Definições e Conceitos de Data Warehouse e Data Mining

O Data Warehouse é um sistema utilizado para armazenar dados, de uma maneira organizada. Considera-se o Data Warehouse (DW) a base para o Business Intelligence (BI).

O DW pode guardar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão.

Surgimento do Conceito de Data Warehouse / Armazém de Dados

Os data warehouses surgiram como conceito acadêmico na década de 80. Com o amadurecimento dos sistemas de informação empresariais, as necessidades de análise dos dados cresceram paralelamente. Os sistemas OLTP não conseguiam cumprir a tarefa de análise com a simples geração de relatórios. Nesse contexto, a implementação do data warehouse passou a se tornar realidade nas grandes corporações.

O mercado de ferramentas de data warehouse, que faz parte do mercado de Business Intelligence, cresceu então, e ferramentas melhores e mais sofisticadas foram desenvolvidas para apoiar a estrutura do data warehouse e sua utilização. Atualmente, por sua capacidade de sumarizar e analisar grandes volumes de dados, o data warehouse é o núcleo dos sistemas de informações gerenciais e apoio à decisão das principais soluções de business intelligence do mercado.

Grandes Depósitos De Dados

Os Data Warehouse/Armazém de Dados, são grandes depósitos de dados que armazenam as informações de empresas de forma consolidada. Sua origem data dos anos 80 em instituições acadêmicas que inspiraram os sistemas de data warehouse corporativos. Tal solução evoluiu e hoje integra as funcionalidades essenciais de sistemas de Business Intelligence. Seu grande princípio é integrar dados de diferentes sistemas em atualização periódica de longo prazo que possibilita a visualização de relatórios de períodos de duração mais prolongada.

Data Warehouse / Armazém de dados – Aplicação

As informações organizadas em um Data Warehouse possibilitam a produção de relatórios e análises em séries históricas, dentre outras funcionalidades. Com base nos dados produzidos a partir de uma base confiável é possível tomar decisões gerenciais assertivas com embasamento em relatórios precisos e amparados por informações sistematizadas.

O DW é a base para montagem de um sistema de dados, onde a corporação pode unificar todos os seus sistemas para ter uma base única para montagem de relatórios, posteriormente atividades de Data Mining/Mineração de Dados também podem ser aplicadas a esse banco de dados.

Segurança e Integridade dos Dados

A informações de um Data Warehouse estão disponíveis apenas para leitura. Seus dados não podem ser modificados, exceto em casos onde tais dados tenham sido inseridos de modo incorreto. A possibilidade de fazer apenas a leitura das informações assegura a integridade do conteúdo armazenado.

Data Warehouse / Armazém de Dados – Principais Ferramentas

As OLAP (Online Analytical Processing) são conhecidas como Processo Analítico em Tempo Real. Trata-se de ferramentas que são utilizadas como interface de um Data Warehouse. Surge em complemento aos sistemas OLTP que não se configuram como soluções suficientes para análises detalhadas de grandes quantidades de dados. O Processo Analítico em Tempo Real é capaz de manipular grandes volumes de dados com versatilidade, permitindo a criação de análises comparativas que autorizem decisões qualificadas cotidianamente.

Os dados encontrados nas OLAP têm sua origem nos sistemas transacionais(OLTP – processamento de transações on-line) que são atualizados com mais frequência e compreendem ações mais regulares e cotidianas como lançamentos de notas, atualização de dados e geração de recibos e boletos.

Data Warehouse / Armazém de Dados – Benefícios

A centralização dos dados é o grande atrativo da ferramenta. Tal organização resulta em maior agilidade para captação e utilização dos dados. Uma dimensão mais ampla do ambiente também é um valor central do sistema que assegura a tomada de decisões de modo acertado, com menos complicações e imprecisões. As informações de períodos maiores obtidas do sistema são capazes de oferecer análises históricas que preveem tendências e auxiliam no planejamento a longo e curto prazo. Soma-se a estas características a melhor usabilidade dos dados de análises de sistemas transacionais.

Usando o google como fonte, vocês podem encontrar diversas escritas para o termo Data Warehouse como : Data Wharehouse, data werehouse, dataware house, data warehous, mas todos se referem a mesma técnica.

