MINISTÉRIO DA DEFESA NACIONAL

INSTITUTO DE ESTUDOS SUPERIORES MILITARES

CURSO DE PROMOÇÃO A OFICIAL GENERAL
2005 / 2006

EMD N.º 10 – ESTUDO DE MÉDIA DURAÇÃO

DOS DADOS AO CONHECIMENTO: A IMPORTÂNCIA DO *DATA MINING*

CMG ECN RAPAZ LÉRIAS

18ABR2006

RESUMO

O trabalho procura abordar a temática do *Data Mining*, designado genericamente por processo de descoberta de informações relevantes tais como, padrões, associações, mudanças, anomalias e estruturas, existentes em grandes quantidades de dados, armazenados nos bancos de dados, nos depósitos de dados, nas bases de dados ou noutros repositórios de informação.

A aproximação ao tema é feita tendo por base uma pesquisa bibliográfica da qual se extraiu diversa matéria de cariz conceptual, que se tentou organizar e estruturar de modo a construir um documento genérico capaz de despertar o interesse daqueles que pouco ou nada sabem sobre o assunto. Embora se traga à colação algumas evidências da importância do *Data Mining* e da sua aplicação, não foi possível reunir as condições suficientes para enveredar por uma investigação aplicada, nem tampouco pela procura fundamentada de conhecimento novo sobre o tema em causa.

O *Data Mining* enquadra-se num processo mais alargado de aquisição do conhecimento que visa a sua obtenção a partir do sucessivo tratamento de conteúdos básicos de informação que são os dados em bruto.

Tendo por base os conceitos relevantes de informação, conhecimento e inteligência, associados ao *Data Mining*, o trabalho procura também identificar e descrever a aplicabilidade prática desta tecnologia da informação em áreas actuais e relevantes de actividade, entre as quais se contam “a gestão de relações com o cliente”, “a compilação automática de recursos da *Web*” e ainda “a segurança e o exercício do contra-terrorismo”.

ÍNDICE

Nº. Pág.

1. INTRODUÇÃO

O tema sobre o qual incide o presente documento enquadra-se numa vasta área científica que se distribui não só pelo campo das ciências sociais e humanas, mas também pela esfera das ciências exactas, puras e aplicadas.

Por si só, o título – dos dados ao conhecimento: a importância do *Data Mining* – abre um rol de pistas e de oportunidades ao estudo e à investigação e coloca, à partida, um problema sério de delimitação do trabalho, que seja minimamente compatível com os propósitos, com a extensão e com o tempo disponível estipulados na publicação de referência para a sua execução. Por outro lado, logo ressalta a dificuldade de orientar o estudo de modo a privilegiar as vertentes com interesse para a Defesa Nacional e, em particular, para as Forças Armadas e para a Marinha, suscitando dúvidas quanto à amplitude e incidência da pesquisa e quanto à opção a tomar em termos do binómio extensão/minúcia do trabalho. Em último lugar, é conveniente sublinhar tratar-se de uma acção académica não acompanhada, nem supervisionada, sobre uma matéria que pouco se prende com a formação do autor e para a qual não dispõe, nem de anteriores conhecimentos básicos, nem de experiência de utilização prática, acção essa elaborada em paralelo com a frequência das várias áreas curriculares do curso em que se insere.

Explicado o enquadramento temático e circunstancial do trabalho, optou-se por tratar o assunto de forma relativamente rudimentar, recorrendo à pesquisa bibliográfica como veículo de obtenção dos elementos que permitissem cobrir os tópicos designados no título do trabalho, focalizando a atenção na observação do processo de *Data Mining*, mas sem descurar, de todo, a sua utilidade em termos de aplicação a diversas áreas de actividade. Para não alimentar as expectativas dos incautos ou dos mais generosos, constitui obrigação do autor clarificar, desde já, que se limitou a coligir o conhecimento desenvolvido por especialistas na matéria, tentando apresentá-lo de forma estruturada, equilibrada, sistematizada e lógica, numa monografia que não passa de abordagem simples, despretensiosa e muito limitada sobre o tema.

Para além desta introdução e das conclusões, o documento contém um corpo principal, no qual o assunto é desenvolvido segundo duas vertentes distintas. Na primeira faz-se uma aproximação aos conceitos relevantes de informação, conhecimento e inteligência, numa perspectiva das ciências puras, para depois

dedicar uma atenção substancial ao conceito de *Data Mining* e às técnicas e metodologias que o suportam, nomeadamente aos tipos de modelos e de algoritmos aplicáveis, isto sem perder de vista o processo global de aquisição de conhecimento a partir de bases de dados, no qual o próprio *Data Mining* tem de estar inserido. Na segunda vertente explora-se conceptualmente a utilidade (e menos a aplicação) do *Data Mining* em três áreas distintas de actividade profissional, que se entendeu serem aquelas que melhor ilustram o potencial desta tecnologia; em primeiro lugar, numa área actualmente muito difundida no mundo empresarial, que é a de gestão das relações com o cliente; em segundo lugar, na área do *Web Mining*, como ferramenta de suporte dos processos automatizados de recolha e de exploração de recursos na *Web*[1]; e, por fim, mais num plano de suporte argumentativo, como instrumento a ter em consideração no exercício de actividades ligadas à segurança, designadamente na de contra-terrorismo.

De uma forma perceptível e muito simples é inevitável caracterizar o *Data Mining* como uma tecnologia de informação que permite prever o futuro, afinal, um dos desideratos que mais desperta interesse e curiosidade no ser humano e cujo carácter premonitório confere a ilusão de ser possível alterar o futuro e o destino das coisas. Na realidade, há algum fundamento de verdade nesta asserção, na medida em que acções e atitudes, que decorrem do processo de adivinhação que o *Data Mining* consubstancia e que, de outro modo, não seriam tomadas, vão condicionar o comportamento futuro de pessoas, de instituições, de empresas, assim como a aquisição do conhecimento e a sua divulgação.

Nas empresas, o *Data Mining* é um processo que está a ser usado simultaneamente para permitir reduzir custos e aumentar as receitas. O potencial deste processo é enorme, estando a ser mundialmente utilizado pelas empresas mais inovadoras com o objectivo de localizar clientes de elevado valor comercial ou para reconfigurar os seus produtos de modo a conseguir aumentar o volume de vendas, seja através de maior procura, seja de procura mais valiosa e focalizada.

Na pesquisa criteriosa da *Web*, o *Data Mining* está a ser utilizado como técnica de compilação automática dos recursos aí existentes, suportando o ajustamento automático dos modelos de consulta, através de mecanismos de aprendizagem inerentes aos próprios modelos.

[1] *World Wide Web Consortium*. http://www.w3c.org.

Na segurança, o *Data Mining* pode ajudar a prioritizar os esforços e a fornecer pistas sobre os elementos de informação em que é preciso focar a atenção, libertando os analistas e os investigadores para se dedicarem ao tipo de análise que requer inevitavelmente a capacidade crítica do ser humano. Trata-se essencialmente de detectar ligações, padrões e erros em quantidades astronómicas de dados, que o ser humano nunca conseguirá encontrar sem a ajuda dos computadores e das aplicações informáticas baseadas nesta e noutras tecnologias.

2. DOS DADOS E DA INFORMAÇÃO AO CONHECIMENTO

2.1 DADOS, INFORMAÇÃO, CONHECIMENTO E INTELIGÊNCIA

Shannon (1948) define uma medida do conteúdo informacional de um evento como sendo a probabilidade de esse evento ocorrer. Esta definição segue a noção intuitiva de que quanto mais improvável for um acontecimento, mais se tem a aprender com ele. Por outras palavras, a frase o "sol nasceu esta manhã" tem muito menos conteúdo informacional que a frase "um boeing 767 caiu esta manhã" (Navega, 2002). A formulação matemática mais simples dessa definição é a que se segue, onde I é a medida da informação e p_i é a probabilidade de ocorrência do "inésimo" evento: $I = - \log (p_i)$

Através desta noção, pode-se definir informação a partir de um qualquer sinal sem ter de saber propriamente de que trata esse sinal. Na verdade, esse seria o primeiro tipo de análise que se precisaria fazer caso se recebesse na terra uma mensagem proveniente de uma civilização extraterrestre. Como nada se sabe sobre essa eventual civilização, a principal referência seria iniciar uma análise da probabilidade de cada evento discreto que compõe esse sinal, em busca de padrões e de sequências. No caso de sinais complexos, com padrões e sequências de padrões, requer-se a presença de um organismo inteligente capaz de acumular em memória uma série de segmentos de informação, a partir dos quais consiga estimar a probabilidade de ocorrências futuras. É precisamente na sequência da figura do agente inteligente que se parte para a definição dos conceitos de informação, de agente, de conhecimento e de inteligência.

- Informação [2] – é um conjunto de descrições simbólicas de mudanças de estado de um sistema; o conteúdo informacional duma mensagem qualquer é dado pela avaliação da probabilidade de ocorrência dos símbolos que compõem a mensagem, nos termos definidos por Shannon (1948);
- Agente – é um sistema (organismo ou máquina) que pode trocar informação com o meio ambiente que o cerca e que tem estados internos que se alteram com o tempo; algumas dessas alterações internas de estado devem ser função da troca de informações que o agente executa com o ambiente que o circunda; exemplos típicos são os gatos, os elefantes, os seres humanos, mas também, surpreendentemente, os computadores, os leitores de discos compactos e mecanismos de busca [3] da *Web*, etc;
- Conhecimento - é um conjunto de informações situadas no interior de um agente que o habilitam a actuar no meio ambiente com maior eficácia do que se esse agente não dispusesse dessa informação; pode dizer-se que um rato tem conhecimento sobre ratoeiras na medida em que disponha de uma série de informações experimentais que lhe permitam evitar a ratoeira;
- Inteligência – é a habilidade (ou medida de uma habilidade) de um agente para gerar (criar) conhecimento; quando não há formação de novo conhecimento não se pode falar de inteligência.

2.1.1 Inteligência Artificial

A expressão Inteligência Artificial (IA) remonta à passada década de 50, tendo dado os seus primeiros passos com a implementação de um sistema primitivo apoiado em máquinas para a resolução de problemas, simulando os métodos humanos - o desenvolvimento mais significativo foi o *General Problem Solver* [4].

Durante a década de 80, surgiu outra variante da IA que ficou conhecida pelos sistemas inteligentes ou especialistas, em que um certo número de dados de um domínio era codificado nas chamadas regras de produção. Foi também durante esta década que as pesquisas em redes neuronais floresceram, tendo-se produzido

[2] Apesar de existir diferença entre dado e informação, entendeu-se poder usar de modo indiferenciado os dois termos no âmbito deste trabalho. Todavia, é bom precisar que um dado é uma sequência de símbolos e um ente totalmente sintáctico que não envolve semântica como na informação. Os dados podem ser representados com sons, imagens, textos, números e estruturas.

[3] Tradução livre do autor relativa a *web browser*.

[4] Criado por Allen Newel e Herbert Simon. Cf. Russel (1995).

algumas aplicações práticas de relativo sucesso comercial, entre elas algumas aplicações de *Data Mining*.

Todavia, faltava algo que pudesse dispor de uma base de conhecimento mais fundamental e mais alargada que se sobrepusesse às áreas dos domínios específicos, em suma, algo que dispusesse de senso comum. Trata-se de uma questão cuja resolução se encontra ainda um pouco no domínio da ficção.

Uma das maiores dificuldades da disciplina de IA tem sido produzir computadores que "raciocinem" com o chamado senso comum. O senso comum requer um volume de conhecimentos, interligado, sobre o mundo real, que permita o desempenho de funções de forma adequada e de acordo com o esperado.

Qualquer pessoa sabe que para entrar numa sala é necessário abrir antes a porta. Porém, isso não é óbvio para os computadores. É esta distinção de obviedade que causa intriga acerca da capacidade das máquinas.

2.2 PROCESSO DE *DATA MINING*

O *Data Mining* é uma das novidades da Ciência da Computação que veio para ficar. Com a acumulação de um volume cada vez maior de informação é essencial conseguir retirar algum proveito da crescente capacidade tecnológica para manter e manusear essa informação.

Hoje em dia, as bases de dados podem atingir uma capacidade superior a um *terabytes* – mais que 1,000,000,000,000 de *bytes* de dados. No seio de um conjunto tão alargado de dados pode jazer escondida informação que tenha importância estratégica. Mas como descobri-la nessa floresta de dados? A resposta reside precisamente no *Data Mining*, um processo que utiliza uma variedade de ferramentas de análise para detectar padrões e relações que possam, à posteriori, ser usados para fazer previsões válidas de comportamentos.

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

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

Na verdade, o *Data Mining* é apenas uma fase (talvez a mais importante) de um processo mais longo e abrangente que vai desde a informação (os dados - bases de dados) até ao conhecimento propriamente dito, conforme se ilustra nas figuras nº 1 e

nº 2.

Para efeitos do presente trabalho, optou-se por apresentar apenas os passos fundamentais desse processo [5]. Tendo como ponto de partida as fontes de dados (bancos de dados, bases de dados, tabelas, relatórios, transacções, etc) procede-se ao seu tratamento preliminar, que consiste numa operação de "limpeza" em que basicamente se eliminam inconsistências, redundâncias e ruído e se acrescentam informações em falta. Estas operações transformam as fontes originais de dados nos chamados repositórios organizados de dados, conhecidos por *Data Marts* ou *Data Warehouses*, os quais podem já ser úteis de diversas formas. É a partir destes repositórios que se extraem alguns conjuntos de registos para os sujeitar ao processo de *Data Mining*. Este processo pode funcionar de modo interactivo e usar interfaces de visualização gráfica que permitam, ao analista que o conduz, refiná-lo até que se detectem padrões efectivamente úteis. Com efeito, o processo global é claramente hierarquizado, começando por incidir em instâncias elementares, embora volumosas, e acabando na obtenção de um ponto relativamente concentrado e muito valioso. O princípio fundamental subjacente ao conceito de *Data Mining* reside no facto deste processo incidir na simplificação sistemática dos dados em bruto, de modo a ignorar aquilo que é específico e a privilegiar aquilo que é genérico. Para que o processo seja eficaz, é necessário desprezar os eventos particulares (*outliers*) para só manter aquilo que é genérico, pois é de padrões que efectivamente se trata.

Para usar o exemplo das empresas, pode-se dizer que estas recebem informação do meio ambiente e que também actuam sobre ele. Durante essas actividades, é importante distinguir os vários níveis de informação, conforme se mostra na figura nº 3. A pirâmide da informação ilustra o aumento da abstracção à medida que se sobe de nível (diagrama da esquerda). A sensível redução de volume, que ocorre cada vez que se sobe mais no nível da pirâmide (diagrama da direita), é uma consequência directa do processo de abstracção.

Neste sentido, abstrair significa representar informação através dos correspondentes simbólicos genéricos. O processo de *Data Mining* localiza padrões através da judiciosa aplicação de processos de generalização, algo que é conhecido por indução. Padrões são unidades de informação que se repetem, ou então são sequências de informações que dispõem de uma estrutura que se repete. A tarefa

[5] Para informação mais detalhada vd. Groth (1998) e Han, Chen & Yu (1996).

de os localizar não é exclusiva do *Data Mining*, visto o cérebro utilizar processos semelhantes de localização de padrões para adquirir o conhecimento que é mantido nas mentes dos humanos. A verdadeira diferença reside na quantidade de informação que o cérebro humano tem capacidade para analisar e padronizar, que é hoje muito inferior àquela que as actuais tecnologias de suporte da informação têm.

