Universidade do Minho
Escola de Engenharia

Ana Cristina Wanzeller Guedes de Lacerda

Selecção de Planos de Mineração de Dados de Utilização da Web

Dezembro de 2006

Selecção de Planos de Mineração de Dados de Utilização da Web
Ana Cristina Wanzeller Guedes de Lacerda

UMinho | 2006

Universidade do Minho
Escola de Engenharia

Ana Cristina Wanzeller Guedes de Lacerda

Selecção de Planos de Mineração de Dados de Utilização da Web

Tese de Doutoramento em Informática, na Especialidade de Inteligência Artificial

Trabalho efectuado sob a orientação do
Professor Doutor Orlando Manuel de Oliveira Belo

Investigação subsidiada por:

Dezembro de 2006

À memória do meu Pai

À minha Mãe

Agradecimentos

A realização de um trabalho desta natureza é, sem dúvida, um grande desafio e, em certos momentos, chega a ser vista como uma missão impossível. Vários foram os que contribuíram, de uma forma directa ou indirecta, para que tal missão se tornasse possível. Cumpre-me pois expressar o meu reconhecimento e gratidão às pessoas e instituições cuja colaboração e apoio foram determinantes para a concretização deste trabalho de doutoramento e da respectiva dissertação.

O meu mais profundo agradecimento é dirigido ao Professor Doutor Orlando Belo, meu supervisor, não só pela orientação do trabalho de investigação, colaboração na elaboração de artigos científicos e revisão da dissertação, como pela apresentação de algumas comunicações e pelas valiosas sugestões, disponibilidade, incentivo e empenho. A sua forma rigorosa, crítica e criativa de arguir as ideias debatidas também foram determinantes para a prossecução deste trabalho. Quero ainda agradecer a oportunidade ímpar que me facultou de levar a cabo esta missão sob a sua orientação e o privilégio de um ambiente de trabalho de grande cordialidade. Em suma, as suas diversas contribuições e o apoio constante foram inquestionavelmente preponderantes para a viabilidade de tal missão.

Agradeço o apoio concedido no âmbito do Programa de Desenvolvimento Educativo para Portugal (PRODEP), através da concessão da bolsa de Doutoramento PRODEP III, acção 5.3 – Formação Avançada no Ensino Superior, Concurso Nº 02/PRODEP/2003, a qual financiou as propinas de doutoramento durante três anos e, principalmente, permitiu a dispensa integral de serviço docente pelo período de dois anos e nove meses.

Ao Instituto Politécnico de Viseu, à Escola Superior de Tecnologia e ao Departamento de Informática agradeço a oportunidade que me proporcionaram de usufruir da bolsa de doutoramento PRODEP III e o apoio financeiro auferido para inscrição em três das conferências, nas quais apresentei artigos científicos produzidos ao longo do doutoramento.

Aos amigos e colegas estou grata pela amizade e apoio que me prestaram em diferentes momentos. Ao Ronny Alves agradeço, particularmente, as indicações e trocas de ideias acerca de WUM. Cabe-me, ainda, salientar o Jorge Loureiro, igualmente em processo de doutoramento no mesmo período, pelo seu espírito empreendedor, sugestões pertinentes, agradável companhia e companheirismo demonstrado ao longo de várias missões.

Finalmente, agradeço à minha família a compreensão e apoio, manifestados ao longo deste período difícil. Ao meu marido Pedro e à minha filha Sofia devo um agradecimento especial, pelo apoio permanente e incondicional e pelos muitos momentos revitalizadores de felicidade, destacando, ainda, o papel de "super pai" que o Pedro tem desempenhado, tão exemplarmente, de forma a compensar as minhas frequentes indisponibilidades.

Resumo

Selecção de Planos de Mineração de Dados de Utilização da Web

A descoberta de conhecimento em dados de *clickstream*, relativos à interacção de indivíduos com sítios Web, está a assumir um papel, cada vez mais, preponderante, englobando uma audiência crescente de agentes de decisão ao longo da organização. A intenção subjacente reside em auxiliar as organizações a atingir as metas estabelecidas para os sítios que promovem e a maximizar as oportunidades emergentes da Web, explorando dados recolhidos, por inerência e de forma implícita, que, apesar de serem complexos e vastíssimos, constituem uma fonte extremamente rica e abrangente acerca do comportamento dos visitantes. No entanto, o desenvolvimento e a aplicação desses processos de mineração de dados são actividades que se revestem de grande complexidade, especialmente para utilizadores sem experiência e conhecimentos profundos neste domínio. Uma forma de combater este desafio consiste em proporcionar ferramentas consentâneas, capazes de assistirem os utilizadores na condução desses processos, procurando, deste modo, contribuir para a simplificação e acréscimo dos níveis de eficácia e de produtividade destas iniciativas. A estratégia defendida, para este efeito, desenrola-se em torno da gestão e reutilização, ao nível da organização, do conhecimento adquirido a partir da experiência prática, referente à resolução de problemas concretos que facultaram, no passado, processos bem sucedidos de mineração de dados de *clickstream*. O âmbito organizacional de tal estratégia visa, principalmente, fomentar um uso sinergético de recursos da organização, integrando os contributos de vários colaboradores e colocando as potencialidades deste tipo de mineração ao alcance e ao serviço de todos os seus membros, inclusive dos utilizadores mais inexperientes.

O trabalho apresentado nesta dissertação descreve um sistema fundamentado no paradigma de raciocínio baseado em casos, o qual foi concebido com o propósito de assistir os utilizadores em duas formas primordiais: (i) captura, organização e armazenamento, num repositório de casos partilhado, do conhecimento acerca de exemplos úteis e bem sucedidos de processos de mineração de dados de *clickstream*; (ii) selecção dos planos de mineração alternativos e mais adequados, para solucionar um problema específico de análise de dados neste âmbito, dada uma descrição de alto nível desse mesmo problema. O sistema proposto foi implementado através de uma aplicação Web protótipo, a ser explorada ao nível da organização, consolidando o conhecimento respeitante a exemplos de exercícios de mineração de utilização da Web, numa base de casos centralizada. O sistema integra e retira benefícios de recursos relacionados da organização, suportando uma abordagem semi-automática de aquisição de conhecimento, a partir dos seguintes tipos de origens: fontes de dados da organização; documentos normalizados em formato PMML, produzidos por ferramentas de extracção de conhecimento e representativos de actividades de mineração concretizadas; informação complementar, obtida por meio de interacção com o utilizador. No apoio à resolução de problemas, o sistema actua a partir de um conjunto de requisitos da análise e de características dos dados de *clickstream* disponíveis, e, com base no conhecimento relativo à aplicação de métodos de mineração e de outras operações, sugere planos de mineração alternativos e apropriados para os dados em causa e para o fim a que a análise se destina. Tais planos são apresentados ao utilizador através de descrições gerais, acompanhadas por informação suplementar e por referências para detalhes explicativos da sua implementação pragmática.

Abstract

Web Usage Data Mining Plans Selection

Discovering knowledge from clickstream data, related to the interaction of individuals with Web sites, is playing an increasingly important role, reaching a growing number of decision makers across the organization. The intention behind this is helping organizations to achieve the goals of the promoted sites and to maximize the latent opportunities of the Web, exploring data inherently and implicitly collected, which are huge and complex, yet a very rich and comprehensive source of visitants' behavior insights. However, developing and applying such mining processes are very complex tasks, especially to users without deep knowledge and experience in this domain. One way to tackle this challenge is by building tools, capable of assisting users within such processes realization, in order to simplify these initiatives and to increase theirs efficacy and productivity levels. The defended strategy regarding such assistance relies on managing and reusing, at corporative level, the knowledge acquired from the practical experience in solving concrete problems, which had provided successful clickstream data mining processes in the past. This corporative-wide perspective mostly aims at favoring an synergetic use of the organization resources, bringing up together the contributions of distinct collaborators and making available the potentialities of this kind of mining to all members, including the inexperienced users.

The work presented in this dissertation describes a system founded on the case based reasoning paradigm. This system was devised with the purpose of assisting users in two main ways: capturing, organizing and storing, on a shared case repository, the knowledge about successful and useful clickstream data mining processes; selecting the most suited and alternative mining plans, to solve a specific clickstream data analysis problem, given an high level description of such problem. The proposed system was implemented as a prototype Web-based application, to be

explored at corporate level, consolidating the knowledge about Web usage mining processes examples on a centralized case base. The system integrates and takes advantage from related corporative resources, supporting a knowledge acquisition semi-automated approach from the following types of origins: corporative data sources; standard documents in PMML format, supplied by knowledge extraction tools and representing the mining activities accomplished; complementary information, obtained through user interaction. When advising problem solving, the system acts, taking the characteristics of the available clickstream data and the analysis requirements, and based on the acquired knowledge about applying data mining and other operations, suggests the most appropriate alternative mining plans to the data and the analysis at hands. The plans are deployed as overviews, complemented by additional information and by links to practical implementation details.

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 A *World Wide Web*

A influência proeminente da *World Wide Web* no quotidiano do mundo contemporâneo e na dita revolução digital é um facto incontestável. A Web é vista por muitos como um manancial para obter resposta a todas as questões e capaz de assegurar, mais agilmente, a realização de várias das actividades diárias da população mundial. A imensidão de informação que aloja e a profusão de bens e serviços que disponibiliza, a qualquer indivíduo munido de um simples navegador Web, vulgo "browser", não são com certeza alheios a esta realidade. Também não é de estranhar que a promoção de sítios Web pelas organizações seja encarada, com insistência crescente, como uma forma promissora de criar novas perspectivas e oportunidades de negócio. Se do lado de alguns tipos de instituições, como, por exemplo, certos organismos da Administração Pública, se anseia pela redução de custos operacionais e pela eficiência e desburocratização de serviços a prestar ao cidadão comum, outras motivações de índole mais globalizada e estratégica, ou mesmo comercial, justificam esta premência por parte das organizações empresariais. Não obstante, muitas organizações, de todos os tipos e dimensões, sentem que é imprescindível possuir um sítio Web, para que possam ser aceites e entendidas como organismos profissionais.

A facilidade e rapidez que caracterizam as transacções desenvolvidas através da Web têm contribuído, de modo significativo, para o acelerado crescimento da oferta de serviços diversificados no seu seio, cuja utilização está à distância de um simples "clique", num dos muitos

ambientes sofisticados que os navegadores disponibilizam aos seus utilizadores. A Web representa, indubitavelmente, o exemplo mais autêntico da actual economia à escala mundial, a qual permite uma exposição permanente ao nível planetário e, consequentemente, retirar todos os benefícios que daí advém, mas também padecer com os efeitos de uma competitividade, igualmente global e deveras aguerrida. O excesso da oferta e a sobrecarga da Web contribuem para que o mesmo simples "clique" possa ser, também, um caminho fácil de saída do sítio. No actual estado de maturidade atingido pela Web e com tanta variedade de informação, produtos, serviços e opções alternativas, os utilizadores sentem-se muitas vezes desorientados e tendem a ser progressivamente mais exigentes e divergentes, fazendo com que a utilização real de um sítio não decorra exactamente como se previa ou almejava. Neste cenário, facultar um sítio Web não é suficiente, nem é tão pouco uma incumbência simples. A complexidade do desenho, implementação e administração de sítios Web aumenta constantemente, em sintonia com o vertiginoso acréscimo da utilização da Internet. Cativar a atenção dos visitantes e suportar serviços, atendendo a requisitos, cada vez, mais rebuscados, são desafios difíceis de vencer, num ambiente fértil, no qual também imperam a heterogeneidade, o dinamismo e a imediaticidade, para além da influência dos sucessivos avanços tecnológicos. Adicionalmente, o investimento, na construção e contínua manutenção de um sítio Web, corresponde a um esforço avultado de recursos financeiros e humanos que, pelo menos a médio prazo, se repercute em elevados níveis de expectativas.

A análise dos dados que reflectem os referidos "cliques", vulgarmente designados dados de *clickstream* ou de utilização, surge como um meio indispensável, não só para avaliar a eficácia das iniciativas e das estratégias subjacentes, como também para identificar formas mais responsivas e pró-activas de tirar partido das imensas oportunidades emergentes da Web. Estes dados são registados, por inerência, nos históricos de servidores Web, podendo ser também explicitamente recolhidos, de maneira optimizada, aproveitando as funcionalidades construídas para assegurar os serviços prestados. Trata-se de uma fonte inesgotável, muito rica e abrangente, acerca do comportamento dos visitantes, notavelmente, sem precedentes comparáveis, apesar de também ser denotada como imperfeita, aquando no seu estado bruto. Os dados de *clickstream* têm a virtude de captar as acções de navegação, na interacção de indivíduos com sítios Web, não se consignando apenas às transacções efectivadas, mas cobrindo também muitos aspectos relativos à totalidade do processo de interacção.

A exploração de dados de *clickstream*, por intermédio de ferramentas de mineração de dados – *Data Mining* (DM) –, pode ajudar a derivar conhecimento efectivamente relevante para as

organizações. Está-se, claramente, a indicar a valia do processo de extracção de conhecimento, neste caso, a partir de dados de *clickstream*, a fim de lidar com o problema da transformação de dados complexos e vastíssimos, em novos factos e respostas para questões elaboradas e estratégicas, com valor inestimável para os agentes de decisão. Este tipo concreto de processo adquiriu um estatuto próprio, sendo denominado, especificamente, mineração de utilização da Web – *Web Usage Mining* (WUM). A WUM visa descobrir padrões de utilização interessantes, que possam ajudar as organizações a entender e melhor servir as necessidades dos visitantes, com vista a atingir os propósitos e metas dos seus sítios Web, bem como a maximizar as potencialidades inerentes da Web.

A tendência actual de globalização do acesso à informação, e a tecnologias consentâneas, contribui para a eliminação de obstáculos que, anteriormente, restringiam o uso de ferramentas de DM a utilizadores com um nível razoável de conhecimentos sobre o seu uso. Esta tendência, aliada às grandes expectativas depositadas em torno das capacidades da DM e, também, à indispensabilidade da constante optimização, a diversos níveis, dos serviços prestados, apontam para um aumento e vulgarização do recurso a estas ferramentas, permitindo que cada utilizador dentro da organização possa ser, em termos gerais, um analista informal. Os benefícios que podem ser retirados, recorrendo a estas ferramentas, relacionam-se com diversos problemas de decisão, envolvendo diferentes tipos de analistas, tipicamente, exemplificados pelas figuras do administrador do sítio e do especialista (analista) de negócios. Porém, as ferramentas de DM e WUM e os paradigmas subjacentes são demasiado complicados para serem usados sem o auxílio de especialistas na área. Os impedimentos à ampla exploração da WUM contrastam com a abrangência da sua aplicação e utilidade ao longo da organização, criando um cenário de exclusão, o qual não se coaduna com os tempos actuais. Neste sentido, urge inverter esta situação, procurando as formas que poderão contribuir para uma utilização mais facilitada da WUM, colocando as suas potencialidades ao alcance e ao serviço de todos aqueles que as possam vir a usufruir, em prol dos mesmos, de outros membros da instituição e da própria organização.

1.2 Motivação do Trabalho

A extracção de conhecimento é uma área em constante desenvolvimento, quer no que se refere às técnicas e ferramentas de DM, quer às tecnologias complementares que contribuem para a persecução de projectos neste âmbito. Esta área encontra-se também em franca expansão, sendo usada em inúmeros domínios de aplicação e na resolução de problemas de decisão, cada vez mais

sofisticados. Apesar dos sucessivos progressos tecnológicos e do aprofundamento de competências dos peritos, a concretização de projectos de descoberta de conhecimento continua a ser muito complicada e a exigir um conjunto bastante diverso de conhecimentos, acções e decisões. Um dos muitos desafios que emerge neste processo consiste na selecção dos métodos de DM mais apropriados, a aplicar na resolução de um determinado problema de análise de dados, de modo a se obter resultados com utilidade para um fim específico. Esta escolha tem, naturalmente, grande impacto nos resultados alcançados, não existindo critérios simples e gerais que suportem sistematicamente tal decisão, já que a adequação das técnicas de DM varia em função de diferentes tipos de factores, entre os quais alguns complexos (e.g. pressupostos dos métodos em termos de características dos dados) e mesmo subjectivos (e.g. simplicidade na interpretação de resultados) e, por vezes, conflituosos (e.g. nível de precisão versus facilidade de interpretação de resultados).

O desenvolvimento e aplicação de processos de WUM não é uma excepção, pelo contrário, comporta dificuldades acrescidas. Antes de mais, os dados de *clickstream* além de muito ricos, são também problemáticos, devido a aspectos como as suas imperfeições, complexidade intrínseca e grande volume e dimensionalidade. Os problemas típicos de WUM e as formas pelas quais estes podem ser solucionados têm vindo a ser discutidos, mas ainda não foram suficientemente estudados e estruturados. Pode-se mesmo afirmar que os vectores subjacentes, sob os quais assenta a WUM, se encontram em constante evolução, fazendo com que este tipo de mineração surja como um alvo móvel e difícil de disciplinar. Evidencia-se ainda a maior pressão de eficácia e celeridade nestas iniciativas, pois, por um lado, os dados de *clickstream* são uma fonte inesgotável, carecendo de um tratamento metódico contínuo e, por outro lado, as descobertas são facilmente accionáveis e, usualmente prementes, quando está em causa a Web. Perante o aumento da importância e uso da WUM, englobando uma audiência estendida de analistas, com vários níveis de conhecimentos na área, a assistência na condução destes processos ganha maior preponderância. Esta assistência também assume contornos distintos e conjuga novas dimensões, tais como as características do sítio Web subjacente. Para além de se acentuar a pertinência de abstracção de complexidade a múltiplos níveis, a WUM reveste-se de características específicas, as quais requerem um tratamento dirigido, especialmente devotado aos tipos de problemas que se podem levantar neste âmbito.

A actividade de descoberta de conhecimento, nomeadamente a sua essência exploratória e interactiva, entre vários outros aspectos, condiciona o tipo de apoio que pode ser prestado à sua persecução. As metodologias de desenvolvimento destes exercícios concedem linhas de orientação

úteis, para situações gerais e abstractas, remetendo para o analista os detalhes da sua implementação para situações concretas. Contudo, é precisamente em torno destes detalhes que surgem os maiores obstáculos a uma extracção de conhecimento mais expedita. A necessidade de um nível elevado de abstracção coloca-se, realmente, mas em função dos requisitos específicos do analista, tendo em mente os problemas a resolver, sob o ponto de vista prático da actividade ou organização, em detrimento de uma perspectiva voltada para as questões de índole mais técnica. O grau de abstracção das metodologias também cria uma distância demasiado acentuada, em relação a situações particulares e ao próprio ambiente das ferramentas aplicáveis, a qual é proeminente esbater. De facto, constata-se frequentemente que o sucesso obtido nestes processos depende fortemente da experiência e conhecimento adquirido ("*know-how*"), detidos pelos especialistas na área, mais do que qualquer outro aspecto, como, por exemplo, a metodologia usada ou as noções de carácter técnico. Efectivamente, no âmbito da DM e, principalmente, da WUM os problemas reais são, normalmente, bastante complexos e possuem muitas particularidades. Os próprios peritos na área não conseguem definir um conjunto de regras ou princípios gerais e consistentes para sustentar a resolução de problemas.

Uma forma de lidar com os desafios referidos consiste em construir um repositório ao nível da organização, capaz de manter e gerir o conhecimento acerca da experiência obtida a partir de processos de WUM bem sucedidos. A documentação explícita de experiências conduz à estruturação e memorização do conhecimento adquirido em situações passadas. Outra vantagem de tal repositório advém da possibilidade de partilhar e reutilizar o seu conhecimento ao longo da organização e, naturalmente, dos benefícios decorrentes de uma utilização mais eficiente e sinergética dos recursos organizacionais. Um repositório desta natureza, concebido de acordo com as necessidades e requisitos da organização, pode actuar como uma memória consolidada, com carácter didáctico e conducente à adopção de boas práticas, no que diz respeito à exploração de dados de *clickstream*. O acesso a este repositório poderá permitir que outros analistas, para além dos especialistas, retirem proveito das capacidades da WUM. Esta abordagem de assistência requer, no entanto, um mecanismo ágil para ajudar os analistas a relacionar os novos problemas com as soluções promissoras existentes.

O paradigma de raciocínio baseado em casos – *Case Based Reasoning* (CBR) – vem ao encontro do tipo de solução defendida para assistir exercícios de WUM, pois pode proporcionar um suporte consentâneo à operacionalização do repositório de conhecimento, relativo a experiências de desenvolvimento de processos de WUM. Este paradigma é uma abordagem de aprendizagem e de resolução de problemas [Kolodner 93] [Aamodt e Plaza 94], que coloca especial ênfase no papel

da experiência [Mantaras et al. 05], representada sob a forma de casos. O CBR inspirou-se no tipo mais comum de raciocínio humano, simulando o comportamento humano na resolução de problemas da vida real, de modo ainda mais sistemático [Kolodner 91]. Em termos gerais, um novo problema é solucionado encontrando um caso prévio similar e reutilizando-o na situação actual. O novo problema resolvido pode ser retido, viabilizando uma aprendizagem incremental sustentada. Por conseguinte, o paradigma vem dar resposta a requisitos essenciais de um sistema para apoiar iniciativas de WUM, com base em experiências prévias. O recurso ao paradigma CBR permite o seu funcionamento como um sistema de gestão e reutilização de casos de WUM bem sucedidos, favorecendo, não só a estruturação do conhecimento para efeitos de reutilização, como também a aplicação deste conhecimento em situações semelhantes, e propiciando a extensibilidade do sistema, face à capacidade inerente de aprendizagem.

Com base no exposto, a motivação do trabalho desenvolvido e da presente dissertação tem, como ponto de partida, a reconhecida dificuldade em estabelecer planos de WUM apropriados e a relevância de assistir este tipo de exercícios, desenrolando-se em torno de uma abordagem, entendida como adequada, para lidar com tal dificuldade e especificamente devotada a este âmbito, para colocar a WUM ao serviço de utilizadores não especialistas nesta área.

1.3 Objectivos do Trabalho

Os objectivos estratégicos do trabalho realizado no âmbito desta dissertação residem em contribuir para a simplificação e acréscimo dos níveis de produtividade e de efectividade, em iniciativas de desenvolvimento e aplicação de processos de WUM, assistindo o analista na escolha e utilização dos métodos conducentes à resolução de um problema específico de análise de dados, considerando a natureza desse problema e as particularidades do seu contexto. A prossecução destes objectivos envolve a concepção, desenvolvimento e implementação de um sistema de selecção de planos de WUM, capaz de alcançar as seguintes metas:

- Viabilizar a aquisição, partilha e reutilização do conhecimento adquirido na resolução de problemas de WUM, promovendo a adopção de práticas efectivas na organização, no que se refere à exploração de capacidades de mineração sobre dados de utilização da Web.
- Suportar a formulação de problemas de WUM, abstraindo a complexidade inerente e auxiliando o utilizador a estabelecer a correspondência entre o problema actual e as definições existentes de tipos de problemas, dando prioridade às necessidades de utilizadores sem experiência neste domínio.

- Sugerir soluções plausíveis para solucionar problemas no âmbito da WUM, seleccionando o conjunto de soluções mais ajustadas ao problema actual, discriminando abordagens alternativas, quanto a métodos e ferramentas disponíveis, e conjugando critérios para apoio na escolha entre essas abordagens.

Com base nas metas anteriores, é possível traçar um conjunto de objectivos, com um carácter mais pragmático, em torno das actividades de investigação e de desenvolvimento associadas. Estes objectivos consistem nos seguintes:

- Identificar os aspectos ou vectores que melhor caracterizam os problemas de WUM.
- Discriminar as categorias de elementos que melhor descrevem e explicam a aplicação de métodos de mineração e a realização de outras actividades, inseridas no processo de extracção de conhecimento.
- Identificar os diferentes tipos de factores com maior influência na selecção de métodos e abordagens de DM, em termos gerais e no âmbito específico da WUM.
- Conceber um modelo conceptual capaz de suportar a representação adequada e detalhada de processos de WUM, no que concerne a problemas de análise de dados e às respectivas soluções, combinando vocabulário e orientações relacionados com o paradigma CBR, domínio da sua aplicação e padrões da área.
- Implementar um repositório de casos de desenvolvimento e aplicação de processos de WUM, de acordo com o modelo conceptual de representação concebido e recorrendo às tecnologias de gestão de dados mais usuais em ambientes organizacionais.
- Desenhar e implementar um sistema de selecção de planos de WUM, fundamentado na exploração do paradigma CBR e com capacidade para realizar as seguintes funções:
    - Caracterizar conjuntos de dados em estudo, com base nas propriedades mais relevantes, para efeitos de selecção de métodos de mineração, a partir da especificação da fonte do conjunto de dados e de informação complementar.
    - Auxiliar a formulação de problemas de WUM, recolhendo e sistematizando os requisitos da análise, respeitantes a restrições explícitas, conducentes ao refinamento da descrição desse problema.
    - Recuperar os casos prévios mais promissores, para coadjuvar a resolução do problema actual, e construir recomendações de estratégias de mineração de dados, em função desses casos.
    - Conciliar descrições de processos de WUM, fornecidas via interacção com o utilizador, com representações normalizadas de resultados de mineração, produzidas por ferramentas de DM ou de WUM, no sentido de integrar formas

expeditas de aquisição de conhecimento acerca de processos previamente desenvolvidos.

- Reter os novos casos de aplicação de WUM, suportando as formas alternativas e viáveis de submissão de descrições, imprescindíveis para garantir a aquisição completa de todos os elementos requeridos, e, ainda, a organização e registo dos mesmos.

- Suportar a exploração do sistema em ambiente distribuído, assegurando flexibilidade na acessibilidade e na interacção com o utilizador.
- Optimizar os serviços de interacção, a fim de garantir as funcionalidades essenciais, com a robustez requerida, entre as quais se salientam:
  - a especificação do problema de análise de dados, o mais simplificada e fidedigna possível;
  - a descrição de processos de DM já concretizados, viabilizando a introdução facilitada de dados, designadamente, quando meios mais ágeis para esse efeito não estão disponíveis.
- Preparar uma amostra de casos, baseada em conjuntos de dados, necessidades e processos de WUM reais, com vista a aferir a efectividade e pertinência do sistema.

1.4 Estrutura do Documento

O conteúdo deste documento foi organizado em seis capítulos. O primeiro e presente capítulo apresenta o contexto em que se insere o trabalho, as motivações que justificam a sua elaboração, os objectivos almejados e considerados ao longo da sua concretização e, ainda, a estrutura deste documento. No *capítulo 2, Exploração de Dados de Clickstream*, introduz-se os aspectos envolvidos no estudo deste tipo peculiar de dados, contemplando, especialmente, o tema da mineração de utilização da Web. Este capítulo visa enquadrar a WUM, já que o trabalho desenvolvido incide, justamente, em torno desta actividade e da assistência na sua condução. O *capítulo 3, Assistência à Mineração de Dados*, centra-se em iniciativas de apoio à descoberta de conhecimento, englobando uma breve análise das categorias de vertentes subjacentes nestas iniciativas e, ainda, a caracterização do tipo de abordagem adoptada neste trabalho, com o propósito de a posicionar e de estabelecer a base de suporte aos capítulos subsequentes.

Os dois capítulos seguintes são devotados ao sistema proposto – o sistema *Selector de Planos de Mineração* (SPM) –, cuja finalidade é assistir o utilizador no desenvolvimento e aplicação de

processos de WUM. O *capítulo 4, Sistema Selector de Planos de Mineração*, descreve a arquitectura funcional do sistema e os seus módulos constituintes fundamentais, dando particular ênfase aos mecanismos de raciocínio do sistema e às tarefas de apoio na resolução de problemas e na aprendizagem a partir de experiência, sobretudo, sob o ponto de vista do modo como estas são tratadas e levadas a cabo pelo sistema. O *capítulo 5, Implementação e Demonstração do Sistema SPM*, é dedicado à perspectiva da implementação do sistema, materializando os módulos concebidos e as ideias defendidas, com o intuito de comprovar a viabilidade pragmática do sistema proposto. Este capítulo cobre os aspectos relacionados com as características essenciais do protótipo do sistema e das suas operações primordiais e, seguidamente, ilustra a forma pela qual o sistema actua, à luz de exemplos concretos demonstrativos.

Por último, o *capítulo 6, Conclusões e Trabalho Futuro*, finaliza a dissertação. Após uma síntese do trabalho desenvolvido, centrada no percurso estabelecido desde o problema investigado até à solução proposta, avaliam-se os resultados obtidos. Em seguida, enumeram-se os contributos do trabalho e apresentam-se as principais conclusões retiradas sobre este, sob a forma de considerações finais. A terminar o capítulo, delineiam-se as direcções de investigação e trabalho futuro, para dar continuidade às actividades encetadas e realizadas ao longo desta dissertação.

Este documento inclui, ainda, um anexo, *Anexo A – Exemplificação do Cálculo de Similaridade entre Casos de Aplicação de WUM*, no qual se procede à demonstração detalhada do procedimento de estimativa do nível de similitude entre dois casos concretos de aplicação de mineração de utilização da Web, recorrendo, para o efeito, a um dos exemplos introduzidos no capítulo 5 (secção 5.3.1).

Capítulo 2

Exploração de Dados de *Clickstream*

2.1 Ambiente Analítico Genérico

A quantidade de dados acumulados pelas organizações cresce velozmente, na sequência de sucessivos progressos tecnológicos, da redução de custos de armazenamento e da tendência instalada de informatização de operações, visando o acréscimo da sua eficiência. Estes dados alcançam volumes, outrora inimagináveis, cuja magnitude ultrapassa as capacidades de análise e síntese fundamentais para tomar decisões informadas, em tempo útil, e num ambiente de actividade, cada vez, mais turbulento. Ao mesmo tempo, sabe-se que um imenso volume de dados, associados à actividade da organização, equivale a um elevado potencial de informação e conhecimento, e estes, por sua vez, são, naturalmente, muito valorizados. O suporte à decisão, neste contexto, coloca um conjunto vasto de desafios, estimulando avanços em variados domínios e conduzindo à subdivisão dos sistemas de informação em duas categorias essenciais, para, assim, atender adequadamente a tipos de requisitos substancialmente distintos [Codd et al. 93] [Chaudhuri e Dayal 97]: os sistemas operacionais, destinados a assegurarem o funcionamento das aplicações tradicionais de processamento transaccional; os sistemas de apoio à decisão, dedicados às solicitações analíticas. A Figura 1 ilustra o ambiente típico de suporte à exploração de dados, segundo [Chaudhuri e Dayal 97].

A necessidade da separação dos ambientes operacional e de suporte à decisão conduziu muitas organizações à implementação de sistemas de armazéns de dados – *Data Warehousing* (SDW). Um

dos alicerces dos SDW reside em repositórios de dados integrados, orientados por assuntos e apenas de leitura, mantidos à parte das várias fontes ou dos sistemas operacionais da organização, a partir dos quais os dados são obtidos, tipicamente para o propósito específico de análise. Tais repositórios de dados são vulgarmente designados armazéns de dados – *Data Warehouse* (DW). Já os *data marts* denotam subconjuntos lógicos e físicos de DW mais abrangentes, frequentemente ao nível mais atómico de dados possível, sendo passíveis de conjugação, por meio de dimensões conformes. As vantagens imediatas da implementação de um SDW são a consolidação de informação, relativa a períodos de tempo estendidos e proveniente de diversas fontes heterogéneas, com o intuito de satisfazer as necessidades particulares dos agentes de decisão, normalmente muito consumidoras de recursos computacionais, sem degradar a capacidade de resposta dos sistemas operacionais. Os SDW são vocacionados para lidar com um elevado volume de dados, estão associados a ferramentas de gestão de dados (e.g. limpeza e povoamento de estruturas de dados), sustentam ferramentas tradicionais de análise e síntese de dados (e.g. interrogação e elaboração de relatórios) e disponibilizam, ainda, o ambiente ideal a ferramentas mais elaboradas de exploração de dados, como os sistemas de processamento analítico em tempo real – *On-Line Analytical Processing* (OLAP).

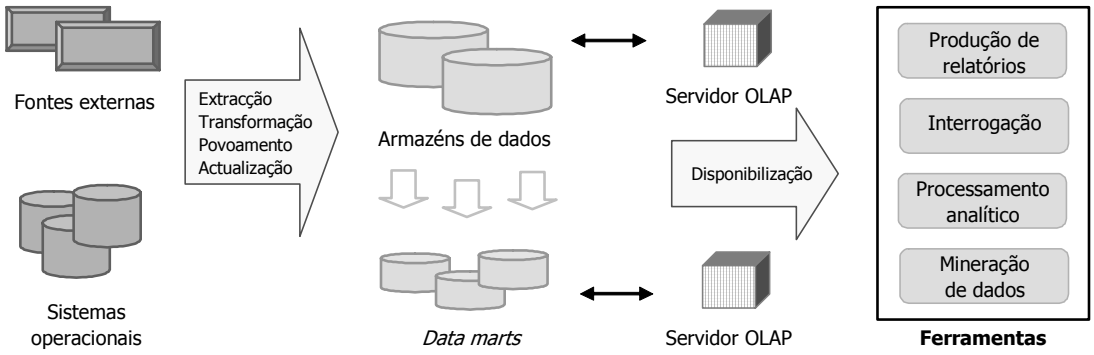


Figura 1 – Ambiente típico de suporte à exploração de dados

O processamento analítico tem emergido como um paradigma poderoso, na análise estratégica de SDW e, sobretudo, em ambientes de negócio, tendo surgido com o objectivo de suprir as limitações impostas pelas linguagens de interrogação e pelos sistemas de bases de dados convencionais, na organização e acesso aos dados. O conjunto de necessidades funcionais e de desempenho das ferramentas OLAP requer estruturas multi-dimensionais de cubo, ou hipercubo de

dados, cujas células contêm instâncias de valores correspondentes ao cruzamento das várias medidas (e.g. unidades vendidas) e dimensões (e.g. data da vendas). Para além dos valores das medidas, determinados ao nível de detalhe considerado para as dimensões, o cubo materializa agregações sucessivas das mesmas, garantindo, assim, rapidez na resposta às consultas mais frequentes. As agregações podem ser calculadas pelo resumo de linhas e colunas do cubo e em função de hierarquias definidas para as dimensões. O cubo faculta grande flexibilidade para manipular os dados, permitindo a aplicação de operações OLAP, tais como, a selecção e projecção de dados (*slice-and-dice*), agregação de dados (*drill-up/roll-up*), detalhe de dados (*drill-down/roll-down*) e rotação do cubo (*pivoting*), a fim de visualizar os dados em múltiplas perspectivas e níveis de abstracção e, ainda, de calcular estatísticas ao longo de diferentes dimensões e níveis conceptuais. A capacidade de processamento de grandes quantidades de dados, aliada a um ambiente altamente interactivo, amigável e baseado numa interacção gráfica destas ferramentas, facilita significativamente a compreensão dos dados e das respectivas relações e tendências.

As ferramentas OLAP oferecem funcionalidades analíticas poderosas, céleres e flexíveis. Porém, os resultados produzidos com elas, usualmente, não são passíveis de generalização, nem concedem as vantagens estratégicas adicionais que as organizações anseiam obter a partir dos dados. Os agentes de decisão ambicionam descobrir novos factos e respostas para questões, ainda mais elaboradas e prementes, com o intuito de detectar novas oportunidades de negócio, traçar estratégias e tomar decisões sustentadas e atempadas, perante situações diversas e, possivelmente, inesperadas. Esses factos e questões prendem-se, por exemplo, com a antecipação de necessidades dos clientes ou com a previsão da sua insatisfação, recorrendo, para tal, aos dados do passado, de modo a reagir convenientemente, por meio de recomendações que possam interessar ao cliente e de incentivos para tentar evitar a perda do cliente. Neste sentido, torna-se indispensável estudar os dados históricos, procurando factos e respostas para questões, as quais, geralmente, não foram especificamente contempladas durante a recolha de dados, nem tão pouco estão explicitamente disponíveis nos dados.

