Cadernos da Controladoria

Nova série Ano VII, nº 2 - junho de 2007

Cérebro, Cognição e Data Mining

Apresentação

Boa tarde. Iniciamos mais uma palestra dos Seminários da Controladoria e gostaria de agradecer a presença de todos e, em especial, a presença do representante da Secretaria de Controle Externo, do Tribunal de Contas da União, pela curiosidade sobre o assunto Data Mining, que é a minha também.

O Dr. Sergio Navega, nosso convidado de hoje, é formado em Física pela Universidade de São Paulo, foi consultor em diversos projetos de organizações públicas e privadas, como a TeleCeará, a Gazeta Mercantil, o Banco Sudameris, o Tribunal Superior Eleitoral, entre outras, onde desenvolveu projetos de software de sistemas transacionais, compiladores de linguagem de programação em português e emuladores gráficos multimídia.

Ele também participou da CEBIT 88, na Alemanha, onde demonstrou seus sistemas transacionais para paises do Leste Europeu. Atualmente é consultor em inteligência artificial, Data Mining e engenharia do conhecimento. Ele profere seminários e conferências sobre criatividade, pensamento crítico, decisão racional, argumentação e ciência cognitiva. Sobre esses assuntos já fez palestras para a Caixa Econômica Federal, OPP Petroquímica, Ericsson, Nestlé, Embrapa, Dow Química, Amanco, Informal Informática, TS-Tech Brasil e na pós-graduação da Unicamp, da Uniuberaba e de outras universidades.

O Dr. Sergio participa de diversas sociedades científicas internacionais, entre as quais a AAAI - American Association for Artificial Intelligence e Cognitive Science Society. É diretor fundador da Intelliwise, empresa paulista de pesquisa em ciência cognitiva e seminários corporativos, e coordena o Grupo I3P, que desenvolve sistemas inteligentes de análise de imagem. Ele foi convidado para proferir a palestra de hoje sobre Data Mining. E o nosso convite decorreu, como explicava a ele antes de iniciar este encontro, do fato de a Controladoria e a Prefeitura terem um razoável conjunto de dados. Agora queremos saber o que fazemos, como extraímos as informações do banco de dados e trabalhamos com o que está ali. Se extrairmos tudo, certamente teremos páginas e páginas de relatórios que não caberiam nesta sala. A grande dificuldade, para aqueles que trabalham em auditoria e em sistemas de controle, é saber exatamente como pinçar, de um conjunto enorme de dados, aquilo que nos interessa, que será útil para que o trabalho na área de controle e de todas as demais áreas seja eficiente. Assim, passo a palavra ao Dr. Sergio Navega, para a sua palestra que tem o título de "Cérebros, Cognição e Data Mining".

Lino Martins da Silva

Controlador Geral do Município


Cérebros,
Cognição e
Data Mining


Cérebro, Cognição e Data Mining

Sergio Navega

Rio de Janeiro, 27 de junho de 2007

Muito obrigado, Dr. Lino. É um grande prazer estar aqui com vocês. Vamos falar hoje sobre pelo menos três tipos de assunto: em um deles vamos nos deter mais - o Data Mining. Para entender um pouco desse assunto temos dois tipos de abordagem: podemos ingressar pelo aspecto da informática e podemos entrar pelo aspecto mais filosófico. É a partir deste último que vou tentar seguir.

Para chegar ao Data Mining começo refletindo sobre o nosso pensar, sobre como as pessoas, os animais que têm cérebro, desenvolveram algo que podemos chamar de forma genérica de inteligência. A partir deste caminhar, por meio deste entendimento do que seria inteligência, chegamos ao potencial do Data Mining. Com isto quero deixar mais ou menos claro que o Data Mining é conseqüência de um processo de acumular dados de um lado e de gerar padrões do outro lado. Vamos ver de que forma fazemos isso na vida diária - não apenas nós, mas outros tipos de animais que disponham de cérebro. Como um camundongo ou até mesmo uma abelha consegue fazer um pouco desse processo. Entendendo isso, que é algo próximo da nossa vida diária, vamos conseguir chegar mais perto da idéia do Data Mining e, com o conceito na cabeça, vamos conseguir entender como integrar isso em termos computacionais.

Começamos dando uma olhada no mundo de assuntos que devemos considerar. Algumas vezes terei que ser um pouco rápido por questão de tempo, pois cada um desses tópicos poderia ser elaborado em palestras de três horas. Primeiro a definição do que seria o regular e o caótico, porque tem uma ligação com a realidade do universo, é um ponto de partida bastante filosófico. Depois vamos ver níveis de análise: de que forma este tópico sobre o qual não se fala muito, mesmo em ambiente universitário, é essencial quando queremos entender assuntos complexos.

Vamos nos deter um pouco em organismos inteligentes e, então, falar de processos cognitivos. A partir deste detalhe dos processos cognitivos vamos introduzir outro aspecto ligado à filosofia, mais precisamente à epistemologia, que é conceituar dedução e indução. E vamos ver como são correlatas a dedução e a indução, principalmente a dedução na área de informática, que é o nosso gancho para entrar em Data Mining e explicar o que seria a idéia desse processo. Talvez muitos de vocês já conheçam o que é Data Mining em si, mas não sabem bem como opera ou de que maneira produz o que queremos. Vamos dar alguns exemplos de processos com o Data Mining e todos serão de mais ou menos superficiais. Não entraremos no mérito das técnicas em particular, mas veremos de que maneira elas funcionam em um aspecto mais abrangente e tratar de alguns conceitos essenciais da questão da mineração. Vamos comentar um pouco as regras que podemos usar para manipular as informações, depois falaremos de alguma divisão de processos supervisionados e não supervisionários. Falaremos de regras de evolução temporal e, finalmente, um pequeno exemplo prático, bastante simples, mas que dá uma idéia do que é tudo isso.

Começamos com esta idéia bem filosófica de que se pararmos um pouco para observar o universo em nosso entorno vamos verificar dois tipos de coisas relativamente distantes uma da outra, vamos ver que o universo é cheio de processos bastante caóticos. Um exemplo típico de exemplos caóticos que costumo dar é quando você pega um material radioativo e chega com um contador geiger próximo deste material radioativo: vão aparecer aqueles cliques que são típicos de uma emissão randômica. Ninguém consegue prever quando vai ocorrer o próximo estalido no contador geiger, é um processo essencialmente randômico. Isto faz parte não só de materiais radioativos, mas de muitas realidades do universo onde há eventos randômicos ocorrendo. Do outro lado dessa moeda temos as coisas regulares: o que está nesta tela, por exemplo, é o que acontece com o fluxo de ar no meio de uma gota de glicerina; ao lado temos um depósito metálico cristalino; aqui temos o crescimento bacteriano em uma chapa de Agar e aqui uma fagulha elétrica, uma quebra dielétrica em plásticos.

Todos esses processos podem ser vistos como regulares de alguma espécie. Posso verificar que cada braço desta árvore (imaginem a quebra dielétrica em plástico) tem similaridades com outros braços, então existe uma semelhança. No entanto, cada braço é único em um aspecto, nenhum deles é exatamente igual. Ao mesmo tempo estamos vendo algo que é caótico de um lado e regular de outro. Isto nos posiciona como elementos inteligentes no universo. De um lado temos o que é randômico, o que é ruído, o que não faz sentido; e de outro lado temos o que é a regularidade total. Inseri um desenho exemplificando o que é esta regularidade, mas temos diversos outros exemplos de coisas regulares. O nascimento diário do sol é um deles: todos os dias ele se levanta e se põe, essa é uma regularidade intensa, como as coisas que acontecem regularmente em termos de meteorologia, não é mesmo?

Dentro deste aspecto randômico e regular temos a mente humana trabalhando. Esta é uma primeira dica da necessidade de cérebros. Os cérebros são elementos que de alguma maneira procuram obter, dentro de um universo caótico, coisas que sejam regulares, que nos permitam predizer as coisas. Predizemos diariamente: hoje, por exemplo, quando estava em São Paulo vindo para cá, o pessoal do aeroporto informou que o Santos Dumont estava fechado por questão de neblina. Era uma questão de tempo: sabemos, pela regularidade das coisas, que isso em geral é uma coisa temporária, que em seguida vai mudar, o tempo não vai ficar eternamente fechado, logo se abre, como realmente se abriu. Então, sem termos noção, contamos com essa previsibilidade das coisas, capturamos o que é regular do universo e tentamos usar na nossa vida diária.

