

COMUNICAÇÃO COORDENADA

JORNALISMO INTELIGENTE (JI) NA ERA DO DATA MINING

Em 13 de dezembro de 1973, o então iniciante na profissão, o conhecido jornalista esportivo Juca Kfoury, que trabalhava no Departamento de Documentação da Editora Abril (Dedoc), escreveu uma matéria para o Jornal da Abril explicando como um repórter poderia complementar uma matéria utilizando o arquivo analógico.

“Supondo que um repórter deseje complementar uma matéria com uma foto e dados sobre um desconhecido Vicente Saldes, que, subitamente, se tornou assunto, ele recorre ao Dedoc. Há duas atendentes, que põem à sua disposição as pastas (recortes e fotos em cores e b&p) de personalidades. Mas, como Vicente é um desconhecido, só há um dado sobre ele, colhido pelo repórter: “ Uma vez, saiu uma reportagem sobre famílias numerosas e ele estava lá, numa ponta, porque tem 14 filhos”. Então, Vicente é pesquisado na pasta Famílias, isto é, por assunto. Se ele não for encontrado ainda assim, dá-se uma busca no arquivo remissivo, que indica onde Vicente poderia estar: ao lado de Pelé, ou na inauguração da agência do Banco do Brasil em Nova Iorque – quem sabe. Se for o caso, as pastas mencionadas no arquivo remissivo são consultadas. Se nem assim for encontrado – bem, o problema é do repórter: afinal, Vicente quase não existe. Pois até o índice de Veja foi pesquisado.”

Walter Teixeira Lima Junior¹

¹ Doutor em Jornalismo Digital pela Escola de Comunicações e Artes (USP). Pós-graduado em Consultoria em Internet (Ciências Exatas) e certificado em Internet/Intranet System Programmer Analyst (ISPA) e Adobe Digital Video Convergence.

Resumo

O paper apresenta um ensaio sobre o uso do processo de Data Mining na mineração de dados no processo jornalístico. A técnica já é utilizada em outros campos da atividade humana e bem formatada pode ajudar o jornalismo na melhora da qualidade da informação pesquisada em bancos de dados e na obtenção de relações ‘invisíveis’ de temas e contextos.

Palavras-chaves

Jornalismo, Banco de dados, Data mining e Busca

Introdução

Desde as descobertas de figuras rupestres desenhadas em cavernas, vem se constatando que o armazenamento de informações é uma condicionante do ser humano. Há 5 mil anos, os distantes sumérios, na região onde hoje é o Iraque, esculpiram em plaquetas de argila os primeiros sinais, nascendo a escrita cuneiforme. Passando pelas escritas em paredes de pedras encravadas nas pirâmides zapotecas ou egípcias e pelas grandes bibliotecas, como a da lendária Alexandria, o homem parece ter necessidade de querer guardar (arquivar) informações. Isto também é visto nas culturas que somente dominam a linguagem oral: o compromisso de perpetuar o antigo por intermédio da oralidade.

Na atualidade, com o advento da tecnologia de armazenamento digital, quase tudo que produzimos de informação passou a ser colocado diretamente no mundo de bits e bytes e o que existe no meio físico, como em livros, revistas e jornais, para citar alguns exemplos, estão sendo transmutados para os discos rígidos ou para as memórias digitais.

A agilidade e eficiência de um banco de informação de um jornal são fundamentais para assegurar a atualidade e credibilidade do próprio jornal. Usando tecnologia avançada, os sistemas de processamento, armazenamento, controle, recuperação e disseminação da informação permitem gerenciar as bases de dados e material informacional em texto e imagem.²

Vejamos o audacioso projeto do New York Times, finalizado em 2002. A ProQuest³, empresa contratada pelo jornal americano, digitalizou todas as edições do Times de capa a capa. Todas as matérias, editoriais, fotografias, cartuns e publicidade estão incluídas no processo. O sistema usa uma poderosa ferramenta de busca em arquivos e os leitores podem ver o material como eles originalmente foram impressos.

