2021-2022学年春季学期

网络空间安全态势感知

Cyber security situation awareness

授课团队：刘宝旭 卢志刚 刘玉岭

助　　教：李　宁

网络空间安全态势感知

Cyber security situation awareness

[第11次课] 态势要素融合与归一化技术

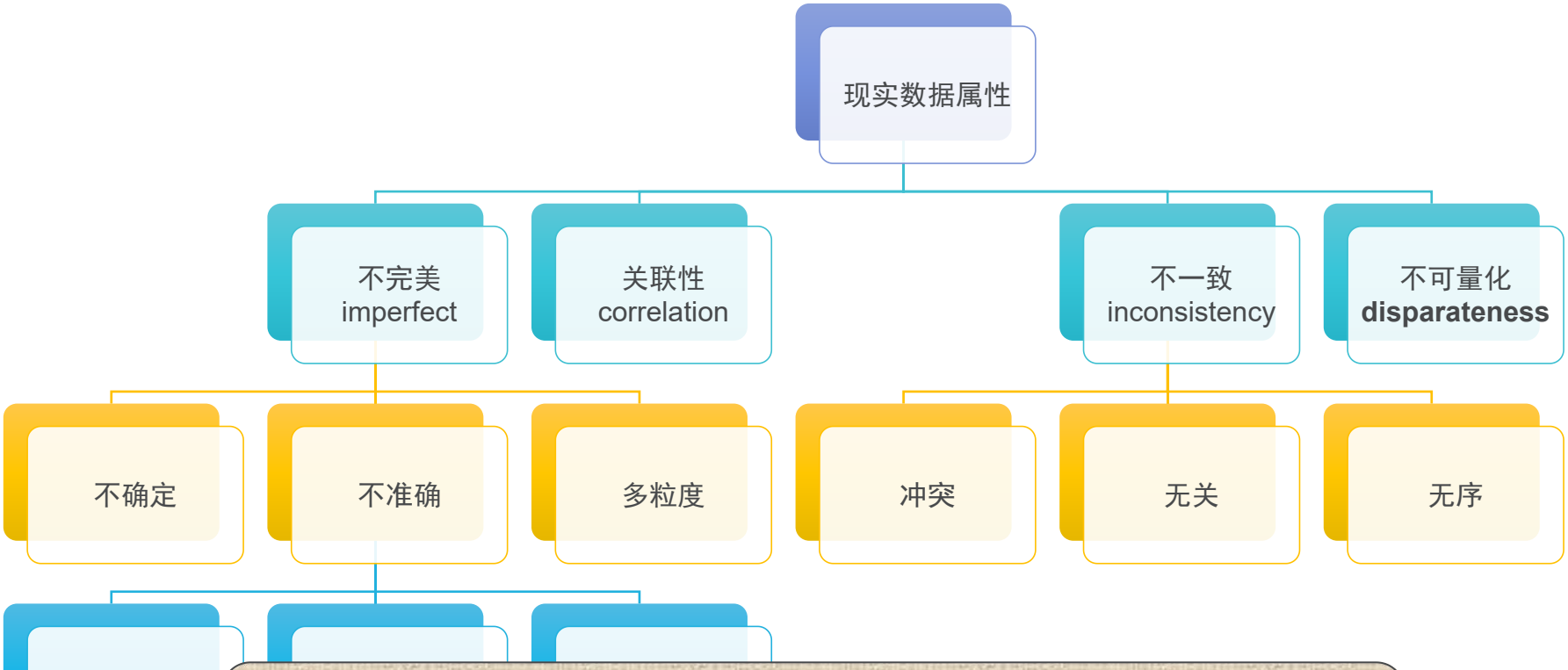
授课教师：刘玉岭

授课时间：2022.3.28

内容概要

一、安全态势要素融合与归一化意义

- 现实中的数据对于业务处理来讲是“脏”的



数据质量与态势的感知结果密切相关
亟需开展要素融合与归一化处理

一、安全态势要素融合与归一化意义

安全态势要素融合与归一化作用	攻防模式发掘
	安全异常检测
	辅助安全决策

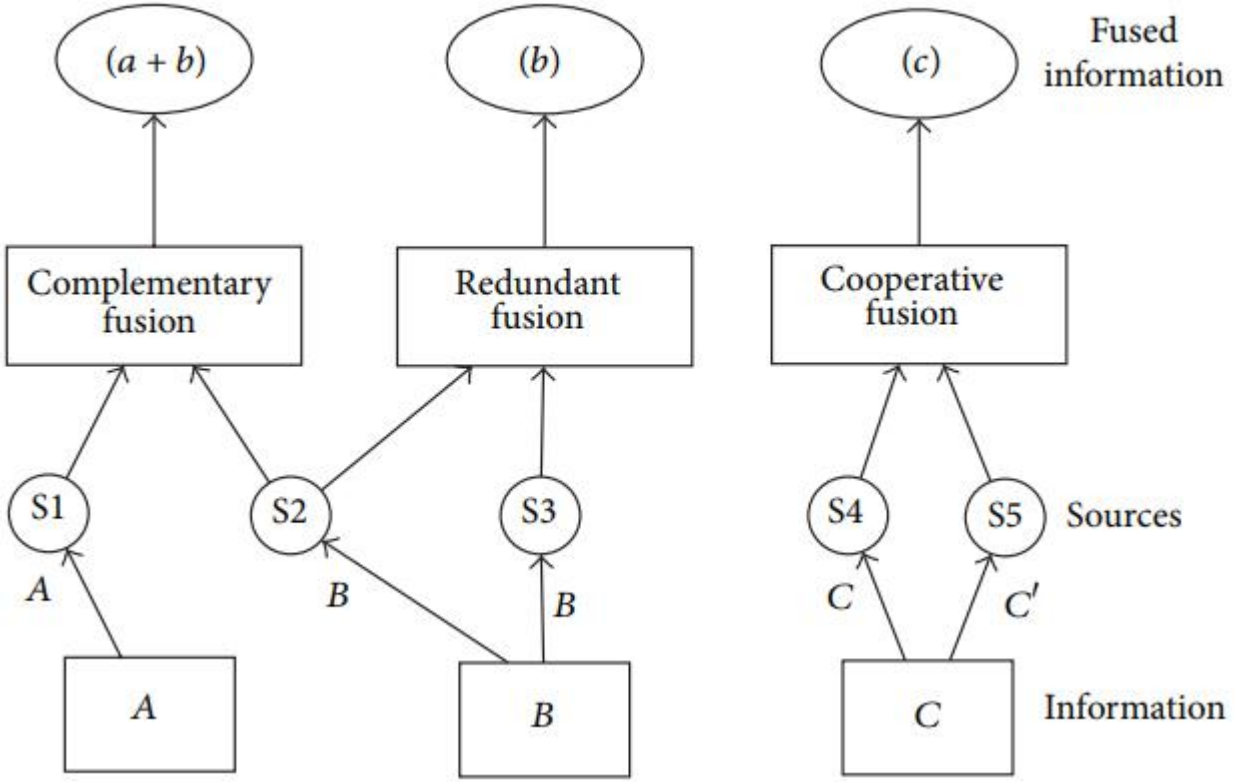
