

# WEB INTELLIGENCE – INTELIGÊNCIA ARTIFICIAL PARA DESCOBERTA DE CONHECIMENTO NA WEB

Stanley Loh  
Ramiro Saldaña Garin

## Resumo

Este tutorial apresenta a área de Web Intelligence (WI), cuja característica principal é a aplicação de técnicas e ferramentas de Inteligência Artificial (IA) na Web. Neste trabalho, será dada ênfase aos aspectos da WI relacionados à busca de conhecimento na Web. Serão apresentadas as principais técnicas e estratégias de IA que podem apoiar o processo de descoberta de conhecimento. As diferentes aplicações de WI serão discutidas, e os desafios ainda existentes e que exigem maior atenção da comunidade de pesquisa serão delineados ao final.

## 1 Introdução

A Web é uma coleção heterogênea de documentos distribuídos, inter-relacionados, estruturados, semi-estruturados e não-estruturados, contendo textos, imagens e sons (Yao et al, 2001). Garofalakis e outros (1999) prevêm que a maior parte do conhecimento humano estará disponível na Web em 10 anos.

Entretanto, ter disponível tanto conhecimento não garante às pessoas que irão encontrar o que desejam num tempo razoável. Este fenômeno já é notado pela maioria dos usuários freqüentes na Web, mesmo existindo tantos sites e ferramentas de busca na Internet. Este problema é conhecido como “sobrecarga de informações” (*information overload*), e acontece quando o usuário tem muita informação ao seu alcance, mas não tem condições de tratá-la ou de encontrar o que realmente deseja ou lhe interessa (Chen, 1994).

Para complicar a situação, segundo Upchurch e outros (2001), o crescimento de *page views* (medida para quantificar o número de páginas vistas na Web) é exponencial: a média de páginas vistas dobra por usuário a cada 30 meses. Isto quer dizer que as pessoas estão visitando mais páginas na Web.

Some-se a isto, o crescimento no volume de informações publicadas. Bharat & Broder (1998) estimaram que, em novembro de 1997, havia 200 milhões de páginas na Web, sendo que apenas uma parte destas (80% ou 160 milhões de páginas) estavam indexadas nos 4 maiores sites de busca (AltaVista, HotBot, Excite e Infoseek).

Além disto, o crescimento também se dá nas transações online, principalmente para comércio eletrônico. Sendo que 79% dos usuários que navegam com este fim, buscam algum tipo de informação sobre produtos ou serviços (Cadê/IBOPE, 1999).

Tais problemas evidenciam a necessidade de mecanismos para auxílio às pessoas em processos de busca de informações ou conhecimento. E estes mecanismos devem ser

inteligentes, pois não há pessoas representando as organizações na Web, mas somente computadores (Spiliopoulou & Pohle, 2001).

Com tal objetivo como foco, surgiu a área de **Web Intelligence**. Segundo Zhong e outros (2000), Web Intelligence é um novo campo de pesquisa que explora a Inteligência Artificial (IA) e a Tecnologia de Informação avançada para o desenvolvimento de sistemas inteligentes para a Web (*Web-based intelligent information systems*).

Tais sistemas devem realizar funções relacionadas à inteligência humana (raciocínio, aprendizado e auto-melhoria). Desta forma, WI pode ser vista como a aplicação de IA na Web (Yao et al, 2001).

Segundo Kohavi (2001), WI é mais que a simples análise do *log* de servidores Web. Apesar de conter informações sobre o uso da Web, o *log* ainda é muito restrito. Por exemplo, não aparecem produtos deixados de lado num processo de compra ou que conteúdos foram lidos pelo usuário. A WI deve trabalhar com informações mais complexas. Sua finalidade principal é permitir entender e interpretar as regularidades no caos da Web (Yao et al, 2001).

Neste tutorial, será dada ênfase aos aspectos da WI relativos a processos de descoberta de conhecimento na Web. Serão detalhadas técnicas e estratégias de Inteligência Artificial que podem ser empregadas para WI. Depois serão apresentadas algumas aplicações de WI e por fim discutidos alguns pontos que ainda necessitam maior desenvolvimento.

## 2 Dados da Web

Há muitos e diferentes dados trafegando pela Web que podem ser usados para WI. Os dados podem ser coletados, segundo Srivastava e outros (2000), em:

- a) bancos de dados da organização: armazenando transações online;
- b) servidores Web: analisando arquivos de log de acesso, por “packet sniffers” (que monitoram o tráfego e extraem dados diretamente de pacotes TCP/IP) ou por login do usuário cadastrado;
- c) clientes Web: através de cookies, plug-ins, agentes remotos (Javascripts e Java applets) e componentes de software instalados no usuário (modificando o código do browser, por exemplo); e em
- d) servidores proxy: principalmente para analisar páginas armazenadas em cache.

Os dados podem dizer respeito a (Schafer et al, 2001; Kohavi & Becher, 2001):

- atributos demográficos do usuário (sexo, idade, localização): fornecidos pelo próprio, descobertos no mundo real (ex: endereço de entrega de produtos comprados online) ou pela rede (ex: domínio de origem);
- navegação explícita: páginas requisitadas ou visitadas, links seguidos, escolhas de atributos pelo usuário ou parâmetros fornecidos como entrada, tempo gasto;
- navegação implícita: itens escolhidos e o que está sendo visto pelo usuário;
- palavras-chave usadas em buscas;
- histórico de relacionamento do usuário com o site: compras feitas, páginas visitadas, documentos ou elementos baixados por download, revisitas;
- feedback do usuário (ratings): preferências, críticas, opiniões e comentários;
- conteúdo visitado;

- descartes do usuário: produtos colocados no cesto e tirados, páginas não carregadas totalmente.

Srivastava e outros (2000) afirmam que os dados sobre páginas Web podem ser classificados quanto:

- ao conteúdo : significado de textos e imagens nas páginas;
- à estrutura intra-página: arranjo de tags (HTML, XML) dentro da página, como uma árvore;
- à estrutura inter-páginas: hiperlinks que relacionam as páginas;
- ao uso: endereços IP, data da visita, tempo gasto nas visitas, requisições de URL's, operações no browser (botões clicados como reload e back/forward, salvar, adicionar a bookmarks).

### 3 Técnicas de IA para Descoberta de Conhecimento

Esta seção apresenta as principais técnicas e abordagens estudadas em Inteligência Artificial. Será dada ênfase às técnicas relativas ao processo de descoberta de conhecimento.

#### 3.1 Data Mining

Segundo Fayyad e outros (1996) Descoberta de Conhecimento em Bases de Dados (*KDD*) ou Mineração de Dados (*Data Mining*) é um processo não-trivial de identificar, em bases de dados, padrões que sejam válidos, novos, potencialmente úteis e compreensíveis ao usuário. *KDD* inclui técnicas e ferramentas inteligentes e automáticas para auxiliar pessoas em analisar grandes volumes de dados para garimpar conhecimento útil.

Ainda segundo os mesmos autores, a Mineração de Dados é a parte do processo de *KDD* responsável pela aplicação de algoritmos para extrair padrões nos dados, enquanto que a Descoberta de Conhecimento é um processo maior, envolvendo também a interpretação dos resultados.

A seguir, são descritas rapidamente as principais técnicas existentes para Data Mining.

- classificação

A técnica de classificação tem por objetivo alocar elementos em classes pré-existentes. Segundo Fayyad e outros (1996), o objetivo é utilizar uma função para mapear os elementos em classes pré-definidas (associar elementos a classes).

As classes devem ter sido definidas através de suas características. O mecanismo de classificação então compara as características do elemento candidato com as características das classes existentes. Como resultado, tem-se um *ranking* das classes mais prováveis, dado pelo grau de pertinência do elemento na classe. É possível associar o elemento em mais de uma classe ou então tomar a primeira do *ranking* como a classe resultante.

- indução de regras de classificação

Esta técnica está associada à técnica de classificação e é por muitos autores classificada também como um tipo de classificação. Entretanto, enquanto que a classificação trabalha com definições já pré-existentes para as classes, a técnica de indução

procura justamente descobrir as características que definem elementos de uma classe. O objetivo é encontrar regras que possam ser usadas posteriormente por técnicas de classificação.

Os algoritmos mais conhecidos são o ID3 e o C4.5, cujo resultado são árvores de decisão (Ingargiola, 1996). Uma árvore de decisão é um conjunto de regras que analisa cada característica e determina a classe mais provável. A criação das árvores é feita por técnicas de aprendizado supervisionado de máquina, que analisam atributos de casos exemplos ou de treino, selecionados manualmente por pessoas para representar classes. A idéia é identificar características que existam em alguma classe e não em outras. Também podem ser usados exemplos negativos, que servem para a identificação de características que afastem elementos de uma classe.