Data Warehousing, pode ser considerado o processo de se fazer Data Warehouse, ou seja, as atividades que envolvem esse tipo de projeto de DW.

Data Mining

Consideramos Data Mining ou Mineração de Dados o processo de explorar grandes quantidades de dados à procura de padrões consistentes. Como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Data mining é formada por um conjunto de ferramentas e técnicas que através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística. Estes são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. O conhecimento em Data Mining pode ser apresentado por essas ferramentas de diversas formas: agrupamentos, hipóteses, regras, árvores de decisão, grafos, ou dendrogramas.

O ser humano sempre aprendeu observando padrões, formulando hipóteses e testando-as para descobrir regras. A novidade da era do computador é o volume enorme de dados que não pode mais ser examinado à procura de padrões em um prazo de tempo razoável. A solução é instrumentalizar o próprio computador para detectar relações que sejam novas e úteis. Data Mining (DM) surge para essa finalidade e pode ser aplicada tanto para a pesquisa cientifica como para impulsionar a lucratividade da empresa madura, inovadora e competitiva.

Diariamente as empresas acumulam grande volume de dados em seus aplicativos operacionais. São dados brutos que dizem quem comprou o quê, onde, quando e em que quantidade. É a informação vital para o dia-a-dia da empresa. Se fizermos estatística ao final do dia para repor estoques e detectar tendências de compra, estaremos praticando business intelligence (BI). Se analisarmos os dados com estatística de modo mais refinado, à procura de padrões de vinculações entre as variáveis registradas, então estaremos fazendo Data Mining.

Para que Serve o Data Mining

Buscamos com a MD (Mineração de Dados) conhecer melhor os clientes, seus padrões de consumo e motivações. Data Mining resgata em organizações grandes o papel do dono atendendo no balcão e conhecendo sua clientela. Esses dados agora podem agregar valor às decisões da empresa, sugerir tendências, desvendar particularidades dela e de seu meio ambiente e permitir ações melhor informadas aos seus gestores.

Pode-se então diferenciar o business intelligence do Data Mining (MD) como dois patamares distintos de atuação. O primeiro visa obter a partir dos dados operativos brutos, informação útil para subsidiar a tomada de decisão nos escalões médios e altos da empresa. O segundo busca subsidiar a empresa com conhecimento novo e útil acerca do seu meio ambiente. O primeiro funciona no plano tático, o segundo no estratégico.

Conteúdos de Data Mining

Hoje em dia a Internet possui muito conteúdo sobre Data Mining, como Vídeos ou Tutoriais de Data Mining, também é possível encontrar excelente Livros sobre DM com Tutoriais e Matérias bem Completas.

Data Mining é uma das novidades da Ciência da Computação que veio para ficar. Com a geração de um volume cada vez maior de informação, é essencial tentar aproveitar o máximo possível desse investimento. Talvez a forma mais nobre de se utilizar esses vastos repositórios seja tentar descobrir se há algum conhecimento escondido neles. Um banco de dados de transações comerciais pode, por exemplo, conter diversos registros indicando produtos que são comprados em conjunto. Quando se descobre isso pode-se estabelecer estratégias para otimizar os resultados financeiros da empresa. Essa já é uma vantagem suficientemente importante para justificar todo o processo.

Definição e Objetos no Data Mining

Data Mining consiste em um processo analítico projetado para explorar grandes quantidades de dados (tipicamente relacionados a negócios, mercado ou pesquisas científicas), na busca de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis e, então, validá-los aplicando os padrões detectados a novos subconjuntos de dados. O processo consiste basicamente em 3 etapas: exploração, construção de modelo ou definição do padrão e validação/verificação.

A premissa do Data Mining é uma argumentação ativa, isto é, em vez do usuário definir o problema, selecionar os dados e as ferramentas para analisar tais dados, as ferramentas do Data Mining pesquisam automaticamente os mesmos a procura de anomalias e possíveis relacionamentos, identificando assim problemas que não tinham sido identificados pelo usuário.

