atitit.数据挖掘的attilax总结

了数据挖掘，涵盖了五个主题：数据、分类、关联分析、聚类和异常检测

1. 关联规则和序列模式12

第3章 监督学习45

第一部分 数据挖掘基础.  
第1章 概述3  
1.1 什么是万维网3  
1.2 万维网和互联网的历史简述4  
1.3 web数据挖掘5  
1.3.1 什么是数据挖掘6  
1.3.2 什么是web数据挖掘7  
1.4 各章概要8  
1.5 如何阅读本书10  
文献评注10  
第2章 关联规则和序列模式12  
2.1 关联规则的基本概念12  
2.2 apriori算法14  
2.2.1 频繁项目集生成14  
2.2.2 关联规则生成17  
2.3 关联规则挖掘的数据格式19  
2.4 多最小支持度的关联规则挖掘20  
2.4.1 扩展模型21  
2.4.2 挖掘算法22  
2.4.3 规则生成26  
.2.5 分类关联规则挖掘27  
2.5.1 问题描述27  
2.5.2 挖掘算法28  
2.5.3 多最小支持度分类关联规则挖掘31  
2.6 序列模式的基本概念31  
2.7 基于gsp挖掘序列模式32  
2.7.1 gsp算法33  
2.7.2 多最小支持度挖掘342.8.1 prefixspan算法38  
2.8.2 多最小支持度挖掘39  
2.9 从序列模式中产生规则41  
2.9.1 序列规则41  
2.9.2 标签序列规则41  
2.9.3 分类序列规则42  
文献评注42  
第3章 监督学习45  
3.1 基本概念45  
3.2 决策树推理48  
3.2.1 学习算法49  
3.2.2 混杂度函数50  
3.2.3 处理连续属性53  
3.2.4 其他一些问题54  
3.3 评估分类器56  
3.3.1 评估方法56  
3.3.2 查准率、查全率、f-score和平衡点(breakeven point)57  
3.4.1 序列化覆盖59  
3.4.2 规则学习：learn-one-rule函数61  
3.4.3 讨论63  
3.5 基于关联规则的分类63  
3.5.1 使用类关联规则进行分类64  
3.5.2 使用类关联规则作为分类属性66  
3.5.3 使用古典的关联规则分类66  
3.6 朴素贝叶斯分类67  
3.7 朴素贝叶斯文本分类70  
3.7.1 概率框架70  
3.7.2 朴素贝叶斯模型71  
3.7.3 讨论73  
3.8 支持向量机73  
3.8.1 线性支持向量机：可分的情况74  
3.8.2 线性支持向量机：数据不可分的情况78  
3.8.3 非线性支持向量机：核方法80  
3.9 k-近邻学习82  
3.10 分类器的集成83  
3.10.1 bagging83  
3.10.2 boosting84  
文献评注84  
第4章 无监督学习87  
4.1 基本概念87  
4.2 k-均值聚类89  
4.2.1 k-均值算法89  
4.2.2 k-均值算法的硬盘版本91  
4.2.3 优势和劣势92  
4.3 聚类的表示95  
4.3.1 聚类的一般表示方法95  
4.3.2 任意形状的聚类95  
4.4 层次聚类96  
4.4.1 单链接方法97  
4.4.2 全链接方法98  
4.4.3 平均链接方法98  
4.4.4 优势和劣势98  
4.5 距离函数99  
4.5.1 数值的属性(numeric attributes)99  
4.5.2 布尔属性和符号属性(binary and nominal attributes)99  
4.5.3 文本文档101  
4.6 数据标准化101  
4.7 混合属性的处理103  
4.8 采用哪种聚类算法104  
4.9 聚类的评估104  
4.10 发现数据区域和数据空洞106  
文献评注108  
第5章 部分监督学习110  
5.1 从已标注数据和无标注数据中学习110  
5.1.1 使用朴素贝叶斯分类器的em算法111  
5.1.2 co-training114  
5.1.3 自学习115  
5.1.4 直推式支持向量机116  
5.1.5 基于图的方法117  
5.1.6 讨论119  
5.2 从正例和无标注数据中学习119  
5.2.1 pu学习的应用120  