O *Data Mining* inicia-se com uma actividade analítica simples de descrição dos dados disponíveis, resumindo os seus atributos estatísticos (tais como médias e desvios padrão), revendo-os graficamente e procurando encontrar ligações óbvias entre as várias variáveis. É durante este primeiro passo que podem ter lugar actividades subsidiárias tais como o *clustering* [6] e a análise de ligações [7]. Para se ser bem sucedido neste processo é fundamental conseguir reunir, explorar e seleccionar os dados correctos. Para isso, há que criar um modelo de previsão baseado em padrões que tenham sido obtidos a partir da realidade, isto é, a partir de um conjunto de dados ou eventos (dados de treino) que correspondam a situações que efectivamente ocorreram no passado. Criado o modelo, é preciso depois testá-lo num outro conjunto de dados (dados de teste), também reais e do passado, que permita aferir a adequabilidade desse modelo. É importante realçar que um bom modelo nunca deve ser confundido com a realidade, mas sim utilizado como um guia útil que permite perceber essa realidade e, de alguma forma, antecipar o futuro.

O último passo do *Data Mining* passa pela verificação empírica do modelo, aplicando-o numa situação concreta. As previsões geradas pelo modelo são então utilizadas para orientar a actividade que se pretende servir e os resultados obtidos a partir dessa actividade são avaliados em função de um referencial.

2.2.1 Hierarquia das Prioridades

Para evitar estabelecer confusão entre os diversos aspectos do *Data Mining*, entende-se ser útil hierarquizar as opções e as decisões a tomar antes de se iniciar este processo, como se segue:

- Determinar o objectivo da actividade;
- Seleccionar o tipo de previsão mais adequado;
- Seleccionar o tipo de modelo ajustado ao tipo de previsão;

[6] Esta actividade de agrupamento permite dividir uma certa base de dados em vários grupos que são muito diferentes uns dos outros, mas cujos elementos, dentro de cada grupo, são muito parecidos.

[7] *Link analysis* – actividade que permite encontrar relações entre os elementos da base de dados; assenta na associação e na sequência de elementos da base de dados.

- Escolher o algoritmo a utilizar no modelo;
- Seleccionar o *software* apropriado.

No topo da cadeia, está naturalmente o objectivo da actividade: qual o objectivo último da aplicação do *Data Mining* a um certo conjunto de dados? O conhecimento dos interesses e dos objectivos da actividade e da organização onde esta se desenvolve deve servir de guia para a formulação do propósito dos modelos a criar.

Segue-se a escolha do tipo de previsão mais adequado: (1) classificação – prever em que categoria ou classe se enquadra um certo evento/registo; (2) regressão – prever que valor numérico adquire uma dada variável (se for uma variável dependente do tempo, então trata-se de uma previsão seriada no tempo).

Pode-se então depois seleccionar o tipo de modelo: talvez uma rede neuronal para executar a regressão e uma árvore de decisão para executar a classificação; há também modelos estatísticos que podem ser utilizados, tais como a regressão logística, a análise discriminante ou os modelos lineares gerais.

Quanto ao algoritmo, há uma grande variedade disponível. O modelo de rede neuronal pode usar algoritmos baseados em funções de *back propagation* ou de *radial basis*. No caso das árvores de decisão, pode-se seleccionar algoritmos designados por CART, C5.0, QUEST ou CHAID [8].

2.2.2 Particularidades dos Modelos de Previsão

Nos modelos de previsão, os valores numéricos ou as classes que resultam da previsão são designados por “resposta”, “variáveis dependentes” ou “variáveis alvo”. Os valores a partir dos quais se lança a previsão são designados por “*predictors*” ou por “variáveis independentes”. Os modelos de previsão são gerados (ou “treinados”) utilizando dados para os quais já se conhece os valores das respostas, isto é, das variáveis dependentes. Este tipo de treino do modelo é conhecido por aprendizagem com supervisão, visto as respostas dadas pelo modelo poderem ser comparadas com resultados conhecidos. Pelo contrário, técnicas como o *clustering* estão compreendidas nas técnicas de aprendizagem sem supervisão, visto não se usar quaisquer resultados já existentes para “guiar” o sucessivo ajustamento (aprendizagem) do algoritmo.

Com a classificação, pretende-se identificar os elementos que caracterizam os

[8] CART - *Classification and regression trees*, Quest - *Quick, Unbiased and Efficient Statistical Tree*, CHAID - *Chi-squared Automatic Interaction Detector*.

vários conjuntos pelos quais os dados/registos/eventos (*instances*) se distribuem. O padrão destas características pode permitir não só entender os dados existentes, mas também prever o comportamento dos futuros eventos. Ao examinar dados existentes, previamente classificados, o *Data Mining* detecta indutivamente padrões de previsão, gerando assim modelos de classificação.

A regressão usa os valores já existentes para prever outros valores. Na sua forma mais simples, a regressão utiliza técnicas simples de estatística, tal como a regressão linear. Contudo, a grande maioria de problemas não consegue ser resolvido através da mera projecção linear dos valores conhecidos, visto depender de muitas variáveis e de complexas interacções entre elas. São necessárias técnicas mais complicadas, como a regressão logística, as árvores de decisão ou as redes neuronais, para fazer previsões adequadas de valores futuros. Alguns modelos podem ser utilizados tanto em classificação como em regressão. É o caso do já aludido algoritmo CART, que pode construir árvores de classificação, classificando variáveis de resposta (discretas) por classes e que pode também construir árvores de regressão, prevendo o valor numérico de variáveis de resposta (contínuas).

A previsão seriada tendo por base o tempo (*times series*) é um caso particular da regressão, que prevê valores futuros das variáveis que dependem do tempo. Os modelos que se apoiam neste tipo de previsão têm em consideração as propriedades relevantes do tempo, tais como a hierarquia dos períodos, a sasonalidade, os efeitos de calendário (feriados), a aritmética das datas, etc.

2.2.3 Modelos e Algoritmos de Previsão[9]

Os modelos mais utilizados para efeitos práticos de *Data Mining* assentam em variantes de algoritmos publicados na literatura da especialidade (ciências da computação ou estatística), especificamente parametrizados no sentido de satisfazer os objectivos de quem os comercializa. Muitos dos modelos e dos algoritmos, a seguir sucintamente descritos, resultam da generalização do modelo de regressão linear. A característica comum a muitas das tecnologias emergentes de *Data Mining* é o facto de adoptarem mecanismos de pesquisa de padrões que estão orientados mais para os dados que para o utilizador, isto é, o relacionamento do modelo com os

[9] Vd. detalhes sobre os modelos e algoritmos enunciados em Two Crows Corp.(1999, p. 11-21).

dados é indutivamente feito pelo próprio *software*, tendo por base os dados sobre os quais incide, e não tanto a interacção nem as funções especificadas pelo utilizador. Interessa reter que os modelos ou os algoritmos não são exclusivos nem suficientemente abrangentes. Num qualquer problema, a natureza dos dados em si afecta a selecção de modelos e de algoritmos. Não há modelos nem algoritmos perfeitos; como em tudo na vida, é necessário obter uma variedade de ferramentas e tecnologias que permitam encontrar a melhor solução para os problemas que se colocam.

2.2.3.1 Redes Neuronais

As redes neuronais têm um interesse particular por constituírem um meio de modelar eficientemente problemas complexos que compreendam elevado número de variáveis independentes que interagem. Uma rede neuronal inicia-se através de uma camada de entrada (*input layer*), cujos nós correspondem às variáveis independentes. Por sua vez, estes nós ligam-se a outros nós pertencentes a camadas intermédias (*hidden layers*), as quais se ligam à camada de saída (*output layer*), que contém as variáveis de resposta produzidas pelo modelo.

2.2.3.2 Árvores de decisão

As árvores de decisão são criadas segundo um processo iterativo de subdivisão de dados em conjuntos discretos, tendo por objectivo maximizar a "distância" entre grupos à medida que se progride nessa subdivisão. Uma dos factores mais relevantes que permite distinguir métodos de árvore de decisão é a forma como caracterizam esta "distância". As árvores de decisão que são usadas para prever variáveis discretas (relativas a categorias) são designadas por árvores de classificação, enquanto as usadas para prever variáveis contínuas são designadas por árvores de regressão.

2.2.3.3 *Splines* de regressão multivaríavel e adaptativa [10] (MARS)

O algoritmo MARS foi concebido com o objectivo de poder obviar algumas das desvantagens do algoritmo CART. Veio assim substituir a descontinuidade existente nos nós, modelando-a por intermédio de um par de linhas rectas, linhas essas que, na base do modelo, são substituídas por linhas curvas, desempoladas, chamadas *splines*. O resultado é uma ferramenta de regressão não linear que actua por

[10] Tradução livre do autor relativa a *multivariate adaptative regression slines* (MARS).

patamares[11]. O MARS, tal como a maioria dos algoritmos de rede neuronal e de árvore de decisão, tem tendência para se sobre-ajustar[12] aos dados de treino.

2.2.3.4 Indução de Regra

A indução de regra é um método que assenta na derivação de um conjunto de regras para levar a cabo a classificação de dados. Contrariamente às árvores de decisão, este algoritmo consegue gerar um conjunto de regras independentes que não formam necessariamente uma árvore. Visto o indutor de regras poder avançar sem forçar a ramificação obrigatória em cada nível, ele é capaz de encontrar padrões diferentes, por vezes melhores, de classificação.

2.2.3.5 *K-nearest neighbour* (KNN) e *Memory-based Reasoning* (MBR)

O *K-nearest neighbour* é uma técnica de classificação que usa uma versão do método com o mesmo nome e que permite atribuir a uma dada classe, ou categoria, um novo registo em função de um determinado número – o número "k" no *K-nearest neighbour* – dos registos já classificados que são mais semelhantes, ou que estão mais próximos, conforme ilustrado e explicado na figura nº 4.

O cerne deste método passa por estabelecer um parâmetro que permita medir a "distância" entre os atributos dos dados/eventos/registos em geral e efectuar o seu cálculo. Embora isto seja relativamente fácil no caso de variáveis numéricas, requer bastante mais cuidado quando se trata de variáveis de categoria ou de classe.

2.2.3.6 Regressão Logística

A regressão logística é uma generalização da regressão linear e é utilizada, em particular, na previsão de variáveis binárias de resposta. Por essa resposta ser discreta, não é possível modelá-la directamente por intermédio de regressão linear. Por isso, em vez de prever se o evento (resposta) vai ocorrer, o modelo é construído para prever o logaritmo da probabilidade do evento vir a ocorrer.

2.2.3.7 Análise discriminante

Esta técnica matemática de classificação é das mais antigas. Determina hiper-planos (e.g., linhas no plano, ou planos no espaço tridimensional) que separam as

[11] *Non-linear step-wise regression tool*.

[12] Tradução livre do autor relativa a *to overfit*. O *overfitting* resulta de um ajustamento excessivo da função de regressão aos dados de treino, fazendo com que o modelo perca o carácter generalizado e, assim, a sua utilidade.

várias classes. O modelo que daí resulta é de fácil interpretação, dado o utilizador se limitar a determinar de que lado da linha ou do plano se deve colocar o registo. O treino do modelo é simples e a técnica é muito sensível a padrões que existam nos dados.

2.2.3.8 Modelos Aditivos Generalizados (MAG)

Há uma classe de modelos que assume ser possível adicionar as funções não lineares associadas a cada variável independente, dando assim outra dimensão aos modelos de regressão linear e de regressão logística. O MAG, em vez de estimar um número elevado de parâmetros, como acontece com as redes neuronais, dá um passo mais à frente, prevendo o valor da resposta para cada valor de entrada e seleccionando o nível de complexidade do ajustamento do modelo em função da natureza dos dados.

2.2.3.9 *Boosting*

Em termos muito simples, *boosting* retira aleatoriamente alguns grupos de dados da base de dados e constrói um modelo de classificação para cada um desses grupos. À medida que se passa de modelo para modelo, é feito variar o conjunto dos dados de treino em função dos resultados obtidos pela aplicação do modelo anterior. A classificação final, atribuída aos eventos/dados/registos que se pretende analisar, é a classificação que apresente maior frequência, depois de aplicados todos os modelos.

2.2.3.10 Algoritmos genéticos

Os algoritmos genéticos não são usados na pesquisa de padrões tal como acontece com os outros já referidos, servindo antes para conduzir o processo de aprendizagem dos algoritmos de *Data Mining*, tal como o de redes neuronais. No essencial, estes algoritmos funcionam como um método para guiar a pesquisa de modelos adequados no espaço das soluções possíveis. São designados por genéticos porque seguem aproximadamente um padrão de evolução biológica, no qual os membros de uma geração (de modelos) competem para passar as suas características à próxima geração (de modelos), até que o melhor (modelo) seja apurado.

3. APLICAÇÕES DO *DATA MINING*

3.1 GESTÃO DE RELAÇÕES COM O CLIENTE

A forma como as empresas interagem com os seus clientes mudou radicalmente durante os últimos anos. A fidelização tradicional dos clientes deixou de estar garantida e as empresas começaram a perceber que precisam de compreender bem os seus clientes, de modo a poder responder prontamente às suas necessidades. Para além disso, o espaço de tempo disponível para dar estas respostas tem vindo sucessivamente a contrair-se, reduzindo a oportunidade para o fazer. Deixou de ser possível esperar pelo aparecimento de sinais de descontentamento por parte do cliente antes de se tomarem as devidas providências. Para se ter sucesso, é cada vez mais preciso ser-se pró-activo e conseguir antecipar as tendências dos clientes.

Há uma série de elementos que contribuem para aumentar a complexidade do relacionamento com os clientes:

- A compressão do ciclo de vida do produto – a abrangência da atenção do cliente decresceu substancialmente e a lealdade é uma coisa do passado;
- O aumento dos custos de promoção do produto – tudo custa mais; impressão, expedição, ofertas grátis (se não o fizer, fá-lo-ão os competidores);
- A grande variedade da oferta – os clientes pretendem produtos que satisfaçam com exactidão as suas necessidades e não apenas produtos que se lhes adaptem;
- Os nichos de mercado – os melhores clientes de uma empresa são uma fonte de cobiça por parte das empresas competidoras, que terão tendência a focar a sua atenção em pequenos e rentáveis segmentos do mercado dessa empresa, na tentativa de lhe subtrair aquilo que tem de melhor.

3.1.1 *Data Mining* aplicado à Gestão de Clientes

As empresas de sucesso são aquelas que têm a capacidade para reagir a todos e a cada um destes elementos, de forma rápida e adequada. O *Data Mining* aplicado à gestão de relações com clientes permite uma melhor compreensão dos clientes no sentido em que permite perceber as suas necessidades e as suas tendências, abrindo o caminho para a obtenção de benefícios tangíveis e para um melhor retorno do investimento (Berson, 2000). Todavia, não pode esquecer-se que o *Data*

Mining é apenas uma fase de um processo de aquisição de conhecimento, muito mais alargado, e não o contrário. Como já se viu, a utilização do *Data Mining* tem de ser feita em conjunto com outras tecnologias da informação (o *data warehousing* ou o *marketing* automatizado) e com as técnicas de *marketing*.

