Mineração de Dados

1. **Introdução**

Atualmente as organizações têm se mostrado extremamente eficientes em capturar, e organizar e armazenar, grandes quantidades de dados, obtidas de diversas fontes de dados, porém, a maioria, ainda não usam adequadamente essa gigantesca massa de dados, para transformá-las em conhecimentos, para que possam ser utilizados, na tomada de decisões, em suas próprias atividades, sejam elas comerciais ou científicas (Han, 2011).

Estamos vivendo na era da informação. Portanto, *terabytes* ou *petabytes* de dados, são jogados nas redes de computadores, na World Wide Web (WWW), e vários outros dispositivos de armazenamento de dados oriundos de diversas fontes, tais como negócios, sociedade, de pesquisas científicas, medicina, e de quase todos os aspectos de nossa vida. Segundo Han (2011), este crescimento explosivo do volume de dados disponíveis é o resultado da informatização de nossa sociedade e, do rápido desenvolvimento de poderosas ferramentas de coleta e armazenamento de dados. Além do mais, as empresas comerciais no mundo, geram gigantesca quantidade de dados, incluindo transações de vendas, negociações de ações, descrição de produtos, promoções, perfis de empresas, feedbacks de clientes, e assim por diante (Han,2011).

Redes globais de telecomunicações (*Global Backbone*), transportam dezenas de *petabytes* de tráfico de dados todos os dias. A indústria médica e da saúde, geram enorme quantidade de dados: registros médicos, monitoramento de clientes, e imagens medicas. Bilhões de pesquisas suportadas pela Web, pelos mecanismos de busca, processam dezenas de *petabyte*s de dados diariamente. As redes sociais tornaram-se, cada vez mais, grandes fontes de dados, produzindo: imagens, vídeos, blogs, comunidades Web, e vários tipos de redes sociais. A lista de fontes que geram enorme quantidade de dados é interminável (Han, 2011).

Segundo Han (2011), este crescimento explosivo, amplamente disponível, de gigantescos conjuntos de dados, faz de nosso tempo, a era do dado. Ferramentas poderosas e versáteis, são extremamente necessárias para descobrir automaticamente, informações valiosas, da enorme quantidade de dados e transformá-las em conhecimento (Han, 2011). Essas necessidades, torna a mineração de dados (*data mining*), uma ferramenta cada vez mais forte, na busca de novos conhecimentos ocultos, em grandes massas de dados.

Segundo Zhou (2003), a área de Mineração de Dados (Data Mining) surgiu no final da década de oitenta, e focaliza a extração de conhecimento a partir de grandes volumes de dados usando o computador. Devido a sua natureza interdisciplinar, a pesquisa e desenvolvimento da área de Mineração de Dados, têm estreitas relações com as contribuições oferecidas por diversas áreas como Banco de Dados, Aprendizado de Máquina, Estatística, Recuperação de Informações, Computação Paralela e Distribuída. Essas são as áreas de conhecimento e pesquisa que estão contribuindo mais efetivamente para o desenvolvimento e o estabelecimento da área de Mineração de Dados (Zhou, 2003).

1. **Breve História**

Conforme Han (2011), na atualidade, o *Data mining (DM),* é usado em vasta áreas de conhecimento, e estão disponíveis, numerosos sistema comerciais de mineração de dados. No entanto, muitos desafios, ainda permanecem. No entanto, este é um campo de pesquisa que tem crescido rapidamente e, esse crescimento está diretamente ligado, a evolução natural da tecnologia da informação.

Os sistemas de banco de dados têm evoluído muito, a partir da década de 80, onde surgem os sistemas potentes de Bancos de Dados Relacionais, e com eles surgem também, os modelos avançados como relacional estendido, orientados a objeto, objeto-relacional, e os sistema de banco de Dados dedutivos. Essa evolução levou também, a uma evolução na coleta, na criação e gerenciamento de dados (incluindo data warehouse e data mining) [Han, 2011].

