Mineração de dados

Origem: Wikipédia, a enciclopédia livre.

Prospecção de dados (português europeu) ou **mineração de dados** (português brasileiro) (também conhecida pelo termo inglês ***data mining***) é 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.

Esse é um tópico recente em ciência da computação mas utiliza várias técnicas da estatística, recuperação de informação, inteligência artificial e reconhecimento de padrões.

Índice

Visão geral

A mineração de dados 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, são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta de conhecimento. Esse conhecimento 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. A mineração de dados (MD) 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 inteligence (BI). Se analisarmos os dados com estatística de modo mais refinado, à procura de padrões de vinculações entre as variáveis registradas, então estaremos fazendo mineração de dados. Buscamos com a MD conhecer melhor os clientes, seus padrões de consumo e motivações. A MD resgata em organizações grandes o papel do dono atendendo no balcão e conhecendo sua clientela. Através da MD, 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 inteligence (BI) da mineração de dados (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.

Etapas da mineração de dados

Os passos fundamentais de uma mineração bem sucedida a partir de fontes de dados (bancos de dados, relatórios, logs de acesso, transações, etc.) consistem de uma limpeza (consistência, preenchimento de informações, remoção de ruído e redundâncias, etc.). Disto nascem os repositórios organizados (*Data Marts* e *Data Warehouses*).

É a partir deles que se pode selecionar algumas colunas para atravessarem o processo de mineração. Tipicamente, este processo não é o final da história: de forma interativa e frequentemente usando visualização gráfica, um analista refina e conduz o processo até que os padrões apareçam. Observe que todo esse processo parece indicar uma hierarquia, algo que começa em instâncias elementares (embora volumosas) e terminam em um ponto relativamente concentrado.

Encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados" de forma a desconsiderar aquilo que é específico e privilegiar aquilo que é genérico. Faz-se isso porque não parece haver muito conhecimento a extrair de eventos isolados. Uma loja de sua rede que tenha vendido a um cliente uma quantidade impressionante de um determinado produto em uma única data pode apenas significar que esse cliente em particular procurava grande quantidade desse produto naquele exato momento. Mas isso provavelmente não indica nenhuma tendência de mercado.

Localizando padrões

Padrões são unidades de informação que se repetem. A tarefa de localizar padrões não é privilégio da mineração de dados. O cérebro dos seres humanos utiliza-se de processos similares, pois muito do conhecimento que temos em nossa mente é, de certa forma, um processo que depende da localização de padrões. Para exemplificar esses conceitos, vamos propor um breve exercício de indução de regras abstratas. Nosso objetivo é tentar obter alguma expressão genérica para a seguinte seqüência:

```
Seqüência original: ABCXYABCZKABDKCABCTUABEWLABCWO
```

Observe atentamente essa seqüência de letras e tente encontrar alguma coisa relevante. Veja algumas possibilidades:

Passo 1: A primeira etapa é perceber que existe uma seqüência de letras que se repete bastante. Encontramos as seqüências "AB" e "ABC" e observamos que elas ocorrem com freqüência superior à das outras seqüências.

Passo 2: Após determinarmos as seqüências "ABC" e "AB", verificamos que elas segmentam o padrão original em diversas unidades independentes:

```
"ABCXY"
"ABCZK"
"ABDKC"
"ABCTU"
"ABEWL"
"ABCWO"
```

Passo 3: Fazem-se agora induções, que geram algumas representações genéricas dessas unidades:

```
"ABC??" "ABD??" "ABE??" e "AB???",
```

onde '?' representa qualquer letra

No final desse processo, toda a seqüência original foi substituída por regras genéricas indutivas, o que simplificou (reduziu) a informação original a algumas expressões simples. Esta explicação é um dos pontos essenciais da mineração de dados, como se pode fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução de informação, esse processo nos permite gerar formas de predizer futuras ocorrências de padrões.

Exemplo prático

Vamos observar aqui apenas um pequeno exemplo prático do que podemos utilizar com as expressões abstratas genéricas que obtivemos. Uma dessas expressões nos diz que toda vez que encontramos a seqüência "AB", podemos inferir que iremos encontrar mais três caracteres e isto completaria um "padrão". Nesta forma abstrata ainda pode ficar difícil de perceber a relevância deste resultado. Por isso vamos usar uma representação mais próxima da realidade.

Imagine que a letra 'A' esteja representando um item qualquer de um registro comercial. Por exemplo, a letra 'A' poderia significar "aquisição de pão" em uma transação de supermercado. A letra 'B' poderia, por exemplo, significar "aquisição de leite". A letra 'C' é um indicador de que o leite que foi adquirido é do tipo desnatado. É interessante notar que a obtenção de uma regra com as letras "AB" quer dizer, na prática, que toda vez que alguém comprou pão, também comprou leite. Esses dois atributos estão associados e isto foi revelado pelo processo de descoberta de padrões.

Esta associação já nos fará pensar em colocar "leite" e "pão" mais próximos um do outro no supermercado, pois assim estaríamos facilitando a aquisição conjunta desses dois produtos. Mas a coisa pode ir além disso, bastando continuar nossa exploração da indução.

Suponha que a letra 'X' signifique "manteiga sem sal", e que a letra 'Z' signifique "manteiga com sal". A letra 'T' poderia significar "margarina". Parece que poderíamos tentar unificar todas essas letras através de um único conceito, uma idéia que resuma uma característica essencial de todos esses itens. Introduzimos a letra 'V', que significaria "manteiga/margarina", ou "coisas que passamos no pão". Fizemos uma indução orientada a atributos, substituímos uma série de valores distintos (mas similares) por um nome só.