Com isso conseguimos um entendimento do universo, observar esse universo, predizer algumas coisas e escrever certas leis, como as leis de Newton, que são expressões da regularidade do universo. Essas expressões de regularidade nos fazem ter uma noção do que é esse mundo. Nesse ponto introduzo uma idéia muito importante chamada "níveis de análise" ou "níveis descritivos". Isto é essencial sempre que estudamos alguma coisa, que percebemos em que nível de análise estamos observando algo. Vamos ver alguns exemplos onde esses níveis são mais claros. Aqui nesta imagem tenho basicamente o nível de análise de uma folha - eu poderia descer mais ainda, mas vamos subir. Tenho então o nível de análise de uma árvore, na qual as folhas são componentes, mas tenho a árvore como indivíduo único e, a partir da árvore, tenho a floresta quando subo de nível de análise. Quando deixamos nossa mente percorrer esses diversos níveis de análise, fazemos um ciclo, que é análise-síntese e voltando de novo para a análise. O que é a análise? É pegar algo que é complexo, que tem um todo, e buscar observar cada parte, cada componente. A síntese é obter justamente um desses elementos e aglutiná-los de forma a obter um novo todo, uma nova visão do global. Quando fazemos isso de forma cíclica estamos gerando conhecimento. Basicamente toda atividade de geração de conhecimento envolve circular entre a análise e a síntese.

Temos, então, exemplos de níveis de análise. O nível mais baixo que hoje conhecemos é o nível da Física, das partículas subatômicas, os quarks, os glúons, os múons, os bárions, etc. São as inúmeras partículas subatômicas descobertas nos últimos 30 ou 40 anos, bem como os elétrons, os nêutrons, os prótons, todas aquelas coisas que aprendemos no Ensino Médio. Depois vamos ao nível das ligações de átomos e temos a química, a ligação entre átomos formando moléculas complexas, temos a catálise. Subindo mais um nível, temos o nível onde falamos de pressão, de gases, até um nível onde não consideramos individualmente as moléculas, mas sim aspectos mais globais. Chegamos então a um fenômeno chamado emergência: quando compramos um botijão de gás estamos comprando uma emergência da pressão que o gás faz, mas se analisamos cada um dos elementos veremos lá dentro apenas moléculas. Novamente subimos de análise e

temos como componentes desses elementos a formação de organelas, temos as mitocôndrias, as paredes celulares. Daqui subimos até chegar às próprias células (temos processos da mitose, da meiose, a genética, os cromossomos, a interação entre células para gerar organismos) e aqui já temos a biologia, as inúmeras formas dos organismos, até chegarmos aos ecossistemas. Percebam que não adianta procurarmos teorizar sobre um nível de análise usando o conceito de outro nível de análise.

Assim passamos, por exemplo, para os sistemas sociais e econômicos. Não é muito interessante analisar sistemas econômicos ou sociais baseado, por exemplo, no que ocorre durante a combustão da gasolina. Sabemos que a combustão da gasolina é o que move os motores e os motores fazem o tráfego, que acaba sendo um problema social e econômico. É aquele negócio dos quarks, glúons e múons gerando átomos, gerando moléculas e subindo de nível. Este é um nível que tem as suas regras, a sua própria dimensão. Temos também os sistemas fisiológicos, estou analisando o indivíduo isoladamente: o sistema nervoso central, o controle de músculos, o cérebro, etc. Se eu baixar mais um pouco mais o nível encontro o sistema celular e a seguir a interação entre as células, o crescimento, a diferenciação celular. Depois vêm os sistemas biomoleculares: o DNA, a transcrição de DNA e a geração de proteínas, até chegar novamente a um nível físico-químico. A nossa mente consegue percorrer dentro deste universo, mas não simultaneamente. Conseguimos até mesmo detectar algumas influências, mas é muito raro detectar alguma influência em alguns sentidos, por isso é preciso muito cuidado ao analisar um nível de análise: temos que nos fixar no vocabulário que esse nível de análise nos apresenta.

Agora vou propor um desses níveis para fixarmos a atenção: o nível de análise de um organismo (desenhei de uma forma circular justamente para dizê-lo de forma genérica - não estou falando que é um camundongo em particular ou um ser humano em particular, estou dizendo um organismo qualquer). Desenho também de forma bem abstrata o que eu chamaria de meio-ambiente, aquilo que está circundando esse organismo, e desenho outros organismos, até mesmo de espécies diferentes. Fazendo a análise nesse nível já identifico três partes: o próprio ambiente, um estudo da meteorologia por exemplo, e o agente, que é o nível do que está acontecendo aqui dentro. Tenho também o conjunto ambiente/agente - existem algumas disciplinas que estudam essa interação e a ecologia é uma delas.

Considerando todos os aspectos observamos de perto esse organismo, mas vamos começar com um organismo simples, aquela mosquinha que de vez em quando cai na nossa sopa. Essa mosquinha, se vocês observarem, tem dentro dela algo que poderia ser visto em três níveis. Um nível é o de impulsos, aquilo que é praticamente gerado por interação genética, ou seja, durante milhões e milhões de anos as mosquinhas foram moldadas de forma a ter um tipo de comportamento, um tipo de necessidade. As mosquinhas que não se alimentaram pereceram, por isso todas desenvolveram um impulso para a alimentação. Além desse impulso, identifico também a percepção: as mosquinhas conseguem perceber certas coisas, se deslocam perceptivamente no ambiente. Disto vem uma ação: a mosquinha, por meio do que ela quer e da percepção que recebe do meio-ambiente, determina uma ação e essa ação não costuma ser vista por nós como inteligente. É o caso típico de uma mariposa, ao percebermos o que acontece com a mariposa presa numa janela. Você, de fora, observa a mariposa, percebe que a solução para este problema é muito simples, pois tem uma abertura lá em cima e ela poderia escapar por lá, mas ela não tem essa percepção. Tudo nela é reduzido, ela não tem essa inteligência. No entanto, se pensarmos bem qual seria a atitude racional de uma mariposa que se bate contra o vidro e não vê a saída, pensaríamos que seria ficar parada, porque não tem solução: ela está batendo no vidro e não há solução. No entanto, o impulso, a carga genética desta mariposa, a faz lembrar e executar atitudes aleatórias: começa andar de lá para cá ate que, por acaso, encontra a abertura e consegue escapar. Esta é uma demonstração de que muitas vezes a inteligência de um organismo não está em um aprendizado eficaz com o meio ambiente, mas já está incrustado numa área genética, que providenciou uma espécie de esperteza para a mosquinha ou a mariposa escapar.

Vamos agora para um sistema um pouco mais complexo, que seria o camundongo. Aqui já conseguimos investigar um pouco mais a percepção que existia no caso da mosca, os impulsos que continuam a existir, só que no caso do camundongo tenho uma memória, ou seja, tenho um acumular de experiências que o camundongo pode desenvolver. Posso treinar o camundongo para resolver um problema, como o clássico sair de um labirinto. Isto faz com que, além dos impulsos para ação efetiva, eu tenha a possibilidade de

selecionar várias opções. Quando o camundongo vê uma barreira naquele labirinto, mas foi treinado, ele tem diversas opções e pode escolher uma opção derivada da memória, o que faz com que tenha maior chance de acerto. O camundongo, portanto, já subiu de nível na inteligência.

Vamos ver finalmente, é óbvio, aquele que consideramos o mais inteligente da parada, o que está no topo das coisas: o ser humano. Tinha que ser o mais complicado da história, cheio de ligações entre impulsos, desejos, objetivos, ações, seleção, meio ambiente, percepção, conhecimento e memória. Temos os mesmos impulsos, a mesma percepção (quando eu digo os mesmos, não estou querendo dizer que sejamos similares ao camundongo ou à mosca, mas que temos o mesmo núcleo destinado a resolver impulsos e percepção), o mesmo conhecimento que o camundongo, só que este conhecimento está em estreita relação com a memória. Aqui teríamos alguns ciclos que vêm da reflexão. O camundongo não tem uma noção de reflexão, mas nós temos. Podemos sentar e ficar parados imaginando coisas que afetam a nossa memória e que geram mais conhecimento, como é o caso de pensarmos sobre o que estudamos. Ocorre então um encadeamento de coisas que costuma ser chamado de inteligência emocional: os impulsos promovem desejos, esses desejos são filtrados por meio de objetivos e finalmente tenho as ações que seleciono. Um caso típico, por exemplo, é o meu impulso de fome, o meu desejo de comer um bife à parmegiana, mas há o meu objetivo de me manter "na estica", então corto essa ação de comer o bife à parmegiana, pelo menos todos os dias, certo? Posso fazer isso uma vez por mês. Então temos um circuito onde podemos lidar melhor com esta questão. Aqui não tratei, obviamente, da reflexividade do ser humano.

Nesta tela temos mais ou menos uma idéia do que seria o diagrama anterior, visto de maneira um pouco diferente. Aqui temos em um órgão dos sentidos, a visão, uma primeira área neural que seria, no caso humano, a occipital, instalada quase na nuca. Então nossos olhos têm circuitos que atravessam o cérebro todo e caem numa área neural da nuca. Depois temos algumas figuras que não são esse circuito sensório em especial. Temos essa área neural definida de forma inata - os nossos olhos são definidos de forma inata, a região neural que recebe essa informação é inicialmente definida de forma inata - mas a partir dela temos algo com uma característica chamada perceptual, desenvolvida por experiência. Assim aprendemos a enxergar. Faces, por exemplo, são imagens para as quais já temos uma ligeira pré-disposição genética, mas é refinada por meio da experiência. Tanto que teríamos, nos próximos 20 ou 30 anos, alguma coisa para aprender em relação a faces.