² ROZADOS, Helen Beatriz Frota. **O jornal e seu Banco de Dados: uma simbiose obrigatória**. In: DoIS (Documents in Information Science), Issue 1, Volume 26, Ano 1997. Acessado dia 20 de setembro de 2004 em <<http://dois.mimas.ac.uk/DoIS/data/Articles/juljqbfchy:1997:v:26:i:1:p:2805.html>> Acessado em 20 de setembro de 2004

³ MAYFIELD, Kendra. **Read All About it**. In: Revista Wired, 29 Jul 2002. <www.wired.com/news/business/0,1367,54030,00.html> Acessado em 20 de setembro de 2004

Os usuários do sistema podem pesquisar eventos históricos de 1851 a 1999. O Times foi o primeiro jornal a ser totalmente digitalizado pelo projeto da ProQuest Historical Newspaper, que converteu eletronicamente as edições completas de grandes jornais, incluindo The Wall Street Journal, The Washington Post and The Christian Science Monitor.

Com mais de 3 milhões de páginas, mais de 25 milhões de matérias em 148 anos de história e 4 terabytes de dados, a conversão do Times é um esforço sem precedentes. A ProQuest desenvolveu um software para facilitar a transformação do texto analógico em ASCII. O reconhecimento óptico de caracteres chegou a 99,5% de precisão.

O jornalista do veículo impresso americano passou a ter muito mais opções do que o acesso ao antigo caderninho contendo números de telefones, realizar uma busca eficiente no departamento de pesquisa do jornal e freqüentar as grandes bibliotecas da cidade, por exemplo.

Ele também já contava com a Internet, ferramenta que aumentou ainda mais as possibilidades de pesquisa. Obteve-se, então, a oportunidade de vasculhar em websites de buscas e acessar banco de dados de organizações governamentais ou não.

Portanto, imensos volumes de informação, que têm sido sistematicamente coletados e armazenados, ultrapassam a capacidade humana, principalmente a do jornalista, na tarefa de levantar dados em pesquisas complexas e realizar os cruzamentos das informações para posterior análise.

Para ajudar nessa tarefa de peneirar dados, surgiu há 50 anos a técnica Computer-assisted Reporting (CAR). Apesar da sua constante evolução, o conceito é bastante amplo, pois compreende qualquer ferramenta (software) que ajuda no processo de obtenção de informação através de um computador.

Nesse mar de informação digital que está se formando, com diferentes configurações de bases de dados e de acesso, o jornalista da atualidade vê sua tarefa se tornar cada vez mais complexa na busca de obtenção de informações, apesar da aparente facilidade mostrada por esses dispositivos, mas que torna complexo o trabalho de se obter informações consolidadas e contextualizadas.

Esse artigo é uma tentativa de avançar em um conceito conhecido como Data Mining, já utilizado em outras atividades. A técnica é uma ferramenta para mineração de dados e descoberta de conexões complexas, que são quase impossíveis de serem encontradas em um mar de informações, por exemplo, utilizando apenas buscas na Internet ou técnicas como o CAR.

A pesquisa de informações na atualidade

Os microcomputadores eram usados para processar texto e tomaram o lugar das máquinas de escrever. Porém, essas máquinas só se tornaram poderosas ferramentas quando foram conectadas a redes internas para acesso a bancos de dados, ajudando na produção de material jornalístico.

Então, os bancos de dados surgem, portanto, nos veículos de comunicação, principalmente nos impressos, como grandes ferramentas para a pesquisa de informações que auxiliam o jornalista a contextualizar, complementar e checar informações, além de reduzirem drasticamente o tempo de busca (pesquisa por informações).

Os bancos de dados tinham como tarefa guardar velhos pedaços (clips) em uma biblioteca computadorizada, para serem utilizados no embasamento de matérias. Algumas redações desenvolveram base de dados para tópicos específicos, além de analisar registros do governo e de ajudarem em reportagens investigativas.

Visando a obtenção, tratamento, produção, empacotamento e distribuição da informação jornalística, fases do processo da notícia, cada veículo e/ou jornalista começa a criar a sua própria estrutura e técnica para realizar a primeira fase: a da obtenção de dados.⁴

Citando os autores WARD, Jean e HANSEN, Kathleen. (Search strategies in mass communication. Second Edition. New York: Longman, 1993), o pesquisador BASTOS (2000, p. 84) propõe um esquema de orientação para o profissional de jornalismo, que tem habilidade na pesquisa on-line e pode lidar com maior eficiência e eficácia com esses sistemas de fontes digitais. Ele descreve cinco níveis:

Primeiro, análise da questão (refere-se ao passo de restringir e definir a informação pretendida); Possíveis contribuintes (indica os três tipos de fontes de informação que podem ser utilizados, que incluem fontes informais, fontes institucionais e fontes de bibliotecas e base de dados, entre as quais as fontes on-line); Entrevistas (discussão de informação encontrada no nível precedente para trazer mais informação e significado sobre o assunto); Seleção e síntese (tornar a informação inteligível, juntando os fatos, idéias, interpretações e pontos de vista)⁵.