Segundo Han (2011), não podemos esquecer dos sistemas de banco de dados orientados pela aplicação, incluindo o espacial, o temporal, os de multimídia, os de fluxo de dados, os sistemas de banco de dados heterogêneos e os sistemas de informações globais baseados na Internet tais como a World Wide Web (WWW), que tem também emergido e executado um papel vital na indústria da informação.

Com todas essas tecnologias em evidencia, pode-se observar que, os dados podem ser armazenados em muitos diferentes tipos de repositórios de informações. No entanto, uma arquitetura de dados que tem emergido foi o **data warehouse (**Um repositório de muitas fontes de dados heterogêneas sobre um esquema unificado em um único local, para facilitar a tomada de decisões). Esta tecnologia inclui, a limpeza dos dados, integração de dados, e *on-line analytical processing* (**OLAP**). Uma técnica de análise de dados, com funcionalidades tais como: sumarização, consolidação, e agregação, bem como as facilidades de se visualizar as informações de ângulos diferentes.

Embora as ferramentas ***OLAP*** suportem a análise multidimensional e tomada de decisão, são necessárias ainda, as ferramentas de análise de dados para análises em profundidade, tais como: classificação de dados, caracterização, e agrupamento de dados (*clusting*), dentre outras (Han, 2011).

Como resultado, temos grandes repositórios de dados que raramente são consultados (Han, 2011). Consequentemente, importantes decisões são tomadas não com base em informações, mais com base nos dados armazenados nos repositórios, pelo simples fato de que, o gestor da instituição, não ter as ferramentas apropriadas para extrair os valiosos conhecimentos da vasta quantidade de dados. Então, neste contexto, as ferramentas de *Data Mining* podem ser utilizadas para realizar análise de dados, afim de encontrar importantes padrões nos dados, contribuindo enormemente para a estratégias de negócio, bases de conhecimentos, e pesquisa científicas, dentre outras [Han, 2011].

Resumindo, de acordo com Han (2011), temos dados em abundância e também a necessidade de ferramentas poderosas de análise de dados, ou seja, encontramo-nos em uma situação onde “o mundo está rico em dados, mas pobre em informações”. Então, as ferramentas de Data mining, que podem transformar, dados em informações, podem ser usadas, para diminuir o fosso entre dado e informação (Han, 2011).

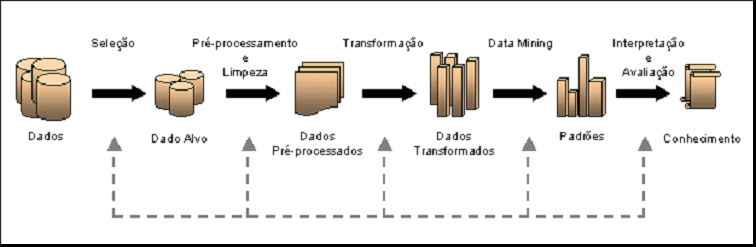
1. **Definições**

Não é surpresa que, Data Mining, como um assunto verdadeiramente interdisciplinar, possa ser definida de diferentes maneiras. Então, diversas definições de Mineração de Dados (*data mining*), podem ser encontradas na literatura. Entre as diversas definições, destaco as seguintes:

* Mineração de dados é a pesquisa por conhecimentos em grande massa de dados, onde o dado pode estar armazenado em diversas fontes de dados, tais como: Bancos de Dados, *Data Warehouse* (DW), ou outros repositórios de informações [Han, 2011].
* Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelo homem [Weis, 1999].
* Mineração de dados é a exploração de análise de dados, por meios automáticos ou semi- automáticos, em grandes quantidades de dados, com o objetivo de descobrir regras ou padrões interessantes [Berry, 1997].
* Mineração de dados é o processo de proposição de várias consultas e extração de informações úteis, padrões e tendências, frequentemente desconhecidos, a partir de grande quantidade de dados, armazenados em bancos de dados [Thuraisingham, 1999].
* Mineração de dados, de forma simples, é o processo de extração ou mineração de conhecimento, em grandes quantidades de dados [Han,2001].
* Mineração de dados, é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados Fayyad, Piatetsky-Shapiro & Smyth (1996ª).
* Mineração de dados, refere-se às atividades que analisam os dados, descobre problemas e oportunidades ocultos em seus relacionamentos, formam modelos computacionais com base nessas descobertas e, então, utilizam esses modelos para prever o comportamento do negócio – exigindo a mínima intervenção do usuário final [Rob, 2011].