2.2 Mineração de Dados e Extracção de Conhecimento

A mineração de dados – *Data Mining* (DM) – consiste, basicamente, no processo semi-automático de exploração de grandes quantidades de dados, com o propósito de descobrir padrões úteis e significativos para as organizações. A ideia subjacente que fundamenta a sua aplicação é a existência de informação implícita e valiosa nos dados acumulados pelas organizações, a qual não

é possível antever, nem identificar por meio de inspecção ou de análises superficiais desses dados. A crescente atenção em torno da DM deve-se às muitas aplicações práticas e potencialidades da extracção de informação útil (conhecimento), a partir de dados muito vastos, e, também, à premência de assistir essa tarefa, através de técnicas e ferramentas capazes de maximizar o nível da sua automatização, na medida do possível. Outro factor que contribui, em muito, para este interesse é a proliferação de métodos oriundos de diferentes domínios, como a aprendizagem automática, entre outros, cuja aplicação tornou a descoberta de conhecimento em dados viável e atractiva. Em suma, o objectivo subjacente é a plena exploração das potencialidades das imensas quantidades de dados acumulados, aproveitando as capacidades consentâneas para esse efeito.

O processo de DM tem recebido várias denominações, nem sempre consideradas equivalentes, nomeadamente, extracção ou descoberta de conhecimento em bases de dados – *Knowledge Discovery in Databases* (KDD). [Fayyad et al. 96] definem a KDD como o processo não trivial de identificar, em dados, padrões válidos, previamente desconhecidos, potencialmente úteis e compreensíveis. A DM corresponde à etapa central da KDD, sendo responsável pela aplicação de algoritmos para procurar padrões nos dados, enquanto a KDD é uma área mais ampla, cobrindo também outras actividades cruciais. Independentemente da designação adoptada, que nesta dissertação se usa indiferenciadamente, existe um consenso de que se trata de um processo complexo e subjectivo, levado a cabo de modo iterativo e interactivo, não sendo, portanto, passível de automatização completa. A sua persecução envolve diversas etapas, nas quais é requerida a participação do analista, por meio de decisões e acções. Estas podem levar ao avanço para uma etapa seguinte ou ao retorno para uma fase anterior e ao seu refinamento, com vista a optimizar os resultados correntes, ou, mesmo, a enveredar por outras direcções de análise, entretanto equacionadas. A Figura 2 ilustra as etapas básicas da KDD ou DM, que podem ser descritas da seguinte maneira:

1. **Compreensão** (problema e dados) – O entendimento do problema prende-se com a sua percepção, sob o ponto de vista da actividade (negócio), e com a sua tradução num problema de DM. Esta tarefa engloba também a identificação dos dados preponderantes para o problema em causa, da meta de DM e de critérios de avaliação a observar ao longo do processo. A compreensão dos dados contempla a imprescindível prospecção preliminar dos mesmos, para familiarização com o seu conteúdo, propriedades e nível de qualidade e, ainda, para determinar direcções plausíveis de estudo.
2. **Transformação** – Genericamente, esta fase consiste no melhoramento e na conversão dos dados, provenientes das várias fontes, nas abstracções apropriadas para a descoberta de padrões. Assim sendo, inclui acções como a transferência e integração de dados das

várias fontes e a limpeza, filtragem, projecção e redução dos dados que serão alvo de mineração. Estas acções são realizadas diversas vezes durante o estudo e visam, essencialmente, incrementar a qualidade dos dados, proporcionar as variáveis e registos que melhor representam o problema em questão, bem como atender a propriedades e pressupostos específicos, em que se baseiam os métodos de mineração a serem aplicados.

3. **Modelação** – Nesta etapa decorrem as seguintes tarefas:
    - eleição de funções de DM (e.g. agrupamento) e selecção entre os respectivos modelos (ou algoritmos[1]), em consonância com os requisitos estabelecidos, na fase de compreensão, e com as características dos dados resultantes do passo de transformação;
    - escolha e ajuste de parâmetros de configuração de modelos;
    - aplicação efectiva de modelos de DM, para a procura de padrões nos dados, e a sua disponibilização numa ou mais formas de representação particular.
4. **Avaliação** – Este passo abrange a interpretação, análise e apreciação dos resultados obtidos, verificando se estes vão ao encontro dos objectivos do processo e se constituem conhecimento válido, novo e relevante. Durante esta etapa também se determinam as próximas tarefas a serem executadas. De acordo com os resultados alcançados, decide-se pela continuidade e avanço para a fase seguinte ou se deverão ser efectuadas correcções.
5. **Operacionalização** – Finalmente, prepara-se a utilização do conhecimento extraído, procedendo-se à sua incorporação em algum sistema ou, simplesmente, à documentação e apresentação do mesmo às partes interessadas.

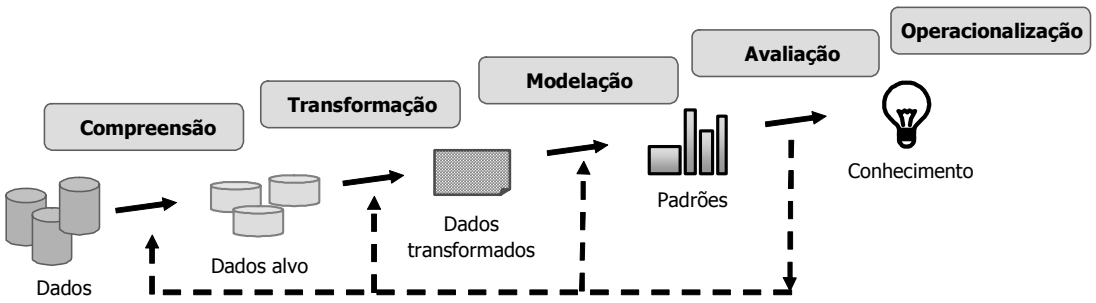


Figura 2 – Principais etapas do processo de extracção de conhecimento

[1] Prefere-se recorrer ao termo modelo, ao invés de algoritmo, já que o primeiro é mais comum em ferramentas de KDD, embora também seja usado para designar o resultado da modelação

A KDD evoluiu, e continua a progredir, retirando vantagens da intersecção de vários domínios e da respectiva investigação. Os SDW, anteriormente referidos, são um exemplo evidente desta constatação, pois propiciam um ambiente efectivo e largamente aproveitado na aplicação de processos de DM. Para além de suportarem o acesso a dados com elevada qualidade, uma vez que estes já foram alvo de procedimentos como os de limpeza e integração, estes sistemas oferecem meios convenientes, não só para o estudo preliminar de dados, como também para o pós-processamento dos seus resultados, com o intuito de ajudar o analista a discernir aqueles que constituem conhecimento pertinente.

2.3 Mineração de Dados da Web

A Web é a maior e a mais importante fonte de dados e informação, mas encontrar informação efectivamente útil e relevante não é tarefa fácil para os utilizadores. Este espaço privilegiado originou novos tipos de aplicações da DM, com o propósito de deduzir conhecimento a partir de um volume muito superior, mais complexo, mais dinâmico e menos estruturado de dados, do que os existentes nos sistemas de informação das organizações. Esses tipos de aplicações incluem-se no domínio amplo da mineração da Web (*Web Mining*), definida como a descoberta de padrões interessantes e potencialmente úteis, e de informação implícita, a partir de artefactos ou actividades relacionados com a Web [Zaïane 99]. A mineração da Web envolve implicitamente as etapas usuais de um exercício de KDD, podendo ser considerada uma extensão deste, aplicada a dados da Web [Kosala e Blockeel 00]. O recurso a métodos de DM no contexto da Web tem sido alvo de profundo interesse, em numerosos trabalhos de investigação e projectos de aplicação prática, nos quais são reconhecidas diferentes categorias de mineração. A classificação mais comum para a mineração da Web distingue entre mineração de conteúdo (*Web Content Mining*), mineração de estrutura (*Web Structure Mining*) e mineração de utilização (*Web Usage Mining* - WUM).

A mineração de conteúdo da Web consiste na extracção de conhecimento no próprio conteúdo dos documentos da Web e das suas descrições. Enquadram-se neste âmbito a mineração de dados textuais ou multimédia da Web e a mineração baseada em indexação de conceitos [Zaïane 99]. A investigação neste âmbito pode ser distinguida sob dois pontos de vista: recuperação de informação e bases de dados [Cooley et al. 97] [Kosala e Blockeel 00]. Na primeira perspectiva o objectivo é assistir a pesquisa ou filtragem de informação, geralmente, com base em perfis ou

preferências induzidas ou especificadas pelos utentes. O ponto de vista das bases de dados pretende, principalmente, modelar ou integrar, em níveis mais elevados, os dados semi-estruturados da Web, para que possam ser aplicados métodos de DM e mecanismos de interrogação mais sofisticados, do que os baseados em palavras chave.

A mineração de estrutura da Web corresponde ao processo de inferir conhecimento, com base na própria organização dos documentos Web e nas ligações existentes entre estes. Este tipo de mineração tenta desvendar o modelo subjacente da organização de ligações da Web, recorrendo à topologia de hiperligações, com ou sem atender à descrição das mesmas. Tal modelo pode ser usado para categorizar páginas Web e é útil para gerar indicações, como a similaridade e relacionamentos entre diferentes sítios Web [Kosala e Blockeel 00].

Por último, a mineração de utilização da Web [Cooley et al. 97] foca o estudo do comportamento dos indivíduos, quando navegam num sítio ou na Web, para extrair padrões de acesso interessantes, a fim de obter resultados, eventualmente, úteis para a optimização de serviços e recursos Web. Enquanto a mineração de conteúdo e de estrutura estão voltadas para os dados primários de recursos da Web, a mineração de utilização centra-se nos dados secundários, derivados a partir da interacção dos utilizadores com a Web [Kosala e Blockeel 00].

Apesar da clara distinção entre os tipos de mineração da Web – conteúdo, estrutura e utilização – especificamente, no que respeita às respectivas partes da Web que são o alvo primordial do processo de mineração, na prática existem aspectos comuns e interligações entre as suas aplicações. A mineração de conteúdo pode recorrer a ligações e a perfis depreendidos ou sugeridos pelo utilizador. O mesmo se constata na aplicação da mineração de estrutura, dado que esta pode usar informação sobre ligações. Na WUM, o conteúdo das páginas e a estrutura do sítio fazem parte do conhecimento de domínio, dado que este é requerido pelos métodos de preparação de dados, procura e análise de conhecimento acerca do uso dos sítios, para que seja possível retirar sentido dos padrões descobertos. Assim sendo, os três tipos de mineração podem ser usados isoladamente ou, mais frequentemente, combinados numa aplicação [Kosala e Blockeel 00].

2.4 Mineração de Utilização da Web

A interacção de indivíduos com sítios Web, através de navegadores, pode facultar indicações valiosas às organizações sobre a utilização dos seus sítios. A captura das acções de navegação dessa interacção, em sistemas designados de *clickstream*, é uma fonte de dados muito rica e

abrangente, acerca do comportamento dos utilizadores na Web. Os sistemas de *clickstream* registam a sequência de gestos efectuados pelos visitantes, proporcionando uma oportunidade para os acompanhar durante as suas visitas e conhecer as trajectórias de acções que levaram a um determinado comportamento. É possível saber o que visualizaram e durante quanto tempo, o que seleccionaram e o que rejeitaram, assim como captar sinais directos de satisfação ou insatisfação. A recolha de dados é, assim, naturalmente ampliada, não se consignando, apenas, ao tradicional registo das transacções efectivamente realizadas com o organismo, mas cobrindo também múltiplos aspectos relacionados com a globalidade do processo de interacção.

O registo de dados em sistemas de *clickstream* vem permitir a realização de análises, anteriormente inviáveis, em particular aquelas que poderão conduzir à compreensão e caracterização do comportamento dos utilizadores. O conhecimento desse comportamento pode ser usado na construção de serviços, mais efectivos e ajustados a cada perfil de utilizador, e na maximização do potencial da Web. Deste modo, é possível personalizar a experiência em tempo real, conduzir os visitantes a áreas que reflectem os seus interesses, ajustar dinamicamente a oferta de serviços mais apelativos e profícuos para os utilizadores individuais, bem como, eventualmente, fidelizar os utentes e permitir às organizações fornecedoras de serviços um maior retorno dos seus investimentos. É a imensa oportunidade, traduzida, por um lado, numa recolha intrínseca de um manancial de dados comportamentais e, por outro lado, nas vantagens de interactividade dos sítios na Internet, em contraste com qualquer outro meio de comunicação de massas, que revestem a WUM de potencialidades e especificidades. Por exemplo, ao contrário do comércio tradicional, é possível, não só, adquirir um conhecimento mais profundo e sistemático relativo aos interesses do utente, sem ter de o questionar, como também efectuar recomendações de maneira automática e quase implícita, designadamente, sem a intervenção de um vendedor.

O seguimento da navegação de cada visitante pressupõe a captura de uma série vasta e diversa de dados, retractando todas as suas acções, desde o instante em que este "entra" no sítio até ao momento em que "sai" do mesmo, individualizando cada sessão para enquadrar o registo dessas acções. Uma sessão corresponde a uma visita de um utilizador, englobando todos os eventos relacionados e consecutivos, decorridos durante essa visita, formando, assim, uma unidade lógica. A identificação de sessões é crucial para o processo exploratório, uma vez que as potencialidades do estudo de eventos isolados, sem atender ao seu contexto, são limitadas. Também se usa vulgarmente o termo episódio, para denotar um subconjunto de acontecimentos de uma sessão, considerados significativos e relevantes para uma determinada tarefa analítica. A identificação ou diferenciação de utilizadores poderá ser igualmente necessária, para distinguir visitantes que

voltam ao sítio daqueles que o acedem pela primeira vez. Além disso, através do reconhecimento de visitantes, é possível interligar e usar toda a informação que possa existir acerca dos mesmos, para, assim, estabelecer um relacionamento mais significativo com estes.

Segundo [Zaïane 99], existem duas orientações na WUM, direccionadas pela aplicação das descobertas – extracção de padrões de acesso gerais e individuais. Atendendo ao âmbito da WUM, é possível subdividir a primeira orientação em duas, discriminando, assim, três abordagens. A primeira possui um âmbito de aplicação global (na prática apenas vários sítios) e procura obter conhecimento sobre o comportamento da generalidade dos utentes, com o intuito de compreender padrões de acesso e tendências gerais, para melhorar o agrupamento de recursos e a estrutura da Web. A segunda situa-se ao nível do sítio e pretende derivar padrões de utilização e tendências comuns, no que concerne ao acesso ao sítio, a fim de aumentar a sua eficácia geral. A terceira foca-se em comportamentos e tendências individuais, com o propósito de optimizar o sítio para cada visitante individual, ao longo do tempo e em função dos seus padrões de acesso. A primeira orientação está associada a investigadores interessados em estudar e caracterizar o uso da Web, enquanto as outras duas estão voltadas para indivíduos e organizações envolvidos no projecto e implementação de sítios Web, retractando as abordagens alvo do presente trabalho. Os exercícios de WUM podem ainda ser subdivididos em duas vertentes distintas:

- a utilização de métodos de DM, usualmente complicados e morosos, em dados registados no passado;
- a aplicação de técnicas de DM a dados correntes, com base no conhecimento inferido anteriormente, facultando os resultados para o controlo e alteração da interacção com o visitante em tempo real.

Tendo presente o contexto e as formas de exploração da WUM, esta pode ser definida como a aplicação de processos de KDD sobre fontes de dados baseadas em *clickstreams*, normalmente conjugadas com outros dados relacionados, para descobrir padrões de utilização úteis. Para as organizações detentoras de sítios Web, o uso de WUM visa, essencialmente, a obtenção de um conhecimento mais profundo acerca do comportamento dos visitantes e a avaliação do resultado dos variados tipos de iniciativas subjacentes, para poderem actuar de forma sustentada, no sentido de optimizar a efectividade das mesmas e de satisfazer as metas dos respectivos sítios.

De acordo com [Cooley et al. 99] [Srivastava et al. 00] [Cooley 00], as principais etapas constituintes do processo de WUM consistem no pré-processamento, descoberta de padrões e análise dos padrões derivados. Por comparação com os passos discutidos para o processo genérico de KDD, depreende-se que essas etapas primordiais são antecedidas por uma fase de

"compreensão", a partir da qual se define os requisitos do processo de WUM. A conclusão de tais etapas também dá lugar a uma fase de "operacionalização" do conhecimento obtido.

A etapa de pré-processamento relaciona-se com o passo de "transformação", mas assume contornos específicos, devido às particularidades do ambiente da Web. O grau de complexidade desta etapa depende, expressivamente, do nível de fiabilidade das fontes de dados disponíveis, que, no caso dos dados de *clickstream*, varia consideravelmente. Estas fontes podem tratar-se de dados brutos, obtidos a partir dos registos de servidores Web, obrigando, eventualmente, à realização de actividades peculiares e laboriosas, tais como a identificação de sessões e utilizadores e a reconstituição das acções dos visitantes, para, deste modo, dotar os dados de maior significado. A descrição de técnicas e heurísticas para lidar com estes desafios e das tarefas de pré-processamento a levar a cabo, nestas circunstâncias, são abordada em [Cooley et al. 99] [Cooley 00]. Alternativamente, as fontes de dados de utilização podem ser enriquecidas ou produzidas por mecanismos conducentes ao registo directo e optimizado das acções dos visitantes, permitindo uma maior aproximação, em termos de uniformidade, detalhe e precisão, aos dados apropriados para WUM. Para essas fontes o pré-processamento é nitidamente simplificado.

A etapa de descoberta de padrões corresponde à fase de "modelação". Os métodos de DM mais usados, no contexto da WUM, e os desafios subjacentes, nesta fase do processo, serão abordados posteriormente, pois é justamente sobre estes que o trabalho realizado incide mais directamente.

A etapa de análise de padrões é equivalente ao passo de "avaliação", apesar de dar mais ênfase à selecção de padrões úteis, entre resultados triviais, numerosos e difíceis de interpretar. Estas dificuldades acentuam-se no contexto da WUM devido a características dos dados de utilização (e.g. volume ou complexidade) e a outras questões, como a simplicidade dos indicadores de importância convencionais para filtrar padrões de navegação. Nomeadamente, as referências a páginas de topo e de interligação tendem a ser mais frequentes do que os acessos a outras páginas e, ainda, muitos dos padrões dominantes, quase sempre, apenas confirmam o óbvio, pois reflectem a estrutura do sítio. Por conseguinte, os mecanismos capazes de auxiliar o analista a filtrar, interpretar e avaliar os padrões extraídos, são de grande utilidade. Algumas das propostas específicas neste âmbito englobam linguagens de mineração para especificar o foco da análise (e.g. sistema *Web Utilization Miner* [Spiliopoulou e Faulstich 98]), mecanismos de filtragem automática de padrões (e.g. sistema *Web Site Information Filter* [Cooley 00]) e linguagens de interrogação de conhecimento (e.g. sistema *WebMiner* [Mobasher et al. 96]).

2.4.1 Potencialidades e Áreas de Aplicação

A complexidade da implementação e administração de sítios Web tem aumentado, acompanhando o vertiginoso acréscimo da utilização da Internet, do volume de informação publicada e das transacções que se concretizam através deste canal. Captar a atenção dos utentes e suportar serviços, atendendo a requisitos, cada vez, mais sofisticados, são desafios difíceis de vencer, num ambiente fértil, de exposição e competitividade à escala planetária, onde imperam a sobrecarga, a heterogeneidade, o dinamismo e a imediaticidade, para além da influência dos sucessivos avanços tecnológicos. A adopção de soluções baseadas numa visão puramente tecnológica não garante resultados efectivos, pelo facto de estas não estarem traçadas em termos de oportunidades e motivações. A eficácia de um sítio é determinada pelos objectivos de negócio a atingir e depende de um conjunto vasto de factores e recursos, os quais, interligadamente, afectam os níveis de interesse e satisfação despertados nos visitantes. O excesso de informação e de oferta, a diversidade de preferências e de características dos utentes e a própria heterogeneidade dos sítios criam desorientação nos utentes, levando a divergências acentuadas entre a forma real e desejada de uso dos mesmos. Adicionalmente, as mudanças constantes e velozes do mercado e dos próprios utilizadores, exigem mecanismos capazes de detectar e reagir imediatamente a alterações. Assim sendo, é indispensável monitorizar continuamente os visitantes e recorrer a meios para conhecer e compreender o seu comportamento, dentro do próprio ambiente do sítio.

A constante avaliação da eficácia de um sítio e a sua contínua optimização são, obviamente, preponderantes para enfrentar vários desafios. Também é indubitável que os dados de utilização contêm respostas para muitas das questões levantadas pelos agentes de decisão. Falta, todavia, precisar de que modo o estudo destes dados, e principalmente a WUM, podem contribuir para os fins enunciados. A exploração dos dados de *clickstream* pode ser usada para uma ampla variedade de finalidades e no apoio a muitos tipos de decisões, as quais, no âmbito da administração de um sítio, podem ser sistematizados nas áreas de aplicação de melhoramento de sistemas, alteração de sítios, personalização e "inteligência" para o negócio (*Business Intelligence*) [Cooley 00] [Srivastava et al. 00].

O melhoramento de sistemas centra-se na optimização do desempenho e de outros atributos de qualidade de serviço (e.g. disponibilidade), claramente fulcrais para garantir o bom funcionamento dos serviços, a eficiência do sítio e a satisfação básica dos visitantes. A WUM pode proporcionar a chave para o entendimento do fluxo de tráfego Web, o qual é requerido para orientar o desenvolvimento de políticas de *caching*, *prefetching*, transmissão de rede, equilíbrio de carga e distribuição de dados. A WUM também pode ser útil no apoio da melhoria da segurança dos

sistemas e redes, ao deduzir padrões úteis para detectar intrusões, fraudes e tentativas de entrada não autorizadas [Cooley 00] [Srivastava et al. 00].

A alteração de sítios é, obviamente, uma área de aplicação muito usual da WUM. O uso dos sítios nem sempre se processa de acordo com as expectativas, sendo difícil de examinar e perceber perante quantidades colossais de dados de *clickstream*. Indicadores como a frequência de acesso a páginas, muitas vezes, revelam-se, não só insuficientes, como também enganosos, para desvendar o verdadeiro papel e a importância de recursos. A extracção de padrões de utilização, via WUM, faculta detalhes do uso real do sítio e pode derivar relações mais complexas e informativas, permitindo avaliar a adequação da sua estrutura e conteúdo. Estes padrões podem apoiar decisões de reestruturação e de afinação da estrutura de ligações e de melhoramento e reorganização do conteúdo das páginas do sítio, para assim optimizar a sua atractividade e usabilidade, melhor servir as solicitações dos visitantes e criar uma presença mais efectiva na Web.

A alteração de sítios pode ser conduzida de modo estático, afectando a interacção com todos os visitantes, ou de forma dinâmica, transparente e em tempo real. O ajuste dinâmico ("silencioso") [Mobasher et al. 01] [Koutri et al. 05] é um tipo de adaptação que procura combater os problemas do excesso de informação e do desenho da interacção na criação de sítios complexos, fundamentando-se em duas motivações essenciais:

- as preferências dos visitantes podem ser divergentes;
- é difícil conciliar os diferentes requisitos, designadamente, à medida que um sítio cresce e evolui.

A adaptação dinâmica de um sítio é uma abordagem concertada, uma vez que reflecte várias perspectivas, sem impor alterações que possam ter efeitos negativos em certos utentes.

Segundo [Mobasher et al. 00], a personalização na Web pode ser descrita como qualquer acção que ajusta a experiência de interacção com um sítio Web, a um visitante ou conjunto de visitantes. A comunidade de utentes da Web é diversa, possuindo conhecimentos, interesses e desígnios distintos. Trata-se, portanto, de uma estratégia de diferenciação, actualmente muito aproveitada, com vista a facilitar a interacção, melhorar o uso do tempo dos indivíduos e aumentar a relevância do conteúdo, à medida que a sessão progride, tentando cativar os utentes e atingir as metas do sítio. As acções de personalização podem variar entre tornar a apresentação mais agradável, a antecipar as necessidades do indivíduo e disponibilizar a informação ou os serviços apropriados. A personalização pode ser apontada como o tipo de adaptação mais proeminente, mas, neste caso, claramente visível. Tipicamente, e mais especificamente sob a designação de recomendação

pessoal [Mobasher et al. 01], pressupõe uma adaptação dinâmica e assumida ao visitante, por meio de sugestões dos itens adequados, com base no seu modelo de utilizador.

Para levar a cabo um processo de personalização na Web pode recorrer-se a mecanismos diversos, em função de dados comunicados explicitamente pelos utentes ou induzidos a partir da interacção. Estes dados podem ser obtidos apenas durante a sessão corrente ou também em sessões anteriores. Em termos genéricos, este processo não pressupõe a identificação dos visitantes, nem tão pouco o recurso a WUM. Contudo, a WUM é uma abordagem excelente para este fim, uma vez que a personalização poderá ser conduzida de maneira mais implícita, efectiva e eficiente, com base no perfil e comportamento do visitante, inferidos a partir de dados de utilização. A WUM pode reduzir a premência da obtenção de indicações subjectivas, por processos manuais e intrusivos, e a dependência em relação a dados sujeitos a desactualização [Mobasher et al. 00], contribuindo para a construção de perfis de visitantes mais precisos e actualizados, assim como para o aperfeiçoamento dos resultados de sistemas de personalização ou de recomendação existentes. Além disso, evita-se o processamento em tempo real dos dados volumosos já registados, aplicando-se, em sua substituição e no contexto dos dados correntes, o modelo ou padrões gerados (*off-line*) a partir dos dados do passado [Mobasher et al. 01].

A "inteligência" para o negócio é outra área de aplicação que pode retirar grandes benefícios da WUM, principalmente, em ambientes de comércio electrónico. Esta área também possui interligações com sistemas de personalização e recomendação, no âmbito do seu uso como uma estratégia orientada para retirar vantagens em torno da actividade na Web. No contexto da "inteligência" para o negócio a finalidade da WUM é, basicamente, auxiliar as organizações a atingir os seguintes objectivos:

- caracterizar e compreender os visitantes, estabelecendo as fundações para a condução de iniciativas diversificadas e diferenciadas, no sentido de os servir melhor, valorizar a sua experiência, incrementar a eficácia das iniciativas e construir relações mais sólidas e profícuas com estes;
- ganhar um conhecimento profundo do seu mercado, visando a efectivação de transacções da forma mais optimizada possível e a maximização das oportunidades de negócio, garantindo, assim, uma posição mais consolidada;
- avaliar o resultado de processos e iniciativas, em função do impacto, eficácia e retorno do investimento, para poder agir em conformidade e saber se essas iniciativas e o próprio sítio estão a atingir as metas de negócio estipuladas.

O primeiro objectivo indicado centra-se no visitante e contempla a captura de dados sobre este, no que concerne a padrões de navegação, interesses indicados pelas suas escolhas ou comunicados explicitamente, reacções a iniciativas, atributos demográficos e historial do seu relacionamento com o sítio Web. Os utentes podem ser organizados em função de distintas características, sendo um exemplo a classificação com base em métricas, como a recência, frequência de visitas e os valores quantitativos que traduzam a finalidade básica do sítio. A informação recolhida pode ser usada na construção de perfis e na segmentação dos visitantes, para determinar o seu valor em termos de ciclo de vida e para definir alvos de marketing, com o intuito de adequar estratégias previamente definidas de recomendações, promoções, publicidade e direccionamento de informação.

Para além de ser requerido compreender o cliente, é imprescindível entender o próprio negócio na Web, captando a dinâmica do mercado e antecipando tendências (e.g. procura de novos produtos). Através da exploração de dados de *clickstream*, uma organização pode obter um esboço do comportamento de toda a população de visitantes, não apenas dos clientes reais, e determinar indicadores que lhe permitam prever riscos, reduzir custos e identificar oportunidades. [Kohavi 01] observa que a Web é um centro de investigação e um "laboratório experimental", no qual muitas iniciativas podem ser testadas (e.g. a introdução de novos produtos ou o uso de diferentes tipos de anúncios) e generalizadas para outros canais, proporcionando um sistema de alerta para detectar padrões emergentes. A WUM também pode ser usada para derivar as recomendações a aplicar, envolvendo transacções adicionais dirigidas de maior valor *(up-sell)*, transacções cruzadas de produtos ou serviços complementares *(cross-sell)* e transacções associadas para visitantes com afinidades.

O terceiro objectivo está voltado para a avaliação do negócio na Web. Para tal, é indispensável examinar e mensurar os resultados de iniciativas (e.g. eventos críticos de processos, transacções concretizadas, campanhas de publicidade, promoções, recomendações e o próprio sítio Web), em consonância com a estratégia que está subjacente (e.g. atracção ou contacto, conversão e retenção), e também em função da associação a produtos ou serviços, eventos e grupos de visitantes individuais, cujo comportamento se conhece. Este objectivo evidencia vários requisitos, tais como captar a semântica subjacente nos recursos (e.g. produtos ou serviços em causa), integrar dados de utilização da Web com dados operacionais e suportar análises poderosas e versáteis.

2.4.2 Fontes e Categorias de Dados Envolvidas

A fonte primária de dados de *clickstream* reside em ficheiros de histórico (*log*), gerados automaticamente por servidores Web, devido à facilidade de obtenção dos mesmos. Embora estes históricos sejam uma fonte de dados rica, não reflectem, com fiabilidade, as acções dos utentes. Os históricos foram originalmente concebidos para fornecer estatísticas de administração, possuindo várias imperfeições (e.g. ausência de identificação única de sessões e visitantes). Por este motivo, usualmente são requeridas operações rebuscadas de tratamento, capazes de produzir dados com qualidade superior, a partir dos dados brutos do histórico. Por outro lado, para além da captura de elementos representativos da navegação de visitantes, é pertinente incluir outras acções e dados relacionados, os quais permitirão enriquecer o perfil da sessão e o conhecimento do visitante. Entre esses dados destacam-se os pertencentes às seguintes categorias e fontes:

- dados do sítio Web, respeitantes à estrutura e conteúdo das páginas, oriundos, sobretudo, de servidores Web, servidores de conteúdo ou servidores de aplicações;
- detalhes de outros eventos de interacção, tais como eventos críticos (e.g. adição de item a cartão de compras e acções como recarregamento) e submissão de recursos, em que os pares nomes e valores são removidos do URI (e.g. expressões de pesquisas) [Kohavi 01], geralmente, recolhidos ou enriquecidos através de mecanismos optimizados;
- informação da actividade, relativa a produtos ou serviços e transacções concretizadas, entre outros, tipicamente proveniente de sistemas operacionais;
- perfil de visitantes, englobando todos os registos explícitos de elementos acerca de grupos ou visitantes individuais, usualmente, mantidos em sistemas contendo dados consolidados.

Como os históricos vulgarmente não cobrem as categorias anteriores, pode ser necessário recorrer a outras fontes e a meios complementares de enriquecimento dos dados. Em acréscimo, a complexidade e vulnerabilidade de heurísticas de tratamento de dados e a dificuldade na integração de dados dispersos por diversas fontes heterogéneas, estão na origem de mecanismos alternativos de recolha. Oportunamente, a colecta de dados pode ser efectuada ao nível do servidor de aplicações. Este método é considerado o mais efectivo [Kohavi 01] [Hu e Cercone 04], uma vez que o servidor de aplicações controla todas as actividades de interacção com o utilizador, estando em posição de criar um conjunto de dados consistente (e.g. identificadores comuns entre diferentes categorias de dados), consolidado (e.g. múltiplos servidores), abrangente (e.g. vários tipos de eventos) e mais preciso (e.g. identificação de sessões e informação sobre conteúdo dinâmico). Na realidade, a optimização da captura de dados pressupõe um esforço significativamente inferior neste contexto, pois as operações ou serviços que os originam já se

encontram automatizadas. Este facto abre novas perspectivas ao estudo de dados – cria-se a oportunidade de desenvolver capacidades de aquisição de dados para o propósito específico de análise e mineração, ao invés de ter de lidar com as dificuldades introduzidas pela sua recolha para outros fins [Kohavi e Provost 01].

De qualquer modo, levanta-se o desafio de atender aos requisitos das solicitações analíticas dos agentes de decisão, já que tais dados atingem volumes extremamente elevados, mesmo para sítios de reduzida dimensão. Uma vez que os SDW têm sido usados com sucesso no suporte da exploração de dados convencionais, constituem uma solução que pode ser adaptada ao âmbito dos dados de *clickstream*. Por outro lado, com o alargamento da audiência de sistemas de DW, faz todo o sentido a universalização da sua disponibilização em ambiente Web, para permitir que todos os tipos de dados sejam publicados de forma integrada, através de interacções baseadas em navegadores. O cruzamento destas tendências alterou a natureza do DW e originou uma nova figura – o sistema de *Data Webhousing* (SDhW) –, um SDW capaz de conciliar também as fontes de *clickstream* e de oferecer os seus serviços em ambiente Web [Kimball e Merz 00]. Nestas circunstâncias, de integração adicional dos dados de *clickstream*, o DW passa a designar-se *Data Webhouse* (DhW). Segundo a abordagem sugerida em [Kimball e Merz 00], a implementação de um SDhW pressupõe o projecto e construção de um *data mart* de *clickstream*, baseado num esquema em estrela, cujas características centrais se descreve seguidamente.

O conjunto de dimensões propostas para o *data mart* de *clickstream*, geralmente também encontrado no contexto de DW tradicionais, engloba:

- as duas dimensões temporais Data e Hora, típicas em qualquer DW;
- a dimensão Produto ou Serviço, habitual em "inteligência" para o negócio;
- a dimensão convencional Cliente ou Visitante;
- a dimensão Causal, relativa às condições presentes na época de um facto e com influência no comportamento do mercado (e.g. promoções);
- a dimensão Entidade de Negócio, acerca dos agentes ou papéis, dentro ou fora da organização, associados aos factos (e.g. parceiros).

Já nas dimensões mais específicas do *data mart* de *clickstream* incluem-se as seguintes:

- a dimensão Página descreve o contexto e detalhes de páginas Web;
- a dimensão Referência contém informação sobre o local de origem da navegação para a página corrente;
- a dimensão Evento retracta acontecimentos particulares que decorreram em torno de determinadas páginas e num dado momento;

- a dimensão Sessão proporciona um ou mais níveis de classificação de cada sessão, que a caracterizam e lhe conferem significado, sob a perspectiva da análise.

Quanto às tabelas de factos fundamentais são indicadas duas:

- a tabela de factos de sessão, na qual cada registo representa uma sessão de um visitante e as medidas contemplam a duração da sessão, total de páginas visitadas, total de pedidos efectuados, unidades e valores desses pedidos (se aplicáveis);
- a tabela de factos de eventos de página, vulgarmente designada tabela de factos de *click*, em que cada registo se refere a um evento individual, respeitante a uma dada página, e as medidas abarcam a duração do evento, número de pedidos efectuados, unidades e valores desses pedidos.

Dado o elevado volume das duas tabelas anteriores, é sugerida uma tabela de factos agregados, para melhorar o desempenho das questões que podem ser respondidas a partir de dados acumulados, tais como totais de sessões para cada mês, por grupos de visitantes.

Uma vantagem inerente na implementação de um SDhW reside na conjugação com tabelas de factos já existentes, noutros *data marts* da organização, por meio de dimensões e factos coerentes. O *data mart* mais relevante e comum é o de transacções, onde são armazenados os registos de cada transacção concretizada, através da Web ou fora desta, recorrendo a uma dimensão Canal para identificar o meio utilizado. A Tabela 1 mostra as principais dimensões aplicáveis às tabelas de factos dos *data marts* de *clickstream* e de transacções.

Tabela 1 – Principais dimensões e tabelas de factos de *data marts* de *clickstream* e transacções

	Página	Sessão	Evento	Referência	Data	Hora	Cliente/ Visitante	Produto/ Serviço	Causal	Entidade de negócio	Canal
Sessão	√[1]	√		√	√	√	√		√	√	
Click	√	√	√	√	√	√	√	√	√	√	
Transacção					√	√	√	√	√		√

[1] Refere-se à página inicial e poderá contemplar também a página final

A implementação de sistemas de DW envolve actividades complicadas de planeamento, modelação e contínua actualização e refinamento. Contudo, uma solução bem projectada permite implementar modelos mais sofisticados, satisfazer as solicitações analíticas dos utilizadores e monitorizar o histórico de operações, além de antever situações futuras. Ao transformar, consolidar e racionalizar

os dados dispersos em várias fontes e localizações, de uma forma individual e coerente, um sistema de DW ou DhW permite que sejam realizadas análises estratégicas com grande eficácia, em dados antes inacessíveis ou subaproveitados. Deste modo, um sistema de DhW surge, não só, como uma vantagem de longo termo, para sustentar o processamento analítico e a WUM, mas também, como uma possibilidade de aproveitamento de investimentos realizados na tecnologia de DW, dada a contínua expansão da sua adopção no seio das organizações.