O *Data Mining* utiliza técnicas bem comprovadas de estatística e de aprendizagem de máquina para construir modelos que consigam prever o comportamento dos clientes. A tecnologia existente nos dias que correm já permite automatizar o processo de *Data Mining*, integrá-lo com *data warehouses* comerciais e apresentar os resultados de forma a serem úteis aos responsáveis pelo marketing das empresas. As aplicações mais evoluídas de *Data Mining* são já hoje bastante mais que simples mecanismos de modelação que empregam poderosos algoritmos, tendo-se tornado em ferramentas capazes de abordar tecnicamente assuntos de largo espectro, integrando-os no complexo ambiente das tecnologias da informação.

No passado, tudo parecia indicar que o *Data Mining*, ao emergir, iria eliminar o necessidade dos analistas estatísticos construírem os próprios modelos de previsão. Porém, o valor acrescentado que o analista trás ao processo não pode ser eliminado, porque não pode ser automatizado. Os analistas continuarão a ser necessários para avaliar os modelos e para validar a plausabilidade dos resultados por eles gerados. Por outro lado, serão também indispensáveis para avaliar a relevância dos padrões automaticamente detectados pelo processo de *Data Mining* – o padrão pode existir e, no entanto, não ter qualquer importância para as actividades, acções, comportamentos ou quaisquer outros elementos ligados às circunstâncias que consubstanciam esse padrão.

3.1.2 Relevância para o Exercício da Actividade

Para que o *Data Mining* tenha impacte numa actividade empresarial, precisa naturalmente ser relevante para os processos que suportam essa actividade. O *Data Mining* tem de estar colocado entre a empresa e os seus clientes, embora a forma como influencia a actividade esteja dependente dos processos que a suportam e não do processo de *Data Mining* em si.

A maneira de abordar o assunto passa por compreender que o processo de *Data Mining* permite extrair conhecimento relacional da informação que, residindo nas bases de dados, não é evidente, fazendo assim a diferença relativamente a outros processos suportados apenas na informação. O *Data Mining* extrai da base

de dados informação que o utilizador não sabia existir, transformando-a em conhecimento, de forma mais ou menos automatizada. As relações entre as variáveis e os comportamentos não intuitivos dos clientes são os alvos que o processo de *Data Mining* pretende detectar. E porque o utilizador não sabe antecipadamente o que é que o processo de *Data Mining* vai ser capaz de encontrar, está-se perante um desafio ainda maior, que é o de transformar os resultados do processo em algo que sirva para melhor rentabilizar a actividade da empresa.

A menos que o gestor de marketing seja capaz de compreender qualitativamente os resultados do *Data Mining*, este processo poderá não ter qualquer utilidade. Ambos os casos estão intimamente ligados. O utilizador precisa de obter resultados do processo de *Data Mining* num contexto que lhe permita compreendê-los. Doutra forma, terá tendência a desinteressar-se e a não confiar na sua utilização. Este problema suscita dois desafios: 1) apresentar os resultados do processo de *Data Mining* num formato que seja perceptível e que tenha significado prático e, 2) permitir que o utilizador interaja com os resultados de modo a que possa obter resposta para as questões mais simples.

3.1.3 *Data Mining* e as Relações com o Cliente

A Gestão das Relações como Cliente (GRC) é um processo que gere as interacções entre a empresa e os seus clientes (Berson, 2000). Para obter sucesso, os profissionais têm de identificar primeiro os segmentos de mercado que contenham clientes, ou futuros clientes, que potenciem actividades muito lucrativas. Depois, dedicam-se à execução de campanhas que causem impactes favoráveis no comportamento da população alvo.

A primeira tarefa de identificação dos segmentos de mercado requer a existência de um conjunto de dados significativo sobre os clientes potenciais e os seus comportamentos de aquisição de bens e serviços. Fazendo incidir o processo de *Data Mining* sobre os dados disponíveis, é possível obter resultados que sirvam de dados de entrada para o *software* de gestão de campanhas, que, como nome indica, visa permitir gerir a campanha publicitária dirigida aos segmentos de mercado alvos.

No passado, esta ligação entre o *software* de *Data Mining* e de gestão das campanhas era feito, sobretudo, manualmente. Conseguir integrar estas duas disciplinas é uma oportunidade que se coloca presentemente às empresas, como forma de adquirirem vantagem competitiva.

3.1.4 Valor do Ciclo de Vida do Cliente

Como já se referiu, o *Data Mining* cria modelos a partir de bases de dados que permitem prever o comportamento dos clientes. A previsão fornecida pelo modelo é tipicamente um valor numérico, é calculada para cada um dos registos da base de dados sobre a qual o modelo actua, e indica, para cada cliente, quão provável é apresentar um determinado comportamento. Depois de calculados os resultados de um conjunto de clientes, estes valores numéricos são utilizados para seleccionar os clientes a quem a empresa deve dirigir uma dada campanha, por exemplo, para evitar perdê-los.

3.1.5 Função do Software de Gestão de Campanhas

O *software* de bases de dados comerciais permite às empresas entregar aos seus actuais, e eventuais futuros clientes, as mensagens e as propostas adequadas, no devido tempo e de forma coordenada. O actual *software* de gestão de campanhas já evoluiu um pouco mais, permitindo gerir e monitorar as comunicações dos clientes através dos vários canais e nós de contacto, tais como o correio convencional, o *marketing* à distância, o serviço ao cliente, os pontos de venda, o *Data Mining* interactivo, a agência da empresa, e por aí adiante.

A gestão das campanhas automatiza e integra o planeamento, a execução, a avaliação e a refinação de inúmeras campanhas muito segmentadas, que são desencadeadas com periodicidade variável, ou mesmo intermitentemente. Quanto mais o *Data Mining* e a gestão de campanhas funcionarem de modo integrado, melhores serão os resultados da actividade.

3.1.6 Avaliação de Benefícios

Um outro aspecto decisivo é ser-se capaz de avaliar os benefícios resultantes da aplicação das técnicas de *Data Mining* à GRC. Para isso, podem ser construídos gráficos de resposta que permitam quantificar o êxito de uma dada campanha, consoante seja dirigida a um grupo de clientes aleatoriamente seleccionados ou a um conjunto de clientes, previamente seleccionados, através de um modelo específico de *Data Mining* que antecipe o comportamento dos clientes em geral. O benefício inerente a essa diferença numérica de respostas à campanha pode ser quantificado em termos de custos e de retornos esperados, de modo a permitir aferir aquilo que confere ao modelo verdadeira qualidade: o lucro que advém da sua

aplicação.

3.2 COMPILAÇÃO AUTOMÁTICA DE RECURSOS DA *WEB*

A *World Wide Web*, ou simplesmente a *Web*, pode ser definida como uma enorme colecção de documentos ou de páginas, produzidas livremente por um vasto número de pessoas, sem qualquer controlo editorial significativo. Trata-se, provavelmente da forma mais democrática (e anárquica), abrangente e aberta para qualquer pessoa poder expressar os seus sentimentos, comentários, convicções e ideias, independentemente da cor da pele, do sexo, da religião ou de outro qualquer atributo da raça humana.

Recolher informação a partir da *Web* não é uma tarefa fácil, já que as suas características intrínsecas colocam muitas dificuldades aos utilizadores que a pretendam usar como fonte de informação. Por outro lado, a busca da informação é apenas um primeiro passo no processo de aquisição do conhecimento. Como já se referiu, a *Web* é muito extensa e compõe-se de múltiplas colecções de documentos. Organizar esta informação de modo conveniente é uma maneira de aumentar a eficiência do passo final de aquisição de conhecimento através da exploração da informação criteriosamente reunida. Assim, torna-se necessário construir ferramentas que facilitem a busca, a organização, a exploração e a análise da informação relevante.

Nesta parte do trabalho, procura-se fazer uma sucinta descrição de metodologias que servem o propósito de busca criteriosa na *Web* de recursos de informação e da sua exploração, aproveitando o *Data Mining* como técnica de compilação automática de recursos da *Web*. O Capítulo 1 do Anexo contém elementos suplementares sobre a matéria.

3.2.1 Web Mining

A *Web* é também um serviço público constituído por um conjunto de aplicações destinadas à extracção de documentos existentes em computadores acessíveis na Internet, que é uma rede de computadores. A *Web* pode também ser definida como um repositório de informação distribuída por mais de dois milhões de computadores interligados através da Internet (Baldi et al, 2003). Os termos documento ou página da *Web* podem ser utilizados para significar qualquer ficheiro que se retire da *Web*

através do seu URL [13]. Os mecanismos de busca apresentam aos utilizadores uma série de documentos da *Web*, a cada um dos quais corresponde um URL. Um sítio [14] da *Web* define-se como um local específico na Internet, identificado por um endereço electrónico IP [15] que reverte uma página da *Web* em resposta a uma solicitação http [16]. Um sítio, por si só, é um conjunto de todas as páginas interligadas da *Web*, armazenadas no mesmo endereço IP. O endereço IP é o endereço de rede do computador de Internet que permite a ligação ao servidor através de TCP [17].

Web Mining é o processo geral de detecção e de extracção, a partir de documentos e serviços da *Web*, de informação potencialmente útil e até então desconhecida. Este processo pode ser caracterizado pela sua composição em termos de fases (Kosala et al, 2000), conforme consta no Capítulo 2 do Anexo.

3.2.2 Áreas de Desenvolvimento do *Web Mining*

É normalmente aceite que o desenvolvimento do *Web Mining* se faz em três direcções principais, que estão relacionadas como tipo de dados a tratar: o tratamento dos conteúdos, o tratamento da estrutura e o tratamento da utilização (Kosala et al, 2000).

Há outro tipo de dados que também tem sido objecto de investigação e desenvolvimento: trata-se da alteração dos documentos, da idade das páginas *Web* e do carácter recente da informação; esta matéria relaciona-se com a dimensão temporal e permite analisar o crescimento e a dinâmica da *Web* ao longo do tempo.

O tratamento de conteúdos prende-se com a detecção de informação útil a partir dos dados existentes na *Web* nos mais variados formatos – texto, metadados [18], ligações (*links*), objectos multimédia, páginas dinâmicas ou escondidas e dados semânticos.

O tratamento de estruturas procura inferir conhecimento a partir da estrutura de

[13] URL – *Uniform Resource Locator*, ou URI – *Uniform Resource Identifier*, são uma sequência de caracteres que se referem a um único endereço de um dado documento na *Web*.

[14] Tradução livre do autor relativa *site.*

[15] IP – *Internet protocol*; endereço electrónico de acordo com os padrões do *Internet Protocol*.

[16] http – *Hyper Text Transport Protocol* – protocolo de transporte usado na Internet para transferir documentos de hiper-texto bem como outros tipos de ficheiro. Hiper-texto refere-se a texto organizado em forma de rede de ítens ou módulos de informação (nós) interligados entre si (ligação) permitindo ao utilizador "navegar" seguindo sua própria sequência.

[17] TCP – *Transmission Control Protocol*. Verifica se os dados são enviados de forma correcta, na sequência apropriada e sem erros, pela rede.

[18] Tradução livre do autor relativa a *metadata* – dados que descrevem outros dados.

ligações na *Web*. Os documentos da *Web* apontam tipicamente para outros documentos relacionados, através de uma ligação ou hiper-ligação [19], consistindo numa unidade de texto [20], denominada texto de ancoragem, e num URL associado.

Por fim, o tratamento da utilização da *Web* segue uma das duas aproximações comuns: numa delas, extrai dados dos ficheiros de registo [21] do servidor e usa-os directamente após prévio pré-processamento; noutra, começa por carregar os dados de utilização para uma espécie de base de dados relacional ou multi-dimensional pré-definida, a qual é sujeita a técnicas de *Data Mining* e de OLAP [22], para ajudar os utilizadores a extrair padrões relevantes.

3.2.3 Tarefas do Utilizador

Quando têm necessidades específicas de informação, os utilizadores interagem com *Web* de duas maneiras distintas, denominadas por "navegação" e por "busca"[23]. A "navegação" é o processo de obtenção de informação a que os utilizadores recorrem quando as suas necessidades não estão bem definidas ou quando admitem que os seus objectivos possam alterar-se durante esse processo. Caso contrário, os utilizadores optam naturalmente pelo processo de "busca", recorrendo a um pequeno conjunto de palavras-chave que, na sua perspectiva, caracterizem os objectivos de informação visados. Esta matéria é objecto de desenvolvimento no Capítulo 3 do Anexo.

3.2.4 Serviços de Busca na *Web*

A *Web* tornou-se no maior repositório de informação. Apesar da variabilidade do seu conteúdo em termos de qualidade e de rigor, tem vindo a ser progressivamente usada como fonte privilegiada de informação para as mais diversas actividades.

Seria relativamente fácil encontrar na *Web* a informação que se pretende se existisse um índice geral. Como não existe, a única forma de conhecer os conteúdos da *Web* passa pela busca e análise dos documentos nela contidos.

---

[19] *Link* ou *hyperlink*.

[20] Tradução livre do autor relativa a *text string*.

[21] *Log file* – registo da actividade na *Web* que grava automaticamente a utilização e os respectivos dados tais como a data, a hora, o endereço IP, o estado HTTP, os *bytes* enviados e recebidos, etc.

[22] OLAP – *On line analytical* processing refere-se à tecnologia que permite o rápido acesso a grande volume de dados que está organizado em várias dimensões e está agregado de maneira a pôr o enfoque sobre os factos relevantes e a ignorar os restantes (Agrawal et al., 1997).

[23] Tradução livre do autor para *browsing* e para *retrieval* ou *searching*.

As características da *Web*, designadamente a sua dimensão, a sua dinâmica, a sua heterogeneidade e a qualidade dos conteúdos, são factores adversos às tarefas de busca, que colocam vários problemas e que exigem resolução. Estes problemas são referidos no Capítulo 4 do Anexo, em conjunto com outros tópicos associados à busca na *Web*, tais como a arquitectura, a abrangência e a taxonomia dos mecanismos de busca.

3.2.5 Compilação Automática de Recursos

A *Web* enferma de algumas características adversas à busca de informação, a saber: contém um enorme volume de dados, é dinâmica por natureza, é constituída basicamente por dados semi-estruturados ou não estruturados e é irregular em termos de coerência e de qualidade.

Sempre que um utilizador pretender fazer o seguimento de um dado tópico e organizar as suas referências de acordo com uma estrutura que reflicta o seu ponto de vista, então um sistema de compilação automática de recursos pode desempenhar um papel de grande valia.

Dado o âmbito do presente trabalho optou-se for focar a atenção nas fases do *Web Mining* que mais estrita relação têm com o *Data Mining* e que são as fases de pré-processamento, de aprendizagem automatizada e de análise, na medida em que contribuem para a captação de padrões de busca e de navegação de utilizadores da *Web*. Estas três fases, sucintamente descritas nos sub-capítulos seguintes, estão desenvolvidas em maior detalhe, respectivamente nos Capítulos 5, 6 e 7 do Anexo.

3.2.6 Pré-processamento

A fase de pré-processamento compreende a preparação dos dados e o seu tratamento, de forma a conferir-lhes uma representação que seja válida e adequada para ser reconhecida pelos processos de aprendizagem automatizada.

Finda a fase de preparação, cada documento fica reduzido aos seus atributos representativos – palavras e outros termos relevantes que tenham sido considerados – seguindo-se a sua codificação num formato específico que possa ser utilizado na fase de aprendizagem. Os algoritmos de aprendizagem requerem que a representação dos documentos se faça de acordo com um modelo específico que facilite a sua aplicação.

Nos modelos clássicos, assume-se que cada documento é descrito por um

conjunto de termos indexados, designados por palavras-chave. Um termo indexado é uma palavra ou uma frase que aparece no documento, cuja semântica contribui para identificar os assuntos principais desse documento. Os termos indexados são considerados independentes.

