网络安全中的数据挖掘

主讲人：张静
jzhang@njust.edu.cn
jz81.github.io

第一章 简介

网络安全事件

- **网络开放性为各类网络安全事件提供了可乘之机**

- 2014年，CNCERT/CC协调处置涉及基础电信企业的漏洞事件1578起，是2013年的3倍。
- 2014年我国境内感染木马僵尸网络的主机为1108.8万余台。
- 2014年针对我国域名系统的流量规模达1Gbit/s以上的拒绝服务攻击事件日均约187起，约为2013年的3倍。
- 2014年通报处置通用软硬件漏洞事件714起，较2013年增长1倍。

其它网络安全事件

- **网络安全是关系国计民生的大问题**

- 国内通用顶级域的根服务器忽然出现异常，导致DNS解析故障
- 比特币交易平台Mt.Gox由于系统漏洞，比特币失窃导致破产
- Heartbleed漏洞波及网银及各大门户网站
- BadUSB漏洞
- Ebay遭遇黑客密码窃取，要求用户全部重置密码
- . . . . . .

目录

网络安全概念

1. 网络安全定义
2. 网络安全面临的挑战
3. 网络安全的重要性

- **网络安全**是指网络系统的硬件、软件及其系统中的数据受到保护，不因偶然的或者恶意的原因而遭受到破坏、更改、泄露，系统连续可靠正常地运行，网络服务不中断。主要强调了保密性、完整性、可用性、可控性、可审查性等主要特性。

网络安全定义

- **网络空间**
- **网络空间安全**

- **网络空间**（Cyberspace）是通过全球互联网和计算系统进行通信、控制和信息共享的动态虚拟空间。
- **网络空间安全**（Cyberspace Security）研究网络空间中的安全威胁和防护问题，即在有攻击者的对抗环境下，研究信息在产生、传输、存储、处理的各个环节中所面临的威胁和防御措施、以及网络和系统本身的威胁和防护机制。

网络安全面临的挑战

- **网络安全面临不同层次、多种多样挑战和威胁**

- 自然威胁（自然灾害、场地环境遭受破坏、设备老化等）；
- 信息泄露（如商业间谍、窃听、流量分析等）；
- 非授权访问（如非授权用户进行入侵）；
- 操作系统缺陷（如操作系统楼梯、后门、I/O非法访问等）；
- 软件漏洞（如数据库的安全漏洞、TCP/IP协议的安全漏洞、网络软件与网络服务的漏洞）；
- 病毒和木马；
- 拒绝服务；
- 甚至还包括网络舆情威胁、网络色情、网络欺诈、网络暴力等

网络安全的重要性

- **习总书记指出：“没有网络安全就没有国家安全”，并要求“加强网络空间安全人才建设，打造素质过硬、战斗力强的人才队伍”。**

- 国际上围绕网络安全的斗争愈演愈烈，夺取网络空间控制权是战略制高点
- 网络安全人才已成为国家竞争的核心所在
- 网络安全技术作用日益彰显
  - 保护个人隐私、
  - 保障经济发展、
  - 维持社会稳定、
  - 保障国家安全

网络空间（信息）安全学科

- 学科概况
- 学科培养目标
- 主要研究方向
- 主要研究内容

- 学科概况
  - “网络空间安全”为“工学”门类下一级学科，学科代码为“0839”，授与“工学”学位。
  - 网络空间由互联互通网络、网络节点和系统及数据组成，可分为物理层、逻辑层和行为体层。
  - 网络空间涉及数学、计算机科学与安全、信息与通信工程等学科，已形成独立教学和研究领域。

学科培养目标

- **通过网络空间安全学科培养，力求让学生**
  - 掌握网络空间安全基础理论和技术方法
  - 掌握信息系统安全、网络基础设施安全、信息内容安全和信息对抗等相关专门知识
  - 能够承担科研院所、企事业单位和行政管理部门网络空间安全方面的科学研究、技术开发及管理工作

主要研究方向

- **安全基础**
- **密码学及应用**
- **系统安全**
- **网络安全**
- **应用安全**

- 为其他方向提供理论、架构和方法学指导
- 为其他方向提供密码体制机制
- 保证网络空间中单元计算系统安全、可信
- 保证连接计算机的网络自身安全和传输信息安全
- 保证网络空间中大型应用系统安全