2.4.3 Abordagens de Suporte à Exploração de Dados

A proliferação da oferta de serviços em ambiente Web desencadeou a rápida disponibilização de variadíssimas formas de exploração de sistemas de *clickstream*. As respectivas abordagens de suporte são apresentadas na Figura 3, podendo ser caracterizadas considerando duas dimensões:

- os tipos de ferramentas;
- as categorias de fontes ou estruturas de dados subjacentes.

A primeira dimensão corresponde a distintos níveis de exploração dos dados, com diferentes papéis no processo de WUM, seja este crucial ou complementar, como no estudo preliminar dos dados ou na análise dos padrões descobertos, enquanto a segunda traduz, em certa medida, as linhas primordiais de investigação em WUM, conforme se descreve seguidamente.

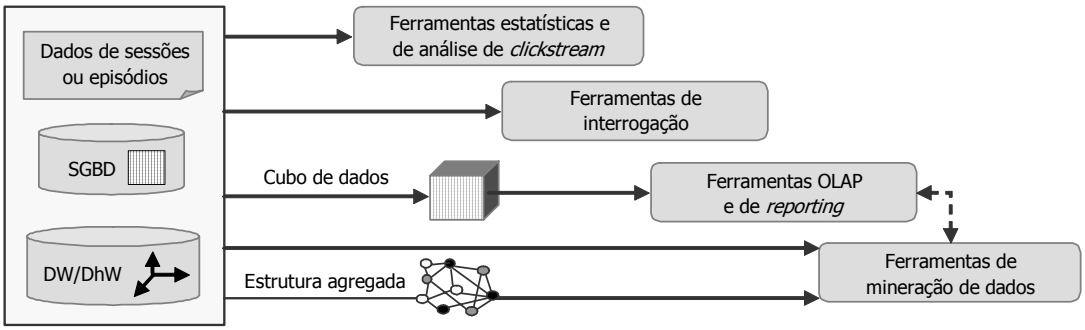


Figura 3 – Abordagens de suporte à exploração de dados de *clickstream*

As técnicas estatísticas são o método mais comum de prospecção de dados de utilização. Dado o seu carácter genérico, surgiram adaptações destas técnicas voltadas para estes dados, sob a forma

de ferramentas de análise de *clickstream* (ou de análise de acessos a páginas Web). Muitas destas ferramentas proporcionam relatórios periódicos, contendo informação desta natureza, como, por exemplo, páginas mais acedidas em determinado período. Estas ferramentas detêm as vantagens de serem simples de utilizar e pouco exigentes, em termos de recursos computacionais, bem como de facultarem mecanismos para examinar o tráfego e para reportar a efectividade do sítio e tendências de uso, concedendo uma visão geral de como o sítio está a ser usado. A informação extraída pode ser útil, mas possui limitações no que concerne a abrangência e profundidade. Outras desvantagens de tais ferramentas são a incapacidade para lidar com dados ao longo de períodos extensos e a rigidez da configuração dos relatórios que podem ser produzidos.

As linguagens de interrogações também se afiguram apropriadas para examinar dados e descobertas, nomeadamente por não restringirem as consultas que podem ser efectuadas. Todavia, é requerida experiência no uso destas linguagens, para que seja possível retirar benefícios da sua utilização, algumas questões são difíceis de formular, através de linguagens convencionais, e o tipo de apresentação de resultados torna-se, ainda, limitativo e pouco intuitivo.

Uma alternativa mais poderosa e intuitiva, do que as linguagens de interrogação, e mais sofisticada e exigente, do que as ferramentas de análise de *clickstream*, são as ferramentas OLAP. Os próprios dados obtidos do histórico podem ser guardados em estruturas de cubos para aplicação de OLAP, em função de dimensões baseadas em campos disponíveis (e.g. domínio, recurso solicitado, agente do utilizador e referência) [Zaïane et al. 98]. No entanto, a fonte de dados ideal e mais usual para OLAP é um SDhW. As dimensões do cubo podem descrever dados acerca do sítio, do seu uso e de conhecimento de domínio (e.g. produtos), enquanto as medidas podem ser relativas a contagens do acesso a recursos ou de seguimento de ligações (*clickthrought*) e, ainda, a quantidades e valores transaccionados. As ferramentas OLAP podem ser usadas para discernir e mensurar tendências associadas a diferentes objectos e para distinguir padrões de uso individuais. Estas ferramentas são ainda largamente aplicadas no tratamento de resultados de KDD. Porém, tais ferramentas, por si próprias, não descobrem automaticamente padrões nos dados, sendo requeridas abordagens mais sofisticadas como a DM.

A primeira abordagem de DM reside na aplicação destes métodos, directamente, aos dados alvo. Outra abordagem distinta, adoptada no sistema *WebLogMiner* [Zaïane et al. 98], envolve a utilização de técnicas OLAP, em conjunto com as de DM, recorrendo, para tal, ao cubo de dados. Esta abordagem pretende ser mais poderosa e flexível, simplificando o processo de KDD e contribuindo para que este se torne mais produtivo. A intenção é integrar e intercalar a aplicação de operações OLAP e de métodos de DM e conduzir a KDD nos moldes OLAP, para que esta se

processe, tanto quanto possível, de modo interactivo e seja realizada em distintas partes da base de dados, ou do DhW, e a diferentes níveis de abstracção [Han 98].

Quanto à segunda dimensão de suporte à exploração de dados de *clickstream*, é possível distinguir três formas essenciais de fontes ou estruturas de dados alvo: dados de sessões ou episódios, SDhW (e sistemas de base de dados) e estruturas agregadas. Outros tipos de fontes de dados não abordados incluem documentos XML (*eXtensible Markup Language*) [WWW1]. Uma linha importante de investigação em WUM contempla mecanismos para lidar directamente com ficheiros de histórico, produzindo sessões ou episódios como o conjunto de dados alvo. Esta linha pode ser exemplificada por trabalhos como [Cooley et al. 99] [Cooley 00] e [Srivastava et al. 00].

A segunda orientação procura retirar benefícios dos avanços em diferentes domínios, focando-se em tecnologias de base de dados e de DhW, e na sua adaptação aos problemas da WUM. Os SDhW são explorados em muitos trabalhos, tais como os indicados a seguir. [Büchner et al. 99] propõem um processo para derivar "inteligência" para marketing, a partir de um SDhW, utilizando um esquema em floco de neve. Já as abordagens seguidas por [Gomory et al. 99], [Ansari et al. 01] e [Hu e Cercone 04] recorrem a SDhW baseados num esquema em estrela. Contudo, o uso de SDhW também possui desvantagens, tais como a ineficiência da extracção de sequências extensas de eventos. Apesar destes sistemas proporcionarem um nível elevado de desempenho no acesso aos dados, esta ineficiência surge para tabelas de factos de *click* longas, em virtude de serem requeridas várias operações de junção ou auto junção. Uma optimização possível consiste na construção de tabelas de factos de subsequências de eventos, de todos os tamanhos. Não obstante, esta solução impõe uma capacidade elevada de armazenamento, no caso de sessões extensas [Jespersen et al. 02].

A terceira direcção de trabalho concentra-se no desenvolvimento de novos modelos, estratégias e métodos de DM, capazes de capturar directamente a semântica da interacção na Web [Borges 00]. Esta orientação pode fundamentar-se na criação de estruturas agregadas, recorrendo a representações compactas de elevados volumes de dados, e em métodos de modelação e de DM dedicados às mesmas, sendo exemplificada por trabalhos como [Spiliopoulou e Faulstich 98] [Borges e Levene 99] [Pei et al. 00]. Em termos gerais, as desvantagens desta abordagem residem, sobretudo, na perda de detalhes, acerca de cada evento individual e da respectiva sessão, e na sua incapacidade para representar eficazmente dados adicionais.

Por último, a constatação das vantagens e desvantagens das duas últimas linhas de trabalho em WUM suscitou o surgimento de abordagens híbridas, como a proposta em [Jespersen et al. 02].

Esta abordagem centra-se na utilização de um SDhW, baseado numa tabela de factos de *click*, e na criação dinâmica de uma estrutura agregada, em função das solicitações analíticas.

2.4.4 Principais Métodos de Mineração de Dados e suas Aplicações

Existem vários tipos de métodos de DM com aplicação no contexto da WUM. A Figura 4 ilustra um conjunto de finalidades típicas no recurso a WUM e as respectivas técnicas aplicáveis, com base em [Berendt et al. 02]. Cada método é implementado recorrendo a diferentes modelos, podendo estes ser adaptados ou concebidos especificamente para WUM. A necessidade de algoritmos específicos deve-se a características peculiares destes dados, tais como a sua magnitude e carácter sequencial, que limitam a eficácia, eficiência e escalabilidade de alguns modelos genéricos. Entre os métodos mais usados em WUM incluem-se a descoberta de regras de associação, a extracção de padrões sequenciais e de padrões de navegação, a classificação e o agrupamento. Com base no referido, prossegue-se com a apresentação destes métodos, salientando o propósito da sua aplicação no presente contexto.

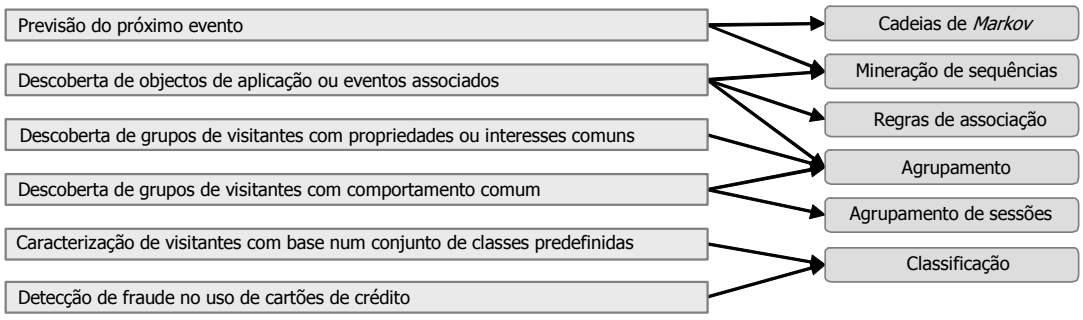


Figura 4 – Exemplos de finalidades de mineração de dados e respectivos métodos aplicáveis

A descoberta de regras de associação procura deduzir relacionamentos entre a ocorrência simultânea de conjuntos de itens, na mesma transacção, sob a forma $X \Rightarrow Y$. Podem também ser derivados, como parte da mineração destas regras, padrões menos restritivos intitulados conjuntos frequentes, relativos às séries de itens que ocorrem frequentemente juntas na mesma transacção. Não obstante, a informação inerente nas regras (e.g. antecedente) conduz a padrões mais informativos. No âmbito da WUM, este método é usado para descobrir associações entre eventos

ou recursos, sendo vulgarmente aplicado para encontrar páginas habitualmente acedidas em conjunto, na mesma sessão ou episódio. A presença ou ausência destas regras é útil no apoio a decisões de reestruturação do sítio (e.g. adição de ligações entre páginas). As regras extraídas também podem conceder indicações para o ajuste dinâmico e melhoramento do desempenho de sistemas, servindo, neste caso, como heurística para políticas de *prefetching* e de *caching*. Em acréscimo, a determinação de conjuntos frequentes e de regras de associação possui grande aplicação na área de "inteligência" para o negócio. Os padrões acerca de itens transaccionados ou acedidos simultaneamente podem ajudar as organizações a estabelecer estratégias de marketing efectivas, envolvendo acções como a identificação de recomendações e promoções, planeamento da disposição de produtos, serviços e informação e na avaliação de iniciativas.

A descoberta de padrões sequenciais pode ser considerada uma extensão da mineração de regras de associação, na medida em que tenta induzir padrões de co-ocorrência, na mesma ou entre diferentes sessões ou episódios, incorporando a ordem de ocorrência dos eventos. Os padrões sequenciais encontrados denotam sequências significativas de eventos, permitindo inferir tendências úteis e prever os próximos pedidos, durante a visita corrente ou em visitas futuras. Estes padrões podem, ainda, assumir a forma, mais específica, de padrões sequenciais contíguos. Nestes padrões os itens que surgem na sequência têm de ser adjacentes com respeito à ordenação subjacente. Em contraste, os itens que aparecem em padrões sequenciais (normais), apesar de preservarem a ordenação subjacente, não necessitam de ser adjacentes. Por conseguinte, reflectem padrões de navegação mais gerais, enquanto os padrões sequenciais contíguos são mais restritivos, podendo ser usados para propósitos como o estudo dos percursos seguidos pelos visitantes e a captura de caminhos de navegação frequentes.

A mineração de padrões sequenciais assume particular relevância em WUM, dada a diversidade das suas aplicações e a própria essência da navegação na Web, sendo alvo de múltiplas iniciativas que procuram optimizar a sua realização. A natureza sequencial dos eventos de sessões ou episódios permite adoptar tipos de modelos distintos, a fim de representar a actividade de navegação dos visitantes e de examinar e desvendar os padrões subjacentes. Conforme já foi dito, uma abordagem proeminente baseia-se na criação de estruturas que agregam dados de várias sessões, com a intenção de produzir representações compactas e passíveis de manipulação mais eficiente.

Um dos modelos muito explorados trata-se de cadeias de *Markov*, em que os estados correspondem a páginas vistas (ou subsequências de páginas em modelos de ordem $k>1$) e as transições a ligações seguidas, às quais são associadas probabilidades proporcionais à frequência

de acesso. Os modelos de *Markov* são especialmente apropriados em modelação preditiva, baseada em sequências contíguas de eventos. Consequentemente, têm sido propostos, em muitos trabalhos, como o modelo subjacente para a implementação de funcionalidades conducentes à redução de carga de servidores e do tempo de serviço. Adicionalmente, estes modelos facultam uma infra-estrutura útil para tarefas como a análise de caminhos, assim como para extrair padrões de navegação, recorrendo a algoritmos voltados para estes modelos. Um exemplo é a infra-estrutura proposta por [Borges e Levene 99], em que as sessões dos visitantes são modeladas como uma gramática probabilística de hipertexto. A mineração desta infra-estrutura permite determinar as *strings* com probabilidade mais elevada, correspondendo estas às subsequências de páginas ou aos percursos do sítio preferidos pelos utilizadores.

Outra forma de representação da actividade de navegação dos visitantes recorre a estruturas em árvore. Um exemplo desta abordagem contempla a utilização de uma árvore agregada, na qual se baseia o sistema de mineração de padrões de navegação *Web Utilization Miner* [Spiliopoulou e Faulstich 98]. Esta estrutura é construída combinando os prefixos comuns dos caminhos percorridos pelos utentes, de modo a que cada nó da árvore retrate uma subsequência de navegação distinta, a partir da raiz para uma página. A proposta de [Pei et al. 00] também usa uma estrutura similar, designada árvore WAP *(Web Access Pattern)*, na qual apenas são inseridas sequências frequentes, após ou durante a mineração de padrões sequenciais.

Os padrões sequenciais fornecem indicações mais precisas, do que as regras de associação, nomeadamente, para efeitos de compreensão do comportamento de navegação, controlo da navegação, avaliação do desenho do sítio e personalização ou ajuste dinâmico da interacção. No entanto, os padrões sequenciais são mais difíceis de obter e possuem, normalmente, um nível de cobertura inferior. Segundo [Mobasher et al. 02], padrões mais restritivos, como os sequenciais contíguos, são mais apropriados para tarefas de prognóstico, em que a principal meta é prever, com precisão, as próximas acções imediatas do visitante. Esta adequação confirma-se, sobretudo, em sítios com elevado número de páginas dinâmicas, nos quais, muitas vezes, um caminho de navegação contíguo traduz uma sequência de acções do visitante, com significado semântico, cada dependente dos actos anteriores. Porém, padrões menos restritivos, como conjuntos frequentes, regras de associação e padrões sequenciais (não contíguos), são alternativas mais efectivas no contexto da personalização da Web e de sistemas de recomendação, em que o alvo reside em oferecer um conjunto de sugestões com cobertura ampla.

Quanto à classificação, refere-se ao processo de atribuir um novo item, a uma de várias classes predefinidas, com base no conhecimento prévio acerca das mesmas. Para esse efeito, é necessário

deduzir um modelo para as classes e regras de classificação, tentando desvendar as características que melhor descrevem e distinguem os elementos de cada classe. As regras derivadas são aplicadas na classificação de novos elementos e são usadas para melhorar o entendimento das classes. Em WUM, as classes podem consistir em diferentes perfis ou segmentos predefinidos de visitantes e determinados comportamentos alvo (e.g. compra), cujo reconhecimento e caracterização interessa efectuar, em função de diversas variáveis, como, por exemplo, padrões de navegação ou atributos pessoais dos visitantes. Este método também pode ser aplicado para classificar páginas ou sequências de páginas, no sentido de facilitar a análise de padrões de navegação e de permitir a comparação destes com os perfis de visitantes.

Ao contrário da classificação, o agrupamento é um método de mineração usado para formar grupos de itens semelhantes entre si, sem qualquer conhecimento prévio de quais serão os grupos formados, nem das propriedades que os caracterizam. Os grupos extraídos serão objecto de uma análise, conducente à determinação das afinidades que melhor representam os elementos de cada grupo. Para além do agrupamento de visitantes, em função de características próprias, interesses demonstrados ou comportamentos exibidos, existem dois tipos preponderantes de agrupamento, realizados com base em dados de *clickstream* e esquematizados na Figura 5:

- agrupamento de sessões ou episódios;
- agrupamento de páginas visitadas.

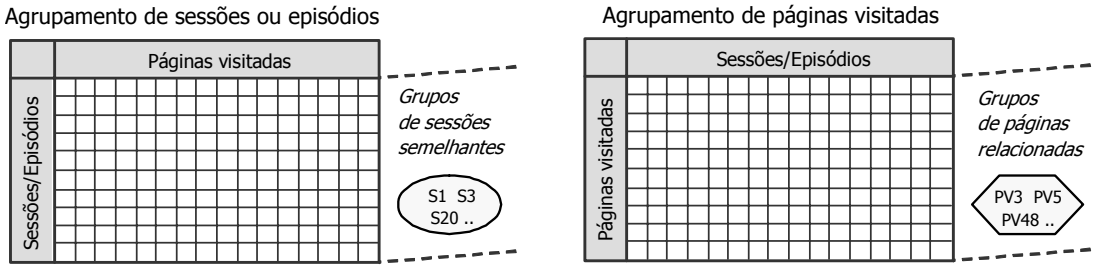


Figura 5 – Esquema de agrupamento de sessões e de páginas visitadas

O agrupamento de sessões ou episódios tende a estabelecer grupos significativos de tipos de visitas ou visitantes, com comportamento comum. Estes grupos reflectem as formas alternativas de navegação adoptadas pelos visitantes e podem prestar indicações sobre as suas metas e motivações. A identificação e caracterização de tais segmentos são especialmente úteis para levar

a cabo tarefas de personalização e no âmbito da "inteligência" para o negócio. O agrupamento de páginas visitadas procura determinar grupos de páginas que parecem estar relacionadas, segundo o uso ou percepção dos visitantes. Geralmente, os grupos obtidos resumem interesses comuns de utentes, apesar de estes poderem ter comportamentos de navegação distintos. O resultado deste tipo de agrupamento é comparável à extracção de regras de associação, embora omita a informação quantitativa (suporte e confiança) e qualitativa (antecedente e consequente) adicional que estas concedem por inerência [Koutri et al. 05]. Por conseguinte, o uso mais comum do agrupamento de páginas visitadas corresponde à alteração estática de páginas ou do sítio.

O agrupamento é um método que tem recebido grande atenção em muitos trabalhos. Em termos gerais, em contraste com a descoberta de regras de associação ou de padrões sequenciais, o agrupamento proporciona um mecanismo flexível para tarefas como a personalização, apesar de nem sempre conduzir a um nível superior de precisão nas recomendações. A flexibilidade provém do facto de muitos atributos inerentes às páginas acedidas poderem ser considerados no processo de mineração, tais como a duração do acesso e atributos dos objectos subjacentes [Mobasher 04]. Acresce o facto de o agrupamento, e também a classificação, serem métodos que podem ajudar o planeamento e execução de estratégias de marketing futuras, tipicamente mais sofisticadas ou direccionadas, tanto *on-line* como *off-line*.

A apresentação dos métodos de DM, vulgarmente explorados em WUM, evidencia as múltiplas aplicações e sobreposição dos mesmos, relativamente aos tipos de problemas a que podem dar resposta. O nível de adequação de cada método varia, no entanto, em função dos vários factores implicados, introduzindo várias questões subtis, passíveis de influenciar substancialmente a metodologia a adoptar, no que respeita à selecção de métodos de DM a usar.

2.4.5 Desafios no Desenvolvimento de Processos de WUM

Tal como qualquer outro processo de KDD, a WUM envolve, também, as diversas dificuldades provenientes da complexidade das ferramentas de DM, dos paradigmas subjacentes e das actividades a realizar. Um desafio frequente, crucial e emergente nestes exercícios é a selecção dos métodos a aplicar, na resolução de um determinado problema de análise de dados, de modo a alcançar resultados com utilidade para um dado fim específico. Esta escolha tem, naturalmente, grande impacto nos resultados atingidos, não existindo princípios simples e gerais que auxiliem essa decisão, já que a apropriação dos métodos de DM varia em função de distintos tipos de factores, podendo estes ser complexos, subjectivos e contraditórios.

Para que seja possível tirar proveito das capacidades de DM é necessário possuir um conhecimento técnico profundo acerca dos métodos. A escolha de funções de DM, apesar de ser mais intuitiva e de abranger um número mais reduzido de opções, não é tão trivial como se poderia pensar, dependendo, antes de mais, da correcta reformulação de um problema de negócio num problema de DM. Normalmente, os analistas de negócios sabem o que pretendem, mas não têm experiência na tradução dos objectivos de negócio na combinação certa de técnicas de DM. A dificuldade fulcral advém da sobreposição dos métodos (funções e modelos), relativamente aos tipos de problemas que podem solucionar. A eleição de modelos prende-se, ainda mais, com as características dos dados alvo e os requisitos da análise, a observar ao longo do estudo. Cada modelo de DM reveste-se de propriedades específicas e baseia-se em diferentes pressupostos, em relação às características dos dados a que pode ser aplicado. A sua adequação varia também em consonância com os requisitos subjacentes, os quais poderão ser subjectivos e, mesmo, contraditórios. Por exemplo, no contexto da função de classificação, as árvores de decisão são modelos descritivos, sendo mais indicados para efeitos de interpretação da influência das variáveis explicativas, do que modelos de redes neuronais. O primeiro modelo será, portanto, preferível se a finalidade da análise for entender os factores influentes, mesmo que a precisão do segundo modelo seja superior. Todavia, esta afirmação é discutível, se a precisão também for importante para o analista, o que é vulgar em tarefas de classificação. Nestas circunstâncias os dois critérios de avaliação não são convergentes, sendo necessário estabelecer prioridades. Além disso, também se constata que a interpretabilidade não é um critério de avaliação objectivo.

Mesmo pressupondo a disponibilidade de fontes de dados ideais, têm de ser levadas a cabo algumas transformações complementares, para que as ferramentas de DM possam actuar da melhor forma. Tais transformações vão ao encontro de requisitos genéricos e específicos de ferramentas, funções e modelos a aplicar, assim como das próprias restrições das aplicações ou problemas a que se pretende dar resposta. A execução dessas operações influencia substancialmente as propriedades dos dados utilizados e os resultados do processo. A aplicação dos métodos seleccionados contempla ainda a configuração de uma série de parâmetros, geralmente, ajustados iterativamente, até que se atinjam os valores considerados óptimos.

Dada a sobreposição dos métodos de DM e a incerteza quanto ao melhor deles para resolver um determinado problema, num certo contexto, é prática comum, neste domínio, experimentar várias funções e modelos, com vista a testar e comparar os seus resultados. A prossecução sucessiva da maioria das actividades referidas, em conjunto com todas as dificuldades inerentes, contribui, em muito, para a morosidade, ineficácia e insucesso de numerosos exercícios de KDD.

No âmbito da WUM a extracção de conhecimento comporta dificuldades acrescidas. Os dados de *clickstream* são, geralmente, mais volumosos, contêm um grande número de variáveis, são intrinsecamente complexos e registam aspectos comportamentais, assumindo, por este motivo, um carácter subjectivo e tornando-os passíveis de interpretações subtis. O conteúdo rico destes dados, os muitos atributos potencialmente úteis e, ainda, as diferentes abstracções que os objectos de análise podem adoptar, fazem com que a própria escolha do conjunto certo de atributos e registos, relevantes e representativos, para o problema em questão, seja mais problemática. À necessidade de conhecer a área de estudo e de perícia técnica, junta-se ainda outra dimensão imprescindível: o conhecimento sobre o arranjo físico e lógico do sítio, requerido ao longo de todo o processo de WUM. O desafio da selecção de métodos de DM é, consequentemente, ampliado nestas circunstâncias. A profusão de algoritmos e de abordagens de exploração, as interpretações mais diversas e elaboradas dos resultados e a variabilidade das particularidades das suas aplicações accionáveis contribuem, também, para este facto. Adicionalmente, toda a estratégia subjacente na escolha de métodos de DM encontra-se menos estudada e estruturada no âmbito da WUM. Nesta estratégia enquadram-se a identificação sistemática dos tipos de problemas de WUM e a determinação dos respectivos elementos de dados chave, actividades de mineração consentâneas e categorias de aplicações práticas das descobertas. Pode-se mesmo afirmar que estes vectores se encontram em constante evolução, fazendo com que a WUM surja como um alvo móvel, difícil de disciplinar. O próprio sítio Web é um elemento em constante mutação e a origem de um volume colossal de dados que exigem grande atenção e um tratamento sistemático. Finalmente, as descobertas são facilmente accionáveis e o período temporal da pertinência das mesmas é curto, impondo uma pressão acentuada na celeridade e eficácia do processo de WUM, a qual não se compadece com desafios como os reportados.

2.5 Especificação PMML de Representação de Resultados de Processos de KDD

A extracção de conhecimento e o seu uso posterior decorrem em ambientes, tipicamente, muito heterogéneos e requerem a cooperação entre tipos distintos de aplicações. A representação normalizada e portável dos resultados produzidos por ferramentas de KDD é, por conseguinte, necessária e desejável para que estes resultados possam ser convenientemente analisados, visualizados e utilizados por vários tipos de aplicações, independentemente da plataforma de operação e do sistema que gerou tais resultados. A especificação *Predictive Model Markup*

Language (PMML) [WWW2] oferece um mecanismo universal e proeminente de descrição, representação e operacionalização de conhecimento induzido em processos de KDD, sendo significativamente explorada, de diferentes formas, ao longo desta dissertação. Trata-se de uma linguagem padronizada e extensível baseada na XML, para descrever e partilhar modelos estatísticos e de DM, entre aplicações aderentes à mesma. O padrão visa fornecer uma infra-estrutura, de modo a permitir que uma aplicação possa produzir um modelo e outra aplicação o consiga consumir, tão somente por meio da leitura e interpretação do documento PMML [Grossman et al. 02]. Tal como qualquer outro documento XML, um documento PMML é disponibilizado em formato texto portável e legível. Em acréscimo, a definição declarativa de modelos visa capturar a informação preponderante e conducente à operacionalização ou aplicação desses modelos.

A Figura 6, adoptada de [Kotásek e Zendulka 02], mostra as partes mais importantes de um documento PMML e os seus relacionamentos. Na Figura 7 apresenta-se um exemplo de um documento, contendo algumas dessas partes constituintes e a descrição de um modelo de regras de associação. Este exemplo torna-se útil para ilustrar os aspectos abordados ao longo da breve exposição acerca das características primordiais da especificação PMML. Um documento PMML é, por inerência, um documento XML possuindo um elemento raiz do tipo PMML. Alguns dos itens de informação geral, presentes na Figura 7, são a versão PMML e a descrição da aplicação ou ferramenta que gerou o documento, indicados, respectivamente pelo atributo *version* e pelo elemento *Application* (atributos *name* e *version*), pertencente ao cabeçalho *(Header)*.

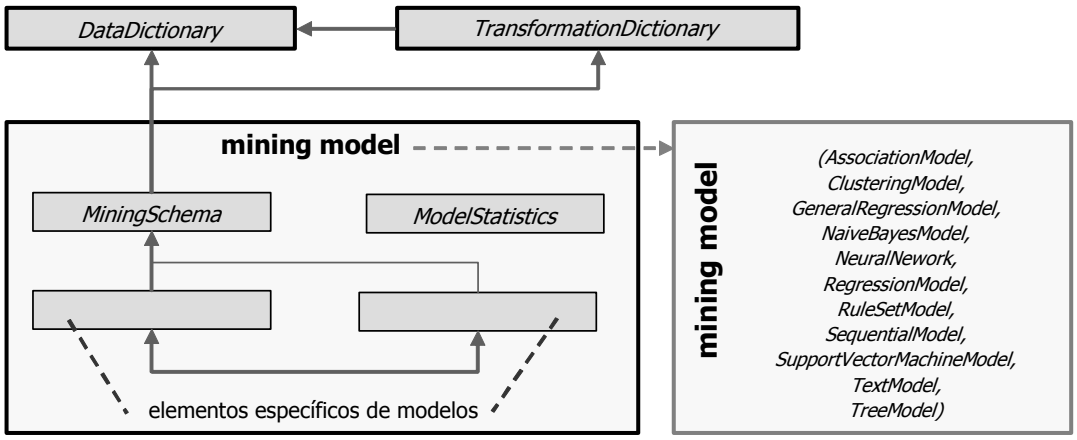


Figura 6 – Principais elementos constituintes de documentos PMML

Um documento PMML pode integrar vários modelos, partilhando os elementos *DataDictionary* e *TransformationDictionary*, relativos aos dicionários de dados e de transformação. Cada modelo incluído é representado por um elemento específico, consoante o seu tipo de modelo (e.g. elemento *AssociationModel* para modelos de regras de associação), e é genericamente referido pelo termo modelo de mineração *("mining model")*. O dicionário de dados *(DataDictionary)* especifica os campos ou variáveis *(DataField)* de entrada dos modelos de mineração e os seus tipos de dados e intervalos de valores. As operações aplicáveis aos campos do dicionário de dados dependem directamente do valor assumido pelo respectivo atributo *optype*. Os campos categóricos admitem o operador de igualdade (=), enquanto os numéricos possuem o operador adicional menor (<) e os contínuos permitem, ainda, o uso de operadores aritméticos. Estas definições são gerais e, portanto, independentes de modelos de mineração individuais. O dicionário de transformação *(TransformationDictionary)* reporta transformações simples de variáveis de entrada para campos derivados, concedendo maior flexibilidade na representação de esquemas de entrada de modelos. Cada modelo pode ainda incorporar as suas próprias variáveis derivadas. As funcionalidades de transformação disponíveis abrangem apenas operações básicas, envolvendo mapeamentos de valores, automaticamente executados pelo sistema de mineração, a fim de melhorar a geração e aplicação de modelos. Um exemplo típico consiste na normalização dos valores de entrada de um modelo de rede neuronal.

A especificação PMML faculta vocabulário para retractar vários tipos de modelos, cobrindo descrições de aspectos genéricos e específicos. Estas descrições são englobadas pelo elemento correspondente ao tipo de modelo (e.g. *AssociationModel*). As propriedades do modelo contemplam os seguintes atributos comuns:

- *modelName*, uma designação única, no âmbito do documento, do nome atribuído ao modelo;
- *functionName*, indica uma função de DM entre as possíveis (e.g. *associationRules*, *sequences*, *classification*, *regression*);
- *algorithmName*, uma descrição livre do algoritmo que produziu o modelo.

Já no que concerne a propriedades específicas, por exemplo, um modelo de regras de associação abarca atributos, tais como:

- número de transacções *(numberOfTransactions)* e de itens *(numberOfItems)*, contidos nos dados de entrada;
- número de conjuntos frequentes *(numberOfItemsets)* e de regras *(numberOfRules)* do modelo;

- valores mínimos de suporte *(minimumSupport)* e de confiança *(minimumConfidence)* que todas as regras satisfazem.

```
<?xml version="1.0" ?>

<PMML version="3.0" >

    <Header copyright="www.dmg.org" description="exemplo de modelo para regras de associação">
        <Application name="……." version="……"/>
    </Header>

    <DataDictionary numberOfFields="2" >
        <DataField name="transaction" optype="categorical" dataType="string" />
        <DataField name="item" optype="categorical" dataType="string" />
    </DataDictionary>

    <! -- elemento mining model -->
    <AssociationModel
        modelName="exemplo_RA" functionName="associationRules" algorithmName="Apriori"
        numberOfTransactions="4" numberOfItems="3" numberOfItemsets="3" numberOfRules="2"
        minimumSupport="0.6" minimumConfidence="0.5">

        <MiningSchema>
            <MiningField name="transaction" usageType="group" />
            <MiningField name="item" usageType="active"/>
        </MiningSchema>

        <! -- Existem: -->
        <! --          três itens nos dados de entrada -->
        <Item id="1" value="Cracker" />
        <Item id="2" value="Coke" />
        <Item id="3" value="Water" />

        <! --          dois conjuntos frequentes com um único item -->
        <Itemset id="1" support="1.0" numberOfItems="1">
            <ItemRef itemRef="1" />
        </Itemset>
        <Itemset id="2" support="1.0" numberOfItems="1">
            <ItemRef itemRef="3" />
        </Itemset>

        <! --          um conjunto frequente com dois itens -->
        <Itemset id="3" support="1.0" numberOfItems="2">
            <ItemRef itemRef="1" />
            <ItemRef itemRef="3" />
        </Itemset>

        <! --          duas regras que satisfazem os requisitos -->
        <AssociationRule support="1.0" confidence="1.0" antecedent="1" consequent="2" />
        <AssociationRule support="1.0" confidence="1.0" antecedent="2" consequent="1" />
    </AssociationModel>

</PMML>
```

Figura 7 – Exemplo de um documento PMML

O elemento genérico *MiningSchema* lista o subconjunto de campos do dicionário de dados usado no modelo, bem como o papel que cada um desempenha, com vista a enumerar as variáveis a

proporcionar para que este possa ser executado. Os papéis desempenhados são indicados pelo atributo *usageType* e residem nas seguintes possibilidades:

- activo *(active)*, para campos efectivamente usados como entrada;
- predito *(predicted)*, relativo a variáveis alvo, cujo valor o modelo visa predizer;
- suplementar *(supplementary)* respeitante a atributos que contêm informação adicional, usualmente utilizados para calcular variáveis derivadas;
- agrupamento *(group)* ou ordenação *(order)*, para denotar atributos que realizam funções de organização dos dados.

Por exemplo, no modelo ilustrado, o campo *transaction* foi usado para agrupar os itens presentes no atributo de entrada *item* em transacções. Já o elemento *ModelStatistics* também é genérico e contém estatísticas globais univariadas de um subconjunto dos campos de mineração. Cada tipo de modelo possui ainda elementos específicos, tais como, itens *(Item)*, conjuntos frequentes *(Itemset)* e regras *(AssociationRule)*, para modelos de regras de associação.

A especificação PMML é mantida pelo *Data Mining Group* (DMG) [WWW2], um consórcio voltado para o desenvolvimento de padrões de DM e formado por várias empresas com um papel proeminente neste mercado. O padrão recebeu grande aceitação e é suportado por muitos sistemas comerciais e académicos, assim como por outros padrões da área.

2.6 Facetas dos Dados de *Clickstream* e da WUM

Criar e consolidar uma presença efectiva na Web é um anseio vulgar das organizações detentoras de sítios Web, mas que se revela difícil de atingir, num meio sobrecarregado e dotado de exposição e competitividade global. A Web faculta novas formas promissoras para conduzir as actividades e divulgar informação, porém também requer meios apropriados para avaliar e melhorar o retorno dos esforços humanos e financeiros investidos. Uma reacção rápida a todos os sinais e estímulos parece ser a única forma viável para enfrentar os múltiplos obstáculos colocados neste contexto, os quais condicionam a maximização das oportunidades potenciais da Web.

