文章编号:1005-9679(2018)05-0055-06

基于本体关系的群体事件网络关注度影响因素数据挖掘研究

巩晓敏[1]　沈惠璋[1]　邓莎莎[2]

(1. 上海交通大学 安泰经济与管理学院,上海　200030;
2. 上海外国语大学 国际工商管理学院,上海　200083)

摘　要: 随着社会的发展,我国群体性事件逐渐增多,且发生得日益频繁,这极大地影响到了广大人民群众和相关企事业单位的正常生活和生产。尤其是伴随着互联网的日益普及,越来越多的群体事件以网络为依托反映个人诉求,群体事件网络关注度成为众多学者关注的焦点。在分析群体事件特征的基础上,运用基于本体关系的数据挖掘算法,对群体事件网络关注度进行分类,获取了分类规则,并对分类规则进行了分析,为我国群体事件预警和治理提供了决策依据。

关键词: 本体关系;群体事件;网络关注度;数据挖掘

中图分类号: C 94　　**文献标志码:** A

Research on the Data Mining of the Influential Factors of Group Event Network based on Ontology Relation

GONG Xiaomin[1]　*SHEN Huizhang*[1]　*DENG shasha*[2]

(1. Antai College of Economics & Management, Shanghai Jiao Tong University, Shanghai 200030, China;
2. College of International Business, Shanghai International Studies University, Shanghai 200083, China)

Abstract: Recently, group event has brought great distress to people's daily life. With the development of the Internet, a growing number of people use web to come up with personal demands, thus the degree of concern on group event become the focus of many scholars. Based on the analyzes on the characteristics of group event, this paper used the ontology-based data mining algorithms to classify the degree of network concern on group events, and acquired classification rules. The classification rules were analyzed, and come up with some suggestion for the early warning and management decision-making on the group event.

Key words: ontology relationship; group event; degree of concern; data mining

1　群体事件网络关注度影响因素的选取

1.1　群体事件的特征提取

一般而言,在突发性的危机事件中,会采用一种叫作元本体 EMM 的模型,本文就是基于该模型展开相关研究的。在研究中,本体库的核心内容则是群体性事件。该事件又被分为三个方面的内容,分别是群体事件所处的状态、决策以及产生的效果等。这三部分,又能被细分为事件的基本信息、过程信息和结果信息等。

(1)事件的基本信息。在本文的研究中,决策变量为事件中的四个基本属性,分别为事件所发生的时间、地点、类型以及受到事件影响的人数。

收稿日期:2018-06-12

基金项目:国家社科基金重大项目"群体行为涌现机理及风险辨识研究"(11&ZD174)

作者简介:巩晓敏(1987—),女,山东枣庄人,博士研究生,主要研究方向为危机决策、数据挖掘;沈惠璋(1958—),女,天津人,博士,教授,主要研究方向为群决策、数据挖掘。

事件的类型：对于群体性事件而言，其类型和覆盖的范围都很大，种类也不一而足。而且，各种不同类型的群体性事件，其发生的原因、发展的阶段和经过，以及最终导致的后果，也互有差异。鉴于此，这些事件能够吸引到的群体也千差万别，引起的网络关注度也各有高低。

事件的发生地点：相比于世界上其他国家，我国幅员辽阔，人口众多，各地区人口数量和分布，素质和文化区别很大。所以，群体性事件也会随着地区的差异而有所变化，比如事件发展的经过、与之相对应的解决方案，等等。换言之，群体事件如果发生在不同地区，其引起的网络关注程度也会因为地区的差异而发生一定的变化。

事件的发生时间：对群体性事件而言，其发生的时间也会在一定程度上影响该事件的网络关注度。比如，如果一起群体性事件如果发生在周末，那么该事件所能引起的网络关注度就会较高，而且参与该事件的人数也会较多。但是，如果该事件发生在工作日内，那么关注该事件的人就会少一点，参与的人数也不会很多。同样，如果群体性事件发生在白天，其受到的网络关注度跟发生在晚上所受到的网络关注度也互有差异。

受到影响的人数：在群体性事件中，围观者等也对群体性事件的进程产生了显著的影响。譬如，围观者越多，受影响的人数也就越多，那么和群体性事件相关的信息就会被更广泛、迅速地传播开，从而对网络关注度产生相当的影响。

