数据仓库与数据挖掘习题答案

## 第 1 章 数据仓库的概念与体系结构

1. 面向主题的，相对稳定的。
2. 技术元数据，业务元数据。
3. 联机分析处理 OLAP。
4. 切片（Slice），钻取（Drill-down 和 Roll-up 等）。
5. 基于关系数据库。
6. 数据抽取，数据存储与管理。
7. 两层架构，独立型数据集市，依赖型数据集市和操作型数据存储，逻辑型数据集市和实时数据仓库。
8. 可更新的，当前值的。
9. 接近实时。
10. 以报表为主，以分析为主，以预测模型为主，以营运导向为主。
11. 答： 数据仓库就是一个面向主题的（Subject Oriented）、集成的（Integrate）、相对稳定的（Non-Volatile）、反映历史变化（Time Variant）的数据集合，通常用于辅助决策支持。 数据仓库的特点包含以下几个方面： （1） 面向主题。操作型数据库的数据组织是面向事务处理任务，各个业务系统之间各自分离；而数据仓库中的数据是按照一定的主题域进行组织。主题是一个抽象的概念，是指用户使用数据仓库进行决策时所关心的重点领域，一个主题通常与多个操作型业务系统或外部档案数据相关。 （2）集成的。面向事务处理的操作型数据库通常与某些特定的应用相关，数据库之间相互独立，并且往往是异构的。而数据仓库中的数据是在对原有分散的数据库数据作抽取、清理的基础上经过系统加工、汇总和整理得到的，必须消除源数据中的不一致性，以保证数据仓库内的信息是关于整个企事业单位一致的全局信息。也就是说存放在数据仓库中的数据应使用一致的命名规则、格式、编码结构和相关特性来定义。 （3）相对稳定的。操作型数据库中的数据通常实时更新，数据根据需要及时发生变化。数据仓库的数据主要供单位决策分析之用，对所涉及的数据操作主要是数据查询和加载，一旦某个数据加载到数据仓库以后，一般情况下将作为数据档案长期保存，几乎不再做修改和删除操作，也就是说针对数据仓库，通常有大量的查询操作及少量定期的加载（或刷新）操作。 （4）反映历史变化。操作型数据库（OLTP）主要关心当前某一个时间段内的数据，而数据仓库中的数据通常包含较久远的历史数据，因此总是包括一个时间维，以便可以研究趋势和变化。数据仓库系统通常记录了一个单位从过去某一时点(如开始启用数据仓库系统的时点)到目前的所有时期的信息，通过这些信息，可以对单位的发展历程和未来趋势做出定量分析和预测。
12. 答： （1）两层架构（Generic Two-Level Architecture）。（2）独立型数据集市（Independent Data Mart）。 （3）依赖型数据集市和操作型数据存储（Dependent Data Mart and Operational Data Store）。 （4）逻辑型数据集市和实时数据仓库（Logical Data Mart and Real-Time Data Warehouse）。
13. 答： 数据仓库技术的发展包括数据抽取、存储管理、数据表现和方法论等方面。在数据抽取方面，未来的技术发展将集中在系统集成化方面。它将互连、转换、复制、调度、监控纳入标准化的统一管理，以适应数据仓库本身或数据源可能的变化，使系统更便于管理和维护。在数据管理方面，未来的发展将使数据库厂商明确推出数据仓库引擎，作为数据仓库服务器产品与数据库服务器并驾齐驱。在这一方面，带有决策支持扩展的并行关系数据库将最具发展潜力。在数据表现方面，数理统计的算法和功能将普遍集成到联机分析产品中，并与 Internet/Web 技术紧密结合。按行业应用特征细化的数据仓库用户前端软件将成为产品作为数据仓库解决方案的一部分。数据仓库实现过程的方法论将更加普及，将成为数据库设计