No caso de dois grandes grupos midiáticos brasileiros, o Grupo Abril e Folha de S. Paulo, apostaram na construção de bancos de dados. Segundo o site do grupo, “o banco de dados Folha é um acervo jornalístico que contém mais de oito décadas da história recente do Brasil. Seu objetivo é dar suporte aos jornalistas do Grupo Folha da Manhã e propiciar o atendimento a pesquisadores, estudantes e empresas na realização de pesquisas. O acervo inclui a coleção de jornais editados pelo grupo, arquivo de recortes com cerca de 100 mil pastas temáticas e 20 milhões de imagens em arquivos físico e digital.”⁶

Já o Grupo Abril tem o seu DeDoc, inaugurado em 1968. Tudo era manual. Em 1984, iniciou-se o processo de informatização, iniciando com um banco com a Veja, com acesso ao resumo de todas as matérias e pesquisa de palavras-referência.

⁴ LIMA JR. Walter Teixeira. **Mídia digital: o vigor das práticas jornalísticas em um novo espaço**. São Paulo, 2003. Tese (Doutorado em Jornalismo) – Escola de Comunicação e Artes da Universidade de São Paulo.

⁵ BASTOS, Helder. **Jornalismo Eletrônico: internet reconfiguração de práticas nas redações**. Coimbra: Livraria Minerva Editora, 2000.

⁶ < <http://www1.folha.uol.com.br/folha/bd/>> Acessado em 20 de setembro de 2004

Atualmente todas as revistas do grupo estão num banco de dados chamado Fólio News . “A Veja, carro-chefe da editora, por exemplo, tem 43.687 matérias; Anamaria, 19.587; Exame, 12.958 e Cláudia, 11.262”, afirma a pesquisadora do Dedoc há nove anos, Vera Lucia Lucas Pinto⁷.

Um grande avanço, mas há problemas no uso de banco de dados no jornalismo

A ProQuest reconhece que pesquisar em banco de dados com matérias (historical databases) é um desafio para os usuários. A empresa detecta pelo menos três problemas:

- Mudança na grafia da palavra: com o passar dos anos, pela a língua ser viva, grafias de algumas palavras vão mudando;
- Mudança de terminologia: as terminologias de algumas palavras também mudam. Por exemplo, Lula em 1968 tem um sentido, hoje, no Brasil, pelo menos dois e
- Imperfeições nos dados: datas erradas, troca de letras em nomes, dados imprecisos e outros.

Para não ter tantos problemas na consolidação de informações, os programadores têm inserido controladores, como o palavras-chave.

Com as palavras-chaves controladas, o sistema acusa se não é controlada. Para não ter outro significado: Governo Lula e controlado, Lula presidente. O PT é Partido dos Trabalhadores e não é sigla de avião. Nomes, normalmente, são controlados, por erros de grafia. A matéria no banco de dados, como a da Folha, é a que saiu no jornal e, de repente, é publicado o nome errado. Se você não tem o termo controlado, não irá achá-lo.⁸

Portanto, o surgimento das fontes digitais, como as on-lines, não representou um passe de mágica para a melhoria da qualidade na produção do jornalismo. As tecnologias on-lines não são uma panacéia que magicamente transformará as notícias com um razoável objetivo social.

A sua utilização como ferramenta de auxílio à profissão, a pesquisa em fontes digitais facilita o trabalho do jornalista na tarefa de localização da informação. Por outro lado, um profissional não bem preparado para lidar com esse tipo de processo encontrará problemas na verificação dos dados.

Sobre os jornalistas que procuram o DeDoc da Abril, Vera Lucia afirma que o tipo de procura varia muito e que os profissionais possuem muitas dificuldades para utilizar o sistema de busca. “Eles não colocam palavras-chave, não têm paciência e não têm tempo. Muitos não têm habilidade para pesquisar e se perdem, isso é muito comum. Também existe muita gente boa, que o consegue extrair uma pesquisa mais apurada, mas precisam de ajuda, pois não tem tempo”, afirma.