Quando nascemos, ainda bebês, observamos a nossa mãe e o nosso pai, observamos sempre a face humana no que chamamos de up right position, uma posição vertical: dois olhos em cima, um nariz no meio e a boca embaixo. Não trouxe uma imagem para esta apresentação que mostra como essa percepção up right nos engana se virarmos a imagem de cabeça para baixo. Uso a foto de uma moça, na qual eu distorci os olhos propositalmente e inverti a cabeça de ponta-cabeça - você olha para aquela foto e não vê nada de estranho, aparentemente, naquela foto que está de cabeça para baixo. Não tem nada de estranho, mas quando você a vira recebe um choque: é uma cara totalmente deformada, distorcida.

Isso acontece porque temos um aprendizado perceptual, obtido da experiência e por questão inata. Localizamos as faces e assimilamos o seu padrão exato. No dia em que bebês nascerem no espaço, onde não teremos essa questão fixa de posição vertical da face: no espaço você pode estar conversando com uma pessoa e ela estar de cabeça para baixo, não existe para cima e para baixo lá. Então esse bebê vai perceber as faces de uma maneira superior à qual percebemos, não será submetido à ilusão que nós seríamos.

Essa parte perceptual toda desemboca em um negócio que chamo controle de atenção. Um exemplo: estão todos me observando e ouvindo o que estou falando, mas não estão considerando as luzes que estão nos iluminando. Agora desviei a atenção de vocês para observar que existem algumas características dessas luzes, são luminárias distribuídas de certa forma. Temos condição de desviar a nossa atenção e a cada desvio de atenção damos uma oportunidade de obtenção de conhecimento. Essas duas coisas funcionam da seguinte maneira: a parte inicial, perceptual, é toda paralela, ou seja, enquanto estou falando vocês estão percebendo um ligeiro ruído do sistema de ar-condicionado, mas não estão atentos a isto, a não ser no momento em que chamo a atenção, que se torna serial e vocês momentaneamente se desconectam do que estou falando para perceber este ligeiro som do ar-condicionado ao fundo.

Esse circuito que nos permite identificar e reconhecer coisas vem do aprendizado perceptual. A criança, um pouco antes de aprender a ler, identificará o que chamamos de componentes, os features, que são as linhas verticais e as linhas horizontais, em cima, embaixo, e isso tudo formará as letras. Quando a criança aprende a ler irá entender como esses componentes anteriores se ligam para formar estruturas mais complexas. Até que chega o momento em que essa criança começa a juntar pares de letras para formar os bigramas e finalmente formar uma palavra. O aprendizado de palavras pelas crianças possui essa característica de top bottom, de bottom up, de baixo para cima, o que desenvolve áreas neurais especializadas no reconhecimento de cada uma dessas características.

Há um exemplo no qual esse tipo de elemento perceptual tem uma finalidade muito importante na cognição, que é corrigir erros. No início eu disse que o universo tem aspectos ruidosos de um lado e regulares de outro. O que detectamos como regular, porém, nunca é regular realmente, apresenta ligeiras variações devido a esse ruído. As palavras que enxergamos por escrito, por exemplo, têm sempre ligeiras variações. É o caso do conjunto de letras C Q R N: para um americano ou para quem está familiarizado com o inglês, em uma leitura rápida a pessoa enxerga CORN, que significa milho. O que se observa é uma correção desta incompetência, deste desvio, simbolizado pelo Q. Isto é feito pelas áreas que recebem ativações sucessivas (poderia gastar o seminário inteiro falando apenas sobre como isto funciona). O que desenvolvemos por percepção tem como finalidade corrigir erros e esse tipo de correção se dá porque as áreas neurais não se preocupam com detalhes, mas se preocupam com essências. Esse é um ponto que vocês verão repetido no Data Mining.

Outro exemplo de como a nossa percepção corrige "erros": quando há falha de impressão a letra **i** se assemelha à letra **r** - esta é uma impressão que fiz anos atrás, quando ainda usava impressora jato de tinta. (Aliás, eu posso ser considerado um torturador de impressoras jato de tinta: eu as usava até a última fagulha, comprava esses botijões de tinta de um litro para recarga, e as impressoras costumavam reclamar muito - gastei umas 6 ou 7 impressoras com isso). O fato é que, nesta impressão, se percebe que existe uma grande similaridade entre a letra **i** e a letra **r** sem o topo. A mesma coisa ocorre com a letra c e a letra e, que são idênticas, mas contextualizando corrigimos a letra **c** pela letra **e**. Essa é a finalidade fundamental da percepção: desprezar o ruído e manter a essência. A essência, assim, é algo que precisa ser mantida porque tem uma vantagem.

Há alguns anos foi lançada em outdoor a publicidade de um curso de inglês, não sei se aqui no Rio houve, no qual a idéia que se queria transmitir era a de fazer um bom curso de inglês, pois do contrário você estaria falando barbaridades. Uma das barbaridades possíveis era você dizer que já foi "asfaltado". Fizeram paródia com uma campanha da rádio Jovem Pan de anos anteriores - um selinho escrito "já fui assaltado", que se colocava no carro. (Eu nem sei se isso é vantagem, porque isso pode animar os ladrões a pensar: "já que ele já foi assaltado, tem experiência nisso, me deixe continuar"...). A idéia era, a partir de uma leitura rápida, dizer "já fui assaltado", por causa desta campanha da Jovem Pan, mas na verdade o que estava escrito era "já fui asfaltado". Corrigimos o erro em um primeiro momento, mas depois identificamos a diferença que dá um sentido novo.

Isto mostra que temos uma hierarquia de níveis cognitivos, que vem desde o inato até o nível superior. Temos como representantes fundamentais desse nível inato os insetos, que atuam basicamente movidos por questões genéticas. O nível perceptual é onde existe o aprendizado do que acontece no mundo, mas é limitado ao que se percebe, não existe reflexão. Finalmente chegamos ao nível que estamos usando agora, o lingüístico-posicional, no qual temos a expressão, a transferência de informação por meio da linguagem em si.

O mesmo diagrama aparece em uma forma diferente e existe aqui algo que chamo de nível sub-simbólico, que é o das sensações. Sempre uso o mesmo exemplo: queria que levantasse a mão quem conhece a fábrica de chocolates Kopenhagen, a marca Kopenhagen? Muitos levantaram as mãos, mas creio que um ou outro não conheça. Imaginem quem conhece a Kopenhagen tentando descrever para alguém que não a conhece o que significa experimentar um chocolate seu. Você pode passar o resto da vida falando o que quiser, mas não conseguirá transmitir a sensação de experimentar um chocolate Kopenhagen. Isto é o que eu chamo de nível sub-simbólico, no qual só se apreende do meio ambiente a partir da experiência. Sem a experiência não há conhecimento. Subindo de nível: a palavra Kopenhagen é um símbolo, remete a algumas sensações para quem conhece, enquanto quem não conhece veria um símbolo solto, sem

nenhuma conexão com o sensório. Mas em cima da palavra eu coloco "chocolate Kopenhagen é muito bom" - e não estou fazendo comercial. Trata-se de uma proposição, uma relação entre palavras, a maioria delas suportada em sensações, tenho uma comunicação lingüístico-proposicional, assim subimos de nível.

Em uma dessas subidas de nível, ou seja, pensando o que seria este nível lingüístico-proposicional, temos a idéia de que muito do que falamos - não tudo, mas uma boa parte - pode ser categorizada como argumentação. Argumentação é uma espécie de proposição, ou seja: falo alguma coisa sobre alguma característica e dou uma conclusão. Sustento essa conclusão por meio de premissas, o que faz com que eu esteja trabalhando essencialmente nesse nível lingüístico-proposicional.

Aqui começamos a entortar o assunto cognição para chegarmos ao Data Mining. A forma usual de se fazer argumentação é usando um tipo de raciocínio chamado dedução. Gostaria de tratar aqui de um ponto importante para todo o restante que vamos falar, estamos quase no domínio do Data Mining. Existe uma diferença essencial entre dedução e indução. Vou categorizar as duas, acredito que muitos de vocês conheçam o que é indução, mas boa parte pode não saber.

O que é dedução? É uma forma de arranjar uma série de alegações que inferem uma verdade a respeito de certa conclusão que garante a verdade. São conclusões certeiras a partir do instante em que aquelas premissas, que estão sustentando a conclusão, sejam verdadeiras e aceitáveis: o que elas nos fazem inferir é certo, não há como errar. A dedução é algo que está ligado a um aspecto chamado racionalismo, ou seja, por meio de premissas que sei que são verdadeiras tenho como inserir coisas verdadeiras.