的一个明确分支，成为管理信息系统设计的必备。

14. 答：　（1）IBM 公司提供了一套基于可视化数据仓库的商业智能（BI）解决方案，包括：Visual Warehouse（VW）、Essbase/DB2 OLAP Server 5.0、IBM DB2 UDB，以及来自第三方的前端数据展现工具（如 BO）和数据挖掘工具（如 SAS）。其中，VW 是一个功能很强的集成环境，既可用于数据仓库建模和元数据管理，又可用于数据抽取、转换、装载和调度。Essbase/DB2 OLAP Server 支持“维”的定义和数据装载。Essbase/DB2 OLAP Server 不是 ROLAP（Relational OLAP）服务器，而是一个（ROLAP 和 MOLAP）混合的 HOLAP 服务器，在 Essbase 完成数据装载后，数据存放在系统指定的 DB2 UDB 数据库中。它的前端数据展现工具可以选择 Business Objects 的 BO、Lotus 的 Approach、Cognos 的 Impromptu 或 IBM 的 Query Management Facility；多维分析工具支持 Arbor Software 的 Essbase 和 IBM（与 Arbor 联合开发）的 DB2 OLAP 服务器；统计分析工具采用 SAS 系统。　（2）Oracle 数据仓库解决方案主要包括 Oracle Express 和 Oracle Discoverer 两个部分。Oracle Express 由四个工具组成：Oracle Express Server 是一个 MOLAP(多维 OLAP)服务器，它利用多维模型，存储和管理多维数据库或多维高速缓存，同时也能够访问多种关系数据库；Oracle Express Web Agent 通过 CGI 或 Web 插件支持基于 Web 的动态多维数据展现；Oracle Express Objects 前端数据分析工具（目前仅支持 Windows 平台）提供了图形化建模和假设分析功能，支持可视化开发和事件驱动编程技术，提供了兼容 Visual Basic 语法的语言，支持 OCX 和 OLE；Oracle Express Analyzer 是通用的、面向最终用户的报告和分析工具（目前仅支持 Windows 平台）。Oracle Discoverer 即席查询工具是专门为最终用户设计的，分为最终用户版和管理员版。　在 Oracle 数据仓库解决方案的实施过程中，通常把汇总数据存储在 Express 多维数据库中，而将详细数据存储在 Oracle 关系数据库中，当需要详细数据时，Express Server 通过构造 SQL 语句访问关系数据库。

（3）Microsoft 将 OLAP 功能集成到 SQL Server 数据库中，其解决方案包括 BI 平台、BI 终端工具、BI 门户和 BI 应用四个部分，如图 1.1。　① BI 平台是 BI 解决方案的基础，包括 ETL 平台 SQL Server 2005 Integration Service(SSIS)、数据仓库引擎 SQL Server 2005 RDBMS 以及多维分析和数据挖掘引擎 SQL Server 2005 Analysis Service、报表管理引擎 SQL Server 2005 Reporting Service。　② BI 终端用户工具，用户通过终端用户工具和 Analysis Service 中的 OLAP 服务和数据挖掘服务进行交互来使用多维数据集和数据挖掘模型，终端用户通常可使用预定义报表、交互式多维分析、即席查询、数据可视化、数据挖掘等多种方法。　③ BI 门户提供了各种不同用户访问 BI 信息的统一入口。BI 门户是一个数据的汇集地，集成了来自不同系统的相关信息。用户可以制定个性化的个人门户，选择和自己相关性最强的数据，提高信息访问和使用的效率。　④ BI 应用是建立在 BI 平台、BI 终端用户工具和 BI 统一门户这些公共技术手段之上的满足某个特定业务需求的应用，例如零售业务分析、企业项目管理组合分析等。

## 第 2 章　数据仓库的数据存储与处理

1. 企业级数据仓库（EDW）。
2. 单一的，详细的。
3. 最初填充数据仓库。
4. 越高，越低，越多。
5. 提高，预处理，事实表。
6. 自然键（Natural Key），代理键（Surrogate Key）。
7. 星型模式。
8. 早期细节级，轻度综合级。

