Publicado nos Anais do Infoimagem 2002, Cenadem, Novembro

Princípios Essenciais do Data Mining

Sergio Navega
snavega@attglobal.net
Intelliwise Research and Training
R. Rosa S. Capelli, 114 - São Paulo - SP - 04725-050
http://www.intelliwise.com/snavega
Agosto de 2002

Resumo

Desde que a informática tomou conta de nossas vidas, imensos volumes de informação têm sido sistematicamente coletados e armazenados. A simples armazenagem e recuperação dessa informação já traz um grande benefício, pois agora já não é mais necessário procurar informação em volumosos e ineficazes arquivos de papel. Contudo, apenas recuperar informação não propicia todas as vantagens possíveis. O processo de Data Mining permite que se investigue esses dados à procura de padrões que tenham *valor* para a empresa. Neste pequeno artigo pretendo expor alguns dos principais conceitos que estão por trás dessa importante tecnologia.

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. Contudo, embora essa idéia básica seja facilmente compreensível, fica sempre uma dúvida sobre como um sistema é capaz de obter esse tipo de relação. No restante deste artigo vamos observar alguns conceitos que podem esclarecer essas dúvidas.

O Que É Data Mining?

Talvez a definição mais importante de Data Mining[1] tenha sido elaborada por Usama Fayyad (Fayyad et al. 1996):

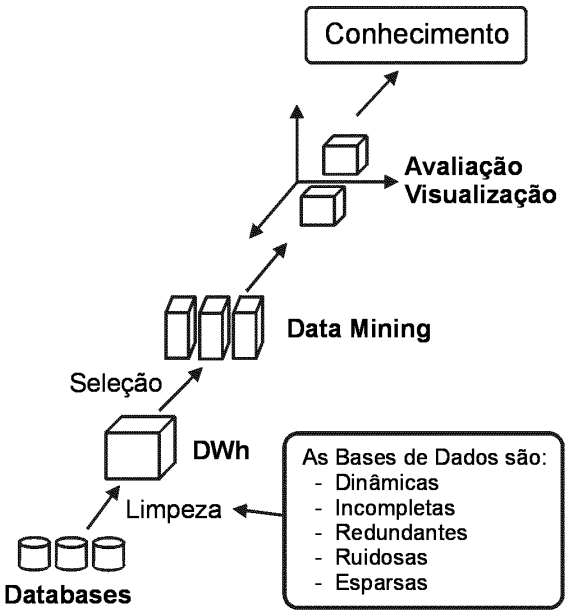
> "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis"

Esse processo vale-se de diversos algoritmos (muitos deles desenvolvidos recentemente) que processam os dados e encontram esses "padrões válidos, novos e valiosos". É preciso ressaltar um detalhe que costuma passar despercebido na literatura: embora os algoritmos atuais sejam capazes de descobrir padrões "válidos e novos",

ainda não temos uma solução eficaz para determinar padrões *valiosos*. Por essa razão, Data Mining ainda requer uma interação muito forte com analistas humanos, que são, em última instância, os principais responsáveis pela determinação do valor dos padrões encontrados[2]. Além disso, a condução (direcionamento) da exploração de dados é também tarefa fundamentalmente confiada a analistas humanos, um aspecto que não pode ser desprezado em nenhum projeto que queira ser bem sucedido.

Os Passos do Data Mining

A literatura sobre o assunto trata com mais detalhes todos os passos necessários ao Data Mining. Veja, por exemplo, Groth (1998) e Han, Chen & Yu (1996). Para o escopo do que pretendemos neste artigo é suficiente apresentar os passos fundamentais de uma mineração bem sucedida (veja figura à direita). A partir de fontes de dados (bancos de dados, relatórios, logs de acesso, transações, etc) efetua-se uma limpeza (consistência, preenchimento de informações, remoção de ruído e redundâncias, etc). Disto nascem os repositórios organizados (Data Marts e Data Warehouses), que já são úteis de diversas maneiras. Mas 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 valiosos 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, mas muito valioso.