- modelos de predição

A técnica de modelos de predição é semelhante à anterior. Entretanto, ao invés de procurar alocar elementos em classes, aqui o objetivo é descobrir uma função matemática que descreva o comportamento de um sistema (calcular valores em função de outros) (Goebel & Gruenwald, 1999). A finalidade é poder prever valores futuros (por exemplo, unidades vendidas em função do tempo).

A mesma técnica também pode ser aplicada para gerar uma função média entre diferentes comportamentos (ex: desempenho médio de filiais).

- detecção de desvios

Esta técnica utiliza uma função média, representando o comportamento normal de um sistema, para avaliar possíveis desvios.

- *clustering*

Agrupamento ou *clustering* é o processo inverso da classificação, pois parte de uma situação em que não existem classes, somente elementos de um universo (não se sabe quais são as classes, nem quantas, muito menos as características de cada uma). A partir dos elementos, as técnicas de *clustering* são responsáveis por definir as classes e enquadrar os elementos (Han et al, 1996; Agrawal, 1995).

O objetivo então é identificar automaticamente grupos de afinidades, avaliando a similaridade entre os elementos e colocando os mais semelhantes no mesmo grupo e os menos semelhantes em grupos diferentes (Willet, 1988).

Em geral, a avaliação de similaridade entre os elementos é feita através de uma função de similaridade, analisando as características que representam os elementos.

- análise de *cluster*

Esta técnica complementa a anterior, pois procura identificar as características comuns nos elementos de cada grupo ou classe. Na maioria dos mecanismos, o fundamento está em identificar um centróide para a classe, ou seja, um vetor de características médias para representar os elementos de uma classe (Han et al, 1996; Agrawal, 1995).

- associação ou correlação

É a mais conhecida das técnicas de Data Mining e muitas vezes se confunde com o processo. Foi esta regra que permitiu descobrir um dos achados mais famosos na cadeia de

supermercados Wall-Mart: “nas sextas-feiras, quem comprava fraldas também comprava cerveja”. A técnica de associação ou correlação verifica se existe algum controle ou influência entre atributos ou valores de atributos (Han et al, 1996; Agrawal, 1995).

O objetivo então é encontrar dependências entre atributos ou valores através da análise de probabilidades condicionais. Em geral, os resultados são apresentados na forma de regras  $X \rightarrow Y$ , que significa que “se X está presente, então Y tem chances de estar presente também”. O primeiro elemento X pode ser uma combinação de atributos ou valores, formando assim regras mais complexas.

As regras possuem dois graus associados: a confiança e o suporte. O suporte é o número de casos onde a regra foi encontrada (onde X e Y aparecem juntos). A confiança é a probabilidade condicional da regra, ou seja, quais as chances do segundo elemento estar presente. É calculada pela divisão do número de casos onde X e Y aparecem juntos (o suporte absoluto), pelo número de casos onde somente X aparece.

- análise de séries temporais

Esta técnica procura encontrar padrões na repetição seguida de valores. Por exemplo, analisando-se as ações de uma empresa na bolsa de valores, pode-se notar que depois de tantos meses subindo, o valor das ações diminui em tantos pontos percentuais.

- evolução ou seqüência de tempo

As técnicas de evolução ou seqüência de tempo buscam descobrir regras de associação ou correlação entre eventos ocorridos em momentos diferentes. Por exemplo, uma loja pode identificar que clientes que compram um sapato voltam depois de um mês para comprar uma camisa (Han et al, 1996; Agrawal, 1995).

### **3.2 Text Mining ou Descoberta de Conhecimento em Textos**

O termo Descoberta de Conhecimento em Textos foi utilizado pela primeira vez por Feldman e Dagan (1995) para designar o processo de encontrar algo interessante em coleções de textos (artigos, histórias de revistas e jornais, mensagens de *e-mail*, páginas *Web*, etc.). Hoje em dia, sinônimos como *Text Mining* ou *Text Data Mining* também são utilizados para o mesmo fim (Tan, 1999).

Pode-se então definir **Descoberta de Conhecimento em Textos (KDT)** ou **Text Mining** como sendo o processo de extrair padrões ou conhecimento, interessantes e não-triviais, a partir de documentos textuais (Tan, 1999).

A seguir, são apresentadas as principais técnicas para Text Mining (maiores detalhes em Loh, 1999).

- extração

Uma técnica clássica é a de Extração de Informação (EI), cujo objetivo é encontrar informações específicas dentro dos textos (conforme Sparck-Jones e Willet, 1997). Riloff e Lehnert (1994) afirmam que o objetivo da área de EI é diferente do objetivo da área de processamento de linguagem natural (PLN), porque é mais focado e mais bem definido, visando extrair tipos específicos de informação. A técnica de EI procura converter dados não-estruturados em informações explícitas, geralmente armazenadas em bancos de dados estruturados. Isto pode ser feito isolando-se partes relevantes do texto, extraindo

informação destas partes e transformando-as em informações mais digeridas e melhor analisadas (Cowie & Lehnert, 1996). Em geral, os métodos utilizados são direcionadas para extrair características do domínio (objetos, entidades, relações), servindo apenas para aplicações específicas (Croft, 1995).

A estratégia mais utilizada é analisar “tags” (marcas) nos textos que possam indicar a presença de um dado. Por exemplo, o termo “anos” pode indicar que o numeral que o precede é a idade de alguém.

- categorização

Outra técnica básica é a categorização de textos, cujo objetivo é associar categorias (assuntos, classes ou temas) pré-definidas a textos livres (Yang & Liu, 1999). Há muitos trabalhos neste área, apresentando diversos métodos para categorização de textos (Apté e outros, 1994; Cohen & Singer, 1996; Yang & Chute, 1994; Liddy et al, 1994).

Yang e outros (Yang e Pedersen, 1997; Yang & Liu, 1999) fazem análises de vários métodos de categorização. Mas também há a preocupação com a escolha das características textuais que serão usadas no método. Yang e Pedersen (1997) comparam métodos para fazer esta seleção. Em geral, os trabalhos de categorização de textos procuram encontrar o tema central de um texto (ou temas, se houver mais de um).

- análise de características ou descrição de conceitos (centróide)

A idéia é apresentar uma lista com os conceitos principais de um único texto (geralmente, os conceitos são termos ou expressões extraídos por análises estatísticas). Moscarola e outros (1998), por exemplo, sugerem uma lista de termos próximos (antes e depois), os quais permitem a análise do conteúdo por quase-frases. Uma técnica similar, discutida em (Maarek, 1992), apresenta cadeias de palavras relacionadas por afinidades léxicas (relações sintáticas).

- análise lingüística

A abordagem por análise lingüística procura descobrir informações e regras analisando sentenças da linguagem a nível léxico, morfológico, sintático e semântico. Analisando padrões sintáticos (*tags*), as técnicas permitem descobrir generalizações escondidas, inferências de relações de coerência em textos (por exemplo, causa e efeito), relações de tempo e relações conceituais (definições, exemplos, partições e composição) através de *tags* no texto (maiores detalhes sobre trabalhos relacionados ver Loh, 1999).

- resumos ou sumarização

A abordagem de descoberta por *Sumarização* ou resumos utiliza as técnicas dos tipos anteriores, mas com ênfase maior na produção do resumo ou sumário. Segundo Sparck-Jones e Willet (1997), *sumarização* é a abstração das partes mais importantes do conteúdo do texto. Miike e outros (1994) apresentam um trabalho de geração automática de resumos em tempo de execução através de interações com o usuário. Já McKeown e Radev (1995) apresentam técnicas e ferramentas para analisar diversos artigos sobre um mesmo evento e criar um resumo em linguagem natural. Em (Hersh et al., 1995) é apresentada uma ferramenta para *sumarização* com dois componentes principais: um planejador de conteúdo (que seleciona informações de uma base) e um componente lingüístico (para gerar as frases de saída em linguagem natural).

- associação entre textos

A descoberta por associação entre textos procura relacionar descobertas presentes em vários textos diferentes. Por exemplo, Swanson e Smalheiser (1997) fizeram descobertas na área médica relacionando textos que não se referenciam e que aparentemente não continham assuntos comuns. Em (McKeown & Radev, 1995), é apresentada uma ferramenta que analisa diversos artigos sobre um mesmo evento e cria um resumo único em linguagem natural. São extraídas informações de partes dos textos e analisadas para encontrar similaridades e diferenças de informações.

- *clustering*

A descoberta por Agrupamento (*clustering*) procura separar automaticamente elementos em grupos por afinidade ou similaridade (não há classes pré-definidas). A técnica de agrupamento é diferente da classificação, pois a primeira visa criar as classes através da organização dos elementos, enquanto que a segunda procura alocar elementos em classes já pré-definidas (Willet, 1988). O agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes.