Alguns autores e/ou pesquisadores, vêm *Data Mining*, simplesmente, como um dos passos do processo de descoberta de conhecimento. O processo de descoberta de conhecimento é mostrado na **Figura** 1 e consiste da sequência iterativa dos seguintes passos:

**Figura 1** O processo de descoberta de conhecimento em bases de dados. Fonte: Autor



1. **Integração de dados:** integrar dados de múltiplas fontes.
2. **Seleção de dados:** nesta fase é escolhido o conjunto de dados contendo todas as possíveis variáveis ou atributos que farão parte da análise. Normalmente essa escolha dos dados fica a critério de um especialista do domínio.
3. **Limpeza dos dados:** a fase do Pré-processamento e limpeza dos dados é uma parte crucial no processo de KDD, uma vez que, a qualidade dos dados, determinar a eficiência dos algoritmos de mineração. Nesta etapa deverão ser realizadas tarefas que eliminem dados redundantes e inconsistentes, recuperem dados incompletos e avaliem possíveis dados discrepantes ao conjunto, chamados de outliers.Também nesta fase**,** é importante o auxílio de um especialista do domínio, pois o mesmo é quem sabe se o valor é fruto de um erro de digitação ou um outliers.
4. **Transformação de dados:** nas grandes empresas, é comum ter diferentes gerenciadores de banco de dados, rodando em diferentes sistemas operacionais. Portanto, é fundamental para a mineração de dados, que estes dados que estão dispersos, sejam consolidados em um repositório único. Além disto, nesta fase, se necessário, é possível obter dados faltantes através da transformação ou combinação de outros, são os chamados “dados derivados”. Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um aluno, que pode ser encontrada a partir de sua data de nascimento.
5. **Data Mining:** é nesta fase onde acontece a exploração e análise, de forma automática ou semi-automática, dos dados com objetivo de descobrir padrões e regras.
6. **Avaliação dos padrões:** identificar os padrões verdadeiramente interessantes que representam conhecimento baseado em algumas medidas.
7. **Representação do conhecimento:**  nesta fase usa-se as técnicas de visualização, para representar o conhecimento obtido, para o usuário.
8. **Uma visão geral da mineração de dados**

Sobre uma visão multi-dimensional da mineração de dados, pode-se ter uma ideia geral da mineração de dados, em função dos seguintes aspectos: dados a serem minerados, conhecimento a ser extraído, técnicas utilizadas na mineração de dados e aplicabilidade da mineração de dados.

* **Dados**: os dados a serem minerados são oriundos de Dados do Banco de Dados: relacional, orientado a objeto, heterogêneo, data warehouse, dados transacionais, Streams, espaço-temporal, de séries temporais, texto e web, multimídia, gráficos e redes sociais e de informação.
* **Conhecimento extraído ou funções de mineração de dados**: o conhecimento extraído pode ser:
  + Caracterização, discriminação, associação, classificação, clustering, tendência / desvio, análise outlier, etc.
  + Descritivo ou preditivos
  + Pode-se integrar essas funções em vários níveis
* Técnicas utilizadas: as técnicas utilizadas para extração de conhecimento são as mais diversas, as mais utilizadas são:
  + data warehouse (OLAP), aprendizagem de máquina, estatísticas, reconhecimento de padrões, visualização de alto desempenho, etc.
* Aplicação: existe uma grande área de aplicabilidade da mineração de dados. Como exemplo, pode-se citar:
  + Varejo, telecomunicações, bancário, análise de fraude, mineração bio-dados, análise de mercado de ações, mineração de texto, mineração Web, etc.