一、安全态势要素融合与归一化意义

- 多感知器：提高可探测性和可信度， 扩大时空感知范围
- 多源数据：增加目标特征维数，提高异常发掘能力
- 多维数据：提高推理准确程度，增强态势感知精度
- 分布部署：增强态势感知系统容错能力和自适应性

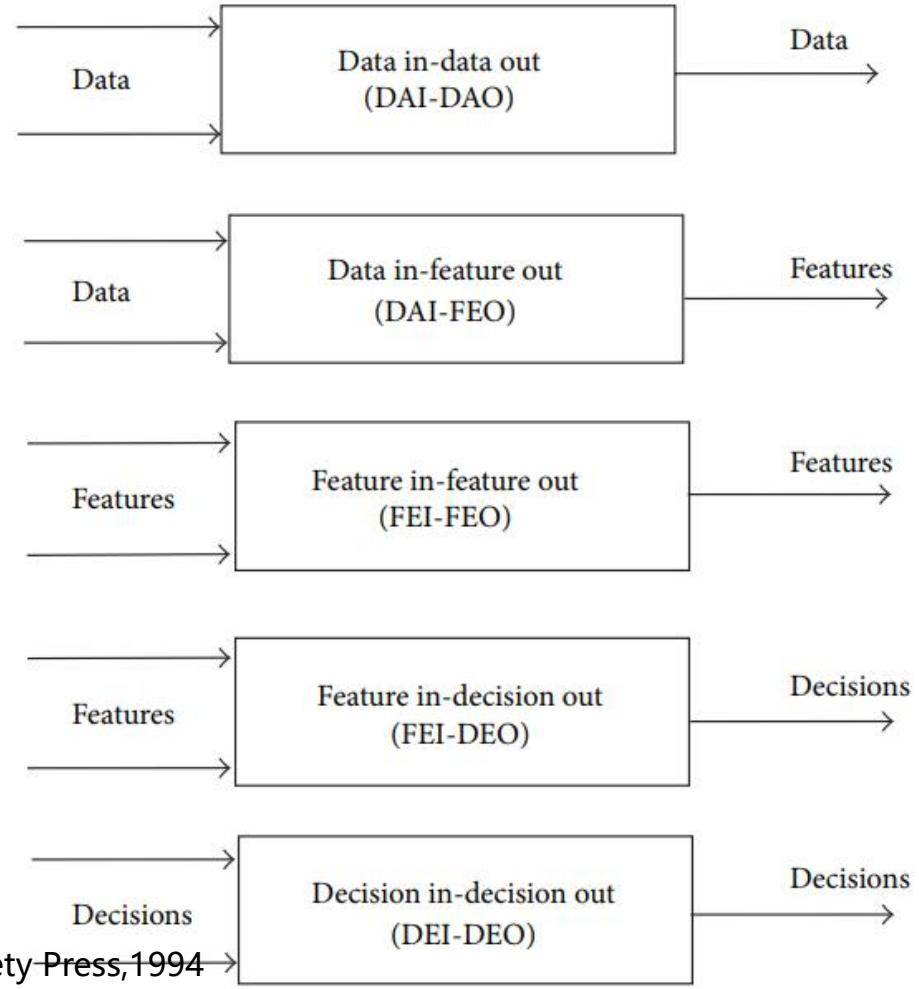
内容概要

- 一、安全态势要素融合与归一化意义与作用
- 二、安全态势要素融合与归一化技术分类
- 三、安全态势要素融合与归一化主要技术
- 四、未来挑战

- Whyte的基于数据源关系的分类方法
  - 互补式
  - 冗余式
  - 合作式



- Dasarathy的基于数据输入输出的分类方法
  - “数据进-数据出”
  - “数据进-特征出”
  - “特征进-特征出”
  - “特征进-决策出”
  - “决策进-决策出”



