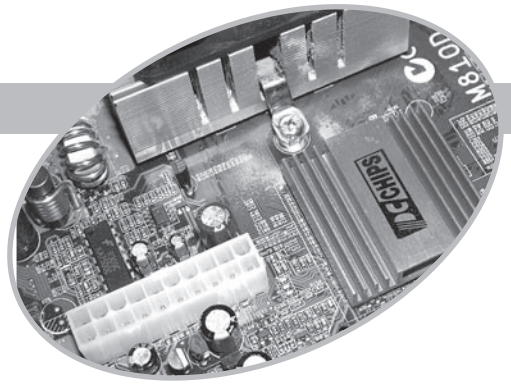
Jornalismo inteligente na era do *data mining*



Walter Teixeira Lima Junior

Doutor em Jornalismo Digital (ECA-USP)
Professor titular do programa de Pós-graduação da Faculdade Cásper Líbero
E-mail: digital@walterlima.jor.br

Resumo: O texto apresenta um ensaio sobre o uso do *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 melhoria da qualidade da informação pesquisada em bancos de dados e na obtenção de relações "invisíveis" de temas e contextos. A introdução de tecnologias no fazer jornalístico não é uma novidade. Porém, as novas formas de armazenar informações em ativos digitais e o desenvolvimento de programas computacionais munidos de certa "inteligência" abrem uma nova perspectiva no trabalho de obtenção e tratamento da informação jornalística.
Palavras-chave: jornalismo, banco de dados, *data mining* e busca.

Periodismo inteligente en la era del data mining
Resumen: El texto presenta un ensayo sobre el uso *data mining* en la extracción de datos en el proceso periodístico. La técnica ya se utiliza en otros campos de la actividad humana y, bien formateada, puede ayudar al periodismo en la mejora de la calidad de la información investigada en bases de datos y en la obtención de relaciones "invisibles" de temas y contextos. La introducción de tecnologías en el quehacer periodístico no es una novedad. Sin embargo, las nuevas formas de almacenar informaciones en activos digitales y el desarrollo de programas informáticos dotados de cierta "inteligencia" abren una nueva perspectiva en el trabajo de obtención y tratamiento de la información periodística.
Palabras clave: periodismo, base de datos, *data mining* y búsqueda.

Inteligent journalism in the data mining *era*
Abstract: The text presents an essay about the using of the *data mining* at the database mining toward the journalistic process.The technique has already been used in other fields of human being activity and, well sharped, can help journalism to achieve better results in the searching of the information quality researched in database and also for getting of invisible relations of themes and contexts. The introduction of technologies in order to make journalism is not something new. Therefore, the new ways of keeping information in digital actives and the development of computer programs filled of some inteligence open up a new perspective in the work of getting and dealing with the journalistic information.
Key words: journalism, database, *data mining* and searching.

Desde as descobertas de figuras rupestres desenhadas em cavernas vem-se constatando que o armazenamento de informações é condicionante do ser humano. Há cinco mil anos, os distantes sumérios, na região onde hoje é o Iraque, esculpiram em plaquetas de argila os primeiros sinais, em escrita cuneiforme. Passando pelas inscrições 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. Também nas culturas ágrafas revela-se o compromisso de perpetuar o antigo por intermédio da oralidade.

Na atualidade, com o advento da tecnologia de armazenamento digital, quase toda a informação produzida passou a ser colocada diretamente no mundo de *bits* e *bytes*, e o que existe no meio físico, como em livros, revistas e jornais, para citar alguns exemplos, está sendo transmutado para discos rígidos ou para 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. Com tecnologia avançada, os sistemas de processamento, armazenamento, controle, recuperação e dissemi-

nação da informação permitem gerenciar as bases de dados e material informacional em texto e imagem.

Digitalização completa de um jornal

Vejamos o audacioso projeto do *New York Times*, finalizado em 2002. A ProQuest, 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 originalmente impresso. Os usuários do sistema podem pesquisar eventos históricos de 1851 a 1999.

Os bancos de dados, como ferramentas de pesquisa, ajudam a contextualizar, complementar e checar informações, reduzindo o tempo de busca

O *Times* foi o primeiro jornal a ser totalmente digitalizado pelo projeto da ProQuest *Historical Newspaper*, que convertera eletronicamente as edições completas de outros grandes jornais, incluindo *The Wall Street Journal, The Washington Post* e *The Christian Science Monitor*.

