Data Mining no suporte à construção de Conhecimento Organizacional

Isabel Ramos

Departamento de Sistemas de Informação, Universidade do Minho, Guimarães, Portugal

iramos@dsi.uminho.pt

Maribel Yasmina Santos

Departamento de Sistemas de Informação, Universidade do Minho, Guimarães, Portugal
maribel@dsi.uminho.pt

Resumo

O processo de descoberta de conhecimento em bases de dados automatiza a descoberta de relacionamentos e outras descrições a partir dos dados. Os padrões extraídos durante este processo são considerados conhecimento, em relação aos conceitos teóricos que sustentam esses padrões. Neste artigo é apresentada uma outra perspectiva, na qual se defende que estes padrões não podem ser considerados conhecimento, uma vez que o conhecimento apenas pode residir na mente humana. Em vez de conhecimento, os padrões devem ser apresentados como representações do conhecimento ou informação. Através da apresentação de um caso de teste, no qual se evidenciam as diversas fases do processo de descoberta de conhecimento na análise de uma base de dados organizacional, é possível constatar que ao construir os modelos, a ferramenta de descoberta de conhecimento utilizada permite trazer para a memória consciente da organização a sua experiência passada, contribuindo para um reforço ou reformulação da sua própria identidade e do seu papel no mercado para o qual se direcciona.

Palavras chave: *data mining*, conhecimento organizacional, construção de conhecimento, descoberta de conhecimento.

1. Introdução

O processo de descoberta de conhecimento em bases de dados automatiza a descoberta de relacionamentos e outras descrições a partir dos dados. A fase de *Data Mining* inclui a aplicação de algoritmos para extracção de padrões nos dados sem a realização dos passos adicionais do processo de descoberta de conhecimento tais como incorporação de conhecimento anterior e interpretação de resultados.

Os padrões extraídos durante o processo de descoberta de conhecimento são considerados conhecimento, em relação aos conceitos teóricos que sustentam esses padrões. Neste artigo é apresentada uma outra perspectiva. As autoras defendem que estas regras não podem ser consideradas conhecimento, uma vez que o conhecimento apenas pode residir na mente humana em contínua ligação com a realidade interna e externa. Em vez de conhecimento, estes padrões são apresentados como representações do conhecimento ou informação.

A interpretação dos resultados obtidos na fase de *data mining* pode conduzir a um entendimento inicial sobre a experiência passada da organização. Este primeiro entendimento deve ser aprofundado experimentando-o em situações organizacionais específicas tais como

processos de decisão, projectos organizacionais, e reformulação de práticas. Desta forma os actores organizacionais criam novas ideias e modelos mentais, reformulam modelos anteriormente criados, redefinem ligações conceptuais, desenvolvem respostas emocionais em relação a essas ideias e modelos, i.e., os actores organizacionais criam novas experiências e ligações com a realidade circundante.

Este artigo encontra-se organizado da seguinte forma. Na secção 2 descrevem-se os conceitos associados ao conhecimento organizacional abordados ao longo deste artigo. A secção 3 apresenta a exploração de uma base de dados organizacional, seguindo as diversas etapas do processo de descoberta de conhecimento. Na secção 4 sistematizam-se os principais resultados obtidos, assim como se salienta como os mesmos podem ser utilizados na construção de *novo* conhecimento organizacional. A secção 5 conclui com uma síntese do trabalho realizado e com propostas de trabalho futuro.

2. Informação, Conhecimento e Conhecimento Organizacional

Nesta secção são apresentados os conceitos que servem de suporte à interpretação que as autoras fazem do processo de descoberta de conhecimento em bases de dados e dos resultados que se obtêm nesse processo.

Neste trabalho considera-se que o conhecimento é socialmente criado pela acção e interacção humana em contextos sócio-culturais específicos [Kafai e Resnick 1996]. Estes contextos, organizações e sociedade, são responsáveis pelos padrões estáveis da acção e interacção humana. A aparente estabilidade da interacção humana confere uma dimensão de objectividade às realidades de trabalho. Esta dimensão de objectividade contribui para a noção de que existe um conhecimento objectivo que pode ser armazenado fora da mente humana ou descoberto independentemente de um enquadramento interpretativo, como seria o caso da descoberta de conhecimento através da utilização de técnicas de *data mining*.

Em (Ramos e Carvalho, 2003), a primeira autora deste artigo questiona de forma detalhada a validade desta visão. Neste artigo faz-se apenas referência a algumas ideias fundamentais sobre a construção individual e social do conhecimento.

No entanto, a construção de conhecimento individual – aprendizagem individual – envolve a criação de imagens mentais de um objecto de interesse (pessoa, evento, lugar, melodia, dor, etc.) que possa ser percebido pelas capacidades sensoriais humanas. Esta aprendizagem individual pode ainda implicar a criação de novas relações entre imagens mentais, ou seja, novas inter-relações entre objectos mentais [Damásio 1999].

A construção de conhecimento individual é um processo constituído por quarto elementos fundamentais: (i) cognição, (ii) emoção, (iii) acção e (iv) interacção [Damásio 1999]. As funções cognitivas como linguagem, memória, razão, e atenção interagem para produzir objectos mentais – símbolos, esquemas, ideias, planos, etc. – e seus relacionamentos. Estas construções mentais ajudam-nos a atribuir sentido à realidade que nos rodeia e ao nosso papel nessa realidade. Estes objectos existem na nossa mente como imagens, as quais incluem as suas características, as respostas emocionais adequadas, os planos que fazemos para eles e as suas inter-relações com outros objectos.

A aprendizagem é inseparável da nossa acção e da interacção com outras pessoas ou objectos do mundo exterior. Através da nossa acção e interacção construímos e reconstruímos os nossos espaços mentais pela experiência com os objectos físicos e pela comunicação com os outros [Kafai e Resnick 1996].

É neste contexto que as Tecnologias de Informação (TI) e suas aplicações se podem tornar num importante factor de construção de conhecimento individual na medida em que facilitem a

construção de novas imagens mentais ou a reconstrução das imagens anteriormente formadas. No entanto, isso só é possível no contexto de um conhecimento prévio da organização e das realidades de trabalho nela integradas – os seus elementos essenciais, a resposta emocional adequada aos acontecimentos, problemas e conflitos, e o lugar que a organização ocupa nos contextos mais abrangentes do sector de negócio em que se situa e da sociedade em geral. É ainda necessário que haja um conhecimento do papel organizacional que o indivíduo desempenha e da decisão ou acção que determinada aplicação das TI deve apoiar [Ramos e Carvalho 2003].