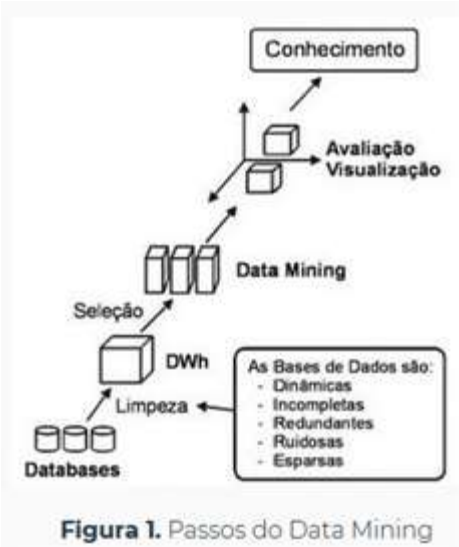

Figura 1. Passos do Data Mining

Em outras palavras, as ferramentas de Data Mining analisam os dados, descobrem problemas ou oportunidades escondidas nos relacionamentos dos dados, e então diagnosticam o comportamento dos negócios, requerendo a mínima intervenção do usuário. Assim, ele se dedicará somente a ir em busca do conhecimento e produzir mais vantagens competitivas.

Como podemos ver, as ferramentas de Data Mining, baseadas em algoritmos que forma a construção de blocos de inteligência artificial, redes neurais, regras de indução, e lógica de predicados, somente facilitam e auxiliam o trabalho dos analistas de negócio das empresas, ajudando as mesmas a conseguirem serem mais competitivas e maximizarem seus lucros.

Principais técnicas no Data Mining

O Data Mining (DM) descende fundamentalmente de 3 linhagens. A mais antiga delas é a estatística clássica. Sem a estatística não seria possível termos o DM, visto que a mesma é a base da maioria das tecnologias a partir das quais o DM é construído.

A segunda linhagem do DM é a Inteligência Artificial (IA). Essa disciplina, que é construída a partir dos fundamentos da heurística, em oposto à estatística, tenta imitar a maneira como o homem pensa na resolução dos problemas estatísticos.

E a terceira e última linhagem do DM é a chamada machine learning, que pode ser melhor descrita como o casamento entre a estatística e a Inteligência Artificial. Enquanto a Inteligência Artificial não se transformava em sucesso comercial, suas técnicas foram sendo largamente cooptadas pela machine learning, que foi capaz de se valer das sempre crescentes taxas de preço/performance oferecidas pelos computadores nos anos 80 e 90, conseguindo mais e mais aplicações devido às suas combinações entre heurística e análise estatística. Machine learning é uma disciplina científica que se preocupa com o design e desenvolvimento de algoritmos que permitem que os computadores aprendam com base em dados, como a partir de dados do sensor ou bancos de dados. Um dos principais focos da Machine Learnig é automatizar o aprendizado para reconhecer padrões complexos e tomar decisões inteligentes baseadas em dados.

O Data Mining é um campo que compreende atualmente muitas ramificações importantes. Cada tipo de tecnologia tem suas próprias vantagens e desvantagens, do mesmo modo que nenhuma ferramenta consegue atender todas as necessidades em todas as aplicações.

Existem inúmeras ramificações de Data Mining, sendo algumas delas:

Redes neurais: são sistemas computacionais baseados numa aproximação à computação baseada em ligações. Nós simples (ou "neurões", "neurônios", "processadores" ou "unidades") são interligados para formar uma rede de nós - daí o termo "rede neural". A inspiração original para esta técnica advém do exame das estruturas do cérebro, em particular do exame de neurônios. Exemplos de ferramentas: SPSS Neural Connection, IBM Neural Network Utility, NeuralWare NeuralWork Predict.

Indução de regras: a Indução de Regras, ou Rule Induction, refere-se à detecção de tendências dentro de grupos de dados, ou de “regras” sobre o dado. As regras são, então, apresentadas aos usuários como uma lista “não encomendada”. Exemplos de ferramentas: IDIS da Information Discovey e Knowledge Seeker da Angoss Software.

Árvores de decisão: baseiam-se numa análise que trabalha testando automaticamente todos os valores do dado para identificar aqueles que são fortemente associados com os itens de saída selecionados para exame. Os valores que são encontrados com forte associação são os prognósticos chaves ou fatores explicativos, usualmente chamados de regras sobre o dado. Exemplos de ferramentas: Alice d’Isoft, Business Objects BusinessMiner, DataMind.