Ao fazer isso estamos perdendo um pouco das características dos dados originais. Após essa transformação, já não sabemos mais o que é manteiga e o que é margarina. Essa perda de informação é fundamental na indução e é um dos fatores que permite o aparecimento de padrões mais gerais. A vantagem desse procedimento é de que basta codificar a seqüência original substituindo a letra 'V' em todos os lugares devidos. Assim fica essa seqüência transformada:

```
ABCVYABCVKABDKCABCVUABEWLABCVO
```

Daqui, o sistema de mineração de dados irá extrair, entre outras coisas, a expressão "ABCV", que irá revelar algo muito interessante:

```
A maioria dos usuários que adquiriram pão e leite desnatado
também adquiriram manteiga ou margarina.
```

De posse desta regra, fica fácil imaginar uma disposição nas prateleiras do supermercado para incentivar ainda mais este hábito. Em linguagem mais lógica, pode-se dizer que pão e leite estão associados (implicam) na aquisição de manteiga, isto é, $Pao, Leite \rightarrow Manteiga$.

Exemplos Reais

Wal-Mart

Embora recente, a história da mineração de dados já tem casos bem conhecidos. O mais divulgado é o da cadeia estado-unidense Wal-Mart, que identificou um hábito curioso dos consumidores. Ao procurar eventuais relações entre o volume de vendas e os dias da semana, o software apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas. Crianças bebendo cerveja? Não. Uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer as reservas de cerveja para o final de semana.

```
Esta é uma história muito divulgada, porém não apresenta nenhuma confirmação real no processo de descobe
```

Vestibular PUC-RJ

Utilizando as técnicas da mineração de dados, um programa de obtenção de conhecimento depois de examinar milhares de alunos forneceu a seguinte regra: se o candidato é do sexo feminino, trabalha e teve aprovação com boas notas no vestibular, então não efetivava a matrícula. Estranho, ninguém havia pensado nisso. Mas uma reflexão justifica a regra oferecida pelo programa: de acordo com os costumes do Rio de Janeiro, uma mulher em idade de vestibular, se trabalha é porque precisa, e neste caso deve ter feito inscrição para ingressar na universidade pública gratuita. Se teve boas notas provavelmente foi aprovada na universidade pública onde efetivará matrícula. Claro que há exceções: pessoas que moram em frente à PUC, pessoas mais velhas, de alto poder aquisitivo e que voltaram a estudar por outras razões que ter uma profissão, etc.. Mas a grande maioria obedece à regra anunciada.

Ligações externas

- Mineração de dados (http://dmoz.org/Computers/Software/Databases/Data_Mining/) no Open Directory Project
- Programa de Mineração de Dados, Universidade da Florida Central (http://dms.stat.ucf.edu)
- Tutoriais e recursos em mineração de dados (http://www.eruditionhome.com)
- Tutoriais de Andrew Moore da Universidade Carnegie Mellon (http://www.autonlab.org/tutorials)
- Data Mining Blog - Conceitos e Exemplos Práticos (http://www.fp2.com.br/datamining)

Software

- Enterprise Miner (http://www.sas.com/technologies/analytics/datamining/miner/) , ferramenta de data mining do SAS (http://pt.wikipedia.org/wiki/SAS_(inform%C3%A1tica))
- Microsoft SQL Server (http://www.microsoft.com/sql/) , ferramenta originalmente de banco de dados que a cada nova versão tem ganho novas funcionalidades de Business Intelligence. Possui 8 algorítmos na versão do SQL Server 2008 e sua plataforma é extensível para integração de outros algorítmos desenvolvidos.
- IlliMine (http://illimine.cs.uiuc.edu) Projeto de mineração de dados escrito em C++.
- InfoCodex (http://www.infocodex.com) Aplicação de mineração de dados com uma base de dados

linguística.
- KDB2000 (http://www.di.uniba.it/~malerba/software/kdb2000/) Uma ferramenta livre em C++ que integra acesso à bases de dados, pre-processamento, técnicas de transformação e um vasto escopo de algoritmos de mineração de dados.
- KXEN (http://www.kxen.com/) Ferramenta de mineração de dados comercial, utiliza conceitos do Profesor Vladimir Vapnik como Minimização de Risco Estruturada (*Structured Risk Minimization* ou SRM) e outros.
- KNIME (http://www.knime.org) Plataforma de mineração de dados aberta que implementa o paradigma de *pipelining* de dados. Baseada no eclipse (http://www.eclipse.org)
- LingPipe (http://www.alias-i.com/lingpipe) API em Java para mineração em textos distribuída com código-fonte.
- MDR (http://www.epistasis.org/open-source-mdr-project.html) Ferramenta livre em Java para detecção de interações entre atributos utilizando o método da *multifactor dimensionality reduction* (MDR).
- Orange (http://www.ailab.si/orange) *Tookit* livre em Python para mineração de dados e aprendizado de máquina.
- Pimiento (http://www.ee.usyd.edu.au/~jjga/pimiento) Um ambiente para mineração em textos baseado em Java.
- PolyaAnalyst (http://www.megaputer.com/polyanalyst.php) Ambiente que permite a montagem de fluxos para mineração de dados e texto.
- Tanagra (http://chirouble.univ-lyon2.fr/~ricco/tanagra/) Software livre de mineração de dados e estatística.
- WEKA (http://www.cs.waikato.ac.nz/ml/weka/) Software livre em java para mineração de dados.
- Cortex Intelligence (http://www.cortex-intelligence.com) Sistema de PLN para mineração de textos aplicado à Inteligência Competitiva

Obtido em "http://pt.wikipedia.org/wiki/Minera%C3%A7%C3%A3o_de_dados"
Categorias: Palavras que diferem em versões da língua portuguesa | Mineração de dados | Data warehouse | Business intelligence | Bancos de dados