填空，选择

Web挖掘的目标是从（Web的超链接结构）（网页内容）（使用日志）中探寻有用的信息

Web挖掘任务的三种类型（Web结构挖掘）（web内容挖掘）（web使用挖掘）

关联规则挖掘的目标是在数据项目中找出所有的（并发关系）

WEB数据挖掘的目标 从WEB超链接、网页内容和使用日志中探寻有用的信息

常见的数据挖掘任务：监督学习（分类），无监督学习（聚类），关联规则挖掘，序列模式挖掘

KDD的三个步骤：（预处理），（数据挖掘），（后续处理）

KDD流程包括数据清理、数据集成、数据选择、转换、数据挖掘、模式评估和知识表示

有监督学习假设：训练示例的分布与测试示例的分布（包括未来看不见的示例）的分布相同。

决策树的思想：分而治之

树以自上而下的递归方式构造的

当数据变得越来越纯净时，熵变得越来越小

过拟合：一棵树可能过拟合训练数据q对训练数据精度好，但对测试数据精度差

症状：树太深，分支太多，有些可能反映了由噪声或异常值引起的异常

两种方法避免过拟合

预修剪：早期停止树构建n很难决定，因为我们不知道如果我们继续生长树会发生什么。

修剪后：从“完全生长”的树上去除树枝或子树。n该方法是常用的方法。C4.5使用一种统计方法来估计每个节点的剪枝误差。n一个验证集也可以用于修剪。

评判标准：精度，效率，鲁棒性，可解释性

评估方法：留出法，交叉验证法，验证集

ROC曲线：受试者工作特征曲线，真阳率比假阳率的曲线。使用它我们希望知道哪个分类器是更好的。

提升曲线：又叫提升图，是关于二分类问题、目标是正例且通常非常稀少这种情况的评估方法。

Precision：查准率：被正确分类的正例数量除以被分类为正例的数据数量

Recall：查全率被正确分类的正例数量除以测试集中实际正例数量。

两种聚类：划分聚类，层次聚类

层次聚类的方法：合并（自上而下）聚类，分裂（自下而上）聚类

EM算法，co-training，自学习是一种LU算法

用户查询的主要形式：关键词查询，布尔查询，短语查询，近邻查询，全文搜索，自然语言查询

信息检索模型：布尔模型（文档表示法，布尔查询，文档检索），向量空间模型，统计语言模型

结构化数据抽取的主要方法：手工方法，包装器归纳，自动抽取

大部分web数据可以被建模成嵌套关系

主要的匹配形式：模式层匹配，域和实例层的匹配，模式、域和实例的综合匹配

模式匹配时的数据预处理包括：分词，扩展，移除停用词和词干提取，单词的标准化

构建统一的全局查询界面要求：结构恰当，词汇恰当，实例恰当

观点的两种主要形式：常规观点，比较性观点

判断

PageRank算法基于web的自然特性

名词解释

无监督学习：有监督学习不是告诉计算机怎么做，而是让机器自己去学习怎样做。无监督学习常用聚类的方法，把相似的东西聚在一起，并且我们并不关心这一类是什么。通常数据无标注

有监督学习：是指从标注数据中学习预测模型的机器学习方法，其本质是学习输入到输出的映射的统计规律。每个实例都是由一个输入对象和一个期望的输出值组成。监督学习算法是分析该训练数据，并产生一个推断的功能，其可以用于映射出新的实例。

部分监督学习：用一组有标记的例子和一组未标记的例子来学习（LU学习），从正例和无标注数据中学习（PU学习）半监督学习是监督学习和无监督学习相互结合的一种学习方法，通过半监督学习的方法可以实现分类、回归、聚类的结合使用。

LU学习：把那些已标引的数据用于训练分类器，把未标引的数据中的有价值信息挖掘出来加以充分利用。每个类别的数据包含少量的已标注数据和大量的无标注数据。目标在于利用无标注的数据来提高学习的质量。

PU学习：从正例和无标注数据中学习。训练数据是由正例和无标注数据组成，其中不包含有反例数据。目的是在不进行反例标注的情况下学到一个精确地分类器。

KDD：数据库知识发现

IR：信息检索

BI：business intelligence [财]商业智能

MIS：最小项目支持度

CAR：分类关联规则挖掘

Precision：查准率：被正确分类的正例数量除以被分类为正例的数据数量

Recall：查全率被正确分类的正例数量除以测试集中实际正例数量。

true positive rate (TPR) ：真阳性率

false positive rate (FPR)：假阳性率

ROC：受试者工作特征曲线

SSE：误差平方和

合并（自上而下）聚类（Agglomerative (bottom up) clustering）：聚类过程从树状图的最底层开始，每一次通过合并最相似的聚类来形成上一层中的聚类。整个过程当全部数据点都合并到一个聚类中时停止。

分裂（自下而上）聚类（Divisive (top down) clustering）：聚类过程从一个包含全部数据点的聚类开始。然后把根节点聚类分裂成一些子聚类。每个子聚类再递归地继续往下分裂直到出现只包含一个数据点的单节点聚类出现

支持向量：距离超平面最近的且满足一定条件的几个训练样本点被称为支持向量

特征空间：原始的输入数据变换到另一个空间，这样变换后的空间中可以用线性决策边界分割正例和负例。

输入空间：原始的数据空间，输入所有可能取值的集合

松弛变量：若所研究的线性规划模型的约束条件全是小于类型，那么可以通过标准化过程引入M个非负的松弛变量。松弛变量的引入常常是为了便于在更大的可行域内求解。若为0，则收敛到原有状态，若大于零，则约束松弛。

惩罚因子：把约束优化问题转化为非受限优化问题。

对称属性：当布尔属性的两个状态有相同的重要性并且具有相同的权重时，称这个布尔属性是对称的。

非对称属性：当一个布尔属性的其中一个状态比其他状态重要时

数据标准化有何作用：能够避免出现聚类结果完全由某个具有很大变化范围的属性主导的情况。

介绍了哪几类数据标准化:区间度量属性（范围标准化，z-score标准化。），比例度量属性（对数函数对每个属性值进行转化），名词性属性，顺序属性

聚类的评估方法：用户验证，真实数据，熵，纯度

混合属性的处理方法有哪几种？

处理混合属性的一种常见方法是决定主导属性类型，然后将其他类型转换为这种类型。

另外一个处理混合属性的方法是首先分别根据各个不同类型的属性计算两个数据点之问的距离，然后把这些不同类型属性的距离结合起来得到一个最终的距离。

数据空洞：如果我们把数据实例当作是r维空间中的点，数据空洞就是指那些不包含或者只包含很少的数据的区域。

信息检索：帮助使用者从大量的数据集信息中发现需要的资料。根据用户的查询信息得到相应的一组文档，得到的结果根据其与用户查询信息的相关程度排序。信息检索研究了信息的获取、组织、存储、检索和分布。