As técnicas podem ser aplicadas sobre palavras ou sobre conceitos. Conceitos permitem trabalhar com sinônimos e variações léxicas (maiores detalhes em Loh, 2000).

### **3.3 Web Mining**

Web Mining se refere à identificação de padrões no comportamento de uso da Web. Srivastava e outros (2000) sugerem aplicar as técnicas de data mining sobre dados coletados na Web. Para tanto, pode-se fazer análise estatística sobre páginas visitadas (page views), tempo gasto entre páginas, páginas mais frequentemente vistas. Pode-se também descobrir associações entre páginas frequentemente vistas na mesma sessão e que não estão relacionadas por hiperlinks. Podem ser analisados padrões sequenciais (eventos que precedem outros, por exemplo, páginas vistas antes de certas ações) e padrões transversos (páginas não diretamente ligadas por hiperlinks mas relacionadas por meio de outras).

Spiliopoulou e Pohle (2001) sugerem analisar associações entre páginas vistas na mesma seção ou caminhos mais frequentes (para sugerir atalhos ou criar links novos), realizar agrupamento (*clustering*) de páginas requisitadas (correlacionadas mas sem links entre elas) e analisar hierarquias de conceitos sobre páginas (criar manualmente por conteúdo ou por objetivo da página ou criar automaticamente por classificação de conteúdo).

Já Spiliopoulou e outros (2000) discutem estratégias para analisar seqüências de páginas (*sequence mining*) para entender o comportamento dos usuários. Podem ser abstraídas páginas do meio (como curingas) já que os caminhos podem ser longos. Eles também sugerem comparar padrões de compradores e não compradores ou procurar identificar diferença entre usuários visitantes rápidos, usuários investigadores e usuários compradores. O conhecimento descoberto pode servir para reprojeter páginas com o objetivo de maximizar a eficiência de contato (razão entre investigadores e todos usuários) e a eficiência de conversão (razão entre compradores e investigadores).

Berendt (2001) comenta que os padrões de comportamento estão ficando longos e complexos (como seqüências de DNA). Para tanto propõem uma técnica alternativa utilizando hierarquias de conceitos. Assim, páginas são tratadas como instâncias de um conceito de mais alto nível, que pode ser um tipo de conteúdo ou serviço requisitado (parâmetros fornecidos pelo usuário). Os prefixos são tratados como tipos mais gerais.

### **3.4 Interfaces Inteligentes e Cooperativas**

As interfaces cooperativas ou inteligentes utilizam técnicas de inteligência artificial para auxiliar o usuário em suas tarefas. Segundo Belkin (*apud* Anick, 1997), a interface inteligente deve simular um intermediário humano, devendo possuir conhecimento sobre a terminologia usada, sobre o sistema em questão e sobre técnicas de elicitação.

A finalidade da interface inteligente é ajudar o usuário a alcançar seu objetivo. Por exemplo, o usuário, ao consultar vôos do aeroporto A para o aeroporto B, deve também ser informado de vôos entre outros aeroportos na mesma cidade de A e B.

Conforme Frainer (1993), a interface inteligente deve entender 4 tipos de informação sobre o usuário:

- objetivo do usuário: estado que o usuário quer atingir;
- seu plano: a seqüência de ações ou eventos que levam ao estado desejado; as ações são atos que a interface permite que o usuário efetue;
- habilidades do usuário: físicas e mentais;
- comportamento e preferências: estilo de interação e comodidades.

Os objetivos do usuário podem ser explicitamente declarados ou então inferidos automaticamente por mecanismos de inteligência da interface. Para este último caso, podem ser empregadas técnicas de aprendizado de máquina (*machine learning*) para analisar o comportamento do usuário, o histórico da interação ou de contatos anteriores e as características do ambiente (por exemplo, observando o que foi recuperado, lido, ignorado, gravado, excluído, enviado a outros, etc.).

Geralmente também, as interfaces inteligentes ou cooperativas procuram estabelecer modelos de usuários para poder auxiliá-los. Isto pode ser conseguido descobrindo-se um perfil comum em grupos de usuários.

Também podem ser utilizados diálogos para que a interface entenda a necessidade do usuário. Uma das áreas de estudo da Inteligência Artificial é o processamento da Linguagem Natural. Este processamento pode ser feito a nível

- léxico: analisando os elementos ou unidades da linguagem, por exemplo, termos e palavras;
- sintático: analisando as relações entre os elementos;
- semântico: analisando o significado das frases e construções; ou
- pragmático: analisando o uso da linguagem no contexto e as conotações decorrentes.

As técnicas de processamento de linguagem natural são utilizadas por interfaces cooperativas ou inteligentes para entender entradas fornecidas pelo usuário ou para formular respostas compreensíveis ao usuário. Também servem para eliminar ambigüidades na interação com usuários (quando tentando elicitar a necessidade do usuário). Eliminar ambigüidades significa determinar o sentido ou significado das palavras e frases, uma vez

que estas podem possuir mais de um significado ou podem levar a interpretações errôneas dependendo da construção formada.

### **3.5 Agentes Inteligentes**

Agentes inteligentes são sistemas automatizados (hardware ou software), embutidos de mecanismos de Inteligência Artificial, capazes de tomar decisões e auto-melhorar seu desempenho (Yao et al, 2001). O objetivo é permitir que a inteligência seja distribuída remotamente ou que indivíduos possam tomar decisões de forma autônoma, aumentando assim a eficiência de sistemas computacionais.

Segundo Fernandes (1998), agentes inteligentes devem possuir as seguintes características:

- autonomia: devem trabalhar sem intervenção humana;
- habilidade social: devem saber interagir com humanos ou outros agentes;
- reatividade: devem poder receber estímulos do ambiente e responder em tempo hábil;
- pró-atividade: devem ter comportamento direcionado a um objetivo, tomando a iniciativa da ação sem precisar esperar estímulos;
- mobilidade: devem locomover-se para outros ambientes; e
- continuidade temporal: devem funcionar continuamente.

Na Internet, os agentes inteligentes são chamados de *Intelligent Web Agents* e servem principalmente para explorar serviços na Web e entender regularidades geradas pela Web (Yao et al, 2001).

Quando existem vários agentes inteligentes atuando de forma integrada e cooperativa, o sistema chama-se de sistema de multiagentes. Geralmente, cada agente inteligente possui conhecimentos próprios e diferentes. Este indivíduos interagem entre si, compartilhando informações e conhecimento para soluções de problemas mais complexos, os quais dificilmente seriam resolvidos por qualquer um dos indivíduos de maneira isolada.

### **3.6 Integração Inteligente de Informações**

Uma das áreas da Inteligência Artificial (IA) que vem-se destacando nos últimos tempos e que pode ajudar em muito na Descoberta de Conhecimento é a área que trata da Integração Inteligente de Informações ( $I^3$  ou *Intelligent Integration of Information*).

Dao e Perry (1996) defendem que a  $I^3$  objetiva abstrair, integrar, fundir, reduzir e acrescentar valor a grandes massas de informação distribuída. Wiederhold (1996) comenta a importância desta área ao afirmar que a integração aumenta o valor das informações, principalmente quando estas se originam de múltiplas fontes e são relacionadas ou combinadas.

A integração de informações e conhecimentos, quando feita com apoio automatizado, geralmente ocorre em algum nível simbólico, no qual as informações ou conhecimentos estão representados. O objetivo não é criar um modelo unificado (não modifica os modelos individuais existentes), gerando apenas uma visão unificada para responder a consultas ou para descobrir algo novo.

Em geral, utiliza-se um modelo mínimo de domínio em mais alto nível, sendo as consultas mapeadas para modelos do nível abaixo. Este modelo mínimo deve representar cada esquema individual sem perdas (Arens et al., 1996).

Wiederhold (1996) explica que as fontes não são combinadas, mas somente os resultados selecionados derivados das fontes, a um nível compreensível pelo usuário (combina níveis de baixo em nível mais acima).

Uma maneira muito comum de fazer tal integração é utilizar uma representação canônica que possa representar de um único modo diferentes formalismos.

Um tipo de representação bastante versátil, capaz de representar inúmeros formalismos diferentes, são os grafos. Croft e Turtle (1992) discutem a integração de informações por combinação de grafos canônicos (cada modelo individual e diferente é mapeado para a representação canônica antes de ser feita a comparação).

Wiederhold (1996) identifica alguns elementos neste processo de integração:

- facilitadores: acessam dados nas fontes;
- processadores de consulta: reformulam as consultas;
- mediadores: combinam dados vindos das fontes e os sintetizam ou resumem para o usuário;
- mineradores de dados: procuram encontrar informações interessantes.