⁷ PINTO, Vera Lúcia Lucas. Entrevista concedida ao autor em 9 de setembro de 2004.

⁸ Ibid

Já o repórter investigativo e professor da ECA/USP, Cláudio Júlio Tognolli , que trabalhou no Dedoc em 1995, diz que tudo era feito à mão.

“Eu lembro que pessoas que lêem filosofia até a Revista Caras eram os melhores pesquisadores. Tinham o que denomino de ‘Cultura inútil’ mais completa. Conseguiram atacar os assuntos de lado. Que tipo de sapato usa o político até que tipo de perfume. Chamo isso de ‘cultura inútil’.”⁹

O duro refino na Internet

O surgimento da Internet no seu modo gráfico (www) e a possibilidade da busca de URLs e arquivos por programas como o Google, por exemplo, facilitaram muito o trabalho do jornalista na busca de mais informações. Mas existem as questões da imprecisão dos dados, credibilidades das fontes e a enorme quantidade de informações não-solicitadas que aparecem na tela do computador quando é realizada uma pesquisa em mecanismos de busca.

“Hoje com a Internet se tem acesso a bancos de dados, mas eles ainda não são bons. A busca na Internet, busca específica é eficiente. Mas se você for fazer, por exemplo, um perfil de governo em quatro anos, acha 10 mil registros”.¹⁰

Júlio Tognolli é um dos primeiros usuários do Google no Brasil. A informação do surgimento do mecanismo de busca foi trazida por uma amiga jornalista americana que visitava o País. Tognolli afirma que “hoje, vem a certeza: ninguém pode investigar um caso sem antes ter passado pelo menos duas horas em um desses sites de busca.”

Livres-associações

Sendo uma das suas ferramentas de obtenção de informação, Júlio Tognolli, até por ter trabalhado no Dedoc, no meio da década de 80, criou uma técnica de pesquisa na Internet que ele nomeia de ‘Livres Associações’.

No ano de 1993, eu ganhei um curso da Folha de S. Paulo para o Investigative Reporters and Editors (IRE), - www.ire.org, nos EUA. Fiz um curso de CAR. Em 93 era um ano que não se falava nisso porque não tinha Internet em quase nenhum lugar. Até porque o Philip Meyer tinha lançado o livro dele (Precision Journalism) em 1991, então, era com conceito, há doze anos isso era absolutamente desconhecido. A partir dali, comecei a me preocupar de nunca sair à rua sem fazer uma grande pesquisa.¹¹

⁹ TOGNOLLI, Cláudio Júlio. Entrevista concedida ao autor em 10 de setembro de 2004.

¹⁰ TOGNOLLI, Cláudio Júlio. **Investigação na Internet: sonho dirigido ou delírio controlado.** <<http://observatorio.ultimosegundo.ig.com.br/artigos/eno130220021.htm>> 13 abr de 2002. Acessado em 20 de setembro de 2004.

¹¹ TOGNOLLI, Cláudio Júlio. Entrevista concedida ao autor em 10 de setembro de 2004.

A técnica de Tognolli se baseia em sempre começar procurando pelo Google Imagens e nunca pelo Google Texto, pois segundo o jornalista, o mecanismo fornece um ‘substrato caótico’ de imagens mais interessante do que o outro. “Portanto, se eu tenho um determinado repertório baseado em livres associações sobre uma pessoa e eu quero pesquisar essa pessoa na Internet, eu penso por alguns minutos nela e a associo a uns vinte ou trinta vocábulos. Bem simples. E coloco o nome ‘dela e And Crime’, ‘And Carro’, ‘And Guitarra’, mas baseado na minha visão daquela pessoa. Então, eu faço um esquema booleando, usando And, com livres-associações.

Mas Tognolli ressalta que a técnica é eficaz porque ela se utiliza da vivência dele. Usa as suas informações e as joga numa busca caótica, porque é aberta. “Fiz a livre-associação baseada na minha experiência. Só eu tenho aquela informação (exclusiva). Fazia parte da minha vida”, afirma. Para exemplificar, Tognolli conta um episódio onde utilizou o Google para obter um furo jornalístico.

