《数据仓库与数据挖掘》试题与答案整理

2013 级智能系 高飙

1.名词解释 5x4

（1）主题

主题（Subject）：宏观分析领域所涉及的分析对象。是在较高层次上将企业信息系统中的数据进行综合、归类和分析利用的一个抽象概念，每一个主题基本对应一个宏观的分析领域。

面向主题的数据组织方式：在较高的层次上对分析对象的数据的一个完整、一致的描述。

（2）事实（P 联机分析）

事实是数值度量的； 存储一个多维数据，表达期望分析的主题（目的、感兴趣的事情、事件或者指标等）；具有一定的粒度，粒度的大小与维层次相关；

一个事实中通常包含一个或者多个度量

一个事实的两个组件：数字型指标 、聚集函数

（3）数据归约（P 数据预处理）

在可能获得相同或相似结果的前提下，对数据的容量进行有效的缩减

数据归约的方法：

1 数据立方体聚集：聚集操作作用于立方体中的数据

2 减少数据维度（维归约）：可以检测并删除不相关、弱相关或者冗余的属性或维

3 数据压缩：使用编码机制压缩数据集

4 数值压缩：用替代的、较小的数据表示替换或估计数据

5 数据离散化以及概念层次的建立：属性的原始值用区间值或较高层的概念予以替换

（4）兴趣度（P 数据挖掘）

一个数据挖掘系统的挖掘结果可能会产生成千上万个模式，但是并不是所有的模式都有意义。兴趣度度量用于将不感兴趣的模式从知识中分开。他们可以用于指导挖掘过程，或在挖掘之后，评估发现的模式。不同类型的数据需要不同的兴趣度量。

兴趣度的度量：一个模式是否感兴趣，取决于它是否容易被用户所理解，是否有效可信，是否潜在有用，是否新颖等

兴趣度的度量:

客观的度量: 从模式的角度出发，基于模式结构的某些统计的结果，如：支持度（support）、置信度（confidence）等。

主观的度量: 从用户的角度出发，对模式的信任程度，如：新颖性、可操作性等。

（5）数据分区（片）（P 数据仓库设计）

把逻辑上统一的数据分割成较小的、可以独立管理的物理单元（分片）进行存储。

可按时间、按地区、按业务类型进行数据分片

（6）数据挖掘

数据挖掘是识别数据中有效的、新颖的、潜在有用的和最终可被理解的模式（Pattern）的非平凡过程。

（7）关联分析

是数据挖掘的分析方法之一，发现数据库中数据间的相互关联。关联分析发现关联规则，这些规则展示属性-值频繁地在给定数据集中也一起出现的条件。

（7'）关联规则

形如 X=>Y,即“A1∧…∧Am=>B1∧…∧Bn”的规则，其中 Ai,Bj 是属性-值对。关联规则 X=>Y 解

释为“满足 X 中条件的数据库元组多半也满足 Y 中的条件”。

发现海量数据中项集之间有趣的关联；

在交易数据、关系数据或其他信息载体中，查找存在于项目集合或对象集合之间的相关性或因果结构

（8）维度（P8）

数据仓库中的每一维对应于模式中的一个或一组属性。

或者（P 联机分析）：对数据进行分类的一种结构，以用于从特定的角度观察数据。（例如：时间、地区、产品）

（9）度量（指标）

数据的实际意义，一般是一个数值度量指标

2.简答 10x4

（1）描述 ROLAP、MOLAP、HOLAP 的特点（P46+P 联机分析）

MOLAP：

1 数据存储容量较 ROLAP 少，往往利用 RDB 存储细节数据，MDB 存储综合数据

2 元数据以内在方式处理，元数据描述了层次关系、时间序列信息、报表项、安全存取控制、数据源以及预综合等等。

3 利用多维查询语言直接访问 MDB（不借助附加程序）

ROLAP：

1 以关系数据库系统方法进行数据存储和管理；安全控制和存取控制基于表； 封锁基于表、页面或行；

2 多维概念下的安全及存取控制，RDBMS 不支持，需由 OLAP Server 实现

3 数据存储容量大（因为 RDB 技术成熟）但为了提高性能，须建中间表（预综合），数据冗余大

4 元数据作为应用的一部分，由 ROLAP Server 管理

5 用户的分析（查询）请求，需 SQL 和附加的应用程序共同完成，可以直接在细节数据上提供 OLAP 的功能

（2）数据粒度的概念及它在 DW（数据仓库）建模中的作用

粒度：数据的综合程度。例如：细节 — 轻度综合 — 高度综合。数据越详细，粒度越小，层次级别就越低；数据综合度越高，粒度越大，层次级别就越高。

作用：合理的粒度划分是提高数据仓库性能的途径之一。粒度影响着数据仓库中数据量的大小，同时影响数据仓库所能回答的查询类型。粒度大小需要数据仓库在设计时在数据量大小和查询的详细程度之间做出权衡。

一张表的数据量很大时，就需要两个级别的粒度。粒度的划分，主要考虑行数。因为按行组织索引，索引依赖于行数，索引大小直接影响 I/O 次数。

（3）最大频繁项集与闭合项集的区别与联系（P 关联规则挖掘）

最大频繁项集：自身是频繁项集，任何直接后继超集都不是频繁项集

闭合项集：自身是频繁项集，所有直接后继超集项集的支持度均小于当前的频繁项集

（4）多维数据模型的概念及优势（P 联机分析）

*概念：*多维数据模型又称多维概念视图，通常用 Cube 来表示。多维数据模型的基本组成：维、度量（变量、指标）

多维数据模型是为了满足用户从多角度多层次进行数据查询和分析的需要而建立起来

的基于事实和维的数据库模型，其基本的应用是为了实现 OLAP

*优势：*多维数据模型最大的优点就是其基于分析优化的数据组织和存储模式。多维数据模型可以更加直观的表示现实中的复杂关系；

（5）数据挖掘的分类

针对的数据源不同

关系数据库、对象数据库、空间数据库、时序数据库、文档数据库、多媒体数据库、Web 等

采用的不同的分析方法

关联分析、分类分析、聚类分析、趋势分析、偏差分析以及异常点分析等

采用的不同技术

利用数据库或数据仓库的方法、机器学习的方法、统计的方法、神经网络的方法等。

不同的应用领域

金融、电信、商业、DNA 分析、 ……、股市分析等

（6）置信度,支持度的概念和联系（P 关联规则挖掘）

支持度（Support）：规则X⇒Y在交易数据库D中的支持度是交易集中包含X和Y的交易数与所有交易数之比，记为support(X⇒Y)，即support(X⇒Y)=|{T:X∪Y⊆ T，T∈D}|/ |D|，它是概率P（X，Y），具体表示为：

$$S = \frac{\text{同时包含项集X 和 Y 的交易数}}{\text{总交易数}}$$

置信度（Confidence）：规则X⇒Y在交易集中的置信度是指包含X和Y的交易数与包含X的交易数之比，记为confidence(X⇒Y)，即confidence(X⇒Y)=|{T: X∩Y⊆T，T∈D}|/|{T:X⊆T，T∈D}|，它是条件概率P（Y|X），具体表示为：

$$C = \frac{\text{同时包含X和Y的交易数}}{\text{包含X的交易数}}$$

他们都是关联规则有效性和确定性的度量值，或者说是模式兴趣度的客观度量。

（7） 数据仓库和数据集市的区别与联系（P9）

数据集市包含企业范围数据的一个子集，对于特定的用户是有用的。其范围限于选定的主题。是数据仓库的三种模型之一。

数据仓库收集了整个组织的主题信息，因此它是企业范围的。数据集市是数据仓库的一个部门子集。它聚焦在选定的主题上，是部门范围的。

一般来说，数据仓库更倾向于是一个战略，但不是一个未完成的概念；而数据集市更倾向于战术，它的目标在于满足企业客户营销即时的需求。

比较对象	数据仓库	数据集市
数据来源	ODS统一信息视图区	数据仓库
数据范围	面向企业级	一般是部门级
数据结构	第三范式	雪花型或星型结构
历史数据	大量的历史数据	一部分历史数据
索引	高度索引	高度索引

补充：P45

（8） 聚类分析和分类分析的区别和联系（P 数据挖掘）

他们是数据挖掘的不同分析方法。

聚类分析：

1 描述型。了解数据中潜在的规律、规则。以简洁概要的方式描述数据，并提供数据的有趣的一般性质

2 每个类的标识事先不确定，把一组对象按照相似性归成若干类别，即“物以类聚”。基本的原则：属于同一类别的个体之间的距离尽可能的小而不同类别上的个体间的距离尽可能的大。