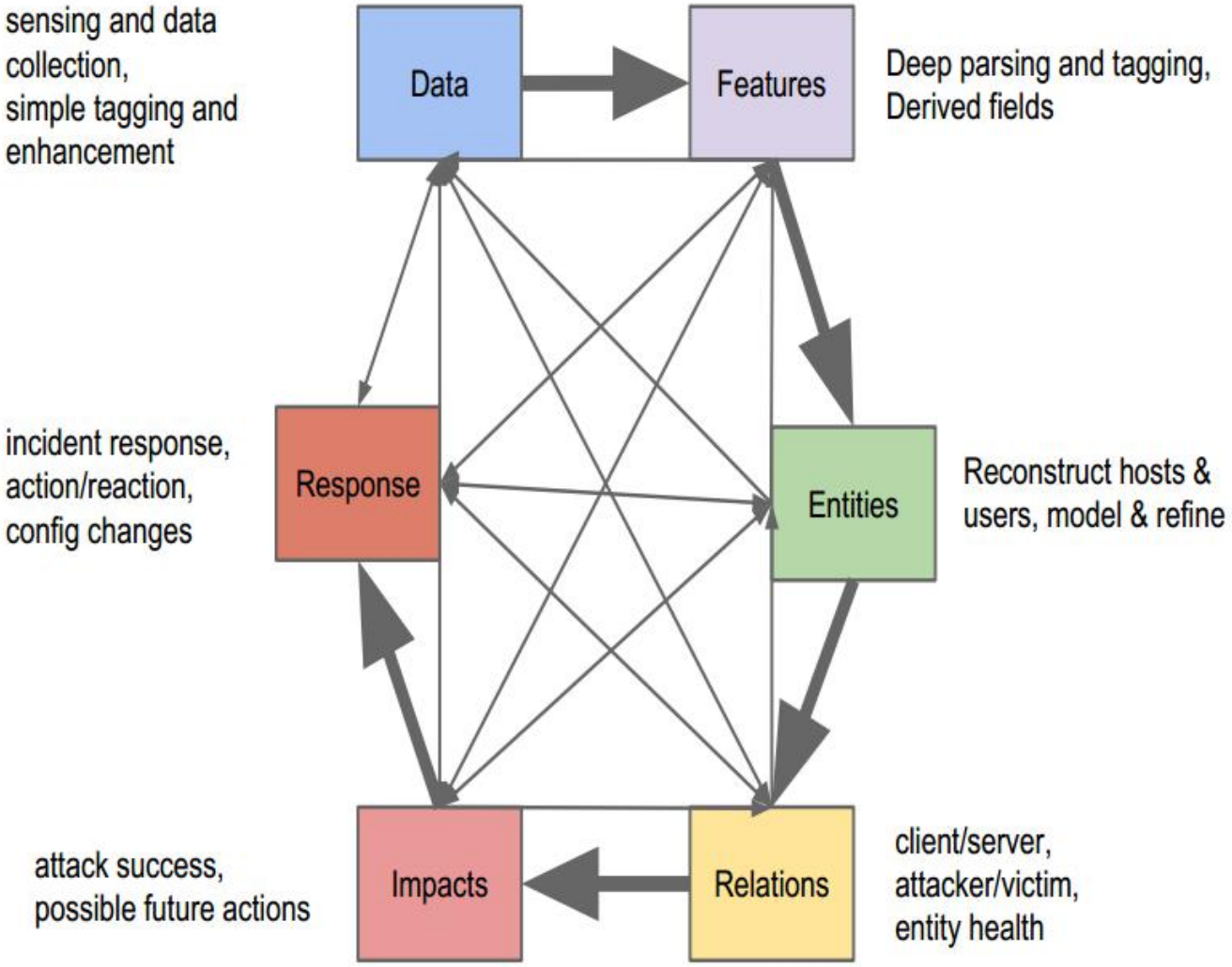
Dasarathy, B., *Decision Fusion*, IEEE Computer Society Press,1994

- Dasarathy的基于数据输入输出的分类方法(拓展)

Inputs \ Outputs	Data	Features	Entities	Relations	Impacts	Responses
Data	Signal Detection	Feature Extraction	Gestalt-Based Entity Extraction	Gestalt-Based Situation Assessment	Gestalt-Based Impact Assessment	Reflexive Responses
Features	Model-Based Detection/ Feature Extraction	Feature Refinement	Entity Characteriza-tion	Feature-Based Situation Assessment	Feature-Based Impact Assessment	Feature-Based Responses
Entities	Model-Based Detection/ Estimation	Model-Based Feature Extraction	Entity Refinement	Entity-Relational Situation Assessment	Entity-Based Impact Assessment	Entity-Relation Based Responses
Relations	Context-Sensitive Detection/ Estimation	Context-Sensitive Feature Extraction	Context-Sensitive Entity Refinement	Micro/Macro Situation Assessment	Context-Sensitive Impact Assessment	Context-Sensitive Responses
Impacts	Cost-Sensitive Detection/ Estimation	Cost-Sensitive Feature Extraction	Cost-Sensitive Entity Refinement	Cost-Sensitive Situation Assessment	Cost-Sensitive Impact Assessment	Cost-Sensitive Responses
Responses	Reaction-Sensitive Detection/ Estimation	Reaction-Sensitive Feature Extraction	Reaction-Sensitive Entity Refinement	Reaction-Sensitive Situation Assessment	Reaction-Sensitive Impact Assessment	Reaction-Sensitive Responses