(2)在事件的过程信息中，本文选用的关键变量是群体性事件所能持续的时间。

事件的持续时间：一般而言，如果群体性事件的持续时间很长，那么网民们的猜测不仅会增多，还会失控，甚至产生各种各样的谣言。由此可见，对于网络关注度，事件的持续时间也起到了一定的作用。

事件解决的方案：在大部分的群体性事件中，研究发现如果对群体性事件应对不当，比如采用的解决方案不积极、回避问题，甚至增加冲突等，不仅不能平息事件，反倒会激起更大的反弹，让群体行为变得更加恶性。与此用时，在解决群体性事件时，网民往往对解决方所持的态度，以及采用的解决方案，都极为看重。

(3)事件的结果信息：在本文中，研究所选取的主要变量为伤亡人数，依此对群体性事件所导致的后果进行评定。

事件的伤亡人数：在任何一起群体性事件中，其导致的结果，最直观的表现就是事件中所产生的伤亡人数。在一些研究中，通过研究和分析网民们的心理。可以发现，伤亡人数越多，会吸引更多的网友关注事件进程并参与讨论。

1.2　群体事件网络关注度的计算

本文主要从三个方面对群体性事件所引起的网络关注度展开了相关的衡量和评价。这三个方面分别为群体性事件所引起的新闻数量、评论数量和参与的人数。但是，对这三个方面的相关数据进行统计时，它们各自的数量级和要用到的界面，不仅差异大，而且各不相同。因此，在进行统计运算前，要先对各个数据展开标准化变换，接着对它们进行权重的平均分配，最后再开展相关计算。在群体性事件的网络关注度方面，经常采用的量化方案如下所示：

群体事件网络关注度＝(新闻媒体关注度＋网络参与关注度＋网络评论关注度)/3

(1)新闻数量：在网上，网络新闻媒体对于群体事件的关注，可通过和群体性事件相关的新闻报导数量直接或间接地表现出来。其计算公式如下所示：

某事件的新闻媒体关注度＝(原始数据-min{新闻数量})/(max{新闻数量}-min{新闻数量}

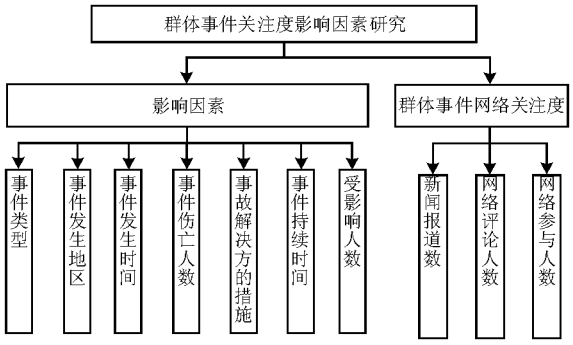

图 1　群体事件网络关注度及其影响因素结构

(2)新闻评论数：该条意思是指，新闻网页上对新闻进行评论的人数之和。通常，该人数之和可以反映出网民对某事件的关注程度。其计算公式如下所示：

某事件网络评论关注度＝(原始数据-min{网络评论数量})/(max{网络评论数量}-min{网络评论数量}

(3)新闻参与人数：所谓的新闻参与人数是指，有一些网民不会留言对事件进行评论，他们更乐于观看其他网民的评论，所以，这些只观看不留言的网民，也是一种能够观察到的变量。其计算公式如下所示：

某事件网络参与关注度＝(原始数据-min{网络新闻参与人数})/(max{网络新闻参与人数}-min

{网络新闻参与人数}

2　基于本体关系的 ID3 算法描述

2.1　基于本体关系的数据挖掘算法研究

利用属性的本体关系进行分类的算法已有不少，大体上主要有建立本体规则的方法和对属性值分类的方法。本文所采用的方法是借鉴 Zhang Jun 的利用本体关系进行分类思想的基础上，与 ID3 算法相结合的方法。

2.1.1　算法的优点

传统的决策树算法主要通过数据库二维表对群体事件案例进行表示，语义表达能力较弱，同时每个变量实例都最终会产生一个节点，形成的决策树规模大，复杂性高，不便于理解和操作。

基于本体的决策树算法通过对本体的运用，首先能够增强群体事件案例的语义表达能力，提升决策树检索和分类的有效程度。其次，用决策树表示更加简单，容易理解，同时在分类方面更加精确、可信。分析和统计有限的数据时，从具体数据层面入手的方式，并没有从抽象概念层面入手的方式显得准确和可靠。最后，基于本体的决策树算法还为解决数据挖掘中的过度拟合现象提供了一个新的解决思路。