A mineração de utilização da Web é um processo de extracção de conhecimento, focado nos dados de *clickstream* ou de utilização da Web, respeitantes aos registos derivados a partir da interacção de visitantes com a Web. A mineração destes dados, usualmente conjugados com outros dados relacionados, vem corresponder às necessidades dos promotores de sítios Web, permitindo descobrir padrões de utilização da Web, cujas aplicações se prendem com o apoio a agentes de

decisão, em áreas como o melhoramento de sistemas, alteração de sítios, personalização da Web e "inteligência" para o negócio.

A acessibilidade a um manancial de dados, capturados sem ter de questionar directamente os utilizadores, pelo contrário, de modo automático e implícito, cobrindo todo o processo de interacção, que antecede e envolve a eventual concretização de transacções é, sem dúvida, uma oportunidade inestimável para os métodos de exploração de dados. Noutras circunstâncias, como o comércio tradicional, dados com detalhe e abrangência semelhantes seriam muito difíceis, senão impossíveis, de obter. A abundância de dados, com descrições ricas, satisfaz, por inerência, pré-requisitos da extracção de conhecimento, criando um ambiente fértil para a aplicação de métodos de mineração como os abordados. A elevada magnitude e a complexidade destes dados também impõe um desafio, reforçando a premência da WUM como o único instrumento verdadeiramente capaz de o enfrentar, transformando tais dados em conhecimento de valor e accionável, para basear melhoramentos do sítio que conduzam a retornos.

Os dados de *clickstream* também se revestem de limitações e características específicas, estando na origem de actividades de transformação e integração de dados peculiares e laboriosas, de algoritmos de mineração baseados nestes dados, bem como de adaptações de tecnologias, abordagens e ferramentas aos problemas da WUM. Vários modelos e paradigmas têm sido adoptados para sustentar a WUM, no sentido de responder aos diferentes problemas colocados neste âmbito. As abordagens baseadas em tecnologias de base de dados, em particular, em sistemas de *Data Warehousing* e *Webhousing*, vêm ao encontro de exigências estratégicas, tais como proporcionar um conjunto de dados integrado, detalhado, com qualidade e organizado para níveis de acesso de elevado desempenho, no que concerne ao propósito específico da análise de dados. Outras abordagens estão voltadas para a captura directa da semântica da navegação na Web, fundamentando-se na criação de infra-estruturas, que permitam derivar métodos eficientes e mais direccionados, aos dados de *clickstream* e aos requisitos de estudos mais específicos.

O desenvolvimento de processos de WUM engloba as dificuldades consideradas inerentes a qualquer processo de descoberta de conhecimento, comportando ainda contratempos acrescidos. Um desafio geral, conhecido e comum nestes processos consiste na selecção de métodos a usar para alcançar uma determinada meta de análise e gerar resultados úteis para esse fim. Este desafio agrava-se no âmbito da WUM, dado o nível bem inferior do estudo e estruturação das respectivas categorias de problemas, tipos de actividades de mineração, aplicações práticas relacionadas e elementos de dados chave. Destaca-se ainda toda a problemática que advém do

elevado volume e dimensionalidade dos dados e a natureza mais rica e subtil dos mesmos, os quais dificultam substancialmente todo o processo de extracção de conhecimento.

Os modelos resultantes de processos de WUM, tal como de exercícios genéricos de mineração de dados, podem ser expressos através de documentos no formato padrão PMML. Este tipo de representação é preponderante, sobretudo pelo facto de ser normalizado, portável e produzido automaticamente pelas ferramentas de mineração aderentes ao padrão, sendo utilizado ao longo da dissertação, conforme se discute nos próximos capítulos.

Pelo exposto, conclui-se que a mineração de utilização da Web pode ser um instrumento de grande utilidade e valia para as organizações empenhadas na optimização dos seus sítios, existindo, contudo, diversos desafios a vencer para a sua aplicação efectiva. Mesmo supondo a disponibilidade de fontes de dados ideais, a complexidade da exploração de capacidades da WUM requer a ajuda de especialistas na área. Subsistem os obstáculos, sem solução fácil e aparente, de colocar a WUM ao serviço dos agentes de decisão, de forma mais eficaz e expedita, para fazer face às solicitações de apoio à decisão em tempo útil, o qual é manifestamente curto no âmbito da Web.

Capítulo 3

Assistência à Mineração de Dados

3.1 Abordagens de Apoio no Desenvolvimento e Aplicação de Processos de Mineração

O desenvolvimento e aplicação de processos de KDD são actividades complicadas, intensivas, interactivas, iterativas e morosas. Os procedimentos envolvidos requerem a aplicação sucessiva de métodos distintos e a sua avaliação em função dos resultados obtidos para diferentes parâmetros de configuração, consoante o fim a que a análise de dados se destina. Inclui-se ainda nestes procedimentos a utilização de operações de transformação de dados, conducentes à obtenção de melhores resultados, considerando os requisitos dos métodos a aplicar e os ditados pela própria análise. Uma vez que grande parte destas actividades não é passível de automatização, uma forma para lidar com esta complexidade consiste em assistir o analista na condução destes processos. O intuito dos sistemas de apoio ao desenvolvimento e aplicação de processos de KDD é, tipicamente, abstrair a complexidade e evitar a execução exaustiva destes procedimentos, designadamente para todos ou um subconjunto significativo de métodos de mineração aplicáveis. O objectivo subjacente é conduzir o analista ao sucesso no seu estudo, sem um esforço excessivo, o qual, muitas vezes, se revela infrutífero. As abordagens seguidas por estes sistemas assumem formas distintas, as quais podem ser sistematizadas ao longo de duas dimensões essenciais (Figura 8): âmbito do apoio e principal orientação.

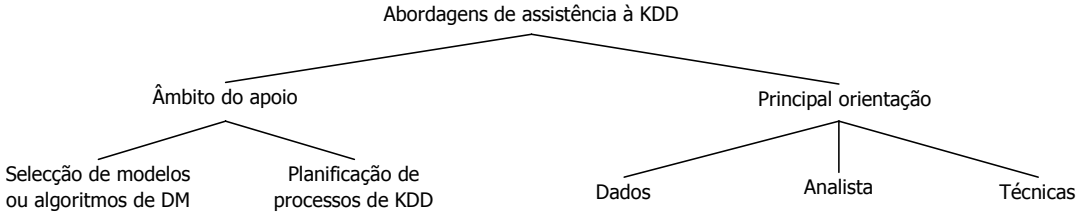


Figura 8 – Principais dimensões de classificação de abordagens de apoio a processos de KDD

O âmbito do apoio prende-se com o nível de abrangência das tarefas assistidas e com o propósito da ajuda prestada, sendo as suas vertentes fundamentais o auxílio específico na selecção de modelos ou algoritmos de DM e o auxílio, mais geral, na planificação de processos de KDD. A selecção de modelos foca, directamente e somente, parte da fase de modelação, partindo de uma situação mais concreta e previamente estabelecida, com a finalidade primordial de recomendar um ou mais modelos adequados alternativos, entre um conjunto de modelos de DM. A planificação de processos de KDD possui um âmbito mais alargado, podendo contemplar várias das suas fases e a aplicação de múltiplos métodos de DM e outras operações, situando-se em níveis diversos e mais elevados de abstracção. O seu ponto de partida poderá ser coincidente com o estado inicial do processo e a sua missão consiste, normalmente, em fornecer um ou mais planos de mineração, em que cada sugere uma sequência de actividades apropriadas a levar a cabo. É importante frisar que podem ser recomendadas diversas soluções alternativas em ambas as vertentes. Não obstante, o que está em causa na selecção de modelos é a aplicação individual de um único modelo, enquanto a planificação de processos cobre situações de selecção mais complicadas, podendo dar lugar à aplicação interligada de vários modelos de DM e outros tipos de operações.

A orientação da abordagem de ajuda relaciona-se com os factores influentes, nos quais se baseia a assistência e o respectivo modelo de decisão subjacente. As orientações predominantes centram-se nos dados, no analista e nas técnicas, mais propriamente:

- nas características dos dados disponíveis, as quais traduzem restrições implícitas;
- nos requisitos explícitos da análise, obtidos a partir do analista;
- nas propriedades específicas de uma série de técnicas, as quais condicionam a sua aplicação.

Uma vez que as abordagens de assistência podem abarcar, simultaneamente, diversas orientações, evidencia-se, vulgarmente, a orientação central ou mais proeminente.

Outra dimensão de caracterização dos sistemas de apoio à actividade de KDD corresponde à estratégia que suporta o sistema na realização da função a que se propõem, nomeadamente na construção de recomendações. Estas são variadas e serão brevemente discutidas nas duas secções que se seguem.

3.1.1 Auxílio na Selecção de Modelos

O âmbito do apoio na selecção de modelos ou algoritmos é entendido como a recomendação de um ou mais modelos (e.g. C5.0, ID3 e MLPN), pertencentes a diferentes ou à mesma classe de modelo (e.g. árvores de decisão e redes neuronais), mas num contexto restrito, no que concerne a outros aspectos. Neste nível de apoio, vários passos do processo já foram definidos [Verdenius e Engels 97] e determinadas questões, como, por exemplo, a escolha de funções de DM (e.g. classificação e agrupamento), já não se colocam. A individualização deste âmbito deve-se, especialmente, à preponderância dos problemas de classificação e à diversidade de modelos aplicáveis aos mesmos e, por conseguinte, aos numerosos esforços de investigação devotados à sua resolução e assistência.

A recomendação de modelos é uma linha muito activa de investigação em meta-aprendizagem, incidindo, sobretudo, no contexto de problemas de classificação e regressão e em orientações centradas nos dados. Alguns trabalhos conjugam também a influência dos requisitos da análise. Muitas das iniciativas neste domínio estão relacionadas com os projectos STATLOG [Michie et al. 94] e MetaL [WWW3]. O projecto STATLOG envolveu a aplicação sistemática de vários algoritmos de classificação, visando determinar as circunstâncias que favorecem a utilização de certos modelos, em detrimento de outros. [Brazdil et al. 94] e outros, em trabalho independente [Aha 92] ou subsequente [Gama e Brazdil 95], usaram meta-regras, derivadas de estudos experimentais, para ajudar a prever os algoritmos mais ajustados. As regras atendiam a características mensuráveis dos dados, incluindo medidas simples, estatísticas e de informação teórica. De acordo com a formulação de meta-aprendizagem adoptada neste projecto, dado um novo conjunto de dados, é sugerida uma série de modelos aplicáveis, em função do nível de desempenho esperado, definido em termos da precisão prevista. A vantagem básica da abordagem utilizada reside no procedimento automático para produzir conhecimento acerca de novos algoritmos e, consequentemente, na possibilidade de actualização do modelo de recomendação e na independência em relação a uma série específica de técnicas. Todavia, são requeridos estudos experimentais exaustivos para testar os novos algoritmos em todos os conjuntos de dados usados.

São ainda apontadas várias limitações aos resultados produzidos e à própria abordagem adoptada no projecto, tais como:

- número limitado de exemplos para efeitos de generalização;
- uso da precisão como único critério de desempenho;
- formulação de meta-aprendizagem de subdivisão dos algoritmos entre aplicáveis ou não, ao invés da indicação do(s) melhor(es).

No seguimento, foram conduzidos múltiplos trabalhos, procurando estender e refinar a abordagem do projecto STATLOG, para ultrapassar as suas limitações, designadamente no contexto do projecto MetaL.

O projecto MetaL prendeu-se com o desenvolvimento de métodos e ferramentas para assistir utilizadores de tecnologias de aprendizagem automática e de DM. Um dos resultados deste projecto foi o protótipo do sistema de assistência *Data Mining Assistant* (DMA) [WWW3]. Esta ferramenta proporciona recomendações sob a forma de uma lista ordenada de algoritmos candidatos, segundo uma combinação ponderada de vários critérios de desempenho. A associação da caracterização de conjuntos de dados a medidas de desempenho é efectuada combinando esquemas de ordenação com variações do algoritmo vizinho mais próximo. Outras abordagens de construção de recomendações recorrem a métodos de classificação e regressão.

Entre os muitos esforços de investigação e desenvolvimento associados ao projecto MetaL, destacam-se alguns contributos mais relacionados com o sistema proposto nesta dissertação. [Lindner e Studer 99] sugeriram o uso de um sistema baseado no paradigma CBR, para representar os problemas de meta-aprendizagem e realizar o processo indutivo, a fim de seleccionar o modelo mais apropriado. O sistema proposto engloba um componente de análise de restrições de aplicação [Engels et al. 97] e uma ferramenta para extracção automática das características dos dados, denominada DCT *(Data Characterization Tool)* [Engels e Theusinger 98]. Esta ferramenta estendeu o conjunto de elementos de caracterização de dados, definido no projecto STATLOG, tendo sido usada em muitos outros trabalhos da área. De acordo com a perspectiva defendida em [Lindner e Studer 99], um caso é descrito pelas restrições do estudo, metadados do conjunto de dados e conhecimento sobre a aplicação de modelos, materializado sob a forma de metadados, relativos a modelos e à experiência da sua utilização. A inclusão de características de algoritmos e de restrições da análise traduz um contributo, em relação a outros trabalhos, baseados apenas em caracterizações dos dados. Adicionalmente, este trabalho antevê a necessidade de refinar o modelo de representação de casos e de acrescentar metadados respeitantes à configuração de modelos.

Outro trabalho, igualmente baseado no paradigma CBR e no contexto do projecto MetaL, trata-se de um assistente para selecção de modelos [Hilario e Kalousis 01], o qual integra duas características relevantes. A primeira consiste na combinação de conhecimento acerca de modelos com características de conjuntos de dados, com vista a simplificar a selecção de modelos, focando as técnicas mais promissoras, face a restrições e preferências definidas pelo utilizador. A segunda corresponde à incorporação de um mecanismo para distinguir entre diferentes parâmetros dos modelos de cada ferramenta de KDD, estendendo a abordagem de recomendação, para contemplar a selecção de modelos e as configurações específicas de parâmetros. O sistema foi implementado recorrendo a uma ferramenta CBR comercial e explora uma base de casos (multi-) relacional. Uma particularidade interessante do modelo de representação de casos reside na inclusão de informação sobre variáveis individuais de conjuntos de dados. Este detalhe, e o seu tratamento nos procedimentos de cálculo, permite ultrapassar as limitações de formulações proposicionais que, até então, obrigavam a simplificações, como o uso de valores médios, e condicionavam a meta-aprendizagem. Alternativamente, podem ser usados algoritmos indutivos de primeira ordem [Todorovski e Dzeroski 99], com a finalidade de garantir o uso pleno das caracterizações ao nível de variáveis individuais. Apesar dos seus contributos, tal como no trabalho anterior, este sistema incide no problema da selecção de modelos de classificação, não abarcando operações de transformação, nem a aplicação composta de vários modelos, referentes a diferentes funções de mineração.

O sistema pericial Consultant [Craw et al. 92] (ou MLT-Consultant), desenvolvido no âmbito do projecto MLT (*Machine Learning Toolbox*) [Kodratoff et al. 92], enquadra-se numa perspectiva de selecção de um modelo entre os disponíveis, sendo um exemplo da orientação centrada e dependente de técnicas [Verdenius e Engels 97]. O sistema complementa uma arquitectura integrada de dez ferramentas (modelos) de aprendizagem automática, tendo sido construído com o propósito de guiar a utilização destas ferramentas. Este apoio é proporcionado aplicando regras, as quais descriminam entre o conjunto possível de modelos aplicáveis, e questionando o utilizador acerca de diversos aspectos, entre os quais, a tarefa que pretende resolver, a natureza dos dados e conhecimento de fundo. As recomendações baseiam-se em conhecimento diferencial ou comparativo entre os modelos existentes, significando que a adição de um novo algoritmo, ao conjunto inicial, requer a redefinição da base de conhecimento, condicionando a extensibilidade do sistema. Além destas limitações, o sistema assume que uma boa definição do problema é explicitamente conhecida pelo analista, não contemplando o entendimento do mesmo no seu domínio de aplicação. Por este motivo, depende fortemente da interacção com o utilizador para a escolha de modelos ajustados.

3.1.2 Auxílio na Planificação de Processos

A questão da selecção de modelos, anteriormente abordada, é proeminente, mas é limitada. Alguns aspectos de desenho de alto nível e dos estágios iniciais do desenvolvimento do processo não são abrangidos [Verdenius e Engels 97]. Além disso, os problemas reais são complexos e nem sempre podem ser resolvidos pela aplicação isolada de uma técnica, exigindo, por vezes, a aplicação sucessiva de múltiplos métodos de DM e de outras operações. A planificação de processos cobre o seu desenvolvimento, neste contexto e a diversos níveis, podendo auxiliar diferentes etapas e tarefas da KDD. Por exemplo, o projecto *Mining Mart* [Morik e Scholz 04] diz respeito ao passo de transformação, englobando a aplicação de algoritmos de aprendizagem automática utilizados neste âmbito. A meta primordial do projecto é a reutilização de processos de pré-processamento, considerados bem sucedidos e construídos por utilizadores experientes, recorrendo, para tal, a um repositório de metadados baseado em casos. Os metadados que descrevem os dados e os processos de pré-processamento são organizados em ontologias, com o intuito de aumentar o seu nível de generalização e, deste modo, facilitar e promover a sua reutilização. A exploração destes processos envolve a pesquisa na base de metadados, daqueles que parecem ser mais adequados para o problema actual, mas sem tirar partido das potencialidade do meta-modelo, nem de funcionalidades típicas de sistemas CBR, para ajudar os utilizadores a estabelecer a correspondência entre o processo actual e os existentes. Esta procura é conduzida pelo utilizador que, em seguida, descreve o mapeamento entre o problema actual e os problemas prévios. Com base neste mapeamento, o sistema gera passos de pré-processamento, podendo estes ser executados automaticamente. Este sistema possui ainda um nível de integração elevado com sistemas de base de dados, focando-se em conjuntos de dados reais e volumosos.

[Engels 96] [Engels et al. 97] [Engels et al. 98] apresentam uma abordagem (*User Guidance Modelling/Module* - UGM) que visa suportar a selecção e combinação de métodos de DM, assim como todo o processo de KDD, guiando o analista na especificação da descrição do problema e na construção de planos de processos de DM. A abordagem fundamenta-se na decomposição sistemática de tarefas, em sub-tarefas mais simples, ao longo de várias etapas de refinamento de um processo de DM de alto nível, ajudando o analista a desenvolver o melhor plano, com base num conjunto de operações. Algumas pré e pós condições são usadas para caracterizar (sub)tarefas, métodos e componentes reutilizáveis, definindo a sua funcionalidade e viabilizando o mapeamento entre estes. A orientação predominante da abordagem é o analista, mas conjugando,

também, técnicas para descrever as características dos dados disponíveis, através de uma ferramenta DCT [Engels e Theusinger 98]. A caracterização dos dados é explorada na sugestão de passos de pré-processamento, desempenhando, ainda, um papel complementar na selecção de métodos aplicáveis, de acordo com as diferentes tarefas que podem ser realizadas. A execução dos planos concluídos é assegurada, através da ferramenta comercial de KDD *SPSS Clementine* [WWW4].

O sistema *Intelligent Discovery Electronic Assistant* (IDEA) [Bernstein e Provost 01] oferece assistência, em certa medida, semelhante à abordagem anterior, mas seguindo uma direcção bem diferente. O analista define, basicamente, a tarefa pretendida (e.g. classificação), as preferências relativamente a critérios de desempenho e os dados a analisar, a partir dos quais são extraídos metadados relevantes. O IDEA inclui dois componentes essenciais. O primeiro gera planos de KDD, recorrendo a uma ontologia, para auxiliar a construção de uma lista de processos válidos e apropriados para a tarefa solicitada, com base em metadados que descrevem os dados e as operações. O segundo componente procede à ordenação dos processos válidos, através de heurísticas e atendendo às preferências do analista. Ao contrário do trabalho anterior, o sistema baseia–se no argumento de que é muito complicado discernir qual é o melhor plano de DM, dado que os resultados de DM são imprevisíveis e os requisitos da análise podem ser difíceis de conciliar e especificar, nomeadamente de maneira completa e precisa e no início do processo. Neste sentido, o sistema incorpora, exaustivamente, todos os métodos potencialmente úteis, acrescentando operações de transformação para gerar sequências válidas de operações. A lista de planos válidos é apresentada ao analista, ajudando-o a escolher entre estes, recorrendo às heurísticas de ordenação.

3.2 Abordagem de Assistência Adoptada

Os desafios que motivaram o presente trabalho, tal como se discutiu no capítulo anterior, requerem uma resposta abrangente, sugerindo a adopção de uma abordagem de apoio ao nível da planificação de processos, integrando, simultaneamente, as virtudes das orientações centradas nos dados e no analista. Não obstante, ao invés de uma abordagem generalista, é requerido um tratamento dirigido, especialmente voltado para a WUM e para os tipos de problemas que se levantam neste contexto. A principal missão desta iniciativa é a recomendação dos planos de mineração mais adequados para resolver um determinado problema de WUM. Para este efeito, é necessário aceitar, como entradas, descrições de problemas de WUM, em termos dos dados alvo e

dos requisitos de análise. Os dados alvo têm de ser convertidos em caracterizações importantes, para propósitos como a selecção de métodos de mineração. A descrição de requisitos da análise deve ser simplificada, nomeadamente, de forma a poder ser efectuada por intermédio de abstracções relacionadas com os problemas reais a resolver, em detrimento da especificação das tarefas de DM que têm de ser levadas a cabo. Nos sistemas com desígnios afins a tarefa de DM pretendida é tipicamente definida indicando uma função de DM. De acordo com a perspectiva defendida, por um lado, esta indicação faz parte da solução, dada a conhecida sobreposição de funcionalidades ao nível das próprias funções de DM e, por outro lado, pode ser necessário recorrer a mais do que uma função para resolver um problema. As saídas ou recomendações a produzir consistem em planos de WUM, contendo, basicamente, cadeias de métodos de DM e operações de transformação, acompanhados de informação complementar. Os aspectos a priorizar no conteúdo das saídas englobam todos os detalhes que possam explicar e guiar a implementação prática dos planos, bem como indicações da sua utilidade e aplicação prática.

A abordagem seguida coloca especial ênfase no passo de modelação, abrangendo não só a selecção de modelos, como a escolha das próprias funções de DM, podendo ser denotada como selecção de métodos. A fase de transformação também é contemplada, mas de modo simplificado, designadamente quando comparada com o projecto *Mining Mart*, e supondo a disponibilidade de fontes de dados pré-processados e com qualidade. Quanto às restantes etapas do processo de KDD, considerou-se algumas das suas actividades, como, por exemplo, a identificação dos dados relevantes para a análise, da fase de compreensão, e a apreciação dos resultados obtidos, do passo de avaliação. Entende-se ainda que a perspectiva de abstracção da complexidade, usada na descrição de problemas e de planos de WUM, e os aspectos evidenciados, quer nos planos como na informação suplementar, poderão contribuir significativamente para a simplificação de todo o estudo.

No que concerne à estratégia subjacente na construção de planos de WUM, argumenta-se que, para além de ser necessário fornecer várias alternativas plausíveis, os factores incertos ou problemáticos de especificar não devem afectar profundamente o conteúdo dos planos, pois a sua criação a partir de descrições imprecisas dificilmente conduzirá a processos de WUM excelentes. A exploração de uma grande variedade de processos válidos também não vai ao encontro do desígnio de proporcionar um número razoavelmente reduzido de opções efectivamente promissoras. Por um lado, a direcção seguida visa diminuir, e não ampliar, o número de possibilidades que devem ser testadas pelo utilizador, discriminando as melhores opções, no sentido de minimizar esforços, tipicamente já muito avultados. Por outro lado, processos válidos

não são, necessariamente, processos bem sucedidos, entendidos como soluções excelentes e com resultados comprovados. Defende-se que as potencialidades de gerar um processo válido, através da adaptação de um processo bem sucedido, são maiores do que o inverso. Isto porque, o desenvolvimento de um processo eficaz é um desafio reconhecido, enquanto a implementação de um processo válido, que não produz resultados úteis, para o fim a que se destina, é uma situação, indesejavelmente, demasiado comum, especialmente no que respeita a utilizadores inexperientes. Em suma, um processo bem sucedido, que produziu resultados comprovadamente úteis, retrata detalhadamente uma solução concreta para um determinado problema e, portanto, uma abordagem efectivamente promissora para resolver eficazmente novos problemas afins.

Constata-se que as actividades de KDD são, usualmente, realizadas recorrendo a experiência prévia no domínio. Na realidade, grande parte do sucesso obtido por especialistas, quando lidam com problemas de WUM, advém, directamente, da sua experiência e conhecimento adquirido (*know-how*), na resolução de problemas práticos concretos. Consequentemente, o ideal seria poder dispor de exemplos de processos de WUM bem sucedidos, desenvolvidos por utilizadores experientes e enquadrados num contexto de actividade comum. O acesso a exemplos de problemas similares resolvidos pode auxiliar todas as decisões envolvidas, pela reaplicação das suas melhores práticas a novas situações. Os tais factores influentes tornam-se sim úteis para encontrar e filtrar o subconjunto de processos mais plausíveis para solucionar o problema corrente, entre os exemplos de processos excelentes existentes, e, ainda, para fundamentar a decisão final da escolha nesse subconjunto. O recurso a tais processos excelentes também permite redefinir a noção de plano de WUM, para uma sequência de métodos de DM e outras operações, passíveis de serem instanciados com exemplos concretos da aplicação desses métodos e operações, e acompanhados pelos respectivos resultados e outra informação, a fim de facultar múltiplos pormenores informativos, tais como, contexto, configurações e descobertas. A Tabela 2 apresenta um resumo das principais características da abordagem defendida e adoptada.

Tabela 2 – Resumo das características primordiais da abordagem adoptada

Âmbito do apoio	Planificação de processos
Principais orientações	Características dos dados e requisitos explícitos do analista
Estratégia de suporte	Reutilização de processos úteis e bem sucedidos da organização
Ênfase da abordagem	WUM, fase de modelação, âmbito organizacional e abstracção da complexidade

Pelo exposto, a ideia defendida para enfrentar muitos dos desafios do desenvolvimento e aplicação de processos de WUM reside, particularmente, em gerir o conhecimento obtido a partir da experiência na resolução de problemas reais de WUM, no âmbito de um ambiente organizacional, com vista a criar a base para a partilha e reutilização de tal conhecimento, ao longo da organização. Para este efeito, é necessário construir um repositório de conhecimento, acerca de processos de WUM bem sucedidos da organização, constituindo-se, assim, uma memória centralizada e consolidada, com carácter didáctico e conducente à adopção de práticas efectivas, no que se refere à aplicação de DM sobre dados de *clickstream*. O âmbito organizacional permite, não só, enquadrar o conhecimento segundo os requisitos e especificidades da organização, como também, viabilizar o seu uso por parte de uma audiência alargada. Assim, torna-se, também, indispensável disponibilizar um mecanismo efectivo para ajudar os utilizadores a relacionar os novos problemas com as soluções promissoras existentes. Uma ferramenta de assistência baseada nesta ideia promove um uso mais eficiente e sinergético dos recursos da organização, permitindo, ainda, que outros utilizadores, para além dos especialistas, recorram à WUM de forma mais eficaz. Esta ideia foi concretizada explorando o paradigma CBR, uma vez que o mesmo concede, por inerência, o alicerce ideal para o tipo de apoio que se almeja oferecer e a resposta apropriada a vários requisitos essenciais, conforme se discute nas duas secções que se seguem.

3.2.1 Requisitos Primordiais na Assistência à WUM

A principal missão estabelecida para o sistema proposto é assistir os utilizadores na resolução de problemas de WUM, através da sugestão dos planos de mineração mais ajustados e com base em exemplos ou casos de aplicação de WUM bem sucedidos da organização. Para cumprir esta missão, o sistema deve ser adequado para utilizadores com níveis variáveis de experiência e conhecimento sobre WUM. O sistema deve permitir que os utilizadores mais inexperientes obtenham uma ideia geral de toda a actividade de desenvolvimento de processos de WUM e da sua utilidade, captando todas as acções fundamentais que permitiram derivar o conhecimento resultante, a partir dos dados iniciais, assim como o respectivo curso de tomada de decisão. Neste sentido, torna-se necessário, antes de mais, assegurar outra capacidade imprescindível. Esta corresponde a assistir a aquisição, organização e armazenamento sistemáticos de conhecimento relativo a exemplos úteis de processos de WUM bem sucedidos. Este conhecimento deve ser mantido num repositório centralizado para viabilizar a sua partilha ao longo de toda a organização.

A condução de operações de captura, estruturação, armazenamento e partilha de conhecimento desta natureza, ao nível da organização, levanta vários desafios e três requisitos imediatos. O primeiro prende-se com o recurso a padrões em uso no domínio, com o intuito de obter indicações para estruturar convenientemente processos de DM e vocabulário estabelecido a adoptar, os quais devem ser considerados em conjunto com as orientações de princípios CBR. O padrão PMML, introduzido no capítulo anterior, vem ao encontro deste desígnio, em virtude de ser largamente aceite e suportado por ferramentas de KDD.

O segundo requisito consiste na interoperacionalidade com tecnologias de gestão e exploração de dados ao nível da organização, de maneira a retirar vantagens das capacidades disponíveis e a maximizar o seu potencial. Esta inter-operação pode ser realizada, pelo menos, a dois níveis:

- para aceder a fontes de dados da organização, contendo os dados alvo de análises;
- para gerir o próprio conhecimento obtido a partir de processos de WUM.

A utilização de tais tecnologias, para gerir os dados de casos de WUM, permite usufruir das suas capacidades inerentes, garantindo, simultaneamente um aproveitamento melhorado das experiências registadas, por meio da combinação de funcionalidades típicas de sistemas CBR.

O terceiro requisito refere-se à optimização do processo de povoamento do repositório de casos, dada a diversidade das fontes implicadas e o montante considerável de informação de entrada. Contar unicamente com o utilizador para a introdução de toda a informação é suficiente para levantar objecções ao uso do sistema. Por conseguinte, um pré-requisito é a automatização, tanto quanto possível, da aquisição de informação e conhecimento, assegurando uma extracção efectiva e uniforme (consistente), ao longo de todos os tipos de fontes heterogéneas. As três origens essenciais de informação e conhecimento são as fontes de dados, documentos PMML e os especialistas ou utilizadores experientes. No que concerne a fontes de dados, importa, sobretudo, produzir caracterizações de dados sistemáticas e consistentes, a partir de distintas categorias de fontes, solicitando, apenas, a informação que não pode ser derivada. Os documentos PMML criam a oportunidade de adquirir automaticamente informação acerca das actividades de mineração, apesar de apenas representarem o resultado de processos e de outras limitações de vária ordem (e.g. versões PMML suportadas, omissão de transformações e de alguns parâmetros de configuração de modelos). Assim sendo, torna-se oportuno incorporar facilidades para editar a informação colectada, com vista a encorajar o fornecimento de informação suplementar.

A recomendação de planos de mineração, baseada em exemplos de processos de WUM conduz, ainda, a constrangimentos mais específicos. É necessário confrontar novos problemas com casos prévios de aplicação de WUM, para assim identificar as estratégias mais plausíveis para tratar o

problema corrente. No entanto, procurar um processo de WUM, exactamente, coincidente com o problema corrente, frequentemente, descrito de forma incompleta e imprecisa, corresponde a uma funcionalidade limitada e a um uso dos casos prévios aquém das suas potencialidades. Torna-se, portanto, imprescindível dispor de grande flexibilidade para compactuar com definições difusas de problemas e com novas situações. Tal flexibilidade poderá ser manifestada permitindo especificações parciais, viabilizando padrões de interrogação variados, tais como critérios exactos, inexactos ou baseados em semelhança, e, especialmente, procurando os processos mais próximos ou com um nível de correspondência superior, pelo que, não obrigatoriamente, completamente coincidentes. O paradigma CBR atende, intrinsecamente, a estes anseios. O seu uso propicia uma comparação flexível, baseada na semelhança entre casos, mesmo quando as características de similaridade não são objectivas e claramente definidas. Além disso, as potencialidades de resposta às necessidades reais dos analistas são elevadas, dada a possibilidade de descrever apenas as características relevantes e de atribuir níveis de proeminência específicos às mesmas.

Por último, para lidar com a incerteza em KDD, é prática comum sugerir múltiplas alternativas. Assim sendo, a apresentação de soluções alternativas deve combinar indicações para auxiliar a decisão do utilizador e, também, facilidades de acesso conveniente a informação relacionada, sucessivamente mais detalhada.

3.2.2 Método de Suporte à Construção de Recomendações

A estratégia adoptada neste trabalho, para assegurar a selecção de planos de WUM a recomendar aos utilizadores, fundamenta-se, essencialmente, no paradigma de raciocínio baseado em casos - *Case Based Reasoning* (CBR). Trata-se de uma abordagem de aprendizagem e resolução de problemas [Kolodner 93], capaz de utilizar o conhecimento específico de situações respeitantes a problemas concretos, previamente experimentadas e explicitamente documentadas (casos) [Aamodt e Plaza 94]. O pressuposto subjacente é que problemas similares possuem soluções semelhantes [Kolodner 93]. Um novo problema é solucionado encontrando um caso passado similar e reutilizando-o na situação actual. O CBR também se evidencia e distingue como uma abordagem de aprendizagem incremental sustentada, pois uma nova experiência pode ser retida, cada vez que um problema é resolvido, tornando-se imediatamente disponível para aplicação a problemas futuros [Aamodt e Plaza 94]. Este paradigma tem sido usado, com sucesso, em circunstâncias em que é difícil, senão impossível, especificar princípios gerais, a partir dos quais as soluções podem ser criadas [Kolodner 93].

A exploração de CBR baseia-se em dois aspectos fulcrais. O primeiro reside, naturalmente, na ênfase colocada em torno da experiência prévia de resolução de problemas do domínio, estruturada e guardada numa base de casos, vulgarmente denominada memória de casos, a qual o utilizador interroga, quando se depara com um novo problema. O segundo aspecto prende-se com o próprio processo de aplicação do paradigma e os princípios que estão subjacentes [Aamodt 95]:

- memorizar, como forma primordial de aprendizagem;
- relembrar e reutilizar situações similares, como os principais métodos de raciocínio.

O modelo de aplicação do paradigma mais aceite e divulgado consiste no proposto em [Aamodt e Plaza 94]. Este processo é descrito, ao nível mais elevado de abstracção, por um ciclo envolvendo quatro tarefas, frequentemente, designadas pelos quatro Rs [Aamodt e Plaza 94]:

1. *Retrieve*, que recupera o(s) caso(s) mais similar(es) ao problema actual;
2. *Reuse,* que reutiliza a informação e conhecimento sobre esse(s) caso(s) para resolver o problema actual;
3. *Revise*, que revê a solução proposta;
4. *Retain*, que retém as partes da experiência que poderão ser úteis na resolução futura de novos problemas.

A Figura 9 ilustra o ciclo CBR, segundo [Aamodt e Plaza 94].

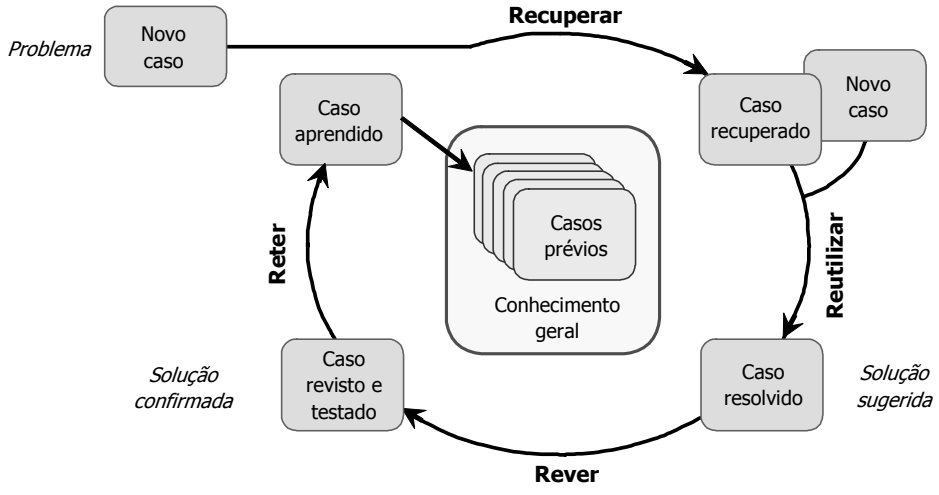


Figura 9 – Ciclo de raciocínio baseado em casos

Cada processo útil de WUM pode ser representado por uma caso, descrito em termos de um problema de domínio e da respectiva solução aplicada. A resolução de problemas de WUM relaciona-se com dois tipos de factores:

- características dos dados e requisitos da análise;
- experiência acerca da extracção de conhecimento em dados de *clickstream*.