5.2.2 理论基础121  
5.2.3 建立分类器：两步方法122  
5.2.4 建立分类器：直接方法127  
5.2.5 讨论128  
附录：朴素贝叶斯em算法的推导129  
文献评注131  
第二部分 web挖掘  
第6章 信息检索与web搜索135  
6.1 信息检索中的基本概念136  
6.2 信息检索模型138  
6.2.1 布尔模型138  
6.2.2 向量空间模型139  
6.2.3 统计语言模型141  
6.3 关联性反馈142  
6.4 评估标准143  
6.5 文本和网页的预处理147  
6.5.1 停用词移除147  
6.5.2 词干提取147  
6.5.3 其他文本预处理步骤148  
6.5.4 网页预处理步骤148  
6.5.5 副本探测149  
6.6 倒排索引及其压缩150  
6.6.1 倒排索引150  
6.6.2 使用倒排索引搜索151  
6.6.3 索引的建立152  
6.6.4 索引的压缩153  
6.7 隐式语义索引157  
6.7.1 奇异值分解158  
6.7.2 查询和检索159  
6.7.3 实例160  
6.7.4 讨论163  
6.8 web搜索163  
6.9 元搜索引擎和组合多种排序165  
6.9.1 使用相似度分数的合并166  
6.9.2 使用排名位置的合并..166  
6.10 网络作弊168  
6.10.1 内容作弊169  
6.10.2 链接作弊169  
6.10.3 隐藏技术170  
6.10.4 抵制作弊171  
文献评注172  
第7章 链接分析174  
7.1 社会关系网分析175  
7.1.1 中心性175  
7.1.2 权威177  
7.2 同引分析和引文耦合178  
7.2.1 同引分析178  
7.2.2 引文耦合179  
7.3 pagerank179  
7.3.1 pagerank算法180  
7.3.2 pagerank算法的优点和缺点185  
7.3.3 timed pagerank185  
7.4 hits186  
7.4.1 hits算法187  
7.4.2 寻找其他的特征向量189  
7.4.3 同引分析和引文耦合的关系189  
7.4.4 hits算法的优点和缺点189  
7.5 社区发现191  
7.5.1 问题定义191  
7.5.2 二分核心社区192  
7.5.3 最大流社区193  
7.5.4 基于中介性的电子邮件社区195  
7.5.5 命名实体的重迭社区196  
文献评注197  
第8章 web爬取199  
8.1 一个简单爬虫算法199  
8.1.1 宽度优先爬虫201  
8.1.2 带偏好的爬虫201  
8.2 实现议题202  
8.2.1 网页获取202  
8.2.2 网页解析202  
8.2.3 删除无用词并提取词干204  
8.2.4 链接提取和规范化204  
8.2.5 爬虫陷阱206  
8.2.6 网页库206  
8.2.7 并发性207  
8.3 通用爬虫208  
8.3.1 可扩展性208  
8.3.2 覆盖度、新鲜度和重要度209  
8.4 限定爬虫210  
8.5 主题爬虫212  
8.5.1 主题本地性和线索213  
8.5.2 最优优先变种217  
8.5.3 自适应219  
8.6 评价标准223  
8.7 爬虫道德和冲突226  
8.8 最新进展228  
文献评注230  
第9章 结构化数据抽取：包装器生成231  
9.1 预备知识231  
9.1.1 两种富含数据的网页232  
9.1.2 数据模型233  
9.1.3 数据实例的html标记编码235  
9.2 包装器归纳236  
9.2.1 从一张网页抽取237  
9.2.2 学习抽取规则238  
9.2.3 识别提供信息的样例242  
9.2.4 包装器维护242  
9.3 基于实例的包装器学习243  
9.4 自动包装器生成中的一些问题245  
9.4.1 两个抽取问题246  
9.4.2 作为正则表达式的模式246  
9.5 字符串匹配和树匹配247  
9.5.1 字符串编辑距离247  
9.5.2 树匹配249  
9.6 多重对齐252  
9.6.1 中星方法252  