分类分析：

1 预测型。用历史预测未来。分析数据，建立一个或一组模型，并且试图预测新的数据集的行为

2 在数据库的一个对象集中发现公共的属性，并根据分类模型把这些对象分成不同的类的过程。

例如：根据不同的气候环境，对不同地区进行分类；根据不同的成绩，对学生进行分类。

方法表述: 决策树、分类规则、神经网络等

（9） 简述数据仓库建模中数据项集(DIS)的概念

数据仓库设计的 Inmon 方法中，数据建模的三级数据模型中的中级数据模型（称为数据项集 DIS），一个 dis 与 E—R 中的一个主题域（实体）对应。另外两层模型是高级数据模型（采用 E-R 方法）和低级数据模型（物理模型）。

3.论述 15x2

（1）有一个事务集 T 如下，最小支持度为 62.5%，求其 1-3 阶频繁项集

001	**ABCD**
002	**BCDE**
003	**ABCDE**
004	**ADE**
005	**BDE**
006	**ACEF**
007	**BCDE**
008	**BCDEF**

（2）K-means 算法和 K-中心点算法的详细步骤与特点，并比较两种算法

步骤：

K-means:

给定 k，算法的处理流程如下:

1. 随机的把所有对象分配到 k 个非空的簇中；
2. 计算每个簇的平均值，并用该平均值代表相应的簇；
3. 将每个对象根据其与各个簇中心的距离，重新分配到 与它最近的簇中；
4. 回到第二步，直到不再有新的分配发生。

K-中心点：

用真实的数据对象来代表簇

随机选择 k 个对象作为初始的中心点；

Repeat

对每一个由非中心对象 h 和中心对象 i, 计算 i 被 h 替代的总代价 Tcih

对每一个有 h 和 i 组成的对象对

If TCih < 0, i 被 h 替换

然后将每一个非中心点对象根据与中心点的距离分配给离它最近的中心点

Until 不发生变化。

特点：

K-means:

优点

相对高效的: 算法复杂度 O(tkn), 其中 n 是数据对象的个数, k 是簇的个数, t 是迭代的次数，通常 k, t << n.

算法通常终止于局部最优解；

缺点

只有当平均值有意义的情况下才能使用，对于类别字段不适用；

必须事先给定要生成的簇的个数；

对“噪声”和异常数据敏感；

不能发现非凸面形状的数据。

K-mediods:

a) K-mediods 算法具有能够处理大型数据集，结果簇相当紧凑，并且簇与簇之间明显分明的优点，这一点和 K-means 算法相同。

b) 同时，该算法也有 K-means 同样的缺点，如，必须事先确定类簇数和中心点，簇数和中心点的选择对结果影响很大；一般在获得一个局部最优的解后就停止了；对于除数值型以外的数据不适合；只适用于聚类结果为凸形的数据集等。

c) 与 K-means 相比，K-mediods 算法对于噪声不那么敏感，这样对于离群点就不会造成划分的结果偏差过大，少数数据不会造成重大影响。

d) K-mediods 由于上述原因被认为是对 K-means 的改进，但由于按照中心点选择的方式进行计算，算法的时间复杂度也比 K-means 上升了 O(n)。

