Universidade Federal de Santa Maria – Centro de Tecnologia
Peso: 5
Disciplina: Mineração de Dados (ELC1098)
Professor: Dr. Joaquim Assunção

7,6

Atenção: Leia atentamente as questões e escreva de forma clara e legível.

Nome: Lana Renato

Curso: Ciência da Computação Data: 06 / 09 / 18

1) Mineração de dados está inserida no contexto de um processo maior, KDD, cujas etapas são prévias e posteriores a mineração em si. Embora no dia-a-dia sejam tratados como sinônimos, o processo de KDD envolve tarefas como seleção e pré-processamento de dados, que ainda não são, tecnicamente, considerados como mineração. Descreva o porquê da existência dessas etapas prévias. Quais são os problemas e como essas etapas são úteis para a mineração em si. (3)

A etapa de Seleção serve para selecionar os dados de um banco de dados, como o MySQL por exemplo. A etapa de Pré-processamento serve para transformar esses dados brutos em dados mais simples de se procurar futuramente. Uma última etapa existe a fase de Limpeza, onde os dados faltantes são completados para que exista uma conformidade entre eles, mas muitas vezes pode haver dúvida quanto ao método a ser utilizado: média, valor mínimo, máximo, mediana dos dados.

2,8

2) Quais são as três grandes categorias de dados (quanto a estruturação)? Destas, qual é a mais próxima de um formato comumente minerado e por quê? (2)

Estruturados - formato mais tabulado Semi-estruturado - XML, HTML
Não estruturado - fotos e vídeos
O formato mais próximo é o estruturado, pois tabulam facilitam a distribuição e mineração dos dados, mas dos não estruturados também podem ser minerados.

3) Considerando conjuntos de dados estruturados "tidy". Por que não é comum armazenar dados nesse formato e porquê pode ser útil transformar os dados para esse formato? (2)

Não é comum pois a tabela fica muito grande se tiver muitos dados. Mas é útil porque não apresentados com mais clareza e na hora de juntar tabelas com os mesmos tipos de dados, eles não ficam separados ou confundidos em linhas/colunas diferentes.

2

4) Analise o seguinte conjunto e responda:

4.1) Caso nosso projeto de mineração tenha como objetivo encontrar correlação entre tamanho de tela, valores e lucro. Quais informações podemos descartar? (1,5)

A coluna Produto e o resto da Descrição que não fala sobre a tela

-0.5

ID

4.2) Considerando esse objetivo (4.1), devo alterar (adicionar ou editar) algo nesse conjunto? Por que? (1,5)

Não é útil pois como no ID índice

Deixar a coluna ID, retirar a coluna Produto, retirar as informações que não forem sobre tela da coluna Descrição (se não for possível, no deixar), juntar as colunas Valor compra e Valor venda subtraindo-as e deixar a coluna de Unidades Vendidas. Assim a tabela ficaria muito mais clara e objetiva, deixando apenas as informações necessárias

-0.3

ID	Produto	Descricao	Valor compra	Valor venda	Unidades vendidas
1	Monitor LG 29p	Entradas 1,2,3. 29p. Full HD IPS 29UM68	250	290	29
...	...		...	...	...
999	Monitor Sansung 30p	Entr. 1,2,3,4. 30p. 4K	500	600	45

(Descrição)

ID	Tela	Dif Venda	Unidades Vendidas