2.1.2　算法描述

第一步：构建测试属性本体。

第二步：依次构建各测试属性的本体关系。

第三步：频数统计。

(1)按照词频进行统计；

(2)自下而上，将子节点的频数加到父节点上；

(3)自上而下，将抽象节点的频数按照子节点分布规律分配到子节点上；

第四步：生成决策树。

(1)构造向量 p，p 向量的各分量指向各属性的一个节点；

(2)计算 p 向量所指的属性的熵增，将最大信息熵增 p 分量作为决策属性；

(3)构造 P 向量组，由决策属性的子节点代替 p 向量中的父节点，形成新的 p 向量；

(4)循环以上三步，直至 p 向量分项值均为空。

2.1.3　算法实现

算法实现通过.NET 平台的进行开发，运用的 C＃语言实现了基本算法功能。

2.2　群体事件特征本体的构建

本研究采用的算法是一种基于本体的 ID3 算法，该算法要求对上述群体事件中所选取的特征建立起各自相应的领域本体。同时，还要对本体领域中存在的知识结构进行相当程度的考量，然后再对群体事件中的七个特征做如下的本体表述：

(1)事件的类型：我国的群体性事件，按发生的起因和所在的领域，可以具体地被划分为下列十种类型：一、公共卫生事件引发的群体行为；二、公关管理与执法不公冲突引发的群体行为；三、劳资冲突引发的群体行为；四、人为事故灾害引发的群体行为；五、社会安全事件引发的群体行为；六、征地拆迁冲突引发的群体行为；七、资源与环境冲突引发的群体行为；八、自然灾害引发的群体行为；九、族群矛盾与境外势力冲突引发的群体行为；十、以网络与微博为载体的群体行为。上述分类较为详细，包容面广，在案例信息的搜集和整理方面能起到很大的贡献。具体见图 2。

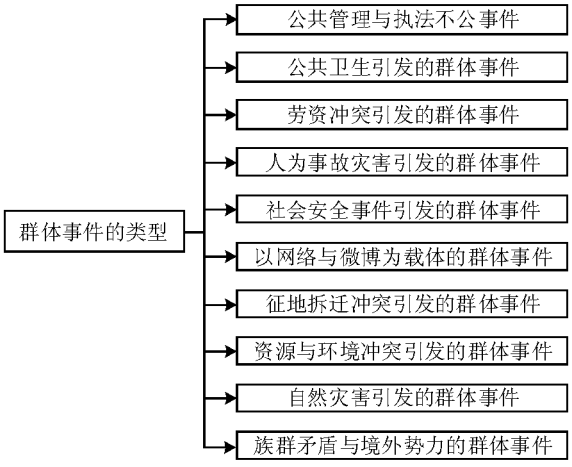

图 2　群体事件类型的本体关系分解

(2)事件的发生地点：本文主要对我国境内发生的群体性事件进行研究，并着重于它们的一般特征。所以，在事件的发生地点上，本研究构建了一种具有三层结构的继承关系。比如，先按照地域，如东北、西北、西南、华中、华北、华南、华东、华西等，将群体性事件的发生地点划分为四大类型；接着，再根据地域所在的省市细分这四大类型，最终获得本体。

(3)事件发生的时间：根据群体性事件中参与的人数特点，以及该特点在事件进程各时间段的不同表现，还对群体传播中时间所发挥的影响展开了相关分析。在本文中，研究者为事件发生的时间构建了具有三层结构的继承关系，这是一种本体结构。该本体的构建分为两个步骤：第一，根据工作日和非工作日对群体事件进行划分；第二，基于白天和黑夜两个时段中群体性事件的不同特点，对事件进行更加详细的分类，如非工作日白天、非工作日夜晚、工作日白天、工作日夜晚。具体见图 3。