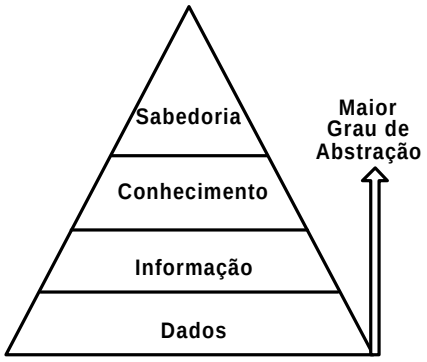


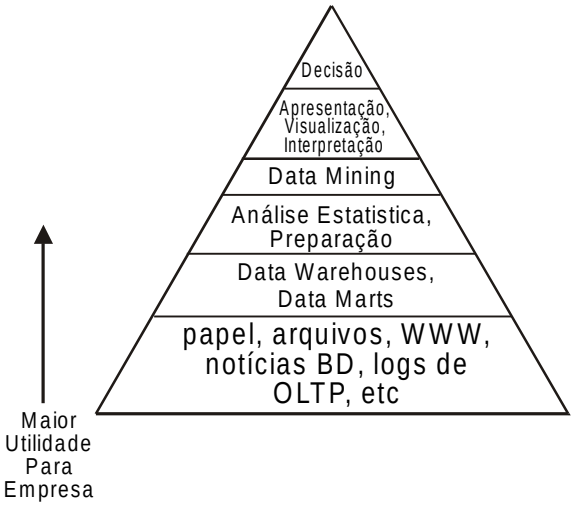
Este é um dos conceitos importantes para nós neste artigo: encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados" de forma a desconsiderar aquilo que é específico e privilegiar aquilo que é genérico. Faz-se isso porque não parece haver muito conhecimento a extrair de eventos isolados. Uma loja de sua rede que tenha vendido a um cliente em particular 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.

Em outras palavras, não há como explorar essa informação em particular para que no futuro a empresa lucre mais. Apenas com *conhecimento genérico* é que isto pode ser obtido. Por essa razão devemos, em Data Mining, controlar nossa vontade de "não perder dados". Para que o processo dê certo, é necessário sim desprezar os eventos particulares para só manter aquilo que é genérico.

Dos Dados à Sabedoria

Assim como um organismo vivo, as empresas recebem informação do meio ambiente e também atuam sobre ele. Durante essas atividades, é necessário distinguir vários níveis de informação. O diagrama à esquerda apresenta a tradicional pirâmide da informação, onde se pode notar o natural aumento de abstração conforme subimos de nível. Traduzido para uma empresa atual, esse diagrama fica como apresentado abaixo. O fundamental a se perceber neste diagrama é a sensível redução de volume que ocorre cada vez que subimos de nível. Essa redução de volume é uma natural consequência do processo de abstração. Abstrair, no sentido que usamos aqui, é representar uma informação através de correspondentes simbólicos e genéricos. Este ponto é importante: como acabamos de ver, para ser genérico, é necessário "perder" um pouco dos dados, para só conservar *a essência* da informação. O processo de Data Mining localiza padrões através da judiciosa aplicação de processos de generalização, algo que é conhecido como *indução*. Na próxima seção vamos ver este processo um pouco mais de perto.





Localizando Padrões

Padrões são unidades de informação que se repetem, ou então são sequências de informações que dispõe de uma *estrutura* que se repete. A tarefa de localizar padrões não é privilégio do Data Mining. Nosso cérebro utiliza-se de processos similares, pois muito do conhecimento que temos em nossas mentes é, de certa forma, um processo que depende da localização de padrões[3]. Por essa razão, muito do que se estuda sobre o cérebro humano também pode nos auxiliar a entender o que deve ser feito para localizar padrões. Mas o que é mesmo localizar padrões? O que é indução?