1. **Funcionalidades**

As funcionalidades de Mineração de Dados (*data mining*), são usadas para especificar os tipos de padrões, que queremos encontrar, em uma tarefa de mineração de dados. Em geral, tais tarefas, podem ser classificadas em duas categorias: **descritiva** e **preditiva** [Han, 2011].

* **Descritiva**: descreve de forma concisa os dados disponíveis. Tarefas de mineração descritivas, caracterizam as propriedades gerais dos dados em um conjunto de dados de destino [Han, 2011].
* **Preditiva**: fazem inferências sobre os dados, objetivando predições [Han, 2011].

Dentre as muitas funcionalidades de mineração de dados, vamos destacar as seguintes:

* 1. **Associação e análise de correlação**

Um **Padrões frequentes**, como o nome sugere, são padrões que ocorrem frequentemente nos dados. Um padrão, tipicamente refere-se a um conjunto de itens que frequentemente aparecem juntos em um conjunto de dados transacional – por exemplo, leite e pão, que são frequentemente comprados juntos em mercearias, por muitos clientes [Han, 2011].

**Exemplo 1 Análise de Associação**. Suponha que você seja o gerente de marketing de uma organização, você gostaria de saber quais itens são frequentemente adquiridos juntos em uma mesma transação. Um exemplo de uma regra, para extrair, essas informações de um banco de dados transacional, é

**buys(X, “computer”) => buys(X, “software”) [support = 1%, confidence = 50%]**

Onde x é uma variável representando um cliente. Uma confiança (*confidence***)**, ou certeza, de 50% significa que, se um cliente compra um computador, há 50% de chance que ele irá comprar software também. Suporte (*Support****)*** igual a 1%, significa que 1% de todas as transações analisadas, mostram que computador e software são vendidos juntos. Esta regra de associação envolve um simples atributo ou predicado (i. e., buys) que repete. Regras de associação que contêm um simples predicado são referidos como ***single-dimensional association rules***. Tirando o predicado, a regra acima pode ser escrita simplesmente como “**computer => software [1%, 50%]**”.

* 1. **Classificação e Regressão**

Classificação é o processo de encontrar modelos (ou funções) que descreve e distingue classes de dados ou conceitos, com a finalidade de ser capaz de utilizar o modelo para prever a classe de objetos, cuja classe é desconhecida. O modelo derivado é baseado na análise do conjunto de dados de treinamento ***(training data***) (i.e., objetos de dados, cuja classe é conhecida) [Han,2011].

O modelo derivado pode ser apresentado de várias formas, tais como regras de classificação (***IF-THEN***), árvores de decisão, fórmulas matemáticas, ou redes neurais (**Figura** 1.5). Uma árvore de decisão (***decisiom tree***) é como um fluxograma em estrutura de árvore, onde cada nó denota um teste em um valor de atributo, cada ramo representa um resultado do teste, e as folhas da árvore representam as classes ou distribuição de classes. Árvore de decisão, pode facilmente, ser convertida para regra de classificação. Uma rede neural (***neural network***), quando usada para classificação, é tipicamente, uma coleção de neurônios com unidades de processamento, com conexões ponderadas entre as unidades. Existem muitos outros métodos de construção de modelos de classificação, tais como classificação ***naïve Bayesian***, máquina de vetor de suporte (***support vector machines***), e classificação do vizinho mais próximo (***k-nearest neighbor***).

**Figura 2** Um modelo de classificação pode ser representado de várias formas: (a) regras IF-THEN, (b) uma árvore de decisão, ou (c) uma rede neural.

**age(X, “youth”) AND income(X, “high”) 🡪 class(X, “A”)**

**age(X, “youth”) AND income(X, “low”) 🡪 class(X, “B”)**

**age(X, “middle\_aged”) 🡪 class(X, “C”)**

**age(X, “senior”) 🡪 class(X, “C”)**

1. Tendências
2. Conclusões

**(a)**



**(b)**



**(c)**

**5.3. Análise de Cluster**

Ao contrário de classificação e previsão, que analisam objetos de dados e classifica-os em classes, análise de agrupamento analisa objetos de dados e agrupa-os sem consultar um rótulo conhecido.