Com mais de 3 milhões de páginas, mais de 25 milhões de matérias em 148 anos de história e quatro 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 alcançou 99,5% de precisão.

O jornalista do veículo impresso americano passou a ter muito mais opções do que acessar o antigo caderninho com números de telefones, realizar uma busca eficiente no departamento de pesquisa do jornal ou 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, assim, a oportunidade de vasculhar em *websites* de buscas e acessar banco de dados de organizações, governamentais ou não.

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 C*omputer-assisted Reporting* (CAR). Apesar da sua constante evolução, o conceito é bastante amplo, pois compreende qualquer ferramenta (*software*) que ajude no processo de obtenção de informação através de um computador.

No 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 tornar-se cada vez mais complexa na busca de informações, apesar da aparente facilidade mostrada por esses dispositivos. É complexo, no entanto, o trabalho de obtenção de informações consolidadas e contextualizadas.

Este 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, quase impossíveis de serem encontradas, nesse mar de informações, por exemplo, através apenas da utilização de 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 novas máquinas só se tornaram poderosas ferramentas quando conectadas a redes internas para acesso a bancos de dados, ajudando na produção de material jornalístico.

Os bancos de dados surgem, portanto, nos veículos de comunicação, principalmente nos impressos, como grandes ferramentas para a pesquisa que auxilia o jornalista a contextualizar, complementar e checar informações, reduzindo drasticamente o tempo de busca.

Os bancos de dados tinham como tarefa guardar velhos pedaços ou recortes de jornais (*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. Com o tempo, visando à obtenção, tratamento, produção, empacotamento e distribuição da informação jornalística – como fases do processo da notícia –, cada veículo e/ou jornalista começa a criar a sua própria estrutura e técnica para realizar a primeira dessas fases, a da obtenção de dados.

Citando a obra *Search strategies in mass communication*, de Jean Ward e Kathleen Hansen, Bastos (2000:84) propõe um esquema de orientação para o profissional de jornalismo com habilidade na pesquisa on-line e que é capaz de lidar com maior eficiência e eficácia com esses sistemas de fontes digitais. Ele descreve cinco níveis:

1) Análise da questão (refere-se ao passo de restringir e definir a informação pretendida);
2) 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 *online*);
3) Entrevistas (discussão de informação encontrada no nível precedente para trazer mais informação e significado sobre o assunto);
4) Seleção;
5) Síntese (tornar a informação inteligível, juntando os fatos, idéias, interpretações e pontos de vista).

Dois grandes grupos midiáticos brasileiros, o Grupo Abril e o grupo que inclui a *Folha de S.Paulo*, apostaram na construção de bancos de dados. Eis o que informa o site de um desses grupos:

> ...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[1].

Já o Grupo Abril tem o seu Dedoc, inaugurado em 1968. Antes, tudo era manual. Em 1984, iniciou-se o processo de informatização. Primeiro foi a vez da revista *Veja*, com acesso ao resumo de todas as matérias e pesquisa de palavras-referência. 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", informa Vera Lúcia Lucas Pinto (2004), pesquisadora do Dedoc há nove anos.

Avanços e problemas

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: a) mudança na grafia da palavra: com o passar dos anos, uma vez que a língua é viva, a grafia de algumas palavras modifica-se; b) mudança de terminologia: as terminologias de algumas palavras também mudam. Por exemplo, Lula em 1968 tem um sentido e, hoje, no Brasil, pelo menos dois; e c) 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 costumam inserir controladores, como o de pala-

[1] Disponível em http://www1.folha.uol.com.br/folha/bd/. Acessado em 12/3/2006.

vras-chave. Com as palavras-chave controladas, o sistema acusa se não for o caso. Exemplos: Governo Lula, Lula presidente. 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 o da *Folha*, é a que saiu no jornal e, de repente, sai com um nome errado. Se o termo não for controlado, não irá ser encontrado.

Portanto, o surgimento das fontes digitais *online* não representou um passe de mágica para a melhoria da qualidade na produção do jornalismo. As tecnologias *online* não são uma panacéia que magicamente transformará as notícias, carregando-as com alto teor de relevância social. Como ferramenta de auxílio à profissão, a pesquisa em fontes digitais facilita o trabalho do jornalista na tarefa de localização da informação. Um profissional não bem preparado para usar esse tipo de processo encontrará problemas na verificação dos dados.