Para exemplificar esses conceitos, proponho um breve exercício de uma indução de regras abstratas[4]. Nosso objetivo é tentar obter alguma expressão genérica para a seguinte sequência:

Sequência original: ABCXYABCZKABDKCABCTUABEWLABCWO

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

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

Passo 2: Após determinarmos as sequências "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 sequência original foi substituída por regras genéricas indutivas[5] que simplificou (reduziu) a informação original a algumas expressões simples. Se você compreendeu esta explicação até aqui, então você acaba de conhecer um dos pontos essenciais do Data Mining: como se pode fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução (compressão) de informação, esse processo nos permite gerar formas de *predizer* futuras ocorrências de padrões. Este é exatamente o ponto onde este processo começa a mostrar o seu valor.

Um Exemplo Prático

Existem muitas técnicas utilizadas pelo Data Mining, muitas delas desenvolvidas na disciplina Aprendizado de Máquina (Machine Learning, veja, por exemplo, Mitchell 1997). Vamos observar aqui apenas um pequeno exemplo prático do que podemos utilizar. Lembre-se das expressões abstratas genéricas que obtivemos na seção anterior. Uma dessas expressões nos diz que toda vez que encontramos a sequência "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. É o que faremos a seguir.

Indução Orientada a Atributos

Continuando com nosso exemplo acima, suponha que a letra X queira dizer "manteiga sem sal", e 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"[6]. Fizemos uma *indução orientada a atributos*, substituímos uma série de valores distintos (mas similares) por *um nome só*.

Observe que 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*.

Qual a vantagem de assim proceder? Basta codificar nossa sequência original substituindo a letra V em todos os lugares devidos. Assim fica essa sequência transformada:

ABCVYABCVKABDKCABCVUABEWLABCVO

Daqui, nosso sistema de Data Mining irá extrair, entre outras coisas, a expressão "ABCV", que nos irá revelar de pronto 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[7]. Em linguagem mais lógica, pode-se dizer que pão e leite estão associados (implicam) na aquisição de manteiga:

$$\text{Pão, Leite} \Rightarrow \text{Manteiga}$$

O lado da esquerda desta expressão (Pão, Leite) é chamado de *Antecedente*, e o lado da direita de *Consequente*.

Mais Técnicas

Introduzimos os exemplos anteriores apenas para dar uma idéia do tipo de pensamento que está por trás da mineração de dados. Faz-se certas induções e descobre-se alguns padrões. Vamos agora ver algumas outras técnicas que se utilizam de princípios similares.

Regras Caracterizadoras

Obtém-se regras que caracterizam um conceito satisfeito por todos (ou pela maioria) dos exemplos disponíveis. Assim, é possível descobrir formas de *sumarizar* certas características que podem revelar padrões nos dados. Exemplos:

a) Sintomas de uma doença específica podem ser sumarizados por uma regra caracterizadora

b) Geração de regras que caracterizem quais os estudantes de graduação que se decidiram por prosseguir com uma carreira acadêmica (MBA, doutorado).

Regras Discriminantes
Neste caso, o que se almeja é obter regras que discriminem (separem) um conceito alvo em relação a outros conceitos (classes contrastantes). Exemplo:

a) Para distinguir uma doença, procura-se por regras que sumarizem as características que separam esta doença das outras.
b) Tenta-se achar as regras que discriminem uma loja bem sucedida de várias outras não tão bem sucedidas.

Regras Associativas
Este é o caso que analisamos anteriormente. Aqui procura-se estabelecer regras que interliguem um conceito a outro. A utilidade deste procedimento é muito grande, conforme pode ser visto nos exemplos abaixo:

a) Achar todas as regras que tenham "coca-cola dietética" como *consequentes*. Isto irá auxiliar no planejamento de lojas para vender melhor este produto (privilegiam-se os *antecedentes* dessas regras).
b) Achar todas as regras que tenham "iogurte" no antecedente. Isto irá auxiliar na determinação do impacto nas receitas, caso este produto seja retirado das prateleiras.
c) Achar todas as regras com "salsicha" no antecedente e "mostarda" no consequente. Isto irá auxiliar na obtenção de melhores regras para determinar quais os itens que devem ser vendidos em conjunto com salsichas para aumentar as vendas de mostarda.