Segundo Dao e Perry (1996), mediadores catalogam e produzem informação, com uso de ontologias distribuídas e fragmentadas para compartilhar conhecimento e eliminar ambigüidades. Os conceitos representados na ontologia única (contexto único integrado) são mapeados para esquemas locais. Pode haver hierarquias de esquemas ou ontologias. Para realizar as minerações de dados, os conceitos são integrados por equivalência, generalizações, especializações, incompatibilidades, etc.

Um dos maiores problemas de se ter um esquema único é que alterações nos modelos individuais implicarão em refazer este esquema global único (Arens et al., 1996).

Finin e outros (1998) apresentam o formalismo KIF (*knowledge interchange format*), que se trata de uma linguagem declarativa para troca de conhecimento entre agentes. Esta linguagem é uma forma de representação de ontologias, onde expressões ou sentenças são arbitrariamente descritas no cálculo de predicados. KIF é uma versão da lógica de primeira ordem com extensões para suportar raciocínio e definições não-monotônicos. A finalidade da linguagem KIF é funcionar como um mediador na tradução entre linguagens ou formalismos.

### **3.7 Semântica da Web**

A semântica da Web (*Web Semantics*) procura compreender o significado mais do que o conteúdo presente nela (Yao et al, 2001).

Para tanto, é necessário representar o conhecimento disponível na Web em um formato canônico (que possa ser entendido por todos). Tazi (1994) defende que o conhecimento pode ser representado com os chamados Grafos Conceituais de Sowa (que é um modelo geral para representar conhecimento). Esta abordagem segue a idéia de Aristóteles de que cada conceito é representado por uma palavra ou símbolo e um conjunto de referenciais do mundo. Atuando como uma rede semântica, os nodos representando conceitos são relacionados entre si.

Outra maneira de representar conhecimento é através de Ontologias. Segundo Studer e outros (1998), uma ontologia é um entendimento comum e compartilhado de algum domínio que pode ser comunicado entre pessoas e computadores; é uma especificação formal (deve ser capaz de ser lida e entendida por máquinas) e explícita de uma conceitualização compartilhada (de um grupo e não individual; deve ser um consenso).

Segundo Gruber (1998), ontologias são baseadas no nível do conhecimento (*knowledge level*) de Newell (1982). Este nível descreve o conhecimento independente da representação simbólica interna utilizada.

Wiederhold (1996) vê uma ontologia como um conjunto de termos e relacionamentos usados num domínio, denotando conceitos e objetos. Já Dao e Perry (1996) caracterizam uma ontologia como contendo definições (específicas de um domínio) sobre conceitos, classificações de tipos, hierarquias, etc.

Segundo Gruber (1993), as ontologias fornecem descrições sobre conhecimento. Studer e outros (1998) comentam que uma das finalidades das ontologias é permitir a interoperabilidade de fontes heterogêneas de informação e assim realizar o compartilhamento de conhecimento.

Representações de conhecimento podem ser armazenadas na Web explicitamente sob forma de documentos eletrônicos ou através de tags em XML. Mas também é possível descobrir conhecimento implicitamente armazenado nas relações entre páginas. Chakrabarti (2000) discute o aspecto social da rede (redes sociais), que pode ser entendido pelas relações de co-autoria, orientadores, citações e hiperlinks. O mesmo autor cita o mecanismo de busca Google ([www.google.com](http://www.google.com)), o qual utiliza um índice de popularidade para estabelecer o ranking da páginas resultantes. Neste mecanismo, não basta às páginas ter palavras. A posição no ranking será dada pelas relações de links. São beneficiadas as páginas caracterizadas como autoridades (que recebem muitos links ou são apontadas por várias páginas) e as páginas do tipo hubs (que apontam para várias páginas do tipo autoridade) (Garofalakis et al, 1999).

## **4 Aplicações da WI**

A seguir, serão discutidas algumas aplicações de WI. A lista não é completa, mas dá ênfase a aplicações mais discutidas na literatura.

### **4.1 Marketing e Merchandising**

A WI pode ajudar marketeiros na difícil tarefa de entender quem são os clientes, como se comportam e quais suas preferências.

Técnicas de Data Mining aplicadas sobre bases de dados de clientes ou de transações online (vendas, pedidos, operações bancárias, etc) permitem extrair padrões estatísticos. Por exemplo, pode-se ter uma análise completa das distribuições de valores por atributos de clientes (bairro, cidade, idade, sexo) com a finalidade de entender o perfil do cliente e assim direcionar a propaganda. Ou então pode-se descobrir associações entre produtos adquiridos na mesma compra (“quem compra fraldas, também compra cerveja”). Conhecimento como este pode ser utilizado em campanhas para venda cruzada (“cross-sales”) ou promoções.

Em especial a técnica de clustering permite encontrar grupos de clientes afins, para segmentar o mercado e assim elaborar campanhas diversas para cada segmento. Os clientes

também podem ser segmentados por comportamento (hábitos de compra ou produtos adquiridos). Lawrence e outros (2001) sugerem utilizar técnicas de clustering de mais alto nível, analisando classes de produtos comprados porque muitas vezes as marcas não se repetem mas sim os tipos de produtos.

Técnicas de Web Mining ajudam a analisar o impacto de campanhas de marketing online, permitindo entender como usuários chegam até as compras, ou seja, de onde vem (links de serviços de busca ou banners de propaganda) e que caminho fazem até a compra de um produto (seqüência de páginas vistas). Também serve para melhorar páginas muito visitadas mas que não levam a objetivos e entender o que os clientes fazem após atingir o objetivo.

Técnicas de Data Mining e Web Mining integradas permitem comparar o comportamento de usuários que compraram dos que não compraram e extrair características comuns em usuários que compraram determinado produto ou visitaram determinadas páginas.

Lee & Podlaseck (2001) afirmam que é possível avaliar a efetividade do site, por exemplo estudando onde se perdem clientes durante o processo de compra (em geral, usuário olha, clica no produto, coloca no cesto e finaliza a compra).

Técnicas de Web Mining ajudaram Kohavi e Becher (2001) a entender o perfil de compradores de um conjunto de sites de comércio eletrônico. Por exemplo, descobriram que visitantes que gastam grandes quantias (*heavy purchasers*) são mais velhos, chegam ao site por notícias ou indicação, possuem propriedades e carros de alto valor, visitam áreas específicas do site e repetem a compra 4 vezes ou mais. Também conseguiram descobrir que certos eventos do mundo real aumentam o tráfego em alguns sites (ex: guerras).

Por sua vez, Kohavi (2001) utilizou técnicas de Web Mining para entender o comportamento do usuário de sites de comércio eletrônico. Descobriu que um visitante médio acessa 10 páginas, gasta 5 minutos no site e gasta 35 segundos entre páginas, enquanto que um comprador médio acessa 50 páginas e gasta 30 minutos no site. Também foi possível identificar as taxas de conversão de usuários:

- de visitante para cadastrado: 5%
- de cadastrado para comprador: 33%
- de visitante para comprador: 1,7%.

Outra descoberta com Web Mining foi de que a taxa de abandono de produtos (tirar produto do cesto) é de 34%. Srivastava e outros (2000) sugerem utilizar WI para entender por que visitantes abandonam o site (o que fizeram ou viram imediatamente antes ou durante a sessão).

A área de marketing também pode-se valer das técnicas de WI para melhorar a publicidade e ajudar a vender produtos. A área de merchandising se preocupa com a apresentação de produtos. Schafer e outros (2001) explicam uma ferramenta que pode decidir que banner mostrar baseado em palavras-chave usadas pelo usuário ou em que sub-conjunto da hierarquia de páginas foi visitado.

## 4.2 Suporte ao Usuário

Técnicas de WI podem ajudar na implementação de conversadores ou assistentes digitais, que interagem com o usuário através de linguagem natural (perguntas e respostas de ambos os lados). Estes agentes inteligentes podem servir como vendedores

(recomendando produtos ou serviços), como facilitadores (para busca de informações ou para ajudar usuários a completarem transações) e como mecanismos de ajuda para resolução de problemas (substituindo chamadas ao call center). Um exemplo é o assistente virtual da Sharp ([www.sharp.com.br](http://www.sharp.com.br)) que conversa com o usuário, fornece informações e apresenta produtos da empresa.

### **4.3 Personalização**

Jeff Bezos, CEO da Amazon.com afirmou certa vez: “*se eu tenho 3 milhões de clientes na Web, eu deveria ter 3 milhões de lojas na Web*” (Schafer et al, 2001).

Peppers e Rogers (2000) afirmam que é possível realizar personalização em massa (mass customization) utilizando sistemas automáticos. Schafer e outros (2001) dividem o processo em duas etapas:

- aprendizado (*learning phase*): para analisar dados e construir um modelo de predição do comportamento do usuário;
- uso (*use phase*): para aplicar o modelo a situações diferentes.