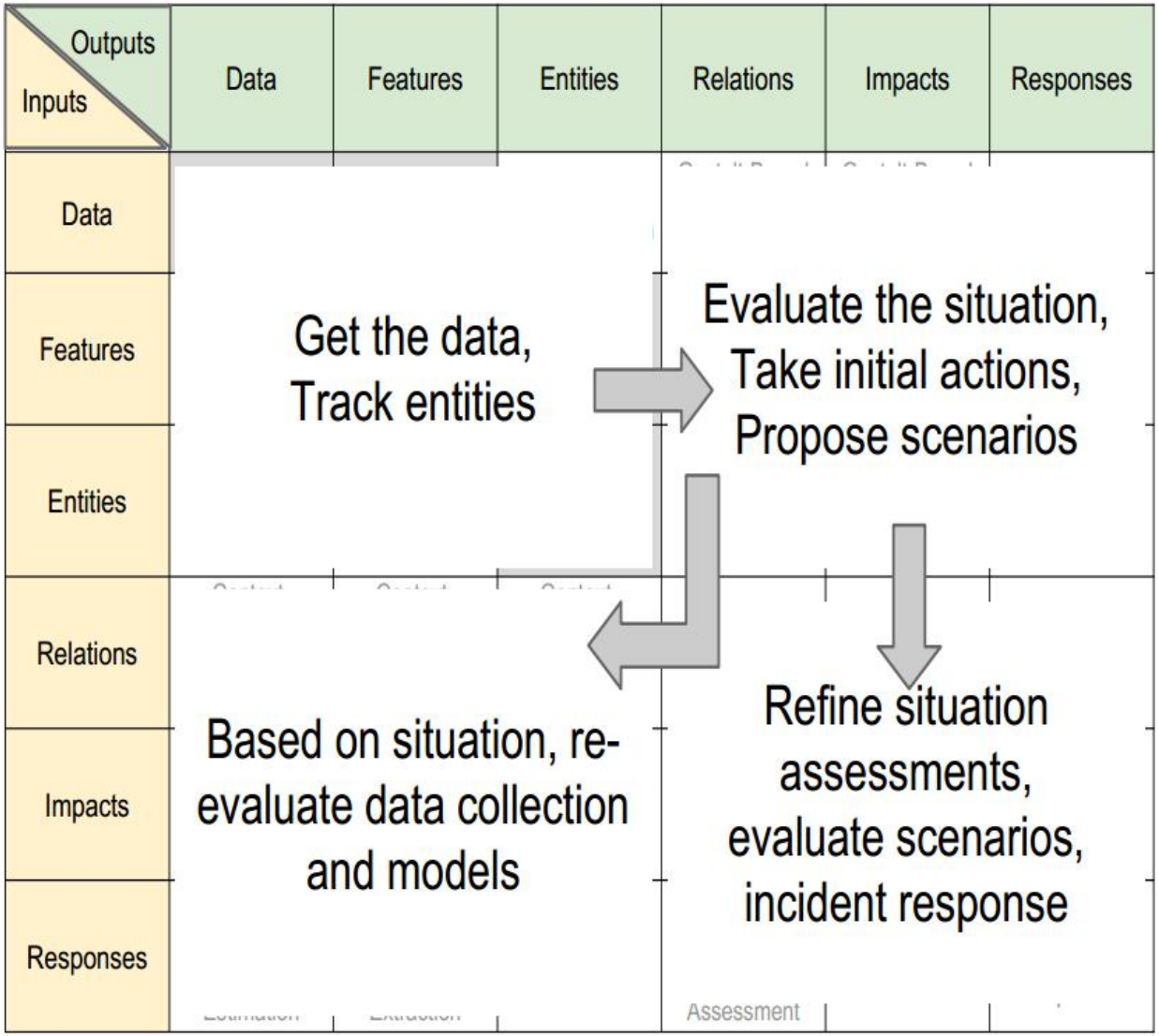
二、融合与归一化技术分类

- Dasarathy的基于数据输入输出的分类方法(拓展)
  - 各类“数据”的作用分析
  - 各要素的关系分析



二、融合与归一化技术分类

- Dasarathy的基于数据输入输出的分类方法(拓展)
  - 典型的分析步骤

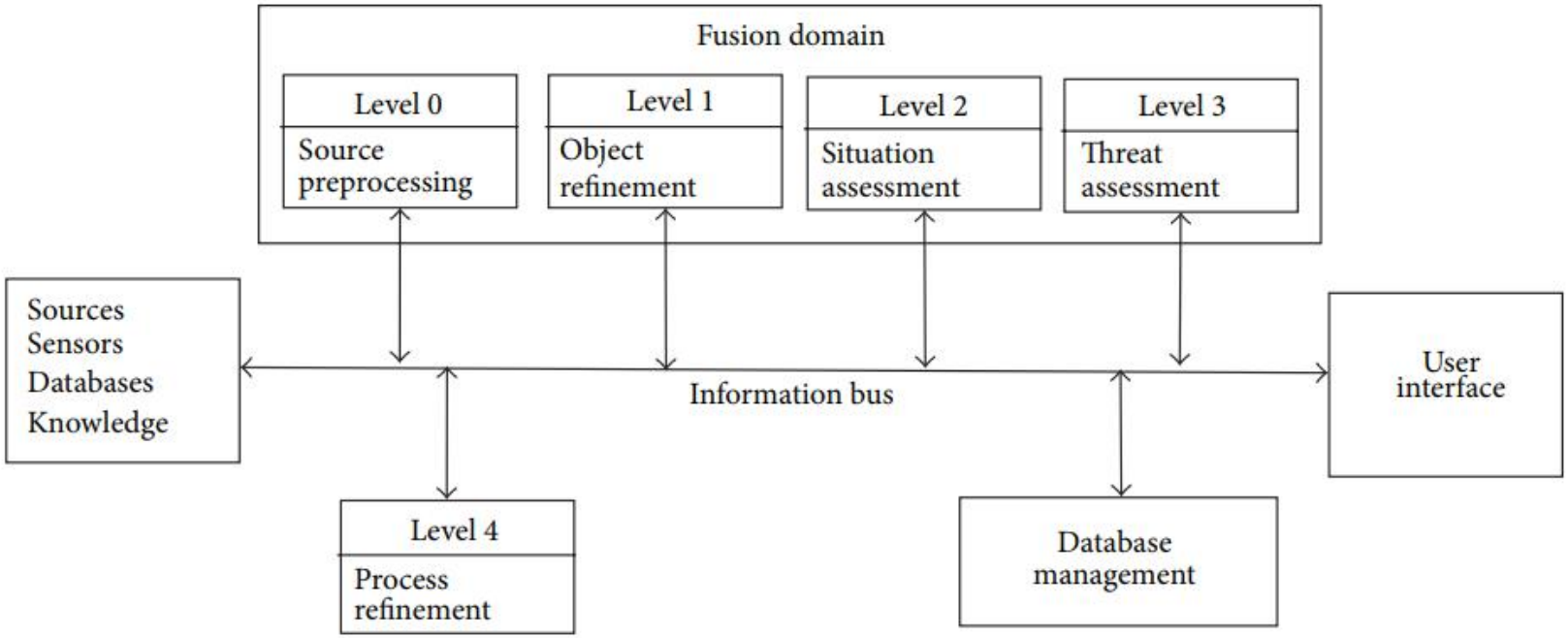


二、融合与归一化技术分类

- Dasarathy的基于数据输入输出的分类方法(拓展)
  - 已经在做的工作

Inputs \ Outputs	Data	Features	Entities	Relations	Impacts	Responses
Data	**PCAP, flow, syslog**	**DPI, Log parsing**	**DHCP, Auth logs**	**DDoS Detection**	Gestalt-Based Impact Assessment	Reflexive Responses
Features	Model-Based Detection/ Feature Extraction	**SIEM**	Entity Characteriza-tion	**IDS**	**SIEM**	**IPS**
Entities	Model-Based Detection/ Estimation	Model-Based Feature Extraction	Entity Refinement	Entity-Relational Situation Assessment	Entity-Based Impact Assessment	Entity-Relation Based Responses
Relations	Context-Sensitive Detection/ Estimation	Context-Sensitive Feature Extraction	Context-Sensitive Entity Refinement	Micro/Macro Situation Assessment	Context-Sensitive Impact Assessment	Context-Sensitive Responses
Impacts	Cost-Sensitive Detection/ Estimation	Cost-Sensitive Feature Extraction	Cost-Sensitive Entity Refinement	Cost-Sensitive Situation Assessment	Cost-Sensitive Impact Assessment	Cost-Sensitive Responses
Responses	Reaction-Sensitive Detection/ Estimation	Reaction-Sensitive Feature Extraction	Reaction-Sensitive Entity Refinement	Reaction-Sensitive Situation Assessment	Reaction-Sensitive Impact Assessment	Reaction-Sensitive Responses