TF：词频率表，文档dj中ti的权值就是在dj中ti出现的次数，被定义为fij，缺点在于，没有考虑如果一个词出现在一个数据集里的许多文档中，那么这个词将没有判别力。

TF-IDF：词逆向文档频率表：最有名的权值表，令N在系统或文件集里的文档的总数，dfi为其中含有至少一次词ti的文档数目，fij为词ti出现在文档dj中的次数。Dj中的ti的正规化词频率

无用词（stop words）：在语言中常常出现，但是可以忽略的单词

词干提取（stemming）简化单词，直到变成它们的词干或者词根

倒排索引：增强了检索文档的性能，最简单的形式是一张列表，包含了每一个不同的词和包含改词的文档列表。

网络作弊：指的是故意误导搜索引擎将某些页面的排名高于其应得的水平。

社会网络分析 （Social Network Analysis）：对于 web 非常有意义，因为实际上 Web 就是一个虚拟的社会，在这个虚拟的社会关系中，每张网页可以被看作是一个参与者，而每个超链接这可以被看作是一个关系。

社会网络：是研究社会实体（一个组织中的人，称为参与者），以及他们的互动和关系。

重要的参与者：广泛与其他参与者连接或者发生关系的参与者。

中心性：是判断网络中节点重要性/影响力的指标，社会网络中，每个节点代表一个参与者，每个链接代表线段两边的两个参与者互相有关系。凭直觉我们就能感觉出参与者i是最中心的参与者因为他/她与其他人的链接最多。在参与者之间，存在着各种不同的链接或者关系。

同引分析：被用来度量两篇论文（出版物）之间的相似性。如果论文i与论文j都被同一篇论文人引用，那么我们就可以说它们两个之间有某种程度上的关系，即便它们互相之问没有直接引用。

引文耦合：度量两篇出版物的相似性，将引用同一篇其他文章的两篇论文联系起来，也就是说，如果论文i和j同时引用了文章人，那么我们就认为它们之间有某种关系，即便它们互相之问不存在直接引用。两篇论文引用的相同文章数目越多，它们之间的相似程度就越高。

悬垂页：没有链出链接的网页。

网页解析：一旦一张网页被下载以后，爬虫就可以解析它的内容（也就是 HTTP Payload 数据）了；并且可以进一步将解析出的内容用来支持调用爬虫的主程序（对于搜索引擎而言，爬得的网页被用来建立索引）和保持爬虫继续不断运行 （将爬得网页中的链接添加到队列中）。

列表页：每个这种页面都含有几个对象列表。在每一个区域中，数据记录根据同一种模板被格式化。

详情页：这种网页侧重于一个对象。注意，当我们说一张网页侧重于一个特定的对象(或者一些对象列表）时，并不表示这张网页不含任何其他信息。事实上，它几乎肯定含有其他信息。对于列表页，通常很容易使用启发式规则来识别主要数据区域。

地标：每一个地标都是一个连续的标志序列，并且被用于定位一个目标项的开头或结尾。

提纯:通过添加更多的终结符，是一个标志或与其匹配的一个通配符，来特化一个析取项。

协同测试：充分利用了通常可以用多种方式抽取同一个数据项这一事实。系统可以学习不同的规则，即前向和后向规则来定位同一个数据项。

多重对齐：一张网页通常包含两个以上的数据记录，所以需要匹配两个以上的字符串或树。产生一个对所有字符串或树的全局的对齐是至关重要的。

信息集成：系统中各子系统和用户的信息采用统一的标准，规范和编码，实现全系统信息共享，进而可实现相关用户软件间的交互和有序工作。

模式匹配：它是指对于两个或者更多个数据库的模式，我们需要在这些模式之问产生映射，把具有相同语义的元素（或属性）映射到一起。这样做的目的是把多个模式整合为一张全局的统一的模式。

简单领域：个值域中的实例值都是单一成分，也就是非合成的。

复合域：一个k元的复合域(composite domain)d是一个有序的k元组，其中第之部分是d 的第i个子域的值，我们用 d;来表示。每个d,都是一个简单域。一个域的元数表示为 arity(d)（=k）。如果一个属性的值域是复合的，那么我们称这个属性是复合的。

观点挖掘：是一种研究领域，用于分析人们对诸如产品，服务，组织，个人，问题，事件，主题及其属性等实体的观点，情绪，评价，评估，态度和情感。

实体：实体e是一个产品、一项服务、一个人、一种事件。。。它与一个序对相关联，e（T,W），T是一个层次化的不见，W是实体e的属性集合。每个部件或子部件也拥有各自的属性

web 使用挖掘是指自动发现和分析模式，这些模式来自于收集的点击流、用户事务数据和因用户与一个或多个网站交互web 资源所产生或者收集其他相关数据。

性化推荐系统：个性化推荐系统是互联网和电子商务发展的产物，它是建立在海量数据挖掘基础上的一种高级商务智能平台，向顾客提供个性化的信息服务和决策支持。

协同过滤：简单来说是利用某兴趣相投、拥有共同经验之群体的喜好来推荐用户感兴趣的信息，个人通过合作的机制给予信息相当程度的回应（如评分）并记录下来以达到过滤的目的进而帮助别人筛选信息，回应不一定局限于特别感兴趣的，特别不感兴趣信息的纪录也相当重要。