Os modelos de busca, que combinam a informação subjacente ao texto propriamente dito com a informação contida na estrutura do documento, são conhecidos por modelos de busca de texto estruturado.

Sempre que o utilizador não pretenda colocar uma consulta específica ao sistema, mas, ao invés, queira explorar a globalidade dos documentos ou encontrar referências interessantes, pode dizer-se que o utilizador exerce uma actividade de navegação – não de busca.

Os modelos virados para a navegação centram-se na organização da globalidade dos documentos e não na representação dos documentos propriamente dita.

3.2.7 Aprendizagem automatizada

As técnicas automatizadas de aprendizagem, que correspondem normalmente a adaptações de técnicas aplicadas na aprendizagem de máquina e na estatística, podem melhorar significativamente o desempenho e a funcionalidade da compilação automática de recursos das mais variadas formas, desde a classificação de documentos até à modelação do comportamento do utilizador, passando pela busca de informação. A classificação de documentos (categorização) é uma tarefa de atribuição ao documento de uma ou de mais categorias. A aproximação clássica a este problema passa pelo uso de um conjunto de classificadores binários, cada um deles responsável por determinar a relevância do documento para uma dada categoria. Combinando os resultados individuais de cada um desses classificadores pode obter-se um valor de relevância global do documento.

Os classificadores de páginas *Web* podem ser categorizados de acordo com os dados que exploram (Quek,1998): 1) os que usam apenas os dados contidos na própria página - chamados algoritmos de *text mining*, aplicados apenas ao conteúdo textual da página *Web* [24] - e os que tentam explorar as estruturas de marcação (*tags*), tais como cabeçalhos e títulos; 2) os que se servem das hiperligações entre as páginas *Web*; e 3) os que examinam os metadados da página.

[24] São exemplo disso os algoritmos Naïve Bayes and K-nearest-neighbours.

A avaliação do nível de desempenho do classificador de documentos é um elemento relevante das tarefas de classificação de documentos. Esta avaliação pode basear-se em vários índices de medição, cada um dos quais avalia um aspecto particular da acção do classificador.

A aplicação de técnicas de aprendizagem de máquina às tarefas de classificação requer o desenvolvimento de duas fases distintas:

- a fase de aprendizagem propriamente dita, quando o algoritmo de classificação cria um modelo do conceito a ser aprendido, tendo por referência o conjunto de dados de treino e o conjunto de dados de teste;
- a fase de classificação, quando o modelo criado é aplicado a dados desconhecidos com o propósito de os classificar.

Sempre que os conjuntos de dados de treino e de teste são previamente classificados na totalidade (pelo utilizador) está-se na presença de uma metodologia com supervisão de aprendizagem. No caso contrário, em que nenhum deles é previamente classificado, está-se na presença de uma metodologia sem supervisão de aprendizagem (ou de *clustering*). Se o conjunto dos dados de treino for parcialmente supervisionado, então está-se em presença de uma metodologia semi-supervisionada de aprendizagem (Bennet et al, 1998).

3.2.8 Análise dos Recursos obtidos a partir da *Web*

Os conjuntos de resultados obtidos são analisados e validados pelo utilizador, que implicitamente analisa o desempenho do sistema de busca da informação à medida que explora esses conjuntos. É razoável assumir que, se um utilizador descarrega e arquiva muitos dos documentos contidos nos conjuntos, isso quer dizer que estes devem ser relevantes para o seu interesse. Por outro lado, se o utilizador não descarregar um dado documento, é provável que isso queira dizer que a informação nele contida não é relevante.

O utilizador pode até ser induzido a indicar ao sistema quais os documentos que lhe interessam num determinado conjunto de documentos. Este retorno (*feedback*) do utilizador, explícito ou implícito, pode ser usado pelo sistema de busca da informação de modo a melhorar o seu desempenho, em particular no que se refere à relevância dos documentos extraídos. Esta problemática é denominada por retorno

relevante [25]. As técnicas de retorno relevante provocam habitualmente a alteração da consulta, adicionando, removendo ou modificando alguns dos termos, isto é, acabando por originar consultas internas que são diferentes das originais, elaboradas pelo utilizador, e das quais se espera possam ser mais representativas das necessidades de informação desse mesmo utilizador.

3.3 *DATA MINING*, A SEGURANÇA E O CONTRA-TERRORISMO

Está hoje em dia muito claro que a ameaça terrorista que o mundo enfrenta é bastante diferente das ameaças existentes ao tempo da Guerra Fria, requerendo uma adaptação significativa das metodologias de recolha e análise de informações. Contrariamente às potências em confronto durante a Guerra Fria, a actividade terrorista está vagamente organizada numa estrutura difusa e não hierarquizada. Deixou de se poder confiar na realidade anterior em que a informação relevante residia num reduzido, mas valioso, número de fontes, a partir das quais era possível conhecer as capacidades, as tácticas e os planos. Embora os métodos tradicionais ainda mantenham a sua importância, compreender a actividade dos terroristas e prever as suas acções, requer necessariamente uma diferente aproximação à matéria, que passa pela capacidade de utilizar e sintetizar pequenas informações dispersas.

Durante a Guerra Fria, muito do conhecimento sobre o opositor resultava da aquisição clandestina de significativos fragmentos de informação crítica e da aptidão para evitar a sua divulgação. No presente, esse tipo de informação é praticamente impossível de obter. Os atentados de 11 de Setembro de 2001 são uma ilustração clara desta asserção. Na verdade, não foi possível localizar uma única fonte de informação [26] onde estivessem concentrados todos, ou uma parte significativa, dos constituintes do cenário que estava a ser planeado. Todavia, foi possível, à posteriori, identificar um determinado número de pistas, que, se tivessem sido reconhecidas, combinadas e analisadas, poderiam ter gerado o conhecimento suficiente para conseguir capturar os terroristas e evitar a concretização do seu plano. Por conseguinte, embora seja de manter o esforço na melhoria da capacidade

[25] Tradução livre do autor relativa a *relevance feedback*.
[26] Talvez e apenas um elemento humano extraordinariamente bem colocado o pudesse fazer.

para obter conhecimento através das pessoas [27] e de outras fontes tradicionais, tudo indica ser necessário colocar o enfoque no espectro [28] de obtenção de informação e na qualidade da análise. No exercício da actividade de contra-terrorismo, é forçoso ser-se capaz de detectar os dados fragmentados que se encontram no domínio global das informações e de, com eles, conseguir construir uma credível previsão dos cenários futuros.

3.3.1 Aplicação do *Data Mining* ao Contra-terrorismo

O uso das técnicas de *Data Mining* e de análise automática de dados [29] não é uma solução completa para o problema. Contudo, estas técnicas são ferramentas poderosas a utilizar para ir ao encontro deste novo requisito de aquisição de conhecimento. Não obstante a intuição e a colocação de hipóteses continuarem a ser partes insubstituíveis do processo de análise, as referidas técnicas podem auxiliar os analistas e os investigadores na automatização de algumas funções de baixo nível que, de outra forma, teriam de ser executadas manualmente. Estas técnicas podem também ajudar a prioritizar a atenção e a fornecer pistas sobre os elementos de informação em que é preciso focar essa atenção, libertando assim os analistas e os investigadores para se poderem dedicar ao tipo de análise que requer inevitavelmente a capacidade crítica do ser humano. Para além disto, *Data Mining* e outras técnicas relacionadas são óptimas ferramentas para levar a cabo tarefas repetitivas e volumosas, anteriores ao processo de análise, que seriam impossíveis de executar por analistas. Trata-se essencialmente de detectar ligações, padrões e erros [30] em quantidades astronómicas de dados que o ser humano nunca conseguiria encontrar sem a ajuda dos computadores e das aplicações informáticas baseadas nessas técnicas. Uma das potenciais vantagens do uso do processo de análise de dados prende-se com o facto dessa análise incidir sobre enormes bases de dados que, contendo informação pessoal, favorecem o rigor da identificação. A existência de mais informação torna mais fácil descobrir se dois ou mais registos dizem respeito à mesma ou a diferentes pessoas. A tarefa de resolução de identidade é seguramente mais fácil de executar sempre que estiver disponível um

[27] Designada HUMINT (*human intelligence*) – recolha de informações através de espiões.

[28] Largura de banda no acesso à informação por forma a tornar o processo mais abrangente em termos de cobertura de todas as fontes possíveis.

[29] *Automated Data Analysis*

[30] Tradução livre do autor relativa a l*inks, patterns and anomalies*.

grande conjunto de dados sobre o qual se possa fazer incidir essa tarefa. A isto acresce o facto de também ser menos provável determinar quando é que uma pessoa, apesar de identificada, não é efectivamente um instrumento da prática de actividades suspeitas [31], reduzindo assim os inconvenientes e o embaraço para essa pessoa.

Uma ferramenta simples e muito útil de análise de dados para efeitos de contra-terrorismo é "análise de ligações baseada em tópicos" [32]. Esta técnica utiliza um conjunto de registos públicos, ou qualquer outro vasto conjunto de dados, para detectar ligações entre um atributo - que pode ser um nome, um endereço ou outro dado relevante qualquer – e outras pessoas, lugares ou coisas. Isto pode dar aso a novas pistas que os investigadores e os analistas estejam compelidos a seguir. A ferramenta "análise de ligações" consubstancia-se em aplicações informáticas que estão hoje em dia disponíveis e que são utilizadas sobretudo nos EUA, entre outros aspectos, para proceder à verificação da adequabilidade dos candidatos a cargos específicos, e, como ferramenta de investigação em matérias de segurança nacional e de imposição da lei (DeRosa, 2004, p.6). Uma aproximação cuidadosa aos atentados de 11 de Setembro de 2001 permite avaliar quão útil poderia ter sido o uso das ferramentas de "análise de dados baseada em tópicos" no apoio à investigação ou à análise dos planos terroristas (DeRosa, 2004, p.7).

A "análise de dados baseada nos padrões" [33] é uma outra ferramenta que, no longo prazo, também tem um enorme potencial para apoiar a acção de contra-terrorismo, em particular se continuar a desenvolver-se investigação aplicada a este tipo de actividade. A investigação na área do *Data Mining* que mais interessa neste assunto é aquela que conduz à determinação de padrões que permitam caracterizar uma actividade, na verdade, extremamente rara – o planeamento e o ataque terrorista. Deverá também permitir distinguir entre o que é uma identificação do padrão e o que é o "ruído" de actividade patente no conjunto dos dados sobre os quais se opera a busca de informação. Uma das possíveis vantagens da procura de informação através da análise de dados baseada em padrões é que essa procura pode evidenciar pistas de eventuais futuros terroristas, adormecidos, que nunca

[31] As pessoas que, apesar de identificadas através destes processos, se conclui nada terem a ver com a matéria em apreço, são apelidadas de "falsos positivos".

[32] Tradução livre do autor relativa a s*ubject-based link analysis*.

[33] Tradução livre do autor relativa a p*attern based data analysis*. Como já se viu anteriormente o *Data Mining* em si cinge-se à detecção de padrões.

executaram qualquer acção desse tipo, mas que estão ligados a outros que já o fizeram e são, por isso, conhecidos.

Os tipos de busca, através da análise de padrões que podem mostrar-se mais úteis, referem-se à procura de combinações específicas de actividades de baixo nível que, no seu conjunto, induzem a previsão de actividade terrorista. Por exemplo, o padrão de um potencial terrorista (adormecido) poderá ser o de um estudante estrangeiro autorizado, que tenha adquirido um livro de "como fazer engenhos explosivos" bem como uma quantidade significativa de fertilizantes. Ao invés, se a ameaça for que os terroristas usem camiões para efectuar um ataque, a análise de padrões poderá ser dirigida de forma regular para a identificação de pessoas que tenham alugado grandes camiões, que tenham usado hotéis ou caixas postais como endereço e que se enquadrem em certas gamas de idade ou apresentem outros atributos que sejam parte de um padrão conhecido de natureza terrorista. Esta técnica poderá também permitir descobrir padrões significativos no tráfego de correio electrónico que revelem actividade terrorista e que permitam descobrir os cabecilhas dessa actividade (Taipale, 2003). As buscas baseadas nos padrões são também muito úteis para a gestão de consequências e da resposta [34].

3.3.2 Processo de Aplicação

Não obstante se identifiquem inúmeros benefícios na utilização do *Data Mining* e das técnicas automáticas de análise de dados, é importante ter uma compreensão aprofundada do processo que é usado na aplicação dessas técnicas e dos riscos de erro e de intrusão na privacidade das pessoas em geral. Do ponto de vista do contra-terrorismo, o processo pode ser esquematizado e resumido em três grandes fases: a recolha e tratamento dos dados, a identificação de modelos de busca e a tomada de decisão. Relativamente à primeira, é importante referir que importa sempre saber em primeira mão a que tipo de descoberta se destina a análise e que tipo de dados será útil recolher para esse efeito. O passo final desta primeira fase que importa referir é o que transforma os dados para um formato útil, passo que é conhecido por "agregação de dados" [35]. A actividade inerente a este passo centra-se na junção de dados, na sua depuração, para eliminar dados inúteis ou redundantes, e na sua normalização, para tornar as buscas mais certeiras. Quando bem

[34] Tradução livre do autor relativa a *response and consequence management*.
[35] Tradução livre do autor relativa a *data aggregation*.

conduzida, esta fase tem um efeito benéfico no êxito da fase seguinte em termos de redução de erros e de identificação de falsos positivos e falsos negativos.

A segunda fase do processo prende-se com a aplicação de modelos fiáveis, criados antecipadamente, que detectem a existência de padrões pré-determinados nas bases de dados preparadas na fase anterior. Relativamente ao *Data Mining*, o processo inicia-se com o desenvolvimento de algoritmos por parte dos investigadores e termina, como já se disse, na validação dos modelos de previsão criados a partir da aplicação desses algoritmos e de acções de ajustamento e correcção que poderão ser introduzidas por meio de inteligência artificial [36]. A validação dos modelos é um aspecto crítico já que, determinando a sua qualidade intrínseca, permite conter o número de falsos negativos dentro de proporções aceitáveis e produzir falsos positivos que tenham reduzido impacto nas liberdades civis dos inocentes. O contra-terrorismo deita sobretudo mão a modelos que permitam encontrar padrões em dados relacionais [37], dados em que os elementos chave sejam as relações entre pessoas, organizações e actividades, obtidos a partir de uma variedade de diferentes tipos de dados. Os terroristas operam em redes mais ou menos livres, de forma que os modelos seleccionados têm de ser capazes de detectar ligações entre actividades de baixo nível, pessoas, organizações e acontecimentos, a partir dos quais se possa inferir a prática de actividades e organizações terroristas de alto nível. A terceira e última fase do processo de *Data Mining* e de análise de dados envolve a condução das buscas, a interpretação dos resultados e a tomada de decisão quanto ao uso a dar a tais resultados. Uma questão decisiva é até que ponto estas decisões podem ter lugar de forma automática, sem intervenção humana, baseadas apenas na análise automática de dados. Estas técnicas são úteis como ferramentas destinadas a enformar a análise e as decisões tomadas pelos seres humanos, mas não se destinam a substituí-los. Os analistas podem e devem utilizar estas técnicas para avaliar a extensão de pistas ou de suspeitas, para gerar essas pistas e para estruturar as investigações, mas os resultados da aplicação das técnicas, não devem, por si, tornar-se na única fonte de apoio à decisão, de análise ou de conclusão da investigação.