9.6.2 部分树对齐253  
9.7 构建dom树257  
9.8 基于列表页的抽取：平坦数据记录258  
9.8.1 有关数据记录的两个观察结果258  
9.8.2 挖掘数据区域259  
9.8.3 从数据区域中识别数据记录263  
9.8.4 数据项对齐与抽取263  
9.8.5 利用视觉信息264  
9.8.6 一些其他技术264  
9.9 基于列表页的抽取：嵌套数据记录265  
9.10 基于多张网页的抽取269  
9.10.1 采用前几节中的技术270  
9.10.2 roadrunner算法270  
9.11 一些其他问题271  
9.11.1 从其他网页中抽取271  
9.11.2 析取还是可选272  
9.11.3 一个集合类型还是一个元组类型273  
9.11.4 标注与整合273  
9.11.5 领域相关的抽取273  
9.12 讨论274  
文献评注274  
第10章 信息集成276  
10.1 什么是样式表匹配277  
10.2 样式表匹配的预处理工作278  
10.3 样式表层次的匹配279  
10.3.1 基于语言学的算法279  
10.3.2 基于样式表中限制的算法280  
10.4 基于领域和实例层次的匹配280  
10.5 不同相似度的联合282  
10.6 1:m匹配283  
10.7 其他问题284  
10.7.1 重用以前的匹配结果284  
10.7.2 大量样式表的匹配285  
10.7.3 样式表匹配的结果285  
10.7.4 用户交互285  
10.8 web搜索界面的集成285  
10.8.1 基于聚类的算法287  
10.8.2 基于互关系的方法289  
10.8.3 基于实例的方法290  
10.9 构建一个全局的搜索界面292  
10.9.1 结构上的正确性和合并算法293  
10.9.2 词汇的正确性294  
10.9.3 实例的正确性295  
文献评注295  
第11章 观点挖掘296  
11.1 意见分类297  
11.1.1 基于意见短语的分类297  
11.1.2 采用文本分类方法进行意见分类299  
11.1.3 基于评分函数进行分类299  
11.2 基于特征的观点挖掘和摘要300  
11.2.1 问题定义301  
11.2.2 对象特征抽取305  
11.2.3 格式1中正面和负面评价部分的特征抽取306  
11.2.4 符合格式2和3的评审上的特征抽取308  
11.2.5 观点倾向分类309  
11.3 比较性句子和比较关系挖掘310  
11.3.1 问题定义311  
11.3.2 等级比较性语句的识别312  
11.3.3 比较关系的抽取314  
11.4 观点搜索315  
11.5 观点欺诈316  
11.5.1 观点欺诈的目标和行为317  
11.5.2 欺诈和欺诈者的种类317  
11.5.3 隐藏技巧318  
11.5.4 欺诈检测318  
文献评注320  
第12章 web使用挖掘322  
12.1 数据收集和预处理323  
12.1.1 数据的来源和类型323  
12.1.2 web使用记录数据预处理的关键元素326  
12.2 web使用记录挖掘的数据建模331  
12.3 web用法模式的发现和分析334  
12.3.1 会话和访问者分析334  
12.3.2 聚类分析和访问者分割334  
12.3.3 关联及相关度分析337  
12.3.4 序列和导航模式分析340  
12.3.5 基于web用户事务的分类和预测342  
12.4 讨论和展望343  
文献评注344  
参考文献…345

有两本比较基础的书，属于经典教材，当然难度不算深，不过内容比较全面而且讲解很细，适合初学者使用  
1、数据挖掘导论，[美]Pang-Ning Tan，Michael Steinbach，Vipin Kumar 著，译者： 范明 范宏建，人民邮电出版社  
2、数据挖掘：概念与技术，作者：（加）韩家炜，堪博　著，范明，孟小峰　译，机械工业出版社  
  
因为都是中文版，阅读起来基本没有障碍，而且这两个出版社也是计算机领域的传统出版社了，质量还是很能保证的

Web数据挖掘(Bing Liu著) PDF扫描版[46MB] 电子书 下载-脚本之家.htm