- JDL的数据融合模型
  - 由美国国防部的数据融合联合指挥实验室提出



Steinberg, A. and Bowman, C. Revisions to the JDL Data Fusion Model, in *Handbook of Multisensor Data Fusion*, 2001

内容概要

三、安全要素的融合与归一化技术

- 数据层处理技术
  - 清洗
  - 转换
  - …
- 特征层处理技术
  - 聚类
  - 关联
  - …
- 决策层处理技术
  - 贝叶斯方法
  - D-S方法
  - 决策树
  - 知识推理
  - …

- 数据层处理技术
  - 核心作用：数据层面的预处理，即清洗、整形、归约等
  - 数据清洗Data Cleaning
    - 填充缺失数据、消除噪音数据、去除外部无关数据、确保一致性等
  - 数据集成Data integration
    - 集成多源的数据集、数据文件等
  - 数据归约Data reduction
    - 降维归约
    - 减量归约
    - 数据压缩
  - 数据转换Data transformation
    - 数据规范化

● 数据层处理技术-数据清洗技术

忽略元组

通常当在缺少类标号时，通过这样的方法来填补缺失值

用属性的均值填充缺失值

数据属性分为数值属性和非数值属性进行处理，通过利用已存数据的多数信息来推测缺失值

人工填写缺失值

数据偏离的问题小，但该方法十分费时,不具备实际的可操作性

填充缺失值

用同类样本的属性均值填充缺失值

利用均值替换缺失值

使用一个全局常量填充缺失值

大量采用同一属性值，可能会误导挖掘程序得出有偏差甚至错误的结论

使用最可能的值填充缺失值

数据属性分为数值属性和非数值属性进行处理，通过利用已存数据的多数信息来推测缺失值

数据层处理技术-数据清洗技术

- 分箱方法
  - 按照属性值划分子区间， 通过考察同一个子区间内相邻数据来确定最终的值
  - 等深分箱法、 等宽分箱法、 最小熵法和用户自定义区间法
- 聚类方法
  - 依据对象特征属性的距离来将一组对象按照距离指标划分为特征相似的不同类别， 并将孤立于所有类别的数据作为离群点(或噪声)清除
  - Kmeans、 分层聚类、 基于密度的聚类等
- 回归分析方法
  - 通过构建相应的数学模型， 从而用一个组函数关系来描述特征变量和目标变量之间的关联关系， 通常被用来做预测分析
  - SVM、 人工神经网络、 决策树等

- 数据层处理技术-数据集成技术
- Why?
  - 数据多源
  - 多通道
- Mode?
  - 纯数据集成：将多来源数据集合成一个数据
  - 数据模式集成：集成多来源的元数据，比如同一威胁在不同安全设备中的表示名称