[36] Numa perspectiva de *machine learning*.
[37] Tradução livre do autor relativa a *relational data*.

3.3.3 Riscos

De um modo geral, os cidadãos de qualquer das nacionalidades ocidentais nunca encaram com bons olhos que o respectivo Estado saiba demasiado sobre as suas vidas privadas. Sabem que as acções que podem resultar da acção do Estado em actividades de vigilância baseadas em informação pessoal podem, muitas vezes, ter consequências negativas, tais como a busca domiciliária ou a detenção injustificadas que se traduzem, quer se queira quer não, em perda de reputação e de potencial para gerar proveitos próprios. Por outro lado, os benefícios potenciais do *Data Mining* e da análise automática de dados na actividade de contra-terrorismo são claramente muito significativos. Uma das causas de preocupação por parte dos cidadãos tem a ver com o reduzido conhecimento público de como as técnicas são usadas pelos Estados. Esta falta de transparência não só torna os Estados menos responsabilizáveis e mais propensos a fazer um indevido uso das práticas de análise de dados, como prejudica o uso benéfico, na actividade de contra-terrorismo, de técnicas que já incorporem mecanismos de protecção de informação e de dados pessoais.

A maior preocupação na aplicação das técnicas de *Data Mining* e de análise automática de dados é que esta identifique pessoas inocentes [38], às quais atribui o estigma de terroristas, apenas por apresentarem padrões de comportamento característicos desse tipo de actividade ou porque estabeleceram qualquer ligação fortuita com algum potencial terrorista que nem sabem quem é. O maior desafio que se põe hoje em dia à aplicação das técnicas enunciadas é precisamente o de conseguir evitar que, de deficientes bases de dados ou modelos de busca, se possam extrair resultados que correspondam a falsos positivos. O que isto quer efectivamente dizer é que se os modelos de *Data Mining* não conseguem, quando aplicados, separar o chamado "ruído" do comportamento inocente do "sinal" [39] da actividade terrorista, então o comportamento inocente passará a ser visto como suspeito.

[38] Designados por "falsos positivos".

[39] O "ruído" refere-se a uma generalidade de casos que apresentam um padrão, que, nalguns dos seus segmentos se assemelha ao padrão procurado e que com ele pode ser confundido; o "sinal" refere-se ao padrão procurado.

3.3.4 Privacidade e Tecnologia

Uma das formas de resolver este problema dos falsos positivos, reside, pelo menos em parte, na tecnologia. São referidas quatro novas categorias de tecnologia, desenvolvida com o objectivo de proteger a privacidade e impedir o abuso de utilização dos dados pessoais armazenados em bases de dados (DeRosa, 2004, p.16). Trata-se, em primeiro lugar, de tecnologia destinada a lidar directamente com a falta de exactidão dos dados e os falsos positivos, procedendo ao aperfeiçoamento do modelo a utilizar, por recurso a auto-aprendizagem artificial, baseada no funcionamento do próprio modelo [40], durante a sua aplicação a dados cujo resultado é previamente conhecido. Em segundo lugar, trata-se de tecnologia destinada a garantir o anonimato das pessoas cujos dados pessoais sejam objecto de identificação por parte de um dado modelo de busca. O modelo mais conhecido de protecção da privacidade é conhecido por modelo *K-anonymity* [41]. A encriptação prévia dos dados pessoais é outra via para evitar a divulgação da identidade das pessoas. A terceira tecnologia prende-se com a actividade de auditoria (Lunt, 2003). Esta tecnologia destina-se a proceder à gravação de toda actividade de que as bases de dados relevantes e as redes de dados são objecto, permitindo assim que uma entidade fiscalizadora possa conduzir verificações aleatórias ou dirigidas a essas actividades. Por fim, a quarta tecnologia assenta na aplicação de regras de processamento, integradas nos próprios modelos de busca, que condicionem o acesso aos resultados e aos elementos de identificação em função do grau de permissão do utilizador e da própria natureza dos dados em causa.

As quatro tecnologias atrás referidas têm a possibilidade de ser implementadas isoladamente ou de modo conjugado, contribuindo cada uma delas para tornar mais transparente o acesso a dados pessoais e o seu uso para os fins em vista, sem com isso prejudicar as garantias, as liberdades e a privacidade das pessoas em geral.

4. CONCLUSÕES

Tendo por base o tema proposto, fez-se uma incursão bibliográfica em áreas relacionadas com a informação, com o conhecimento e com a inteligência, sem perder de vista a sua inserção nas modernas tecnologias da informação e numa das

[40] Aperfeiçoamento do modelo através de “machine learning”.
[41] Modelo desenvolvido por Latanya Sweeney da Carnegie Mellon University.

suas metodologias mais promissoras, o *Data Mining*. Rapidamente se entendeu que, sendo vasto, o assunto requeria uma abordagem específica e segmentada que permitisse delimitar a incidência do trabalho e ajustá-lo às expectativas estabelecidas no âmbito do Curso em que se enquadra. A partir das matérias estudadas, fez-se um levantamento prévio dos conceitos de informação e de conhecimento, procedendo-se a uma abordagem um pouco mais detalhada do processo de *Data Mining*, caracterizando-o e descrevendo sucintamente alguns dos modelos que o utilizam e alguns algoritmos que podem ser postos ao seu serviço.

No que respeita à aplicação prática do *Data Mining*, optou-se por abordar três das áreas mais significativas da actualidade, tendo-se apurado, em cada uma delas, diversas conclusões que, sem serem genuínas, se reputam de interessantes.

No âmbito empresarial e no que se refere à eficácia da actividade e à gestão do relacionamento com os clientes, conclui-se que o *Data Mining* é uma ferramenta essencial para detectar não só padrões de comportamento, mas também ligações relevantes entre dados aparentemente não relacionados, que permite prever a variação dos interesses dos clientes, e assim racionalizar despesas e gerar proveitos, com verdadeiro sentido de antecipação e de oportunidade.

No que se refere à *Web*, conclui-se também pelas inegáveis vantagens do aproveitamento do *Data Mining* como técnica de compilação automática de recursos e como parte integrante das metodologias que servem o propósito de busca criteriosa na *Web* de recursos de informação e da sua exploração.

Na área da segurança, conclui-se que as actividades de contra-terrorismo requerem um dispositivo inteligente que recorra às tecnologias avançadas de informação de forma muito mais insidiosa que durante os anos da Guerra Fria e que o *Data Mining* e a análise automática de dados são poderosas ferramentas ao serviço da obtenção do conhecimento e dos agentes de combate ao terrorismo. Conclui-se ainda que a questão da privacidade levanta preocupações que não podem ser descuradas, tudo parecendo apontar para a necessidade de se regulamentar este tipo de actividade, recorrendo, para o efeito, e entre outras medidas, à incorporação nos modelos de tecnologias “defensoras” de privacidade.

BIBLIOGRAFIA

AAS, K., **EIKVIL**, L. (1999), *Text Categorization: A Survey*, Norwegian Computing Center, Report N0 941, 1999. ISBN 82-539-0425-8. Acedido a 5 Jan. 2006 em http://www.nr.no/files/samba/bamg/tm_survey.ps.

ADRIAANS, P., **ZANTINGE**, D. (1996). *Data Mining.* Reading, Mass.: Addison-Wesley Pub.. 1996. ISBN 0201403803.

AGRAWAL, R., **GUPTA**, A., **SARAWAGI**, S. (1997), *Modeling multidimensional databases*. In Proceedings of the 13th ICDE (International Conference on Data Engineering), IEEE Computer Society, Jul-Nov 1997, p. 232-243. ISBN: 0-8186-7807-0.

BAEZA-YATES, R., **NETO**, R. (1999). *Modern Information Retrieval*. ACM Press / Addison Wesley Longman. ISBN0-201-39829-X.

BALDI, P., **FRASCONI**, P., **SMYTH**, P. (2003). *Modeling the Internet and the Web. Probabilistic Methods and Algorithms*. West Sussex, Wiley. ISBN: 0-470-84906-1.

BENNET, K.P., **DEMIRIZ**, A. (1998). *Semi-Supervised Support Vector Machines*. Proceeding of Neural Information Processing Systems, Vol. 11, p. 368-374. MIT Press.

BERSON, A., **SMITH**, S., **THEARLING**, K. (2000). *Building Data Mining Applications for CRM*. New York, McGraw-Hill, 2000. ISBN: 0071344446.

BHARAT, K., **BRODER**, A. (1998), *A technique for measuring the relative size and overlap of public Web search engines*. Proceedings of the 7th World Wide Web Conference, p. 379-388. ISSN:0169-7552.

BRUZA, P., **MCARTHUR**, R., **DENNIS**, S. (2000). *Interactive Internet search: keyword, directory and query reformulation mechanisms compared*. In SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 280-287. ACM Press.

CHAKRABARTI, S. (2003), *Mining the web, Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers. ISBN 1558607544.

CHAKRABARTI, S., **BYRON**, E., **KUMAR**, S., **RAGHAVAN**, P., **RAJAGOPALAN**, S., **TOMKINS**, A., **GIBSON**, D., **KLEINBERG**, J. (1999), *Mining Web Link Structure*, IEEE Computer, Vol. 32, No. 8, p. 60-67.

DeROSA, Mary (2004). *Data Mining and Data Analysis for Counterterrorism*, CSIS Report, CSIS (Center for Strategic and International Studies), Washington, D.C., The CSIS Press, March 2004. ISBN 0-89206-443-9.

ESCUDEIRO, Nuno F. (2004). *Automatic Web Resource compilation using Data Mining*. Tese de Mestrado em Análise de Dados e Sistemas de Apoio à Decisão. Porto: Faculdade de Economia, 2004.

FAYYAD, Usama; **PIATETSKI-SHAPIRO**, Gregory; **SMYTH**, Padhraic (1996). *The KDD Process for Extracting Useful Knowledge from Volumes of Data*. In: Communications of the ACM, Vol. 39, No. 11, p. 27-34, Nov.1996.

GLOVER, E.J., **FLAKE**, G.W., **LAWRENCE**, S., **BIRMINGHAM**, P., **KRUGER**, A., **GILES**, C.L., **PENNOCK**, D.M. (2001). *Improving Category Specific Web Search by Learning Query Modifications*. In Symposium on Applications and the Internet, IEEE Computer Society. San Diego, Saint 2001, p. 23-31.

GROTH, Robert (1998) *Data Mining, a Hands-on Approach for Business Professionals*. Upper Saddle River, New Jersey: Prentice-Hall. ISBN 0137564120.

HAN, Jiawei; **CHEN**, Ming-Syan; **YU**, Philip S. (1996). *Data Mining: An Overview from Database Perspective*. In IEEE Trans. On Knowledge And Data Engineering, Vol. 8, p. 866-883.

JOACHIMS, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In Proceedings of ECML97 (10th European Conference on Machine Learning). Heidelberg, Springer Verlag, p.137-142.

JOACHIMS, T. (1997). *A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization*. In ICML97 (the 14th International Conference on Machine Learning). San Francisco, Morgan Kaufmann Publishers, 1997, p. 143-151.

KOSALA, R., **BLOCKEEL**, H. (2000). *Web Mining Research: A Survey*. SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, Vol. 2, No. 1, p. 1-13.

LUNT, Teresa (2003). *Protecting Privacy in Terrorist Tracking Applications*. Presentation at CSIS Data Mining Rountable. Washington, D.C. Set. 2003. Acedido em http://www.csis.org/tech/2004_counterterrorism.pdf a 16 Dez. 2005.

MITRA, M., **SINGHAL**, A., **BUCKLEY**, C. (1998), *Improving automatic query expansion*. In Proceedings of the 21st ACM SIGIR 98 Conference, p. 206-214. ISBN:1-58113-015-5.

NAVEGA, Sergio C. (2002a). *Projecto CYC: Confundindo Inteligencia com Conhecimento*. In Anais do Workshop Brasileiro de Inteligência Competitiva, Vol. 3. Brasil, S. Paulo. 2002.

NAVEGA, Sergio C. (2002b). *Princípios Essenciais do Data Mining*. In Anais do Infoimagem, Cenadem, Nov. 2002. São Paulo.

QUEK, C.Y. (1998), *Classification of World Wide Web Documents*. Senior Honours Thesis, School of Computer Science, Carnegie Mellon University.

RUSSELL, Stuart; **NORVIG**, Peter (1995). *Artificial Intelligence, a Modern Approach*. Englewood Cliffs, New Jersey: Prentice-Hall Inc..1995. ISBN 0131038052.

SHANNON, Claude (1948). *A Mathematical Theory of Communication*. Bell Systems Technical Journal, No. 27, p. 379-423, 623-656.

TAIPALE, K. A. (2003), *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*. In Columbia Science and Technology Law Review, Vol. 5, No. 2, Dez 2003.

TWO CROWS CORP (1999). *Introduction to Data Mining and Knowledge Discovery*. 3ª. Ed., 1999. Potomac. ISBN: 1-892095-02-5.

WANG, J., **LOCHOVSKY**, F. (2003), *Web Search Engines*, Journal of ACM ComputingSurvey (accepted for revision). Acedido a 20 Jan. 2006 em http://www.cs.cityu.edu.hk/~wangjy/.

YANG, Y., **PEDERSON**, J. (1997). *A Comparative Study of Feature Selection in Text Categorization*. In ICML97 (Proceedings of the 14th International Conference on Machine Learning), p. 412-420. 1997. ISBN:1-55860-486-3.

YANG, Y., **CHUTE**, C. G. (1994). *An example-based mapping method for text categorization and retrieval*. ACM Transaction on Information Systems, pp. 252-277.

LISTA DE ANEXOS

Nº. Pág.

LISTA DE FIGURAS

Nº. Pág.

ANEXO

Compilação Automática de Recursos da *Web*

1. A *Web*

No final de 2002 estimava-se que a *Web* compreenderia, nessa altura, cerca de 4.000 milhões de páginas [42]. De acordo com pesquisas mais recentes, feitas por mecanismos de busca [43], algumas das entidades que recolhem informação relativa à utilização e ao tráfego na *Internet* calculavam que, em 2004, já havia cerca de 900 milhões de utilizadores da *Web*, estimando que se atingiriam os 1350 milhões em 2007 [44]. Dar satisfação às suas necessidades através do ambiente *Web* levanta diversas dificuldades ao utilizador. Algumas delas provêm da própria *Web* em si, designadamente as que de seguida se apresentam:

- Dimensão da *Web* – a quantidade de informação disponível na *Web* torna difícil a busca de informação relevante dentro de um prazo de tempo razoável;
- Dinâmica da *Web* – A *Web* está constantemente em alteração, o que requer um esforço contínuo dos utilizadores que pretendam manter actualizada a informação sobre um dado tópico. A falta de controlo editorial contribui decisivamente para o seu sucesso, mas constitui também um entrave ao processo de busca de informação;
- Heterogeneidade dos documentos – a ausência de normas quanto ao formato e à estrutura de documentos da *Web* permite que qualquer pessoa publique documentos sobre qualquer matéria, em qualquer formato, estrutura ou língua;
- Qualidade do conteúdo – os documentos da *Web* podem conter erros, não ter validade ou ser incorrectos; alguns são mesmo cópia de outros.

Por outro lado, o interface entre o utilizador e a *Web* suscita também outro tipo de problemas, conforme de seguida se refere:

- Especificação das necessidades de informação – a necessidade de informação é normalmente expressa através de um conjunto de palavras-chave que o utilizador julga serem representativas; tal pode constituir uma dificuldade, dadas a ambiguidade e a complexidade da língua usada;
- Apresentação da informação – os resultados das pesquisas efectuadas na *Web* pelos mecanismos de busca são constituídos por conjuntos de documentos. Para os explorar e analisar eficientemente é necessário criar ferramentas informáticas adequadas ao fim em vista.