“Em 1997, tinha sido preso, no Estado de Tocantins, uma pessoa chamada Antonio da Mota Graça, vulgo Curica. Ele estava com sete toneladas de cocaína, dentro de toras e o advogado do Curica, que é advogado em São Paulo do Cartel de Medellín, é uma das minhas fontes. Bom, quando teve o seqüestro da filha do Sílvio Santos, todo mundo começou a fazer uma série de acusações contra o delegado Antonio Bélió. Ninguém sabia nada desse advogado. Um dia a minha fonte me liga e fala: sabia que eu estou advogando para o Bélió? Através dessa informação, fiz uma livre associação. Entrei no Google e digitei ‘Bélió AND Curica’. Uma coisa desconexa. Surgiu uma matéria do Estadão, 13 de maio, falando que esse delegado havia ido à casa de detenção do Carandiru retirar o Curica, dizendo que ele seria testemunha de um grande crime em Taboão da Serra. Quando o delegado removeu o acusado, o Curica foi resgatado pelos comparsas, ou seja, o delegado era acusado de ter facilitado o resgate. Quando eu coloquei no ar essa reportagem, pela rádio Jovem Pan, me ligou o delegado da Corregedoria e falou: o senhor teve acesso à ficha funcional do delegado Bélió. Ela é sigilosa. O senhor pode ser acusado de ter divulgado dados sob sigilo”.¹²

Portanto, Tognolli criou a sua técnica de encontrar informações ‘escondidas’ na internet e, categoricamente, afirma que para isso o jornalista tem que ter o que ele chama de ‘cultura inútil’ e informações privilegiadas. Provavelmente, outros jornalistas investigativos criaram as suas técnicas para obter informações. Mas elas são realmente eficientes e eficazes para todo o tipo de matéria? Talvez a utilização do Data Mining no jornalismo possa ajudar nesse aspecto.

O Que É Data Mining?

Definição importante de Data Mining elaborada por Usama Fayyad

“...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis”¹³

¹² TOGNOLLI, Cláudio Júlio. Entrevista concedida ao autor em 10 de setembro de 2004.

¹³ FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTHI, Padhraic. **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. In: Communications of the ACM, pp.27-34, Nov.1996

Essa definição foi apresentada 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.

Funções do Data Mining

Uma mineração de dados pode iniciar com uma simples descrição e caracterização dos elementos da base de dados ou de um data warehouse. Contudo, as principais tarefas da mineração de dados são:¹⁴

- a) formar grupos relativamente similares (agrupamentos) (Bussab, Miazaki, Andrade, 1990);
- b) visualizar inter-relações de dados multivariados através de gráficos relativamente simples (Johson, Wichern, 1998; Haykin, 2001);
- c) estabelecer modelos ou regras para classificar elementos em categorias previamente definidas (Hastie et al., 2001; Han, Kamber, 2001).;
- d) construir modelos para prever ou prever o valor de uma variável (Haykin, 2001; Neter et al., 1996);
- e) realizar análise de associação (Market Basquet Analysis) ((Berry, Linoff, 1997)

Onde e como é utilizado

São vários os setores da sociedade que trabalham com informação que utilizam a técnica do Data Mining para obter padrões válidos e potencialmente úteis em suas atividades. Há cinco anos, ao procurar eventuais relações entre o volume de vendas e os dias da semana, um software de Data Mining apontou que, às sextas-feiras, as vendas de cervejas, na rede Wal Mart, cresciam na mesma proporção que as de fraldas. Uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana .

Já o Bank of America usou essas técnicas para selecionar entre seus 36 milhões de clientes aqueles com menor risco de dar calote em um empréstimo. A partir desses relatórios, enviou cartas oferecendo linhas de crédito para os correntistas cujos filhos tivessem entre 18 e 21 anos e, portanto, precisassem de dinheiro para ajudar os filhos a comprar o próprio carro, uma casa ou arcar com os gastos da faculdade. Resultado: em três anos, o banco lucrou 30 milhões de dólares.

O governo dos EUA também utiliza o Data Mining há muito tempo: na identificação de padrões de transferências de fundos internacionais que se parecem com a lavagem de dinheiro do narcotráfico.