主要研究内容

- **网络空间安全学科囊括的研究内容包括**

- 可信计算体系、新型密码体制、密码编码与密码分析、网络通信安全、信息安全风险评估、信息安全管理、灾难备份和应急响应、操作系统安全、数据库安全、信息隐藏与检测、内容识别与过滤、信息对抗理论与技术，以及信息安全工程

数据挖掘简介

- 数据挖掘定义
- 数据挖掘流程
- 数据挖掘原因
- 数据挖掘特点

- 数据
  - 人能看到的，听到的，闻到的，能感觉到的事物都是数据
  - 已[illegible]我们人看不见的，听不见的，感觉不到的事物或者关系[illegible]数据，而且很多关键的数据正是隐藏在某些
- 挖掘
  - 一是从众多的数据中提取处[illegible]出有用的数据；
  - 二是从已知的数据中，通过[illegible]的关系来发现总结出隐藏的数据和一般规[illegible]

数据挖掘流程

数据预处理
- ✓ 数据清理
- ✓ 数据集成
- ✓ 数据选择
- ✓ 数据变化

任务相关数据

模式
- ✓ 分类
- ✓ 聚类
- ✓ 密度估计
- ✓ 频繁模式
- ✓ 关联规则

评价
- ✓ 准确率
- ✓ 泛化能力

辅助决策

为什么要进行数据挖掘

- 数据挖掘改变着我们的生活方式
  - 社交软件数据、电子金融数据的增长
  - 云存储技术催化
- 数据挖掘深入各行各业，起着重要作用
  - 商业数据（销售记录、利润和业绩）
  - 医疗数据（诊疗记录，医学图像）
  - 互联网领域（搜索引擎数据）
- 大数据时代的需要
  - 数据指数增量