Os sistemas públicos de obtenção de informação a partir da *Web* assentam essencialmente em dois tipos de métodos: o de "mecanismos de busca" e o de "estrutura de tópicos". Sendo ferramentas genéricas tendentes a satisfazer as perguntas ou consultas [45] dos utilizadores, não estão preparados para dar resposta a uma dada necessidade específica.

2. *Web Mining*

Web Mining é o processo geral de detecção e de extracção a partir de documentos e serviços da *Web* de informação potencialmente útil e até então desconhecida. Este

[42] *Web Characterization Project*. http://wcp.oclc.org .

[43] Tradução livre do autor relativa a *search engines*. (http://searchenginewatch.com).

[44] Nua, (http://www.nua.ie). Nielsen, (http://www.nielse-netratings.com).

[45] Tradução livre do autor relativa *queries*.

processo pode ser caracterizado pela sua composição em termos de fases (Kosala et al, 2000):

- Aquisição – o propósito desta fase é o de descobrir e recolher da *Web* tantos documentos relevantes quanto possível e o de evitar a busca de documentos não relevantes; a relevância é uma medida da utilidade do documento face à necessidade específica do utilizador; nesta fase ocorrem dificuldades com a forma de especificação das necessidades do utilizador, com a medição da relevância dos documentos, com a obtenção de documentos de forma ordenada e com a manutenção de uma imagem actualizada da *Web* relativamente a essa necessidade do utilizador;
- Pré-processamento e Selecção de Informação – compreende qualquer processo de transformação aplicado aos documentos recolhidos; estes processos de transformação são habitualmente aplicados com o objectivo de reduzir o número de termos que constituem os documentos, por eliminação daqueles que são irrelevantes (*stop-words* [46]) e por redução de termos à sua raiz semântica [47], de modo a gerar modelos dos documentos recolhidos, adequados às fases seguintes e também com o objectivo secundário de extracção de informação significativa desses documentos;
- Aprendizagem ou Generalização - esta fase visa a detecção de padrões e a aprendizagem de como satisfazer os objectivos do processo de *Web Mining*; é normalmente implementada recorrendo a técnicas de aprendizagem de máquina (*machine learning*) ou de *Data Mining*, adaptadas às características específicas da *Web*; estas características criam novos problemas a estas técnicas, que requerem uma atenção particular: a falta de estrutura dos documentos *Web*; a dimensão espacial de atributos é muito superior ao número de documentos; rotular (*labelling*) documentos é muito caro e moroso;
- Análise – esta é a fase de validação e de interpretação dos resultados obtidos através do processo de *Web Mining*; por norma, requer a intervenção humana, embora haja já desenvolvimentos no sentido de, na sua interacção com o processo, se obter o retorno (*feedback*) do utilizador que permita manter o processo alinhado com os objectivos materializados nas suas necessidades de informação;
- Apresentação – esta é talvez a fase mais expressiva do *Web Mining*, por duas razões: primeiro, porque compreende a apresentação dos resultados ao utilizador e segundo, porque se trata de uma boa oportunidade para apreender o comportamento desse mesmo utilizador, o qual se reflecte nas acções que ele desencadeia durante esta sua interacção com o sistema. A satisfação de uma determinada necessidade de informação não será devidamente conseguida se não se conseguir disponibilizar ao utilizador um meio simples para apreender e analisar uma colecção de documentos, de maneira que ele possa perceber o seu conteúdo global, a sua estrutura e a sua abrangência.

3. Tarefas do Utilizador

Os dois processos de interacção com a *Web (navegação e busca)* estão subdivididos conforme se esquematiza na figura nº 5. No processo de busca com filtragem, o utilizador coloca as mesmas consultas a uma colecção dinâmica de documentos. Quando o processo sujeita os documentos recolhidos a algum

[46] As palavras que são muito frequentes e sem qualquer significado (tal como "a", "").

[47] Processo que, em língua inglesa, é conhecido por “stemming”; vd. algoritmo de Porter.

mecanismo de escalonamento, está-se em presença de um processo de *"routing"*.
Em termos gerais, o utilizador começa habitualmente por um processo de "navegação", de forma a obter uma ideia genérica sobre o tópico em causa, passando depois para o processo de "busca" do sub-tópico que lhe interessa; localizado o tópico em que efectivamente está interessado, o utilizador terá vantagem em optar em permanência por um processo de "busca" com filtragem. É também sobre este último processo de que mais se falará à frente.

4. Serviços de Busca na *Web*

Estima-se que a *Web* tivesse mais de três milhões de diferentes sítios em 2002 e mais de novecentos milhões de utilizadores em linha (*"online"*) a meio do ano de 2004. Há factores adversos às tarefas de busca que colocam vários problemas e que exigem resolução, como se indica:

- Manutenção de bases de dados dos índices dos mecanismos de busca que estejam actualizadas; a dinâmica do ambiente da *Web*, i.e., a elevada frequência com que as páginas da *Web* são alteradas, acrescentadas ou suprimidas, sem qualquer controlo centralizado, dificulta essa manutenção; dificulta também a manutenção retrospectiva do tópico, i.e., a possibilidade de se dispor do estado de apresentação do tópico ao longo do tempo;
- Melhoria do escalonamento, tornando-o adequado às exigências do utilizador e do momento em que este requer a informação; fazendo parte da compilação de documentos, esta função de escalonamento deverá ser condicionada por três "dimensões" principais – o utilizador, o tópico e o momento;
- Exploração do retorno do utilizador; a informação de retorno do utilizador pode servir para ajustar automaticamente os já referidos modelos de *Web Mining* através de técnicas de aprendizagem de máquina;
- Melhoria da apresentação dos resultados; o interface do sistema com o utilizador é o único meio de comunicação entre os dois; é através deste interface que o utilizador toma conhecimento da colecção de documentos seleccionada pelo sistema e é também através dele que o mesmo utilizador transmite ao sistema as suas necessidades e o seu nível de satisfação, face aos documentos que lhe são apresentados.

4.1. Arquitectura dos mecanismos de busca

Em termos muito sucintos, um mecanismo de busca é um sistema que extrai documentos da *Web* para um repositório local e os indexa de acordo com palavras-chave ou outros termos. A arquitectura centralizada típica dos mecanismos de busca consiste em duas partes funcionais principais (Wang et al, 2003): uma que interage com o utilizador e que é constituída pelo interface de consulta e pelo processamento dessa consulta, e a outra, que é responsável pela busca na *Web*, a extracção de documentos e a sua indexação. Os motores de busca extraem documentos da *Web* à medida que a atravessam, seguindo a sua estrutura de ligações através da execução de um processo de *crawling* [48]. Este processo é executado por *web crawlers*, que são programas que atravessam a *Web* seguindo as ligações identificadas nas página visitadas. O *crawler* começa por um grupo de páginas de raiz (*root set*), determinado pela natureza da consulta a que a *Web* é sujeita. Os

[48] O processo de *web crawling* (também conhecido por *web spider or ant*) é um programa que navega na *Web* de forma metódica e automática. Os *web crawlers* são usados para criar uma cópia de todas as páginas visitadas para que possam ser posteriormente processadas por um mecanismo de busca para efeitos de indexação das páginas extraídas e de pesquisa rápida.

ciclos de *crawling* podem levar semanas ou mesmo meses, o que levanta a questão do grau de "refrescamento" (*freshness*) da informação [49].
A outra vertente do mecanismo de busca, a do interface com as consultas, usa as bases de dados dos documentos e dos índices, sendo responsável pela execução das consultas feitas pelo utilizador e pela apresentação do conjunto de páginas relevantes, escalonadas de acordo com um qualquer critério de importância relativa.

4.2. Abrangência dos mecanismos de busca

Diferentes mecanismos de busca apresentam leituras distintas da *Web* aos utilizadores. É o número de páginas da *Web* que são indexadas e incluídas nas bases de dados de cada um deles que limita a cobertura [50] das suas leituras da *Web*. Os mecanismos de busca comerciais lutam incessantemente por uma melhor posição no nível da cobertura e de actualização da *Web* que conseguem oferecer aos utilizadores. A figura nº 6 apresenta um resumo dos valores dos tamanhos das bases de dados dos mecanismos de busca mais relevantes, em milhões de páginas indexadas, tal como é proclamado pelos detentores de mecanismos, reflectindo a cobertura do mecanismo de busca. A outra vertente da abrangência reportada refere-se ao nível de refrescamento dos índices dos mecanismos de busca. A figura nº 7 reflecte os resultados relativos a este parâmetro de avaliação. A idade das páginas mais antigas existentes em cada uma das bases de dados dos mecanismos de busca reflecte aproximadamente a duração do seu ciclo de *crawling*.

4.3. Taxonomia dos serviços de busca

Como já se referiu, os mecanismos de busca são os serviços de busca da *Web* mais utilizados. Todavia, há outros serviços de busca da *Web* que convém conhecer.
A "estrutura de tópicos" organiza os documentos de forma hierarquizada, representando o conhecimento humano. A sua taxonomia é definida previamente por especialistas de classificação de documentação, o que por um lado garante a validade, a correcção e a precisão [51] associada ao documento e à sua classificação, mas por outro, apresenta um baixo índice de revocação [52] e uma taxonomia estática e invariável. Se o utilizador estiver interessado numa diferente organização dos documentos ou numa categoria específica que tenha sido excluída da taxonomia, então este tipo de serviço de busca por tópicos não é de grande utilidade.
Os "mecanismos de busca" extraem automaticamente documentos da *Web*. A diferença em relação à "estrutura de tópicos" é que os primeiros têm incorporado um processo de *crawling* e não categorizam, com raras excepções, os documentos extraídos de um qualquer modo particular. Os mecanismos de busca genéricos, estando dependentes de demorados processos de *crawling* e possuindo longos índices, não deixam de apresentar baixos níveis de cobertura. Estes mecanismos, podem, em muitas circunstâncias, ser substituídos com vantagem por um conjunto de acessos mais limitados e mais especializados que os dos portais convencionais,

[49] Este problema pode ser minimizado através de uma estratégia adequada de refrescamento, ou por reconstrução periódica do índice ou por actualização incremental. O *crawling* incremental procura visitar apenas as páginas que foram alteradas depois do ciclo anterior (Wang et al, 2003).
[50] Número de páginas da *Web* que estão indexadas na base de dados do mecanismo de busca.
[51] *Precision* - racio entre os documentos relevantes extraídos e a totalidade dos documentos extraídos.
[52] *Recall* -. racio entre os documentos relevantes extraídos e a totalidade dos documentos relevantes.

focando a atenção em parcelas da *Web* em vez do seu todo [53].
Uma dessas abordagem mais especializada é a compilação automática de recursos da *Web*, que pode ser descrita como um conjunto de técnicas de *Web Mining* e de busca de informação, estruturadas para a busca e a organização de colecções de documentos (recursos). Este conjunto pode ser aproximadamente situado a meio caminho entre a "estrutura de tópicos" e o "mecanismo de busca". A utilidade e, por isso, o valor, deste tipo de sistemas pode ser bastante melhorado se permitir uma personalização automática que possua a capacidade de captar automaticamente a variação das necessidades do utilizador. Tendo estas capacidades, um compilador automático de recursos poderá ser usado como um gestor de médio ou longo prazo das necessidades de informação do utilizador desde que consiga fazer o seguimento da evolução dos seus interesses e da evolução natural do tópico em si, na *Web* e ao longo do tempo.

5. Pré-processamento

5.1. Preparação de documentos

O pré-processamento de documentos de texto compreende passos sequenciais que têm por objectivo eliminar os elementos irrelevantes. Pode iniciar-se com a análise lexical que, entre outros aspectos, se destina a eliminar pontuação, acentos, espaços suplementares e a converter maiúsculas em minúsculas. Segue-se a remoção de termos irrelevantes, processo que está muito dependente da língua em que o documento se encontra escrito, bem como da consulta ou dos tópicos de interesse, requerendo normalmente a existência prévia de uma lista de palavras a eliminar (*stop-word list*). Vem depois a redução dos termos à raiz semântica, e por fim a indexação, i.e., o processo que define os termos (atributos) a usar na representação do documento para efeitos da sua modelação e do correspondente índice.
Um outro aspecto com interesse para a compilação automática de recursos e que importa referir, prende-se com o processo de classificação de documentos, já que estes habitualmente contêm um número de atributos (termos) muito superior ao número de documentos disponíveis para fins de "exemplos de treino" [54]. Esta circunstância torna difícil estimar com um grau razoável de confiança a distribuição dos atributos dos documentos e, por maioria de razão, definir o correspondente modelo representativo, sem que se corra o risco de praticar "sobre-ajustamento" [55].

5.2. Representação de documentos

Os documentos obtidos através de mecanismos de busca podem ser organizados e modelados segundo a taxonomia evidenciada na figura nº 8 (Baeza Yates et al., 1999).
Para efeitos da descrição dos modelos de representação de documentos optou-se por usar a notação referida em (Baeza Yates et al., 1999) e que é a seguinte:
$K = \{k_1 , k_2 , ... , k_t\}$ é o conjunto dos termos indexados no sistema, k_j, cujo número

[53] Para este efeito ver *Focused crawling* (Chakrabarti et al, 1999), *Meta-search* (Bharat et al, 1998) e *Interactive search* (Bruza et al, 2000).

[54] Correspondem a um conjunto de dados seleccionados a partir dos quais se procura gerar os parâmetros que permitam criar um modelo capaz de representar esse conjunto de dados e de se tornar também representativo de conjuntos de dados afins.

[55] O *overfitting* ocorre quando não se consegue evitar a inclusão nos dados de treino de dados irregulares ou de dados com regularidade destituída de sentido.

total é *t*. $W_{ij} \geq 0$ é a ponderação associada a cada termo indexado k_i do documento d_j ; $w_{ij} = 0$ quando o termo k_i não aparece no documento d_j .

$\overrightarrow{d_j}$ é o vector de termos indexados associados ao documento d_j ; é representado por $\overrightarrow{d_j} = (w_{1j}, w_{2j}, \ldots, w_{tj})$.

g_i é uma função que retorna a ponderação "*i*" do vector $\overrightarrow{d_j}$; $g_i(\overrightarrow{d_j}) = w_{ij}$.

N é o número total de documentos da colecção.

n_i é o número de documentos nos quais aparece o termo indexado k_i .

$Freq_{ij}$ é a frequência absoluta do termo k_i no documento d_j (número de vezes que o termo k_i aparece no documento d_j) .

F_{ij} é a frequência normalizada do termo k_i no documento d_j, dada por :

$$f_{ij} = \frac{freq_{ij}}{\max_l(freq_{lj})}$$

em que $max_l(freq_{lj})$ é a frequência máxima possível no documento d_j (número de vezes que o termo mais frequente k_l aparece no documento d_j .

5.3. Modelos Clássicos

Modelo boleano

Uma consulta é uma expressão boleana que recorre à utilização de três operadores de ligação – "*e*", "*ou*" e "*não*". As ponderações associadas aos termos indexados k_j são do tipo binário, i.e., $w_{ij} \in \{0,1\}$. A noção de relevância parcial não é aplicável a este modelo. Cada documento d_j é ou não relevante para a dada consulta *g* . A simplicidade é a grande vantagem deste modelo; as desvantagens advêm do facto de permitir a extracção de demasiados (baixa precisão) ou reduzidos (baixa revocação) documentos.