A indução é diferente: garante-nos a consistência, é um tipo de raciocínio que nos dá certa segurança da conclusão, mas não é 100% seguro, é só uma indicação de que aquilo deve ser verdade, mas não posso dizer que é verdade. As conclusões já não são certeiras, são apenas prováveis - aqui cabe bem esse aspecto da probabilidade. Uma indução faz com que identifiquemos a conclusão provável. A indução promove uma redução no número de proposições, enquanto a dedução aumenta. Na dedução eu tenho as premissas iniciais e o processo da dedução resulta em uma nova conclusão, aumentando o universo. Na indução tenho algumas coisas que me dão certa verdade posterior, que engloba um pouco as premissas anteriores, reduzindo o número de proposições.

Vou dar alguns exemplos de deduções: "Todas as baleias são mamíferos, todos os mamíferos têm pulmões, portanto todas as baleias têm pulmões". É uma dedução e está correta, porque se eu aplicasse as fórmulas de inferência dedutivas (das quais passarei bem longe se não vocês irão certamente dormir) me garantiria a verdade desta conclusão. Para outro exemplo criei um novo nome para organismos que são mamíferos, bípedes e onívoros: mapeludos. "Todos os homens são mapeludos, só que todos os ursos são, também, mapeludos, portanto todos os homens são ursos". Esse tipo de construção não deveria nos fazer inferir isto, e realmente não faz, é um silogismo inválido. Existe uma série de regras de inferência que nos permite identificar que este daqui não é válido, é falso. Então a dedução é o quê? "Se você tem boas premissas, se você tem um raciocínio com inferência lógica correta, ela lhe garante a verdade".

Essas são as duas premissas da indução: "a maioria dos cães são amigáveis, tenho aqui perto de mim um cão, vou concluir que este cão é amigável". Na via indutiva reparem que as premissas que estou usando já não afirmam de forma certeira. Estou dizendo que a maioria dos cães é amigável, e estou dizendo que vejo um cachorro. Estou inferindo, por indução, que este cachorro que estou vendo aqui é amigável. Isto é razoável, é uma indução chamada de forte, nos dá certa crença na opinião final. Mas existem algumas formas de dizer que isto falha. É o caso aqui: "Esta pessoa conhece o Sergio, esta pessoa também conhece o Sergio (várias pessoas conhecem o Sergio), então todas as pessoas deste edifício conhecem o Sergio". Esta é uma indução meio estranha: é falsa na verdade, se pensar de forma rigorosa. Então a indução é algo que nos leva a pensar em certa conclusão que pode, eventualmente, ser falsa em casos particulares, mas possui um núcleo, uma essência, e aí novamente vem essa palavra "essência", que é útil, porque posso dizer que todas as pessoas deste auditório conhecem o Sergio, tenho evidências para mostrar isso. O tópico "indução" é cheio de elementos, existem debates filosóficos arrepiantes sobre a justificação ou não da indução, os racionalistas versos os empiristas, é uma festa. Isto é assunto para um sarau filosófico de três dias inteiros, mas não vamos chegar a isso. Vamos usar a idéia da indução apenas para deixar no ar esse aspecto da essência. Usarei alguns conceitos que têm ligação com tudo o que vimos desde o início.

Inteligência não é uma palavra associada com freqüência à quantidade de conhecimento. Conhecimento é uma palavra associada à possibilidade de gerar conhecimento: posso pegar um bebê humano e dizer que ele é inteligente; mas um bebê macaco, por exemplo, talvez não seja comparável em termos de inteligência a um bebê humano: o potencial de geração de conhecimento de um bebê humano é maior.

Ser inteligente não é ser incapaz de errar, pelo contrário: quem é inteligente costuma cometer muitos erros, justamente porque existe a exploração deste universo que está à volta para obter regularidade e isto, às vezes, envolve interação. A inteligência não precisa ser fixa durante nossa vida, o modo de resolvermos problemas se altera conforme aprendemos mais coisas - ganhamos heurísticas, ganhamos forma de pensar e resolver problemas com maior rapidez. Desenvolvemos métodos para desenvolver métodos, e isto significa inteligência. Conhecimento não é apenas formal e teórico, não basta apenas estudar a teoria: é preciso ter essas vivências, essas experiências quase sensoriais sobre o assunto. A inteligência é uma habilidade de gerar conhecimento.

Agora vamos começar a quase a tocar no Data Mining. Como poderíamos enxergar uma empresa inteligente? Como poderíamos imaginar esses conceitos todos aplicados a um organismo imenso chamado empresa ou órgão público ou nação, qualquer organismo maior que uma pessoa? Podemos dizer que a habilidade de gerar conhecimento competitivo é essencial para uma empresa, que a inteligência empresarial parte da criatividade das pessoas, que quando são criativas transpiram essa criatividade e o potencial de gerar conhecimento para o restante da empresa. Obviamente a inteligência precisa melhorar os métodos e os mecanismos. Ser inteligente significa poder potencializar ainda mais a capacidade que a pessoa ou a empresa já possuía.

Será que consigo juntar todo esse apanhado de coisas que vimos desde o começo e transformar isso num gancho para Data Mining? Como vamos entender o Data Mining a partir de tudo que falei? Passemos então ao Data Mining.

Em 1974 houve um acidente na Marinha americana, quando a caldeira de um destróier - ainda hoje se usa caldeiras para destróieres - simplesmente explodiu provocando a morte de marinheiros e uma série de prejuízos. Obviamente disso resultou uma investigação detalhada: o que aconteceu para aquela caldeira explodir, qual foi a causa? Os investigadores acabaram percebendo que a caldeira havia sido reparada diversas vezes, na qual tinha sido aplicado certo procedimento de reparo diversas vezes. Mais interessante ainda é que a informação sobre essa caldeira estava em bancos de dados que continham os registros das diversas instalações de reparo, os diversos tipos de reparo realizados. Mas isso permanecia escondido, a informação estava no banco de dados adormecida, esperando que alguém pudesse enxergá-la.

Marvin Denicoff foi o oficial do Office of Naval Research encarregado de estudar o caso. A partir da experiência de Denicoff a idéia de minerar dados ficou forte: se pudéssemos imaginar o momento do início do Data Mining esse seria um bom principio. Foi o momento em que houve um acidente com implicações sérias, reunidas informações sobre a ocorrência do acidente, e os dados anteriores ao acidente já estavam em bancos de dados. Ficamos a observar aquele acidente ocorrer tendo um banco de dados em nossas mãos que poderia ter ajudado a evitá-lo. Chegamos agora no que é Data Mining.

A definição de Data Mining mais citada pela literatura, de Usama Fayyad, é de que "Data Mining é o processo não trivial da identificação de padrões em dados que sejam válidos, novos, potencialmente úteis e ultimamente compreensíveis". Esse "ultimamente compreensíveis" é um detalhezinho que se não conseguirmos tudo bem, o que importa no Data Mining é que temos um processo que não seja trivial, porque se for trivial já está documentado na literatura, algum teórico já pesquisou, é uma regra que já existe por aí. Deve ser então algo que não seja trivial, relativo a identificação de padrões e dados: significa dados e montes de informações que revelam algum tipo de padrão. Devem ser válidos e novos, isto é: que realmente representem algo que seja consistente, novo, que não conhecemos, e que sejam úteis, que possamos usar para alguma coisa. Essa é a essência do Data Mining: são processos mecânicos - como mecânicos entendam, obviamente, informáticos - que fazem acontecer essas coisas, fazem aparecer esses padrões.

Aqui está outro entre os teóricos mais citados, Jiawei Han. É sua a afirmação de que "estamos nos afogando em dados" e no mundo da internet quem é que não acha que estamos? Blogs, Google,

Webpages, temos informações de todos os lados, mas "estamos morrendo de fome de conhecimento": todos nós buscamos conhecimento, queremos novos conhecimentos, mas estamos muito cheios de informação, de dados, em todo lugar. O Data Mining veio para resolver isso.

A imagem desse diagrama é uma das mais batidas e utilizadas em toda a literatura, talvez porque realmente mostre bem o que ocorre. Temos fontes diversas de dados, transações, bancos de interações. Vocês, mais do que eu, sabem os tipos de informações diferentes que são acumuláveis. Essas informações nós jogamos em bases de dados (Data Mining) se forem departamentais, Data Warehouses se forem abrangentes e nesta passagem, de bases de dados comuns para Data Warehouses temos que fazer uma limpeza, porque essas bases em geral muito dinâmicas alteram muito, têm coisas incompletas (o camarada esqueceu de colocar a idade, mas deixou a data de nascimento, coisas do tipo), são redundantes, incoerentes, ruidosas, possuem erros. É o caso de uma base de dados que provocou dor de cabeça algum tempo atrás: o sujeito estava pedindo crédito para um banco e afirmava que já estava há 18 anos no mesmo emprego, só que ele tinha 20 anos de idade. Esse tipo de incongruência não aparece, e continuará não aparecendo, enquanto não houver sistemas com um negócio chamado senso comum. Ninguém em sã consciência daria crédito a um sujeito que diz estar trabalhando há 18 anos no mesmo emprego, cuja idade seja de 20 anos. É difícil acreditar que ele praticamente começou a falar e já estava empregado. Passada a limpeza temos os Data Warehouses, que são a matéria-prima para que possamos efetuar esse processo milagroso chamado Data Mining.