É ainda importante notar que o conhecimento desenvolvido ao nível individual só pode ser partilhado mediante a implementação de construções sociais adequadas. Estas construções sociais – relacionamentos de trabalho, artefactos físicos, metas e projectos partilhados, normas e tradições culturais – dão sentido às acções planeadas e permitem a criação do chamado conhecimento organizacional, isto é, conhecimento partilhado pelos vários actores organizacionais [Kafai e Resnick 1996]. Elas tornam tangíveis as ideias e significados, apoiam a negociação de interesses, e facilitam a comunicação. No caso específico das técnicas de *data mining*, elas podem apoiar, por exemplo,

i) a coordenação de tarefas (criação e reforço de relacionamentos de trabalho),

ii) a partilha de modelos de comportamento dos consumidores (partilha de artefactos físicos),

iii) o lançamento de campanhas de marketing (partilha de metas e projectos),

iv) a criação de uma cultura de tomada de decisão apoiada pela informação acumulada pela organização.

Neste contexto, é importante realçar que as autoras consideram que os factos, acontecimentos, coisas, conceitos, modelos, e ideias registados e armazenados em bases de dados não podem ser considerados conhecimento, uma vez que este apenas pode residir na mente humana continuamente em ligação com realidades internas e externas. Tudo o que se encontra armazenado nas bases de dados é representações de conhecimento ou informação. Toda a informação que se ajusta aos nossos espaços mentais e sociais possui o potencial de criar novo conhecimento através dos processos de cognição, emoção, e interacção.

Tipos de conhecimento

Tendo em consideração tudo o que ficou acima dito, as autoras consideram que o conhecimento organizacional pode ser classificado em duas grandes categorias: (i) Conhecimento individual, e (ii) Conhecimento partilhado. Para cada uma destas duas categorias, o conhecimento pode ser classificado em: (i) tácito e (ii) explicito (Tabela 1).

Estas categorias são dinâmicas e, ao longo do tempo, o conhecimento é reformulado, novo conhecimento é acrescentado a cada uma delas, e produzem-se transferências de conhecimento entre elas (individual para partilhado, tácito para explícito) [Orlikowski 2002].

Tabela 1 – Conhecimento organizacional

Conhecimento	Individual	Partilhado
Tácito	Resulta da experiência de vida dos indivíduos e condiciona de forma inconsciente a acção dos indivíduos.	Resulta da interacção prolongada de vários indivíduos e condiciona de forma inconsciente a acção dos grupos.
Explícito	Resulta da experiência de vida dos indivíduos, de reflexão sobre essa experiência, e de aprendizagem direccionada. Condiciona de forma consciente a acção dos indivíduos.	Resulta da interacção prolongada de vários indivíduos, da reflexão conjunta sobre essa interacção, e de aprendizagem direccionada. Condiciona de forma consciente a acção dos grupos.

De notar que mesmo quando partilhado, o conhecimento continua a residir apenas na mente humana. Este conhecimento é partilhado na medida em que ele representa um entendimento comum da realidade organizacional e dos papéis que os vários indivíduos desempenham nessa realidade.

A gestão do conhecimento organizacional

De acordo com o que atrás foi dito, o conhecimento organizacional é o conhecimento que cada indivíduo possui de forma consciente ou não consciente bem como aquele que partilha com os restantes actores organizacionais. Este conhecimento organizacional determina a forma como a organização – conjunto de indivíduos que interage de forma estruturada para atingir metas e objectivos comuns e conjunto de recursos utilizados para apoiar ou permitir essa interacção – se compreende a si própria, reage a estímulos externos, decide e planeia a sua acção, reflecte sobre a sua experiência, aprende e corrige a sua acção, comunica e interage com o seu exterior. Neste sentido, podemos identificar os vários elementos da mente humana expressos na "mente organizacional", ou seja, (i) cognição, (ii) emoção, (iii) acção e (iv) interacção.

Assim sendo a gestão do conhecimento organizacional traduz-se numa gestão eficaz do processo de construção e partilha de conhecimento de forma a potenciar a acção organizacional, tornando-a mais eficaz na prossecução das suas metas e objectivos [Maier 2002].

As várias aplicações de TI têm um papel muito importante de amplificadores e/ou inibidores dos vários elementos da mente organizacional e dos processos de aprendizagem. As secções seguintes apresentam uma ilustração prática do enquadramento teórico apresentado (esta ilustração está limitada à análise de um caso de teste). Tendo por base o processo tradicionalmente denominado por descoberta de conhecimento em bases de dados, o artigo faz uma reflexão sobre o papel das técnicas de *data mining* na construção de conhecimento organizacional.

3. *Data Mining* – a análise de uma base de dados organizacional

Esta secção apresenta uma síntese dos principais conceitos associados à descoberta de conhecimento em bases de dados, assim como evidencia a exploração de uma base de dados organizacional, atendendo aos princípios subjacentes à descoberta de conhecimento.

O processo de descoberta de conhecimento

A Descoberta de Conhecimento em Bases de Dados (DCBD), *Knowledge Discovery in Databases*, é definida como *"o processo não trivial de identificação de padrões válidos e potencialmente úteis, perceptíveis a partir dos dados"* ([Fayyad, et al. 1996] pág.6). Os algoritmos utilizados para extrair padrões dos dados são denominados de algoritmos de *Data Mining*. O processo global de DCBD, que se desenvolve em várias fases, inclui a gestão dos algoritmos de *Data Mining* e a interpretação dos padrões encontrados pelos mesmos, os quais serão utilizados posteriormente no suporte à tomada de decisão. Além de *iterativo* (uma vez que pode existir retrocesso à etapas anteriores), este processo é também *interactivo*, já que requer a participação do utilizador sempre que é necessária a tomada de decisão.

Na definição apresentada, um *padrão* pode ser caracterizado por modelos, relações ou estruturas nos dados, que devem ser *perceptíveis*, se não imediatamente, após determinado período de processamento. Os *dados* representam um conjunto de factos armazenados numa base de dados, na qual subconjuntos do mesmo são responsáveis pela caracterização de diversos padrões. O termo *processo* está associado à execução de diversos passos iterativos, que vão desde a selecção dos dados a analisar até à interpretação de resultados. O processo é

assumido como *não trivial* uma vez que pode envolver a procura de estruturas, modelos, padrões ou parâmetros. Os padrões descobertos deverão ser:

- *válidos* quando aplicados a novos dados (isto é, dados não considerados na construção do modelo ou determinação do padrão);
- *desconhecidos*, do sistema utilizado na sua detecção e preferencialmente do utilizador; e ainda,
- *úteis* para o utilizador, auxiliando o processo de tomada de decisão.