Analise de séries temporais: a estatística é a mais antiga tecnologia em DM, e é parte da fundação básica de todas as outras tecnologias. Ela incorpora um envolvimento muito forte do usuário, exigindo engenheiros experientes, para construir modelos que descrevem o comportamento do dado através dos métodos clássicos de matemática. Interpretar os resultados dos modelos requer “expertise” especializada. O uso de técnicas de estatística também requer um trabalho muito forte de máquinas/engenheiros. A análise de séries temporais é um exemplo disso, apesar de freqüentemente ser confundida como um gênero mais simples de DM chamado “forecasting” (previsão). Exemplos de ferramentas: S+, SAS, SPSS.

Visualização: mapeia o dado sendo minerado de acordo com dimensões especificadas. Nenhuma análise é executada pelo programa de DM além de manipulação estatística básica. O usuário, então, interpreta o dado enquanto olha para o monitor. O analista pode pesquisar a ferramenta depois para obter

diferentes visões ou outras dimensões. Exemplos de ferramentas: IBM Parallel Visual Explorer, SAS System, Advenced Visual Systems (AVS) Express - Visualization Edition.

A expressão data mining surgiu pela primeira vez em 1990 em comunidades de bases de dado. A mineração de dados é a etapa de análise do processo conhecido como KDD (Knowledge Discovery in Databases), sendo a sua tradução literal "Descoberta de Conhecimento em Bases de Dado".

O data mining pode ser divido em algumas etapas básicas que são: exploração, construção de modelo, definição de padrão e validação e verificação.

A mineração de dados é uma prática relativamente recente no mundo da computação, e utiliza técnicas de recuperação de informação, inteligência artificial, reconhecimento de padrões e de estatística para procurar correlações entre diferentes dados que permitam adquirir um conhecimento benéfico para uma empresa ou indivíduo. Para uma empresa, o data mining pode ser uma importante ferramenta que potencia a inovação e lucratividade.

A utilização da mineração de dados é bastante usual em grandes bases de dados, e o resultado final da sua utilização pode ser exibido através de regras, hipóteses, árvores de decisão, dendrogramas, etc.

Uma mineração de dados bem executada deve cumprir tarefas como: detecção de anomalias, aprendizagem da regra de associação (modelo de dependência), clustering (agrupamento), classificação, regressão e sumarização. O processo de data mining costuma ocorrer utilizando dados contidos dentro do data warehouse.

Existem várias empresas e softwares que se dedicam à mineração de dados, pois a identificação de padrões em bancos de dados é cada vez mais importante. No entanto, a identificação de padrões relevantes não é exclusivo do mundo informático. O cérebro humano, utiliza um processo semelhante para identificar padrões e adquirir conhecimento.

Nos últimos anos, a mineração de dados tem sido amplamente utilizada nas áreas da ciência e engenharia, tais como bioinformática, genética medicina, educação e engenharia elétrica.

O conceito de data mining é muitas vezes associado à extração de informação relativa ao comportamento de pessoas. Por esse motivo, em algumas situações, a mineração de dados levanta aspectos legais e questões relativas à privacidade e ética. Apesar disso, muitas pessoas afirmam que a mineração de dados é eticamente neutra, pois não apresenta implicações éticas.

Exemplos reais de Data Mining

A mineração de dados é muitas vezes usada por empresas e organizações para a obtenção de conhecimento a respeito de utilizadores / funcionários / clientes. Por exemplo, no setor público é possível fazer o cruzamento de dados entre o estado civil de um funcionário e o salário que ele ganha, para verificar se isso tem influência na sua vida conjugal.

Empresas como cadeias de supermercados podem recorrer a esse cruzamento de dados para determinarem produtos que são comprados em conjunto. Se um cliente que compra o produto X também compra o produto Y, talvez seja uma boa ideia posicionar os dois produtos perto, para facilitar a compra por parte do cliente.

Qual a diferença entre Big Data, Data Warehouse e Data Mining? Embora sejam conceitos relacionados, não é correto afirmar que Data Mining, Big Data e Data Warehouse possuem o mesmo significado. O Big Data é caracterizado pela vasta quantidade de dados aleatórios produzidos a todo minuto no mundo inteiro.