Feito o processo do Data Mining, diversas coisas aparecem, em geral uma espécie de visualização e muitos pacotes fazem de forma gráfica. Isso significa que esse trecho final, próximo das decisões, não está na mão da máquina: está na mão de um analista humano. Data Mining é um processo que vem dos dados lá do fundo, mas que deve terminar na mão de um humano. No futuro isso pode ser alterado, mas hoje é assim. Data Mining é uma confluência entre estatística e ciência da computação. Existe uma interseção que é uma briga dos estatísticos - e os estatísticos, de certa forma, têm um pouco de razão nisto, pois estatística é algo que existe deste de 1800 e as primeiras formulações do tratamento da incerteza de forma matemática é muito antigo. De certa forma o Data Mining usa um pouco destas coisas, mas o Data Mining tem como finalidade principal pegar grandes bancos de dados e daí extrair padrões; os estatísticos têm mais uma idéia de inferência, de pegar dados e localizar certas inferências, além de sumarizar dados.

Este diagrama mostra praticamente como o Data Mining está no meio de uma selva de outras disciplinas muito grandes. Só para ter uma idéia: da ciência da computação temos influências de algoritmos, banco de dados e linguagem de programação (que também tem alguma influência); da área de inteligência artificial, academicamente muito forte mas não muito divulgada, temos desde a representação do conhecimento; da área de Machine Intelligence (faltou um link), que é aprendizado de máquina, temos a influência da recuperação de informação e outras influências; da área da matemática temos fortes influências da estatística, da lógica e da teoria de probabilidade.

O Data Mining usa todas essas coisas. Vamos voltar ao assunto da dedução, que vimos antes. Dedução é o processo no qual tenho duas premissas em que concluo, com certeza, alguma coisa. É com a dedução, no final das contas, que o pessoal de informática trabalha no dia-a-dia: quando se manipula uma base de dados está se fazendo dedução. Por exemplo: tenho uma base de dados, uma tabela, relacionando funcionários e departamentos, e tenho outra tabela relacionando departamentos e gerentes. Nessas duas tabelas aplico um operador Join e obtenho uma terceira tabela, de funcionários contra gerentes. Parece que descobrimos alguma coisa a mais, afinal de contas produzimos uma terceira tabela, de funcionários e gerentes, mas se percebermos bem, essa informação já existia lá em cima, não é nova, só está reorganizada e isto é feito com base de dados - fazemos operações essencialmente dedutivas. Esse é o nosso dia-a-dia de informática, mas não é isso que precisamos fazer no Data Mining. O Data Mining trata de indução.

Vamos rever um bocado desses conceitos que tratei desde o início (obter regularidades de ruídos, etc). Numa base de dados como aquela de funcionários contra departamentos e departamentos contra gerentes, aplico um processo de indução de regras - indução neste caso significa puxar aquelas coisas que capturam alguma espécie de essência desses dados e, dessa essência, obtenho a seguinte regra: "cada funcionário tem um gerente". É uma regra óbvia. Para nós é uma regra óbvia, só que o processo de geração desta regra pode ser aplicado em diversos outros tópicos, diversas outras formas. Uma forma pouco conhecida é

a ILP (Inductive Logic Programming), de obter expressões lógicas, como essa, que dizem exatamente isso. Aqui está dito exatamente o seguinte em linguagem matemática: qualquer que seja o **x** existe um **y** tal que funcionários de x implicam em gerente **yx**. Isto é uma forma, o processo ILP, que gera essa informação por ser alimentado com estas coisas. Este processo pode ser usado para gerar novas inferências.

Já aprendemos que manipulação convencional de banco de dados não vai nos dar Data Mining. O Data Mining precisa de um processo diferente, de um processo indutivo, de um processo onde exista a perda de informação, exista uma espécie de compactação. Este é um assunto que remeto novamente ao começo das idéias. No começo da palestra afirmei que o randômico se transforma no regular por meio de um cérebro inteligente que consegue detectar as essências desse randômico e produzir aquela regularidade. Nesse processo perdemos dados - eu, por exemplo, não posso dizer se alguém trouxer uma maçã por trás de mim: não vi a maçã, não posso descrever todas as características precisas dessa maçã, só posso dizer o que é genérico nela. "Acho que ela terá um cabinho saindo, que será avermelhada, que terá um formato abaloado - posso até desenhar esse formato - e terá uma concavidade em cima e outra embaixo". Bastante coesas essas características, mas se formos comparar com termos precisos, dedutivos, aquela maçã que me apresentarem será bem diferente daquela que descrevi.

No processo indutivo, as pessoas de informática devem se conscientizar de que precisam perder informações, precisam jogar fora. E o que jogamos fora? Detalhes que são irrelevantes, detalhes que não fazem diferença numa expressão mais compacta. Quando falo "maçã" estou falando de uma forma extremamente compacta daquele objeto. Quando descrevo todas as características dessa maçã - avermelhada, arredondada, com cova em cima, etc. - ainda assim estou fazendo uma grande compactação no nível de detalhes que há naquele objeto e perdi, dessa forma, informação. Essa não é uma perda ruim, é uma perda excelente porque nos faz concentrar nas essências.

Voltamos, aqui, a falar dos níveis de análise. Nesse caso tenho em um banco de dados, por exemplo, diversos tipos de análises, o nível do José da Silva (coitado do José da Silva que é sempre usado como nome genérico - ele praticamente é a indução dos nomes brasileiros) que está numa categoria de pessoas, em funcionário, em uma gerência de informática, que é uma divisão operacional, que é de uma filial, etc.

Dentro desta linha de níveis de análise tenho como aplicar a mineração em qualquer um dos níveis. Se aplicar esta mineração no nível das pessoas obterei expressões do tipo padrão: "os Josés têm dois filhos" - é isto que o Data Mining vai me falar. Se aplicar no banco de dados de pessoas terei: "este José tem dois filhos, este outro têm três, este outro dois, outro tem um filho, outro não tem filhos", ele dirá que "os Josés têm dois filhos", uma conclusão indutiva, que não é certeira, mas remete a uma espécie de essência - a essência de que, na média, os casais têm dois filhos, mas ele detectou isso no José. Ele achou uma regra que é muito fraca, sujeita a ruídos.

Na época em que fiz o colegial, há muito tempo, os professores dividiam as salas por ordem alfabética: era um terror, porque eram salas grandes, mas também com um grande número de alunos, e na minha sala só havia gente com S, logo tinha Sergio, Sandra, um monte de Sergios. Tínhamos que inventar apelidos para todo esse pessoal, havia o Jesus, o Pirulito, era a única forma de identificar, pois todos tinham o mesmo nome. Um Data Mining aplicado sobre esta sala do meu colégio diria assim: "pessoa tem nome Sergio", essa é a conclusão que ele dará, é a conclusão que sai dessa base de dados. Se for aplicar a mineração no nível de cima terei a seguinte idéia: "todas as filiais têm uma matriz", é isso que sairá do Data Mining, uma obviedade, uma coisa que não tem muito valor, algo que faz parte da descrição do que é matriz e filial, faz parte da definição. Temos que circular dentro de um nível intermediário e não basta ficar em certo nível: temos que ir para cima, para baixo, mexer um pouco nesse nível de análise da informação.

Os três principais focos da mineração - tentaremos agora entrar um pouco mais no detalhe - são esses três elementos: identificar classes; achar padrões em seqüências (seriam seqüências de classes) e descobrir regras associativas (associando uma dessas categorias à outra). São regras, restrições, que nos permitem afinar a coisa e são padrões de forma estatística. Vamos ver um pouquinho dessas coisas. Diria que esses três processos são fundamentais para lidarmos com o Data Mining. Todos têm diversos algoritmos prontos, há inclusive pacotes de domínio público como o WEKA, se bem que agora o pessoal parece estar querendo fechar uma parte e começar a comercializar - então corram, baixem logo da internet antes que fechem para comercializar. Alguns pacotes anteriores já passaram por isso, viram que a idéia era boa demais e

fecharam, mas ainda existem pacotes de domínio público e a maioria deles possui esses processos, esses algoritmos, que podemos aplicar nos dados. Não importa tanto como funcionam esses algoritmos, veremos muito brevemente, o que importa é saber o que eles estão tentando fazer.

Aqui tenho uma tabela e mostrarei como pode gerar uma árvore de decisão. Tenho casos de empréstimos e tenho o tipo de risco que esses empréstimos representaram, ou seja, tenho uma tabela histórica dizendo o que aconteceu com esse empréstimo. No caso moderado, por exemplo, o camarada atrasou a prestação, ficou quase para escorregar no Serasa, mas pagou. O risco alto é o inadimplente, engrossando a lista do Serasa, que aliás não é mais esse nome, já foi vendido para uma empresa internacional. E há ainda os outros tipos de riscos. Tenho o histórico deste cliente, o que ele representou no passado, não apenas na última operação, mas historicamente. Tenho o tipo de débito que ele fez, a garantia que ofereceu e o salário anual dessa pessoa. Essa é a tabela que tenho, a informação bruta, aquela que faríamos basicamente com dedução em base de dados. Esta é a árvore de decisão construída pelo algoritmo ID3, algoritmo já muito antigo, quase histórico. Esse algoritmo gera coisas, verifica que o ponto mais fundamental para começar a investigação é um ponto onde se analisa o histórico. Como o ID3 acha isso? Faz uma varredura em todos os atributos da tabela e verifica aquele que tem maior ganho informacional, qual é aquele capaz de dar maior possibilidade de informação, maior essência de informação.

