (100.0%)	严重程度 财损 (100.0%)	天气 严重程度 财损 (100.0%)	严重程度 财损 (100.0%)	严重程度 财损 (100.0%)
	日期 星期一 (17.3%)	日期 星期五 (16.3%)	日期 星期日 (18.3%)	日期 星期四 (21.7%)	日期 星期日 (35.6%)	日期 星期一 (51.9%)	日期 星期五 (19.3%)	日期 星期六 (24.1%)	日期 星期一 (18.2%)

图 1　交通事故案例检索库结构

Fig.1　Structure of traffic accident case retrieval base

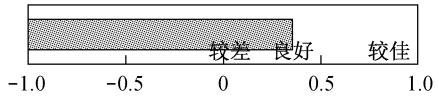

图 2　交通事故案例聚类效果

Fig.2　Clustering effect of traffic accidents case

3　案例检索

K-NN 检索算法是对案例库中每个案例的属性相似度进行加权求和计算案例综合相似度，并检索形成具有一定数目案例的案例集. 决策者需要从最高相似案例以及案例集整体匹配性两方面综合考虑，制定决策措施.

3.1　案例相似度计算

数值型属性相似度计算：设案例 C_i 与案例 C_j 的第 m 个属性为数值属性，其属性值分别记为 V_{im} 和 V_{jm}. 对属性值进行归一化处理，使得 $V_{im}, V_{jm} \in [0,1]$，则案例 C_i 与案例 C_j 第 m 个属性的相似度为

$$\mathrm{sim}(C_{im}, C_{jm}) = 1 - D(C_{im}, C_{jm}) = 1 - |V_{im} - V_{jm}| \tag{4}$$

式中：$\mathrm{sim}(\cdot)$ 为相似函数；$D(\cdot)$ 为属性值之间距离.

枚举型属性相似度计算：枚举型属性相似度的计算比较简单，只要两个案例的属性值相同，则两者的相似度值为 1，否则为 0

$$\mathrm{sim}(C_{im}, C_{jm}) = \begin{cases} 1, & C_{im} = C_{jm} \\ 0, & C_{im} \neq C_{jm} \end{cases} \tag{5}$$

在计算出案例各属性的相似度之后，可以计算两个案例之间的综合相似度为[7]

$$\mathrm{sim}(C_i, C_j) = \sum_{m=1}^{n} w_m \mathrm{sim}(C_{im}, C_{jm}) \tag{6}$$

式中，w_m 为案例第 m 个属性的权重值. 交通事故案例各属性权重值见表 1，n 表示案例属性数量，$\sum_{m=1}^{n} w_m = 1$.

根据式(4)～式(6)对案例检索库中每类案例分别进行案例相似度计算，依据案例相似度的大小及每类案例数量选取出不同的案例数，一般取 $[\sqrt{k_l}]$ 条案例[11]，则共选取 $k = \sum_{l=1}^{9} \sqrt{k_l}$ 条案例形成案例集. 其中，k_l 表示案例检索库第 l 类案例数量，k 表示案例集所包含的案例数量.

3.2　案例集匹配度

当检索出的最高相似案例与目标案例并不完全相似（$\mathrm{sim}_{\max}(C_i, C_j) \neq 1$）时，决策者需要对检索得到的案例集进行综合评估，以期制定完善的交通事故管理决策措施. 本文提出案例集匹配度的概念，结合

案例最高相似度综合评价案例检索效果.

“案例集匹配度”定义为“案例集中的案例数据能够与目标案例各属性值相匹配的程度”.由 k 条案例所组成的案例集中,若案例第 m 个属性为枚举型,此该属性匹配度为

$$p_m=\begin{cases} w'_m, & \sum_{t=1}^{k}\mathrm{sim}_t(C_{im},C_{jm})\geqslant 1 \\ 0, & \sum_{t=1}^{k}\mathrm{sim}_t(C_{im},C_{jm})=0 \end{cases} \tag{7}$$

若案例第 m 个属性为数值型,则此属性匹配度为

$$p_m=w'_m\max(\mathrm{sim}_t(C_{im},C_{jm})) \tag{8}$$

其中 $w'_m=1/n, m=1,2,\cdots,n$,则案例集匹配度为

$$p=\sum_{m=1}^{n}p_m \tag{9}$$

4 试验分析

以 Matlab R2008a 为仿真工具开发交通事故案例检索 GUI(graphical user interface)系统(图 3).计算机配置为 Geenuine Intel(R)处理器,0.99 G 内存,分别从案例最高相似度和案例集匹配度两方面评价案例检索精度,制定案例决策措施.试验过程中记录了检索耗时,验证案例检索的时效性.

图 3 交通事故案例检索系统

Fig.3 Traffic accident case retrieval system

仿真系统随机产生 40 条试验案例,分别在原交通事故案例库和案例检索库上进行案例检索,检索效果如图 4 和图 5 所示.

定义参考案例相似度最小阈值为 0.8.从图 4 可以看出,在 40 次随机案例检索试验中,共有 37 次检索案例最高相似度在 0.8 及其以上,检索精度为 92.5%,案例属性权重设置基本合理.检索平均耗时 0.39 s,满足实时决策的需要.然而,案例 T5,T24 及 T26 最高相似度分别仅为 0.678 8、0.652 6 和 0.788 7,且案例集匹配度较小,显然不能满足在目标案例条件下进行决策选择的需要.

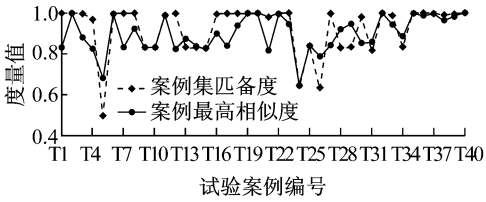

图 4 直接对案例库进行检索效果分析

Fig.4 Case-based analysis of case retrieval base

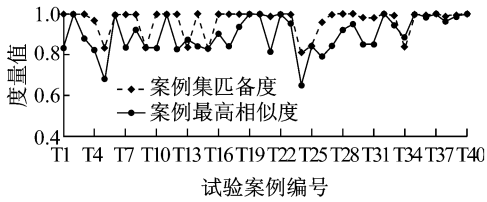

图 5 对案例检索库进行检索效果分析

Fig.5 Case retrieval-based analysis of case retrieval base

从图 5 可知,通过对案例检索库的分类检索,案例集匹配度都有了不同程度的提高,尤其对案例 T5,T24 及 T26 进行检索时,案例集匹配度分别增长了 67.1%,25.9%和 53.6%,弥补了检索案例最高相似度值过小的问题.决策者可以通过对案例检索集的综合评估,制定更加科学、合理的管理措施,保证了案例推理的有效性.检索试验平均耗时0.40 s,同样满足交通事故条件下实时决策管理的需要.

5 结论

本文基于交通事故案例数值型和枚举型数据并存的构成特征,提出采用信息熵方法评价数据离散程度,客观确定交通事故案例属性权重;在数据归类化处理的基础上,应用二阶聚类算法建立了案例检索库,并进行了 K-NN 案例检索试验;提出了从案例最高相似度和案例集匹配度两个方面进行案例检索精度评价的方法.试验分析证明,在客观确定交通事故案例属性权重并建立案例检索库的基础上进行 K-NN 案例检索,案例检索精度提高,弥补了以往仅仅以最高相似案例辅助决策的弊端,优化了案例检索效果,且检索试验平均耗时仅约 0.4 s,符合道路交通事故实时决策管理的需要,验证了该交通事故案例检索优化方法的有效性.

(下转第 750 页)