Regras de Evolução Temporal
Aqui a preocupação é detectar associações entre itens ao longo do tempo. Descobre-se padrões de compras após um evento inicial de aquisição. Exemplos:

a) Consumidor comprou um PC hoje, irá comprar um DVD-ROM em 6 meses. Isto permite que se faça uma oferta desse produto a todos os que estão nesta situação.
b) Um consumidor adquiriu um videocassete, em 4 meses terá muita probabilidade de comprar uma camcorder. Faz-se uma promoção especial para estes clientes.

Conclusões

No breve espaço deste artigo, minha principal missão foi introduzir um pouco do pensamento que está por trás do Data Mining. Obviamente, ainda há muito a se falar sobre o assunto (clustering, redes neurais, métodos genéticos, mineração em textos, roll up/drill down, etc), mas é importante notar que em praticamente todos esses casos o que se deseja é descobrir padrões em volumes de dados. É importante ressaltar também que o Data Mining não é o final da atividade de descoberta de conhecimentos, mas é tão somente o início. É imprescindível (ao menos com a tecnologia atual) dispor de analistas capacitados que saibam interagir com os sistemas de forma a conduzi-los para uma extração de padrões úteis e relevantes.

Referências

Allee, Verna (1997) The Knowledge Evolution. Butterworth-Heinemann Publishing

Buntine, Wray (1995) Learning and Probabilities, in Mlnet Summer School on Machine Learning and Knowledge Acquisition.

Dean, Thomas; Allen, James; Aloimonos, Yiannis (1995) Artificial Intelligence, Theory and Practice. The Benjamin/Cummins Publishing Company, Inc.

Elder IV, John F. (1999) The Interface 98 Conference - A Resource for KDD,in SIGKDD Explorations, Vol 1, Issue 1

Fayyad, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM, pp.27-34, Nov.1996

Groth, Robert (1998) Data Mining, A Hands-on Approach for Business Professionals. Prentice-Hall PTR.

Han, Jiawei; Fu, Yongjian (1995) Discovery of Multiple-Level Association Rules From Large Databases. Proceedings of 21st VLDB Conference.

Han, Jiawei; Fu, Yongjian (1996) Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases

Han, Jiawei; Chen, Ming-Syan; Yu, Philip S. (1996) Data Mining: An Overview from Database Perspective

Han, Jiawei (1996) Mining Knowledge at Multiple Concept Levels. Invited Talk.

Hand, David J. (1999) Statistics and Data Mining: Intersecting Disciplines, in SIGKDD Explorations, Vol 1, Issue 1

Holsheimer, M; Siebes, A. P. (1994) Data Mining: The Search for Knowledge in Databases. CS Dept. Centrum Voor Wiskunde en Informatica CS-R9406

Hosking, Jonathan R. M.; Pednault, Edwin P.; Sudan, Madhu (1997) A Statistical Perspective on Data Mining. IBM Research Report RC20856

John, George H. (1999) Behind-the-Scenes Data Mining: A Report on the KDD-98 Panel in SIGKDD Explorations, Vol 1, Issue 1

Michie, Donald (1982) The State of the Art in Machine Learning. In Introductory Readings in Expert Systems. Gordon and Breach, New York

Mitchell, Tom (1997) Machine Learning. The McGraw-Hill Company, Inc.

Munakata, Toshinori (1998) Fundamentals of the New Artificial Intelligence. Springer-Verlag New York Inc.