Um dos principais problemas com que se deparam as técnicas de *Data Mining* é que o número de possíveis relacionamentos é extremamente elevado, ocultando por vezes os mais importantes. As estratégias de pesquisa têm então de ser inteligentes, para o que se recorre à área da Aprendizagem Automática (*Machine Learning*). Outro problema encontrado com bastante frequência é a existência de dados corrompidos ou desconhecidos, os quais conduzem normalmente à utilização de técnicas estatísticas para avaliar o grau de confiança dos relacionamentos encontrados [Holsheimer e Siebes 1994].

O processo de descoberta de conhecimento é iniciado com a Aprendizagem do domínio de aplicação, o que inclui a percepção do conhecimento relevante sobre o domínio e os objectivos a atingir no processo. Posteriormente, procede-se à execução das diversas etapas que caracterizam este processo, as quais são:

1. *Selecção dos dados*. A selecção dos dados tem como principal objectivo limitar o espaço de pesquisa, eliminando atributos irrelevantes para o processo de descoberta de conhecimento.
2. *Tratamento dos dados*. Entre os procedimentos habituais nesta fase, destaca-se a des-duplicação de registos, normalmente originada por negligência na introdução dos dados, pelo incorrecto fornecimento dos mesmos ou por um erro de digitação. É também frequente o aparecimento de dados com valores omissos, para os quais é necessário definir uma estratégia de actuação.
3. *Pré-processamento dos dados*. Passa essencialmente pela redução do espaço de pesquisa, isto é, pela diminuição do número de linhas/colunas a analisar. Esta redução é conseguida transformando, por exemplo, os atributos com valores contínuos em atributos com valores discretos. É também possível a generalização de atributos, para o qual são utilizadas as hierarquias conceptuais definidas para o domínio de aplicação em causa.
4. *Data Mining*. Esta é a fase de procura, na qual os dados provenientes da fase de pré-processamento são analisados. A verificação do tipo de resultados pretendido, *tarefa a executar* (classificação, segmentação, …), permite a identificação da *técnica a utilizar* (indução de regras, redes neuronais, …). Para atingir os objectivos propostos pode ser necessário utilizar mais do que uma técnica, já que a quantidade e o tipo dos dados disponíveis influência de forma decisiva os resultados que podem ser encontrados. Mais detalhes sobre as tarefas e técnicas podem ser encontrados em [Han e Kamber 2001] [Santos 2001].
5. *Interpretação de resultados*. Nesta fase procede-se a análise dos resultados obtidos na etapa anterior. Os modelos encontrados são aplicados a novos conjuntos de dados, permitindo verificar o desempenho dos mesmos com dados desconhecidos para o sistema. A ocorrência de falhas ao longo do processo de descoberta de conhecimento, originadas por decisões que se revelam inapropriadas, é normalmente traduzida na obtenção de modelos que não satisfazem o interesse do utilizador (subjectivo, já que em termos objectivos os algoritmos utilizados verificam quantitativamente o interesse das regras), ou que apenas retratam o comportamento dos dados analisados, não podendo ser aplicados a dados desconhecidos. Nestes casos, existe a possibilidade de retrocesso a fases anteriores para rectificar as

decisões tomadas ou para incluir novos dados na análise. O processo é então retomado, permitindo identificar novos modelos que resultam das alterações efectuadas.

Em termos de "trabalho", a fase de *Data Mining* representa normalmente 20% do tempo gasto em todo o processo. Esta é também a fase que é melhor suportada automaticamente (por *software*). Todas as outras fases, desde a selecção dos dados até a interpretação dos padrões encontrados, constituem mais uma questão de "arte" do que uma rotina que possa ser automatizada [Andrienko e Andrienko 1998].

A análise de uma base de dados organizacional

Nesta subsecção é apresentada a análise de uma base de dados organizacional com o objectivo de descoberta de conhecimento. A base de dados analisada integra um conjunto de dados fictício, os quais foram preparados com o objectivo de auxiliar o processo de assimilação dos conceitos associados à DCBD e ainda, das diversas técnicas e algoritmos disponíveis no Clementine. Este conjunto de dados permite exemplificar o processo de descoberta de conhecimento e a intervenção do utilizador requerida no mesmo.

O Clementine [SPSS 1999] é uma ferramenta de DCBD que permite executar todas as fases deste processo. É um sistema baseado em programação visual, cuja filosofia de trabalho assenta na construção de *streams*, nas quais cada operação sobre os dados é representada por um nodo. Nodos com funções similares encontram-se organizados em paletas, permitindo ao utilizador seleccionar o nodo mais apropriado para a execução de uma dada tarefa.

O conjunto de dados seleccionado para análise agrupa 3.031 registos que caracterizam os clientes de uma empresa de financiamento, que fornece crédito para a aquisição de bens. Para estes dados, foram definidos os seguintes objectivos:

Objectivo do negócio: minimizar o risco de incumprimento que advém do financiamento concedido aos clientes.

Objectivo do Data Mining: conseguir determinar o perfil dos clientes, de forma a minimizar o risco de investimento da empresa.

Compreensão dos dados

Antes de prosseguir com as diversas fases do processo de descoberta de conhecimento, é necessário analisar os dados a explorar, de forma a compreender o significado de cada um dos atributos, e definir estratégias de análise para os mesmos.

1. Descrição dos dados

Os atributos que integram os dados a analisar são: Identificação, Número fiscal, Estatuto, Nome, Bem financiado, Tipo de contrato, Duração, Rendimento bruto, Valor do crédito, Tipo de pagamento, Crédito à habitação, Valor da prestação, Estado civil, Número de filhos, Idade e Incumprimento.

Globalmente, refere-se que além da identificação dos clientes, à qual é associado o número de filhos, é referido o bem financiado, o tipo de pagamento seleccionado pelo cliente, o valor da prestação e ainda, se o cliente possui um outro financiamento para a habitação. O atributo incumprimento é utilizado para assinalar os clientes que verificaram anomalias no pagamento das respectivas prestações.

2. Exploração dos dados

Nesta fase pretende-se detectar anomalias nos dados, verificando o conjunto de valores que cada atributo armazena e ainda, a sua distribuição. A exploração dos dados foi realizada no Clementine, recorrendo aos nodos Distribution e Histrogram da paleta Graphs, através da *stream* apresentada na Figura 1. Nesta figura é ainda possível verificar a qualidade dos dados que,

excluindo três atributos com valor informativo, Nome, Número Fiscal e Estatuto, se encontram completamente preenchidos.