算法

Apriori算法基于演绎（apriori）原理（向下封闭属性）

关键思想：频繁项集的子集必须为频繁项集。任何不频繁的项集的超集都不能频繁出现。

向下封闭属性：如果一个项集满足某个小的支持度要求，那么这个项集的任何非空子集必须都满足这个最小支持度。

APRIORI算法：（1）生成所有频繁项目集：一个频繁项目集是一个支持高于minsup的项集。

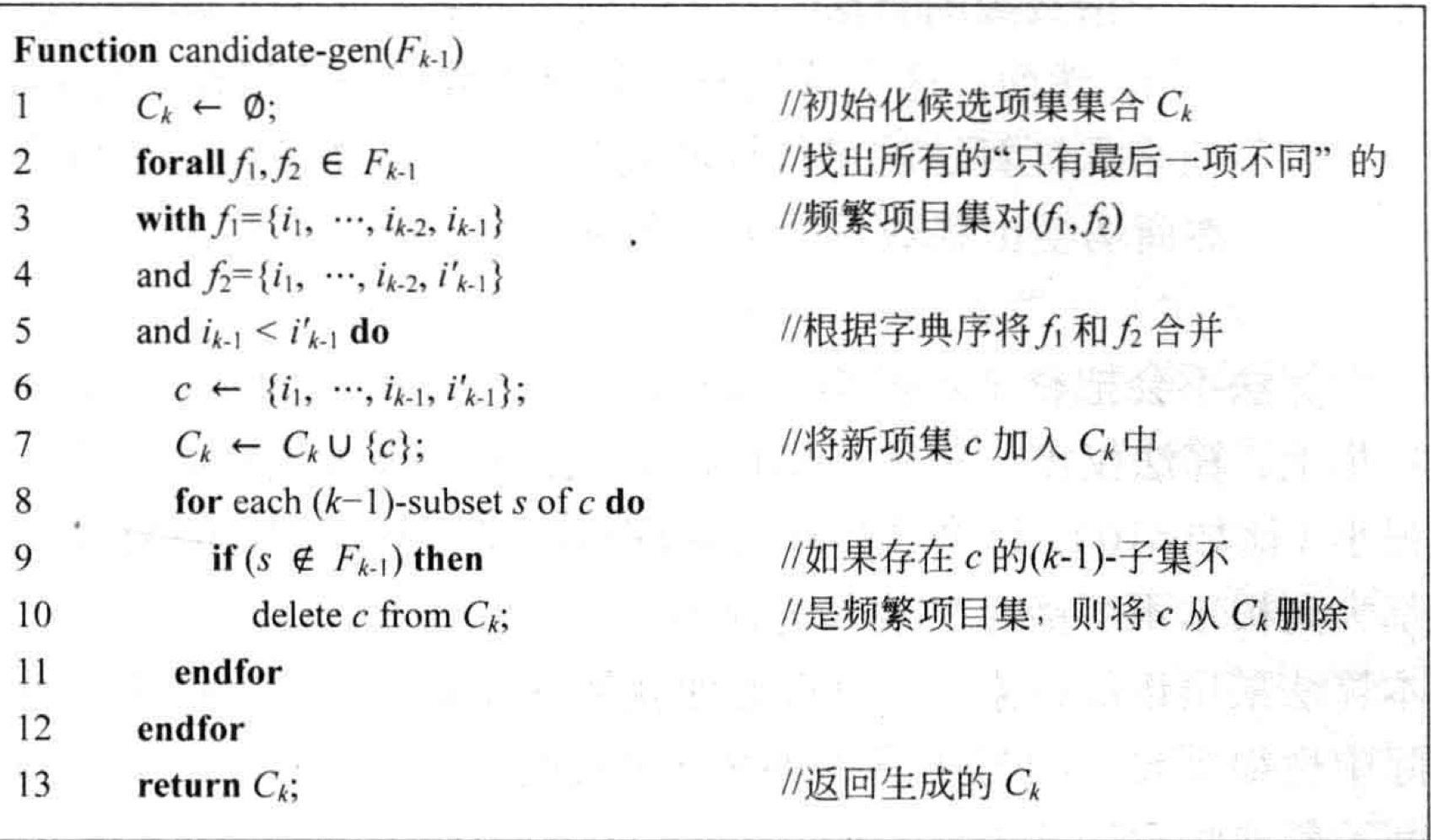
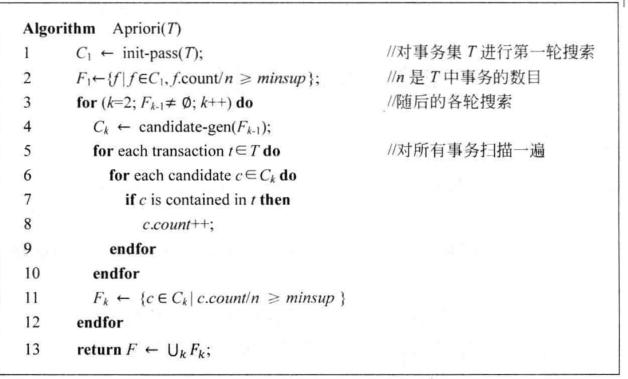
（2）从频繁项目集中生成所有可关联规则：一个可信关联规则是置信度大于minconf的规则

基数：一个项集中项目的个数，基数为k的项集称为k-项集

层次搜索

这个算法的速度很快。在某些条件下，所有的规则都可以在线性时间内找到。

显然，所有关联规则的空间都是指数级的，O（2m），其中m是I中的项目数。

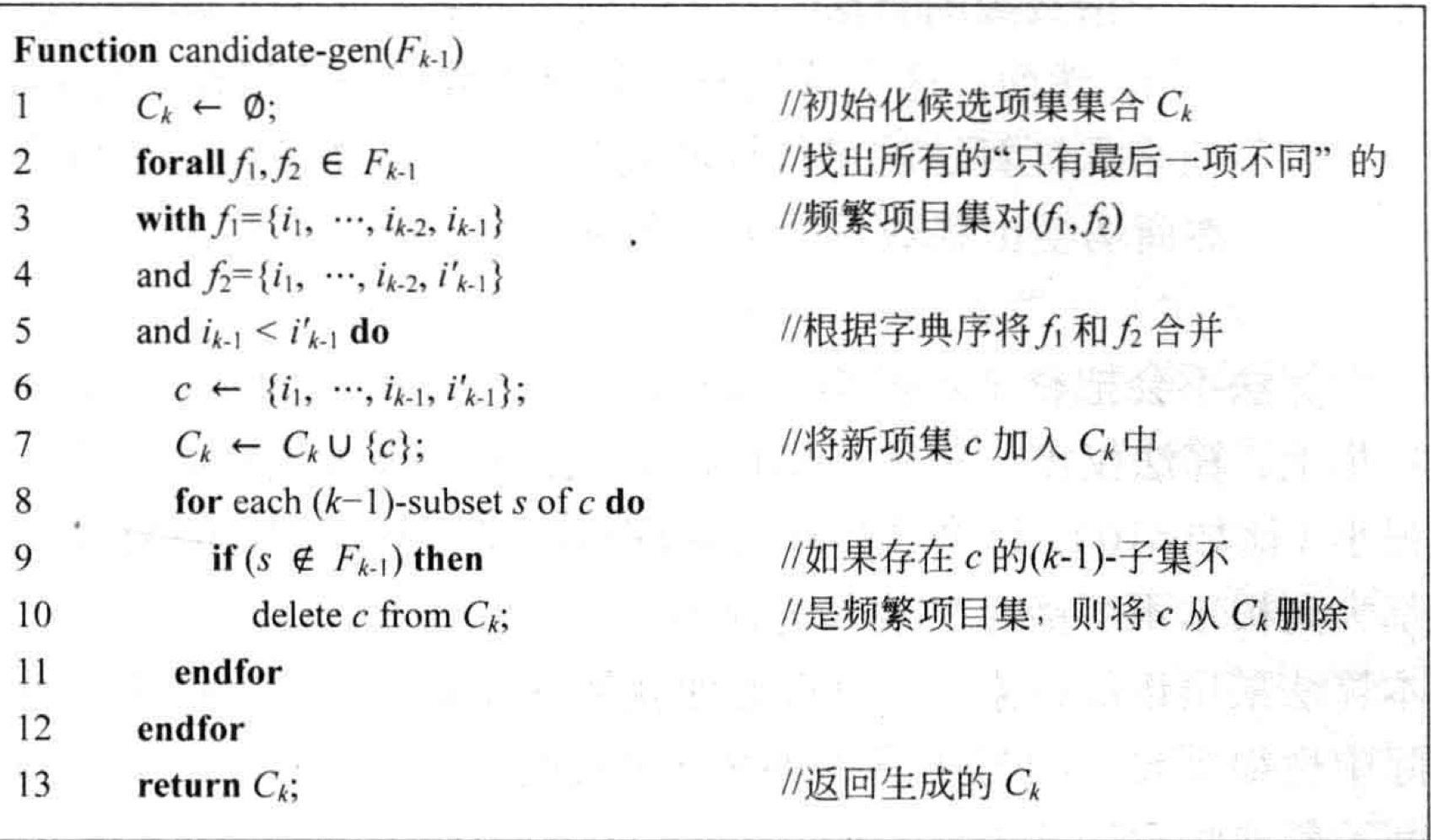


Stepl :令K-1，计算单个商品（项目）的支持度，并筛选出频繁1项集(最小支持度为0.3-2.1/7 )