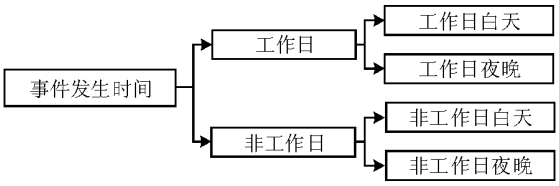

图 3　事件发生时间的本体关系分解

(4)事故解决方采取的解决方案:根据事故解决方对群体性事件所采取的解决态度,解决方案可以被划分为两种类型,即主动反应和被动反应。其中,主动反应又可以依据案例分析的结论,具体细分为 8 种类型的解决方案,分别为:包庇既得利益者、执法不当、不当言论、武力威慑、对抗、封锁消息、调解疏导、协商解决。而被动反应又能被细分为下述 5 种行为,分别为:敷衍民众、故意拖延、反应迟钝、无作为、直接妥协。鉴于此,所获得的事故解决方之继承本体应如下述内容,见图 4。

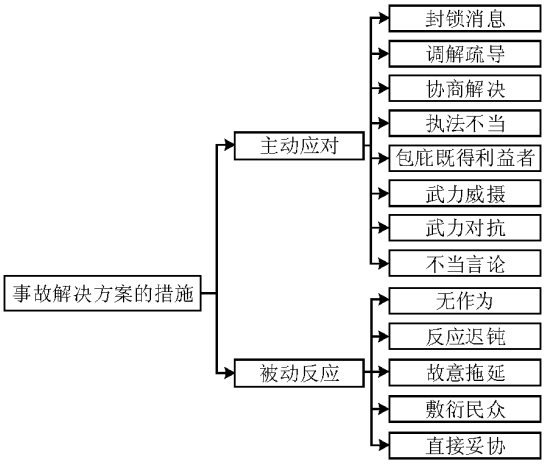

图 4　事故解决方的措施本体关系分解

(5)事件的持续时间:群体性事件引起的网络关注度,受到该事件持续时间的显著影响。本研究中,群体性事件只是简单地根据事件的持续时间被划分为三个类别,分别为:"<1 天"">7 天"和"1～7 天"。

(6)事件中的伤亡人数:根据全部群体性事件中伤亡的人数展开数字排序,把前面三分之一的伤亡人数定义为"大",后面三分之一的伤亡人数定义为"小",中间三分之一的伤亡人数定义为"中"。通过这样的划分,构建起一种具有两层结构的继承关系本体。

(7)受影响的人数:根据全部的群体性事件案例中受到影响的人数展开数字排序,把前面受影响人数的三分之一定义为"多",后面的三分之一定义为"少",中间的三分之一定义为"中"。通过这样的分类,构建起一种适用于受影响人数的具有两层结构的继承关系本体。

3　实验与分析

本实验共搜集有效案例 612 个,其中 80%的案例作为训练集,通过对原始数据的训练,获得初始决策树;另外 20%的案例作为测试集,通过测试判别决策树的有效程度。

3.1　原始数据的搜集

通过查阅相关数据和统计,在我国,每天有超过五百起群体性事件发生,给人民群众的财产带去了巨大的损失。而且,这种群体事件有逐年上升的趋势,从最初的几万起,到现在的二十多万起,严重影响了社会的正常生活和生产秩序。本文所做的研究,案例、素材均取自互联网,共收集和整理了 2011—2013 年,三年间的我国各类群体事件案例共计 612 起。新闻报道也主要取自六家既能发布新闻又能让网民对新闻进行互动评论的网站,比如腾讯新闻、凤凰资讯、新浪新闻、网易新闻、人民网和中国新闻网。本文案例中,所引用的基础数据如表 1 所示。

表 1　群体事件部分案例库

序号	事件类型	发生地点	发生时间	事件持续时间	伤亡人数	受影响人数	事件解决方措施	网络关注度
1	社会安全	浙江	非工作日白天	>7 天	小	多	反应迟钝	中
2	公共管理与执法不公	上海	工作日白天	1～7 天	小	少	调解疏导	中
3	人为事故	江苏	工作日白天	>7 天	小	多	调解疏导	高
4	公共卫生	广东	工作日白天	1～7 天	小	多	主动应对	高
5	以网络与微博为载体	湖南	非工作日白天	1～7 天	小	多	不当言论	高
6	劳资冲突	广州	工作日白天	1～7 天	小	中	武力威慑	中
7	劳资冲突	山西	工作日白天	<1 天	小	少	武力威慑	中
8	自然灾害	湖南	工作日白天	1～7 天	大	多	主动应对	高
9	征地拆迁	安徽	工作日白天	1～7 天	小	多	武力对抗	高
10	资源与环境	山东	非工作日白天	1～7 天	小	少	调解疏导	低
11	社会安全	江西	工作日夜晚	>7 天	大	多	主动应对	高
13	社会安全	浙江	工作日白天	>7 天	大	多	反应迟钝	低
14	征地拆迁	河南	非工作日白天	1～7 天	小	多	反应迟钝	低
15	社会安全	广东	工作日白天	>7 天	小	多	武力威慑	高