As características dos dados e os requisitos da análise definem a descrição do problema, sendo descritores úteis para recuperar casos similares e, por conseguinte, para seleccionar processos de WUM. Os processos mais similares reflectem a experiência de WUM mais preponderante para o problema actual, sendo reutilizados para produzir a descrição da solução.

Vários argumentos sustentam a exploração do paradigma CBR no domínio de aplicação da KDD. Neste domínio, especialmente em WUM, os problemas reais são, normalmente, bastante complexos e possuem muitas particularidades. Os próprios especialistas não conseguem proporcionar regras gerais e consistentes para suportar a sua resolução. A este respeito, os sistemas CBR, usualmente, facultam resultados com qualidade aceitável e podem, ainda, recorrer a casos como um meio apropriado para justificar decisões [Kolodner 91]. A representação de casos concretos possui, só por si, diversas vantagens significativas, mesmo quando as soluções não são facilmente passíveis de reutilização. A documentação explícita de experiências conduz à estruturação e memorização do conhecimento adquirido em situações passadas, viabilizando a sua transferência e partilha ao longo da organização. Os exemplos também são a forma mais útil e convincente de ajuda neste domínio, pois podem:

- simplificar a complexidade subjacente, fornecendo, ao mesmo tempo, detalhes de uma situação resolvida e testada;
- oferecer informação de contexto, tornando possível reportar as soluções em conjunto com as respectivas justificações e descobertas;
- promover o mapeamento do problema corrente para os existentes, quando uma forma mais directa de reutilização não é viável.

Na realidade, a reutilização directa de soluções é possível neste âmbito, pois problemas recorrentes e o uso repetido dos mesmos métodos de DM são comuns. A utilização de CBR favorece ainda a extensibilidade de um sistema que se baseie neste paradigma. Esta característica é proeminente, dada a constante evolução neste domínio e a necessidade de acrescentar conhecimento relativo a novos tipos de problemas, modelos, algoritmos ou ferramentas de KDD e suas características específicas. Além disso, o paradigma disponibiliza um ambiente aberto, para a incorporação de múltiplos tipos de técnicas e tecnologias [Althoff 01].

3.3 Vectores da Abordagem de Assistência a Processos de WUM

A complexidade e os procedimentos de repetição sucessiva de várias etapas do processo de KDD, para diferentes métodos, configurações e transformações de dados, conduzem a um esforço excessivo e, frequentemente, infrutífero. As iniciativas de assistência a estes processos visam, basicamente, incrementar a eficácia e eficiência dos mesmos, reduzindo as dificuldades, o tempo e o esforço requeridos para derivar conhecimento útil e relevante para o problema de análise em questão. Estas iniciativas desenrolam-se em torno de duas vertentes ou propósitos principais – a selecção de modelos e a planificação de processos – e, ainda, com base em duas orientações primordiais – as características dos dados disponíveis e os requisitos da análise. Entende-se que estes dois tipos de factores são ambos imprescindíveis para orientar a assistência a processos de WUM e, também, que a vertente de planificação de processos é a mais ajustada para combater os desafios que motivam o presente trabalho. Esta vertente representa um âmbito de apoio mais alargado, tanto em termos dos diversos níveis de abstracção das actividades a levar a cabo (e.g. conceptual, desenho e implementação), como nas distintas etapas a que pode prestar auxílio e, também, nos tipos de processos que podem ser contemplados (e.g. aplicação individual ou combinada de métodos de DM e outras operações). Portanto, aproxima-se mais dos contratempos e cenários reais da WUM. Considera-se ainda que a assistência a estes processos deve estar voltada para vários aspectos de abstracção da complexidade, dando especial ênfase à facilidade na descrição de problemas de WUM, à simplificação das tarefas da etapa de modelação, aos detalhes conducentes à implementação bem sucedida dos planos sugeridos e à compreensão da sua utilidade e da aplicação prática dos seus resultados.

Grande parte do sucesso conseguido pelos especialistas na resolução de problemas de WUM deve-se, geralmente, à sua experiência e ao conhecimento adquirido no tratamento de situações práticas concretas. Neste domínio, os problemas reais são, tipicamente, muito complicados e dotados de diversas facetas específicas. Os próprios especialistas não conseguem estabelecer princípios gerais e consistentes, capazes de sustentarem a resolução de problemas. Neste sentido, a estratégia adoptada, no suporte à assistência de processos de WUM, consiste em gerir o conhecimento obtido, a partir da experiência na resolução de problemas de WUM, recorrendo a uma abordagem baseada no paradigma CBR e com um âmbito organizacional. Os modelos de representação fundamentados em casos poderão actuar como bases para a estruturação,

exploração e partilha de repositórios de conhecimento e, por conseguinte, fomentar a adopção de práticas efectivas de WUM. Adicionalmente, a aplicação dos mecanismos CBR promove a extensibilidade incremental e sustentada do repositório de conhecimento e a reutilização do mesmo na resolução de novos problemas, dada a flexibilidade da comparação baseada em similaridade e as respectivas potencialidades de ampliação da aplicação dos casos a novas situações. Deste modo, torna-se possível simular, de forma mais sistemática, o comportamento dos especialistas na resolução de problemas reais de WUM. O paradigma CBR faculta ainda outras vantagens acrescidas, entre as quais se salienta o ambiente aberto para a incorporação de várias técnicas e tecnologias. Esta vantagem pode ser especificamente aproveitada para assistir a aquisição de conhecimento de forma semi-automática. Quando aplicado ao nível da organização e integrado com as tecnologias de informação ao seu serviço, o paradigma CBR pode ser usado para consolidar os processos prévios de WUM numa memória comum, enquadrar a captura de conhecimento, consoante os requisitos e particularidades do organismo, e promover a disseminação e reutilização deste conhecimento ao longo de toda a organização.

Capítulo 4

Sistema Selector de Planos de Mineração

4.1 Arquitectura Funcional do Sistema

O sistema *Selector de Planos de Mineração* (SPM) foi especialmente orientado para assistir os analistas, no desenvolvimento e aplicação de processos de WUM, ajudando-os a estabelecer as estratégias mais adequadas (planos de mineração) para minerar dados de *clickstream*, suportado por casos criados a partir de outras experiências de WUM (casos de aplicação), concretizadas com sucesso no passado. Os planos de mineração são sugeridos, atendendo às características dos dados alvo disponíveis e aos requisitos da análise (definição do problema) e com base nos casos mantidos num repositório de conhecimento. O SPM fundamenta-se no paradigma CBR, funcionando como um sistema de gestão e reutilização de casos de aplicação de WUM da organização.

A Figura 10 ilustra a arquitectura funcional básica do sistema SPM. Esta é constituída pela base de conhecimento e por quatro componentes funcionais primordiais, os quais são responsáveis pela realização das actividades mais significativas do sistema, nomeadamente, as que se enumera seguidamente:

- **Caracterizador de dados** – analisa os dados alvo do estudo e produz os respectivos metadados mais relevantes, para o propósito de selecção de métodos e abordagens de mineração de dados, a partir de indicações do utilizador e da fonte de dados envolvida.

- **Construtor de problemas** – auxilia o utilizador a especificar os requisitos da análise de dados em causa e trata esses requisitos, para os sistematizar e gerar um novo problema de WUM.
- **Conciliador de descrições** – fornece descrições de processos de WUM úteis e concluídos com sucesso, a partir de informação concedida por ferramentas de DM ou WUM e, também, de especificações complementares, obtidas por meio de interacção com o utilizador.
- **Motor CBR** – efectua a inferência, implementando os módulos recuperar, reutilizar e reter e recorrendo a um *Sistema de Gestão de Base de Dados* (SGBD) convencional para gerir o conhecimento adquirido. Estes três módulos são, em certa medida, correspondentes às três tarefas recuperar, reutilizar e reter do ciclo CBR básico, abordado anteriormente na secção 3.2.2:
    - o módulo **recuperar** usa o novo problema construído, para o comparar com os casos existentes e devolver aqueles que são mais similares;
    - o módulo **reutilizar** elabora recomendações de planos de WUM, com base nos casos recuperados;
    - o módulo **reter** procede à organização e ao armazenamento dos casos resultantes de processos de WUM bem sucedidos.

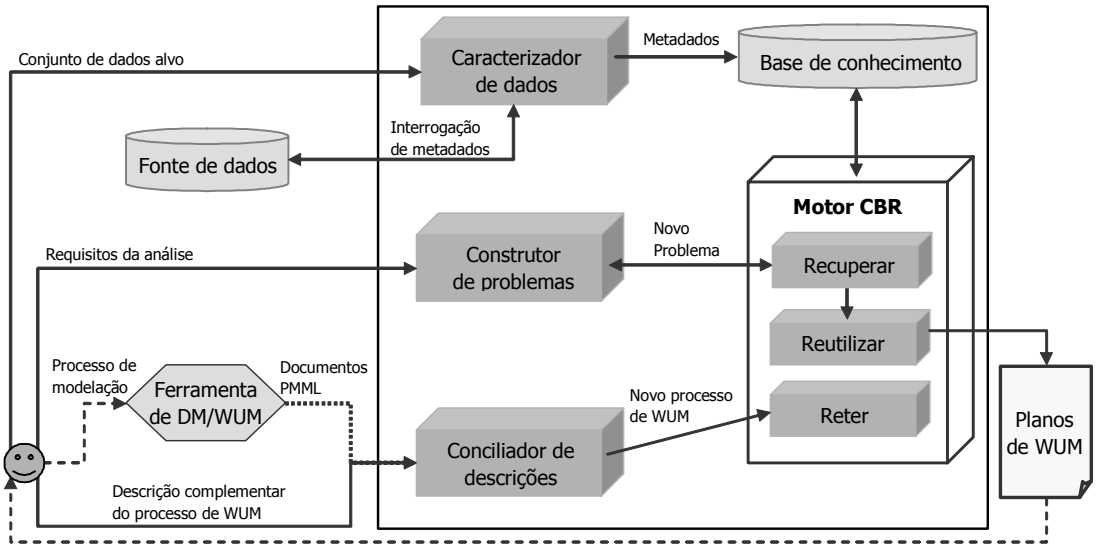


Figura 10 – Arquitectura funcional do sistema Selector de Planos de Mineração

Vários modelos têm sido propostos para orientar a operacionalização do paradigma CBR. Os modelos mais influentes, aceites e largamente conhecidos incluem:

- o modelo dos contentores de conhecimento, introduzido por [Richter 95];
- o modelo de processamento do ciclo CBR, anteriormente apresentado, e a estrutura de decomposição de CBR em tarefas e métodos que as realizam, ambos provenientes de [Aamodt e Plaza 94].

O primeiro modelo foca os diferentes tipos de conhecimento encontrados em sistemas CBR, sugerindo uma abordagem para a organização da representação de conhecimento nestes sistemas, ao longo de quatro contentores:

- vocabulário do domínio;
- medidas de similaridade;
- transformação de soluções;
- base de casos.

Os dois restantes modelos facultam meios para estruturar o próprio sistema, sendo complementares, uma vez que traduzem duas perspectivas de um sistema CBR:

- Uma visão geral, na qual se identificam os sub-processos essenciais e os seus relacionamentos e resultados.
- Uma visão orientada por tarefas, expressando a decomposição hierárquica de tarefas gerais em sub-tarefas e nos métodos relacionados para as levar a cabo. A tarefa de topo desta hierarquia é designada "resolução de problemas e aprendizagem a partir da experiência".

O processo CBR do sistema SPM foi modelado com base em sete tarefas centrais, adaptando o ciclo típico aos requisitos funcionais específicos da corrente aplicação do paradigma. A Figura 11 mostra essas tarefas, as suas principais interligações, correspondentes entradas e saídas e, ainda os contentores de conhecimento determinantes para a sua persecução. No ciclo CBR original as actividades com finalidades afins às das tarefas acrescentadas, designadamente, as duas tarefas caracterizar e construir e a tarefa conciliar, são integradas, respectivamente, como sub-tarefas dos passos recuperar e reter. Contudo, no sistema SPM, essas actividades assumem papéis proeminentes, desempenhando funções, em certa medida, autónomas. No que concerne à base de conhecimento, esta contempla, sobretudo, contentores relativos ao vocabulário, similaridade e base de casos e, ainda um repositório de metadados provisórios, conforme se explica mais à frente.

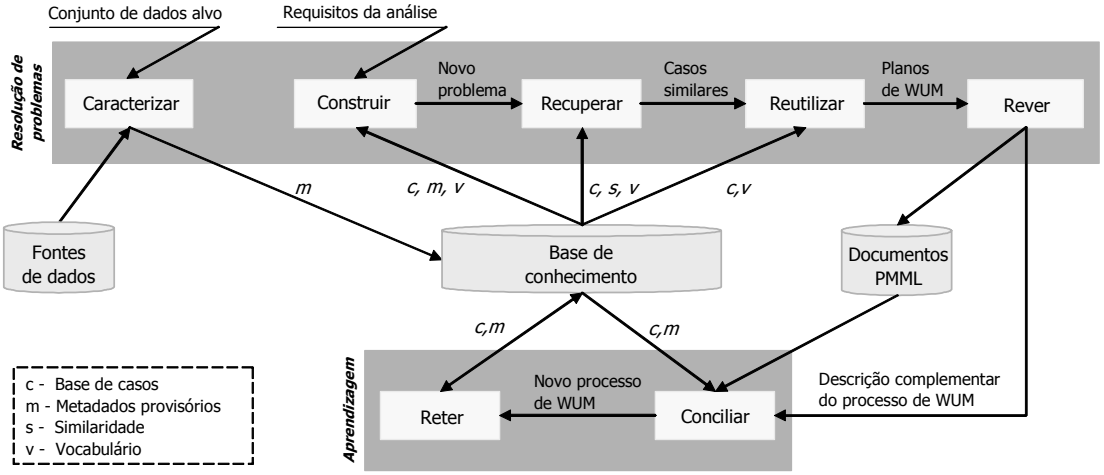


Figura 11 – Tarefas centrais do ciclo de raciocínio baseado em casos adoptado

As seis tarefas com suporte activo no sistema são subdivididas nas sub-tarefas que o mesmo tem de assegurar na Figura 12. A assistência à resolução de problemas de WUM e a aprendizagem a partir de novas experiências neste âmbito são, portanto, conduzidas pelo sistema ao longo de um ciclo de sete passos distintos, nomeadamente:

1) **Caracterizar** dados – Um conjunto de dados alvo é submetido ao sistema, especificando a sua fonte e alguma informação adicional. O sistema **recolhe** as indicações concedidas pelo analista e **extrai** metadados a partir da fonte, para derivar uma caracterização dos dados. Esta caracterização é **preservada** no repositório de metadados provisórios.
2) **Construir** problema – Uma análise é descrita informalmente, indicando, tipicamente, um conjunto de dados alvo (provisório ou prévio e existente na base de casos) e interagindo como o sistema para facultar requisitos explícitos. Estes requisitos baseiam-se nos metadados dos dados alvo e noutros tipos de constrangimentos, em função de descritores e valores previamente **gerados** pelo sistema. Seguidamente, o sistema analisa estes requisitos, para traduzir e sistematizar as restrições subjacentes, **definindo** um novo problema alvo.
3) **Recuperar** – O novo problema, reflectindo restrições implícitas e explícitas, retracta o perfil de **procura** de casos aplicáveis, sendo **comparado** com os casos encontrados, no sentido de **seleccionar**, entre esses, os mais similares ao alvo.

4) **Reutilizar** – Os casos mais similares recuperados são **avaliados** e **organizados**, de acordo com as diferentes soluções técnicas que representam, nível da sua similitude em relação ao alvo e os critérios de avaliação de exercícios de KDD, os quais foram assinalados como relevantes pelo analista, permitindo **produzir** uma solução orientada para a reutilização, contendo uma lista ordenada de planos de WUM e referências para casos exemplificativos.
5) **Rever** – Os planos de WUM seleccionados e instanciados em casos concretos assistem o analista, durante o desenvolvimento do processo actual, recorrendo a uma ferramenta de DM ou WUM. O analista adapta os detalhes dos planos propostos à situação corrente, para obter um processo final revisto. Se a ferramenta de DM ou WUM suportar o padrão PMML pode fornecer uma representação dos modelos resultantes do processo revisto, exportando esses modelos para documentos neste formato.
6) **Conciliar** descrições – O processo revisto pode ser submetido ao sistema, o qual **combina** os documentos PMML (se disponíveis) e a informação complementar numa descrição unificada. Seguidamente, o conteúdo dos documentos PMML é analisado e **transformado** para uma representação interna, mais apropriada para a manipulação do mesmo pelo sistema.
7) **Reter** – Os elementos descritores, prestados pelo passo conciliar, são **integrados**, conjugando também os eventuais metadados provisórios, formando assim um novo caso que, depois de **estruturado**, é **registado** na base de conhecimento, de acordo com o seu esquema interno, finalizando o ciclo.

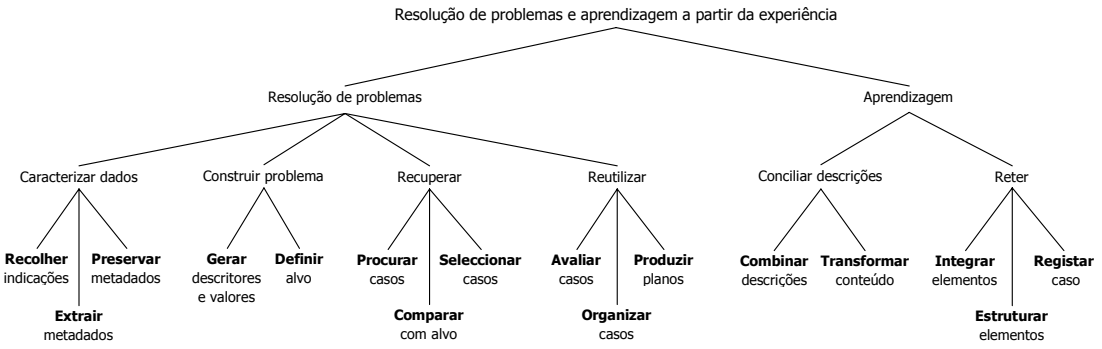


Figura 12 – Decomposição das tarefas centrais do sistema em sub-tarefas

4.2 Base de Conhecimento

Nos sistemas CBR a principal fonte de conhecimento são exemplos de situações experimentadas, convenientemente organizados e armazenados na memória ou base de casos. As descrições concretas de episódios, sobre a resolução de problemas passados no domínio, apesar de tradicionalmente vistas como dados ou informação, afiguram-se como conhecimento para os mecanismos CBR, pelo facto de estes serem capazes de usar casos para raciocinar [Aamodt 95]. Outros tipos e formas mais gerais de conhecimento poderão ser integrados no processo CBR, com diferentes níveis de riqueza, graus de representação explícita e papéis [Aamodt e Plaza 94]. Uma contribuição importante para a questão da estruturação da representação de conhecimento nos sistemas CBR trata-se do modelo dos contentores de conhecimento [Richter 95], já referido. Este modelo preconiza a distribuição da representação de conhecimento, ao longo de quatro contentores, de acordo com os requisitos da aplicação:

- vocabulário, que descreve o domínio de aplicação do paradigma;
- medidas de similaridade, a aplicar para aferir o nível de similitude entre casos, designadamente, entre os problemas prévios e alvo;
- conhecimento de transformação ou adaptação, para ajustar as soluções encontradas ao problema actual;
- base de casos, contendo os exemplos práticos ou casos do domínio, descritos usando o respectivo vocabulário.

Os três primeiros contentores representam conhecimento compilado (mais estável), enquanto os casos contêm conhecimento interpretado (dinâmico). Esta visão dos contentores de conhecimento recebeu grande aceitação, passando a ser uma abordagem natural para basear a organização da representação de conhecimento em sistemas CBR.

A base de conhecimento do sistema SPM pode ser descrita considerando dois tipos de componentes essenciais: a base de casos e o conhecimento de domínio. A base de casos contém os exemplos úteis de desenvolvimento e aplicação de WUM, correspondendo cada exemplo a um caso, estruturado em termos de um problema de domínio e da respectiva solução aplicada. O conhecimento de domínio proporciona itens de conhecimento enquadrados em contentores de vocabulário e de similaridade, assim como noutros tipos de contentores específicos desta área de aplicação, tais como os metadados provisórios. Como o sistema não prevê a transformação de soluções, não possui um contentor de adaptação (ou de transformação). Além disso, o contentor de similaridade é, em certa medida, diferente, pois não se reporta apenas a medidas de similitude, mas também a outros aspectos, conforme se explica mais adiante.

Para além do modelo dos contentores de conhecimento e de princípios CBR orientadores, úteis para fundamentar a estruturação da base de conhecimento e do repositório de casos, o recurso a padrões em uso na área, conforme já se indicou, é imprescindível. Os casos de domínio são complexos e o respectivo vocabulário é diverso, mesmo ao nível dos aspectos mais básicos, dificultando a modelação do conhecimento deste domínio de aplicação. A própria designação mineração de dados possui diferentes conotações para pessoas distintas. Deste modo, procurou-se adoptar vocabulário em conformidade com o padrão PMML, bem como retirar indicações desta especificação para conceber a base de conhecimento, descrita nas duas secções que se seguem.

4.2.1 Base de Casos

A base de casos é o componente fulcral da base de conhecimento do sistema SPM, consistindo num repositório povoado com metadados relativos a processos úteis de desenvolvimento e aplicação de WUM. Estes metadados e, consequentemente, a representação de processos de WUM devem versar em torno das seguintes dimensões primordiais (**D**, **T**, **A**, **K**, **P**):

- Caracterizações de dados alvo de estudos (**D**), ao nível geral (conjunto de dados) e das variáveis individuais, cobrindo as propriedades significativas para a selecção de métodos e abordagens de DM.
- Categorizações de tipos de problemas de WUM (**T**), em termos de abstracções específicas, consoante as necessidades particulares da organização.
- Sequências de actividades importantes e pormenores da sua realização (**A**), incluindo passos de transformação e modelação, os dados envolvidos, a configuração de parâmetros e as respectivas explicações e justificações.
- Conhecimento prévio e derivado (**K**), respeitante a factos precedentes pertinentes, às descobertas alcançadas pelos processos e a relacionamentos das mesmas com tais factos.
- Descrição geral de processos de WUM ($\mathbf{P}_G$) e critérios de sucesso ou de avaliação genéricos ($\mathbf{P}_V$), os quais viabilizam a classificação de processos e dos resultados obtidos.

Cada caso corresponde a um processo de WUM, descrito em termos de um problema do domínio e da respectiva solução aplicada, e em função do vocabulário adoptado. Um problema de domínio é definido, sobretudo, por:

- caracterizações do conjunto de dados utilizado (**D**);
- categorizações do tipo de problema contemplado (**T**);
- critérios de avaliação de processos e resultados de mineração ($\mathbf{P}_V$).

Os novos problemas a resolver são formulados, igualmente, com base nas dimensões anteriores, mas sob a forma de preferências ou expectativas em torno das mesmas. Já a solução aplicada compreende descrições dos seguintes aspectos:

- actividades levadas a cabo ao longo das várias etapas do processo (**A**);
- conhecimento prévio e derivado (**K**);
- informação geral do processo ($\mathbf{P}_G$).

Tais descrições, reportadas aos casos recuperados, são usadas pelo sistema para produzir os planos de WUM a recomendar. A Figura 13 apresenta o modelo conceptual abreviado de representação de casos, de acordo com as dimensões anteriores (**D**, **T**, **A**, **K** e $\{P_G, P_V\} \in \mathbf{P}$), utilizando um diagrama de classes em notação *Unified Modeling Language* (UML) [WWW6] simplificada. Algumas das classes não ilustradas, para evitar a complexidade da figura, assumem papéis com menor relevância, prendendo-se, por exemplo, com os autores de processos e fontes utilizadas pelos mesmos.

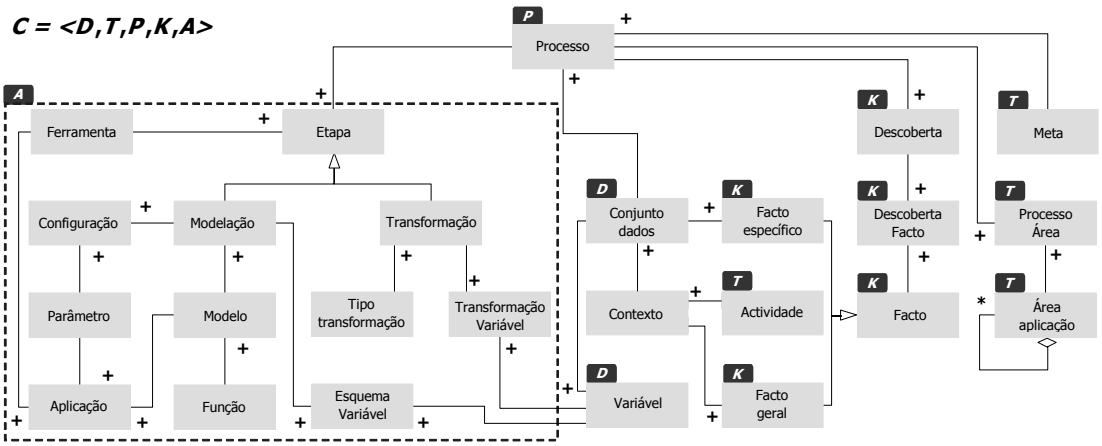


Figura 13 – Modelo conceptual abreviado de representação de casos

O núcleo do modelo conceptual reside na classe *Processo* (**P**). Esta classe contém, basicamente, uma descrição geral de cada exercício de WUM ($\mathbf{P}_G$) e informação acerca da sua avaliação ($\mathbf{P}_V$) e do estado do caso. Além disso, interliga as classes respeitantes à descrição do problema e da solução, contendo, também, atributos destas duas vertentes de um caso.

A categorização do tipo de problema de WUM (**T**) deve reflectir a categoria de situação tratada pelo estudo, podendo, ainda, desempenhar um papel proeminente na identificação de casos úteis para auxiliar a resolução de novos problemas. Por conseguinte, entendeu-se que esta categorização se deveria basear em abstracções associadas aos problemas reais solucionados (ou a solucionar), ao invés de depender de classificações técnicas das funcionalidades dos métodos de mineração. A solução encontrada, para este efeito, foi recorrer ao objectivo da análise, já que este pode ser alvo de abstracção de complexidade e é o factor que melhor indicia o problema que está subjacente. O objectivo de uma análise, neste âmbito, possui duas perspectivas distintas em aplicações reais: a perspectiva de negócio e a de WUM. O ponto de vista de negócio prende-se com as intenções da análise ou com os usos possíveis das descobertas, sendo designado área de aplicação. A perspectiva da WUM retracta o tipo de resultado de mineração obtido (ou que se espera obter) e a categoria de abordagem de DM explorada (ou a explorar). Este ponto de vista foi denominado meta (de análise), uma vez que traduz um tipo de problema (ou meta) de WUM.

Tendo em conta as duas perspectivas mencionadas, por exemplo, uma análise que visa discriminar as páginas infrutíferas de um sítio Web dedicado a comércio electrónico, poderá ser categorizada com base na área de aplicação de avaliação de iniciativas e na meta de distinguir visitas com e sem compras, em função das páginas visitadas. Assim sendo, cada caso deve possuir uma meta específica associada e deve poder ser relacionado com uma ou mais áreas de aplicação. Isto porque, tipicamente, antevêem-se diversas aplicações práticas das descobertas e, embora seja muito plausível ter múltiplas metas em mente, quando se descreve um problema de análise, já não é conveniente classificar o tipo de resultado obtido, simultaneamente, de formas distintas. As áreas de aplicação foram organizadas numa hierarquia, com vista a permitir vários níveis de subdivisão em sub-áreas de aplicação mais específicas e a sua definição consoante as necessidades da organização. Para as metas de análise poderão ser adoptadas as abstracções consideradas mais apropriadas, igualmente de acordo com os requisitos do organismo.

Com base no exposto, as classes *Meta*, *Área_aplicação* e *Processo_Área*, referem-se, respectivamente, a metas de análise, à hierarquia de áreas de aplicação accionáveis das descobertas e à associação entre as classes *Processo* e *Área_aplicação*. Enquadrou-se ainda a classe *Actividade* na dimensão **T**, uma vez que esta pode traduzir vertentes de operação mais específicas, como a área ou a unidade de actividade na qual o sítio Web se insere, sendo, portanto, um aspecto pertinente para qualificar e procurar problemas.

As classes *Conjunto_dados* e *Variável* representam caracterizações dos dados iniciais a minerar e, também, variáveis derivadas durante o estudo. Cada conjunto de dados pertence a um *Contexto*,

uma abstracção (e.g. um sítio ou unidade de negócio) para agrupar conjuntos de dados relacionados, com propriedades e factos gerais (*Facto_Geral)* comuns. A classe *Contexto* é ambígua no que concerne a dimensões (**T** e **K**). Na realidade, o seu papel é evitar redundância na descrição de aspectos partilhados por uma série de conjuntos de dados.

O conhecimento prévio é disponibilizado através das classes *Facto*, *Facto_Geral* e *Facto_específico*, as quais reflectem o ambiente da análise e explicam os factores que poderão afectar o processo de KDD. Estas classes reportam factos relevantes (e.g. eventos, circunstâncias e condicionantes), relativamente a contextos gerais *(Facto_Geral)* ou a conjuntos de dados específicos *(Facto_específico)*. O conhecimento extraído e os seus relacionamentos com factos existentes, ou seja os efeitos desses factos na análise, são incluídos, respectivamente, nas classes *Descoberta* e *Descoberta_Facto*. Apesar de o conhecimento derivado estar representado nos documentos PMML (se disponíveis), nomeadamente, sob a forma de modelos resultantes do processo, pretende-se capturar uma descrição dos mesmos após a sua interpretação e usando a linguagem do analista.

Várias classes representam as actividades concretizadas (**A**). A (super) classe *Etapa* denota cada fase de modelação ou transformação, designadamente, os atributos partilhados pelas suas subclasses, etapas estas que são desenvolvidas explorando uma *Ferramenta* disponível. A subclasse *Transformação* permite retractar operações simples de pré-processamento, tendo um determinado tipo *(Tipo_transformação)* e envolvendo variáveis de entrada e saída *(Transformação_Variável)*. Cada instância de aplicação de um modelo é coberta pela subclasse *Modelação*, abrangendo ainda as configurações de parâmetros *(Configuração)* e as variáveis implicadas e o seu papel na análise *(Esquema_Variável)*. A classe *Aplicação* relaciona a utilização de diferentes parâmetros com os modelos e ferramentas existentes, a fim de captar circunstâncias da sua aplicação. Finalmente, as classes *Modelo* e *Função* contêm as definições de modelos e funções implementados por ferramentas.

4.2.2 Conhecimento de Domínio

O conhecimento de domínio refere-se a aspectos do corrente âmbito de aplicação do paradigma CBR, especialmente, em termos das propriedades requeridas para interpretar, comparar e recuperar casos. Este componente versa, essencialmente, em torno dos atributos descritores de problemas de WUM e abarca quatro tipos de aspectos: metadados provisórios, vocabulário, similaridade e configuração do sistema.

Os metadados provisórios são mantidos num repositório próprio, para que possam ser explorados na construção de problemas alvo e no eventual registo dos mesmos na base de casos. Estes metadados permanecerão neste repositório até que ocorra uma das seguintes situações: remoção do respectivo conjunto de dados provisório; retenção de um caso baseado nesse conjunto de dados.

Quanto à vertente de vocabulário, os seus tipos de itens mais preponderantes são os seguintes:

- Tabelas que definem os valores possíveis de alguns atributos, como, por exemplo, os tipos de dados e as categorias semânticas que as variáveis de conjuntos de dados podem assumir.
- Domínio dos atributos descritores de problemas, para efeitos de interpretação, validação e manipulação dos mesmos, contemplando as seguintes categorias fundamentais:
    - um conjunto predefinido de valores possíveis;
    - uma tabela contendo os valores válidos;
    - um tipo de dados básico (e.g. inteiro) ou mais específico (e.g. inteiro $\in$ [1..5]).
- Especificação de todos os atributos que constituem os descritores de problemas e as suas propriedades.
- Enumeração das diferentes categorias e formas de uso dos atributos descritores de problemas, como, por exemplo, identificação dos descritores que constituem os critérios de avaliação de soluções e que serão alvo de cálculos de indicadores para tipos de soluções.

É importante referir que o componente de conhecimento de domínio é o único ponto do sistema com informação acerca dos campos dos casos que fazem parte da descrição de problemas e os papéis que estes desempenham no âmbito dos mecanismos de raciocínio. Esta característica conduz a múltiplas vantagens, discutidas ao longo da dissertação. Entre estas, evidenciam-se a independência e extensibilidade de mecanismos como os de recuperação e reutilização, as quais viabilizaram um melhoramento gradual destas funcionalidades do sistema, por meio da simples incorporação de mais atributos. Esta característica também contribui para a necessidade de uma tarefa capaz de basear a construção de problemas nestas especificações.

Já a vertente de conhecimento de similaridade encontra-se distribuída por dois tipos de componentes. O primeiro consiste nas medidas ou funções de similaridade, implementadas via programação. O segundo corresponde a um contentor de similaridade pertencente à base de conhecimento. Este contentor refere-se apenas a atributos descritores de problemas e contempla elementos como os seguintes:

- Matrizes de similaridade para atributos categóricos simbólicos, as quais estabelecem o nível de similitude entre todos os pares de valores viáveis, para campos em que não é possível estabelecer uma função de semelhança, como, por exemplo, o tipo de dados de variáveis, a actividade, as metas de análise e as áreas de aplicação.
- Propriedades necessárias para o cálculo de similaridade, englobando a especificação de aspectos como os seguintes:
    - ponderações e funções de similitude dos descritores, a aplicar por omissão;
    - valor mínimo e máximo para atributos numéricos em que é possível predefinir, pelo menos, algum destes valores.

A associação de funções de similaridade a atributos descritores e, também, o seu registo no contentor de similaridade permitem optimizar o código da implementação da funcionalidade de recuperação, e, ainda, preparar a escolha entre várias funções possíveis, durante a exploração do sistema, por meio de configurações a implementar futuramente.

Por último, existem outras propriedades que definem diferentes ângulos do comportamento do sistema, as quais foram enquadradas numa vertente geral de configuração do mesmo. Alguns exemplos são a especificação de valores por defeito do número k de planos a apresentar e do limiar de similaridade dos casos incluídos no resultado, bem como a enumeração de séries de itens, úteis em diversos pontos da aplicação, para assegurar determinadas funcionalidades, tais como, o preenchimento de objectos de interacção. A ideia subjacente é centralizar estes elementos, para evitar redundância e dependência do código de implementação e, consequentemente, os efeitos nefastos de alteração e optimização do sistema nestas circunstâncias.

4.3 Módulo de Caracterização de Dados

As propriedades dos dados em estudo são determinantes na escolha de métodos de mineração e, também, de operações de transformação de dados. A caracterização de dados tem por missão traduzir os dados iniciais de entrada para uma meta-representação sistemática, capaz de:

- capturar as propriedades influentes na selecção de métodos e abordagens de mineração a adoptar;
- substituir os dados originais, na comparação entre conjuntos de dados diferentes;
- assegurar a independência e consistência dos metadados, ao longo de vários tipos de fontes heterogéneas.

Conforme já se referiu no capítulo anterior, no âmbito do projecto STATLOG foi proposta uma abordagem de caracterização de dados que, posteriormente, foi estendida em [Engels e Theusinger 98] e tem sido frequentemente explorada com sucesso em meta-aprendizagem, para recomendar algoritmos apropriados. Esta caracterização baseia-se em medidas gerais, estatísticas (dados numéricos) e informação teórica (dados categóricos), respeitantes ao conjunto de dados e às variáveis individuais. Todavia, estas medidas são numerosas e complexas e têm sido usadas para um subconjunto de funções de DM (e.g. classificação e regressão). Além disso, os dados de *clickstream* possuem características peculiares, as quais condicionam substancialmente o processo de extracção de conhecimento e não são passíveis de extracção automática.