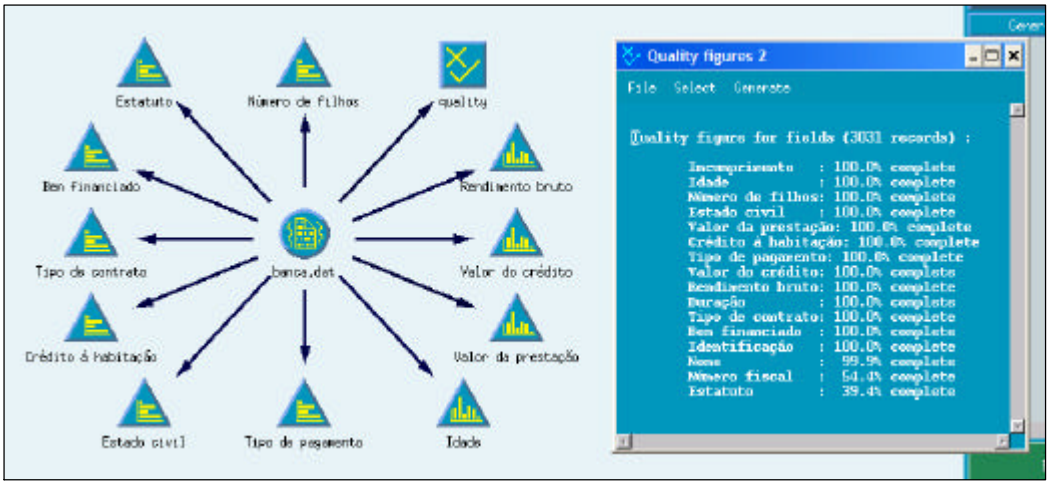


Figura 1 – Exploração dos dados

Os resultados obtidos em cada um dos nodos Distribution, utilizados para verificar a distribuição de atributos com valores categóricos, são sintetizados na Figura 2.

Figura 2 – Distribuição dos atributos categóricos

Pela análise da Figura 2 constata-se que:

- O atributo que indica se o cliente possui ou não crédito à habitação (Crédito à habitação), com os valores 1 ou 0 respectivamente, existe um registo com o valor 2, o qual deverá ser removido uma vez que representa um erro nos dados;
- No atributo Estatuto existem cinco casos de financiamento concedido a empresas, os quais não podem ser analisados em conjunto com os restantes casos de financiamento concedido a particulares. Para além do conjunto de regras que dita a concessão de financiamento a estes dois tipos de clientes ser diferente, o reduzido número de casos disponíveis para o cliente empresa também não permite que os mesmos sejam considerados na análise, e como tal têm de ser removidos da amostra.

Para os restantes atributos não foram detectadas quaisquer anomalias, apresentado os mesmos a distribuição que resulta do normal funcionamento da empresa.

No caso dos atributos com valores contínuos, a Figura 3 apresenta os histogramas que permitem analisar a distribuição dos mesmos, e definir as classes a utilizar na transformação dos atributos com valores contínuos, em atributos com valores discretos. A análise dos histogramas apresentados permitiu adoptar as classes apresentadas na Tabela 2, na transformação dos atributos com valores contínuos em atributos com valores discretos (os limites definidos para as diversas classes visam distribuir homogeneamente os dados pelas mesmas).

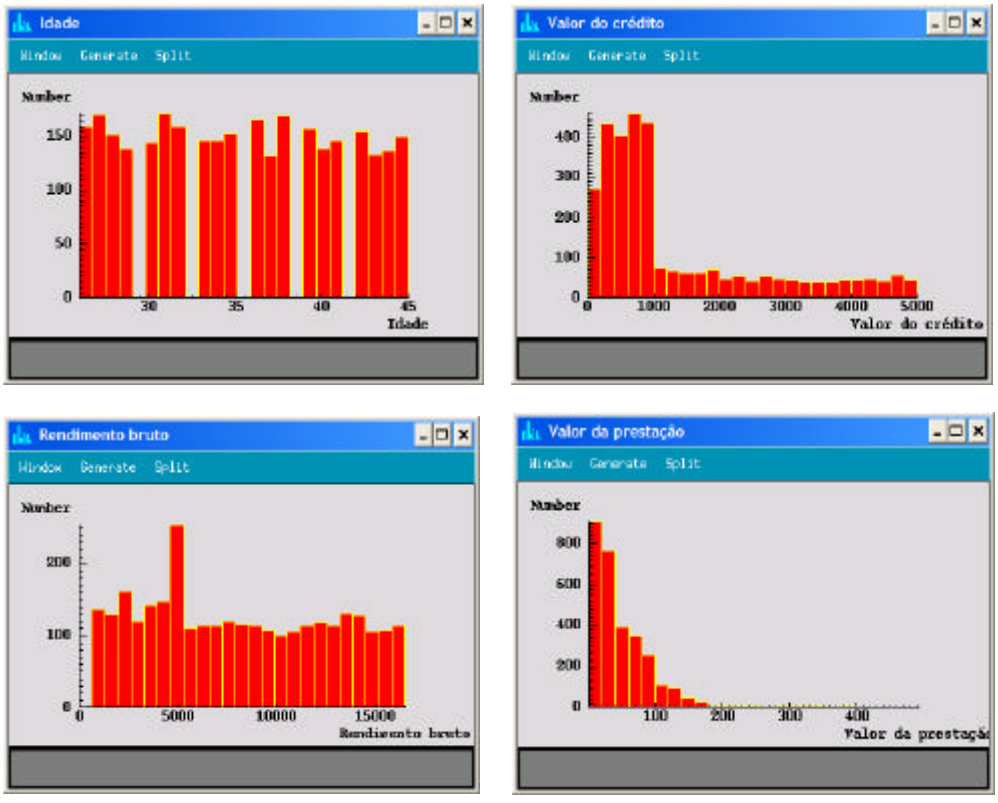


Figura 3 – Histogramas com a distribuição de atributos com valores contínuos

Tabela 2 – Classes para os atributos com valores contínuos

Atributo	Classes
Idade	(25..31] - '**26-31**', (31..38] - '**32-38**', (38..45] - '**39-45**'
Valor do crédito	(0..350] - '**0-350**', (350..650] – '**351-650**', (650..900] - '**651-900**', (900..2500] - '**901-2500**', (2500..5000] - '**2501-5000**'
Rendimento bruto	(0..4500] - '**0-4500**', (4500..8000] - '**4501-8000**', (8000..12500] - '**8001-12500**', (12500..17000] - '**12501-17000**'
Valor da prestação	(0..17] - '**0-17**', (17..30] - '**18-30**', (30..50] - '**31-50**', (50..80] - '**51-80**', (80..500] - '**81-500**'

O exercício de compreensão e exploração dos dados conduziu à identificação dos atributos a analisar e à definição das classes a utilizar na etapa de pré-processamento dos dados. As próximas subsecções apresentam as diversas fases do processo de descoberta de conhecimento, que conduziram à detecção de padrões nos dados.