3.2 构建决策树

基于上述分析和本体的 ID3 算法，本文对决策树进行了构建。其中，选择群体性事件的七个变量，比如事件发生的类型、时间、地点，事件所持续的时间、受到事件影响的人数、出现的伤亡人数，以及解决方所采取的解决方案等，作为测试变量。同时，又将群体性事件所引起的网络关注度当作分类变量，从而使用改进后的决策树算法展开相关计算。本文研究中构造的决策树模型，正如图 5 所示。

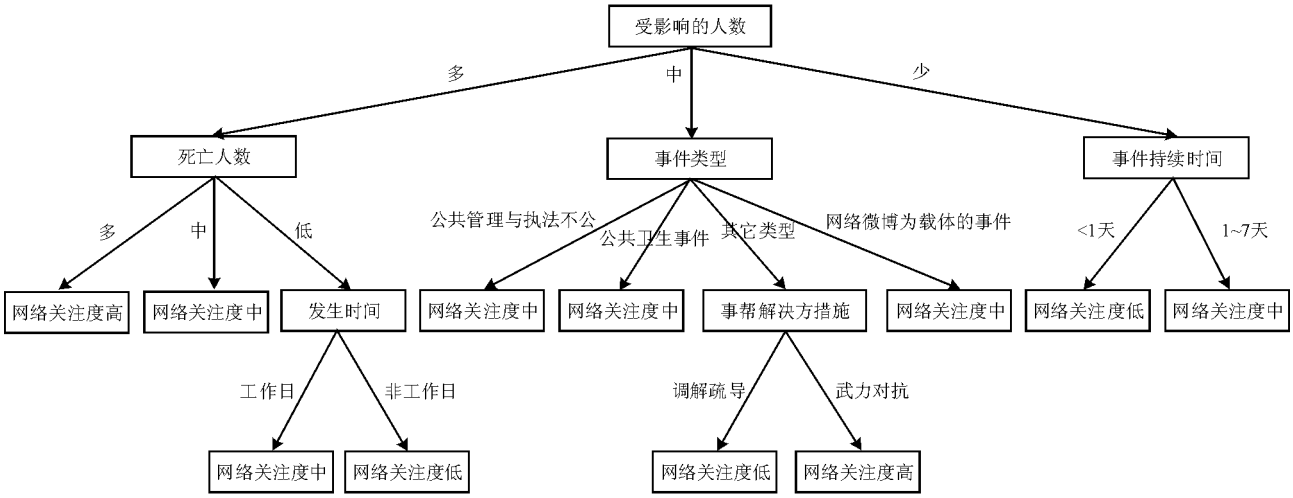

图 5　群体事件网络关注度分类树

3.3 分类规则

根据决策树生成的分类规则得到：

(1)IF 受影响人数＝“多”and 伤亡人数＝“多”THEN 网络关注度＝“高”

(2)IF 受影响人数＝“多”and 伤亡人数＝“中”THEN 网络关注度＝“中”

(3)IF 受影响人数＝“多”and 伤亡人数＝“低”and 发生时间＝“工作日”THEN 网络关注度＝“中”

(4)IF 受影响人数＝“多”and 伤亡人数＝“低”and 发生时间＝“非工作日”THEN 网络关注度＝“低”

(5)IF 受影响人数＝“中”and 事件类型＝“公共管理与执法不公”THEN 网络关注度＝“中”

(6)IF 受影响人数＝“中”and 事件类型＝“公共卫生事件”THEN 网络关注度＝“中”

(7)IF 受影响人数＝“中”and 事件类型＝“以网络与微博为载体的群体事件”THEN 网络关注度＝“中”

(8)IF 受影响人数＝“中”and 事件类型＝“其他类型”and 事故解决方的措施＝“协商解决”THEN 网络关注度＝“低”

(9)IF 受影响人数＝“中”and 事件类型＝“其他类型”and 事故解决方的措施＝“武力对抗”THEN 网络关注度＝“高”

(10)IF 受影响人数＝“少”and 事件持续时间＝“<1 天”THEN 网络关注度＝“中”