Para levar a cabo a caracterização de dados, identificou-se um conjunto simples de atributos descritores, cobrindo metadados obtidos automaticamente pelo sistema e outros cujo valor é indicado pelo utilizador. Estes atributos podem ser recolhidos ao nível do conjunto de dados ou das variáveis, sendo estruturados em duas categorias, tal como se ilustra na Tabela 3:

- características genéricas de DM;
- propriedades específicas de WUM.

Tabela 3 – Principais metadados de caracterização de conjuntos de dados

Nível	**Conjunto de dados**	**Variável individual**
Características genéricas de DM	– Número de linhas ou registos – Número de colunas ou variáveis – Percentagem de variáveis numéricas – Percentagem de variáveis categóricas – Percentagem de variáveis temporais – Percentagem de variáveis binárias	– Tipo de dados – Número de valores distintos – Número de valores nulos
Propriedades específicas de WUM	– Tipo de identificação de visitantes – Tipo de registo de informação de visitantes – Disponibilidade de ordem de acesso a páginas – Disponibilidade de repetição de acesso a páginas – Disponibilidade de duração de acessos – Disponibilidade de data de acessos – Disponibilidade de hora de acessos – Granularidade dos dados	– Categoria semântica

A primeira categoria de metadados reflecte, essencialmente, propriedades básicas com influência na aplicação e desempenho dos métodos de mineração, como a dimensionalidade e volume dos dados, nível de incidência de classes gerais de tipos de dados e variabilidade e qualidade dos valores das variáveis. Já a segunda categoria retracta, maioritariamente, factores que afectam a adequação de abordagens e propósitos de análise dos dados. Por exemplo, o tipo de identificação de visitantes (e.g. identificado via registo explícito ou apenas diferenciado) e a abrangência e natureza do registo de informação acerca de visitantes, condicionam a apropriação de estudos voltados para os visitantes individuais. A granularidade dos dados (e.g. nível da sessão ou de acessos a páginas) é um factor determinante para as formas de análise que poderão ser levadas a cabo. A disponibilidade de determinados dados (e.g. ordem de acesso) também afecta a viabilidade da aplicação de certas análises (e.g. métodos sequenciais). Já a categoria semântica das variáveis permite estabelecer afinidades entre variáveis de diferentes conjuntos de dados que, de outra forma, não seria possível detectar.

A persecução da caracterização de dados pelo sistema, cujos aspectos nucleares se recapitula na Figura 14, contempla três sub-tarefas. A primeira decorre por meio de interacção com o utilizador para recolher três tipos de indicações:

- contexto em que se enquadram os dados disponíveis;
- origem e especificação do conjunto de dados, envolvendo aspectos dependentes do tipo de fonte (e.g. localização no sistema de ficheiros ou base de dados);
- propriedades típicas dos dados de *clickstream* e, por conseguinte, específicas de WUM.

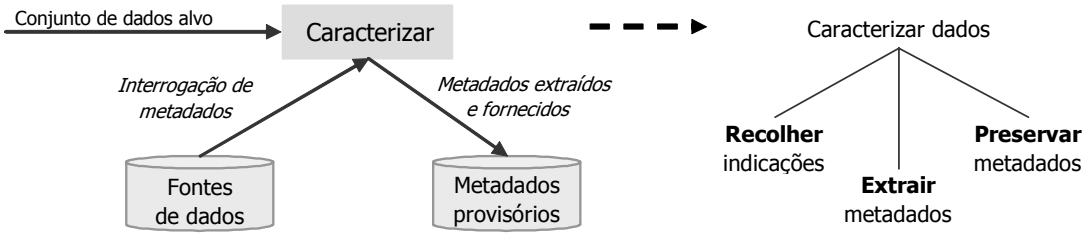


Figura 14 – Entradas, resultados e sub-tarefas do módulo de caracterização de dados

As indicações relativas à fonte são, então, usadas pela segunda sub-tarefa para aceder à mesma e extrair metadados, designadamente as características genéricas de DM. Na terceira sub-tarefa, todos os metadados derivados ou fornecidos são preservados no contentor ou repositório de

metadados provisórios, para viabilizar o seu uso posterior. A recolha de informação pode ocorrer mais do que uma vez, através de operações de edição. Esta situação sucede, por exemplo, para as variáveis, as quais são primeiro extraídas e, só então, algumas das suas propriedades podem ser enriquecidas (e.g. categoria semântica).

A caracterização de dados produz uma representação do conjunto de dados inicial e reflecte requisitos inerentes. Porém, não considera requisitos explícitos da análise, quer no que concerne às próprias propriedades dos dados, como a outras categorias de restrições, as quais poderão melhorar a descrição do problema alvo.

4.4 Módulo de Construção de Problemas

A descrição de requisitos complementa a especificação do problema corrente, com base na natureza, contexto e objectivos da análise e preferências do utilizador. Esta descrição poderá basear-se num caso existente ou num conjunto de dados previamente caracterizado, podendo ainda ser livre e inteiramente definida pelo utilizador, em termos das propriedades que deseja encontrar nos casos a recuperar. De qualquer modo, as categorias de restrições e as capacidades suportadas são similares, residindo a principal diferença no ponto de partida, ou seja, nos itens que instanciam essas categorias. Os três tipos primordiais de requisitos englobam preferências em relação a características de dados, à tarefa de WUM e aos critérios de avaliação de resultados e processos. Podem também ser indicados constrangimentos sobre a data do processo e a actividade do contexto em que os dados se enquadram.

Se o ponto de partida da formulação do problema se tratar de um caso ou conjunto de dados, o utilizador poderá distinguir as variáveis (do conjunto de dados alvo) que entende serem mais relevantes. As propriedades destas variáveis eleitas serão objecto de uma comparação específica, com as variáveis de casos prévios, efectivamente usadas em etapas de modelação, incluindo variáveis originais ou derivadas. Por omissão, esta comparação abrange todas as variáveis iniciais do conjunto de dados alvo. Em situação contrária (especificação "livre"), o utilizador poderá definir as propriedades desejáveis para um conjunto (virtual) de variáveis alvo pretendidas. Todas as restantes descrições são independentes do ponto de partida. É também possível alterar todas as propriedades ao nível do conjunto de dados ou do caso de referência, para efeitos de especificação do problema alvo.

No que respeita à descrição da tarefa de DM pretendida, entendeu-se que esta deveria ser simplificada e baseada em abstracções relacionadas com os problemas práticos a resolver. Uma descrição baseada em funções de DM não permite abstrair a complexidade subjacente e poderia conduzir à exclusão de processos de WUM, envolvendo funções alternativas. Em acréscimo, uma tarefa pode corresponder a mais do que uma função de DM. Na realidade, a recomendação de funções, tal como de modelos de DM, faz parte da missão do sistema SPM. Esta questão foi precisamente a que determinou a introdução de metas de análise e de áreas de aplicação, na representação de casos, de acordo com a explicação da secção 4.2.1 da categorização do tipo de problema de WUM, pertencente à dimensão **T**. Assim sendo, a tarefa de DM desejada é comunicada, de forma simplificada, por meio da selecção de metas de análise e de áreas de aplicação, consideradas significativas pelo analista para o problema em causa.

Quanto a critérios de avaliação de processos e resultados DM ou WUM, procurou-se recorrer aos mais promissores e usuais neste âmbito, tendo-se elegido os seguintes:

- precisão dos resultados;
- tempo de resposta;
- exigência de recursos;
- simplicidade de implementação;
- interpretabilidade.

De modo a permitir a comparação entre casos, em função dos critérios de avaliação enumerados, assim como de lidar com o problema da sua subjectividade, adoptou-se uma escala ordinal, contemplando um número limitado de valores. A utilização da mesma escala para todos os critérios facilita a descrição de requisitos e permite estabelecer prioridades entre estes. O utilizador selecciona os critérios que entende serem os mais relevantes para o estudo, indica os respectivos valores aceitáveis, preenchendo os limites mínimos, e define a preponderância relativa entre os critérios seleccionados.

O sistema disponibiliza ainda formas típicas de focagem da descrição do problema, designadamente:

- a aplicação de critérios de filtragem exacta, para excluir casos irrelevantes;
- a atribuição de níveis de importância específicos a determinados descritores, a fim de aumentar ou reduzir a sua proeminência.

A construção de um problema alvo consiste, basicamente, em recolher uma série de descritores proeminentes do problema corrente e os respectivos valores esperados ou pretendidos, os quais serão, posteriormente, comparados com os casos existentes. Esta tarefa é concretizada, de acordo

com a Figura 15, cobrindo duas sub-tarefas essenciais. A primeira sub-tarefa reside em gerar e apresentar duas categorias de elementos:

- descritores de problemas e suas propriedades, nomeadamente os seus domínios e os tipos de restrições de filtragem exacta aplicáveis, recorrendo, para tal, ao contentor de vocabulário;
- os valores por defeito dos descritores, provenientes de um conjunto de dados ou problema já registado na base de casos ou, ainda, de um conjunto de dados provisório.

Já a segunda sub-tarefa diz respeito à aquisição dos valores finais e eventuais critérios de filtragem exacta e ponderações específicas, após a filtragem dos descritores irrelevantes (importância nula) ou desconhecidos (não preenchidos), para assim definir o problema alvo, estabelecendo o perfil de procura entre os casos existentes.

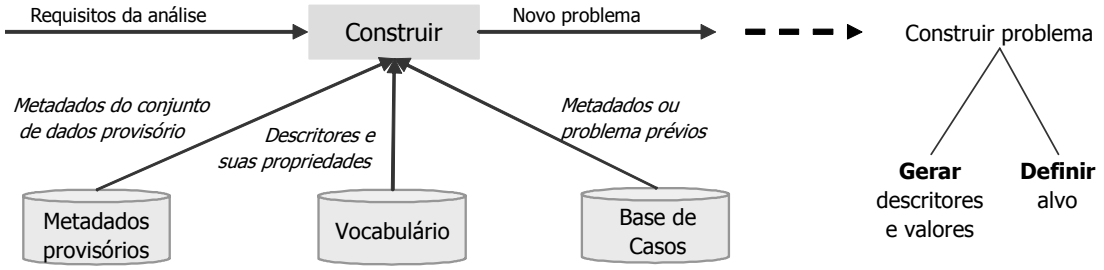


Figura 15 – Entradas, resultados e sub-tarefas do módulo de construção de problemas

Em suma, cabe à construção de problemas a função de recolha e organização de todos as restrições especificadas pelo utilizador, com base numa série de descritores de problemas e suas propriedades, no sentido de conceder ao sistema um problema alvo convenientemente estruturado e detalhado, no que concerne a constrangimentos explícitos.

4.5 Módulo de Recuperação

Uma funcionalidade essencial dos sistemas CBR é a recuperação dos casos mais similares que possam ser úteis para solucionar o problema alvo. O módulo de recuperação recebe como entrada a descrição do problema alvo, contendo os seus atributos descritores e os respectivos valores, ponderações e critérios de filtragem exacta, e devolve os casos que mais se assemelham ao

problema corrente, usando, sobretudo, os contentores de similaridade e da base de casos, conforme se ilustra na Figura 16. Para esse efeito, o módulo concretiza três sub-tarefas:

- usa o problema alvo anteriormente construído para procurar o subconjunto de casos candidatos plausíveis;
- compara os problemas dos casos candidatos com o problema alvo, calculando valores de similitude entre cada par, com base nos descritores de problemas e respectivas medidas de similaridade;
- selecciona os casos candidatos mais promissores, de acordo com um valor limite de semelhança.

Quanto maior for o nível de similaridade, maior será a probabilidade de obtenção de processos de WUM englobando modelos e operações de transformação semelhantes aos que foram aplicados noutros casos e que têm de ser executados no conjunto de dados actual.

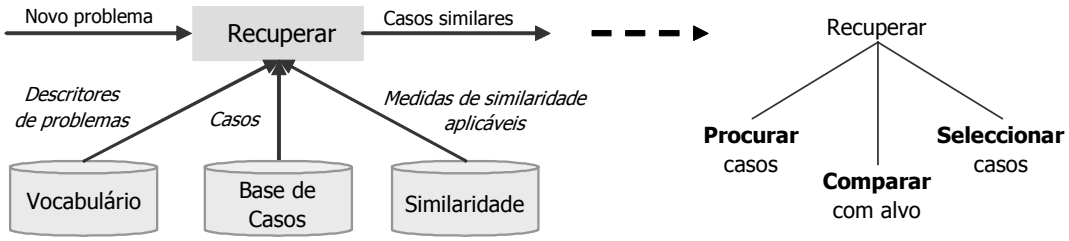


Figura 16 – Entradas, resultados e sub-tarefas do módulo de recuperação

A procura do subconjunto de casos prévios relevantes é tipicamente efectuada recorrendo a mecanismos de indexação. Os índices traduzem as situações em que os casos são significativos, com o intuito de evitar a comparação exaustiva do alvo com todos os casos existentes e, portanto, de tornar a funcionalidade de recuperação mais eficiente. A possibilidade de aceder e inspeccionar apenas os casos indispensáveis é inegavelmente preponderante, particularmente, quando a base de casos é numerosa e em condições de operação do sistema em que é exigido um elevado nível de desempenho. Apesar de se reconhecer este facto, ainda não se implementou nenhum esquema de indexação no sistema SPM. A respectiva sub-tarefa foi isolada das restantes actividades, a fim de simplificar o seu melhoramento futuro e, correntemente, apenas garante a pré-selecção de casos, atendendo a critérios de filtragem exacta.

Uma actividade fulcral para assegurar a funcionalidade de recuperação é a medição do nível de similitude do alvo em relação aos problemas previamente descritos, guardados na base de casos em conjunto com as respectivas soluções. Esta actividade constitui o foco desta secção, pois, para além dos muitos esforços envidados para a sua concretização, a mesma é representativa de um dos principais contributos desta dissertação. Antes de mais, é necessário especificar um modelo de similaridade, em consonância com a representação de casos adoptada e anteriormente discutida. O desafio desta incumbência prende-se com a definição de um modelo de semelhança para processos de WUM, capaz de lidar com uma formulação não proposicional, não abrangida pelas abordagens típicas de condução desta actividade. A representação proposicional clássica, composta por uma série de pares de atributos e valores com dimensão fixa, não é viável neste âmbito devido aos relacionamentos de um para muitos que existem entre as partes da descrição de casos de aplicação de WUM.

A representação estruturada e a determinação de similaridade no contexto de dados complexos, como seria de esperar, são questões proeminentes para uma ampla variedade de domínios de aplicação, sendo consideradas em várias linhas activas de investigação, para além da área do CBR. Neste sentido, torna-se oportuno pormenorizar as dificuldades encontradas na comparação de casos de aplicação de WUM para, após uma breve análise de abordagens alternativas, ser possível estabelecer um modelo de similaridade e medidas para aferir a similitude, consentâneas com esse modelo e com os requisitos particulares desta aplicação do paradigma CBR.

No que respeita à sub-tarefa de selecção final dos casos mais promissores, o sistema permite que o analista especifique o limite mínimo de similitude, procedendo à filtragem dos casos comparados, cujo nível de similaridade não atinge este valor, e conduzindo os restantes para a fase de reutilização.

4.5.1 Características da Comparação entre Problemas

O modelo conceptual adoptado e descrito na secção 4.2.1 revela a natureza multi-relacional da representação de casos de aplicação de WUM. Focando a parte da descrição de problemas mantidos na base de casos (recapitulada na Figura 17), a qual é objecto do mecanismo de recuperação, cada instância da classe central *Processo* está directa ou indirectamente relacionada com:

- uma instância de algumas classes (e.g. *Meta* e *Conjunto_dados*);

- várias instâncias de outras classes, como *Variável* (através da classe *Conjunto_dados*) e *Processo_Área*.

O segundo tipo de relacionamento (um para muitos) é, naturalmente, uma questão mais importante, pois é justamente a sua ocorrência que impede a adopção de uma formulação proposicional. Este tipo de interligação surge também na especificação de novos problemas alvo e, ainda, com maior incidência, devido à estrutura mais flexível do alvo. Por exemplo, ao contrário dos casos prévios, o alvo pode conter múltiplas metas de análise. Além disso, estes relacionamentos originam diversos tipos de comparações, entre os atributos do alvo e de cada problema existente.

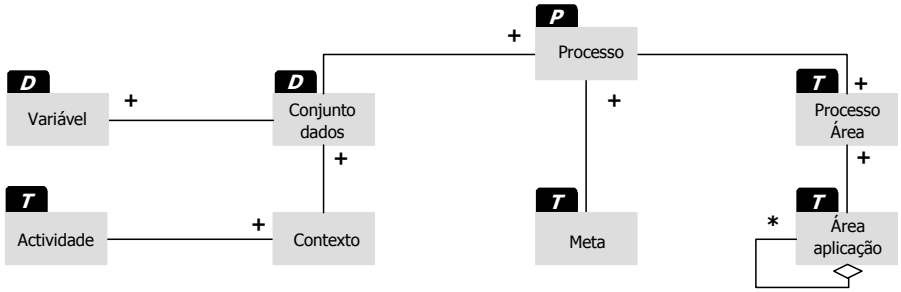


Figura 17 – Modelo conceptual de representação de problemas

A Tabela 4 enumera os atributos do problema alvo, em conjunto com os seus tipos de valores, organizados por categorias de descritores. Os atributos encontram-se ainda estruturados e foram contabilizados, na primeira coluna da tabela, em função das dimensões primordiais de descrição de problemas, anteriormente introduzidas na secção 4.2.1 (e.g. **P**, **T**, **D**). Por exemplo, a dimensão **D** inclui (14) descritores relativos às características consideradas ao nível do conjunto de dados. Já as características ao nível de variáveis, também pertencentes à mesma dimensão **D**, são contabilizadas através de um único descritor, respeitante à comparação entre cada par de séries de variáveis, em virtude de os (4) atributos (internos) indicados se referirem às propriedades das variáveis individuais. Os tipos de valores dos atributos do alvo foram denotados como:

- simples (ou atómicos) para atributos que contêm apenas um valor;
- compostos (ou estruturados), modelando os relacionamentos de um para muitos como uma série de elementos.

Tabela 4 – Lista de categorias, atributos e tipos de valores da descrição do problema alvo

Dimensões e número de descritores	**Categoria**	**Atributos**	**Tipo de valor e *(Tipo de comparação)***	
P	Critérios de avaliação	- Precisão dos resultados - Tempo de resposta - Exigência de recursos - Simplicidade de implementação - Interpretabilidade	Simples ordinal	*(1-1)*
Σ=6	Data do processo		Simples continuo	*(1-1)*
T	Actividade		Composto simbólico	***(N-1)***
Σ=3	Tarefa de DM	- Metas de análise - Áreas de aplicação	Composto simbólico	***(N-1)*** ***(N-M)***
D	Características ao nível do conjunto de dados	- Número de linhas ou registos - Número de colunas ou variáveis - Percentagem de variáveis numéricas - Percentagem de variáveis categóricas - Percentagem de variáveis temporais - Percentagem de variáveis binárias	Simples contínuo	*(1-1)*
14		- Tipo de identificação de visitantes - Tipo de registo de informação de visitantes - Disponibilidade de ordem de acesso a páginas - Disponibilidade de repetição de acesso a páginas - Disponibilidade de duração de acessos - Disponibilidade de data de acessos - Disponibilidade de hora de acessos - Granularidade dos dados	Simples binário ou simbólico	*(1-1)*
+ 1 (com 4 descritores internos) **Σ=15**	Características ao nível de variáveis	(conjunto de:) - Tipo de dados - Número de valores distintos - Número de valores nulos - Categoria semântica	Composto variáveis	***(N-M')***
Σ Total=24				

A Tabela 4 também reporta o tipo de confronto envolvido (e.g. 1-1, 1-N) entre os atributos correspondentes do alvo, em relação aos casos prévios. Os atributos estruturados geram comparações entre conjuntos de elementos finitos, cuja cardinalidade pode ser diferente e variável, evidenciando-se três tipos de confrontos, embora a sua generalização seja possível e desejável:

- N-1, ocorre entre conjuntos de valores simbólicos, para os atributos metas de análise e actividade, pois o alvo pode incluir a selecção de uma ou mais metas de análise e actividades, apesar de cada caso estar associado apenas a uma instância das respectivas classes *Meta* e *Actividade.*
- N-M, surge para séries de valores simbólicos de áreas de aplicação, uma vez que o alvo pode conter múltiplas áreas de aplicação e os casos podem estar relacionados com várias instâncias da classe *Processo_Área.*
- N-M', verifica-se em conjuntos de elementos com os seus próprios atributos (internos), na comparação entre os pares de séries de variáveis do conjunto de dados do alvo e de cada caso prévio.

4.5.2 Modelo de Similaridade

Nas aplicações típicas do paradigma CBR a determinação de similaridade é efectuada com base nos atributos, ditos superficiais, do alvo e dos casos prévios [Mantaras et al. 05]. Os atributos superficiais de um caso são os campos qualitativos e quantitativos, mantidos como parte da sua descrição, tipicamente representados por pares de atributo e valor. A similitude de cada problema de um caso, com um determinado problema alvo, é mensurada atendendo aos valores dos descritores correspondentes e às medidas de semelhança seleccionadas. As medidas de similaridade são um aspecto crítico em qualquer sistema CBR, contendo, só por si, conhecimento acerca da utilidade de uma solução passada reaplicada num novo contexto [Bergmann 01]. Um princípio genérico para orientar a estimativa do grau de semelhança consiste em subdividir a sua modelação, estabelecendo dois tipos de medidas: locais e globais. As medidas de similitude local são definidas e aplicadas sobre os atributos simples e individuais dos casos. As medidas de similaridade global consideram os casos como um todo e proporcionam um valor geral, entre cada par de casos, de acordo com alguma regra que combina os valores de semelhança local (e.g. uma função de agregação) e um modelo de ponderação. Não obstante, este princípio reconhecidamente

útil tem de ser estendido, de modo a lidar com os requisitos de uma representação multi-relacional.

A representação estruturada e a determinação de similitude em dados complexos são questões, hoje em dia, comuns e cruciais para diversas áreas, como o CBR, geometria computacional, aprendizagem automática e DM, especialmente no que concerne a métodos baseados em distância. Tarefas importantes para a última área incluem o agrupamento de objectos e a classificação de novas instâncias. Vários esforços de investigação visam suportar e aprender a partir de representações de dados mais expressivas e poderosas, do que a formulação proposicional. Os domínios da programação lógica indutiva e da DM multi-relacional, tradicionalmente, simbolizam as abordagens para lidar directamente com configurações mais complexas e intuitivas. Alguns exemplos proeminentes são os sistemas RIBL *(Relational Instance Based Learning)* [Emde e Wettschereck 96], RIBL 2 [Bohnebeck et al. 98] e RDBC *(Relational Distance-Based Clustering)* [Kirsten e Wrobel 98]. Por exemplo, o sistema RIBL constrói casos, a partir de dados multi-relacionais, e calcula a similitude entre casos de complexidade arbitrária, através do confronto recursivo de componentes de primeira ordem, até recair em comparações proposicionais, sobre os atributos elementares.

Em acréscimo, tem sido devotada grande atenção à extensão de algoritmos proposicionais de aprendizagem, particularmente, envolvendo linguagens de representação baseadas em tipos de dados estruturados. Tal representação viabiliza uma descrição intuitiva e próxima da proposicional, simplificando, frequentemente, a modelação de um problema [Flach et al. 98]. Nomeadamente, os métodos baseados em distâncias podem ser facilmente estendidos, incorporando medidas de similaridade definidas especialmente para tipos de dados estruturados, como listas, grafos e conjuntos.

O domínio do CBR também dá corpo a uma linha de investigação preponderante e relacionada, para tratar as questões referidas. Os casos são frequentemente representados através de estruturas complexas (e.g. grafos, objectos, termos de primeira ordem), exigindo formas especialmente ajustadas para aferir o grau de proximidade (e.g. similaridade estrutural) [Bergmann e Stahl 98]. Essas medições são, tipicamente, computacionalmente dispendiosas, mas podem ser recuperados casos com um nível de relevância superior. As representações de casos orientadas por objectos, actualmente muito vulgares, generalizam as formulações de atributo e valor, sendo úteis para retratar casos de áreas de aplicação complexas. Este tipo de representação é particularmente adequado quando podem ocorrer casos com estruturas diferentes [Bergmann 01]. O nível de semelhança entre objectos é avaliado recursivamente, recorrendo a

uma abordagem do particular para o geral (*bottom-up*). Este tipo de abordagem foi, ainda, estendido por uma proposta, a qual contempla a comparação de objectos pertencentes a classes distintas, considerando o conhecimento implícito na hierarquia dessas classes [Bergmann e Stahl 98].

Uma vez que a estrutura e os atributos usados para representar processos de WUM são sempre os mesmos para todos os casos, não é necessário lidar com o confronto entre objectos com configurações diferentes e arbitrárias, nem tão pouco mensurar formas mais profundas de similitude ou aplicar mecanismos recursivos, tal como os discutidos em [Bergmann 01], [Bergmann e Stahl 98] e [Emde e Wettschereck 96]. Pode-se optar por uma abordagem mais simples e intermédia, baseada na determinação de similaridade sobre tipos de dados estruturados, designadamente conjuntos de valores ou elementos, contendo os seus próprios atributos. Na realidade, existe um compromisso entre a expressividade das linguagens de representação e a eficiência (complexidade) do método de aprendizagem [Laer 02]. A estratégia de estender métodos proposicionais baseados em distâncias, através de representações estruturadas, capazes de simplificar a modelação do problema, bem como, de recorrer ao tratamento dos atributos e das suas propriedades, nas medidas de similitude, é vantajosa. Para além de ser uma estratégia mais simples, permite retirar benefícios da investigação e eficiência desses métodos, explorando, ao mesmo tempo, a maior expressividade de tais representações. Dado que esta estratégia é apropriada para atender aos requisitos estabelecidos e para enfrentar os respectivos desafios, a opção recaiu sobre a mesma. Neste sentido, a abordagem adoptada engloba a modelação dos seguintes tipos de mecanismos:

- Medidas de similaridade global, definidas através de uma função de agregação e de um modelo de ponderação (Sim_{Global}).
- Medidas de similitude local para atributos simples ou atómicos ($Sim_{Local\ simples}$).
- Medidas de semelhança local para atributos compostos ou estruturados, nomeadamente conjuntos de valores ($Sim_{Local\ conj}$).

O procedimento básico para o cálculo de similaridade entre casos ou problemas, usando o modelo anterior e omitindo o efeito de ponderações, pode ser resumido num algoritmo, apresentado, de forma simplificada, na Figura 18. Este procedimento é, ainda, abordado seguidamente, ao longo da enumeração das medidas de similaridade adoptadas, sendo também demonstrado, mais detalhadamente e conjugando ponderações de atributos, no Anexo A, recorrendo, para tal, a um exemplo concreto, introduzido no próximo capítulo.

Entradas: Alvo **a** e um caso prévio **c**, cujos problemas são descritos por uma série de descritores **d** e um destes descritores possui os seus próprios atributos internos **d'**

Saída: O valor de similaridade entre **a** e **c**

1. Ciclo para descritores

Para cada descritor **d**:

1.1. Similaridade para descritores simples

Se os valores de **d** são simples **então**

$Sim_d \leftarrow Sim_{Local\ simples}$ apropriada aplicada aos valores **d** de **a** e **c**;

1.2. Similaridade para descritores estruturados

Senão

1.2.1. Similaridade para conjuntos cujos elementos são valores simples

Se os elementos dos conjuntos são valores simples **então**

Usar $Sim_{Local\ simples}$ adequada entre cada par de valores de **a** e **c**;

$Sim_d \leftarrow Sim_{Local\ conj}$ aplicada a todos os valores anteriores para derivar um só resultado;

1.2.2. Similaridade para conjuntos cujos elementos são objectos com os seus próprios descritores **d'**

Senão

Usar $Sim_{Local\ simples}$ apropriadas aplicadas aos descritores internos **d'** dos objectos, para cada par de elementos de **a** e **c**;

Aplicar Sim_{Global} para agregar as similaridades dos descritores internos **d'**, derivando um valor de similitude para cada par de elementos de **a** e **c**;

$Sim_d \leftarrow Sim_{Local\ conj}$ aplicada a todos os valores anteriores para derivar um só resultado;

2. Similaridade global entre **a** *e* **c**

Aplicar Sim_{Global} a todos os valores Sim_d.

Figura 18 – Algoritmo explicativo do modelo de similaridade adoptado

4.5.3 Medidas de Similaridade

De acordo com o modelo de similitude anteriormente estabelecido, as medidas de similaridade requeridas, para levar a efeito a comparação entre o problema alvo e os problemas de casos

prévios existentes na base de casos, enquadram-se nas três categorias seguintes: globais (Sim_{Global}); locais para atributos atómicos ou simples ($Sim_{Local\ simples}$); locais para atributos constituídos por um conjunto de valores ou elementos ($Sim_{Local\ conj}$).

As medidas de semelhança global são aplicadas a diferentes níveis, nomeadamente ao nível de casos e de objectos estruturados, constituintes dos mesmos, com vista a agregar valores de similitude local dos seus atributos. A medida de similaridade global adoptada e implementada, até ao momento, trata-se da função tradicional de média ponderada, definida pela equação (1), em que *a* e *c* denotam o alvo e caso (prévio) ou parte destes, *a.d* e *c.d* são os respectivos valores de cada descritor *d*, Sim_{Local} é uma medida de similaridade local, *n* é o número total de descritores usados na comparação e w_d é a ponderação atribuída ao descritor *d*.

$$Sim_{Global}(a,c) = \frac{\sum_{d=1}^{n} Sim_{Local}(a.d, c.d) * w_d}{\sum_{d=1}^{n} w_d} \qquad (1)$$

As medidas de semelhança global são orientadas pela tarefa subjacente e contêm conhecimento acerca da utilidade dos casos [Bergmann 01]. A medida adoptada reflecte a relevância dos descritores de casos, com base nos respectivos pesos atribuídos, permitindo que estes assumam graus variáveis de preponderância. Segundo a perspectiva defendida neste trabalho, a definição de níveis de proeminência dos descritores (e.g. importante, muito importante) é útil para melhor especificar o problema e, consequentemente, pertence à sua descrição. Deste modo, o modelo de ponderação considerado, basicamente estabelece cinco graus de ponderação e representa o mapeamento entre os mesmos e os respectivos valores efectivamente usados internamente.

As medidas de similitude local, por sua vez, contêm conhecimento de domínio, retractando o relacionamento entre os valores de um descritor. A adopção de medidas de similaridade local depende, sobretudo, do domínio dos atributos, para além da semântica pretendida. O grau de semelhança entre descritores categóricos atómicos baseia-se em correspondências exactas, designadamente para atributos binários e de texto, ou pode ser expressa sob a forma de matrizes de similitude, especialmente para atributos simbólicos, as quais estabelecem o nível de proximidade entre cada par de valores possíveis. No que respeita a descritores numéricos atómicos, adoptou-se uma medida de semelhança muito vulgar neste âmbito, baseada na distância

de *Manhattan* normalizada (i.e. Similaridade = 1 – distância). Esta medida é definida na equação (2), na qual *a.d* e *c.d* denotam os valores do alvo e do caso para o descritor *d* e d_{max} e d_{min} são os valores máximo e mínimo observados para o descritor, cuja gama de variação é usada para normalizar o resultado.

$$Sim_{Local\ simples}(a.d, c.d) = 1 - \frac{|a.d - c.d|}{d_{\max} - d_{\min}} \quad (2)$$

A estimativa do grau de semelhança para os descritores relativos a critérios de avaliação é uma excepção, sendo sujeita a uma restrição imposta à equação (2), dada pela regra (3). Esta restrição foi aplicada pois os valores dos critérios de avaliação do alvo são entendidos como limites mínimos. Em acréscimo, maior significa sempre melhor, recorrendo à escala ordinal estabelecida para estes descritores. Por esta razão, a similitude só deve ser calculada quando o valor respectivo do caso é pior (mais baixo), assumindo-se o valor 1 nas restantes situações (i.e. maior ou igual).

$$Sim_{Local\ simples}(a.d, c.d) = \begin{cases} 1 & c.d \geq a.d \\ 1 - \frac{|a.d - c.d|}{d_{\max} - d_{\min}} & c.d < a.d \end{cases} \quad (3)$$

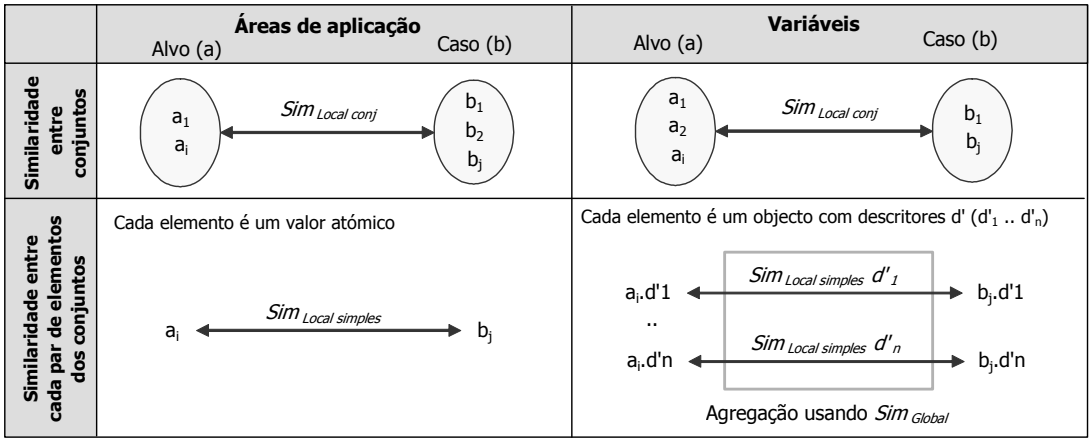
A representação estruturada dos casos de aplicação de WUM conduz a circunstâncias não contempladas pelas formas de determinação de similaridade anteriores. São requeridas medidas específicas para aferir o grau de similitude entre conjuntos de itens, conforme se ilustra na primeira linha da Tabela 5. Estes itens poderão ser valores atómicos (e.g. conjunto de áreas de aplicação) ou objectos com as suas próprias propriedades (e.g. série de variáveis).

Antes de definir medidas de similaridade para conjuntos, é necessário explicar o procedimento para mensurar o nível de semelhança entre cada par de itens dos dois conjuntos (e.g. um par de áreas de aplicação ou um par de variáveis). Esse procedimento já foi estabelecido no algoritmo da Figura 18 e engloba duas vertentes, esquematizadas na segunda linha da Tabela 5. A primeira vertente, exemplificada na coluna da esquerda, lida com valores atómicos, aplicando um dos mecanismos ao nível local, definido previamente. Presentemente, estes elementos atómicos são sempre valores simbólicos, cuja similitude foi modelada através de matrizes, contendo os níveis de

proximidade para todos os pares de valores possíveis. A segunda vertente, presente na coluna da direita da tabela, é devotada a pares de objectos e abrange os seguintes passos:

- uso de um mecanismo local, para estimar a similaridade entre cada par de valores correspondentes, para cada descritor interno dos dois objectos;
- aplicação de uma função de similitude global, para devolver um valor total, o qual combina os valores locais, atendendo às ponderações desses descritores internos.

Tabela 5 – Esquema de cálculo de similaridade entre pares de conjuntos



Pelo exposto, falta, no entanto, precisar a forma pela qual os valores de semelhança entre todos os elementos individuais do conjunto são conjugados para produzir uma valor único final. Existem muitas propostas na literatura para medir a distância ou similitude entre conjuntos de elementos ou itens [Eiter e Mannila 97] [Ramon 02] [Gregori et al. 05] [Hilario e Kalousis 03], algumas das quais se encontram definidas na Tabela 6. Por exemplo, [Eiter e Mannila 97] descrevem várias funções de distância para conjuntos de pontos, entre outras a soma das distâncias mínimas ($Dist_{SDM}$) e a métrica de *Hausdorff* (convertida em similaridade em Sim_H). Já [Hilario e Kalousis 03] abordam precisamente o desafio da comparação entre variáveis de conjuntos de dados, sugerindo a aplicação de três medidas, fundamentadas em noções da técnica de agrupamento, na qual o confronto entre conjuntos é uma tarefa comum. Duas dessas medidas são inspiradas em ideias provenientes de algoritmos de agrupamento hierárquico aglomerativo [Duda et al. 01]. A primeira (Sim_{SL}) é baseada no algoritmo de ligação simples (*single-linkage*), enquanto a segunda (Sim_{AL}) recorre ao algoritmo de ligação média (*average-linkage*). Quanto à terceira, advém do âmbito da

aprendizagem multi-relacional baseada em similaridade [Kirsten et al. 01] e consiste na medida usada no sistema RIBL (Sim_{RIBL}).