Selecção e tratamento dos dados

A fase de *selecção dos dados* permite eliminar todos os atributos que não têm interesse no processo de descoberta de conhecimento. São estes a Identificação, o Número fiscal, o Estatuto e o Nome. Os restantes atributos são seleccionados, com o objectivo de avaliar a sua contribuição na determinação do perfil dos clientes.

A fase de *tratamento dos dados* consiste basicamente no tratamento de dados omissos e dados corrompidos. No exemplo em análise, apenas em dois casos foram detectadas anomalias, como já referido anteriormente, conduzindo à eliminação do registo com o valor 2 no atributo Crédito à habitação, e à remoção do valor empresa no atributo Estatuto.

A Figura 4 apresenta a *stream* construída para as fases de *selecção e tratamento dos dados*, atendendo às tarefas acima especificadas. Como resultado, é criado o ficheiro DadosTratados, com os dados a utilizar nas próximas fases do processo de descoberta de conhecimento. Nesta etapa foi ainda realizada a mudança de nome dos atributos, de forma a facilitar a utilização do CLEM (*Clementine Language for Expression Manipulation*), na construção das expressões necessárias à manipulação de dados.

Pré-processamento dos dados

Na fase de *pré-processamento dos dados* (Figura 5), os atributos com valores contínuos são transformados em atributos com valores discretos, atendendo às classes definidas na Tabela 2. Nesta fase são, ainda, utilizados nodos Web na exploração dos dados. Esta exploração permite identificar associações entre os atributos, que indiciam a relevância dos mesmos na identificação do perfil dos clientes. A última tarefa, efectuada nesta etapa, consiste na divisão aleatória dos dados em dois ficheiros, Treino e Teste, que serão utilizados na construção dos modelos que caracterizam os dados e na sua validação, respectivamente.

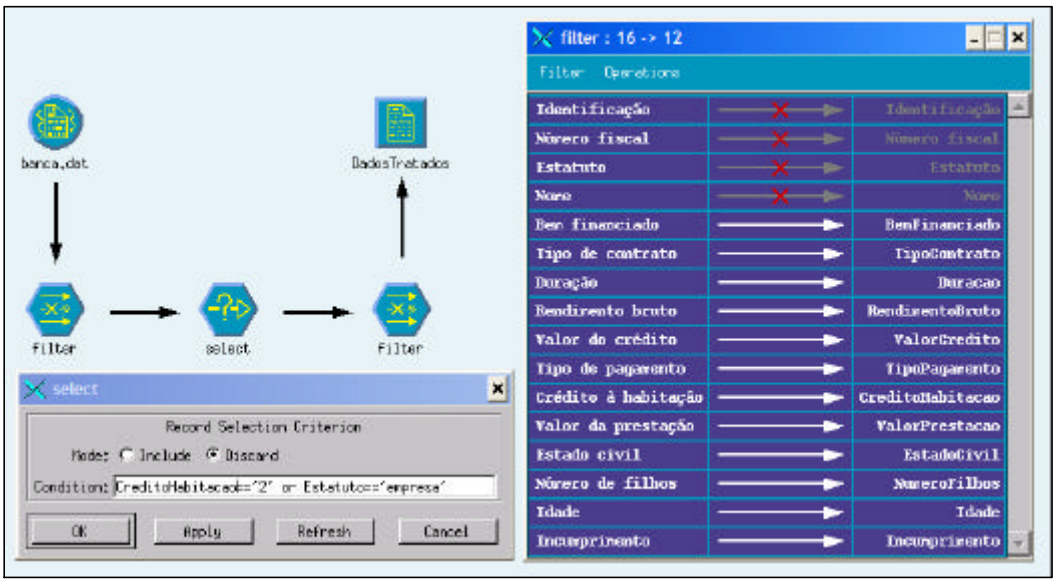


Figura 4 – Selecção e tratamento dos dados

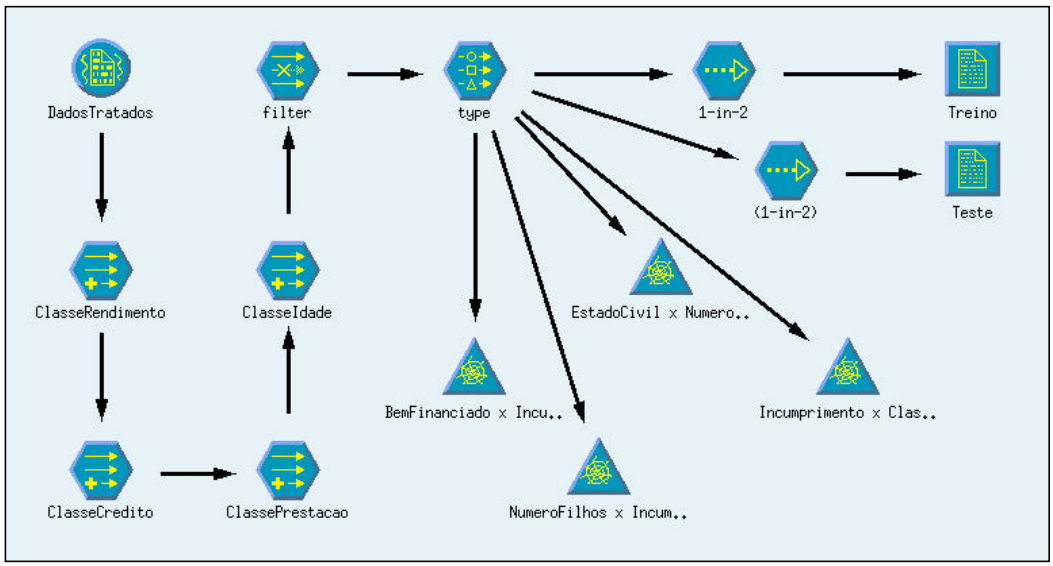
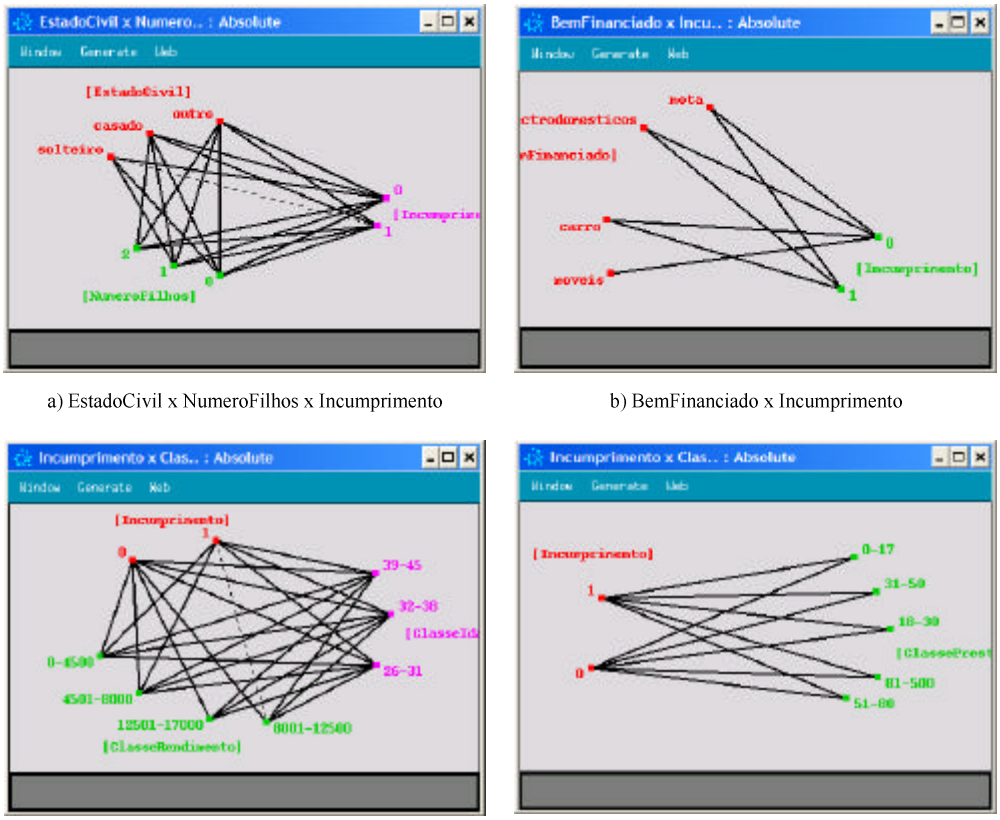