9. 答：　简单地说，数据是从企业内外部的各业务处理系统（操作型数据）流向企业级数据仓库（EDW）或操作型数据存储区(ODS)，在这个过程中，要根据企业（或其他组织）的数据模型和元数据库对数据进行调和处理，形成一个中间数据层，然后再根据分析需求，从调和数据层（EDW、ODS）将数据引入导出数据层，如形成满足各类分析需求的数据集市。
10. 答：　数据的 ETL 过程就是负责将操作型数据转换成调和数据的过程。如上面的 2.3.1 小节所述，这两种数据具有明显的区别，因此，数据调和是构建一个数据仓库中最难的和最具技术挑战性的部分。在为企业级数据仓库填充数据的过程中，数据调和可分为两个阶段：一是企业级数据仓库（EDW）首次创建时的原始加载；二是接下来的定期修改，以保持 EDW 的当前有效性和扩展性。　整个过程由四个步骤组成：抽取、清洗、转换、加载和索引。事实上，这些步骤可以进行不同的组合，如，可以将数据抽取与清洗组合为一个过程，或者将清洗和转换组合在一起。通常，在清洗过程中发现的拒绝数据信息会送回到源操作型业务系统中，然后将数据在源系统中加以处理，以便在以后重新抽取。
11. 答：　在星模式中，事实表居中，多个维表呈辐射状分布于其四周，并与事实表连接。位于星形中心的实体是事实表，是用户最关心的基本实体和查询活动的中心，为数据仓库的查询活动提供定量数据。位于星模式四周的实体是维度实体，其作用是限制和过滤用户的查询结果，缩小访问范围。每个维表都有自己的属性，维表和事实表通过关键字相关联。
12. 答：　因为数据仓库或数据集市的数据总是历史的数据，需要时间维来区别。

## 第 3 章　数据仓库系统的设计与开发

1. 在线分析处理(OLAP)　分析。
2. 信息包图法，维度，类别，度量。
3. 逻辑模型。
4. 事务事实，快照事实，线性项目事实。
5. 聚合。
6. 时间，区域。
7. 退化维。
8. 无变化，缓慢变化，剧烈变化。
9. 索引。
10. 反向规范化，引入冗余。
11. 答：　信息包图法，也叫用户信息需求表，就是在一张平面表格上描述元素的多维性，其中的每一个维度用平面表格的一列表示，通常的维度如时间、地点、产品和顾客等；而细化本列的对象就是类别，例如时间维度的类别可以细化到年、月、日，甚至小时；平面表格的最后一行（代表超立方体中的单元格）即为指标度量值，例如，某年在某销售点的某类产品的实际销售额。创建信息包图时需要确定最高层和最低层的信息需求，以便最终设计出包含各个层次需要的数据仓库　总之，信息包图法是一种自上而下的数据建模方法，即从用户的观点开始设计（用户的观点是通过与用户交流得到的），站在管理者的角度把焦点集中在企业的一个或几个主题上，着重分析主题所涉及数据的多维特性，这种自上而下的方法几乎考虑了所有的信息源，以及这些信息源影响业务活动的方式。
12. 答：　收集、分析和确认业务分析需求，分析和理解主题和元数据、事实及其量度、粒度和维度的选择与设计、数据仓库的物理存储方式的设计等。
13. 答：　（1）收集和分析业务需求；　（2）建立数据模型和数据仓库的物理设计；　（3）定义数据源；　（4）选择数据仓库技术和平台；　（5）从操作型数据库中抽取、清洗及转换数据到数据仓库；　（6）选择访问和报表工具，选择数据库连接软件，选择数据分析和数据展示软件；　（7）更新数据仓库。