O desafio, segundo os mesmos autores, é fazer em tempo real as duas fases.

Srivastava e outros (2000) sugerem criar links dinâmicos em páginas mais freqüentemente vistas juntas. Bruner (2001) afirma que é necessário realizar segmentação dos usuários por conteúdo, tecnologia usada, perfis demográficos (dados pessoais), localização geográfica e por comportamento.

### **4.4 Recomendação**

A recomendação é um tipo especial de personalização que tem como objetivo decidir que produtos ou serviços apresentar para o usuário e não somente modificar a forma de apresentação.

Schafer e outros (2001) admitem que a recomendação pode ser uma sugestão e uma explicação da predição. Os métodos para a recomendação incluem:

- recuperação simples: a partir de uma consulta do usuário;
- filtragem (*pushing* ou *clipping*): enviando por mail, com consentimento do usuário, recomendações segundo uma consulta previamente armazenada e fornecida pelo usuário;
- seleção manual de especialistas (ex: *editor choices*);
- os itens mais vendidos ou mais lucrativos;
- itens associados: descobertos por análise associativa de vendas cruzadas (*cross-sales*);
- filtragem colaborativa: a partir de um item comum, recomendar itens freqüentemente comprados ou vistos por outros clientes;
- relação entre usuários: por exemplo, produtos comprados por pessoas semelhantes (mesmas características, hábitos ou preferências).

O grau de personalização pode variar do modo passivo (ex: banner nas laterais das páginas) para ativo (interromper processo do usuário).

Quando há pouca repetição de itens, Lawrence e outros (2001) sugerem fazer associações não entre produtos, mas entre classes de produtos (tipos ou marcas).

Srivastava e outros (2000) analisam recomendações que oferecem links seguidos por outros usuários com comportamento similar.

#### **4.5 Busca de informações**

A WI é especialmente útil para auxiliar pessoas que buscam informações na Web. Existem métodos para filtragem (*filtering*) de informações ou documentos. Eles consistem em formalizar uma consulta ou perfil de interesse do usuário e então recuperar informação sem que o usuário necessite dar início ao processo. Ou seja, o sistema inteligente notifica o usuário quando alguma informação de interesse é encontrada na Web ou quando algo novo é armazenado na Web e captado pelo sistema.

Uma alternativa é enviar a própria informação (*pushing*), ao invés de apenas avisar o usuário onde ela se encontra, ou então enviar parte dela (*clipping*), selecionando resumos de interesse.

Métodos de filtragem mais complexos podem adaptar os filtros conforme mude o interesse do usuário. Neste caso, métodos automáticos de aprendizagem supervisionada realizam então a modelagem do usuário, para entender seu perfil de interesse sem que o usuário precise manifestá-lo.

A filtragem colaborativa é um caso especial. Ela é utilizada para recomendar conteúdos para um usuário com base em *feedback* fornecido por outros usuários. Este *feedback* pode ser explícito (críticas) ou implícito (documentos lidos, baixados).

Existem também sistemas automáticos extratores de detalhes de informação, conhecidos como *Wrappers* (Mattox et al., 1999; Etzioni, 1996). Estes sistemas de extração de informação analisam páginas da Web à procura de informação específica (nomes, datas, endereços, preços), analisando padrões lingüísticos (estruturas, palavras-chave, relações entre palavras, layout da página). As regras de extração podem ser definidas manualmente por pessoas ou então podem ser utilizados métodos de aprendizado supervisionado.

Perkowitz e outros (1997) apresentam a ferramenta ShopBot, que visita *sites* na Web de lojas e vendedores de CD e Software. A ferramenta extrai informações sobre produtos e *sumariza* resultados para o usuário. Além disto, ela pode fazer compras (preencher campos nas páginas Web) baseada em critérios. A novidade deste trabalho está em que as regras para extração são descobertas automaticamente pela ferramenta com base em análise de *sites* existentes do mesmo tipo (casos de exemplo fornecidos manualmente). A ferramenta então aprende os padrões de nomes próprios, formatos de números, ordem das informações e formato das telas. Uma técnica interessante empregada é preencher campos com nomes de produtos inexistentes e de produtos populares, para avaliar as mensagens e informações sobre os produtos.

#### **4.6 Intermediação de negócios**

Sistemas de WI podem ser utilizados para intermediar transações (negociação e concretização). Hoje em dia existem vários mercados virtuais na Internet (*market places* e sistemas de *e-procurement*), onde empresas realizam negócios entre si (B2B – business-to-business). Um sistema inteligente pode analisar demandas, fazer ofertas, negociar preços e concretizar o negócio.

Também é possível utilizar sistemas inteligentes para comparar ofertas e demandas, por exemplo divulgadas por correio eletrônico, como é o caso do Trade Point. Neste caso, o sistema deve ter conhecimento para cruzar e casar diferentes propostas (*matching*).

#### **4.7 Inteligência Competitiva**

A Inteligência Competitiva (IC - *Competitive Intelligence*) é a área que procura suprir uma empresa com informações estratégicas sobre sua posição no mercado, em relação aos concorrentes e frente a seus clientes (Zanasi, 1998).

As informações importantes para a IC compreendem o mercado em si, os participantes (*players* - empresas concorrentes), suas tendências e estratégias, novas tecnologias de produção, a opinião e a satisfação dos clientes em relação à empresa e às ações da concorrência.

Muitas informações deste tipo estão disponíveis publicamente na Web, podendo ser analisadas por qualquer pessoa de forma lícita. Coletar e analisar essas informações é extremamente importante para que uma empresa adquira um diferencial competitivo.

Existem bases de dados (de patentes, produtos, serviços) disponíveis na Web. Também as páginas Web das empresas descrevem que estratégias estas estão usando para divulgar seu produtos ou serviços. Sites de notícias online também podem ser úteis quando divulgam balanços, fusões, demissões, contratações e ações de empresas concorrentes.

Watts e Porter (1997) utilizam medidas bibliométricas (como número de publicações, citações, patentes, palavras-chave, frequência no tempo, co-ocorrência de termos) para prever inovações tecnológicas.

Outros autores utilizam mecanismos de busca para encontrar termos-chave e a partir de suas frequências na Web estabelecer o grau de disseminação da tecnologia.

#### **4.8 Gerência de redes e sites**

Sistemas inteligentes podem ser utilizados para melhorar o desempenho de redes e sistemas de transmissão de dados. Por exemplo, técnicas de Web mining analisam o tráfego para determinar páginas que serão armazenadas no servidor proxy (*cache*) ou então para baixar páginas frequentemente vistas juntas diretamente e ao mesmo tempo no cliente. Também pode-se utilizar técnicas de mineração para auxiliar na segurança de sistemas de rede, detectando pontos de entradas não-autorizadas ou padrões nos ataques. (Srivastava et al, 2000).

A análise de tráfego na rede também pode ser útil para melhorar o projeto de páginas Web, adicionando links entre páginas relacionadas ou vistas frequentemente juntas (criar atalhos para o usuário).

A mineração também permite descobrir padrões de fraudes em comércio eletrônico ou transações (por exemplo, compras com mesmo cartão ou retirada de extratos bancários, via Web, feitas por domínios bem distantes em tempo curto).

Sistemas de WI também podem auxiliar na detecção de *spams* e realizar a exclusão automática de tais mensagens.

Kohavi (2001) comenta que os próprios sistemas inteligentes podem enganar outros sistemas. Por exemplo, na contagem de page views por sistemas de métricas da Web, não

devem ser levados em conta os agentes inteligentes usados por wrappers e mecanismos de busca (spiders, robots, bots e crawlers).

## **5 Desafios para WI**

A seguir são discutidos alguns pontos dentro da área de WI que merecem melhor atenção ou refinamento.

### **5.1 Adequação a novas tecnologias**

A WI deve adaptar-se às novas tecnologias que estão surgindo. Por exemplo, a comunicação via banda larga pode ser um impulso para métodos melhores que façam análises e dêem respostas em tempo real.

Também a computação móvel, fazendo uso de palmtops, handhelds e telefones celulares, exigirá sistemas com métodos melhores para auxiliar usuários. Entretanto, cada tecnologia exige modos diferentes de interação, o que demanda do sistema inteligente um mecanismo de detecção dos recursos do usuário e a conseqüente adaptação ou personalização.

Também deverá haver um avanço na tecnologia de software cliente, para apoiar melhor o usuário e também permitir identificações mais seguras (por digital, íris, reconhecimento facial, etc). Tal avanço permitirá por exemplo saber por onde o usuário move o mouse ou mesmo para que ponto da tela está olhando.

Um sistema mais avançado que utilizasse a visão computacional poderia analisar expressões faciais e reconhecer se o usuário da Web está triste ou contente.