数据挖掘的特点

- 基于大量数据
- 非平凡性
- 隐含性
- 新奇性
- 价值性

数据量大能全面反映事物本质的

挖掘出的知识应该是不简单的

发现深藏在数据内部的知识，而不是浮于表面的

挖掘出的知识应该是以前未知的

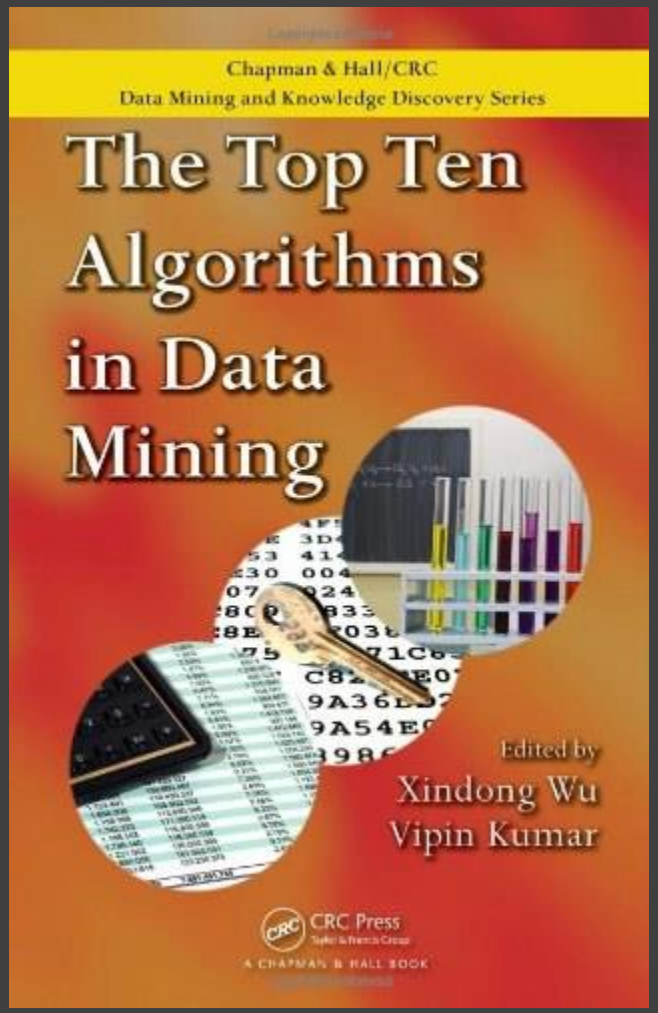
必须能带来直接或间接地效益

数据挖掘算法简介

- **数据挖掘十大算法**
- **国内外数据挖掘发展概况**
- **数据挖掘步骤**

- 数据挖掘十大算法源于ICDM 2006上的一篇论文。
  - 首先，依据引用次数（50次）以上评选出18个算法；
  - 再邀请ACM SIGKDD 2006， IEEE ICDM 2006，SIAM 2006三个国际会议的委员会委员投票选出前十名。

数据挖掘十大算法



排名	算法	简单说明
1	C4.5	决策树分类
2	K-means	K均值聚类
3	Support Vector Machine(SVM)	支持向量机
4	Apriori	关联规则挖掘
5	Expectation Maximization(EM)	最大期望算法
6	PageRank	链接分析
7	AdaBoost	集成算法
8	K-Nearest Neighbors(KNN)	K近邻分类
9	Naive Bayes	朴素贝叶斯分类
10	CART	分类和回归

数据挖掘十大算法

- 十大数据挖掘算法用于处理分类、聚类、关联规则挖掘、概率模型估计和链接分析等任务

排名	算法	简单说明
1	C4.5	决策树分类
2	K-means	K均值聚类
3	Support Vector Machine(SVM)	支持向量机
4	Apriori	关联规则挖掘
5	Expectation Maximization(EM)	最大期望算法
6	PageRank	链接分析
7	AdaBoost	集成算法
8	K-Nearest Neighbors(KNN)	K近邻分类
9	Naive Bayes	朴素贝叶斯分类
10	CART	分类和回归

国外数据挖掘发展概况

- **研究方面：**
  - 数据挖掘算法改进
  - 统计与数据挖掘相结合
- **应用方面**
  - KDD软件由孤立走向系统

- **国外很多知名的软件公司都纷纷加入到数据挖掘工具的研发行列**
- （1）**Knowledge Studio**：由Angoss软件公司开发的能够灵活的导入外部模型和产生规则的数据挖掘工具
- （2）**IBM Intelligent Miner**：自动的实现数据选择、转换、挖掘和结果呈现的一整套数据挖掘操作，支持分类、预测、关联、聚类等算法，并且具有强大的API函数库，可以创建定制的模型
- （3）**SPSS Clem entine**：SPSS是世界上最早的统计分析软件之一，Clem entine是SPSS中的数据挖掘应用工具，它可以把直观的用户图形界面与多种分析技术如神经网络、关联规则和归纳技术结合在一起
- （4）**Cognos Scenario**：该软件是基于树的高度视图化的数据挖掘工具，可以用最短的响应时间得出最精确的结果

国内数据挖掘发展概况

- **国内数据挖掘研究主要集中在高校**

- **我国也有不少新兴的数据挖掘软件**：
- （1）**MSMiner**：有中科院智能信息处理重点实验室开发的多策略通用数据挖掘平台，该平台对数据和挖掘策略的组织有很好的灵活性。
- （2）**DMiner**：由上海复旦德门软件公司开发的自主知识产权的数据挖掘系统，该系统提供了丰富的数据可视化控件来展示分析结果。
- （3）**Scope Miner**：由东北大学开发的面向先进制造业的综合数据挖掘系统。

数据挖掘步骤

1. 问题定义
2. 数据提取
3. 数据预处理
4. 数据挖掘实施
5. 知识表示
6. 结果评估

- 熟悉背景知识，确认需要发现何种知识
- 据目标要求从数据源中提取与挖掘任务相关的数据集
- 检查数据的完整性、剔除数据噪声、数据的一致性处理、对丢失的数据进行填补、数据约简等
- 运用适合的数据挖掘算法进行分析，例如分类、聚类、关联分析、事例推理、决策树、规则推理、模糊集、神经网络、遗传算法等方法，最终得到数据挖掘结果
- 将发现的知识以合理、科学的方法向用户展现
- 数据挖掘结果进行评估分析、发现某种规则、对结果进行优化

Thanks!