Tabela 6 – Definição de algumas medidas de similaridade ou distância para conjuntos

Medida	**Definição**
Soma das distâncias mínimas	$Dist_{SMD}(A,B) = \frac{1}{2}\left(\sum_{1}^{n_A} \min_{a \in A}(sim(a,b)) + \sum_{1}^{n_B} \min_{b \in B}(sim(a,b))\right)$
Hausdorff	$Sim_H(A,B) = \min(\min(\max_{a \in A}(sim(a,b))), \min(\max_{b \in B}(sim(a,b))))$
Single-linkage	$Sim_{SL}(A,B) = \max_{a,b}(sim(a,b))$
Average-likage	$Sim_{AL}(A,B) = \frac{1}{n_A * n_B} \sum_{1}^{n_A * n_B} sim(a,b)$
RIBL	$Sim_{MR}(A,B) = \begin{cases} \frac{1}{n_B} \sum_{1}^{n_A} (\max_{a \in A} sim(a,b)), & n_A < n_B \\ \frac{1}{n_A} \sum_{1}^{n_B} (\max_{b \in B} sim(a,b)), & n_A \geq n_B \end{cases}$
A e B denotam dois conjuntos, tais que a $\in$ A e b $\in$ B, sim(a,b) traduz a similitude entre cada par de elementos dos dois conjuntos e n_A e n_B correspondem à cardinalidade dos conjuntos A e B.	

Dada a diversidade de medidas sugeridas na literatura e em virtude de nenhuma ser apontada como a melhor, para a generalidade das situações, a escolha da mais apropriada não é uma tarefa fácil. De facto, a adequação de uma medida depende, particularmente, da semântica pretendida. Neste sentido, seleccionou-se um subconjunto contendo as medidas consideradas mais promissoras e, seguidamente, procedeu-se à respectiva implementação, a fim de levar a cabo um estudo comparativo entre as mesmas. A realização deste estudo possibilitou averiguar as propriedades, vantagens e desvantagens das medidas testadas. Adicionalmente, este estudo permitiu ganhar maior sensibilidade para as propriedades e semântica almejadas e, ainda, para as formas através das quais estes dois propósitos poderiam ser alcançados. Uma vez que o processo de cálculo subjacente é, por inerência, intensivo e dada a possibilidade de o mesmo envolver um número substancial de elementos, designadamente no que se refere a séries de variáveis, deu-se

prioridade a medidas passíveis de implementação eficiente. Entre os critérios de avaliação de medidas incluíram-se os seguintes requisitos:

- **R1** – identificar conjuntos idênticos, i.e. a propriedade de reflexividade (sim(x,x)=1) deve verificar-se.
- **R2** – distinguir significativamente pares de conjuntos muito diferentes de pares de conjuntos próximos entre si.

As diversas medidas implementadas requerem a determinação prévia do grau de semelhança entre todos os pares de elementos dos dois conjuntos, gerando uma matriz de similaridade, conforme se ilustra, mais adiante, na primeira coluna da Tabela 7, para dois tipos de configuração do alvo e caso comparado, em termos da cardinalidade dos seus elementos. As restantes colunas da tabela esquematizam o procedimento de estimativa da similitude, para um conjunto de medidas testadas, cujos resultados primordiais se reporta seguidamente.

Na métrica de *Hausdorff* a distância entre dois conjuntos A e B é dada pela distância máxima de um conjunto em relação ao elemento mais próximo do outro conjunto. A medida de semelhança baseada nesta distância (Sim_H) não é adequada para os propósitos correntes, pelo facto de se basear demasiado em valores extremos de dissimilitude dos elementos dos dois conjuntos. Contrariamente, pretende-se dar prioridade aos aspectos em comum entre os dois conjuntos. Além disso, esta métrica não usa informação da distância ou similaridade entre os restantes elementos dos conjuntos.

As medidas baseadas em agrupamento hierárquico aglomerativo também não se revelaram apropriadas. Na medida (Sim_{SL}) a similitude é determinada pelos elementos mais semelhantes dos dois conjuntos, falhando o requisito R2, devido a não usar informação dos outros elementos dos conjuntos. Já a medida (Sim_{AL}), pelo contrário, usa demasiada informação, falhando o requisito R1. Esta medida proporciona um nível de similaridade muito baixo, mesmo para dois conjuntos idênticos, pois as diferenças entre todos os pares de elementos são incluídas na média e a sua influência é, geralmente, maior do que o efeito das igualdades. Em acréscimo, o resultado da comparação entre conjuntos muito diferentes e semelhantes não é, usualmente, expressivamente distinto, falhando também o requisito R2.

A medida (Sim_{RIBL}) dá ênfase às melhores correspondências entre os dois conjuntos, mas é propositadamente sensível a diferenças entre cardinalidades, significando que produz valores de similaridade baixos quando o número de elementos dos conjuntos é distinto. A ideia subjacente é atingir a similaridade perfeita (identidade) apenas quando a cardinalidade é a mesma. No contexto presente, esta característica é desvantajosa. Os resultados da aplicação da medida geram uma

ordem inesperada, de acordo com a noção intuitiva de similitude, pois a medida penaliza muito uma diferença que não se deseja acentuar. Por exemplo, no caso da comparação de conjuntos de variáveis, já existe um descritor ao nível do conjunto de dados para reflectir esta divergência. Além disso, a utilização de um número distinto de variáveis não é um factor determinante para a selecção da maior parte dos métodos de DM.

Uma vez que nenhuma das medidas testadas se revelou ideal nem correspondeu às expectativas, combinou-se as melhores ideias de algumas delas para definir duas outras medidas de similaridade, mais ajustadas à abordagem e finalidades correntes. Estas medidas verificam os requisitos estabelecidos e demonstraram ser as mais adequadas para retractar a semântica pretendida, de acordo com os testes realizados.

A primeira medida baseia-se no mapeamento de cada elemento dos dois conjuntos para o elemento mais próximo no outro conjunto. A medida foi designada Sim_{MM} (i.e. Média dos Máximos), dado que considera as similitudes máximas, entre cada elemento de um conjunto em relação ao outro conjunto e nos dois sentidos, calculando a média de todos esses valores. A ideia subjacente é acentuar as maiores correspondências entre os dois conjuntos e assegurar um nível moderado de uso da informação disponível (maior do que Sim_{SL} e Sim_{RIBL} e menor do que Sim_{AL}). Esta medida é similar à soma das distâncias mínimas ($Dist_{SDM}$), após a transformação dos mínimos em máximos, a qual se baseia precisamente no tipo de mapeamento antes indicado. A medida Sim_{MM} é definida na equação (4), em que A e B são dois conjuntos, tais que a $\in$ A e b $\in$ B, sim(a,b) traduz a similitude entre cada par de elementos dos dois conjuntos e n_A e n_B denotam a cardinalidade dos conjuntos.

$$Sim_{MM}(A,B) = \frac{1}{n_A + n_B}\left(\sum_{1}^{n_A} \max_{a \in A}(sim(a,b)) + \sum_{1}^{n_B} \max_{b \in B}(sim(a,b))\right) \quad (4)$$

A segunda medida proposta foi construída com base na anterior, focando as melhores correspondências do conjunto alvo em relação ao conjunto de cada caso confrontado. Esta medida, Sim_{MMA} (i.e. Média dos Máximos do Alvo) é propositadamente assimétrica e usa menos informação do que Sim_{MM}, nomeadamente a parte ou sentido directamente respeitante ao alvo, dando ênfase às similaridades máximas de um dos conjuntos, tal como em Sim_{RIBL}. A sua principal característica é a orientação pelo alvo, o qual reflecte as propriedades relevantes que se espera encontrar. A medida é definida na equação (5), denotando A o conjunto contido pelo alvo, B o

conjunto de cada caso, sim(a,b) a similitude entre cada par de elementos dos dois conjuntos e n_A a cardinalidade do conjunto A.

$$Sim_{MMA}(A,B) = \frac{1}{n_A}\sum_{1}^{n_A}\max_{a\in A}(sim(a,b)) \quad \mathrm{A} \subset Alvo, \mathrm{B} \subset \mathrm{C}aso \tag{5}$$

A medida Sim_{MMA} é mais simples de estimar e, na generalidade, faculta melhores resultados. A medida Sim_{MM} usa mais indicações, mas não necessariamente mais informativas. Uma excepção ocorre quando o conjunto do caso se sobrepõe e excede o conjunto alvo. Nesta situação Sim_{MMA} não reflecte esse excesso, pois todos os elementos constantes no alvo foram encontrados no caso, apesar dos elementos adicionais. Este tipo de comportamento é justamente o desejado, no que respeita a descritores como a área de aplicação. Um utilizador que procura exemplos da área de aplicação ap1 ficará satisfeito com análises referentes, simultaneamente, às áreas ap1 e ap2. Esta questão não se coloca para os descritores da meta de análise e da actividade, pois cada caso apenas contém uma instância das mesmas. A maior adequação e eficiência da estimativa de Sim_{MMA} justificam, por conseguinte, a opção pela mesma para estes dois descritores.

Tabela 7 – Exemplo esquemático de medidas de similaridade para conjuntos

Matriz de similaridade entre A e B	***Sim H***	***Sim SL***	***Sim AL***	***Sim RIBL***	***Sim MM***	***Sim MMA***
Conjunto do alvo / *Conjunto do caso* (b_1 b_2): a_1: u v; a_2: w x; a_3: y z. **Máximo por linha (L_i)**: L_1, L_2, L_3. **Máximo por coluna (C_j)**: C_1 C_2	min (min (L_1,L_2,L_3), min(C_1,C_2))	max (u , v , w , x , y , z)	$\frac{u+v+w+x+y+z}{3*2}$	$\frac{C_1+C_2}{3}$	$\frac{L_1+L_2+L_3+C_1+C_2}{3+2}$	$\frac{L_1+L_2+L_3}{3}$
Conjunto do alvo / *Conjunto do caso* (b_1 b_2 b_3): a_1: u v w; a_2: x y z. **Máximo por linha (L_i)**: L_1, L_2. **Máximo por coluna (C_j)**: C_1 C_2 C_3	min (min (L_1,L_2), min(C_1,C_2, C_3))	=	=	$\frac{L_1+L_2}{3}$	$\frac{L_1+L_2+C_1+C_2+C_3}{2+3}$	$\frac{L_1+L_2}{2}$

O comportamento descrito para a medida Sim_{MMA} não pode, porém, ser considerado correcto quando se compara dois conjuntos de variáveis. O excesso não traduzido na medida, e no resultado da sua aplicação, poderá fazer toda a diferença. Neste tipo de confronto é conveniente recorrer a mais informação. Por exemplo, pretende-se que o sistema seja capaz de distinguir um conjunto de variáveis alvo de outro conjunto, contendo o primeiro e outras variáveis adicionais. Consequentemente, a medida Sim_{MM} foi eleita para a comparação de conjuntos de variáveis, enquanto Sim_{MMA} trata o confronto dos restantes tipos de conjuntos.

4.6 Módulo de Reutilização

Uma vez que a questão central e a missão do sistema SPM residem em recomendar métodos de mineração a aplicar em novas situações, e não apenas processos de WUM concretos, a sugestão de soluções baseada num ou mais casos específicos e nos seus níveis de similitude com o alvo não corresponde inteiramente às expectativas. O módulo reutilizar vai ao encontro de tais expectativas, construindo uma série de planos de WUM alternativos, para lidar com o problema alvo, a partir dos candidatos mais similares, devolvidos pela tarefa recuperar, tal como consta na Figura 19. Inicialmente, estes casos são agrupados, de acordo com os diferentes tipos de soluções técnicas representadas, em termos de métodos de DM. As diferentes soluções técnicas são determinadas pelos modelos de mineração efectivamente utilizados em etapas de modelação, dando origem ao conceito de categoria de modelação (ou categoria_modelação). Uma categoria_modelação trata-se, portanto, de uma combinação distinta de um ou mais modelos de mineração. Estas categorias são criadas automaticamente pelo sistema na fase de aprendizagem, mais precisamente, durante a catalogação de cada novo caso.

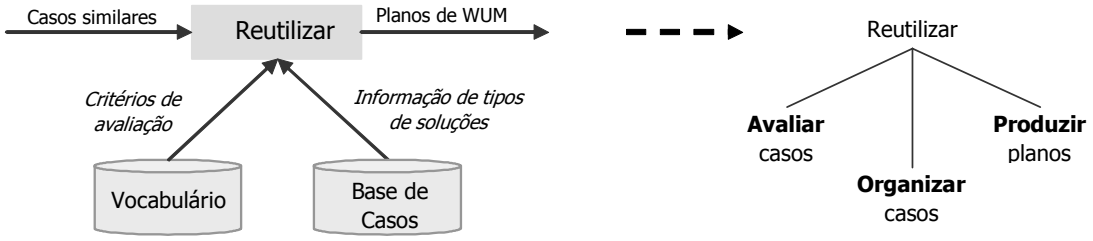


Figura 19 – Entradas, resultados e sub-tarefas do módulo de reutilização

Cada categoria_modelação, presente nos casos seleccionados, é usada para formar grupos e é avaliada, calculando o valor médio de cada critério de avaliação por grupo, de acordo com os critérios de avaliação definidos no contentor de vocabulário. Deste modo, é possível, por um lado, agrupar eventuais soluções parecidas (envolvendo a mesma categoria_modelação), para evitar a apresentação repetida de planos equivalentes, e por outro lado, obter uma perspectiva global da avaliação de cada tipo de solução, independentemente dos restantes descritores, passível de apoiar a decisão final do analista. Seguidamente, os casos candidatos são organizados e ordenados, em função da categoria_modelação do tipo de solução, do nível de similitude máximo do grupo e dos valores médios dos critérios de avaliação, com base na importância relativa anteriormente atribuída aos mesmos pelo analista. Procede-se então à produção de planos, considerando os grupos formados e o número k de soluções alternativas a devolver, previamente especificado pelo analista. Os primeiros k grupos são apresentados, pela ordem estabelecida anteriormente, sendo instanciados no caso mais similar do grupo, após a obtenção de informação geral acerca dos modelos e operações usados nas respectivas etapas de modelação e transformação. Os restantes casos de cada categoria_modelação dos planos apresentados, ou seja, aqueles que possuem um nível de similitude inferior, também são incluídos no resultado para viabilizar o acesso aos seus detalhes.

A organização da apresentação de soluções permite ilustrar a lista de k planos de mineração mais promissores, ordenados com base no nível de semelhança e nos valores dos critérios mais relevantes para o analista. Adicionalmente, esta organização foca as partes centrais dos casos candidatos que podem ser transferidas para o alvo, preparando o processo de reutilização derivacional (i.e. a reutilização do método que gerou a solução). Os casos pertencentes a cada grupo, desde o mais ao menos semelhante, podem ser acedidos para consulta dos detalhes conducentes ao apoio na aplicação dos métodos. Por conseguinte, para além de diversos tipos de soluções (planos) alternativos, podem ser proporcionadas várias instâncias exemplificativas, para cada tipo de solução sugerido.

A adaptação e revisão de planos de mineração decorrem fora do sistema, numa ferramenta de KDD. De facto, o sistema não efectua um ajuste extensivo de casos ao problema actual, no sentido alargado, subentendido pela tarefa reutilizar do ciclo CBR original, uma vez que não transforma soluções. O analista faz parte do processo de raciocínio, escolhendo entre as soluções alternativas e moldando e revendo aquela que elegeu ao problema actual, contando com descrições detalhadas e explicações para apoiar a execução dessas actividades. O processo de WUM final e os respectivos documentos PMML, automaticamente gerados pelas ferramentas de mineração em uso,

constituem elementos de entrada para a fase seguinte de aprendizagem, nomeadamente, do módulo de conciliação de descrições.

4.7 Módulo de Conciliação de Descrições

A descrição de um processo de WUM é uma tarefa árdua, devido à quantidade e diversidade de dados implicados. Uma estratégia especialmente voltada para facilitar esta actividade consiste em explorar novamente o padrão PMML, mas agora viabilizando a submissão de documentos exportados neste formato, para efeitos de automatização, pelo menos, de parte, da aquisição de casos de aplicação de WUM.

A vantagem do recurso a documentos PMML deve-se à aderência ao padrão, por parte de muitas ferramentas de KDD. No entanto, vários constrangimentos restringem uma exploração mais expressiva destes documentos. Primeiro, a adesão é significativa, mas, muitas vezes, apenas para versões mais antigas, com capacidades inferiores. Também se está condicionado pelos itens efectivamente fornecidos no documento exportado, uma vez que muitos são opcionais. Segundo, alguns aspectos do padrão ainda são limitados. As operações de transformação são um exemplo, pois não cobrem todo o conjunto possível de funções que podem ser necessárias para adquirir e preparar os dados para mineração. O padrão só representa expressões criadas autonomamente pelas ferramentas de KDD (e.g. normalização dos valores de entrada de redes neuronais). Tais transformações são irrelevantes, no presente contexto, dado que são executadas automaticamente, não existindo interesse na sua recomendação ao analista. A conjugação de modelos é outra limitação importante. Somente as duas versões mais recentes da especificação suportam a exportação de documentos que combinam a aplicação de vários modelos, e mesmo assim, apenas para alguns tipos de métodos (regressão e árvores de decisão). Contudo, é possível exportar a aplicação de cada modelo para um ficheiro individual e, portanto, gerar vários documentos relativos a um mesmo exercício de KDD. Terceiro, um documento PMML descreve o resultado do processo de mineração. Em contraste, pretende-se capturar o processo de desenvolvimento e aplicação, dando ênfase ao passo de modelação, incluindo as acções que permitiram derivar o conhecimento a partir dos dados iniciais e do ponto de vista explicativo. Esta divergência reflecte-se na omissão de muitos aspectos proeminentes. Por exemplo, alguns parâmetros de configuração são irrisórios para expressar as propriedades dos modelos resultantes, porém, são preponderantes para explicar como o modelo foi produzido. Finalmente, o documento PMML pode não estar disponível.

Mesmo reconhecendo as restrições referidas, a possibilidade de submissão de documentos PMML é compensadora, tratando-se de uma forma expedita de recolha de parte da descrição de casos. Por outro lado, também é imprescindível viabilizar a indicação de informação complementar por interacção com o utilizador. A missão da conciliação de descrições é justamente aceitar uma definição de processos heterogénea, baseada, simultaneamente, nestes dois tipos de submissão, construindo uma representação intermediária, coerente e devidamente sequenciada, passível de ser entendida por outros módulos do sistema. Esta tarefa suporta a descrição de processos de WUM, do ponto de vista da conveniência do analista, de acordo com a Figura 20, abarcando as seguintes sub-tarefas:

- combinação de documentos PMML e de informação prestada pelo analista;
- extracção dos elementos e atributos úteis de documentos PMML, para os transformar numa descrição compatível.

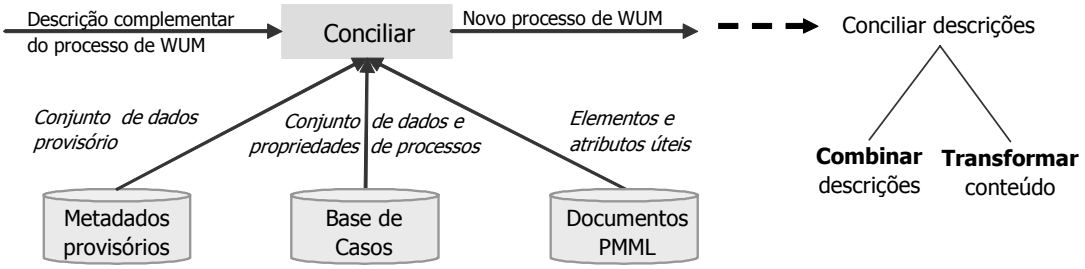


Figura 20 – Entradas, resultados e sub-tarefas do módulo de conciliação de descrições

A sub-tarefa combinar visa, essencialmente, sustentar um ambiente para simplificar a descrição de novos processos de WUM, contemplando dois propósitos divergentes. Por uma lado, pretende-se maximizar a aquisição de elementos sem o esforço do utilizador, designadamente recorrendo a documentos PMML, e, por outro lado, é necessário incentivar a cedência de informação e recolher os elementos que não é possível derivar, por meio de interacção com o utilizador, procurando auxiliar a sua realização. De qualquer forma, é imprescindível viabilizar a conjugação da descrição de etapas do processo via documentos PMML e interacção, assegurando a transparência das questões subjacentes e a correcta sequenciação das etapas definidas. As actividades a levar a cabo incluem as seguintes:

- Proporcionar informação sobre conjuntos de dados provisórios e existentes na base de casos, permitindo a escolha daquele a que o processo de WUM se reporta.

- Facultar categorias de informação geral, como metas de análise e áreas de aplicação, para possibilitar a classificação do processo.
- Permitir a enumeração dos documentos PMML que serão alvo de processamento posterior.
- Compensar a indisponibilidade ou omissões de documentos PMML, suportando a definição de etapas via interacção com o utilizador e, consequentemente, conceder todos os elementos requeridos para a especificação de acções de modelação e transformação.
- Viabilizar a edição de casos criados, para enriquecimento progressivo dos mesmos, até que a sua definição seja dada por concluída.

A primeira actividade torna transparente para o utilizador a questão interna de tratamento de conjuntos de dados provisórios ou já existentes na base de casos e, também, estabelece a interligação de um novo processo com um determinado conjunto de dados. A segunda tarefa diz respeito aos elementos gerais para descrições ao nível do processo. As duas actividades seguintes relacionam-se com a especificação de etapas do processo.

Para colmatar a limitação anteriormente mencionada da exportação de documentos PMML, quanto à aplicação combinada de vários modelos, decidiu-se possibilitar a submissão de múltiplos documentos. Esta funcionalidade também é necessária para processos desenvolvidos em diversas ferramentas. Desta forma, um processo de WUM pode ser representado por vários documentos PMML. Preparou-se, ainda, a evolução futura do padrão, pressupondo que cada documento pode conter mais do que um modelo, dependendo da respectiva versão de PMML, embora, correntemente, não se explore esta capacidade.

Já a especificação de etapas por meio de interacção envolve, em primeiro lugar, o acesso prévio a uma série de elementos descritivos disponíveis, tais como variáveis de conjuntos de dados, tipos de transformações, funções, modelos e parâmetros de configuração, entre outros. Dado o elevado número de alguns destes elementos, especialmente, de parâmetros de configuração, estes devem ser apresentados, salientando os que já foram aplicados a cada modelo, no sentido de facilitar a sua escolha.

Uma desvantagem da submissão de documentos PMML é a indisponibilidade imediata do seu conteúdo, pois, primeiro, tem de se proceder ao seu processamento e à extracção de elementos e atributos. Por conseguinte, poderá ser conveniente definir o processo em várias iterações. A consubstanciação progressiva de casos em diferentes fases responde a este requisito e, também, aos anseios de concluir e alterar a descrição de processos, exigindo, em acréscimo, a obtenção de toda a informação acerca dos mesmos, para efeitos de edição. Em suma, a sub-tarefa de combinação de descrições engloba grande parte dos elementos da base de casos e um suporte

robusto, não só para apoiar a especificação de itens, como também para proceder à recolha de elementos diversos.

A sub-tarefa transformar prende-se, basicamente, com capacidades de processamento e extracção de itens a partir de documentos PMML. Correntemente, dadas as limitações do padrão, apenas abrange etapas de modelação. Para concretizar esta sub-tarefa é conveniente suportar:

- distintas versões do padrão, garantindo a extensibilidade do sistema quanto a versões futuras;
- a extracção de itens específicos de cada modelo, atendendo às particularidades do modelo e de cada versão, uma vez que as evoluções do padrão se reflectem a diversos níveis;
- o processamento genérico subjacente de XML, pois um documento PMML é, antes de mais, um documento XML.

Os principais resultados da tarefa de conciliação são discutidos na secção seguinte, enquanto os meios para os alcançar serão abordados no próximo capítulo, dado que as questões subjacentes se relacionam com aspectos do âmbito da implementação.

4.8 Módulo de Retenção

A ideia defendida de gerir o conhecimento adquirido a partir da experiência, na resolução de problemas de WUM concretos, constituindo casos de aplicação e explorando o paradigma CBR ao nível da organização, afigura-se promissora. Duas forças proeminentes dos sistemas CBR, as quais contribuem substancialmente para tais expectativas, são a natureza mais intuitiva de casos, como representação de conhecimento, e a capacidade sustentada de aprender incrementalmente, através da incorporação de novos casos. Estas vantagens reduzem drasticamente os esforços de aquisição e manutenção de conhecimento [Watson e Marir 94], nomeadamente, em comparação com outros tipos de sistemas baseados em conhecimento. Contudo, estes dois aspectos não podem ser assumidos como garantidos. De facto, a memória de casos é um factor chave para a capacidade de resolução de problemas do sistema e, portanto, a pertinência, evolução e actualização do seu conteúdo são estratégicas.

Relativamente à aquisição de conhecimento, alguns desafios advêm do facto de os casos de domínio serem complexos, de a concepção e estruturação da base de conhecimento não serem incumbências triviais, pelo contrário, e ainda, de a informação vital para construir casos se encontrar dispersa, por diferentes tipos de recursos, ao longo da organização. Neste sentido,

especialistas entendidos na matéria são uma fonte importante de conhecimento e estratégias para permitir a integração de outros meios e recursos, com vista a reduzir estes desafios, tornam-se prementes. Por este motivo, procurou-se reduzir a necessidade de intervenção destes especialistas e, ainda, os esforços avultados de descrição de processos de WUM. Para este efeito, simplificou-se a captura de conhecimento ao longo da organização, recorrendo a uma abordagem semi-automática de aquisição do mesmo. Esta abordagem desenrola-se ao longo de vários passos do ciclo CBR adoptado, contando com a colaboração de alguns dos módulos já apresentados, particularmente, os respeitantes à caracterização de dados e, sobretudo, à conciliação de descrições. O módulo reter completa o ciclo, conjugando resultados produzidos por estes módulos com as suas próprias funcionalidades, para permitir que novos processos de WUM sejam adicionados à base de casos.

A tarefa desempenhada pelo módulo reter foi subdividida em três tipos de sub-tarefas (Figura 21). Primeiro, todos os itens concedidos pelo módulo de conciliação de descrições e pelos metadados provisórios são integrados. Segundo, estes itens são estruturados, de acordo com o esquema interno de representação de casos, e o processo é catalogado, com o intuito de simplificar a sua reutilização futura. Terceiro, o caso é registado na base de casos, testando a existência de diferentes itens, entre os quais funções de DM, modelos e parâmetros de configuração, e acrescentando as novas instâncias, assim como transferindo os metadados provisórios para a base de casos. Subsequentemente, o caso guardado pode ser editado e enriquecido com informação complementar, acerca de aspectos como, por exemplo, as descobertas de processos de WUM e mesmo novas actividades de mineração ou transformação.

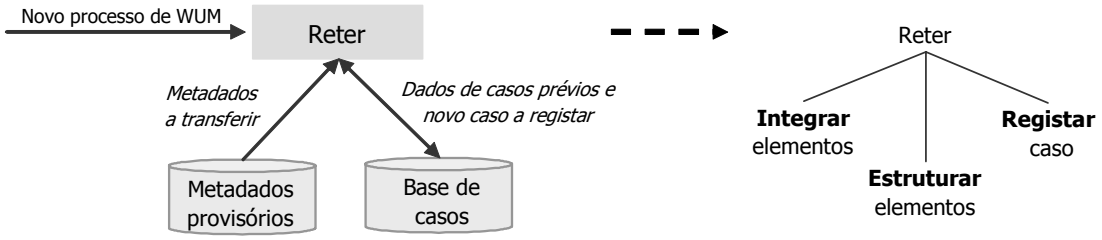


Figura 21 – Entradas, resultados e sub-tarefas do módulo de retenção

A Figura 22 apresenta novamente o modelo conceptual de representação de casos (introduzido na secção 4.2.1), salientando, agora, as classes cujo conteúdo pode ser, em grande medida,

capturado de forma automática (em tom mais escuro). Assinala-se ainda (a itálico) as classes que não são alvo de recolha automática, mas cujo conteúdo tem de ser instanciado (i.e. *Processo*, *Contexto*, *Actividade* e *Processo_Área*, usando também instancias prévias de *Meta* e *Área_aplicação*).

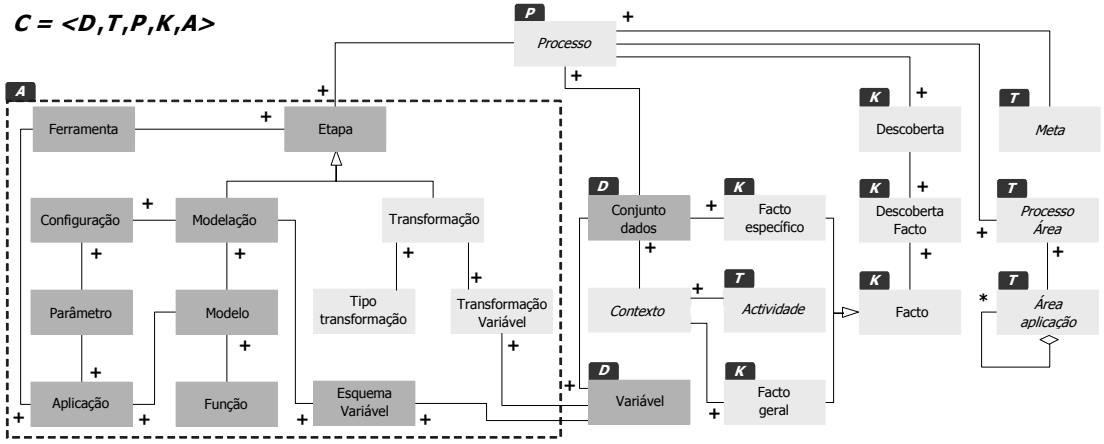


Figura 22 – Classes alvo de povoamento semi-automático

Seguidamente descrevem-se os aspectos primordiais da retenção de casos de aplicação de WUM, considerando três vertentes desta actividade, correspondentes a conjuntos de dados, actividades de mineração, baseadas em documentos PMML, e informação complementar, proporcionada via interacção com o utilizador. Na última vertente, aborda-se, ainda, a questão da manutenção e outras funcionalidades que o sistema deve assegurar.

4.8.1 Criação de Conjuntos de Dados

No ciclo CBR típico a aquisição de casos é iniciada quando um novo problema é submetido ao sistema, durante o seu uso para efeitos de resolução de problemas. A definição do novo problema é usada para basear a constituição de parte de um eventual novo caso que, depois de solucionado e conjugado com a descrição dessa solução, pode ser retido pelo sistema. O sistema SPM não é excepção, apesar de não recorrer à descrição de problemas, pois esta, geralmente, não é fidedigna para retratar um novo caso. O processo de KDD é exploratório, os resultados são imprevisíveis, as

soluções não são propriamente exactas e o papel do sistema é, precisamente, sugerir estratégias não antevistas pelo analista, alterando-se assim, usualmente, muitos dos pressupostos iniciais que basearam a descrição do problema. Por exemplo, para discriminar quais são as variáveis mais relevantes de um estudo é preferível recorrer ao esquema de mineração, o qual representa as variáveis efectivamente usadas em etapas de modelação, em detrimento das que foram indicadas pelo analista na descrição de requisitos. Consequentemente, a parte prévia aproveitada pelo sistema SPM, para a formação de um novo caso, consiste, essencialmente, na caracterização de dados, detida no repositório de metadados provisórios. As caracterizações acerca de dados alvo definitivos, usados em processos prévios, já se encontram na base de casos.

As caracterizações de dados disponíveis, sob a forma de um conjunto de dados provisório ou existente na base de casos, são utilizadas para basear a definição dos esquemas de variáveis de etapas de modelação e transformação. As etapas de transformação podem ainda derivar novas variáveis, as quais terão de ser acrescentadas aos dados, durante a retenção do processo. A especificação de etapas via interacção permite validar a adição e uso de variáveis. O mesmo já não sucede para documentos PMML, cuja verificação só pode ser efectuada após a integração dos itens resultantes do seu processamento. Deste modo, importa, antes de mais, identificar o tipo de relacionamento existente entre a caracterização de dados e documentos PMML, em virtude de os últimos também conterem metadados, designadamente no dicionário de dados.

A utilização de documentos PMML para fundamentar a caracterização de dados é inviável, por diversos motivos. Primeiro, o conteúdo e propósitos são distintos. O dicionário de dados representa um esquema e algumas propriedades gerais das variáveis de entrada de modelos, enquanto a caracterização de dados visa captar mais propriedades, ao nível de todo o conjunto de dados e das suas variáveis. Segundo, a caracterização de dados deve atender a vários tipos de fontes de dados, porém tem de ser independente dessas fontes e de ferramentas de mineração. As definições do dicionário de dados não dependem dos modelos de DM, mas reflectem a interpretação dos dados efectuada pela ferramenta subjacente. A este respeito, a caracterização de dados inspecciona os dados, a fim de verificar propriedades assumidas pelas próprias fontes. Por último, o documento PMML não está disponível durante a fase de resolução de problemas e é conveniente recorrer sempre aos mesmos métodos de caracterização, de forma a assegurar a consistência das comparações entre conjuntos de dados. Deste modo, o único uso de documentos PMML, para efeitos de criação de conjuntos de dados finais, reside na extracção de novas variáveis, ou seja, aquelas que não pertencem ao conjunto de dados inicial, nem tão pouco foram derivadas em etapas prévias de transformação. O surgimento de variáveis nestas circunstâncias é

permitido, dado o empenho em capturar todos os elementos significativos alcançáveis. No entanto, neste contexto, apenas são consideradas as propriedades que constam no dicionário de dados, no que concerne a essas variáveis. A criação de conjuntos de dados culmina com a transferência de todos os metadados provisórios para as classes *Conjunto_dados* e *Variável*. Se o processo se basear num conjunto de dados já existente na base de casos, apenas serão aditadas as novas variáveis, derivadas em etapas de transformação ou extraídas a partir de documentos PMML.

4.8.2 Aquisição Automática de Actividades de Mineração

As actividades de modelação de um exercício de KDD representam a parte central da solução de um problema de análise de dados neste âmbito e, também, a vertente com maior ênfase no sistema SPM. O objectivo da aquisição das actividades de WUM é, precisamente, captar todas as acções de modelação que conduziram à indução do conhecimento resultante, a partir dos dados iniciais. Grande parte destas acções pode ser obtida a partir de documentos PMML, com base nos elementos e atributos extraídos, passíveis de retractar as operações desenvolvidas.

A Tabela 8 apresenta as correspondências primordiais entre as classes do modelo conceptual de representação de casos e os itens PMML extraídos, pelo módulo de conciliação de descrições. As partes mais proeminentes para basear a descrição de casos não são as propriedades dos modelos, sob o ponto de vista da sua operacionalização, mas sim aquelas que permitem reproduzir as acções de modelação subjacentes, realizadas anteriormente para gerar o modelo. Daí advêm diversas diferenças entre estes tipos de representação. Antes de mais, recorre-se ao termo modelo, ao invés de algoritmo, pois, conforme já se referiu, o primeiro termo é mais comum em ferramentas de DM. A ideia é abstrair alguma complexidade, contando com o conceito mais intuitivo de modelo, eventualmente, como um recurso capaz de levar a cabo tarefas de modelação, indo ao encontro da filosofia adoptada, não só de redução de dificuldades, mas também de proximidade em relação ao ambiente de operação encontrado pelos analistas.

A diferença chave entre os dois tipos de representação coloca-se em torno do modelo de mineração *("mining model")*. O seu conteúdo contém uma hierarquia de elementos internos, não reproduzida na classe Modelação. Estes elementos são mapeados para várias classes, tal como se indica na tabela. As propriedades genéricas do modelo PMML, relativas à função e algoritmo aplicados, são usadas para obter as funções e modelos de DM e, mesmo, para registar novas instâncias das respectivas classes. Como algumas das propriedades específicas do tipo de modelo correspondem a parâmetros de configuração, fornecem os valores especificados para os mesmos e

permitem povoar as classes Parâmetro e Configuração. O esquema de mineração (*MiningSchema*), pertencente ao modelo de mineração, é utilizado em conjunto com o dicionário de dados (*DataDictionary*) para definir as instâncias da classe Esquema_Variável, a qual mantém informação acerca das variáveis envolvidas na modelação. O esquema de mineração traduz as variáveis requeridas para a execução do modelo, enquanto o dicionário de dados reflecte a forma como estas foram interpretadas pela ferramenta, podendo não ser coincidente com as propriedades detidas pelas mesmas no sistema (e.g. tipo de dados na classe Variável). Conforme já foi mencionado, o dicionário de dados permite ainda adquirir informação sobre variáveis inexistentes. Por último, os documentos PMML submetidos ao sistema são automaticamente registados como fontes de processos na respectiva classe.