Step2:(从K=2开始）根据K-1项的频繁项目集生成候选K项目集;并进行预剪枝Step3:由候选K项目集生成频繁K项集(筛选出满足最小支持度的k项集)

重复步骤2和3，直到无法筛选出满足最小支持度的集合。(第一阶段结束)

Step4:将获得的最终的频繁K项集，依次取出。同时计算该次取出的这个K项集的所有真子集，然后以排列组合的方式形成关联规则，并计算规则的置信度以及提升度，将符合要求的关联规则生成提出。(算法结束)



单项最小支持度：它假设数据中的所有项都具有相同的性质和/或具有相似的频率。

不正确：在许多应用程序中，一些项目经常出现在数据中，而其他项目很少出现。

如果项目的频率变化很大，我们就会遇到两个问题

如果最小支持度设置得太高，就不会找到那些涉及稀有物品的规则。

要找到涉及频繁和稀有物品的规则，最小支持度必须设置得很低。这可能会导致组合爆炸，因为这些频繁的项目将以所有可能的方式相互关联。

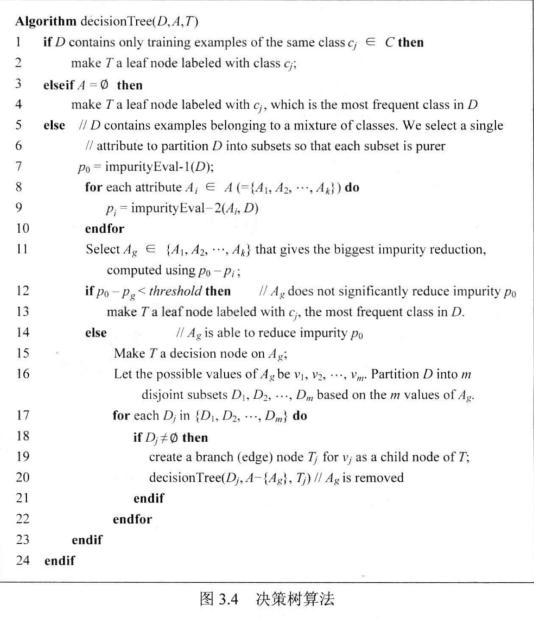
最小项目支持度的问题解决：我们根据I中的MIS值对I中的所有项目进行排序（使其成为一个总顺序）