Com o avanço da multimídia, em breve será possível transmitir pela Web sensações de cheiro, gosto e tato, além das já existentes para visão e audição.

### **5.2 Preparação e seleção de dados para mineração**

Uma instituição bancária na Web descobriu que 5% dos seus clientes tinham nascido na mesma data (1º de janeiro). A razão deste padrão era que os clientes não preenchiam corretamente o cadastro: certos usuários não gostavam de preencher alguns dados obrigatórios, os quais eram armazenados com seus valores default (Kohavi, 2001).

Tal problema evidencia a necessidade de uma melhor preparação para coleta de dados na Web. Dados errados ou inconsistentes podem influenciar um processo de WI e conseqüentemente gerar padrões não verdadeiros.

Outro cuidado que se deve ter é com a seleção dos dados para realizar WI. Por exemplo, para um processo de mineração numa empresa, foram selecionados dez anos de vendas. Descobriu-se um padrão  $X \rightarrow Y$ , significando que quem comprava o produto X também comprava o produto Y, com confiança (probabilidade) de 90%. Mas analisando as vendas ano a ano, verificou-se que, nos 9 primeiros anos, o mesmo padrão tinha confiança de 100% e no último ano o padrão não apareceu (0% de confiança). Se uma campanha de marketing fosse feita em cima deste padrão poderia gerar estratégias distorcidas.

Em outro caso, analisando as vendas numa loja de departamentos, descobriu-se que 81% das mulheres compravam sapatos de mais de 30 reais. Entretanto, uma análise minuciosa por bairro, identificou 9 bairros onde a mesma regra tem confiança de 90% e

somente um bairro onde a regra não aparece. Uma campanha para mulheres continuarem comprando neste nível não deve envolver o último bairro.

Por fim, um terceiro exemplo. Uma farmácia implantou um sistema de comércio eletrônico via Web. Nos primeiros meses, a média das vendas por este sistema foi de 34,2% (figura 1). Entretanto, analisando somente o último mês, pôde-se observar que o sistema novo alcançou 45% das vendas pelo sistema. A análise somente da média poderia levar a uma frustração ou mesmo à decisão de descontinuar o sistema, enquanto que a tendência era de ascendência e portanto sucesso.

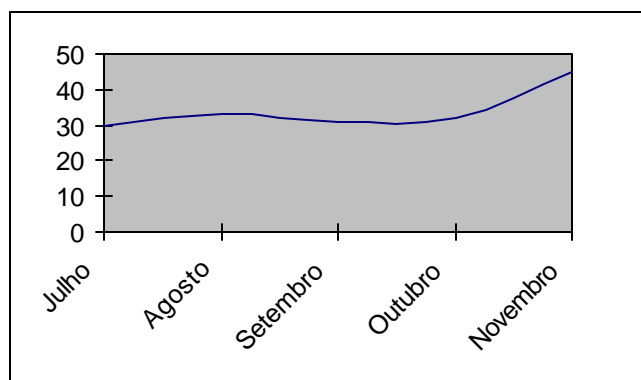


Figura 1: Seleção de dados para mineração

### 5.3 Identificar usuário e seus dados

Um grande desafio na Web é saber identificar o usuário. Pode-se solicitar tais informações, mas nem sempre é de interesse do usuário. Entretanto, respeitados os direitos de privacidade (comentados mais adiante), é importante para as organizações saber quem é o usuário. No caso da Web, a identificação de um visitante pode permitir ao site auxiliar melhor o usuário, através de personalizações ou recomendações, ou mesmo evitar que sejam fornecidas informações que o usuário já conhece.

Alguns autores sugerem o uso de cookies ou a identificação pelo número IP da máquina. Entretanto, pode ocorrer casos de uma mesma máquina ser utilizada por várias pessoas ou a mesma pessoa utilizar máquinas diferentes, e outras variações do gênero (Srivastava et al, 2000).

As técnicas de WI poderiam ajudar a identificar o usuário por seu comportamento (respeitando o direito à privacidade). Por exemplo, pode-se observar se as mesmas páginas são vistas no início do dia (ex: notícias e mail), se o usuário chega no site sempre pelo mesmo caminho (ex: por portal horizontal, pois é o que ele conhece), se uma mesma seqüência de páginas iniciais é seguida à risca (usuário experiente ou leigo) ou se ele inicia em páginas no meio do site (usuário que conhece os atalhos).

Além de ser capaz de coletar dados sobre o usuário, o sistema inteligente deve manter somente dados confiáveis, o que muitas vezes não acontece por erro do usuário ou por intenção. Algumas informações podem ser validadas com dados coletados em ações de tijolo-e-cimento (por exemplo, entrega de um prêmio ou produto em certo endereço).

#### **5.4 Descoberta de fenômenos (Phenomenal Data Mining)**

McCarthy (2000) afirma que o processo de mineração é feito sobre dados que representam a realidade. Neste caso, os padrões descobertos dizem respeito aos dados e não necessariamente (mas por inferência) relatam sobre coisas do mundo real (fenômenos). Em geral, o processo de ligar fenômenos (eventos reais) e dados (representações e observações dos eventos) é feito manualmente.

A WI pode auxiliar neste processo. Observando com cuidado os dados, é possível inferir entidades no mundo real, suas características e inferir relações entre entidades do mundo real. A observação deve funcionar como a criptoanálise, que identifica, por exemplo, padrões no comportamento de quem envia mensagens.

McCarthy (2000) cita o caso de um supermercado que deseja identificar seus clientes (que compras foram feitas pelo mesmo cliente). Neste caso, é necessário observar compras parcialmente comuns (produtos incomuns, marcas, grupos de produtos) e comportamentos semelhantes (datas especiais, dias da semana, horários). As pequenas variações nas regularidades das compras podem ser úteis.

Também é possível descobrir certas características dos clientes, como por exemplo se tem microondas ou freezer. A análise pode ser feita sobre categorias de produtos (ex: alimentação), sobre produtos (ex: chocolates em barra) ou por marcas (ex: chocolate Garoto).

Na Web, a descoberta de fenômenos pode auxiliar a identificar usuário (visitas feitas pelo mesmo usuário) com o objetivo de personalização ou recomendação (para não mostrar mesmas coisas, por exemplo).

Um caso interessante ocorreu com as vendas de laranja num supermercado e exemplifica bem a necessidade da mineração fenomenal. A figura 2 apresenta os níveis de venda deste produto durante alguns meses. No início havia o fornecedor A, mas em outubro houve a troca de fornecedor (para B). O evento que suscitou interesse foi a baixa nas vendas de laranja no mês seguinte (novembro). Entretanto, até hoje o supermercado ainda não identificou a causa, pois poderia ser:

- a) que os clientes não gostaram da volta das laranjas do fornecedor A e preferiam a do B;
- b) que os clientes não gostaram das laranjas de B mas refletiram sua insatisfação nas vendas no mês seguinte, mesmo não sabendo da volta do fornecedor A;
- c) que houve algum fenômeno externo, independente da qualidade das laranjas de A ou de B, que influenciou nas vendas (algum concorrente abriu filial próxima, por exemplo); ou
- d) que havia um problema de sazonalidade, ou seja, em geral no mês de novembro as vendas caem (seria necessário analisar os dados dos anos anteriores).

A mineração fenomenal exige captar e entender os eventos do mundo real e sua ligação com os padrões verificados nos dados. Isto é extremamente útil em situações onde acontecimentos no mundo (guerras, lançamento de livros) ou condições climáticas (chuvas, secas, calor, tempestades) influenciam o comportamento das pessoas ou das organizações.

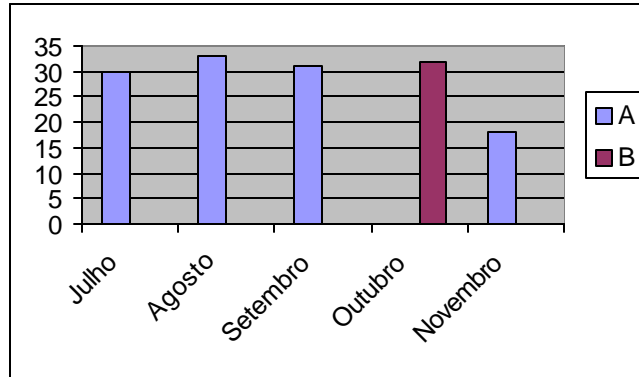


Figura 2: Mineração de Fenômenos

### 5.5 Integração dos vários tipos de mineração

A WI utiliza várias técnicas e métodos de mineração (data mining, web mining, text mining). O melhor desempenho é atingido quando estas abordagens são utilizadas em conjunto. Por exemplo, Web mining analisa o comportamento dos usuários de um site pelo log armazenado no servidor. Seria importante montar um banco de dados com estas informações e ainda acrescentar dados sobre o conteúdo das páginas (text mining). Uma técnica simples seria substituir as URL's por palavras, temas ou conceitos presentes nas páginas Web.