14. 答：参考 3.3 节的过程。

## 第 4 章 关联规则

1. apriori，fp-growth，fp-growth。
2. {{abc}{abd}{acd}}，{{abc}{abd}}。
3. {{a}{b}{c}}，{ac}。
4. 答： 关联规则挖掘最初由 R.Agrawal 等人提出，用来发现超级市场中用户购买的商品之间的隐含关联关系，并用规则的形式表示出来，称为关联规则(Association Rule)。 关联规则除了可以发现超市购物中隐含的关联关系之外，还可以应用于其他很多领域。关联规则的应用还包括文本挖掘、商品广告邮寄分析、网络故障分析等。
5. 答： 关联规则的分类：

（1）基于规则中涉及到的数据的维数，关联规则可以分为单维的和多维的。

（2）基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则。

（3）基于规则中处理的变量的类型不同，关联规则可以分为布尔型和数值型。

关联规则挖掘的步骤：

（1）找出交易数据库中所有大于或等于用户指定的最小支持度的频繁项集；

（2）利用频繁项集生成所需要的关联规则，根据用户设定的最小可信度进行取舍，产生强关联规则。

6. 答：规则：c=>a，a=>c。
7. 答：

| 项 | 条件模式库 | 条件FP-tree |
|---|---|---|
| p | {(fcam:2), (cb:1)} | {(c:3)}\|p |
| m | {(fca:2), (fcab:1)} | {(f:3, c:3, a:3)}\|m |
| b | {(fca:1), (f:1), (c:1)} | Empty |
| a | {(fc:3)} | {(f:3, c:3)}\|a |
| c | {(f:3)} | {(f:3)}\|c |
| f | Empty | Empty |

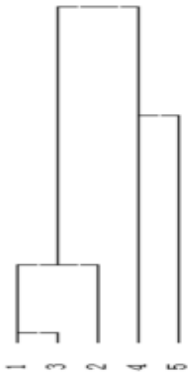
## 第 5 章 数据分类

1. 获取数据，预处理，分类器设计，分类决策。
2. 划分数据集，分类器构造，分类器测试。
3. 精确度，查全率和查准率，F-measure，几何均值。
4. 多项式核函数，径向基核函数，S 型核函数。
5. 答： 分类是指把数据样本映射到一个事先定义的类中的学习过程，即给定一组输入的属性向量及其对应的类，用基于归纳的学习算法得出分类。 分类问题是数据挖掘领域中研究和应用最为广泛的技术之一，许多分类算法被包含在统计分析工具的软件包中，作为专门的分类工具来使用。分类问题在商业、银行业、医疗诊断、生物学、文本挖掘、因特网筛选等领域都有广泛应用。例如，在银行业中，分类方法可以辅助工作人员将正常信用卡用户和欺诈信用卡用户进行分类，从而采取有效措施减小银行的损失；在医疗诊断中，分类方法可以帮助医疗人员将正常细胞和癌变细胞进行分类，从而及时制定救治方案，挽救病人的生命；在因特网筛选中，分类方法可以协助网络工作人员将正常邮件和垃圾邮件进行分类，从而制定有效的垃圾邮件过滤机制，防止垃圾邮件干扰人们的正常生活。
6. 答：求解过程请参考例 5.1。

7. 答： 计算 x 与 x1~x7 的欧氏距离，可以知道 x 的最近邻是 x4，x 的前 3 个近邻是 x4，x1，x2，所以，利用最近邻分类方法对 x 进行分类时 x 的类标号是 y=-1，利用 k-近邻分类方法（k=3）对 x 进行分类时 x 的类标号是 y=+1。