Sobre os jornalistas que visitam o Dedoc da *Abril*, Vera Lúcia 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 nem tempo. Muitos não têm habilidade para pesquisar e se perdem, o que é muito comum. Também existe muita gente boa, que consegue extrair uma pesquisa mais apurada, mas que precisa de ajuda, pois não tem tempo."

O repórter investigativo e professor da ECA/USP Cláudio Júlio Tognolli, que trabalhou no Dedoc em 1995, lembra que na época tudo era feito à mão. "Eu lembro que pessoas que lêem desde filosofia até a revista *Caras* eram os melhores pesquisadores. Tinham o que denomino 'cultura inútil' mais completa. Conseguiam atacar os assuntos de lado: que tipo de sapato usa o político até que tipo de perfume" (Tognolli, 2004).

O 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, da credibilidade das fontes e da 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" (Tognolli, 2002).

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. Ele 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

Tognolli criou uma técnica de pesquisar na internet que chamou de "livres associações. Ele explica:

> 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 (*Computer-assisted Reporting*). Era um ano em 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, onde o conceito foi cunhado. Há doze anos isso era absolutamente desconhecido. A partir dali, comecei a me preocupar de nunca sair à rua, sem fazer uma grande pesquisa (Tognolli, 2004).

A técnica de Tognolli baseia-se 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 sistema:

> 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 (Tognolli, 2004).

Mas Tognolli ressalta que a técnica é eficaz, porque a ela se soma à 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, conta um episódio em que utilizou o *Google* para obter um furo jornalístico.

> Em 1997, tinha sido presa 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 Bélio. Ninguém sabia nada desse advogado. Um dia, a minha fonte me liga e fala: sabia que eu estou advogando para o Bélio? Através dessa informação, fiz uma livre associação. Entrei no *Google* e digitei "Bélio *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 Bélio. Ela é sigilosa. O senhor pode ser acusado de ter divulgado dados sob sigilo (Tognolli, 2004).

Assim como 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 também criaram as suas técnicas para isso. 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.

Vários setores que trabalham com informações utilizam a técnica do data mining *para obter padrões válidos e potencialmente úteis em suas atividades*

O que é *data mining*

Definição importante de *data mining* elaborada por Usama Fayyad (1996): "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis". Essa definição foi apresentada para explicar o termo KDD (*Knowledge Discovery in Databases*), processo que engloba a mineração. Portanto, *data mining* seria apenas um dos passos necessários ao processo todo.

A 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[2] são:

a) formar grupos relativamente similares, ou agrupamentos (Bussab, Miazaki, Andrade, 1990);
b) visualizar inter-relações de dados multivariados através de gráficos relativa-

[2] Disponível em http://www.intelliwise.com/snavega. Acessado em 12/3/2006.

mente 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 predizer ou prever o valor de uma variável (Haykin, 2001; Neter et al., 1996);
e) realizar análise de associação (*Market Basquet Analysis*) (Berry, Linoff, 1997).

São vários os setores 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 cerveja 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 fim de semana.

A tarefa de localizar padrões não é privilégio do data mining. *Segundo Navega, o cérebro humano utiliza-se de processos similares*

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 que, portanto, precisassem de dinheiro para ajudar esses 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 pareçam com lavagem de dinheiro do narcotráfico. Porém, o governo americano está indo além da legalidade nesse campo. Como a técnica visa usar um programa de banco de dados para compilar e peneirar através de grandes quantidades de dados, 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 destinados a 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[3].

Portanto, as ferramentas de *data mining* são utilizadas para prever 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

Há áreas em que o *data mining* ainda é pouco explorado, como na medicina, talvez porque a técnica, uma nova concepção dirigida para pesquisa, ainda seja quase completamente des-

[3] Disponível em http://www.wired.com/news/privacy/0,1848,-63623,00.htm . Acessado em 12/3/2006.

conhecida 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 porque, embora aparentemente inúteis, são exatamente o que o pesquisador de *data mining* procura.