Navega, Sergio C. (2000) Inteligência Artificial, Educação de Crianças e o Cérebro Humano. Publicado em Leopoldianum, Revista de Estudos de Comunicações da Universidade de Santos (Ano 25, No. 72, Fev. 2000, pp 87-102). Disponível em http://www.intelliwise.com/reports/p4port.htm

Navega, Sergio C. (2002) Projeto CYC: Confundindo Inteligência com Conhecimento. In: KMBrasil 2002, 3° Workshop Brasileiro de Inteligência Competitiva. Disponível em http://www.intelliwise.com/reports/kmbscn.htm

Navega, Sergio C. (in press) Pensamento Crítico e Argumentação Sólida. Intelliwise Publicações. Trechos em http://www.intelliwise.com/books

Nonaka, Ikujiro; Takeuchi, Hirotaka (1995) The Knowledge-Creating Company. Oxford University Press

Nilsson, Nils J. (1998) Artificial Intelligence, A New Synthesis. Morgan Kauffmann Publishers, Inc.

Park, Jong Soo; Chen, Ming-Syan; Yu, Philips S. (1997) An Effective Hash Based Algorithm for Mining Association Rules

Russell, Stuart, Norvig, Peter (1995) Artificial Intelligence, A Modern Approach, Prentice-Hall, Inc.

Waltz, David; Hong, Se June (1999) Data Mining: A Long-Term Dream. IEEE Intelligent Systems Vol 14, No. 6.

Notas

[1] Na verdade, a definição que mostro aqui foi apresentada por Fayyad et al. (1996) para explicar o termo KDD (Knowledge Discovery in Databases), um processo que engloba a mineração. Portanto, Data Mining seria apenas um dos passos necessários ao processo todo.

[2] Este tópico é complicado porque requer que as máquinas tenham *senso comum*, uma das pedras no sapato dos cientistas que estudam a Inteligência Artificial. Só poderemos ter máquinas com senso comum quando o problema da aquisição automática de conhecimentos a partir de sinais sensórios for efetivamente resolvido. Isto requer, como ponto de partida, que se redefinam conceitos tradicionais, como *conhecimento* e *inteligência*. Este assunto está melhor explorado em Navega (2002).

[3] Na verdade, é preciso reconhecer que a geração de conhecimento também é feita através do pensamento e da reflexão. Esta posição é conhecida nos meios filosóficos como *Racionalismo*. A proposição de que o conhecimento vem dos sentidos é a posição do *Empirismo*. Este assunto faz parte de um ramo da filosofia conhecido por Epistemologia e está ligado a parte da Ciência Cognitiva. Uma exposição mais abrangente destes assuntos pode ser vista em Navega (in press) e em Navega (2000), acessível em http://www.intelliwise.com/reports

[4] Disponho de um programa que permite testar na prática uma parte deste exemplo. Você pode fazer o download do programa Simple Miner no endereço http://www.intelliwise.com/programs. Consulte o autor (snavega@attglobal.net) caso tenha dúvidas sobre a utilização deste programa.

[5] Essas expressões são indutivas porque elas generalizam *além* do que os dados informam. Em termos dedutivos, a expressão "ABC??", por exemplo, só poderia garantir que os caracteres XY, ZK, TU e WO fizessem parte da expressão, pois esses são os únicos que encontramos na expressão original. Mas a indução leva isso a *todos* os caracteres do alfabeto, de forma que a expressão "ABCNR" também poderia ser esperada a partir da expressão genérica que obtivemos. Aqui pode ocorrer um dos problemas filosóficos relacionados à indução: só podemos esperar *boa probabilidade* de nossas predições, nunca certezas. Este tema está tratado em maior detalhe no livro de Navega (in press).

[6] Esse caso demonstra porque no Data Mining ainda é necessária a presença humana. A introdução desse atributo requer um conhecimento do mundo (ou, mais frequentemente, *um conhecimento do domínio do negócio* da empresa) que as máquinas ainda não dispõe. Talvez o futuro do Data Mining seja associar-se a sistemas de Inteligência Artificial que possam suprir parte dessa deficiência.

[7] Cabe notar aqui que nem sempre os padrões detectados pelo Data Mining precisam ser compreensíveis. Mesmo que não tenhamos uma "explicação lógica" para esta associação, sabemos que se a utilizarmos estaremos melhorando a rentabilidade de nosso negócio. Embora o ideal seja sempre entender a razão disso, por enquanto o simples uso dessa informação já é suficiente.