## 第 6 章 数据聚类

1. 连续型，二值离散型，多值离散型，混合类型。
2. 欧氏距离，曼哈顿距离，明考斯基距离。
3. 选定某种距离作为数据样本间的相似性度量，选择评价聚类性能的准则函数，选择某个初始分类，之后用迭代的方法得到聚类结果，使得评价聚类的准则函数取得最优值。
4. 凝聚型层次聚类，分解型层次聚类。
5. 答： 聚类分析是将物理的或者抽象的数据集合划分为多个类别的过程，聚类之后的每个类别中任意两个数据样本之间具有较高的相似度，而不同类别的数据样本之间具有较低的相似度。 聚类分析是数据挖掘应用的主要技术之一，它可以作为一个独立的工具来使用，将未知类标号的数据集划分为多个类别之后，观察每个类别中数据样本的特点，并且对某些特定的类别作进一步的分析。此外，聚类分析还可以作为其他数据挖掘技术（例如分类学习、关联规则挖掘等）的预处理工作。聚类分析在科学数据分析、商业、生物学、医疗诊断、文本挖掘、Web 数据挖掘等领域都有广泛应用。在科学数据分析中，比如对于卫星遥感照片，聚类可以将相似的区域归类，有助于研究人员根据具体情况做进一步分析；在商业领域，聚类可以帮助市场分析人员对客户的基本数据进行分析，发现购买模式不同的客户群，从而协助市场调整销售计划；在生物学方面，聚类可以帮助研究人员按照基因的相似度对动物和植物的种群进行划分，从而获得对种群中固有结构的认识；在医疗诊断中，聚类可以对细胞进行归类，有助于医疗人员发现异常细胞的聚类，从而对病人及时采取措施；在文本挖掘和 Web 数据挖掘领域中，聚类可以将网站数据按照读者的兴趣度进行划分，从而有助于网站内容的改进。
6. 答：参考图 6.1。
7. 答：参考 6.4.2 节的步骤。
8. 答：根据给定的数据集，设定数据样本之间的距离采用欧氏距离，聚类集合之间的相似性度量采用最小距离，聚类结果如下图所示。



## 第 7 章 贝叶斯网络

1. 答： 由 SA 发生得知，HO 发生/不发生的概率为 0.56566/0.43434；由 PX 发生得知，BT 发生/不发生的概率为 0.0891/0.9109。根据表 7.6 中给出的联合条件概率分布，可得 HA 发生/不发生的概率是 0.4533/0.5467。再根据条件概率公式，P(+BT|+HA) = P(+HA|+BT) P(+BT) / P(+HA) = 0.9509 × 0.0891 / 0.4533 = 0.1869。

2. 答：　事实上，从“有酒精味”到“患脑瘤”之间没有任何的因果关系，也就是说，“有酒精味”并不能影响到脑瘤的产生。但是，“有酒精味”能够影响“患脑瘤”的诊断概率。这是因为，“有酒精味”使得引起“头疼”的更多因素归为喝酒，而不是患脑瘤，从而使得“患脑瘤”的概率大大降低。

3. 答：贝叶斯网络的 3 个主要议题分别是：预测、诊断和历史数据训练。

4. 答：　要训练条件概率 P(B|A)，可以在历史数据中统计 A 发生的次数 T(A)，然后统计在 A 发生的数据中 B 发生的次数 T(A,B)，条件概率 P(B|A) = T(B)/T(A)。要训练联合条件概率 P(C|A,B)，可以在历史数据中统计 A、B 共同发生的次数 T(A,B)，然后在 A、B 共同发生的数据中统计 C 发生的次数 T(A,B,C)，联合条件概率 P(C|A,B)=T(A,B,C)/T(A,B)。以上的符号 A、B、C 可以表示某个事件，也可以表示该事件的相反事件。

5. 答：　可以用两种方式从历史数据中得到各个节点的发生概率：（1）用各节点的发生次数除以总的数据条数，就是各个节点的发生概率。（2）首先，用第一种方法计算原因节点的发生概率，然后计算原因节点到中间节点或结果节点的条件概率，最后根据原因节点的概率和这些条件概率计算结果节点的概率。

## 第 8 章　粗糙集

1. 答：　粗糙集理论是一种新型处理不完整性和不确定性问题的数学工具，它能对不完整资料（数据）进行分析，推理，学习和发现，具有很强的知识获取能力。

2. 答：X 的下近似集合为 $\{x_5\}$。X 的上近似集合为 $\{x_1, x_2, x_1, x_5, x_6, x_7, x_8\}$。

3. 答：

根据上近似的定义，$\bar{R}(X)$是一个非粗糙集，也就是说，$\bar{R}(X)$是一些等价类的并。所以$\bar{R}(X)$取上近似后，仍然是它自身；$\bar{R}(X)$取下近似后，也仍然是它自身。