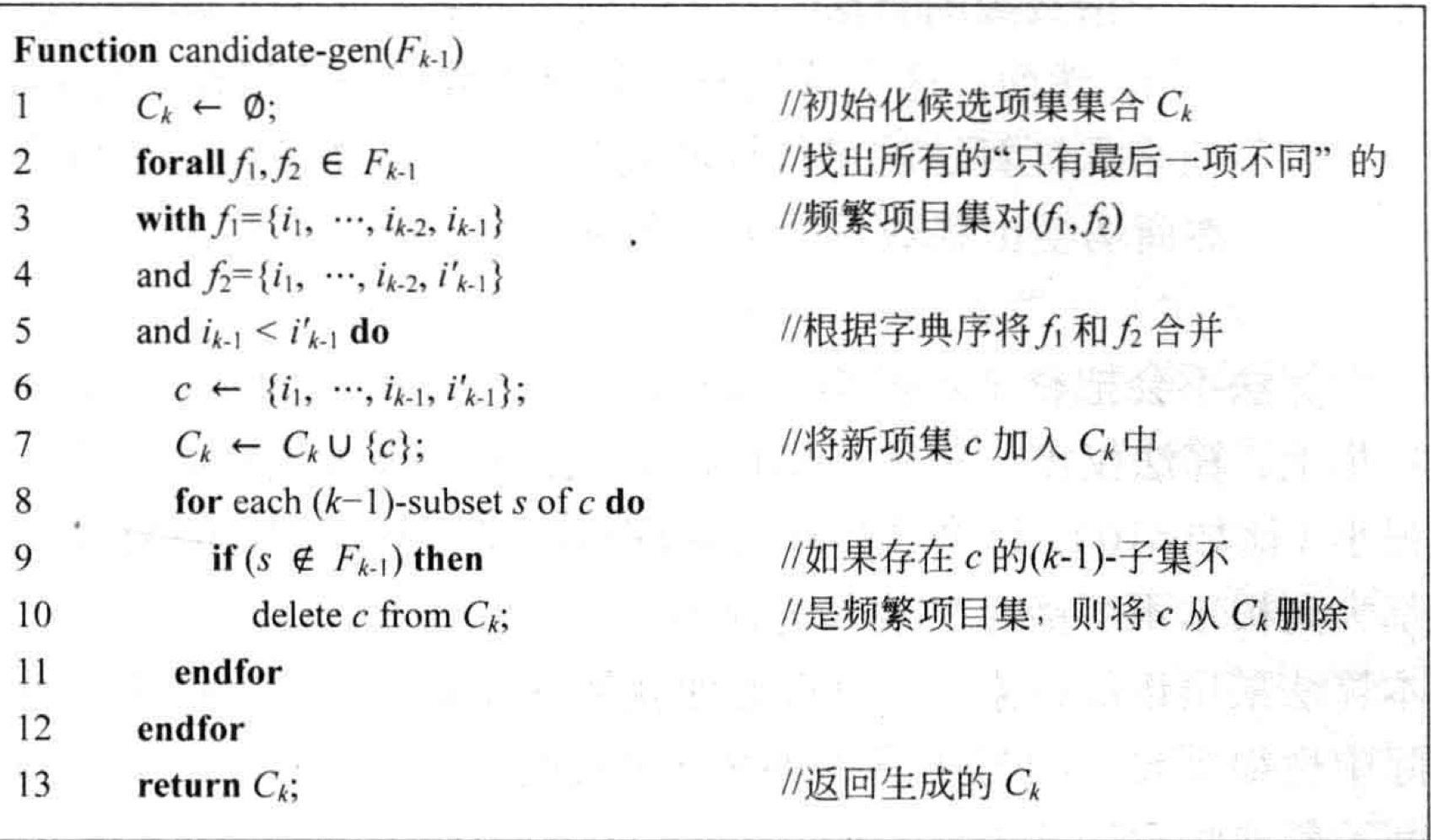
决策树：

当前结点包含的样本全部属于同一类别

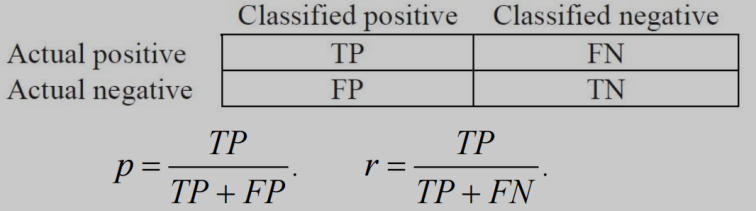
（2）当前属性集为空，或所有样本在所有属性上取值相同

（3）当前结点包含的样本集合为空

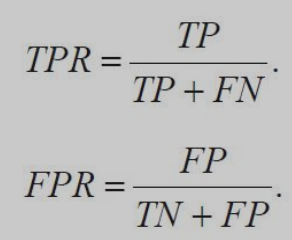




混淆矩阵：



F-score=2pr/p+r



朴素贝叶斯

朴素贝叶斯分类器中最重要的假设是？

条件独立假设。

数值属性：朴素贝叶斯学习假定所有属性都是分类的。数值属性需要被离散化。

优点：易于实现，在许多应用中获得的良好结果

缺点：假设：类条件独立性，因此当严重违反假设时（那些高度相关的数据集）会失去准确性

生成模型基于两个假设：数据文档由一个混合模型生成，混合模型的每一个成分和类别一一对应

混合模型通过以下步骤生成文档d：（1）首先根据类别先验概率（即混合权重）选择一个混合成分/类别

1. 然后根据对应的混合成分生成文档d，由分布决定，

支持向量机是一种线性分类器，它找到超平面来分离正负两类数据。

核函数用于非线性分离。

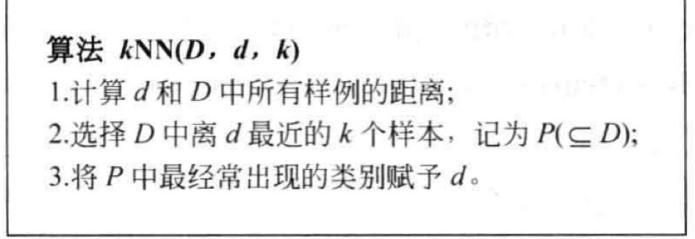
1. 这么多可能的超平面，选择哪一个？

一个的边际幅度最大

1. 我们如何处理需要非线性决策边界的非线性分离或数据集？

使用内核

Knn

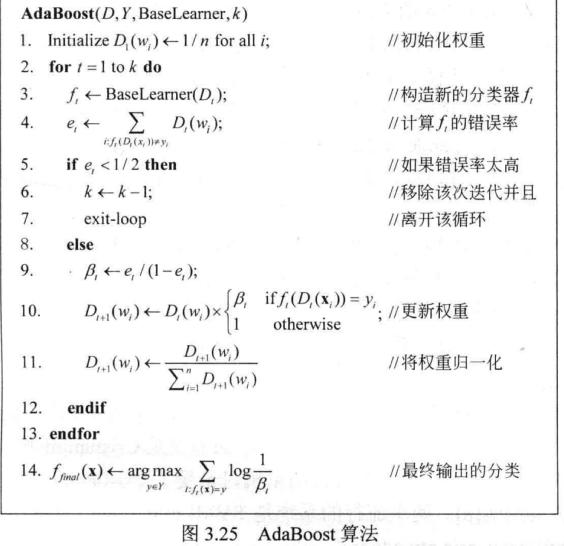


kNN可以处理复杂的和任意的决策边界

kNN不能产生一个可以理解的模型

kNN在分类时速度较慢

Adaboost



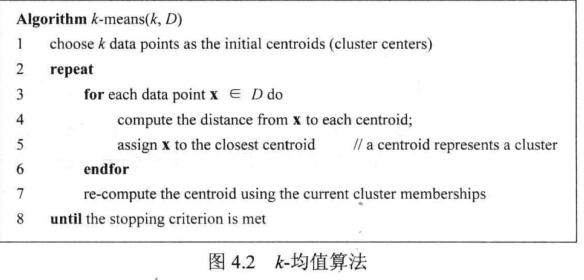
动机：是为了提高单个模型的性能。减少不幸选择一个不良模型的可能性。

聚类：划分聚类，层次聚类

Kmeans聚类

簇内距离最小，簇间距离最大

K-means是一种划分聚类的方法

****

给定k，k-means算法的工作原理如下：1)随机选择k个数据点（种子）作为初始质心，聚类中心，2)将每个数据点分配给最近的质心，3)使用当前的聚类隶属关系重新计算质心。4)如果不满足一个收敛准则，则转到2)。

终止条件：（1）没有（或最小数目）数据点被重新分配给不同的聚类。

（2）没有聚类中心再变化

（3）误差平方和（sse，终止于局部最优）最小

优势：简洁。效率。时间复杂度是O(tkn)其中，n为数据点数，k为簇数，t为迭代次数。

注意：如果使用SSE，它终止于局部最优。由于复杂性，很难找到全局最优值

劣势：该算法仅适用于定义均值时才适用。用户需要指定k。该算法对异常值很敏感

异常值解决方法：一种方法是在聚类过程中删除一些比其他数据点离中心更远的数据点。另一种方法是进行随机抽样。由于在抽样中，我们只选择了数据点的一小个子集，所以选择一个离群值的机会非常小。

表示聚类的常用方法：

(1)使用每个聚类的质心来表示集群。

(2)使用分类模型

(3)利用聚类中最为常见的值来表示聚类

层次聚类的方法：合并（自上而下）聚类，分裂（自下而上）聚类

层次聚类中用来衡量类间距离的方法有哪些？

单连结层次聚类（两个聚类中距离是两个聚类最近的两个数据点之间的距离）

全连结层次聚类（两个聚类中距离是两个聚类中所有数据点之间两两距离的最大值）

平均连结层次聚类（介于全连接防范对于异常值的敏感性和单链接方法形成的长链的趋势的一种折中的方法，两个簇之间的距离是两个簇中数据点之间所有成对距离的平均距离。）

聚类中心法：在该方法中，两个聚类中心之间的距离是它们的聚类中心之间的距离

Ward方法

距离函数：闵可夫斯基距离，欧几里得距离，曼哈顿距离，加权欧几里得距离，平方欧几里得距离，切比雪夫距离

聚类的表示方法：子空间聚类，半监督聚类，谱聚类，协同聚类

聚类的评估方法：用户验证，真实数据，熵，纯度

空洞区域数据作用：空洞区域指一个数据空洞。应用于数字属性的算法也能扩展到离散或者类别属性上去。

EM算法：存在不完整数据的情况下广泛使用的进行最大似然估计的迭代算法。

两个步骤：期望过程（e-step），最大化过程（m-step）

算法大纲1。训练一个只使用已标记的文档的分类器。 2.使用它来对未标记的文档进行概率性分类。 3.使用所有的文档来训练一个新的分类器。 4.迭代步骤2和步骤3，以实现收敛。