（1）模式集成和对象匹配问题

（2）冗余问题

（3）元组重复

（4）数据值冲突的检测与处理问题

三、安全要素的融合与归一化技术

● 数据层处理技术-数据集成技术

- 分箱方法
  - 按照属性值划分子区间， 通过考察同一个子区间内相邻数据来确定最终的值
  - 等深分箱法、 等宽分箱法、 最小熵法和用户自定义区间法
- 聚类方法
  - 依据对象特征属性的距离来将一组对象按照距离指标划分为特征相似的不同类别， 并将孤立于所有类别的数据作为离群点(或噪声)清除
  - Kmeans、 分层聚类、 基于密度的聚类等
- 回归分析方法
  - 通过构建相应的数学模型， 从而用一个组函数关系来描述特征变量和目标变量之间的关联关系， 通常被用来做预测分析
  - SVM、 人工神经网络、 决策树等

三、安全要素的融合与归一化技术

- 数据层处理技术-数据变换技术
  - 将数据变换为另一种数据，利于后续分析使用

1、光滑。去除数据中的噪声

2、聚集。对数据进行汇总或聚集。

3、 数据泛化。使用概念分层，用高层概念替换低层或“原始”数据

4、规范化。将属性数据按比例缩放，使之落入一个小的特定区间

5、属性构造。可以构造新的属性并添加到属性集中，以帮助挖掘过程

- 数据层处理技术-数据归约技术
  - 数据冗余信息多，数据量大，不利于分析效率的提高
  - 在不破坏数据完整性的同时， 通过使用比原始数据规模更小的子集进行融合
  - 常用的数据归约方法
    - 维度归约：去除不重要的属性
      - 小波变换
      - 主成分分析PCA
      - 特征选取
    - 数值压缩
      - 回归分析
      - 采样
    - 数据压缩
    - 离散化
    - 概念分层等

- 特征层处理技术-聚类（1）
  - 将具体或抽象对象的集合分组成由相似对象组成的为多个类或簇的过程
  - 由聚类生成的簇是一组数据对象的集合，簇必须同时满足以下两个条件
    - 每个簇至少包含一个数据对象
    - 每个数据对象必须属于且唯一地属于一个簇

聚类过程

聚类算法的要求

- 特征层处理技术-聚类（2）
- 划分式聚类方法：将给定的数据集初始分裂为*K*个簇，每个簇至少包含一条数据记录，然后通过反复迭代至每个簇不再改变即得出聚类结果，典型方法为K-Means算法

常用距离算法

1）欧氏距离

$$d\left(x_i,x_j\right)=\left|\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2\right|^{\frac{1}{2}}$$

2）曼哈顿距离

$$d\left(x_i,x_j\right)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$$

3）闵可夫斯基距离

$$d\left(x_i,x_j\right)=\left|\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^r\right|^{\frac{1}{r}}$$

4）切比雪夫距离

$$d\left(x_i,x_j\right)=\max_{k\in\{1,2,\cdots,p\}}\left\{\left|x_{ik}-x_{jk}\right|\right\}$$

- 特征层处理技术-聚类（3）
- 具有噪声的基于密度的空间聚类应用（ Density-Based Spatial Clustering of Application with Noise， DBSCAN ）：
  - 从任意对象*P*开始根据阈值和参数通过广度优先搜索提取从*P*密度可达的所有对象，得到一个聚类
  - 若*P*是核心对象，则可以一次标记相应对象为当前类并以此为基础进行扩展，得到一个完整的聚类后，再选择一个新的对象重复上述过程
  - 若*P*是边界对象，则将其标记为噪声并舍弃

如聚类的结果与参数关系较大

- 阈值过大容易将同一聚类分割
- 阈值过小容易将不同聚类合并

固定的阈值参数对于稀疏程度不同的数据不具适应性

- 密度小的区域同一聚类易被分割
- 密度大的区域不同聚类易被合并

- 特征层处理技术-聚类（4）
- 基于模型的聚类：构建一个模型，寻找数据对给定模型的最佳拟合

概念聚类是机器学习中的一种聚类方法，给出一组未标记的数据对象，它产生一个分类模式。

概念聚类除了确定相似对象的分组外，还为每组对象发现了特征描述，即每组对象代表了一个概念或类。

统计学方法(EM和COBWEB算法)