Figura 5 – Pré-processamento dos dados

Os nodos Web gerados (Figura 6) apenas permitem constatar que existe uma associação fraca entre o estado civil solteiro e o incumprimento (atributo Incumprimento com valor 1) (Figura 6 a)), e ainda, entre o Rendimento bruto caracterizado pela classe 8001-12500 e o incumprimento (Figura 6 c)). Para o Rendimento bruto caracterizado pela classe 12501-17000 não existe qualquer associação com o atributo Incumprimento. Estes três casos permitem concluir que nesta empresa de financiamento, os solteiros e as pessoas com maiores rendimentos são os mais cumpridores. No que diz respeito às associações existentes entre os bens financiados e o Incumprimento, constata-se que o bem móveis não tem qualquer associação com o Incumprimento (Figura 6 b)), salientando que não existe qualquer anomalia no financiamento deste bem. Em relação ao Valor da prestação (Figura 6 d)) não é possível tirar qualquer conclusão, uma vez que tanto o *cumprimento* como o *incumprimento*, apresentam associações fortes com as diversas classes que caracterizam os valores das prestações (representado pelo atributo ClassePrestacao).



a) EstadoCivil x NumeroFilhos x Incumprimento

b) BemFinanciado x Incumprimento

c) ClasseIdade x ClasseRendimento x Incumprimento

d) ClassePrestacao x Incumprimento

Figura 6 – Web nodes com os relacionamentos entre atributos

Data Mining

Na fase de *data mining* (Figura 7), o ficheiro Treino é utilizado na construção de dois modelos que caracterizam os dados. O primeiro, construído recorrendo ao algoritmo C5.0 (permitindo induzir uma árvore de decisão), tem como função determinar o conjunto de atributos relevante para a previsão do atributo Incumprimento, e ainda, identificar as regras que caracterizam o perfil dos clientes financiados, principalmente, dos *incumpridores*. O modelo obtido (evidenciado na mesma figura) identifica o perfil dos clientes desta organização. A análise das regras obtidas com o algoritmo C5.0 permite identificar o conjunto de atributos relevante para a caracterização do Incumprimento. São eles EstadoCivil, BemFinanciado, ClasseRendimento, NumeroFilhos, TipoContrato e ClasseIdade. Estes atributos são utilizados no treino de uma rede neuronal (nodo IncumprimentoNN na *stream* da Figura 7), que complementa o modelo obtido com o algoritmo C5.0, na previsão do atributo Incumprimento.

Uma das regras que pode ser extraída da árvore de decisão apresentada na Figura 7, e que caracteriza um conjunto de clientes catalogados como *incumpridores* é:

Se EstadoCivil='casado' e BemFinanciado='mota' e ClasseRendimento='0-4500' **Então** 1