4. 答：

（1）由 $R_1$ 形成的等价类划分是$\{x_1, x_2\}$、$\{x_3, x_6\}$和$\{x_4, x_5\}$；由 $R_2$ 形成的等价类划分是$\{x_1, x_2, x_3, x_6\}$和$\{x_4, x_5\}$；由 $R_3$ 形成的等价类划分是$\{x_1, x_2\}$、$\{x_3, x_5\}$和$\{x_4, x_6\}$。

（2）由 R 形成的等价类划分是$\{x_1, x_2\}$、$\{x_3\}$, $\{x_4\}$, $\{x_5\}$, $\{x_6\}$。（3）X 相对于 R 的粗糙度为 1/3。

5. 答：

辨识矩阵为：

$$\begin{vmatrix} 0 & R2 & R1,R2 & R2 & \phi \\ & 0 & R1,R2 & R2 & R1 \\ & & 0 & \phi & R2 \\ & & & 0 & R1,R2 \\ & & & & 0 \end{vmatrix}。$$

## 第 9 章　神经网络

1. 答：　前馈网络和递归网络的本质区别是网络的某些输出是不是循环作为网络的输入。前馈网络的所有输出都不能作为输入，而递归网络的某些输出可以循环作为网络的输入。

2. 答：　多层前馈网络中隐藏层神经元的作用是增强网络的适应能力。通过隐藏层，多层前馈网络可以逼近系统中任意非线性的成分。

3. 答：　在 BP 算法中，总体误差对网络输出的偏导数和有序导数始终一致。这是因为：总体误差与网络输出变量之间没有中间变量；而总体误差对网络输入的偏导数和有序导数是不一致的。这是因为，总体误差和输入变量之间有中间变量。

4. 答：　第一行的 2 个输出分别是：4.84 和 20.32。第二行的 2 个输出分别是 3.96 和 19.27。

## 第 10 章　遗传算法

1.　4。

2. 1000110101，0101001101。

3. 1001010101，0100101101。

4. 1101111101，0000000101。

5. 答：　（1）函数优化问题　（2）组合优化问题　（3）生产调度问题　（4）自动控制　（5）图像处理　（6）人工生命　（7）遗传编程　（8）机器学习

6. 答：　SGA 的基本流程如下：　（1）初始化，产生初始种群。　（2）个体评价，即计算种群中每个个体的适应度。　（3）按选择概率 Ps，执行选择算子，从当前种群中选择部分个体进入下一代种群。　（4）按交叉概率 Pc，执行交叉算子。　（5）按变异概率 Pm，执行变异算子。　（6）若满足设定的终止条件，则输出种群中适应度最优的个体作为问题的最优解或满意解，否则执行（2）。

7. 答：　（1）确定编码方式，以便对问题的解进行编码，即用个体表示问题的可能解。　（2）确定种群大小规模。　（3）确定适应度函数，决定个体适应度的评估标准。　（4）确定选择的方法及选择率。　（5）确定交叉的方法及交叉率。　（6）确定变异的方法及变异率。　（7）确定进化的终止条件。

8. 答：存在早成熟收敛，收敛速度慢等缺点。

9. 答：

**表 10.6　习题 9 所用表格**

| 个体 | 适应度 | 选择概率 | 累计概率 |
|---|---|---|---|
| 1 | 328 | 0.146037 | 0.146037 |
| 2 | 446 | 0.198575 | 0.344613 |
| 3 | 529 | 0.23553 | 0.580142 |
| 4 | 943 | 0.419858 | 1 |