Disto resultam as três possibilidades possíveis: tenho um histórico desconhecido ou um histórico ruim ou, ainda, um histórico bom. Se o histórico é bom e o débito é baixo, baixo será o risco. Se o débito é alto qual a garantia? Ou seja, verifiquem que, quando o débito é baixo, não preciso perguntar mais nada, mas se for alto tenho que ter uma idéia do que é a garantia. Tem garantia adequada, o risco é baixo. Não tem garantia? Como funciona o salário? Percebam que a árvore de decisão, então, nos apresenta o caminho das pedras. Isto é indução pura. Porque que é indução? Porque é possível que peguemos um sujeito com excelente histórico, que tenha um débito alto, mas não ofereça garantia e tenha um salário grande e acaba dando um risco alto. A indução é algo que diz qual a essência da coisa, mas não impedirá de achar uma exceção. A exceção acontece. Os processos do Data Mining, todos eles, não dão certeza: dão a inferência da essência, a inferência do que é razoável acontecer. Essa é uma das técnicas que podem ser usadas para mineração: colocar o banco de dados de uma forma organizada e jogar num C4.5, que é um algoritmo superior ao ID3, e obter daí informações. Mas existem outras formas, regras, que obtêm dos dados alguns tipos de associações. Veremos algumas dessas regras, não todas.

O primeiro passo seria distinguir o que é uma regra. Neste caso vamos dizer que regra é algo que lembra um pouco a dedução, ou seja: tenho várias condições antecedentes que me darão uma conseqüente, é isto que vamos querer obter. Embora tenha esse gosto dedutivo, a regra aqui é indutiva, não dá certeza. Exemplo de uma regra assim: se o empréstimo for maior que R$ 5.000,00 e o salário menor que R$ 1.500,00, recusa-se o crédito. É até meio chato dar esta regra de forma tão forte e aí entra o fator humano.

Se essas informações aparecem em um Data Mining de bancos, por exemplo, onde você trabalha como gerente, conversando com o cliente recebe esta informação da mineração do banco de dados. Você vai recusar o crédito, mas conversando com a pessoa verifica que o camarada parece ser de boa família, bom pagador, já esteve várias vezes com ele, entram fatores acessórios que não são, digamos, computáveis. A idéia é que o Data Mining forneça uma informação para que o gerente não decida sem um parâmetro.

Outra idéia fundamental em toda geração de regras trata da indução orientada por atributos, feita por Jiawai Han, que significa a tentativa de analistas humanos em obter alguma coluna da tabela e buscar um nome que caracterize algumas destas coisas (como o caso do nome José para caracterizar um grupo de pessoas). Significa restringir conceitos em torno de uma generalização. Essas hierarquias normalmente são feitas por alguém que conhece o domínio, que vai observar a tabela e verificar que este mundo de variação de dados pode ser sumarizado em três categorias, por exemplo.

Entre alguns exemplos, podemos remover atributos que não têm sentido: o nome da pessoa, em geral, não faz sentido em um banco de dados de crédito, com exceção se o nome for Fernandinho Beira-Mar ou coisa do gênero. Mas podemos generalizar atributos que sejam categorizados pelo mesmo nome. Quando usamos uma base de dados de estado, cidade, população e orçamento da saúde será que poderíamos descobrir alguma coisa desta base de dados, deste mundo de informação? Podemos fazer uma indução

orientada a atributos transformando o país, que tem mais de vinte categorias, em Norte e Nordeste, Sudeste, Sul e Centro-Oeste. Isto é óbvio, fazemos normalmente. Mas este processo está transformando uma informação, que antes era imensa, em poucas categorias; com isso estamos fazendo uma indução.

Em relação à população, ao invés de colocar os números da população, vamos fazê-los oscilar em três categorias - população pequena, média ou grande. Só fazendo isto ganhamos uma dimensão nova, porque teremos uma tabela bem menor, onde perceberemos de cara algum tipo de relação. Assim chegamos aos gastos da arrecadação, por exemplo: se o orçamento de saúde representa entre 0 e 5%, ou entre 5 e 15%, entre 15 e 20% da arrecadação. Percebam que fazendo apenas essas modificações transformamos a base de dados em outra base menor, na qual não precisaríamos nem de computador para dizer quais são os padrões, só de bater o olho poderíamos dizer: "as populações grandes que possuem essa faixa de arrecadação da Zona Norte apresentam esse tipo de problema". Nós mesmos podemos fazer isso, o Data Mining pode ser feito sem computador quando os dados forem poucos, mas obviamente não é isso que acontece na maioria das vezes.

Outro exemplo de indução orientada a atributos: uma pessoa lista os seus hobbies, entre os quais tênis, futebol, piano, Nintendo (acho que nem existe mais), ópera e playstation (meu filho entende disso), mas posso transformar esses atributos aqui em apenas três: esportes, música e videogame. Com isto já consigo reduzir o escopo das variações e aí posso tentar obter algo interessante. Isso favorece o aparecimento de atributos.

Agora já podemos usar a indução orientada a atributos para fazer algumas operações. Uma delas é a de "regras caracterizadoras" - obter regras que consigam caracterizar aqueles conceitos satisfeitos pela maioria. Se fizer o orçamento de saúde contraposto à população e outros dados direi que, na Zona Norte, aquela faixa de população tem aquele tipo de orçamento. Podemos usar essa indução orientada a tributos para nos sugerir quais são os campos que possuem essa qualidade de poder resumir, caracterizar, essas coisas. Por exemplo: tenho uma lista de doenças no banco de dados da saúde com diversas outras características e posso verificar quais são os sintomas prévios reportados pelo doente que implicam naquele tipo de doença. Isso tem um valor imenso, porque pode-se receber um doente e verificar os sintomas previamente e o Data Mining direcionar para o especialista: "esse camarada deve ver um infectologista imediatamente". Os médicos sabem disso, mas o Data Mining também pode ajudar no atendimento prévio.

Outro exemplo pode ser relacionado a características típicas dos estudantes de MBA que decidiram em fazer o curso em seguida à graduação: pode-se usar um banco de dados de alunos e verificar quais são as suas notas, a presença em seminários opcionais e uma série de outros dados. Assim é possível descobrir o perfil daqueles que usualmente optam por um MBA.

Outro caso típico de análise de crédito faz uso da regra caracterizadora. Isso também poderia ser passado pela árvore de decisão, mas nesse caso as características são obtidas de cada um destes elementos por meio de um algoritmo especial que revela diretamente as regras, já destinando a sua finalidade. Se as contas em atraso forem maiores que dois e o número de pagamentos em atraso for maior do que um, se não for um cliente rentável recusa-se o crédito. Ou seja: há regras encadeadas, um nível de regras e um segundo nível de regras obtido pelo Data Mining. Em outro caso, se as contas em atraso for zero e o salário maior que R$ 3.000,00 ou mais de três anos como cliente aceita-se a solicitação de crédito. Isto brota dos dados e ajuda a compor um novo tipo de conhecimento.

As regras discriminatórias são aquelas que conseguem separar, delimitar, classes que não contenham tantas informações apreendidas em relação a outras: são as classes contrastantes. Um exemplo típico é conseguir distinguir uma doença por meio de uma regra discriminatória que sumarize sintomas da doença que não sejam comuns em outra doença. Febre, por exemplo, é uma coisa que acontece quando o sujeito tem alguma infecção, ou seja: febre não é um bom elemento nesse aspecto. Mas um tipo de irritação avermelhada da pele, em certo local, pode indicar uma classe mais restrita. Logo, as regras discriminatórias têm esse poder de dizer quais são os objetos que têm a capacidade de direcionar rapidamente para um tipo de conclusão especial. As regras discriminatórias possuem categorias cuja principal preocupação é evitar a sobreposição com classes contrastantes, o objetivo é realmente separar.

As regras associativas são as mais conhecidas do Data Mining e as mais úteis também. É o caso típico do

que é feito por exemplo, no supermercado Pão de Açúcar. O estabelecimento tem um banco de dados que dá inclusive um cartão fidelidade - não sei se aqui no Rio existe, mas em São Paulo sempre que se passa num caixa do supermercado a moça pergunta se tenho ou não o cartão. Qual é a vantagem em você ter essa identificação via cartão? Para a empresa é saber que o cliente fez tais compras, listar as coisas que ele comprou e identificá-lo não de forma individual, mas de forma a permitir processos de mineração. Tenho assim aqueles exemplos típicos: se o sujeito comprou pão, leite, ele irá comprar manteiga também. Esta é uma regra óbvia, que pode brotar dos dados. Mas as regras associativas podem ter um processo que despreza a quantidade: o importante é a extensão feita por Jiawai Han (como sugestão, para quem gosta de pesquisar no Google, procure por Jiawai Han e verá as inúmeras publicações e materiais sobre isso). Pode-se então pesquisar sobre faixas de quantidade e aí começamos a perceber como o Data Mining começa a nos dar um retorno interessante.