神经网络方法将每个簇描述成一个模型。模型作为聚类的一个“原型”，不一定对应一个特定的数据实例或对象。

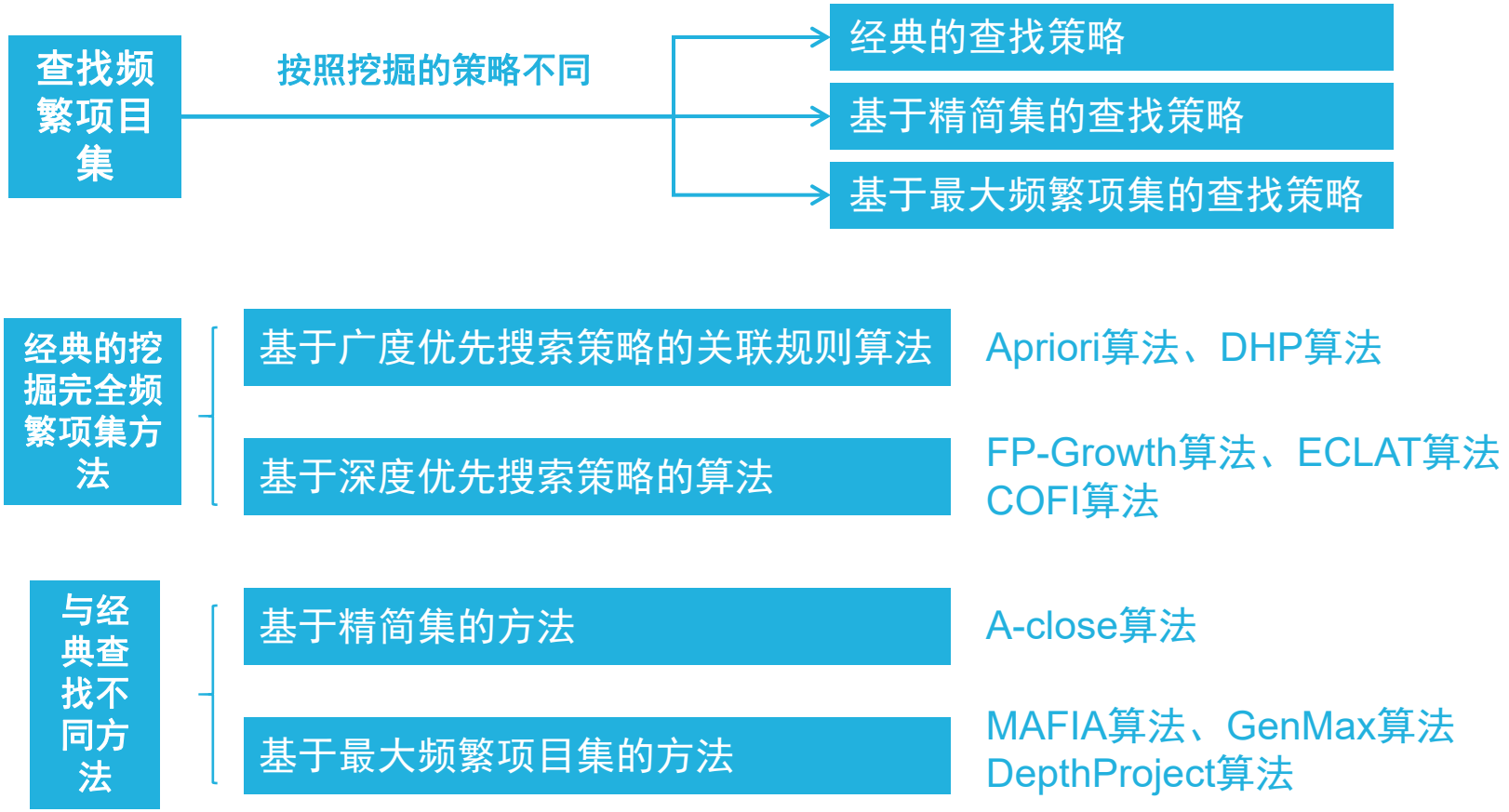
神经网络聚类的两种方法：竞争学习方法与自组织特征图映射方法。神经网络聚类方法存在较长处理时间和复杂数据中复杂关系问题，还不适合处理大数据库。

神经网络方法(SOM算法)

- 特征层处理技术-关联（1）
- 从一个大型的数据集（Dataset）发现有趣的关联（Association）或相关关系（Correlation），即从数据集中识别出频繁出现的属性值集（Sets of Attribute Values），也称为频繁项集（Frequent Itemsets，频繁集），然后利用这些频繁项集创建描述关联关系的规则的过程

发现频繁项集	生成关联规则
发现所有的频繁项集是形成关联规则的基础。通过用户给定的最小支持度，寻找所有支持度大于或等于Minsupport的频繁项集。	通过用户给定的最小可信度，在每个最大频繁项集中，寻找可信度不小于Minconfidence的关联规则。

● 特征层处理技术-关联（2）


查找频繁项目集
按照挖掘的策略不同
经典的查找策略
基于精简集的查找策略
基于最大频繁项集的查找策略
经典的挖掘完全频繁项集方法
基于广度优先搜索策略的关联规则算法
Apriori算法、DHP算法
基于深度优先搜索策略的算法
FP-Growth算法、ECLAT算法
COFI算法
与经典查找不同方法
基于精简集的方法
A-close算法
基于最大频繁项目集的方法
MAFIA算法、GenMax算法
DepthProject算法