Em geral, a classe não é representada nos dados de treinamento, simplesmente porque eles não são conhecidos no início do processo. Entretanto, os agrupamentos (*clusters*) podem ser usados para gerar tais rótulos. Os objetos são agrupados baseados no princípio de máxima similaridade intra-classe e mínima similaridade inter-classe. Isto é, o agrupamento de objetos, é formado de modo que, os objetos dentro do agrupamento, tenham alta similaridade em comparação com um outro objeto, mas que sejam muitos diferentes dos objetos, em outro agrupamento. Cada agrupamento formado pode ser visto como, uma classe de objetos, da qual pode derivar regras.

**Exemplo 2 Análise de agrupamento**. Análise de agrupamento pode ser realizado nos dados dos clientes de uma organização para identificar subpopulações homogêneas de clientes. Esses agrupamentos podem representar grupos-alvos para o marketing.

5.4. **Análise de Outliers**

Um banco de dados pode conter objetos de dados que não estão em conformidade com o comportamento geral ou modelo de dados. Estes objetos de dados são ***outliers***. Muitos métodos de mineração de dados descartam *outliers* como sendo ruído ou exceções. Embora, em algumas aplicações, tais como: detecção de fraude, os eventos raros, podem ser mais interessantes do que os que ocorrem regularmente. A análise de dado *outlier*, é chamada como **outlier mining** [Han, 2011].

Os *Outliers* podem ser detectados, usando testes estatísticos que, assumem uma distribuição ou modelo de probabilidade para os dados, ou usando medidas de distância, onde objetos que estão a uma distância substancial de algum outro agrupamento, são considerados *outliers*. Ao invés de usar estatística ou medidas de distância, os métodos baseados em desvios, identificam valores discrepantes (outliers), examinando as diferenças nas características principais dos objetos nos grupos [Han, 2011].

**Exemplo** 3: Análise de discrepância, pode descobrir, uso fraudulento, em cartões de crédito, por detectar compras de quantidade extremamente grande, para um determinado número de contas, em comparação a compras regulares ocorridas nas mesmas contas. Valores discrepantes, também podem ser detectados, com relação a localização e tipo de compra, ou frequência de compra.

6. **Tecnologias usadas na Mineração de Dados**

Como uma aplicação altamente orientada ao domínio, a mineração de dados incorpora muitas técnicas de outros domínios, tais como estatísticas, aprendizado de máquina, reconhecimento de padrões, bancos de dados e sistemas de *data warehouse*, recuperação de informações, visualização, algoritmos, computação de alta performance, e muitos outros domínios de aplicação (**Figura 3**). A natureza interdisciplinar das pesquisas e desenvolvimento da mineração de dados, contribuiu significativamente com o sucesso da mineração de dados e suas amplas aplicações. A seguir, será dado exemplos de várias disciplinas que têm influenciado em muito, o desenvolvimento dos métodos da mineração de dados [Han, 2011].

**Figura 3** Mineração de dados adota técnicas de muitos domínios. **Fonte**: adaptado de Han (2013).

Data Mining

Estatísticas

Sistemas de Banco de Dados

Armazém de Dados

Recuperação de Informações

Aprendizado de Máquina

Reconhecimento de Padrões

Visualização

Algoritmos

Computação de Alto Desempenho

Aplicações

**6.1. Estatísticas**

Um modelo estatístico, é um conjunto de funções matemáticas, que descrevem as ações dos objetos, em uma classe alvo, em termos de suas variáveis aleatórias e sua distribuição de probabilidades associadas. Os modelos estatísticos são muito usados para modelar dados e classes de dados. Por exemplo, pode-se utilizar os modelos estatísticos para caracterização e classificação de dados em tarefas de mineração de dados. Em outras palavras, tais modelos estatísticos podem ser o resultado de uma tarefa de mineração de dados. Alternativamente, tarefas de mineração de dados podem ser construídas em cima dos modelos estatísticos. Por exemplo, podemos usar a estatística para modelar dados ausentes e ruídos. Então, quando extraído um padrão em uma grande massa de dados, o processo de mineração de dados, pode usar o modelo para ajudar a identificar e manipular dados ausentes ou ruídos [Han, 2011].