Falava do banco de dados óbvio: o camarada comprou leite, vai comprar pão também. Mas se vocês se lembrarem do que eu disse sobre os níveis de análise e descermos um pouco o nível de análise não vou me preocupar com o leite em geral, mas com o leite desnatado, e este tipo de leite me revela que a compra conjunta usual foi pão integral (em geral são os "naturebas", entre os quais me incluo). Os "naturebas", que compram leite desnatado vão pegar um pão integral para acompanhar a filosofia, a idéia. Só que esta regra aqui já não é tão óbvia: leite desnatado Parmalat implicou na compra de pão integral Pullman. Esta regra não é obvia, é uma regra que não saberia deduzir e não importa se sei ou não deduzir, o fato é que veio dos dados, brotou da experiência acumulada do banco de dados. Surge então a idéia do supermercado: colocar ao lado do leite desnatado a prateleira de pacotes de pão integral e com isso fomenta-se a venda daquele produto.

É dessa análise que surge o termo mais conhecido, mais famoso, do Data Mining: o chamado market basket analysis. Quando o pessoal de marketing das empresas vê isso, os olhos deles em geral triplicam de tamanho. A idéia é mexermos com essas informações para sabermos o que fazer em termos de marketing. E temos algumas idéias: achar todas as regras que tenham Diet Coke como conseqüentes, ou seja, todas as regras do meu banco de dados mineradas que acabam saindo em Diet Coke. Isto vai ajudar a planejar como vender melhor este produto. Sei quem comprou Diet Coke, pão Pullman ou quem comprou qualquer outra coisa. Faço esses produtos ficarem próximos ou facilito a sua aquisição. Pronto: o sujeito do marketing ganhou o dia.

Mas descemos para o item B e verificamos todas as regras que têm iogurte como antecedente. Acontece muito que o supermercado esteja brigando com o fornecedor por causa de preço: o fornecedor não quer reduzir o preço, então o supermercado pensa no que acontecerá se continuar batendo o pé e não comprar mais daquele fornecedor, qual será a conseqüência orçamentária. Essa é uma decisão terrível para o marketing. Se resolver brigar com o fornecedor pode ser que perca a venda daquele produto, mas o Data Mining pode informar que esse produto é antecedente de um monte de outras coisas que valem a pena continuar. Então o supermercado pode dizer: "Tudo bem, vamos aceitar o preço do fornecedor". Nesse caso, o Data Mining é uma ferramenta fundamental para varejistas.

Outra idéia seria achar todas as regras com salsichas no antecedente e mostarda no conseqüente: vai ajudar a encontrar os itens que preciso colocar junto com a salsicha para fomentar as vendas de mostarda e esse também é um exemplo criado manualmente, mas com implicação prática muito intensa. Este é o paraíso dos "marqueteiros". Espero que possam supor como se daria no mundo de vocês: de que forma poderiam pensar nisso, que tipo de auditoria poderia ser feita, em que tipo de departamento, como seria possível localizar o departamento que precisa de auditoria - e assim chegamos no mundo de vocês.

As regras classificadoras são usadas praticamente por todos e tratam de dispor de acordo com certas características. Posso classificar carros de acordo com o consumo de combustível. Posso classificar clientes em rentáveis, não rentáveis, muito rentáveis. Assim faço marketing direcionado. Se você é bom cliente de uma empresa costuma receber aquela propaganda com seu nome impresso, com materiais bons. Como gosto muito de livros, compro muito, recebo de vez em quando ofertas de empresas do ramo em material de publicidade são extremamente caros e envolventes: têm selos para colar em diversos lugares, são produtos caros, mas conhecendo esse truque de marketing não vou na conversa do "marqueteiro", jogo fora logo o que é marketing e vou para o livro, o que me interessa, e acabo comprando e, com isso, subindo de nível no universo das editoras. Existe, então, uma categoria de clientes que se descobre por

meio dessas regras classificadoras, que servem inclusive para determinar pontos de lojas: a partir da informação de diversas lojas verifica-se quais são as bem sucedidas e se determina, pela localização, onde elas deveriam ser implantadas prioritariamente. A classificação pode ser feita por meio da árvore de decisões ou do CART, um algoritmo muito conhecido.

Vamos falar agora de um tipo de estratégia da mineração na qual usamos processo um pouco diferente. Até agora temos usado um processo chamado supervisionado: tenho a interação de alguém, que chamo de professor, em relação ao sistema. Vimos que decido como faço a indução orientada a tributos e manipulo, como ser humano, o que o sistema vai me dar de informação. Isto porque eu conheço um pouco este domínio, ou seja, como especialista de determinada área consigo identificar o que é importante e o que não é. Consigo saber que o nome do cliente não é importante, mas o seu salário anual sim. Esses são os chamados sub-métodos supervisionados e os apresentados aqui fazem parte desta categoria. Mas existe um outro tipo de método, o não supervisionado. Aqui o professor, coitado, não sabe o que quer: vê um monte de dados que não possui sentido prático - como valores de transação individuais - e não tem a menor idéia do que pode fazer com valores de transação. Será que existe algum padrão, qualquer tipo de média ou de tendência? Ele não faz a menor idéia, por isso deve usar um processo chamado não supervisionado: aquele no qual o professor não sabe o que procurar, mas sabe que quer alguma coisa boa.

Um dos processos, chamado clustering, é o de detectar aglomerações. Esse processo tem muito em comum com o que vimos desde o início: o nosso cérebro faz clustering o tempo inteiro; fazemos clustering de características que, às vezes, nem temos consciência. Quando fazemos uma viagem para o Norte, para o Sul ou para qualquer lugar do país diferente da nossa origem, existem características fisionômicas nas pessoas - de altura, peso, coloração de pele - que não temos consciência de que são uma espécie de gosto da região em que estamos. Isso é feito através de clustering, porque vamos a estes locais, observamos exemplares de forma inconsciente e fazemos a categorização. Se vamos para Minas Gerais temos uma idéia de como são as pessoas lá, se vamos para o Espírito Santo a mesma coisa. Não há essa noção no inconsciente, o clustering é uma formalização disso e vou mostrar mais ou menos como funciona.

Temos a representação do que poderia ser um espaço de dados, como salários ou valores de transição ou, ainda, número de prestações pagas em casas lotéricas, uma informação qualquer. Eles compõem uma espécie de universo representado por duas dimensões apenas, mas que em geral são multidimensionais - esta é uma palavra que os matemáticos adoram e os leigos detestam. Há uma história que acho engraçada, não sei se todos irão achar, sobre uma grande reunião, nos Estados Unidos, com fazendeiros e profissionais, estes querendo apresentar alguma coisa muito interessante e sofisticada para os fazendeiros e eles, aquele pessoal com cigarrinho de palha na boca, pessoas simples de verdade, ouvindo. Convidaram um físico para dar uma palestra para esses fazendeiros e a primeira coisa que o físico fala quando pega o microfone é: "Senhores fazendeiros, imaginem aqui uma vaca esférica". Pimba, acabou o discurso! Na verdade essa é a piada. Como os fazendeiros, que são práticos, nada teóricos, vão imaginar uma vaca esférica? Isto mostra que o físico pode fazer um raciocínio abstrato, do qual depreenderá que a produção de leite, devido à esfericidade, não tem nada a ver com os fazendeiros... É mais ou menos o que acontece se mencionarmos a questão das múltiplas dimensões. Normalmente vamos falar apenas de duas dimensões justamente para visualizarmos este processo.

Assim, a idéia do clustering é a seguinte: não sabemos o que queremos, mas digamos que queiramos achar duas coisas relevantes, dois clusters, dois agrupamentos. Então chuto dois agrupamentos, fecho os olhos e falo: "Minha mãe mandou bater nesse daqui e neste outro", aleatoriamente. Achei dois centros, chutei dois centros. Agora o que faço com esses dois centros? A partir deles computo uma dimensão, uma distância de cada um desses pontos a partir deste centro e a distância é calculada a partir de algo que conhecemos como, por exemplo, a eficácia de pagamento de um funcionário. Quanto mais eficaz é o pagamento de um funcionário mais esta distância é menor. Fazemos, então, uma regra de estipulação de distância. Com isto consigo classificar todos os elementos desta região entre próximos deste ou próximos daquele e até elementos que não estão próximos de nenhum deles. Com isso obtenho uma divisão onde existe um cluster que circunda o número 1 e outro que circunda o número 2. Não é ainda o meu padrão. Não achei o clustering em si, mas fiz uma primeira tentativa. A segunda tentativa é tentar encontrar entre esses elementos que selecionei como sendo do 1 e recalcular um centro deles: o centro seria a posição mais ou menos eqüidistante dos restantes. No número 2 também faço isso e obtenho uma reposição dos centros. Simplesmente recalculei a posição desses centros que estão mais próximos de todo mundo. Agora

faço o mesmo processo, computo a distância entre cada um deles e jogo fora aquilo que não está dentro daquela distância: com isso obtenho um novo padrão. Em poucos ciclos, aplicando este processo, tenho rapidamente uma concentração em torno dos clusters.