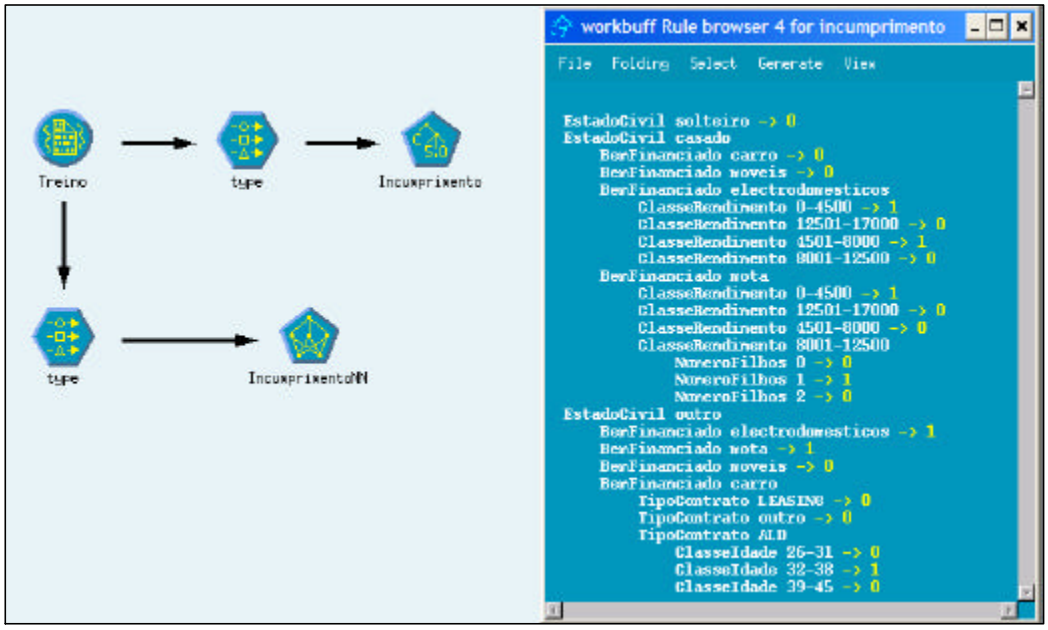


Figura 7 – Fase de *Data Mining*

Interpretação de resultados

Os modelos construídos na fase de *data mining*, recorrendo ao algoritmo C5.0 e redes neuronais, são nesta fase utilizados na classificação de casos desconhecidos, com o objectivo de avaliar o seu desempenho na previsão do Incumprimento dos clientes. O ficheiro Teste, gerado na fase de *pré-processamento dos dados*, é nesta etapa utilizado na verificação da confiança das regras encontradas na fase de *data mining*. A *stream* construída para a fase de *interpretação de resultados* é apresentada na Figura 8. Nesta figura é possível visualizar o resumo do desempenho de cada um dos modelos e ainda, o desempenho dos dois modelos se utilizados integradamente. Neste último caso, o desempenho na previsão apresenta uma percentagem de acerto de 97.44%, contra os 96.23% evidenciados pelo modelo gerado pelo algoritmo C5.0 e os 96.76% obtidos com a rede neuronal.

A utilização conjunta dos dois modelos, na previsão do Incumprimento dos clientes, permite obter resultados mais precisos. Os modelos obtidos apresentam desempenhos diferentes, que dependem do bem financiado. Esta situação, que pode ser confirmada na Figura 9, permite conhecer os bens para os quais os modelos são mais precisos, e as situações que carecem de uma análise mais detalhada ou eventualmente a inclusão de novos atributos no processo de descoberta de conhecimento. Na referida figura é possível verificar que, por exemplo, para o bem carro, a confiança na utilização do modelo obtido com o algoritmo C5.0 é de 91.39%, enquanto que o modelo obtido com a rede neuronal apresenta uma confiança de 92.83%. No caso do bem mota esta situação mantém-se. Para os electrodomésticos e móveis os modelos apresentam desempenhos semelhantes, sendo também mais precisos na previsão. Apesar da análise da figura poder induzir que o modelo obtido com a rede neuronal seria o suficiente para a previsão do Incumprimento, este não exprime (explicitamente ao utilizador, por exemplo através de regras) qualquer informação acerca dos critérios que estão na base da classificação. Tal justifica a utilização da árvore de decisão, uma vez que a mesma transmite ao utilizador os critérios utilizados na previsão do *cumprimento* ou *incumprimento* dos clientes.

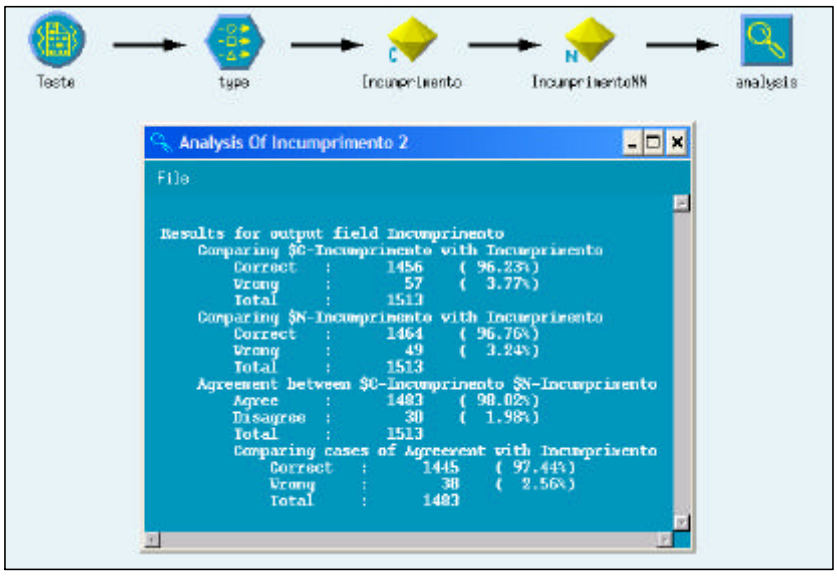


Figura 8 – Validação dos modelos encontrados na fase de *Data Mining*

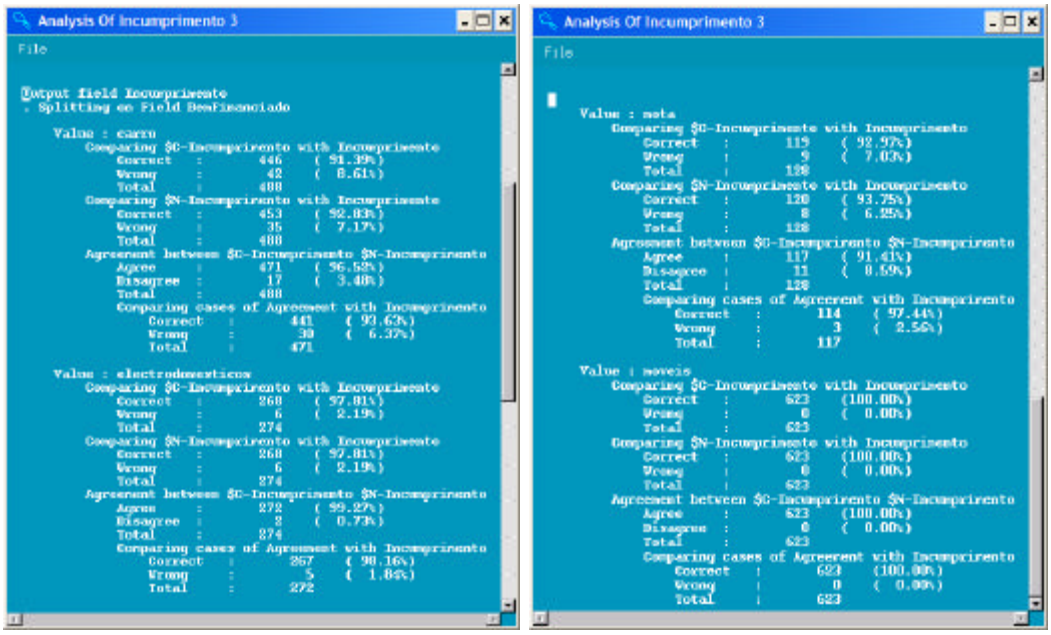


Figura 9 – Desempenho, por bem financiado, dos modelos obtidos

4. Construção de Conhecimento Organizacional

A partir da análise apresentada na secção anterior foi possível, com a ajuda de um sistema de descoberta de conhecimento definir o perfil dos clientes de uma empresa de financiamento que fornece crédito para a aquisição de bens. Para o processo, os actores organizacionais trazem o seu entendimento do negócio, riscos que se pretende minimizar, e a forma como os resultados obtidos podem ser integrados nos processos organizacionais de negociação e decisão. Todo este conhecimento prévio pode residir a um nível consciente ou não consciente, pode revestir-se de contornos subjectivos ou estar bem enraizado em entendimentos partilhados.