Modelo Vectorial

O modelo vectorial associa ponderações de valor nulo ou positivo aos termos indexados dos documentos e às palavras utilizadas nas consultas. O nível de semelhança entre o documento d_j relativamente à consulta q, que é dado por sim (d_j , q), é obtido calculando a correlação entre os vectores $\overrightarrow{d_j}$ e $\vec{q}$, correlação essa que é definida como o co-seno do ângulo formado entre esses vectores:

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \otimes \vec{q}}{|\overrightarrow{d_j}| * |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

e, $0 \leq sim(d_j, q) \leq 1$. No modelo vectorial, as ponderações dos termos indexados são normalmente obtidas a partir de dois factores: o factor de frequência do termo , TF, que é uma medida da semelhança entre documentos do mesmo grupo [56], e o factor de frequência inversa do documento (IDF), que é uma medida da falta de semelhança entre documentos de grupos diferentes.

Nos modelos vectoriais espaciais, os documentos são representados por vectores num espaço euclideano multi-dimensional. Cada dimensão corresponde a um atributo/palavra existente na colecção de documentos. Cada colecção de documentos é representada por uma matriz *W*, com elementos w_{ij} , que representa a

[56] Tradução livre do autor relativa a *intra-cluster similarity*.

ponderação de cada termo k_j no documento d_j. Este modelo requer a identificação dos termos relevantes – os termos indexados – e o cálculo das correspondentes coordenadas w_{ij} (ponderações). Para se proceder ao cálculo destas coordenadas há vários métodos (Aas, K, et al, 1999): a ponderação Boleana, a ponderação por frequência de palavras, a ponderação TF x IDF, a ponderação entrópica e a ponderação TFC.

Modelo Probabilístico

Dada uma consulta, o modelo probabilístico atribui a cada documento d_j o racio,

$$\frac{P(d_j\_relevante\_para\_q)}{P(d_j\_não\_relevante\_para\_q)}$$

como medida da sua semelhança com a referida consulta, q , o qual permite calcular a probabilidade do documento d_j ser relevante para a consulta q . Este modelo atribui ponderações de natureza binária aos termos indexados, i.e., $w_{ij} \in \{0,1\}$, $w_{iq} \in \{0,1\}$. R é o conjunto de documentos que se sabe serem relevantes; $\overline{R}$ é o complementar de R, i.e., o conjunto dos documentos não relevantes. $P(R|\overrightarrow{d_j})$ representa a probabilidade do documento $\overrightarrow{d_j}$ ser relevante para a consulta q e $P(\overline{R}|\overrightarrow{d_j})$ representa a probabilidade do documento $\overrightarrow{d_j}$ não ser relevante para a consulta.

A semelhança $sim(d_j , q)$ do documento $\overrightarrow{d_j}$ relativamente à consulta q é dada pelo racio:

$$\text{Sim}(d_j , q) = \frac{P(R|\overrightarrow{d_j})}{P(\overline{R}|\overrightarrow{d_j})}$$

A expressão final para efeitos de escalonamento de documentos é dada por (Escudeiro, 2004):

$$sim(d_j,q) \cong \sum_{i=1}^{t} w_{iq} * w_{ij} * \left(\log \frac{P(k_i|R)}{1-P(k_i|R)} + \log \frac{1-P(k_i|\overline{R})}{P(k_i|\overline{R})} \right)$$

onde $P(K_i|R)$ é a probabilidade do termo K_i pertencer aos termos do conjunto de documentos relevantes e os restantes termos são auto explicativos em função desta definição. Este modelo faz o escalonamento dos documentos por ordem decrescente de probabilidade de serem relevantes, o que é uma vantagem. As desvantagens prendem-se com a necessidade de requerer uma separação inicial dos documentos em relevantes e não relevantes e com o facto de, como em modelos referidos anteriormente, apresentar ponderações de natureza binária e assumir a independência entre termos.

5.4. Modelos Estruturados

Os modelos de busca estruturados combinam a informação subjacente ao texto propriamente dito com a informação contida na estrutura do documento. O interesse aqui reside na representação e no processamento de documentos de hipertexto. O hipertexto contém uma série de atributos que não se encontram em documentos de texto comum e que se podem tornar relevantes em termos de conteúdo informativo.
Os atributos do hipertexto em páginas da *Web* compreendem o seguinte: marcações de HTML, URLs, endereços IP, nomes de servidores contidos nos URL, sub-

unidades de texto [57] contidas em URLs, ligações da página em causa para outras páginas [58] ou de outras páginas para a página em causa [59].

Modelo de Listas Não Sobrepostas (*non-overlapping lists*)

Neste modelo, o texto de cada documento é subdividido em regiões de texto separadas que são colocadas numa lista própria. Como há inúmeras maneiras de dividir o texto em regiões separadas, acabam por ser geradas múltiplas listas. Estas listas são mantidas como estruturas separadas e distintas de dados. Embora as regiões de texto incluídas numa mesma lista não se interceptem entre si, o mesmo não se passa naturalmente entre regiões de texto de listas diferentes. Por forma a permitir a busca de termos indexados e de regiões de texto é criado um ficheiro invertido singular, no qual cada componente estrutural do texto corresponde também a uma entrada no índice do documento.

Modelo de Nós de Proximidade (*proximal nodes*)

Este modelo facilita a definição de estruturas independentes de indexação hierárquica sobre o mesmo documento. Cada uma destas estruturas corresponde a uma hierarquia no sentido estrito, composta por nós. A cada um dos nós é associada uma região de texto. Duas hierarquias diferentes podem corresponder a regiões de texto que se interceptam. Um índice invertido de palavras/termos é associado a cada hierarquia de componentes estruturais.

Modelo de Pré-fixação de Percurso (*path pre-fixing*)

Uma forma simples de explorar a informação contida na estrutura de páginas escritas em HTML consiste na prefixação de cada termo – palavra ou frase – com a sequência das marcações de HTML associadas ao próprio termo, separadas dele por um qualquer caracter especial, e.g. ".". (Escudeiro, 2004).
Esta simples técnica é, em si própria, suficiente para garantir uma melhoria significativa nalguns casos (Chakrabarti, 2003), embora gere representações do texto que são muito específicas e inflexíveis.

Modelo Relacional

Relações tais como: (a) Classified (domNode, label), (b) Contains Text (domNode, text), podem ser utilizadas para modelar páginas da *Web*. Sobre o modelo relacional é possível aplicar classificadores indutivos tais como "FOIL", que identifica as regras do esquema de relações estabelecido.

5.5. Navegação

Sempre que o utilizador não pretenda colocar uma consulta específica ao sistema, mas, ao invés, pretenda explorar a globalidade dos documentos ou encontrar referências interessantes, pode dizer-se que o utilizador exerce uma actividade de navegação – não de busca.

Navegação Simples

Neste tipo de modelo, os documentos estão organizados de modo uni-dimensional. Os documentos podem ser representados como elementos de uma lista, na qual cada um deles está associado apenas a um atributo. O utilizador visita de forma

[57] Tradução livre do autor relativa a *sub-strings*.
[58] Designados por *out-links* ou ligações exteriores.
[59] Designados por *in-links* ou ligações interiores.

aleatória os documentos constantes na lista.

Navegação Guiada na Estrutura

Neste modelo, os documentos estão organizados de acordo com uma hierarquia de classes, na qual os documentos são agrupados de acordo com os tópicos relacionados – índice dos tópicos.

Modelo de Hipertexto

O modelo de hipertexto assenta numa estrutura de navegação estruturada de alto nível, que permite navegar através do texto de modo não sequencial no monitor do computador. É basicamente constituído por nós, que estão correlacionados e ligados numa estrutura de grafo.

6. Aprendizagem Automatizada

No ambiente *Web* está-se interessado em conjuntos de classes com mais que apenas duas classes. Trata-se pois de um problema multi-classe. O número de propriedades ou atributos incluídos na representação de um documento é muito mais pequeno que a dimensão espacial de todos os atributos possíveis. Neste contexto, não é possível determinar com toda a certeza se um dado documento é relevante para uma dada consulta ou não. Assim, os resultados das consultas são tipicamente constituídos por um conjunto de documentos, escalonados por ordem decrescente de relevância.
Classificadores de texto da área do *text mining* aplicam-se, sobretudo, ao texto dos documentos propriamente dito e não tiram vantagem de qualquer outro atributo de que o documento *Web* eventualmente disponha, tais como ligações ou metadados.
Vários algoritmos de *text mining* derivados do campo de aprendizagem de máquina têm sido aplicados nas tarefas de classificação de documentos *Web*.

6.1 Classificadores de texto simples

Quando aplicados a páginas *Web*, os métodos clássicos de *text mining* tratam individualmente cada página, ignorando não só as ligações entre elas, mas também a distribuição por classes das páginas adjacentes.

Algoritmo Naïve Bayes

O método de Naïve Bayes (Joachims, 1997) utiliza a probabilidade conjunta de palavras/termos K_j e as categorias V_j para estimar as probabilidades de cada categoria num dado documento (constituído por um conjunto de palavras previamente pré-processado) $P(V_j|K_1, K_2, ...K_n)$, ignorando eventuais dependências entre palavras. Assumindo assim a independência entre as palavras do documento, a probabilidade condicional do documento d relativa à classe V_j, pode ser obtida a partir da fórmula:

$$P(d \mid v_j) = \prod_i P(k_i \mid v_j)$$

i.e., é o produto das probabilidades de cada palavra poder ser classificada na categoria V_j. A aplicação do algoritmo requer naturalmente que se defina um método para calcular as probabilidades condicionais $P(k_j|v_j)$, as quais podem ser obtidas a partir da seguinte fórmula:

$$P(k_i \mid v_j) = \frac{n_k + 1}{n + |vocabulary|}$$

Nesta equação, n é o número total de palavras em todos os documentos de treino pertencentes à categoria v_j, n_k é o número de vezes que a palavra k_j aparece nesse conjunto n de palavras e |vocabulary| é o número total de palavras distintas contidas na colecção de documentos de treino.

Algoritmo k-nearest-neighbours

O algoritmo K-nearest neighbours (KNN) destina-se à classificação de registos [60] e tem apresentado um bom desempenho no reconhecimento de padrões e em problemas de categorização de texto (Yang, 1997) (Yang et al, 1994). Esta metodologia classifica um documento tendo em consideração as características (atributos) dos documentos que lhe são vizinhos até ao grau "k". Os documentos terão de ser pré-processados de forma a poderem ser representados na forma tradicional de modelos vectoriais espaciais. A medida da semelhança entre o documento e o um seu vizinho (ou adjacente) é dada pelo valor do co-seno entre os vectores que os representam. As categorias que apresentam um resultado acima de uma determinada fasquia são atribuídas ao documento.

Árvores de Decisão

As árvores de decisão são estruturas de decisão construídas a partir do topo – nó de raiz – que contém todos os exemplos de treino. A colecção de exemplos em cada nó específico é subdividida numa partição representada pelos nós descendentes. Esta divisão é levada a cabo com o objectivo de minimizar a diversidade de categorias presentes em cada nó e é prosseguida até que não subsistam quaisquer hipóteses razoáveis de maior subdivisão. Em cada nó, a subdivisão é feita de modo a garantir que a soma das diversidades dos nós que dele derivam [61] é menor que a diversidade existente no próprio nó antes da sua subdivisão.
O objectivo é o de maximizar a expressão:

$$diversidade(antes\,da\,divisão) - \sum_i diversidade(nó\,derivado_i)$$

Os nós na base da árvore são designados por nós de folha. Cada um dos exemplos de treino pertence a um dado nó de folha. Cada folha é classificada numa determinada categoria, sendo-lhe atribuída uma taxa de erro, que corresponde à probabilidade dos exemplos nela contidos estarem mal classificados. A taxa de erro global da árvore é uma soma ponderada das taxas de erro de todos os nós de folha.

Máquinas de Vectores de Apoio [62]

Ainda no âmbito dos classificadores de texto simples, foi desenvolvido um outro algoritmo designado por Máquina de Vectores de Apoio (MVA) (Joachims, 1998), que se baseia na intuição de que um hiper-plano que esteja perto de vários exemplos de treino tem uma maior probabilidade de conduzir a decisões erradas que um que se encontre tão longe quanto possível de todos esses exemplos.
Como os anteriores, o algoritmo MVA é um classificador binário que estabelece uma margem máxima através da definição de um hiper-plano colocado entre os conjuntos dos exemplos de treino pertencentes a cada classe ou categoria. Esse hiper-plano é aquele que se situa a meia distância entre cada dois conjuntos e é ortogonal à recta de menor comprimento que une esses conjuntos. Este hiper-plano (x) pode ser

[60] Entenda-se semanticamente este palavra como uma tradução do termo da língua inglesa *instance*. Este termo, aplicado a esta matéria, pode significar um documento composto por texto e respectivas estruturas de marcação (*tags*), apenas um fragmento de texto ou ainda um qualquer outro dado ou conjunto de dados.

[61] Designados por *child nodes*.

[62] Tradução livre do autor relativa a *Support Vector Machines* (SVM).

definido em função dos vectores de apoio, que não são mais que os exemplos de treino que dele estão mais perto:

$$x = b + \sum_i \alpha_i c_i \vec{d_i} \otimes \vec{d}$$

onde o somatório corresponde ao do produto vectorial entre os documentos vectorizados de apoio, $\vec{d_i}$, e o documento vectorizado a classificar, $\vec{d}$, e onde *b* e α_i são parâmetros numéricos a determinar e ajustar pelo algoritmo de aprendizagem.

6.2. Medição do Desempenho

No que respeita a classificadores de documentos e a índices de medição, referem-se algumas métricas comuns: revocação, precisão, exactidão (*accuracy*)[63], erro e novidade (*novelty*). Nestas circunstâncias, o índice de revocação é dado pelo racio entre o número de documentos correctamente classificados na categoria e o número total de documentos pertencentes à categoria.
A precisão corresponde ao racio entre o número de documentos correctamente classificados na categoria e o número de documentos classificados na categoria (que é a soma entre os bem classificados e os mal classificados).
O algoritmo de classificação apresenta normalmente uma tendência para estabelecer opções entre precisão e revocação, correlacionando-as negativamente, i.e., a melhoria da revocação é feita à custa da redução da precisão e vice-versa. Há algoritmos de classificação de texto que funcionam no ponto de equilíbrio entre estes dois índices, em que ambos têm o mesmo valor.
Existe também um índice combinado destes dois, que é o seguinte:

$$F_\beta = \frac{(\beta^2 + 1)\, x\, precisão\, x\, revocação}{\beta^2\, x\, precisão + revocação}$$

onde β é um parâmetro que permite estabelecer uma ponderação diferente entre precisão e revocação para efeitos deste índice conjunto. A exactidão e o erro são medidas complementares da probabilidade de erro do classificador numa dada categoria. O índice de novidade é definido como o racio entre os documentos relevantes extraídos que não são do conhecimento do utilizador e a totalidade dos documentos relevantes (os já conhecidos e os que não eram do conhecimento).

$$V = \frac{N_u}{N_u + N_k} = \frac{N_u}{N}$$

onde N_u é o número de documentos relevantes desconhecidos do utilizador e N_k o número dos conhecidos. (Baeza-Yates et al, 1999):

6.3. Aprendizagem com e sem supervisão

A metodologia semi-supervisionada de aprendizagem (Bennet et al, 1998) pode ser subdividida em outras duas, consoante o número de dados de treino previamente classificados pelo utilizador seja pequeno ou grande, quando comparado com a dimensão do conjunto de dados de treino.
Na aprendizagem sem supervisão não há qualquer conhecimento prévio de classificações, nem das classificações de cada documento, nem ainda das próprias classificações. Os algoritmos de agrupamento (*clustering*) organizam os documentos

[63] Numa dada categoria, racio entre o número de documentos bem classificados e o número de documentos classificados.

em grupos homogéneos, tendo por base a semelhança entre eles, formando sub-grupos dentro do conjunto total dos dados, que minimizem a variância entre os dados contidos em cada sub-grupo e que permitam maximizar a variância entre sub-grupos. A metodologia de aprendizagem com supervisão requer que se classifiquem todos os dados dos conjunto de treino e de teste, ou, pelo menos, que se classifique um número significativo de dados pertencentes a cada categoria.
As técnicas de aprendizagem semi-supervionadas despertam bastante interesse nas situações em que o processo de classificação exige muitos recursos, como é o caso dos documentos da *Web*. Estas técnicas fazem apenas a classificação de um limitado número de documentos dos conjuntos de treino e de teste, recorrendo a processos iterativos tendentes a incorporar no modelo de classificação documentos já classificados por esse próprio modelo em iterações anteriores.