Além disto, há informações sobre comportamento que não estão no log, por exemplo, compras e transações bancárias feitas pelo usuário, que aparecem em bancos de dados corporativos. Pode-se ainda acrescentar informações relativas à semântica da Web (links semânticos e comportamento social).

Kimball e Merz (2000) sugerem criar um Data Webhouse, termo análogo a Data Warehouse. O objetivo é manter todos os dados necessários à inteligência do negócio reunidos em uma mesma base e de forma integrada para facilitar a mineração.

### 5.6 Mineração de conteúdo

O próximo passo além da integração das técnicas de mineração é realizar a mineração de conteúdo da Web. Srivastava e outros (2000) sugerem fazer classificação ou clustering de páginas Web ou partes destas (textos, imagens), identificando o assunto ou tipo. Isto permitiria encontrar páginas ou documentos de conteúdo relacionado para apoiar comunidades virtuais. Também é possível realizar clustering de usuários com comportamento similar, para entender padrões na comunidade virtual.

Spiliopoulou e outros (2000) complementam sugerindo generalizar conteúdos de sites para conceitos mais gerais (tipos de produtos, descritores de documentos, parâmetros de consulta). Tal abordagem permitiu a Chakrabarti (2000) descobrir que páginas sobre ciclismo tinham mais links para páginas sobre primeiros socorros do que usualmente (mais que o comum).

A análise de conteúdo permite identificar certos interesses sociais. Por exemplo, as palavras mais usadas em buscas podem indicar os temas mais vistos ou de maior interesse (ver termos mais usados para busca em 50.lycos.com).

Pode-se também realizar análise de conceitos presentes em textos. Loh e outros (2000) conseguiram identificar variações e associações entre temas apresentados em um jornal online sobre determinado político.

Técnicas de WI poderiam ser embutidas em assistentes digitais para dialogar com usuário. Tais assistentes podem ajudar usuários em processos de busca de informações ou operações ou mesmo respondendo perguntas.

Um desafio ainda existente é conseguir gerar uma ontologia ou thesaurus de forma automática. Ontologias e thesauri formam vocabulários controlados ou linguagens para representar conceitos da realidade.

Outra possibilidade é a geração automática de links entre textos (documentos ou partes) ou páginas na Web, como fazem Swanson e Smalheiser (1997).

### **5.7 Descoberta de intenções**

Uma extensão da mineração de conteúdos é a descoberta de intenções dos usuários Web. Srivastava e outros (2000) acreditam ser possível entender as intenções do usuário pela classificação do conteúdo visto.

Interfaces inteligentes e cooperativas podem ajudar nesta tarefa descobrindo objetivos de usuários através de suas ações. As técnicas de Text Mining e Web mining podem ser empregadas para descobrir o conteúdo que o usuário está procurando.

Wiebe (1994) apresenta estudos sobre descoberta de crenças e intenções em diálogos, por inferências sobre palavras-chave (“tags”).

### **5.8 Privacidade**

Um fator que pode inibir o avanço da WI é a questão da privacidade. Hoje em dia é possível descobrir informações sobre pessoas sem que estas forneçam. Estas informações são importantes para as empresas. Entretanto, as ações de coleta e uso destes dados precisam ser regidas por direitos e deveres.

Muitas pessoas não querem que seus dados sejam coletados (preferem o anonimato). Já outras admitem isto mas exigem que os dados não sejam divulgados para outros ou usados para propaganda (tele ou mail-marketing, por exemplo).

Schafer e outros (2001) sugerem que as empresas e sites divulguem explicitamente suas políticas de privacidade (que informação está sendo coletada e para que tipo de uso). Há também empresas de certificação (como a [www.truste.com](http://www.truste.com)) que anexam selos de certificação conforme as diferentes políticas empregadas.

Está em discussão no WWW Consortium (W3C) o protocolo P3P (Platform for Privacy Preferences) que permitirá a negociação automática entre empresas e clientes. Segundo Srivastava e outros (2000), este protocolo permite aos sites publicarem suas políticas em formatos capazes de serem lidos e entendidos por outras máquinas. Desta forma, o browser cliente pode ler e comparar estas políticas com as configurações de segurança do usuário.

## 6 Conclusão

A Web é caótica e dinâmica (Etzioni, 1996). Sua estrutura e linguagem são as mais variadas possíveis. Tal complexidade e a falta de organização formal levam inevitavelmente à necessidade de mecanismos automáticos e inteligentes que ajudem pessoas a diminuir a sobrecarga de informações (information overload) e a se comunicarem de forma mais eficaz e eficiente.

A área de Web Intelligence (WI) está impulsionando o desenvolvimento de tecnologias que tenham estas finalidades. Este tutorial apresentou algumas destas iniciativas e discutiu os primeiros resultados sendo colhidos. Entretanto, como discutido na parte final deste artigo, há ainda muito a ser feito para transformar a Web em uma rede inteligente.

## 7 Bibliografia

- ANICK, Peter G. (1997). Exploiting clustering and phrases for context-based information retrieval. ACM SIGIR Conference on Research and Development in Information Retrieval. Philadelphia, 1997.
- APTÉ, Chidanand et al. (1994). Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, v.12, n.3, Julho de 1994.
- ARENS, Yigal et al. (1996). Query reformulation for dynamic information integration. Journal of Intelligent Information Systems, v.6, n.2/3, Junho de 1996.
- AGRAWAL, Rakesh. (1995). Data mining: the quest perspective. EDBT Summer School on Advances in Database Technology, Gubbio-Itália, Setembro de 1995. [www.almaden.ibm.com/cs/quest](http://www.almaden.ibm.com/cs/quest)
- BERENDT, Bettina. (2001). Understanding Web usage at different levels of abstraction: coarsening and visualising sequences. Workshop de Data Mining na Web (WEBKDD), 2001.
- BHARAT, Krishna & BRODER, Andrei. (1998). A technique for measuring the relative size and overlap of public Web search engines. 7. Conf. on World Wide Web, 1998. [www7.scu.edu.au/programme/fullpapers/1937/com1937.htm](http://www7.scu.edu.au/programme/fullpapers/1937/com1937.htm)
- BRUNER, Rick E. et al. (2001). Marketing on-line – estratégias, melhores práticas e estudos de casos. São Paulo: Futura, 2001.
- Cadê/IBOPE. (1999). 4ª Pesquisa, 1999. No site [www.ibope.com.br](http://www.ibope.com.br)
- CHAKRABARTI, Soumen. (2000). Data mining for hypertext: a tutorial survey. ACM SIGKDD Explorations, v.1, n.2, Janeiro de 2000. [www.acm.org/sigkdd/explorations](http://www.acm.org/sigkdd/explorations)
- CHEN, Hsinchun. (1994). The vocabulary problem in collaboration. IEEE Computer, v. 27, n. 5, Maio de 1994. <http://ai.bpa.arizona.edu/papers/cscw94/cscw94.html>
- COHEN, William W. & SINGER, Yoram. (1996). Context-sensitive learning methods for text categorization. ACM-SIGIR Conference on Research and Development in Information Retrieval, Zurich, 1996. [www.research.att.com/~wcohen/index.html](http://www.research.att.com/~wcohen/index.html)
- COWIE, Jim & LEHNERT, Wendy. (1996). Information extraction. Communications of the ACM, v.39, n.1, Janeiro de 1996.
- CROFT, W. Bruce & TURTLE, Howard R. (1992). Text retrieval and inference. In JACOBS, Paul S. (ed.). Text-based intelligent systems: current research and practice in information extraction and retrieval. New Jersey: Lawrence Erlbaum, 1992.