Porém, o governo americano está indo além da legalidade no uso do Data Mining. Como a técnica visa usar um programa de banco de dados para compilar e peneirar através de grandes quantidades de dados,

¹⁴ NAVEGA, Sérgio. **Princípios Essenciais do Data Mining**. <www.intelliwise.com/snavega>. Ago de 2002. Acessado em 20 de Setembro de 2004.

freqüentemente de natureza pessoal, vários órgãos dos EUA estão produzindo perfis de pessoas, analisando suas atividades e deduzindo padrões de informação.

Segundo a revista Wired, publicação americana de tecnologia e comportamento, a investigação da General Accounting Office (GAO) descobriu uma prática pervasiva em toda parte do governo americano, identificando 52 agências que tinham 199 projetos de Data Mining ativos ou em estágio de planejamento. Desses, o GAO encontrou 122 que usam informações pessoais de americanos.

Das agências envolvidas, o Departamento de Defesa teve o maior número de projetos, mas nem todos eram apontados para achar terroristas ou criminosos. Alguns foram desenhados para rastrear a performance de pessoal ou departamentos militares ou do governo. Outros departamentos usaram o Data Mining para achar fraudes, desperdício e abuso, análise científica ou pesquisa de informação.¹⁵

Portanto, as ferramentas de Data Mining são utilizadas para preverem futuras tendências e comportamentos. Empresas comerciais utilizam esse novo processo nas tomadas de decisão, baseando-se, principalmente, no conhecimento acumulado, que está 'invisível' em seus próprios bancos de dados.

Novo campo de uso: o jornalismo

Porém, há áreas em que o Data Mining ainda é pouco explorado, como na medicina. No momento, o ponto que está emperrando o uso de Data Mining é o fato de que a técnica, sendo uma nova concepção dirigida para pesquisa ainda é quase completamente desconhecida da comunidade médica. Mas a área fornece dados clínicos abundantes e, segundo os especialistas, esses dados são freqüentemente adequados a um estudo de Data Mining por não conterem dados que aparentemente são inúteis mas que são exatamente os que o pesquisador de Data Mining procura.

No jornalismo, como poderá ser visto na proposta a seguir, o Data Mining poderá ser útil, mas para isso, precisamos que os banco de dados sejam precisos e não históricos e que tenham uma certa inteligência artificial para lidar com as modificações semânticas das palavras, por exemplo.

Com o Data Mining podemos extrair padrões válidos, por exemplo, se o índice de desemprego diminui quando se aproxima uma eleição e porque isso acontece.

¹⁵ ZETTER, Kim. **GAO: Fede Data Mining Extensive**. In: Wired Magazine. <www.wired.com/news/privacy/0,1848,63623,00.htm> 27 may 2004. Acessado em 20 de setembro de 2004

Princípios essenciais do Data Mining no jornalismo

Temos no jornalismo um grande volume de dados guardados em arquivos históricos e, na Internet, temos acesso a banco de dados dos mais variados. Segundo Sergio Navega¹⁶, talvez a forma mais nobre de se utilizar vastos repositórios seja tentar descobrir se há algum conhecimento escondido neles.

Nesse ponto, o engenheiro afirma que por não haver uma solução eficaz para determinar padrões válidos, o 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”.

Entendo que essa necessidade de contar com ‘analistas humanos’ seja uma abertura para o trabalho de jornalistas especializados em mineração de dados e padrões válidos e úteis. O profissional para executar essa tarefa terá que possuir ‘conhecimento de mundo’ que as máquinas ainda não dispõe. Segundo Sérgio Navega, “talvez o futuro do Data Mining seja associar-se a sistemas de Inteligência Artificial que possam suprir parte dessa deficiência”.

Um dos conceitos importantes: encontrar padrões requer que os dados brutos sejam sistematicamente "simplificados" de forma a desconsiderar aquilo que é específico e privilegiar aquilo que é genérico. Para que o processo dê certo, é necessário sim desprezar os eventos particulares para só manter aquilo que é genérico.¹⁷

É um processo muito diferente comparado a análise de um grupo de informações jornalísticas, que tem como característica básica extrair dados de eventos isolados. No processo de Data Mining, é necessário se ‘perder’ um pouco de dados só para conservar a essência da informação, só assim existe a possibilidade de encontrar padrões¹⁸ válidos e potencialmente úteis.

