Ciências ULisboa

Integração e Processamento Analítico de Informação

Código: 425124
Ano Letivo: 2015/16
Departamento: Informática
ECTS: 6
Carga horária: T: 2:00 h; TP: 1:30 h;
Área Científica: Informática;

Objetivos da Unidade Curricular

Abordar as especificidades da modelação de data warehouses, nomeadamente a extração, transformação, e carregamento de dados, e a modelação dimensional. Discussão de várias ferramentas e técnicas para processamento analítico (OLAP) e para prospeção de informação (data-mining). É ainda abordada a articulação destas técnicas em sistemas de apoio à decisão para vários domínios de aplicação.

Pré-requisitos

- Sistemas de Informação e Bases de Dados (26726)

Conteúdos

Sistemas operacionais e de apoio à decisão; arquitetura e construção de um data warehouse; modelação dimensional de dados; estudo de casos; extração, transformação, e carregamento de dados; desenho físico de data warehouses; ciclo virtuoso do data mining; aplicações práticas; métodos de agrupamento, classificação, e associação; avaliação de modelos de data mining.

Descrição detalhada dos conteúdos programáticos

Componente Teórica

Introdução ao data warehousing (DW) e business intelligence (BI): sistemas operacionais e de apoio à decisão; diferenças entre online analytical processing (OLAP) e online transaction processing (OLTP); história dos sistemas de apoio à decisão; executive information systems (EIS); benefícios esperados e obtidos dos EIS; dimensão e importância da business intelligence; bases de dados federadas; modelos de dados multidimensionais; interrogações OLAP; processo de construção de um data warehouse.

Modelação dimensional: obtenção de dimensões; metáfora do cubo de dados; navegação num cubo de dados; tabelas de factos e de dimensões; grão das tabelas de factos; desenho lógico; aspetos da concretização de um modelo multidimensional; interrogações OLAP em SQL; exemplo do retalhista; chaves candidatas e dimensões degeneradas; dimensão data e hierarquias fixas; chaves substitutas; snowflaking; análise dimensional; medidas, aditividade, e semiaditividade; tabelas de factos sem factos; dimensões de mudança lenta; técnicas para registar mudanças em dimensões; role-playing de dimensões; dimensões conformadas; bus matrix; relatórios transdepartamentais; dimensões muito grandes; bifurcações e mini-dimensões; hierarquias de profundidade fixa e variável; tabelas de ponte; dimensões multi-valor; tabelas de factos de tipo transação, instantâneo periódico, e instantâneo cumulativo; regras e erros a evitar no modelação dimensional.

Processo de carregamento do data warehouse: propósito da data staging area; sistema de extração, transformação, e carregamento; manutenção de chaves substitutas com tabela de correspondências.

Estudo de casos de data warehousing: gestão de encomendas, com preços em várias moedas; gestão de relação com o cliente, com bifurcações e mini-dimensões; telecomunicações, com revisão de um modelo inicial; comércio electrónico, com factos com vários grãos.

Desenho físico de data warehouses: motivação; índices de árvore B+, função de dispersão, e de mapas de bits; índices agrupados e compostos; vistas materializadas; compressão de dados; partição de dados; armazenamento distribuído de dados.

Introdução ao data mining: propósito e tarefas típicas; ciclo virtuoso; métodos supervisionados e não supervisionados; conjuntos de treino, teste, e avaliação; métodos de agrupamento hierárquico aglomerativo e iterativo k-means; método de classificação com árvores de decisão; método de classificação com redes neuronais; método Apriori para geração de regras de associação; redução da dimensionalidade dos dados; avaliação de modelos de classificação; armadilhas no data mining.

Componente Teórica-Prática

Análise multidimensional em Excel: demonstração de utilização de pivot tables; colocação de atributos em colunas e linhas de um relatório; cálculo de valores agregados tais como somas e médias; aplicação de filtros sobre os dados; criação de novos atributos preenchidos com valores derivados de outros atributos.

Modelação e carregamento de dados multidimensionais: conversão, em folha de cálculo Excel, de tabela única de vendas em tabelas de dimensões com atributos detalhados e tabela de factos com chaves estrangeiras e medidas; criação de tabelas de dimensões e de factos na base de dados relacional SQL Server; carregamento de dados de um ficheiro Excel para o SQL Server usando o Import and Export Data wizard; carregamento de dados avançado com um projeto de Integration Services do SQL Server Business Intelligence Development Studio; execução de interrogações OLAP em SQL.

Criação de cubos de dados com os Analysis Services: criação do projeto e definição de permissões de uso do cubo; identificação da fonte de dados; vista mais inteligível sobre os dados; definição de dimensões e hierarquias de atributos; criação do cubo de dados e composição de relatórios dinâmicos; acesso ao cubo de dados e geração de relatórios em Excel.

Introdução ao data mining no ambiente R: introdução ao ambiente de trabalho R; exercícios de aplicação dos métodos de agrupamento hierárquico aglomerativo e iterativo k-means; exercício de classificação com uma rede neuronal.

Bibliografia

Recomendada

Ralph Kimball e Margy Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Wiley, 2ª edição, 2002, ISBN 0471200247

Ian Witten, Eibe Frank, e Mark Hall, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kauffman, 3ª edição, 2011, ISBN 0123748569

Outros elementos de estudo

Coletânea com 74 exercícios de testes e exames

William H. Inmon, Building the Data Warehouse, Wiley, 4ª edição, 2005, ISBN 0764599445

Ralph Kimball e Margy Ross, The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence, Wiley, 2010, ISBN 0470563109

Hugh Watson, George Houdeshel, e Rex Rainer, Building Executive Information Systems and Other Decision Support Applications, Wiley, 1997, ISBN 0471069302

Raghu Ramakrishnan e Johannes Gehrke, Database Management Systems, McGraw-Hill, 3ª edição, 2003, ISBN 0072465638

Jiawei Han, Micheline Kamber, e Jian Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2ª edição, 2005, ISBN 1558609016

Michael Berry e Gordon Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support, Wiley, 2011, ISBN 0470650931

Trevor Hastie, Robert Tibshirani, e Jerome Friedman, The Elements of Statistical Learning: Data Mining Inference and Prediction, Springer, 2ª edição, 2003, ISBN 0387952845

Métodos de Avaliação

Existem dois métodos de avaliação: a) projeto (35%), 2 testes durante as aulas (27,5% + 27,5%), e participação nas aulas (10%); ou b) projeto (35%), exame (55%), e participação (10%). O projeto é avaliado em três etapas distintas durante o semestre. Os alunos com estatuto de trabalhador-estudante podem substituir a participação nas aulas por respostas a questões semanais.

Língua de ensino

Português