**6.2. Aprendizagem de Máquina**

Aprendizagem de máquina, investiga como os computadores pode aprender, com base em dados [Han, 2011].

Aprendizado de máquina, é uma área de Inteligência Artificial (IA), cujo objetivo, é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática [Rezende, 2003].

A principal área de pesquisa está voltada para programas de computador, aprenderem automaticamente reconhecer padrões complexos, com base em dados, e a partir daí, tomar decisões inteligentes. Por exemplo, um problema típico de aprendizado de máquina, é programar um computador para que ele possa reconhecer automaticamente os códigos postais escritos, depois de aprender a partir com um conjunto de exemplos [Han, 2011].

Aprendizado de máquina, é uma sub-área da Inteligência Artificial (IA), cujo objetivo é desenvolver métodos, técnicas e ferramentas para construir máquinas inteligentes, que se modificam para realizar cada vez melhor, sua tarefa [Mitchell, 1997].

Para aprender, os sistemas, bem como os seres humanos, podem se valer de estratégias de aprendizado. A seguir, será apresentado, algumas dessas estratégicas clássicas de aprendizagem de máquina relacionado a mineração de dados (*data mining*) [Han, 2011]:

* **Aprendizado supervisionado**: é basicamente um sinônimo para classificação. No aprendizado supervisionado o objetivo é induzir conceitos a partir de exemplos que estão pré-classificados, ou seja, exemplos que estão rotulados com uma classe conhecida. Se as classes possuírem valores discretos, o problema é caracterizado como classificação. Caso as classes possuam valores contínuos, o problema é caracterizado, como regressão [Han, 2011].
* **Aprendizado não-supervisionado**: é essencialmente, um sinônimo para agrupamento (*cluster*). No aprendizado não-supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando agrupamentos ou *clusters*. A tarefa do algoritmo é agrupara exemplos não rotulados, i.e., exemplos que não possuem o atributo classe especificado. Nesse caso, é possível utilizar algoritmos de aprendizado para descobrir padrões nos dados a partir de alguma caracterização de regularidade, sendo esses padrões denominados *clusters* [Decker & Focardi, 1995; McCallum, Nigan, & Ungar, 200]. O processo de formação dos clusters é geralmente conhecido por *clustering*. Tipicamente, pode-se usar agrupamento para descobrir classes dentro de dados [Han, 2011].
* **Aprendizado semi**-**supervisionado**: quando existem exemplos rotulados, pode-se utilizar o aprendizado supervisionado para induzir classificadores a partir destes exemplos. Caso contrário, quando os exemplos não estão rotulados, pode-se utilizar o aprendizado não-supervisionado com o objetivo de encontrar os *cluster*s. Já o aprendizado semi-supervisionado, consiste em utilizar algoritmos que aprendam a partir de exemplos rotulados e não rotulados. Ou seja, o aprendizado semi-supervisionado, pode ser aplicável tanto em tarefas de classificação quanto em tarefas de agrupamento [Han, 2011].
* **Aprendizado ativo**: é uma abordagem de aprendizado de máquina, que permite que o usuário execute um papel ativo no processo de aprendizagem. Uma abordagem de aprendizado ativo, pode pedir ao usuário (i.e., o especialista do domínio) para rotular um exemplo, que pode ser a partir de um conjunto de exemplos não-rotulados, ou sintetizado pelo programa de aprendizagem. O objetivo é otimizar a qualidade do modelo, através de aquisição de conhecimento ativamente do usuário humano, a partir da quantidade exemplos a serem rotulados [Han, 2011].