No jornalismo, como na proposta a seguir, o *data mining* poderá igualmente ser útil, mas, para isso, é preciso que os bancos 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* é possível extrair padrões válidos, por exemplo, para investigar se o índice de desemprego diminui quando se aproxima uma eleição e por que isso acontece.

No jornalismo, é grande o volume de dados guardados em arquivos históricos e, na internet, temos acesso a banco de dados dos mais variados. Segundo Sérgio Navega (2002), 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 solução eficaz para determinar padrões válidos, o *data mining* ainda requer "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, precisa ter "conhecimento de mundo" de que as máquinas ainda não dispõem. Segundo 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 (Navega, 2002).

É um processo muito diferente, quando comparado à análise de um grupo de informações jornalísticas, que tem como característica básica extrair dados de eventos isolados. No processo de *data mining*, faz-se necessário "perder" alguns dados para conservar a essência da informação. Só assim existe a possibilidade de encontrar padrões[4] válidos e potencialmente úteis.

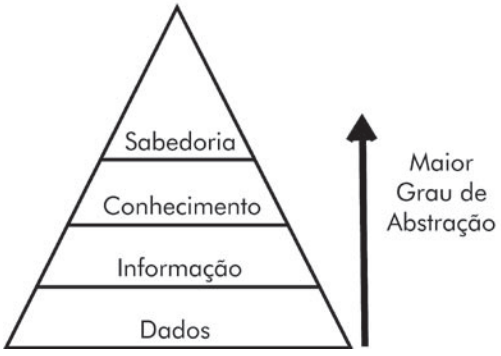
A tarefa de localizar padrões não é privilégio do *data mining*. Ainda segundo Navega (2002), o nosso cérebro utiliza-se de processos similares. "Muito do que se estuda sobre o cérebro humano também pode nos auxiliar a entender o que deve ser feito para localizar padrões".

Pode-se perceber, no diagrama acima, redução sensível no volume, que ocorre cada vez que se sobe um nível. A redução de volume é uma conseqüência natural 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

[4] Padrões são unidades de informação que se repetem ou, então, são seqüências de informações que dispõem de uma estrutura que se repete.

informação. O processo de *data mining* localiza padrões através da judiciosa aplicação de processos de generalização, algo que é conhecido como indução. (Navega, 2002).



No jornalismo, os *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 de ser mais amplos, ou seja, deixar de ser 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*).

Referências

BASTOS, Helder. *Jornalismo electrônico: internet reconfiguração de práticas nas redações*. 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". *Communications of the ACMI*, nov.1996, pp.27-34.

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 multivarate statistical analysis*. 4a. edição, USA: Prentice Hill, 1998.

LIMA JR. Walter Teixeira. *Mídia digital: o vigor das práticas jornalísticas em um novo espaço*. Tese (Doutorado em Jornalismo). São Paulo, Eca-USP, 2003.

MAYFIELD, Kendra. "Read all about it". *Revista Wired*, 29 Jul. 2002. Disponível em http://www.wired.com/news/business/0,1367,54030,00.html. Acessado em 12/3/2006.

NAVEGA, Sérgio. "Princípios essenciais do data mining". Disponível em http://www.intelliwise.com/snavega. Agosto de 2002. Acessado em 12/3/2006.

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

PINTO, Vera Lúcia Lucas. Entrevista concedida ao autor em 9/9/2004.

ROZADOS, Helen Beatriz Frota. "O jornal e seu banco de dados: uma simbiose obrigatória". *DoIS (Documents in Information Science)*, issue 1, vol. 26, 1997. Disponível em http://dois.mimas.ac.uk/DoIS/data/Articles/juljqbfchy:1997:v:26:i:1:p:2805.html. Acessado em 12/3/2006.

TOGNOLLI, Júlio Cláudio. "Investigação na internet: sonho dirigido ou delírio controlado". Disponível em http://observatorio.ultimosegundo.ig.com.br/artigos/eno130220021.htm. 13 abril de 2002. Acessado em 12/3/2006.

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

WARD, Jean e HANSEN Kathleen. *Search strategies in mass communication*. 2ªedição, New York: Longman, 1993.

ZETTER, Kim. "GAO: Fede Data Mining Extensive". *Wired Magazine*, 27 may 2004. Disponível em http://www.wired.com/news/privacy/0,1848,63623,00.htm. Acessado em 12/3/2006.