7. Análise dos Recursos obtidos através da *Web*

A aproximação mais comum ao problema passa por extrair um conjunto inicial de documentos relevantes (e eventualmente alguns não relevantes) e por tentar descobrir entre esses documentos atributos cuja forte correlação possa ser usada para estreitar o campo de incidência da consulta inicial, melhorando assim a precisão. O método de Rochio (Chakrabarti, 2003) reflecte o retorno do utilizador, através de uma técnica em que simplesmente se adiciona ao vector-consulta [64] uma soma ponderada de todos os vectores-documento relevantes e se subtrai a soma ponderada de todos os vectores-documento irrelevantes.

$$\vec{q}' = \alpha\vec{q} + \beta\sum_{D^+}\vec{q} - \gamma\sum_{D^-}\vec{q}$$

onde q e q' representam o vector-consulta antes e depois da sua alteração, respectivamente, D^+ e D^- representam o conjunto dos vectores-documento positivos e negativos [65], respectivamente e α, β e γ são parâmetros ajustáveis.
Mitra et al (1998) descrevem uma aproximação pseudo-relevante ao problema que descobre os atributos que estão correlacionados, mas exclusivamente a partir dos documentos mais relevantes, tendo por base a intuição de que uma alteração mais eficaz da consulta pode ser conseguida usando apenas os documentos que se encontram mais "próximos" [66] da consulta original.
Glover et al (2001) propõem um sistema de retorno relevante que se destina a melhorar a revocação, mantendo um valor mínimo aceitável de precisão e cujos resultados obtidos, para além de relevantes para a consulta em causa, passam a pertencer a uma categoria ou classe específicas.

[64] Por exemplo, um conjunto de palavras simples, constituído por uma linha contendo as várias palavras, umas a seguir às outras, separadas por um símbolo qualquer (uma vírgula ou outro).
[65] Dos documentos extraídos inicialmente, os relevantes pertencem ao conjunto dos positivos e os não relevantes ao conjunto dos negativos.
[66] Por isso são considerados os mais relevantes.

FIGURAS

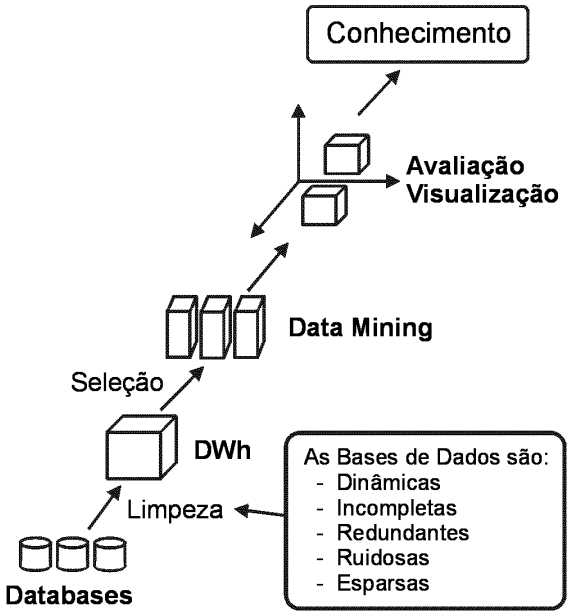

Figura nº 1 - Processo de descoberta de conhecimento a partir de bases de dados (retirado de [Navega, 2002]).

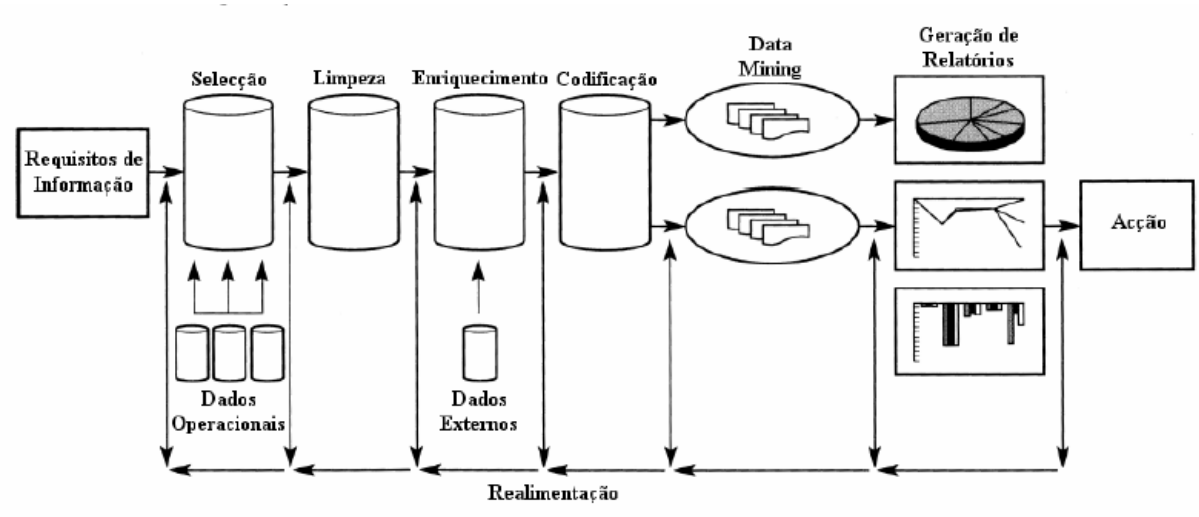

Figura nº 2 - Processo de descoberta de conhecimento a partir de bases de dados (adaptado de [Adriaans1996, p.38]).

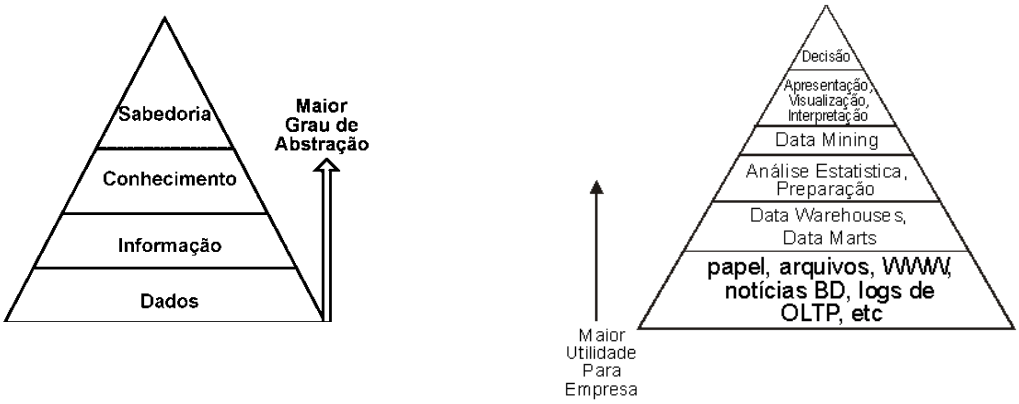

Figura nº 3 – Pirâmide da Informação (à esquerda) aplicada à actividade de uma empresa (à direita) (retirado de [Navega, 2002]).

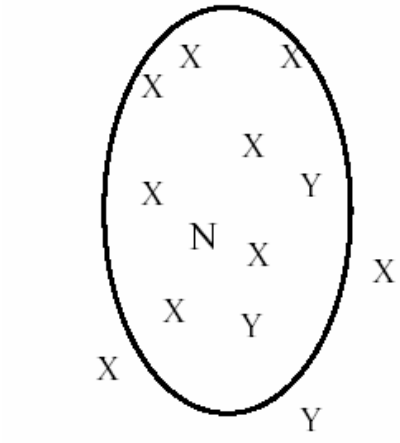

Figura nº 4 – Técnica do *K-nearest neighbour*. Diagrama auxiliar.
(N é o novo registo ao qual é atribuída a classificação "X", por ser esta a classificação predominante dentro da elipse. Para classificar um novo registo, faz a contagem do número de registos de cada classe que se encontram dentro de um determinado perímetro em torno do registo, perímetro esse que é controlado pelo número "k", atribuindo-lhe a classe da maioria desses vizinhos.)

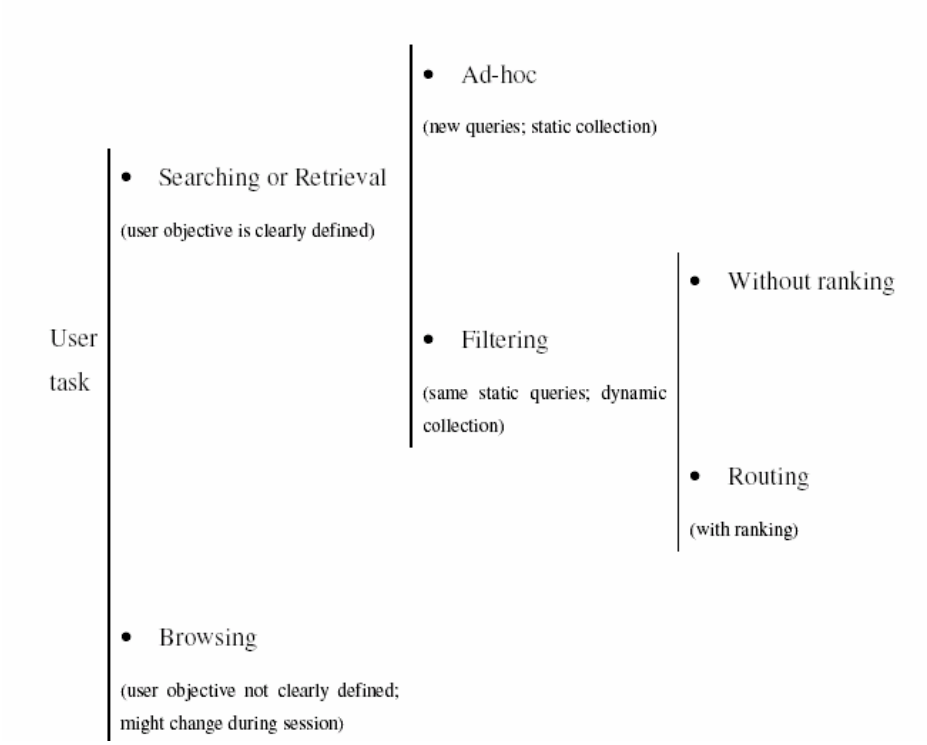

Figura nº 5 – Processo de Interacção com a *Web*.
(*searching or retrieval* – busca; *Browsing* – navegação)

Search Engine	Showdown Estimate (millions)	Claim (millions)
Google	3,033	3,083
AlltheWeb	2,106	2,112
AltaVista	1,689	1,000
WiseNut	1,453	1,500
Hotbot	1,147	3,000
MSN Search	1,018	3,000
Teoma	1,015	500
NLResearch	733	125
Gigablast	275	150

Data from: Dec. 31, 2002

Based on AlltheWeb reported size and percentages from relative size showdown

AlltheWeb: 2,106,156,957 reported

Figura nº 6 – Dimensões das Bases de Dados dos Mecanismos de Busca (*search engines*) mais relevantes (referidas a Dezembro de 2002) (*Search Engine Showdown* retirado de http://searchengineshowdown.com).

Mecanismo de Busca	Idade da pagína mais recente	Idade Média	Idade da pág. mais antiga
MSN (Ink.)	1 day	4 weeks	51 days
HotBot (Ink.)	1 day	4 weeks	51 days
Google	2 days	1 month	165 days
AlltheWeb	1 days	1 month	599 days*
AltaVista	0 days	3 months	108 days
Gigablast	45 days	7 months	381 days
Teoma	41 days	2.5 months	81 days
WiseNut	133 days	6 months	183 days

Figura nº 7 – Nível de Refrescamento (*freshness*) dos Índices dos Mecanismos de Busca (referido Dezembro de 2002) (*Search Engine Showdown* retirado de http://searchengineshowdown.com).

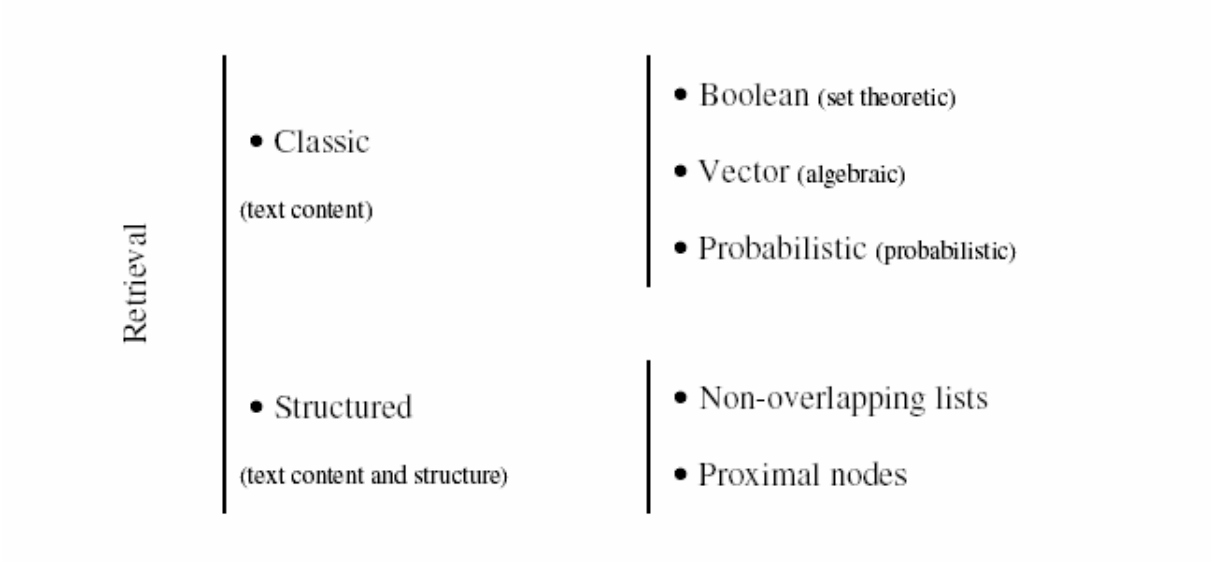

Figura nº 8 – Taxonomia de Documentos (retirado de [Baeza Yates et al., 1999]).
(classical – modelos clássicos; Structured – modelos estruturados)