O Data Mining é o reconhecimento de padrões dentro desses dados. Já o Data Warehouse é o banco de informações no qual todos esses resultados são armazenados. Etapas do Data Mining O processo de Data Mining ocorre através das seguintes etapas: Definição do problema A definição do problema é a primeira etapa do processo de Data Mining. Nessa fase o objetivo é entender o problema e estabelecer qual o objetivo que se deseja atingir com o processo de mineração. Exploração de dados é na exploração de dados que as ferramentas estatísticas básicas começam a ser utilizadas. Esta também é a etapa em que os especialistas coletam, descrevem e exploram os dados.

Além disso, a qualidade de todos os dados também é testada. Preparação de dados A preparação de dados é um processo que depende da origem dos mesmos. Assim, dependendo do estado em que os dados brutos se encontram, é necessário prepará-los através de métodos de filtração, combinação e preenchimento de valores vazios.

Modelagem esta etapa possui relação direta com o objetivo de cada processo de Mineração, pois é necessário escolher uma técnica de modelagem, dentro do Data Mining, que garanta a solução do problema proposto. Avaliação A avaliação é a fase mais crítica do processo, visto que é necessário a participação de um grupo de pessoas especializadas em Data Mining e no negócio alvo de análise para avaliar se a Mineração de Dados alcançou o resultado desejado.

Implementação

A implementação é a etapa final do projeto de Data Mining. É nessa fase que ocorre a importação dos resultados obtidos para os bancos de dados ou para outros tipos de diretórios. Técnicas de Data Mining

A Mineração de Dados é uma área muito extensa, dessa forma não há apenas uma maneira de encontrar padrões dentro um grande volume de dados. Abaixo você vai poder conferir quais são as principais técnicas utilizadas no momento de transformar dados em informações: Descoberta de regra de associação A descoberta de regras de associação é uma das técnicas mais utilizadas para a descoberta de conhecimento no Data Mining, visto que é possível extrair uma solução simples de casos complexos. Esta técnica consiste em analisar a relação entre os itens de um certo conjunto de dados e encontrar tendências e/ou padrões que possam ser utilizados para entender o comportamento desses dados.

Um exemplo muito popular e elucidativo sobre as regras de associações é o do supermercado. Segundo esta explicação, se uma pessoa vai ao supermercado comprar leite e pão ela também comprará manteiga. Dessa forma esta técnica é muito usual nas campanhas de marketing e no controle de estoques de centros comerciais, pois a compra de um produto A pode implicar na venda do produto B. Redes Neurais Artificiais As redes neurais artificiais (RNA) apresentam um modelo matemático baseado no sistema nervoso central. Este tipo de algoritmo busca resolver problemas através da simulação do comportamento e das funções de um neurônio. O seu funcionamento ocorre através de dezenas ou até centenas de unidades de processamento, as quais são interconectadas por canais de comunicação.

Dessa maneira, as entradas são semelhantes aos dendritos e simulam uma área de captação de estímulos. Já a saída de dados é comparada aos neurônios e o contato entre esses elementos formam a sinapse. Em algumas Redes Neurais a saída de um neurônio também pode se tornar um sinal de entrada de outro. Assim, as RNAs são capazes de gerar vários tipos de estruturas distintas. Árvores de Decisão As árvores de decisão funcionam como um fluxograma, porém possuem o formato de uma árvore. Através deste modelo, é possível que o usuário tome decisões a partir de inúmeras possibilidades de escolha.

Estas possibilidades são testadas automaticamente e funcionam da seguinte maneira: O nó representa dados ou problemas e cada ramificação possui um aglomerado de soluções baseadas em custos, probabilidades e benefícios. Áreas de aplicação do Data Mining Atualmente o Data Mining possui milhares de aplicações ao redor do mundo, logo este conceito está mais presente no seu dia a dia do que você pode imaginar.

Que tal dar uma olhada nos exemplos abaixo e descobrir como este artifício faz parte da sua rotina? Captação de clientes: identificação do perfil dos possíveis compradores de um determinado produto; Supermercado: alocação dos produtos nas prateleiras de acordo com o perfil de consumo de seus clientes; Segurança: detecção de atividades criminosas e terroristas; Telemarketing: captação de dados de possíveis clientes; Recursos Humanos (RH): análise das competências de um currículo; Banco: identificação de padrões que possam auxiliar no gerenciamento da relação com o cliente.