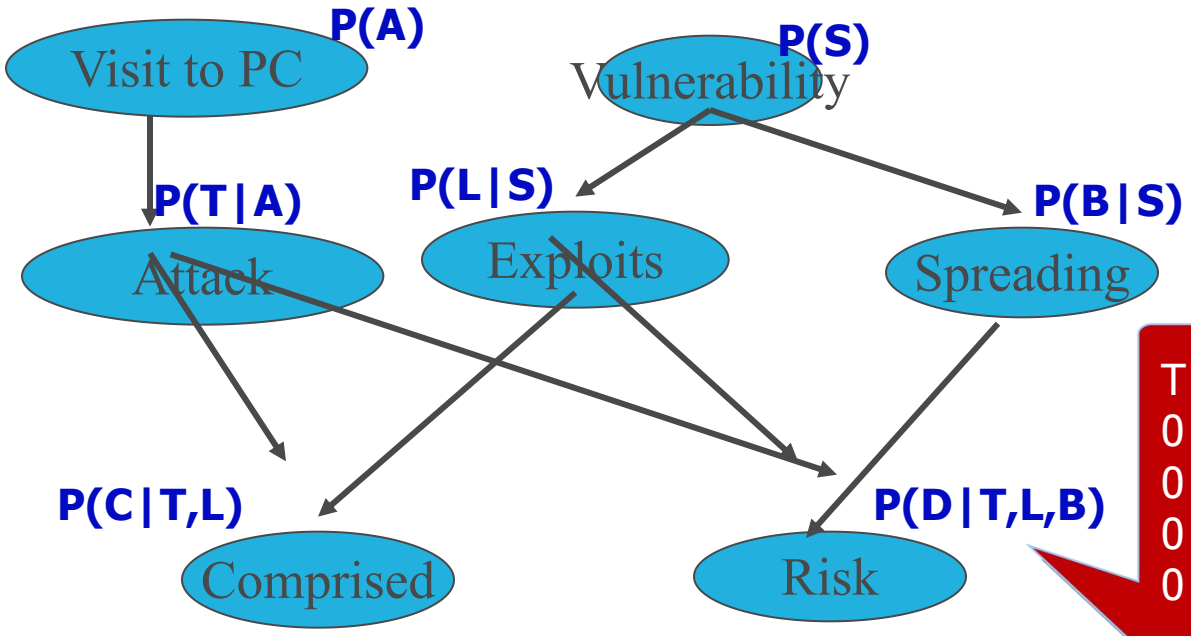
三、安全要素的融合与归一化技术

- 决策层处理技术-贝叶斯方法
  - 贝叶斯学习
  - 贝叶斯网络
  - ...

  提供了一种计算假设概率的方法，基于假设的先验概率、给定假设下观察到不同数据的概率以及观察到的数据本身

- 优势
  - 独特的不确定性知识表达形式
  - 丰富的概率表达能力
  - 综合先验知识的增量学习特性

贝叶斯网络是表示变量间概率依赖关系的有向无环图



$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

先验概率P(h)、P(D)和P(D|h)、后验概率P(h|D)

CPT:

T	L	B	D=0	D=1
0	0	0	0.1	0.9
0	0	1	0.7	0.3
0	1	0	0.8	0.2
0	1	1	0.9	0.1
...				

$$P(A, S, T, L, B, C, D) = P(A)\,P(S)\,P(T|A)\,P(L|S)\,P(B|S)\,P(C|T,L)\,P(D|T,L,B)$$

条件独立性假设 ➡ **有效的表示**

- 决策层处理技术-D-S证据理论（1）
  - 满足比Bayes概率理论更弱的条件，即不必满足概率可加性
  - 具有直接表达“不确定”和“不知道”的能力

Dempster合成规则（Dempster’s combinational rule）也称证据合成公式，其定义如下：

对于$\forall A \subseteq \Theta$，$\Theta$上的两个mass函数$m_1$, m_2的Dempster合成规则为：

$$m_1 \oplus m_2(A) = \frac{1}{K} \sum_{B \cap C = A} m_1(B) \cdot m_2(C)$$

其中， K为**归一化常数**

$$K = \sum_{B \cap C \neq \varnothing} m_1(B) \cdot m_2(C) = 1 - \sum_{B \cap C = \varnothing} m_1(B) \cdot m_2(C)$$

- 决策层处理技术-D-S证据理论（2）

  对于$\forall A \subseteq \Theta$，识别框架$\Theta$上的有限个mass函数$m_1$, m_2, …, m_n的Dempster合成规则为：

$$(m_1 \oplus m_2 \oplus \cdots \oplus m_n)(A) = \frac{1}{K} \sum_{A_1 \cap A_2 \cap \cdots \cap A_n = A} m_1(A_1) \cdot m_2(A_2) \cdots m_n(A_n)$$

其中，

$$K = \sum_{A_1 \cap \cdots \cap A_n \neq \varnothing} m_1(A_1) \cdot m_2(A_2) \cdots m_n(A_n)$$

$$= 1 - \sum_{A_1 \cap \cdots \cap A_n = \varnothing} m_1(A_1) \cdot m_2(A_2) \cdots m_n(A_n)$$

决策层处理技术-决策树方法

构造决策树

根据实际需求及所处理数据的特性，选择类别标识属性和决策树的决策属性集

↓

在决策属性集中选择最有分类标识能力的属性作为决策树的当前决策节点

↓

根据当前决策节点属性取值的不同，将训练样本数据集划分为若干子集

↓

针对上一步中得到的每一个子集，重复进行以上两个步骤，直到最后的子集符合约束的3个条件之一

① 子集中的所有元组都属于同一类。
② 该子集是已遍历了所有决策属性后得到的。
③ 子集中的所有剩余决策属性取值完全相同，已不能根据这些决策属性进一步划分子集。

↓

根据符合条件不同生成叶子节点

修剪决策树

对决策树进行修剪，除去不必要的分枝，同时也能使决策树得到简化。

常用的决策树修剪策略

- 基于代价复杂度的修剪
- 悲观修剪
- 最小描述长度修剪

按照修剪的先后顺序

- 先剪枝（Pre-pruning）
- 后剪枝（Post-pruning）

- 决策层处理技术-知识融合与推理方法
  - 知识融合：将多个数据源抽取的知识进行融合
  - 不同抽取工具通过实体链接和本体匹配可能产生不同的结果，需要考虑本体的融合和实例的融合

模式匹配

模式匹配主要寻找本体中属性和概念之间的对应关系

实例匹配

评估异构知识源之间实例对的相似度，用来判断这些实例是否指向给定领域的相同实体

技术方法：
- 启发式方法
- 概率方法
- 基于图的方法
- 基于学习的方法和
- 基于推理的方法

- 决策层处理技术-知识融合与推理方法
  - 知识推理可以分为基于符号的推理和基于统计的推理
    - 基于符号的推理一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）
    - 基于统计的方法一般指关系机器学习方法，通过统计规律从知识集中学习到新的实体间关系

内容概要

四、未来的挑战

- 数据源
  - 数据采集装置、采集策略的健壮性、完整性：提高源数据质量
- 数据处理方法
  - 自动化的特征抽取、知识学习方法
  - 不确定性数据分析方法的选取和评估
  - 组合方法的使用
- 人机交互
  - 更高效的交互语言、方式：逻辑性更强
- 结果使用
  - 融合结果与安全决策的对应关系


Q&A