Conquanto não seja difícil perceber a necessidade desse conhecimento assentar em entendimentos partilhados sobre metas e objectivos organizacionais bem como práticas institucionalizadas, não deve ser ignorada a importância da experiência pessoal dos actores organizacionais bem como do conhecimento que resulta da reflexão sobre essa experiência. Se por um lado a diversidade de perspectivas e experiências pode fazer abrandar a reacção da organização a estímulos externos, ela é também a força motriz da criatividade e inovação.

No caso específico apresentado neste artigo, a ferramenta de descoberta de conhecimento permite caracterizar os clientes *incumpridores*. Será o conhecimento prévio do utilizador da ferramenta que vai permitir decidir como essa caracterização deverá condicionar decisões de atribuição de crédito a clientes específicos ou que procedimentos deverão ser institucionalizados para garantir que bons clientes não sejam rejeitados à partida.

Uma aplicação directa dos resultados de *data mining* poderia trazer impactos negativos à relação da empresa com os consumidores. Classificar potenciais clientes como *incumpridores* pela utilização de modelos resultantes dos dados armazenados pela empresa significa institucionalizar preconceitos relativos a determinados comportamentos, e situações profissionais ou da vida pessoal. Se bem que estes preconceitos possam ser fundamentados e assentes na experiência passada da organização, eles podem introduzir dificuldades na relação da empresa com o seu mercado.

Para além do conhecimento do negócio, os utilizadores da ferramenta de descoberta de conhecimento precisam conhecer os aspectos técnicos ligados a essa utilização bem como ser capazes de interpretar os resultados que lhe são apresentados. Estes são conhecimentos prévios que devem existir para que possa ser feita uma utilização eficaz da ferramenta. Quer seja por formação técnica quer por aprendizagem individual ou em grupo, é necessário que a organização assegure e valorize a construção dos conhecimentos técnicos necessários.

Para que toda a organização possa beneficiar da informação obtida no processo de descoberta de conhecimento em bases de dados, é necessário garantir condições que permitam a sua partilha. Para tal é necessário que a utilização de técnicas de *data mining* se enquadre no âmbito de construções sociais como criação ou reforço de relações de trabalho (ex: definição do processo organizacional de descoberta de conhecimento em bases de dados), construção partilhada de artefactos físicos (ex: modelos de apoio à decisão), definição de metas e projectos partilhados (ex: definição de campanhas de marketing), normas e tradições culturais (ex: participação na decisão, redefinição das regras de negociação com os clientes).

Os modelos produzidos pelo Clementine traduzem de uma forma compacta a experiência passada da organização. Esta experiência pode ser conhecida por alguns, pode estar integrada de forma não consciente nas práticas institucionalizadas, ou pode ser em grande parte desconhecida dos actores organizacionais. Desta forma, a utilização da ferramenta permite comunicar essa experiência passada (linguagem), reagir a ela (emoção), reflectir, decidir e planear (cognição) e alterar ou reforçar comportamentos de selecção de clientes (interacção). Ao construir os modelos, a ferramenta permite trazer para a memória consciente da organização a sua experiência passada contribuindo para um reforço ou reformulação da sua própria identidade e do seu papel no mercado para que se direcciona.

5. Conclusão

Este artigo procura integrar o processo de descoberta do conhecimento em bases de dados no contexto mais abrangente da criação e partilha de conhecimento organizacional. Para atingir este objectivo é, em primeiro lugar, apresentada a perspectiva das autoras sobre o processo de descoberta de conhecimento e sua integração nos processos sociais de criação de conhecimento nas organizações. Posteriormente é apresentado um caso de teste, sobre uma base de dados

fictícia, que tem como objectivo evidenciar as diferentes fases do processo de descoberta de conhecimento, e ainda, o tipo de resultados que é possível obter com as técnicas de *data mining* utilizadas.

Ao assentar a discussão num caso de teste e não num estudo de caso acabou por reduzir a capacidade para fazer evidenciar a variedade e profundidade dos entendimentos resultantes da integração dos conceitos e práticas associados ao processo de descoberta de conhecimento na perspectiva mais abrangente da construção social da realidade organizacional [Kafai e Resnick 1996]. Proximamente, as autoras esperam poder fornecer estes entendimentos através da realização de estudos de caso e *action research*. As autoras têm um projecto planeado para estudar, num hospital, a contribuição do processo de descoberta de conhecimento para a criação de conhecimento partilhado e reformulação de práticas de trabalho. O objectivo é o de definir linhas de orientação para a concepção e implementação de processos de descoberta de conhecimento em bases de dados que possam contribuir de forma clara para a eficiência organizacional.

6. Agradecimentos

Agradecemos à NTech – Sistemas de Informação, Lda., a cedência da base de dados utilizada neste trabalho.

7. Referências

Andrienko, G. L., e N. V. Andrienko, *Knowledge Extraction from Spatially Referenced Databases: a Project of an Integrated Environment*, Varenius Workshop on Status and Trends in Spatial Analysis, Sta. Barbara, CA, 1998.

Damásio, A. *O Sentimento de Si: o corpo, a emoção e a neurobiologia da consciência*, Publicações Europa-América, 1999.

Fayyad, U., G. Piatetsky-Shapiro, e P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", *Communications of the ACM*, 39, 11 (1996), 27-34.

Han, J., e M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

Holsheimer, M., e A. Siebes, *Data Mining: The search for Knowledge in Databases*, Centrum voor Wiskunde en Informatica, Technical Report, Amsterdam, 1994.

Kafai, Y. and M. Resnick, Eds. (1996). Constructionism in Practice: designing, thinking, and learning in a digital world. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Maier, R. (2002). Knowledge Management Systems: information and communication technologies for knowledge management. Berlim, Springer-Verlag.

Orlikowski, W. J. (2002). “Knowing in practice: enacting a collective capability in distributed organizing.” Organization Science 13(3): 249-273.

Ramos, I., João A. Carvalho (Aceite para publicação). "Towards constructionist Organizational Data Mining (ODM): changing the focus from technology to social construction of knowledge", Em Barko e Nemati (Eds.), *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*, Hershey, PA: Idea Group, livro a ser editado no final de 2003.

Santos, M. Y., *PADRÃO: Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas*, Tese de Doutoramento, Universidade do Minho, 2001.

SPSS, *Clementine, User Guide, Version 5.2*, SPSS Inc., 1999.