Título
Ok, Computer : confluências na informática

Coordenador da Colectânea dos Textos
Luís Mendes Gomes

Autores
Armando Brito Mendes, Bernardo Rodrigues Peixoto,
José Manuel Cascalho, Luís Mendes Gomes,
Gunther Matthias Funk e Rui Sampaio da Silva

Edição
Influir

Direcção Gráfica e Capa
José Júlio Ribeiro

Impressão
Coingra

ISBN
978-989-97107-1-9

Dep. Legal
329360/11

Apoios

Governo dos Açores SRCTE cmati

METODOLOGIAS DE *DATA MINING*

Para se compreender o que é mais este termo criado pela indústria das novas tecnologias, pode começar-se por esclarecer o que não é. Certamente *data mining* não são simples consultas de bases de dados ou as tecnologias de integração de dados em bases de dados multidimensionais ou cubos de dados, ou mesmo um conjunto de algoritmos como redes neuronais, algoritmos genéticos, métodos estatísticos, *etc.*. A prospecção de dados é um processo ou uma metodologia para a descoberta de conhecimento que pode envolver todas as técnicas anteriores em diferentes passos bem definidos.

O modelo processual CRISP-DM[1] (*CRoss Industry Standard Process for Data Mining*) tem-se destacado como um padrão, principalmente por ser não proprietário e pretender ser independente do sector e das aplicações em que é utilizado. Esta metodologia tem sido validada com vários projectos de grande dimensão e resulta do trabalho conjunto de um utilizador intensivo, um fabricante de *software* e um consultor especialista em armazéns de dados (ou *data warehouses*). Actualmente encontra-se em revisão que pode ser participada por todos.

Na versão 1.0, recomendam-se seis fases. Começa-se por compreender o problema e o contexto onde surge, incluindo a definição de objectivos a atingir e um plano de acção. Esta fase gera e difunde entre os vários intervenientes no projecto, frequentemente multidisciplinar, conhecimento de domínio que será utilizado durante todo o processo. O conhecimento do domínio ou da área em estudo define-se como conhecimento não explícito ou tácito sobre o contexto da indústria ou serviço em causa.

A fase dois compreende a recolha, a integração, exploração e compreensão dos dados. Esta fase é igualmente responsável por uma avaliação prévia da qualidade destes. A preparação e pré-processamento incluem tarefas de integração, redução, transformação e limpeza de dados. Nesta fase, tecnologias de armazéns de dados, de integração e cubos, são muito úteis.

Na fase de estimação ou aprendizagem de modelos, são utilizados diversos algoritmos tanto da estatística como da aprendizagem automática, como os referidos no início deste texto. Esta é a fase onde,

[1] Sítio web do projecto europeu que resultou na metodologia descrita, http://www.crisp-dm.org

mais do que nas restantes, se gera conhecimento novo, comparando os resultados obtidos por uma grande diversidade de algoritmos.

Na fase cinco, os resultados são validados, comparados, interpretados e confrontados com conhecimento de domínio, permitindo identificar conhecimento novo.

Na última fase, a divulgação e implementação (*deployment*) pode ser tão simples como a escrita de um relatório, ou tão complexa como a criação de uma aplicação integrada no sistema de informação. Em qualquer dos casos, pretende fazer chegar o conhecimento aos utilizadores e decisores. Apesar das fases bem definidas, o processo não é linear e apresenta imensos ciclos e retornos, a que alguns autores chamam a espiral de modelação e extracção de conhecimento.

Tal como em todos as metodologias, a CRISP-DM não garante resultados mas permite disciplinar o processo e tem como grande finalidade alinhar os objectivos do projecto de *data mining* com os do negócio.

Recursos

Cortes, B. C. (2005). *Sistemas de Suporte à Decisão:* FCA.
Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. e Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*: SPSS Inc..