IR系统的基本架构和工作原理。

查询操作

索引器

检索系统

要查询信息的用户需要通过查询操作 (Query Operations）模块发送一个查询(用户查询）到检索系统(Retrieval System）。检索模块使用文档索引(Document Index)找到包含这些查询词的文档(这些文档很可能与查询相关），并且计算这些文档的相关度分数，然后根据分数给这些文档排序。经过排序的文档返回给用户。文档数据集 (DocumentCollection）也可以称为文本数据库 (Text Database)，该数据库为了有效地检索已经建立了索引。

文本要根据与查询的相关度排序，方式是jisuanchaxunq与文档数据集中的每个文档dj的相似度，衡量：向量夹角余弦相似度。

信息检索算法的几种评估方法。

查准率：dlq到diq中相关文档的数目比上当前的排名i

查全率：dlq到diq中相关文档的数目比上Rq中相关文档的总数。

平均查准率：使用不同的检索算法的效果，可以使用平均查准率。

PR曲线：根据每个文档的查全率和查准率，可以得到一条查准率-查全率曲线

比较不同算法；

利用多次查询评估：

排名查准率：选择一些排名位置上的文档，计算查准率。

关联反馈的目的：为了加强检索的性能

文本和网页需要做哪些预处理。

无用词移除，词干提取，数字处理，连字号处理，辨别不同的字段，辨别锚文本，移除HTML标签，辨别主要内容块。

词干提取：有效性：提高IR和文本挖掘，匹配相似单词，主要提高召回率，减少索引大小，合并相同根的单词可以使索引大小减少40%-50%

网页预处理步骤：辨别不同的字段，辨别锚文档，移除HTML标签，辨别主要内容块，

倒排索引查找相关文档：（1）搜索词汇表，在词汇表中搜索每一个查询词，词汇表包含所有索引中包含的词。（2）结果合并，在倒排索引中查找到每一项之后，需要合并每一项取得它们的交集。（3）计算排名分数，根据相关度函数给出每个文档计算排名分数/相关度。

网页预处理：辨别不同的文本字段，辨别锚定文本，删除HTML标记，辨别主要内容块，重复检测

网络作弊指的是故意误导搜索引擎将某些页面的排名高于其应得的水平。

内容作弊（content spamming）=词组作弊（term spamming）：标题，元标记，文本，锚文本，网址都可以作弊，重复一些重要的术语，大量添加其他不相关的词

链接作弊：链出链接作弊（加一个对外链出到权威页面中，从而增加一个页面的中心程度），链入链接作弊（比较难成功，很难在其他网页中添加超链接）

作弊器：创建蜜罐，在网络目录中添加链接，在用户生成内容中添加链接，交换链接，自发添加

三种中心性：度中心型，接近中心性（基于接近度/距离），中介中心性

权威：是一种比中心性更精确的衡量演员突出地位的方法。

度量权威的方式：度权威，邻近权威，等级权威（如果一个参与者的影响范围内充满了其他有权威的参与者，那么他自己的权威显然也应该很高。）

PageRank条件：从一个网页指向另一个网页的超链接是一种对目标网站权威性的隐含认可，一个网页的链入连接越多则它的权威就越高。指向网页i的网页本身也有权威值。一个拥有高权威值的网页指向i比一个拥有低权威值的网页指向i更加重要。一个网页被其他重要网页指向，这个网页也重要

Pagerank优缺点：优点：防止作弊的能力，从全局出发的度量以及其非查询相关的特性。

缺点：它不能分辨网页在广泛意义上是权威的还是仅仅在特定的查询话题上是权威的。没有考虑时间。

Hits的基本思想：拥有更高权威值和更高中心值的网页表明它们分别是好的权威页和好的中心页。

Hits的优缺点：HITS 的主要优点是它根据搜索内容来为网页评级，这样它就能提供更加相关的权威页和中心页。这种评级方法也可以结合其他基于信息获取的评级方式。

缺点：1.首先，它没有像 PageRank 那样好的反作弊能力。

2.另外一个问题是话题漂移问题（Topie Drift）。在扩充根集的过程中，该算法很容易将一些与所搜索话题无关的网页（包括中心页与权威页）加入到基集中去。

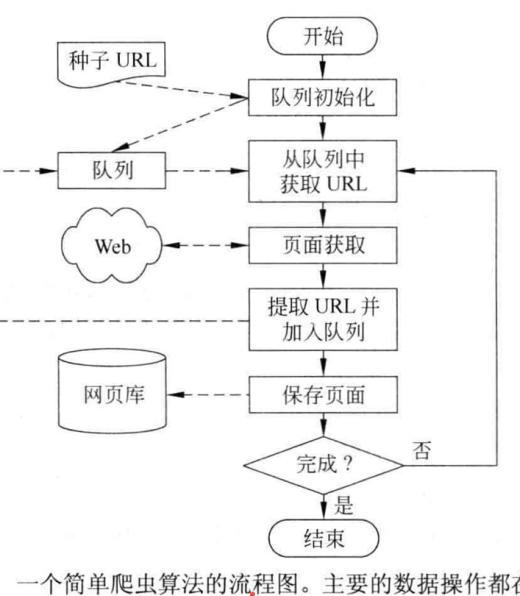
3.搜索时计算也是一个主要的不足之处。寻找根集，扩展根集，然后计算特征向量都是非常花时间的操作。