A tarefa de localizar padrões não é privilégio do Data Mining. Segundo Sérgio Navega, o nosso cérebro utiliza-se de processos similares. “Muito do que se estuda sobre o cérebro humano também pode nos auxiliar a entender o que deve ser feito para localizar padrões”, afirma.

¹⁶ NAVEGA, Sérgio. **Princípios Essenciais do Data Mining**. <www.intelliwise.com/snavega>. Ago de 2002. Acessado em 20 de Setembro de 2004.

¹⁷ Ibid

¹⁸ Padrões são unidades de informação que se repetem ou, então, são seqüências de informações que dispõem de uma estrutura que se repete



Pode-se perceber no diagrama acima que há uma redução sensível no volume, que ocorre cada vez que se sobe um nível. A 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.¹⁹



¹⁹ NAVEGA, Sérgio. **Princípios Essenciais do Data Mining**. <www.intelliwise.com/snavega>. Ago de 2002. Acessado em 20 de Setembro de 2004.

No jornalismo, o Databases (fontes de dados) seriam compostos por bancos de dados com matérias publicadas (históricos), listas de conteúdo ou resumos de CD e DVD's e bancos de dados disponíveis em redes (Internet ou Intranet), mas que tivessem consistência nas informações disponíveis (dados precisos e pertinentes), remoções de ruídos e redundância.

Também teriam que ser mais amplos, ou seja, deixando de serem apenas repositórios de textos e fotos. Poderiam conter vídeo (por palavras-chave controladas, resumos, dados sobre sonoras, offs e videografia) e áudio (palavras-chave controladas, resumos, dados sobre sonoras e offs).



Bibliografia

BASTOS, Helder. **Jornalismo Eletrônico: internet reconfiguração de práticas nas redacções**. Coimbra: Livraria Minerva Editora, 2000.

BERRY, M. J. A., LINOFF, G. – **Data Mining techniques**. USA: John Wiley, 1997.

BUSSAB, A.; MIAZAKI, E. S.; ANDRADE, D. F. – **Introdução à análise de agrupamentos**. São Paulo: IX SINAPE, 1990.

FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTHI, Padhraic. **The KDD Process for Extracting Useful Knowledge from Volumes of Data**. In: Communications of the ACM, pp.27-34, Nov.1996.

HAN, J., KAMBER, M. – **Data Mining: concepts and techniques**. USA: Morgan Kaufmann, 2001.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. – **The elements of statistical learning**. USA: Springer, 2001.

JOHNSON, R. A., WICHERN, D. W. - **Applied multivariate statistical analysis**, 4 ed. USA: Prentice Hill, 1998.

LIMA JR. Walter Teixeira. **Mídia digital: o vigor das práticas jornalísticas em um novo espaço**. São Paulo, 2003. Tese (Doutorado em Jornalismo) – Escola de Comunicação e Artes da Universidade de São Paulo.

MAYFIELD, Kendra. **Read All About it**. In: Revista Wired, 29 Jul 2002 .
<www.wired.com/news/business/0,1367,54030,00.html> Acessado em 20 de setembro de 2004.

NAVEGA, Sérgio. **Princípios Essenciais do Data Mining**. <www.intelliwise.com/snavega>. Ago de 2002. Acessado em 20 de Setembro de 2004.

NETER, J.; KUTNER, M. H.; NACHTSHEIM, C. J.; WASSERMAM, W. – **Applied Linear Regression Models**. London: Richard D. Irwing, Inc., 3. ed., 1996.

ROZADOS, Helen Beatriz Frota. **O jornal e seu Banco de Dados: uma simbiose obrigatória**. In: DoIS (Documents in Information Science), Issue 1, Volume 26, Ano 1997. Acessado dia 20 de setembro de 2004 em <<http://dois.mimas.ac.uk/DoIS/data/Articles/juljqbfchy:1997:v:26:i:1:p:2805.html>> Acessado em 20 de setembro de 2004>

TOGNOLLI, Júlio Cláudio. **Investigação na Internet: sonho dirigido ou delírio controlado**. <<http://observatorio.ultimosegundo.ig.com.br/artigos/eno130220021.htm>> 13 abr de 2002. Acessado em 20 de setembro de 2004.

ZETTER, Kim. **GAO: Fede Data Mining Extensive**. In: Wired Magazine.
<www.wired.com/news/privacy/0,1848,63623,00.htm> 27 may 2004. Acessado em 20 de setembro de 2004.