**6.3. Sistemas de Banco de Dados**

As pesquisas em Sistemas de Banco de Dados, se concentram na criação, manipulação e, uso do banco de dados, pela organização e usuários finais. Particularmente, os pesquisadores em sistemas de banco de dados, têm estabelecido princípios altamente reconhecidos em modelos de dados, linguagens de consultas, processamento de consulta, e métodos de otimização. Sistemas de banco de dados são geralmente bem conhecidos por sua alta escalabilidade no processamento de grandes conjuntos de dados, relativamente estruturados [Han, 2011].

Muitas tarefas de mineração de dados, precisam lidar com grandes massas de dados, ou até mesmo fluxo de dados (*stream*) em tempo real. De acordo com Han (2011), a mineração de dados, pode tirar proveito da escalabilidade das tecnologias de banco de dados, para alcançar alta eficiência e escalabilidade em grandes massas de dados. Além disso, as tarefas de mineração de dados, podem ser usadas para estender as capacidades dos sistemas de banco de dados existentes, para satisfazer as necessidades de análise de dados dos usuários avançados [Han, 2011].

Os Sistemas de banco de dados recentes, têm implementado capacidades de análise sistemática de dados, sobre os dados do banco de dados, utilizando *data warehouse* (DW) e facilidades de mineração de dados. Um *data warehouse* integra dados de muitas fontes. Ele consolida os dados em um espaço multidimensional, para forma cubos de dados. Os cubos de dados, facilitam operações OLAP sobre a base de dados multidimensional, permitindo uma melhor performance no processo de mineração de dados [Han,2011].

**6.4. Recuperação de informações**

É uma ciência de pesquisa sobre busca por informações em documentos. A busca pode ser realazada nos documentos propriamente ditos, através dos [**metadados**](https://pt.wikipedia.org/wiki/Metadado) que descrevam documentos e em [**banco de dados**](https://pt.wikipedia.org/wiki/Banco_de_dados), sejam eles relacionais e isolados ou banco de dados interligados em rede de [**hipermídia**](https://pt.wikipedia.org/wiki/Hiperm%C3%ADdia), tais como a [***World Wide Web***](https://pt.wikipedia.org/wiki/World_Wide_Web). A mídia pode estar disponível sob forma de textos, de sons, de imagens ou de dados [Han, 2011].

De acordo com Han (2011), cada vez mais, grandes quantidades de textos e dados multimídia são acumulados e disponibilizados on-line, devido ao rápido crescimento da Web e aplicações, tais como bibliotecas digitais, dados governamentais e sistemas de informações sobre saúde e, assim por diante. Para Han (2011), as pesquisas e análises nesta área, têm levantado muitas questões desafiadoras em mineração de dados. Han (2011), finaliza, afirmando que, a mineração de texto e a mineração de dados multimídia, integrados com métodos de recuperação de informação, tornaram-se cada vez mais importantes.

1. **Tendências**
2. **Conclusões**

A Mineração de dados, é o processo de descoberta de conhecimento em grande quantidade de dados. Como um processo de descoberta de conhecimento, ele tipicamente envolve, a limpeza dos dados, a integração dos dados, a transformação dos dados, a descoberta dos padrões, a avaliação dos padrões descobertos, a avaliação dos padrões extraídos e, a visualização, por parte dos usuários, do conhecimento extraído. Para isso, a mineração de dados, usa funcionalidades para especificar os tipos de padrões ou conhecimento a ser encontrado em uma tarefa de mineração de dados. Essas funcionalidades incluem caracterização e descriminação; a mineração de padrões frequentes, associações e correlações; classificação e regressão; análise de agrupamento; e detecção de *outlier*.

A mineração de dados (data mining), tem muitas aplicações bem sucedida, tais como *business intelligence* (**BI**), pesquisas na Web, bioinformática, saúde, sistemas financeiros, bibliotecas digitais, e informações governamentais digitais, e assim por diante [Han, 2011].