WEB爬虫：能够自动下载网页的程序。

Web爬虫目的：将多个站点的信息收集起来，并通过在线（网页被下载后）或离线（网页被存储后）的方式，集中进行进一步的分析和挖掘。

Web爬虫作用：商业智能，企业组织可以利用它来收集对手情报以及潜在的合作信息。监视用户感兴趣的网站和网页。

流程：开始-》队列初始化-》从队列中获取url-》页面获取-》提取URL冰加入队列-》保存页面-》完成？-》结束



实现中的议题：网页获取，网页解析，删除无用词并提取词干，链接提取和规范化，爬虫陷阱，网页库，并发性

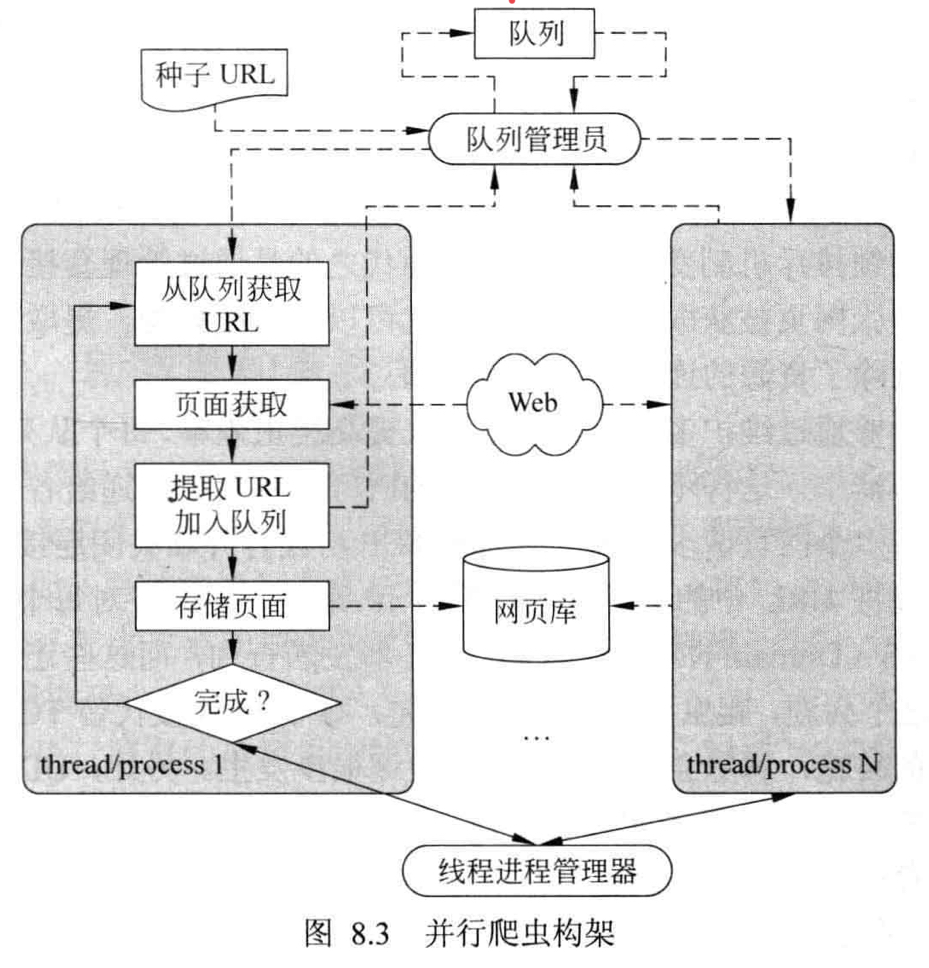
爬虫必须将相对URL转换成绝对URL

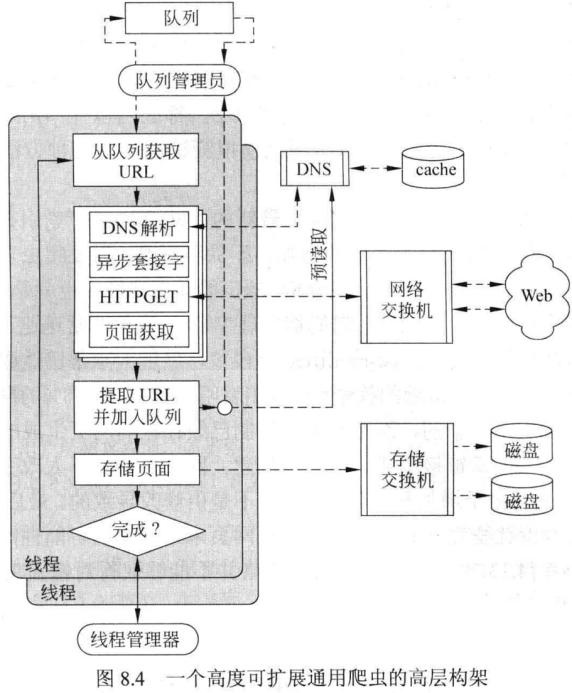
通用爬虫：（1）性能：它们需要能每秒处理成千上万网页抓取工作的规模，这需要在构架上有特别的设计。

（2）策略：它们尽量後盖尽可能多的重要网页，并且尽可能的维持最新状态。这些目标当然是和爬虫原先的均衡设计有所冲突的。

爬虫的主要分类：简单序列爬虫、并发爬虫、通用爬虫（可扩展性，覆盖度，新鲜度，重要性）及主题爬虫，上下文限定爬虫

爬虫实现中的问题：不去取同一个网页两次！队列增长得很快，有一定的长度！网页的获取要鲁棒。确定文件类型以跳过不需要的文件





对爬虫爬取的网页主要的组织方式有哪些形式？各有何利弊？

1）宽度优先性遍历策略(BreathFirst)。

2）OCIP策略(OnlinePageImporteComputation，在线页面重要性计算)。

3）大站优先性策略(LargerSitesFirst)。

1. 衡量爬虫质量的标准有哪些常用的方式？

查全率，查准率，成果率，搜索长度

包装器归纳方法的主要原理是什么？

它是基于有监督学习的。一个包装器归纳系统从一个标注好的训练样例集合中学习数据抽取规则。标注通常是手工完成的，包括标记训练网页或样例中用户希望抽取的数据项。随后学到的规则被用于从其他用相同标记编码或者有相同模板的网页中抽取目标数据。

两种提纯策略：地标提纯，学习析取项

在结构化数据抽取时，主动学习的基本原理是什么？

主动学习是一种帮助自动识别提供信息的未标注样例 的方法。给定一个未标注的样例集 U，在包装器学习中，该方法按如下方式工作：

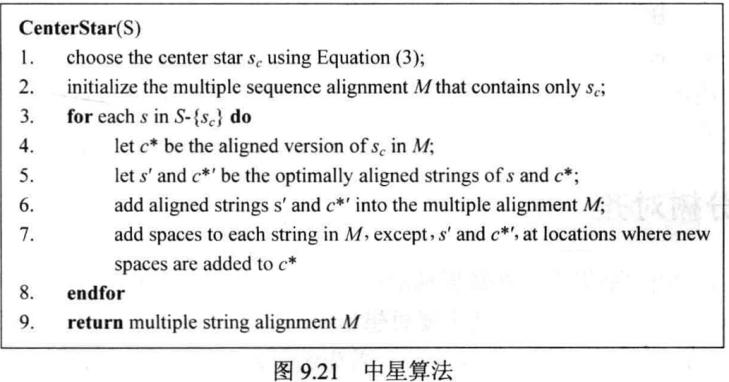
(1)从U中随机选取一个较小的末标注样例子集L；

(2)手工标注工中的样例，并令U=U-L；

(3)基于标注集工学习一个包装器W;

(4)将以应用于U以找到一个提供信息样例的集合L；

(5)如果L为空，则终止，否则跳到第2步。



STM算法的主要原理是什么？试给出算法描述。

STM算法的主要原理是什么？试给出算法描述。

在上述一般情况下，映射可以是跨层次的，例如树 4 中的节点a和树 B中的节点 a。替换也是允许的，例如A 中的节点万和B中的节点h。我们现在定义一种严格的树匹配137]，它叫做简单树匹配 STM，其中不允许节点替换和层次交叉。STM 的目标是找到两棵树间的最大匹配（而不是两棵树的编辑距离)。这一严格的模型己被发现对 Web 数据抽取很有效。

令A和B为两棵树,i和j分别为A和B上的结点.A =(RA,A1,…· ,Am)

B =(RB,B1,……·，Bn)

其中，RA和RB是A和B的根结点，A;和B;分别是A和B的第i个和第j个第1层子树.