## 第 11 章　统计分析

1. 答：参考 11.1.1 节的推导过程。

2. 答　线性回归模型的因变量是连续的，不太适合因变量 Y 为二分变量（例如因变量 Y 的具体取值为 1 时表示购买了产品，因变量 Y 的具体取值为 0 时表示没有购买产品）的场合。在因变量为二分变量时一般采用 Logistic 回归模型（逻辑回归模型）的形式，用极大似然估计法（maximum likelihood estimate）求解模型中参数。

3. 答：　一般来说，建立 ARIMA 模型需要以下几个步骤：　（1）根据时间序列的图形或者其他方法对序列的平稳性进行判断。包含长期趋势和周期性变化的时间序列一定是不平稳的。　（2）对非平稳序列进行平稳化处理，一般使用差分的方法。在差分时需要确定差分的阶数，即 d 的取值。　（3）对于差分后的平稳序列，根据时间序列模型的识别规则建立

相应的模型，也就是确定模型中 p 和 q 的值。模型识别中最主要的工具是自相关函数和偏相关函数。自相关函数描述了时间序列的当前序列和滞后的相关系数；偏相关函数描述了给定中间序列的条件下当前序列和滞后序列的相关系数。自相关函数和偏相关函数的图形可以帮助使用者初步判断时间序列所适合的模型形式和自回归、移动平均的阶数。 （4）确定了模型中 p、d、q 的值，接下来就需要对模型中的 p+q 个参数进行估计了。ARMA 模型的参数估计可以采用最小二乘估计或者极大似然估计等。参数估计的过程比较复杂，但借助于统计软件的帮助在实际应用中这已经不是一个问题了。 （5）估计出模型的参数后，通常需要借助于一些统计方法对模型中参数的显著性、拟合效果等进行检验和分析。对模型残差的自相关函数和偏自相关函数进行分析是检验的重要内容，如果残差序列的自相关系数和偏自相关系数在统计上都不显著，就可以认为模型是可接受的。 （6）通过检验的模型就可以用来进行预测了。预测通常通过统计软件来实现，手工计算对于包含 MA 项的模型来说困难比较大。

4. 答：参考 11.1.5 节的过程。

5. 答：参考 11.2.3 节的过程。

6. 答：参考 11.3.4 节的过程。

## 第 12 章 文本和 Web 挖掘

1. 答：Web 挖掘的 3 个主要类别是 Web 内容挖掘、Web 结构挖掘和 Web 使用挖掘。

2. 答： 文本的向量为

$$v_1 = \{ 1,1,3,2,1,1,1,2,4,2,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0 \}$$
$$v_2 = \{ 0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0 \}$$
$$v_3 = \{ 1,1,3,0,1,0,1,1,2,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0 \}$$

文档之间的相似性为：

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|} = \frac{0}{\sqrt{43} \times \sqrt{5}} = 0$$

$$sim(v_1, v_3) = \frac{v_1 \cdot v_3}{|v_1||v_3|} = \frac{23}{\sqrt{43} \times \sqrt{19}} = \frac{23}{28.58} = 0.80$$

$$sim(v_2, v_3) = \frac{v_2 \cdot v_3}{|v_2||v_3|} = \frac{0}{\sqrt{5} \times \sqrt{19}} = 0$$

3. 答： 查准率是检索到的文档中的相关文档占全部检索到的文档的百分比，它所衡量的是检索系统的准确性。查全率是被检索出的文档中的相关文档占全部相关文档的百分比，它所衡量的是检索系统的全面性。 4. 答：请参看 12.2.5 节中有关分词的内容。 5. 答： 路径分析可用于发现 Web 站点中最经常被访问的路径，从而调整站点的结构。例如，某个网站的主页 A 中有一个链接指向了网页 B，而网页 B 中有一个链接指向了网页 C。经过 Web 路径挖掘发现，凡是从主页开始访问并链接到网页 B 的用户，大都最后链接到了网页 C。根据发现的这条规律，可以在主页中增加一个链接 C，这样可以方便大多数用户的使用。