Quando falamos clusters podemos lembrar destas notícias astronômicas - "Descoberto um cluster de estrelas..." - e quando olhamos a fotografia vemos realmente um monte de estrelas compactadas em torno, mais ou menos, de um centro. É isto que o nosso processo de clustering, com esse estratagema chamado K-Means, consegue executar. Partindo de dados que não tinham nada em comum consegui detectar dois ou mais pontos que têm algum interesse, então posso usar nossas técnicas de dar nomes para coisas: digo que o primeiro é o tipo A e o outro é o tipo B, e a partir daí passo por processos de mineração tradicional que encontram relações entre categorias, que fazem com que isto se distinga do outro, e brotam mais atributos. Em suma, deste processo eu consigo gerar a engrenagem do Data Mining e fazer a informação aparecer.

Há um outro tipo de regra que detecta associações que existem ao longo do tempo, e não ao longo de uma fotografia. Esse tipo de regra de evolução temporal é um exemplo extremamente desatualizado: há seis meses uma pessoa comprou um computador com CD-ROM, mas hoje ela compraria um DVD-ROM e a maioria dos PCs já trazem o DVD-ROM. Ou seja: alguém comprou um computador com certa configuração e detectamos este padrão. Algum tempo depois a pessoa compra outro tipo de acessório, um pen-drive, alguma coisa especial. Então exploramos este padrão: todos que compraram computadores, entre 6 e 5 meses depois, receberam uma carta dizendo: "Oferta especial: compre seu pen-drive por apenas R$ 98,00". O mesmo pode acontecer com impressoras ou qualquer tipo de produto usado com computadores que tenha sido identificado como padrões novos. Videocassete também é algo que não existe mais, hoje seria DVD.

O que interessa na evolução temporal é que as coisas variam em função do tempo, não são estáticas. A maioria é estática, como a obtenção de crédito, mas podemos fazer uma variação temporal. Daqui surge a idéia, por exemplo do departamento em que um funcionário está: obviamente o funcionário pode sair de um departamento e ir para outro, mas é uma evolução temporal, meio inútil; o que importa aqui são coisas que tenham certa exposição ao longo do tempo. Muito da evolução temporal é feito com as técnicas anteriores - a caracterização e a classificação. O Data Mining não trata simplesmente de usar uma técnica ou outra, mas de usar diversas técnicas conforme entendemos que o processo tenha um jeito de prosseguir. A evolução temporal envolve essas outras técnicas. As características das empresas cujas ações em bolsas de valores tiveram crescimento de 20% é uma regra temporal: observamos qual é a sua evolução e temos como predizer isso.

Chegou a hora de observarmos um exemplo prático, na verdade um exemplo trivial. Tenho uma locadora de vídeo, com um banco de dados de fitas numeradas e classificadas por gênero. Esse banco de dados está dividido em tipos de gêneros, previamente feito e fácil de classificar - se o filme é de ficção, de guerra, western, erótico, de aventura, drama, todas as categorias já estão pré-determinadas. Já existe, portanto, uma forma de pré-classificar a informação, que vem do conhecimento do domínio. O que posso obter desse material por mera dedução, se faço o processo das coisas que conhecemos de banco de dados? Posso relacionar, por exemplo, o código da fita com o título e com o gênero e obter as locações por gênero. Tenho assim uma grande tabela que informa os gêneros mais alugados. Isto é um pouco óbvio, mas é uma espécie de manipulação, só que dedutiva. Obtenho uma informação que é a mesma que estava no banco de dados.

Aqui temos uma mineração feita com indução. Pego as tabelas anteriores de locação versus gênero, faço uma tabela de locações por gênero, só que agora de acordo com o cliente, jogando fora o restante do que não está no topo da classificação. Permito, com isso, achar os gêneros preferidos e assim tenho como gerenciar, acessar ou descobrir, os gêneros mais rentáveis - isto é uma indução, um processo no qual jogo dados fora sucessivamente e chego à informação de que o gênero guerra, por exemplo, é muito rentável a partir de uma informação dedutiva. Também posso ter outro tipo de informação, sobre o gênero preferido de cada cliente. Um cliente alugou um filme de guerra, um histórico, um erótico, vários tipos de filmes, só que ele tem preferência por um tipo de estilo e essa preferência significa a lista de preferências de locações dele, mas uso apenas as principais, que têm significação. Isto não quer dizer que a próxima fita que ele vá alugar seja um desses gêneros, mas que se eu fizer um marketing direcionado sobre uma fita

nova deste gênero otimizo o custo de mala direta e atinjo um cliente com interesse.

Agora vamos usar um passo adicional nesta base de dados: pegar os gêneros, que são muito extensos, e fazer uma indução orientada a atributos, ou seja, transformar esses gêneros em cinco elementos. Inventei categorias: filmes leves, filmes fortes, filmes de suspense, filmes de ação e outros filmes (que não se encaixam em nenhumas das categorias anteriores). Como analista de domínio, gerei artificialmente um atributo novo e esse atributo está dizendo para classificar nessas informações. Posso fazer também uma indução orientada a atributos mais forte ainda: transformo as categorias em três tipos: filmes infantis, filmes teens e filmes adultos. Apenas fazendo isso consigo obter informações relevantes, consigo obter dados que dizem as preferências principais dos locadores, consigo otimizar o meu recurso. Percebam que o exemplo é muito simples, trivial até, talvez o locador faça isso na cabeça. Sou sócio de uma locadora no meu bairro e recebo e-mail com lançamentos, só que o dono ainda não descobriu como me jogar no tipo de filme que gosto - ficção científica, alguns de ação e para acomodar a esposa pego alguns romances. Eu me enquadro em três ou quatro categorias: se ele soubesse identificar isso iria me mandar um e-mail ou propaganda já direcionada, dizendo o que eu efetivamente teria interesse imediato. Isto significa rentabilidade maior fazendo poucas substituições já teria uma eficácia. Este é praticamente o final da nossa palestra, para mostrar que as regras de evolução temporal podem dar informações de comportamento.

Ainda sobre esse exemplo da base de dados da locadora se poderia descobrir, por exemplo, que ao longo do tempo houve essa seqüência: ação - romance - ficção - infantil, de certo grupo de clientes. Uma outra seqüência de cliente começou com o romance, o romance esquentou o clima ele foi para um erótico, mas baixou a penitência e ele foi para um bíblico; finalmente, para ficar algo leve, acabou no documentário. Este é o ponto do Data Mining que eu queria ressaltar, claro que consigo criar uma historinha, só que raramente conseguimos fazer isso no Data Mining. Não importa muito explicar essa seqüência, muitas delas são completamente sem pé nem cabeça, não vale a pena gastar tempo pensando nisso por enquanto. Talvez um teórico possa olhar para essas coisas e explicar que "isto ocorre por causa disso", mas existe um processo científico que é algo mais complicado.

O que interessa para quem manipula o Data Mining é obter as seqüências, porque elas dão eficácia comprovada no seguinte sentido: são estatisticamente relevantes, significa que os dados revelaram que existe um padrão e me permite quantificar este padrão. A idéia do Data Mining é obter pedrinhas preciosas que surgem dos bancos de dados e que fazem com que melhoremos a idéia daquilo que conseguimos. Em épocas próximas do Natal posso descobrir que existe uma concentração maior de filmes românticos, de caráter bíblico. Outras épocas do ano, o Pan-americano por exemplo, pode influenciar o consumo de filmes esportivos, o que pode levar quem detém essa informação a usá-la de forma eficaz melhorando o negócio. Sempre falo do negócio, mas temos que olhar pelo lado da descoberta de informações em banco de dados com aquilo que podem manipular: o que importa é ter o banco de dados, pois sem o banco de dados não há como fazer coisa alguma, mas com ele e com esses processos temos certas pepitas a obter. Cheguei ao final e gostaria de agradecer a atenção de vocês, qualquer coisa estou à disposição.

Expediente

Prefeitura da Cidade do Rio de Janeiro
Prefeito: Cesar Maia

Controladoria Geral do Município
Controlador Geral: Lino Martins da Silva
Subcontrolador de Gestão: Vinícius Viana

Assessoria de Comunicação
Assessora: Sonia Virgínia Moreira

Cadernos da Controladoria
Organização de Eventos: Graça Louzada
Administração de Eventos: Vanda Pastro
Edição de Texto: Sonia Virgínia Moreira
Editoração, Capa e Fotos: Gabriel Campano

Transcrição de Áudio: Carolina Orofino
Versão Online: Renato Gomes