(1)当RA和RB的标记相同时, A和B的最大匹配为MAB ＋1,其中MA,B <是A1,A2,…,Am>和<B1,B2,…,Bn>的最大匹配。MA,B可以通过动态规划算法得到。

(2)当RA和RB的标记不相同时，返回0

用法挖掘的主要数据源包括哪些：

Web服务器日志文件（WEB服务器访问日志，应用程序服务器日志），网站内容，关于访问者的数据，从外部通道收集的数据，进一步的应用程序数据

Web使用记录数据预处理的关键步骤：融合和同步数据、数据清理、页面访问识别、用户识别、会话识别、事务识别、数据整合、路径完善

Web使用记录挖掘的主要应用是哪些：产品视图，产品点击通过，购物车的变化，产品购买或投标

解答

KDD流程：

了解应用应用领域

识别数据源并选择目标数据

预处理：清理、属性选择等。

数据挖掘来提取模式或模型

后处理：识别有趣或有用的模式/知识

在现实世界的任务中加入模式/知识

定义问题-》数据收集-》数据准备-》数据建模-》评估-》应用

1. 请简要分析WEB1.0、WEB2.0和WEB3.0的区别,简要说明各自的特点。

Web1.0（1994）：Web1.0是PC互联网，主要是门户网站作为中心，我们用户是被动的去吸收这些信息，他不能互动。

Web2.0（2004）：Web2.0代表了移动互联网，互联网平台它成了中心，也成为了主导。我们用户就可以是双向的互动，存在平台垄断的问题，你的权益得不到保护。

Web3.0（现在）：通常称为价值互联网，它的核心是要去中心化的，或者说它是以用户为中心。但是它的技术的支撑，主要是区块链技术、数字身份认证、数字货币，你提交到网上的作品，所有权属于你自己的，并且价值的权益也是属于你自己的，它是去中心化的。

数据挖掘面临的困难：

数据挖掘算法的效率和可扩展性

并行、分布式、流和增量挖掘方法

处理高维

处理噪声，不确定性，和不完整的数据

结合约束，专家知识，和背景知识在数据挖掘

模式评估和知识集成

挖掘多样化和异构类型的数据：例如。、生物信息学、网络、软件/系统工程、信息网络面向应用程序和特定领域的数据挖掘

隐形数据挖掘（嵌入到其他功能模块中）

数据挖掘中的安全性、完整性和隐私保护

简述下列数据挖掘技术的主要任务：分类、回归、聚类、关联规则挖掘。并举例说明每种技术的应用领域。

分类：分类是一种程序，根据关于一个或多个特征的定量信息（称为变量）和一套以前标记的项目，将单个项目分组。应用：客户流失预测

回归：指把实函数在样本点附近加以近似的有监督的函数近似问题。简单地说，回归是建立因变量关于自变量变化的数学模型的过程。回归分析，就是建立数学模型的分析过程。应用：预测数据

聚类：聚类是将一组观测结果分配到子集（称为聚类），以便同一聚类中的观测在某种意义上是相似的。应用：市场调研

关联规则挖掘：在数据项目中找出所有的并发关系。应用：购物篮数据分析

核函数的作用？

函数的作用简单的说就是由低维度空间向高维度空间作一个映射，使原本线性不可分数据变得在高维度上变得可分，并找到这个分割函数（也并不是所有的线性不可分都能变成可分，能否可分一是要看数据，另一方面要看你找到的核函数是否合适）

异同：

1.HITS算法是与用户输入的查询请求密切相关的，而PageRank与查询请求无关。所以，HITS算法可以单独作为相似性计算评价标准，而PageRank必须结合内容相似性计算才可以用来对网页相关性进行评价；

2.HITS算法因为与用户查询密切相关，所以必须在接收到用户查询后实时进行计算，计算效率较低；而PageRank则可以在爬虫抓取完成后离线计算，在线直接使用计算结果，计算效率较高；

3.HITS算法的计算对象数量较少，只需计算扩展集合内网页之间的链接关系；而PageRank是全局性算法，对所有互联网页面节点进行处理；

4.从两者的计算效率和处理对象集合大小来比较，PageRank更适合部署在服务器端，而HITS算法更适合部署在客户端；

5.HITS算法存在主题泛化问题，所以更适合处理具体化的用户查询；而PageRank在处理宽泛的用户查询时更有优势；

6.HITS算法在计算时，对于每个页面需要计算两个分值，而PageRank只需计算一个分值即可；在搜索引擎领域，更重视HITS算法计算出的Authority权值，但是在很多应用HITS算法的其它领域，Hub分值也有很重要的作用；

7.从链接反作弊的角度来说，PageRank从机制上优于HITS算法，而HITS算法更易遭受链接作弊的影响。

8.HITS算法结构不稳定，当对“扩充网页集合”内链接关系作出很小改变，则对最终排名有很大影响；而PageRank相对HITS而言表现稳定，其根本原因在于PageRank计算时的“远程跳转”。

我们需要爬行整个网络吗？

如果我们覆盖太多，它会变得陈旧。

网页中有大量的页面，声望很低的页面基本上是无用的

目标是什么？一般的搜索引擎：具有很高声望的页面

新闻门户网站：经常发生变化的页面

垂直门户：关于某些主题的页面

1、文档层次、句子层次，以及特称层次的观点挖掘的主要任务分别是什么？

文档层次：评论的情绪分类

句子层次:识别主观的/固执己见的句子，句子的情感分类

特称层次：任务1：识别并提取意见持有人（如审阅人）评论的对象特征。任务2：确定对这些特征的看法是积极、消极的还是中性的。任务3：组特征同义词

1. 什么是比较关系挖掘？比较性句子的比较类型有哪些？比较关系挖掘的基本任务是什么？

比较关系挖掘：直接或间接地表达对某一对象和它的各个方面的正负面观点只是评估的形式之一。将一个对象和其他相似对象进行比较是另一种形式。比较也许是更能让人信服的一种评估方式。比较和普通的观点表述相关但又有所不同。它们有不同的语义和不同的句法形式。比较可以是主观的也可以是客观的。

类型：（1）不相等的等级比较（2）相等的比较（3）最高级比较（4）非等级比较

基本任务：挖掘比较性观点的目标：给定一组带有观点的文档D，在D中发现所有关于比较性观点的六元组：(E1, E2, A, PE, h, t)