- CROFT, W. Bruce. (1995). Machine learning and information retrieval. COLT Conference, Lake Tahoe, Julho de 1995. (palestra convidada)  
<http://www.ee.umd.edu/medlab/filter/>
- DAO, Son & PERRY, Brad. (1996). Information mediation in cyberspace: scalable methods for declarative information networks. **Journal of Intelligent Information Systems**, v.6, n.2/3, Junho de 1996.
- ETZIONI, Oren. (1996). The world-wide web: quagmire or gold mine ? Communications of the ACM, v.39, n.11, p.65-68, Novembro de 1996.
- FAYYAD, Usama M. et al. (1996). From data mining to knowledge discovery: an overview. In: FAYYAD, Usama M. et al. (eds) *Advances in Knowledge Discovery and Data Mining*. Menlo Park, The MIT Press, 1996.
- FELDMAN, Ronen & DAGAN, Ido. (1995). Knowledge discovery in textual databases (KDT). International Conference on Knowledge Discovery, Montreal, 1995.
- FERNANDES, Jorge H. C. (1998). Ciberespaço: modelos, tecnologias, aplicações e perspectivas. In: De MOURA, Hermano P. (ed). XVII Jornada de Atualização em Informática - JAI, vol. 2, 1998.
- FININ, Tom et al. (1998). KIF101: a brief introduction to the knowledge interchange format. [www.cs.umbc.edu/kse/kif/kif101.shtml](http://www.cs.umbc.edu/kse/kif/kif101.shtml)
- FRAINER, Antônio S. (1993). Planos na interação homem-máquina. CPGCC/UFRGS, Porto Alegre, Junho de 1993. (dissertação de mestrado)
- GAROFALAKIS, Minos N. et al. (1999). Data mining and the web: past, present and future. ACM Workshop on Web Information and Data Management, Kansas City, 1999.
- GOEBEL, Michael and GRUENWALD, Le. (1999). A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations, v.1, n.1, June 1999.
- GRUBER, Tom R. (1993). A translation approach to portable ontologies. Knowledge Acquisition, v.5, n.2, 1993.
- HAN, Jiawei et al. (1996). Intelligent query answering by knowledge discovery techniques. IEEE Transactions on Knowledge and Data Engineering, v.8, n.3, Junho de 1996.
- HERSH, William R. et al. (1995). Towards new measures of information retrieval evaluation. ACM-SIGIR Conference on Research and Development in Information Retrieval, 1995.
- INGARGIOLA, G. (1996). Building classification models: ID3 and C4.5. <http://www.cis.temple.edu/~ingargiola/cis587/readings/id3-c45.html>.
- KIMBALL, Ralph & MERZ, Richard. (2000). Data Webhouse: construindo o data warehouse para a WEB. Rio de Janeiro: Campus, 2000.
- KOHAVI, Ron & BECHER, Jon. (2001). E-commerce and clickstream mining tutorial. SIAM International Conference on Data Mining. Abril de 2001.
- KOHAVI, Ron. (2001). Mining e-commerce data: the good, the bad and the ugly. Simpósio Internacional de Gestão de Conhecimento e Gestão de Documentos. Curitiba, agosto de 2001. (palestra convidada)
- LEE, J. & PODLASECK, Mark. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. Journal of Data Mining and Knowledge Discovery, v.5, n.1/2, Janeiro de 2001.

- LIDDY, Elizabeth D. et al. (1994). Text categorization for multiple users based on semantic features from a machine-readable dictionary. *ACM Transactions on Information Systems*, v.12, n.3, Julho de 1994.
- LOH, Stanley. (1999). Descoberta de conhecimento em textos. Exame de Qualificação EQ-29. PPGC/UFRGS, Porto Alegre, Fevereiro de 1999. [atlas.ucpel.tche.br/~loh](http://atlas.ucpel.tche.br/~loh)
- LOH, Stanley. (2000). Concept-based knowledge discovery in texts extracted from the web. *ACM SIGKDD Explorations* v.2, n.1, Julho de 2000. [www.acm.org/sigkdd/explorations](http://www.acm.org/sigkdd/explorations)
- MAAREK, Yoëlle S. (1992). Automatically constructing simple help systems from natural language documentation. In JACOBS, Paul S. (ed.). *Text-based intelligent systems: current research and practice in information extraction and retrieval*. New Jersey: Lawrence Erlbaum, 1992.
- MATTOX, David et al. (1999). Rapper: a wrapper generator with linguistic knowledge. *ACM Workshop on Web Information and Data Management*, Kansas City, 1999.
- McCARTHY, John. (2000). Phenomenal data mining: from data to phenomena. *ACM SIGKDD Explorations*, v.1, n.2, Janeiro de 2000. [www.acm.org/sigkdd/explorations](http://www.acm.org/sigkdd/explorations)
- McKEOWN, Kathleen & RADEV, Dragomir R. (1995). Generating summaries of multiple news articles. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, Seattle, 1995.
- MIIKE, Seiji et al. (1994). A full-text retrieval system with a dynamic abstract generation function. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1994.
- MOSCAROLA, Jean et al. (1998). Technology watch via textual data analysis. *Note de Recherche n° 98-14*, Université de Savoie. Julho de 1998.
- NEWELL, Allen. (1992). The knowledge level. *Artificial Intelligence*, v.18, n.1, Janeiro de 1982.
- PEPPERS, Don & ROGERS, Martha. (2000). Nos conhecemos de algum lugar ? *HSM Management*, n.19, março/abril de 2000.
- PERKOWITZ, Mike et al. (1997). Learning to understand information on the Internet: an example-based approach. *Journal of Intelligent Information Systems*, v.8, 1997.
- RILOFF, Ellen & LEHNERT, Wendy. (1994). Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, v.12, n.3, Julho de 1994.
- SCHAFER, J. Ben et al. (2001). E-commerce recommendation applications. *Journal of Data Mining and Knowledge Discovery*, v.5, n.1/2, Janeiro de 2001.
- SPARCK-JONES, Karen & WILLET, Peter (eds). (1997). *Readings in Information Retrieval*. San Francisco: Morgan Kaufmann, 1997.
- SPILIOPOULOU, Myra & POHLE, Carsten. (2001). Data mining to measure and improve the success of Web sites. *Journal of Data Mining and Knowledge Discovery*, v.5, n.1/2, Janeiro de 2001.
- SPILIOPOULOU, Myra et al. (2000). Improving the effectiveness of a Web site with Web usage mining. *Workshop de Data Mining na Web (WEBKDDO. Lecture Notes on Artificial Intelligence 1836*, Springer-Verlag, Julho de 2000.
- SRIVASTAVA, Jaideep et al (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations*, v.1, n.2, Janeiro de 2000. [www.acm.org/sigkdd/explorations](http://www.acm.org/sigkdd/explorations)

- STUDER, Rudi et al. (1998). Knowledge engineering: principles and methods. *Data & Knowledge Engineering*, v.25, n.1/2, Março de 1998.
- SWANSON, Don R.; SMALHEISER, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, v.91, n.2, Abril de 1997.
- TAN, Ah-Hwee. (1999). Text mining: the state of the art and the challenges. *Pacific-Asia Workshop on Knowledge Discovery from Advanced Databaseses*, Beijing, 1999. *Lecture Notes in Computer Science* v.1574, Springer-Verlag, 1999.  
<http://textmining.krdl.org.sg/publications.html>
- TAZI, Saï d. (1994). Using Sowa's conceptual graphs for enhancing hypertext readers performances. *Intelligent Hypertext Workshop*, Washington, 1994. <http://lis.univ-tlse1.fr/tazi>
- UPCHURCH, Linda et al. (2001). Using card sorts to elicit web page quality attributes. *IEEE Software*, v.18, n.4, Julho/Agosto de 2001.
- WATTS, Robert J. & PORTER, Alan L. (1997). Innovation forecasting. *Technological Forecasting and Social Change*, v.56, 1997.
- WIEBE, Janyce M. (1994). Tracking point of view in narrative. *Computational Linguistics*, v.20, n.2, Junho de 1994.
- WIEDERHOLD, Gio. (1996). Foreword: intelligent integration of information. *Journal of Intelligent Information Systems*, v.6, n.2/3, Junho de 1996.
- WILLET, Peter. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management*, v.24, n.5, 1988.
- YANG, Yiming & CHUTE, Christopher G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems*, v.12, n.3, Julho de 1994.
- YANG, Yiming & PEDERSEN, Jan O. (1997). A comparative study on feature selection in text categorization. *International Conference on Machine Learning*, Nashville, 1997.
- YANG, Yiming & LIU, Xin. (1999). A re-examination of text categorization methods. *ACM-SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, 1999.
- YAO, Y.Y. et al. (2001). Web Intelligence (WI): research challenges and trends in the new information age. In Zhong, N., Yao, Y.Y., Liu, J., and Ohsuga, S. (eds.). *Web Intelligence: Research and Development. Lecture Notes on Artificial Intelligence* 2198, Springer-Verlag, 2001. [kis.maebashi-it.ac.jp/wi01/ps/wi-intro.ps](http://kis.maebashi-it.ac.jp/wi01/ps/wi-intro.ps)
- ZANASI, Alessandro. (1998). Competitive Intelligence though datamining public sources. *Competitive Intelligence Review*, v.9, n.1, 1998.
- ZHONG, N. et al. (2000). Web Intelligence (WI). *Proceedings 24th IEEE International Computer, Software and Applications Conference (COMPSAC)*, 2000.  
[kis.maebashi-it.ac.jp/wi01/ps/wi-ieee.ps](http://kis.maebashi-it.ac.jp/wi01/ps/wi-ieee.ps)