比较：

存在“噪声”或者孤立点数据时，K-中心点的方法比 K-平均方法健壮；

K-中心点方法的执行代价比 K-平均方法高；

四（15'）

列举 4 种面向数据仓库实际需求的索引技术，并说明其特点和适应性

索引	描述	优点	缺点	商业实现
B 树索引	把存储块组织成一棵树来减少 I／O 操作	• 需要少的 I／O 操作； • 适合于高基数的列； • 能自动数据文件大小相适应的索引层次； • 索引空间的需求独立于被索引列的基数； • 易于创建；	• 低基数的列效果不好； • 不支持即席查询；对宽范围查询 I／O 代价相对较高； • 在获取数据之间索引不能合并；	大多数据的商用数据(0raleRed Brick 等)
简单位图索引	为索引属性的每个取值建立一个位向量，向量长度等于元组数。对应的元组中相应的索引位置 1，若该值出现，否则为 0	• 适合于低基数的列； • 利用位操作； • 获取数据之前可合并索引； • 便于在并行机上执行； • 可以高效地完成对属性的数量型函数(如 count) 的查询； • 易于创建； • 易于插入新的索引值； • 适合于 OLAP；	• 高基数的列效果不好； • 更新索引列代价较高； • 不能很好处理稀疏数据	OracleAscen ialSybaseRed Brick DB2
编码位图索引	对属性的域进行二进制编码	• 有效地利用空间； • 可以实现宽的范围查询；	• 对于等值查询效率低； •很难找到好的的编码方案； • 当有新的属性值出现，现有的位数不能满足，需重建；	DB2
位图连接索引	索引的创建通过维表对事实表的约束实现的	• 灵活性较好； • 有效地执行； • 支持星型查询；	• 索引列的次序很重要	OracleAscen ialRed Brick
投影索引	通过实际存储被索引表的列值建索引	• 当只检索表中的几个列时，加速比较高；	• 只能用于检索原数据；	Sybase

五.

就你感兴趣的领域，说说数据仓库和数据挖掘技术的应用及在该领域的应用 10x1

数据挖掘见 PPT

数据仓库技术在金融信息化中的定位和作用（**引言**）

金融业务和信息技术的紧密融合，已经成为金融行业打造核心竞争力的重要途径。随着国内外金融行业竞争的日益加剧，如何利用信息技术提升业务管理水平，增强业务创新能力，为客户提供更优质的服务，成为我国金融行业面临的重大课题。

近年来，我国金融信息化按照“数据集中化、业务综合化、管理扁平化、服务网络化、决策科学化”的理念，构建了两大数据平台：一个是基于数据大集中的策略，面向金融业务

数据处理，构建高效、统一的核心业务数据平台；另一个是面向分析处理，构建完整、一致、反映时间变化的数据仓库平台。

数据仓库平台的建设实现了企业异构数据的集成，企业按照分析主题重组数据库，建立了面向整个企业的、一致的信息视图，提升了数据的利用价值。在此基础上，结合联机分析处理技术（Online Analytical Processing）和数据挖掘技术（Data Mining），为有效控制企业风险、实现金融企业经营资源的优化配置等提供了数据基础，也为相关金融企业的经营决策提供有力支撑，大大增强了决策的科学性。

数据仓库技术是金融信息化发展到一定阶段的必然选择。数据仓库技术在金融行业的应用，将为推进金融业务的发展和创新，促进我国金融行业的改革和发展，起到积极的作用。目前，部分银行、保险、证券等企业的数据仓库建设主要围绕资产负债管理、客户关系管理、风险管理、绩效管理等业务主题展开，实现了对历史数据的集成和重组，为各类分析型应用提供了较好的数据基础。

七 结合你熟悉的领域,说明(1)构建数据仓库和数据挖掘应用的必要性.(2)建设数据仓库涉及的主题及其内容.(3)数据挖掘的主要应用. 15

(1) 构建数据仓库和数据挖掘应用的必要性

数据采集、数据存储、数据处理、数据共享能力的持续增强

数据极大丰富，知识极其匮乏；数据挖掘的动机：在海量数据集中挖掘知识

从金融信息化的角度

- “数据集中化、业务综合化”，极大推动了金融行业的信息化建设进程，提高了劳动生产率，同时也积累了大量的数据
- “管理扁平化、决策科学化”成为金融行业在激烈竞争中创造优势的重要手段

优化客户管理，提供高质量、个性化的服务

客户分类分析、行为分析、满意度分析、忠诚度分析

发挥金融创新，推出创新产品

交叉销售、营销策略

⇨ 构建企业级数据仓库、开发深层次的数据挖掘应用

⇨ 呼唤强大的智能型自动数据分析技术和工具，对数据进行有效预处理；对数据进行智能分析（数据分割、数据分类等）

(2) 建设数据仓库涉及的主题及其内容

从金融信息化的角度

主题：客户、产品、收入

保险公司：

面向主题：客户、保单、保费、理赔（赔款）

(3) 数据挖掘的主要应用. （见 PPT）

数据仓库技术为企业提供企业级的数据存储，数据挖掘技术为企业提供智能