(11)IF 受影响人数＝“少”and 事件持续时间＝“1 ～7 天”THEN 网络关注度＝“少”

3.4 数据检验

我们将获得的 612 份数据分为 10 份，然后将其中 9 份的数据作为训练集，还有一份作为测试集。具体来说，9 个训练集共拥有 540 个案例，测试集内有 62 个案例。使用交叉验证法对数据集分别进行 10 次验证，最终得到表 2 所示的基于决策树的验证准确率，并以此对算法的精度进行估算，得到表 2 所示的对决策树效果的校验结论。

表 2　10 次交叉验证结果

次序	预测正确数	预测错误数	准确率/%
1	47	7	87.04
2	45	9	83.33
3	49	5	90.74
4	46	8	85.18
5	48	6	88.89
6	50	4	92.59
7	47	7	87.04
8	49	5	90.74
9	45	9	83.33
10	55	7	88.12
平均值	48.1	6.7	87.78

由表 2 我们可知，此分类树的查准率较高，为 88.78%，预测准确性很高，能够作为参考依据。

3.5 实验结果分析

本文通过群体事件网络关注度的分类规则深入

了解群体事件网络关注的驱动因素，为群体事件的治理提出了可行性建议。基于以上决策规则，总结群体事件网络关注度影响因素的规律：

(1)受影响人数是群体事件网络关注度的主要影响因素之一，随着受影响人数的增加，群体事件网络关注度普遍升高。因此，建议在处理群体事件的过程中，相关单位需要忧患意识，在群体事件集化程度较低时妥善处置，避免事件影响范围扩大。

(2)与非工作日相比，工作日发生的群体事件的网络关注度普遍偏低。因此，对于非工作日组织或自发的群体行为要重视并做好引导和协调工作，避免事态的扩大升级。

(3)对于参与人数达到 100～1 000 人的，即受影响人数中等的，对于不同类型的事件需要采用不同的处置措施。从决策树规则可以看出，对于劳资纠纷、征地拆迁冲突等事件，采用武力解决会对群体事件网络关注度产生促进作用，吸引媒体和公众的关注，未来在群体事件治理过程中，建议采用协商解决的方式积极表态和处理，避免直接冲突。

(4)对于受影响人数低于 100 人的，侧重于观察事件的持续时间，建议在应对群体事件的过程中，尽量能够在一周内解决群体事件产生的最尖锐矛盾，将群体事件带来的危害降到最低。

4 结论

综上所述，本文的成果和进一步展望如下：

(1)传统数据挖掘方法语义表达能力较弱，基于本体的数据挖掘算法在传统算法的基础上增强了语义表达能力，生成的决策树更加精简，具有更高的检索效率。

(2)得到群体性事件所引起的网络关注度、群体性事件特点的分类规则，并对以后预警群体性事件和对群体性事件进行有效决策，提供了充足的数据支持。

(3)未来可以建立群体事件案例决策支持系统，案例的数量可以通过滚动式的积累进行进一步拓展和补充。

参考文献：

[1] 吕莉. 网络群体性事件的成因与传播模式[J]. 新闻世界，2013(7):164-165.

[2] 阳德青，肖仰华，汪卫. 基于统计模型的社会网络群体关注度的分析与预测[C]// ndbc2010 中国数据库学术会议，2010:000378-384.

[3] 张明军，陈朋. 2014 年度中国社会典型群体性事件分析报告[J]. 中国社会公共安全研究报告，2015(1).

[4] WEI J，ZHOU L，WEI Y，et al. Collective behavior in mass incidents：A study of contemporary China [J]. Journal of Contemporary China，2014，23(88)：715-735.

[5] 瞿琼丹. 基于本体的群体性突发事件案例推理研究[D]. 兰州：兰州大学，2012.

[6] 赵继娣，沈惠璋，刘欢. 突发危机事件管理的元本体模型[J]. 科技管理研究，2013，289(15):240-243.

[7] 徐弋加，韩耀赐，何冠霄，等. 基于本体建模的应急管理决策支持方法及在 MERS 中的应用[J]. 管理评论，28(8).

[8] 耿晓平. 基于本体的决策树算法在应急决策系统中的研究[J]. 机械管理开发，2011：134-136.

[9] ZHANG J，HONAVAR V. Learning naive bayes classifiers from attribute-value taxonomies and partially specified data[C]// International Conference on Machine Learning，2003.