Tabela 8 – Correspondências primordiais entre classes da representação de casos e itens PMML

Classes	**Elementos e atributos PMML**	**Descrição**
Ferramenta	(*Header*) *Application*	Indicação da aplicação que produziu o documento
Função e Modelo	*("mining model") functionName* e *algorithmName*	Propriedades genéricas de modelos
Parâmetro e Configuração	*("mining model")* nomes e valores de algumas propriedades de modelos	Propriedades específicas do tipo de modelo que correspondem a parâmetros de configuração
Modelação	*"mining model"*	Refere o modelo aplicado na etapa
Esquema _Variável	*DataDictionary* (e.g. *dataType* e *optype*) e *MiningSchema* (e.g. *usageType*)	Variáveis usadas no modelo, forma como foram interpretadas pela ferramenta e papel desempenhado no modelo
Variável	*DataDictionary*	Nome e tipo de dados de novas variáveis

No que concerne a omissões dos documentos PMML salientam-se duas. Primeiro, nem todos os parâmetros de configuração de modelos, com um papel significativo para os gerar, fazem parte das propriedades do modelo resultante. Segundo, alguns modelos de DM, como, por exemplo, as árvores de decisão, excluem as variáveis de entrada consideradas irrelevantes para a execução do modelo e, nestas circunstâncias, estas não constam no esquema de mineração. Caso exista interesse em indicar que essas variáveis excluídas também foram contempladas no desenvolvimento do modelo, estas terão de ser explicitamente integradas. Em acréscimo, os

documentos PMML não abrangem detalhes informativos e explicativos que o sistema permite incorporar.

As actividades de modelação constituem ainda a fonte de catalogação do processo, em função do tipo de solução representada por este e com base no conceito anteriormente discutido de categoria de modelação. A catalogação é realizada automaticamente pelo sistema, após a construção de um novo caso, sendo, ainda, actualizada quando um caso existente é alvo de alteração, adição ou remoção de etapas de modelação.

Diante do exposto, apesar de alguns aspectos da descrição das actividades de modelação dependerem da actuação do analista (e.g. parâmetros de configuração omitidos nos documentos PMML e explicações da etapa), muitos outros podem ser obtidos automaticamente, reduzindo expressivamente estes esforços. Por exemplo, a simples indicação das variáveis usadas numa etapa é trabalhosa, pois o seu número pode ser da ordem das centenas. Além disso, grande parte das classes concebidas é povoada, recorrendo a conceitos e vocabulário plenamente aceite e normalizado. Concluiu-se, pois, que a abordagem de aquisição semi-automática, além de possuir grande utilidade prática, diminui a necessidade de recorrer a especialistas da área para orientar a actividade de aquisição de conhecimento no sistema SPM.

4.8.3 Recolha de Descrições Complementares e Outras Operações

Conforme tem vindo a ser mencionado, é necessário proporcionar um mecanismo complementar de aquisição de processos de WUM, por interacção com o utilizador, com vista, por um lado, a colmatar as omissões dos documentos PMML e a acrescentar pormenores e explicações das acções de modelação concretizadas e das respectivas configurações, e, por outro lado, a definir completamente as actividades de modelação, quando os documentos PMML não estão disponíveis. Este mecanismo é, ainda, especialmente requerido para especificar outros aspectos indisponíveis em documentos PMML, tais como, operações de transformação.

A Figura 23 resume os componentes primordiais da descrição de um caso, apresentando a negrito os itens contemplados na captura automática, os quais, no que respeita a conjuntos de dados, também envolvem o fornecimento de informação pelo utilizador e, para os restantes aspectos, poderão ser alvo de descrição explícita, via interacção. Para além das actividades de modelação e transformação, é importante recolher uma categorização alargada de cada processo de WUM, pois torna-se essencial preservar o contexto específico requerido para encontrar, interpretar e avaliar as

soluções de mineração. Em particular, segundo [Mantaras et al. 05], quando os critérios de avaliação de soluções são complexos, a representação de casos deve incorporar informação adicional. Em processos de DM ou WUM estes critérios são complexos e, sobretudo, subjectivos, justificando a premência de uma categorização abrangente dos casos.

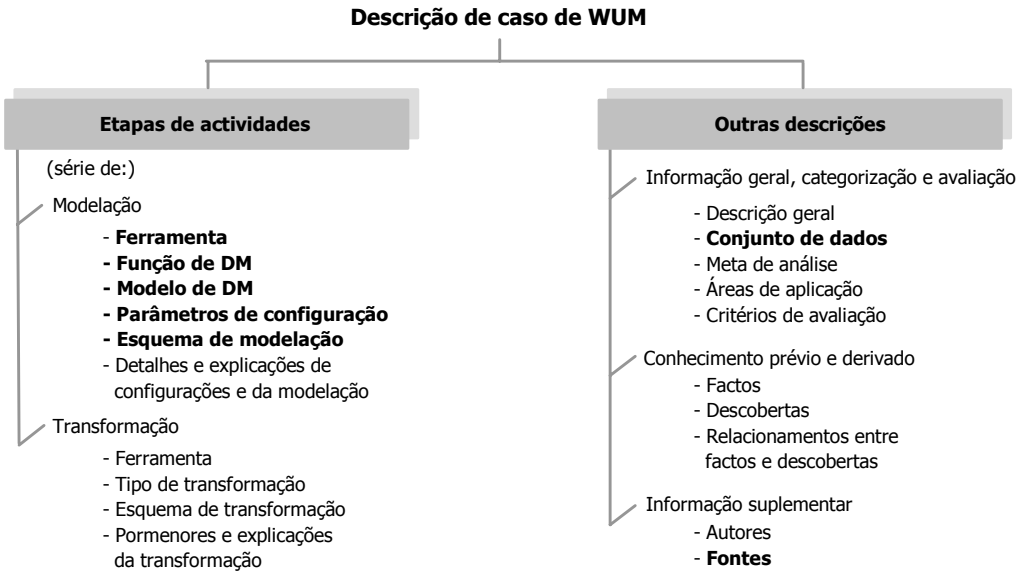


Figura 23 – Principais elementos constituintes da descrição de um caso de WUM

Relativamente a outras operações que o sistema deve assegurar, evidencia-se o acrescento, alteração e remoção de etapas, as quais exigem funcionalidades complexas, semelhantes às da criação de um novo caso e, ainda, a remoção de casos sem utilidade, levantando-se a questão da manutenção da base de casos, abordada seguidamente.

Um caso quando é adicionado possui o estado provisório, até que a sua descrição seja dada por concluída pelo utilizador, passando então ao estado confirmado. Alternativamente, um caso provisório ou confirmado pode ser alterado para o estado inútil, a fim de indicar ao administrador do sistema que o mesmo não possui interesse, sendo um candidato de remoção. Cada caso confirmado pode ainda ser convertido em protegido, pelo administrador do sistema, para distinguir os casos confirmados efectivamente representativos e úteis para a resolução de problemas dos restantes casos. Cabem, portanto, ao administrador do sistema as incumbências de avaliação dos casos disponíveis e de gestão do ciclo de vida dos mesmos, transformando os casos que deixaram

de ser úteis em obsoletos e, mesmo, decidindo, entre estes e os que possuem o estado inútil, aqueles que devem ser removidos. Os casos utilizados para assistir a resolução de problemas correspondem, por omissão, somente aos que se encontram em estado protegido, podendo ainda ser incluídos também os que possuem o estado confirmado, via solicitação explícita do analista, durante a submissão de um problema ao sistema.

A manutenção da base de casos é uma questão relevante e desafiante em sistemas CBR. O crescimento contínuo da memória de casos e a incorporação de casos sem utilidade ou redundantes diminuem o nível de desempenho, a eficácia e a capacidade de solucionar problemas do sistema. Esta questão ainda não foi objecto de um tratamento verdadeiramente consentâneo no sistema SPM, uma vez que ainda não se implementou nenhum mecanismo activo de suporte ao administrador do sistema, para o ajudar a discriminar casos sem interesse. A funcionalidade existente, correntemente, trata-se somente do controlo do estado dos casos, os quais são atribuídos pelos utilizadores e, particularmente, pelo administrador do sistema, atendendo às possibilidades já enumeradas – provisório, confirmado, protegido, obsoleto e inútil. Sucintamente, as principais políticas ou restrições na manipulação de casos são as seguintes:

- um caso protegido não pode ser alterado, podendo apenas ser objecto de mudança de estado, por parte do administrador do sistema;
- somente o administrador pode remover casos e estes têm de possuir o estado inútil ou obsoleto;
- os estados protegido e obsoleto só podem ser atribuídos pelo administrador do sistema.

No âmbito do auxílio ao administrador do sistema, está previsto, no curto prazo, acrescentar informação estatística, acerca da selecção e reutilização de casos, para assim ser possível mensurar o nível de utilidade dos mesmos. A questão da redundância é mais difícil de aferir, pois problemas completamente diferentes podem usar soluções do mesmo tipo e o mesmo problema pode ser resolvido de diferentes formas. Os detalhes de um caso também podem ser tão informativos, de forma a justificar a sua pertinência. De facto, assume-se a existência de alguma redundância. O modo de funcionamento do sistema inclusive contribui, em certa medida, para esta situação, ao permitir, em acréscimo, a integração dos casos que se encontram, apenas, em estado confirmado na resolução de problemas. Todavia, esta situação é gerida pelo sistema e procura-se, além disso, retirar benefícios através desta. Os planos de WUM apresentados ao utilizador agrupam os casos seleccionados por tipo de solução, sendo os eventuais casos adicionais aproveitados para facultar vários exemplos do mesmo tipo de solução.

4.9 Aplicação de CBR na Assistência à WUM

O sistema selector de planos de mineração, denotado abreviadamente por SPM, visa assistir o desenvolvimento e a aplicação de processos de WUM, recomendando as estratégias de KDD mais apropriadas para solucionar um determinado problema de análise de dados de *clickstream*. Estas estratégias são sugeridas sob a forma de planos de WUM, atendendo às características dos dados alvo e aos requisitos da análise, com base em casos úteis de aplicação de WUM, concretizados com sucesso no passado e registados na base de conhecimento do sistema. O SPM explora o paradigma CBR, actuando como uma ferramenta de gestão e reutilização de casos de aplicação de WUM da organização.

A base de conhecimento do sistema foi concebida, conjugando vários princípios orientadores da organização de conhecimento em sistemas CBR e, também, indicações proporcionadas pela especificação PMML, englobando os seguintes tipos primordiais de contentores de conhecimento:

- Vocabulário descritivo do domínio de aplicação, contemplando a definição de propriedades e papéis desempenhados pelos atributos, os quais são usados na interpretação e manipulação dos mesmos.
- Similaridade, respeitante a propriedades e matrizes de semelhança, para alguns atributos, que, em conjunto com as funções de similaridade, implementadas via programação, permitem aferir o nível de similitude entre casos.
- Repositório de metadados provisórios, contendo caracterizações de conjuntos de dados alvo de estudo.
- Base de casos, a qual mantém os exemplos úteis de desenvolvimento e aplicação de WUM, correspondendo cada exemplo a um caso, descrito em termos de um problema de domínio e da respectiva solução aplicada.

Um problema de domínio é, essencialmente, definido através da caracterização do conjunto de dados utilizado, categorizações do tipo de problema de WUM e critérios de avaliação do processo e dos seus resultados. A solução aplicada, por sua vez, é representada por descrições de actividades de mineração e transformação, conhecimento prévio e derivado e informação geral do processo.

Para além da base de conhecimento, o sistema é constituído por seis módulos fundamentais, responsáveis pela realização das principais tarefas do sistema, formando um ciclo de resolução de problemas e aprendizagem, a partir de experiência no domínio. Os módulos de caracterização de dados, construção de problemas, recuperação e reutilização, colaboram entre si para levar a cabo

a assistência na resolução de novos problemas, enquanto os módulos de conciliação de descrições e de retenção asseguram a aprendizagem do sistema.

A missão do módulo de caracterização de dados é derivar uma meta-representação sistemática dos dados alvo, capaz de captar propriedades influentes na selecção de métodos e abordagens de KDD, substituir os dados originais, em comparações entre conjuntos de dados, e garantir a independência e consistência da meta-representação, ao longo de diferentes tipos de fontes de dados. A caracterização de dados abrange propriedades específicas de dados de *clickstream*, fornecidas pelo analista, e, também, metadados com um carácter mais genérico, extraídos, automaticamente pelo sistema, a partir da fonte.

A especificação dos requisitos explícitos da análise é auxiliada pelo módulo de construção de problemas, com base numa série de descritores de problemas e num conjunto de dados ou problema disponível, previamente obtidos pelo sistema. Os requisitos expressam preferências e expectativas do analista, relativamente a características dos dados, critérios de avaliação ou sucesso e à tarefa de DM pretendida, sendo esta identificada por meio de abstracções relacionadas com os problemas reais enfrentados. É ainda possível estipular níveis de importância específicos e critérios de filtragem exacta, para todos os atributos descritores de problemas. Este módulo também trata a especificação de requisitos, para gerar um novo problema, convenientemente estruturado e detalhado a submeter ao sistema.

O módulo de recuperação tem a seu cargo a incumbência fulcral de devolver os casos prévios mais semelhantes ao problema alvo. No sistema SPM, esta incumbência levanta desafios acrescidos, dada a estrutura multi-relacional adoptada de representação de casos. A dificuldade crucial surge no tratamento de relacionamentos de um para muitos, existentes entre algumas partes da descrição de casos, os quais conduzem a comparações entre conjuntos de elementos, com dimensionalidade distinta e variável, pertencentes ao alvo e cada caso confrontado. Para lidar com estes desafios, foi necessário definir um modelo de similaridade consentâneo e estender medidas de similitude propostas na literatura, em ordem a corresponder aos requisitos subjacentes e à semântica de semelhança almejada. Independentemente dos múltiplos esforços envidados para a persecução do cálculo da similaridade entre processos de WUM, acredita-se que os respectivos resultados são representativos de um dos contributos primordiais desta dissertação. Por conseguinte, foi dada maior ênfase a este módulo neste capítulo.

Os casos recuperados são avaliados e organizados pelo módulo de reutilização, no sentido de produzir uma série de planos de WUM alternativos, em função do tipo de solução contida nos casos, nível de similaridade e valores médios de critérios de avaliação, considerados pela sua

ordem de relevância para o analista. Os planos apresentados constituem a solução proposta, ilustrando as estratégias mais promissoras para resolver o problema corrente e salientando os métodos que devem ser aplicados aos dados alvo, preparando, desta forma, a reutilização pragmática das recomendações. O sistema não transforma as soluções, de acordo com o problema actual. O seu objectivo é apoiar, e não substituir, o analista, o qual faz parte do processo de raciocínio, escolhendo entre as soluções alternativas e adaptando e revendo aquela que elegeu.

Os processos de WUM bem sucedidos e úteis para a organização são a fonte de aprendizagem do sistema. Todavia, a informação e conhecimento acerca dos mesmos são diversos e vastos, encontrando-se, ainda, dispersos por diferentes tipos de recursos, ao longo da organização. Para além das fontes humanas e de origens de dados, nas quais as últimas podem ser alvo de extracção automática de metadados significativos, muitas ferramentas de KDD podem exportar resultados de processos para documentos em formato PMML, criando um novo recurso, passível de um tratamento expedito. Ampliam-se, assim, as oportunidades e a pertinência de suportar uma abordagem semi-automática de aquisição de conhecimento, capaz de facilitar e sistematizar a captura do mesmo e de reduzir os esforços avultados de descrição de processos de WUM.

O módulo de conciliação de descrições visa sustentar um ambiente simplificado de submissão de descrições heterogéneas de exercícios de WUM, via documentos PMML ou interacção directa com o utilizador, combinando e compatibilizando essas descrições, através da conversão do conteúdo de tais documentos, para uma representação compreensível para outros módulos do sistema.

A representação concedida pelo módulo de conciliação de descrições é usada pelo módulo de retenção, cuja missão é promover, efectivamente, a capacidade de aprendizagem do sistema, por meio da incorporação de novos casos de aplicação de WUM. Para este efeito, o módulo integra os resultados derivados pelos módulos de caracterização de dados e de conciliação de descrições, estrutura o seu conteúdo, de acordo com o esquema interno de representação de casos, criando um novo caso que, depois de catalogado, é registado na base de casos. A abordagem semi-automática de retenção de casos abrange grande parte das classes de representação de processos de WUM, permitindo que estas sejam povoadas, com instâncias baseadas em conceitos e vocabulário aceite e normalizado, e reduzindo drasticamente os esforços de descrição explícita do seu conteúdo.

Capítulo 5

Implementação e Demonstração do Sistema SPM

5.1 Abordagem de Implementação

Um desígnio do sistema SPM é poder ser explorado num ambiente alargado, viabilizando a aquisição, partilha e reutilização do conhecimento detido sobre processos de WUM, ao longo de toda a organização. Actualmente, o foco do desenvolvimento ao nível organizacional incide em torno de aplicações baseadas na Web, dadas as inúmeras vantagens que advêm desta opção para as organizações. Entre estas vantagens, encontram-se a acessibilidade ou alcance estendido das aplicações, a plena conivência com ambientes heterogéneos, dada a independência de plataforma, e, naturalmente, a flexibilidade da sua manutenção centralizada, ao invés da configuração exaustiva de cada computador individual. A índole mais intuitiva, flexível, poderosa e, supostamente, uniforme do ambiente de interacção com o utilizador é, também, outro mérito apontado a tal categoria de aplicações.

No caso concreto do sistema SPM, o estilo inerente do tipo de interacção de aplicações Web torna-se muito conveniente. O conhecimento manipulado pelo sistema possui uma natureza complexa, introduzindo dificuldades na própria sistematização da sua apresentação. Existem vários tipos de elementos relacionados entre si e com conteúdo extenso, sugerindo uma abordagem de interacção baseada em hipertexto, de forma a favorecer e facilitar a acessibilidade sucessiva aos mesmos e

aos respectivos detalhes. Atendendo, sobretudo, a estes desígnios e condicionantes, decidiu-se implementar o sistema através de uma aplicação Web protótipo, segundo uma arquitectura cliente servidor típica, internamente organizada em três camadas de serviços: interacção, negócio e dados. A Figura 24 ilustra a arquitectura do protótipo do sistema, de acordo com o referido. As opções que orientaram a implementação do protótipo regeram-se por recorrer, preferencialmente, a *software* livre, com código aberto e multi-plataforma, assim como a padrões e a interfaces de programação de aplicações – *Application Program Interfaces* (API) – com grande aceitação.

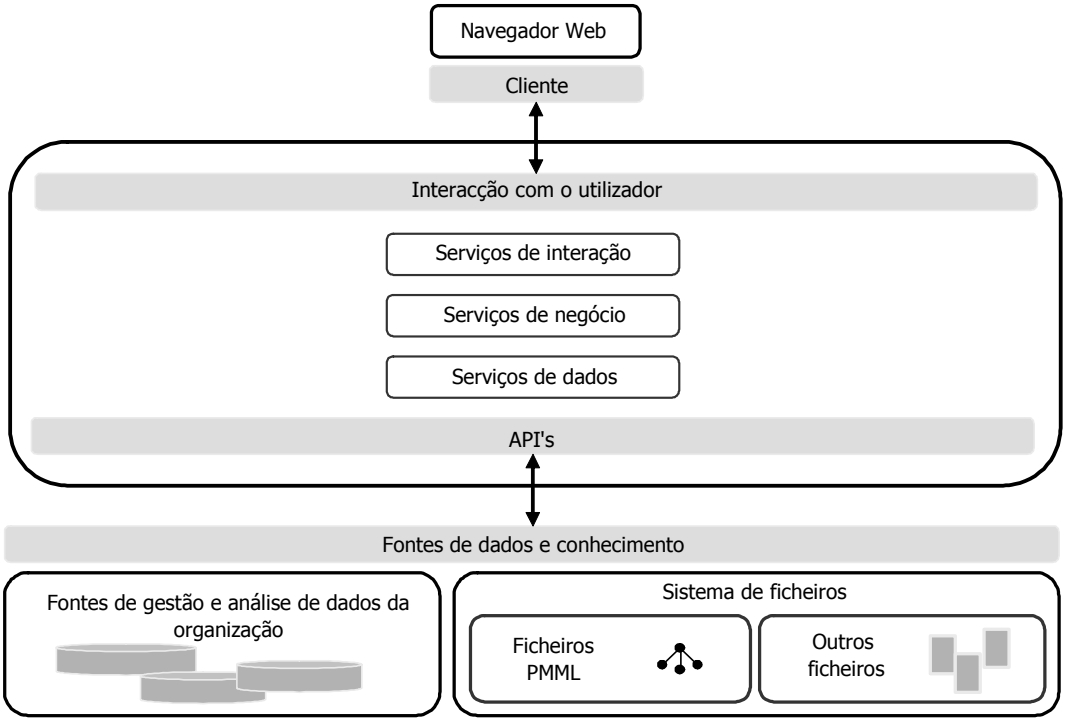


Figura 24 – Arquitectura do protótipo do sistema

As tecnologias aplicadas no lado cliente da aplicação consistiram, essencialmente, na linguagem de marcação de hipertexto HTML *(HyperText Markup Language)* [WWW1], complementada por folhas de estilo CSS (*Cascading Style Sheets*) [WWW1], para efeitos de formatação, e por programação em *Javascript*, para validação e melhoramento do comportamento do navegador e da interacção com o utilizador.

O lado servidor da aplicação foi desenvolvido em ambiente Java [WWW7], mais precisamente, na plataforma *Java 2 Platform Standard Edition* (J2SE versão 1.5.0, também referida por 5.0) [WWW8]. A lógica de negócio está a cargo de componentes Java e os serviços de interacção utilizam a especificação *Java Server Pages* (JSP) [WWW12]. A publicação de serviços é assegurada pelo contentor de JSP e *Servlets Apache Tomcat* (versão 5.5) [WWW13]. Os serviços de dados exploram diferentes API, para sustentar e abstrair o acesso a distintos tipos de fontes de dados e conhecimento, entre as quais, a base de conhecimento do sistema, implementada por meio de um SGBD relacional convencional. O protocolo e API *Java Database Connectivity* (JDBC) [WWW9] viabiliza o acesso e manipulação de diferentes fontes relacionais. O processamento de documentos XML e PMML baseou-se no modelo de objectos da especificação *Document Object Model* (DOM) [WWW1], na interface de programação *Java API for XML* (JAXP) [WWW10] e no analisador sintáctico *Apache Xerces* [WWW11], fornecido, por omissão, na versão da plataforma Java adoptada.

No desenvolvimento e exploração da aplicação usou-se vários tipos de ferramentas de suporte. As páginas Web foram concebidas no *Microsoft FrontPage 2002* [WWW16]. As classes Java foram implementadas no ambiente integrado de desenvolvimento *Eclipse* (versão 3.0) [WWW14]. Para a integração do servidor *Tomcat* no ambiente *Eclipse* recorreu-se ao *plug-in Eclipse Tomcat Launcher* (versão 3.0.0) da *Sysdeo* [WWW15]. Quanto a *software* de DM ou KDD e estatístico, utilizou-se, maioritariamente, a ferramenta livre *Waikato Environment for Knowledge Analysis* (WEKA versão 3.4) [Witten e Frank 05] e as ferramentas comerciais *Statistical Package for the Social Sciences* (SPSS versão 13.0) [WWW5] e SPSS *Clementine* (versão 8.5) [WWW4].

A Figura 25 mostra os principais blocos de construção do protótipo do sistema SPM (bibliotecas Java), em termos das três camadas fundamentais de serviços, acima mencionadas. Os serviços de interacção facultam acesso às funcionalidades do sistema, actuando como clientes da lógica de negócio, encapsulada na camada intermédia de serviços de negócio, e suportando a lógica de apresentação. Para esse fim, estes serviços recebem os pedidos dos utilizadores, analisam esses pedidos e os seus parâmetros, recolhem os dados concedidos e delegam o processamento nas classes de lógica de negócio, devotadas ao tratamento de cada tipo de pedido, para, então, construírem respostas com a apresentação adequada. Estes serviços baseiam-se directamente em dois tipos de componentes: uma biblioteca de serviços de interface (*serviçosinterface*) e páginas Web estáticas e dinâmicas. Os elementos mais relevantes destes dois tipos de componentes serão apresentados ao longo das próximas secções.

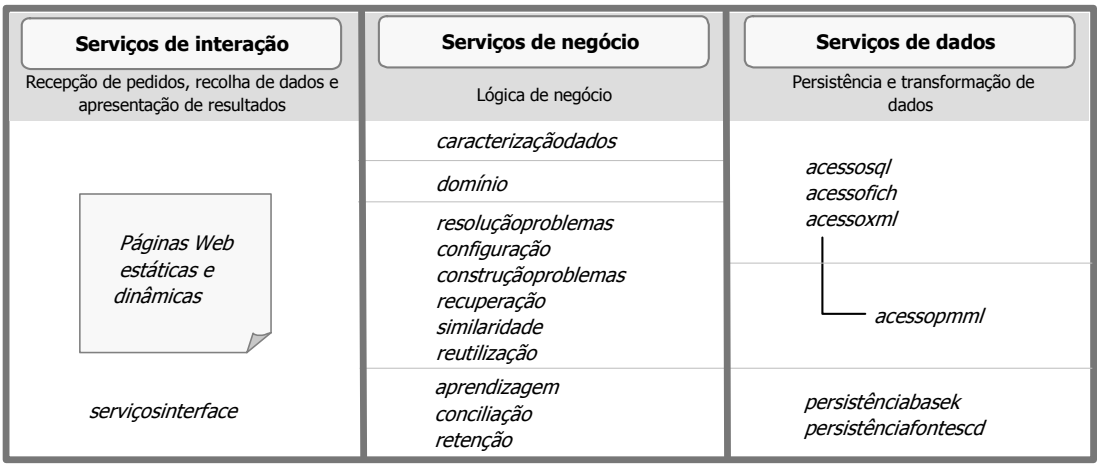


Figura 25 – Principais blocos de estruturação do sistema por camadas de serviços

A camada de serviços de negócio encapsula a funcionalidade nuclear do sistema, tendo sido estruturada nos seguintes componentes primordiais:

- *caracterizaçãodados* – cria caracterizações de conjuntos de dados e permite o acesso e manipulação dos mesmos, explorando, sobretudo, serviços de persistência;
- *domínio* – representa e gere toda a colecção de objectos constituintes de processos de WUM, com a excepção de conjuntos de dados, em colaboração com as camadas de dados e interacção. Esta biblioteca visa, essencialmente, promover a separação entre o modelo de domínio e os métodos que levam a cabo o processo de raciocínio;
- *resoluçãoproblemas* – conjuga e coordena as actividades necessárias para conduzir a resolução de problemas;
- *configuração* – encapsula os itens dos contentores de vocabulário, similaridade e configuração, possibilitando o acesso sistemático às propriedades subjacentes, de forma a contribuir para que os mecanismos de recuperação e reutilização possam assumir um carácter mais genérico e extensível;
- *construçãoproblemas* – assiste a descrição de requisitos, produz definições de problemas alvo e actualiza certas propriedades dinâmicas, requeridas durante os passos de recuperação e reutilização, maioritariamente, com base em capacidades complementares de configuração, persistência e interacção;
- *recuperação* – é responsável pelo passo de recuperação, sendo auxiliado, particularmente, por serviços de similaridade e configuração;

- *similaridade* – proporciona uma série de medidas de similaridade local e global, as quais são seleccionadas durante a recuperação de casos, de acordo com as configurações estabelecidas. Novas medidas de similitude podem ser acrescentadas, sem afectar as restantes bibliotecas, tendo apenas de ser activadas por meio da sua configuração;
- *reutilização* – concretiza o passo de reutilização, usando, basicamente, funcionalidades de configuração e persistência de dados;
- *aprendizagem* – interliga as actividades necessárias para levar a cabo a aprendizagem;
- *conciliação* – compatibiliza as duas formas de descrição de processos de WUM, apoiando-se, fundamentalmente, em serviços de interacção e de transformação de dados;
- *retenção* – implementa o passo de retenção de novos casos, contando, principalmente, com serviços de persistência de dados.

Por sua vez, a camada de dados assegura capacidades de persistência e transformação de dados, podendo ser, conceptualmente, subdividida em três grupos de serviços com diferentes propósitos. O primeiro grupo contempla funcionalidades de persistência, maioritariamente genéricas, orientadas pelo tipo de fonte e categoria de acesso à mesma. As bibliotecas *acessosql*, *acessofich* e *acessoxml* constituem este grupo, integrando serviços devotados aos correspondentes tipos de fontes de dados e conhecimento, consultados e manipulados pelo sistema, nomeadamente:

- bases de dados locais ou remotas;
- diversos tipos ou formatos de ficheiros (e.g. txt ou csv), frequentemente usados como origens de dados de utilização da Web;
- documentos XML, mais propriamente PMML.

Tais funcionalidades genéricas são partilhadas e reutilizadas pelos restantes componentes da camada de dados, determinando os tipos de fontes, presentemente, suportados pelo sistema. O segundo grupo inclui somente a (sub-)biblioteca *acessopmml*, também voltada para um tipo de fonte genérico, mas conferindo-lhe um tratamento particular. *acessopmml* lida com documentos PMML, apoiando-se nas capacidades de processamento de XML, facultadas pela biblioteca *acessoxml*. Contudo, a sua missão é disponibilizar serviços de extracção e transformação do conteúdo destes documentos, particularmente, de acordo com as necessidades específicas deste sistema. O terceiro grupo, formado pelas bibliotecas *persistênciabasek* e *persistênciafontescd* também presta serviços predominantemente específicos. A biblioteca *persistênciabasek* destina-se à manipulação da base de conhecimento. Já *persistênciafontescd* prende-se com o acesso a diferentes tipos de fontes de conjuntos de dados (cd) alvo de análises, abrangendo, correntemente, apenas duas das origens mais comuns de dados de *clickstream*: bases de dados e ficheiros simples. A finalidade destas duas bibliotecas é fornecer mecanismos para lidar com estes

recursos, garantindo o seu encapsulamento. Por exemplo, *persistênciabasek* isola e abstrai os detalhes de implementação do esquema e armazenamento de componentes da base de conhecimento, entre os quais, a base de casos e o repositório provisório de metadados.

5.2 Implementação das Principais Operações do Sistema

As operações ou funções fulcrais do sistema SPM são a resolução de problemas e a aprendizagem, a partir de exemplos de problemas solucionados. Estas duas funções desenrolam-se conjugando as sub-tarefas desempenhadas pelos módulos constituintes do sistema, já abordadas no capítulo anterior e ilustradas na Figura 12, bem como utilizando os vários componentes do protótipo, introduzidos na secção precedente. A exposição que se segue, acerca da implementação destas duas funções, reporta-se, simultaneamente, a essas sub-tarefas e componentes, procurando estabelecer a correspondência entre as primeiras e os elementos desses componentes que as concretizam.

5.2.1 Resolução de Problemas

A função de resolução de problemas engloba as tarefas caracterizar, construir, recuperar e reutilizar. A Tabela 9 mostra a equivalência entre as sub-tarefas incorporadas nessas tarefas e as principais classes que as implementam. A descrição da implementação da resolução de problemas foi organizada em três fases. A primeira fase consiste na caracterização de um conjunto de dados alvo. A Figura 26 apresenta as classes primordiais, envolvidas nesta fase, utilizando um diagrama de classes em notação UML simplificada e acrescentando informação sobre as bibliotecas a que as classes pertencem. A classe *InterfaceCaractDados* recolhe todas as indicações concedidas, através de interacção com o utilizador, e delega o seu processamento na classe central da caracterização de dados, designada *CaractDados*. Esta classe, por sua vez, desencadeia a extracção automática de metadados e o posterior armazenamento persistente dos metadados fornecidos ou derivados automaticamente, recorrendo, intensivamente, aos serviços da camada de dados. A classe *InfVariável* define as propriedades e métodos de tratamento das variáveis extraídas.

A interface <<*IPFonteCD*>> abstrai e confere independência ao acesso e extracção de metadados, em relação ao tipo de fonte subjacente do conjunto de dados, especificando a assinatura do método responsável por esta actividade. As classes *PFonteCDFich* e *PFonteCDBD* implementam o

método declarado nessa interface e utilizam os serviços de persistência de classes genéricas *(FluxoFT* e *FluxoBD, FonteBD)* e, também, os disponibilizados pelas subclasses *ExtractFT* e *ExtractBD*. Estas duas subclasses levam, efectivamente, a cabo a extracção de metadados, por meio de operações específicas, dependentes do tipo de fonte. A persistência de metadados é garantida pela subclasse *PCaractDados* que, por sua vez, explora os serviços da (super) classe geral *PGeralBaseK*, a qual centraliza os serviços gerais e comuns de persistência da base de conhecimento. Esta classe reutiliza os serviços genéricos da biblioteca *acessosql* (classes *FluxoBD e FonteBD)*, no acesso e manipulação do conteúdo relacional da base de conhecimento.

Tabela 9 – Correspondência entre sub-tarefas e classes da resolução de problemas

Tarefa	**Sub-tarefas**	**Principais classes de implementação das sub-tarefas**
Caracterizar dados	**Recolher** indicações	*InterfaceCaractDados*
	Extrair metadados	*IPFonteCD, PFonteCDFich, PFonteCDBD, ExtractFT, ExtractBD*
	Preservar metadados	*PCaractDados*
Construir problema	**Gerar** descritores e valores	*InterfaceRequisitos, GerarAtribVals*
	Definir alvo	*AlvoCbr*
Recuperar	**Procurar** casos	*CasosPréSelecção*
	Comparar com alvo **Seleccionar** casos	*CasoCbr*, classes da biblioteca *similaridade* e *CompararSelFinal*
Reutilizar	**Avaliar** casos	*MédiasCritérios*
	Organizar casos	*ItensOrdenação*
	Produzir planos	*ProduzirPlanos*

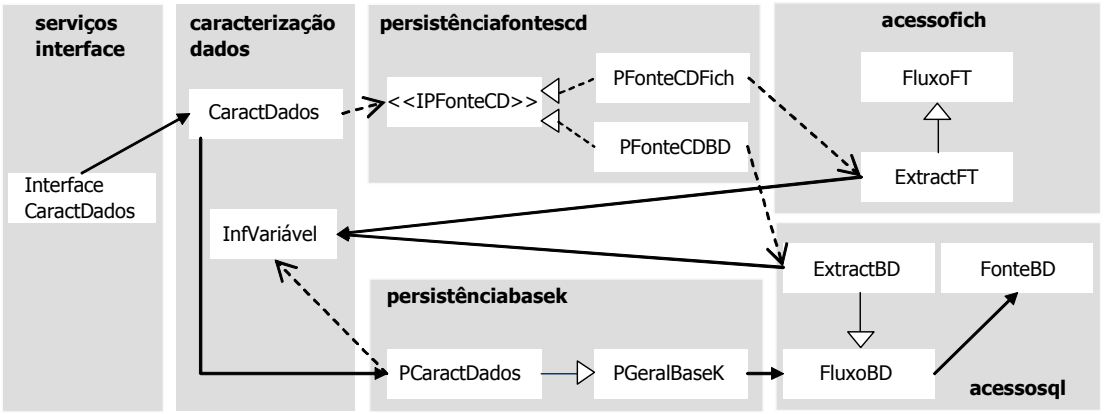


Figura 26 – Diagrama de classes da caracterização de dados

A classe *CaractDados* desempenha, ainda, papéis adicionais, contando com apoio de interacção, prestado pela classe *InterfaceCaractDados*, ou em resposta a solicitações de outras classes da aplicação. A classe *CaractDados* conduz operações suplementares, como a consulta, alteração e remoção de conjuntos de dados, mantidos, quer na base de casos, como no repositório de metadados provisórios, gerindo, portanto, internamente, esta questão e os aspectos comuns e diferentes destas duas formas de conjuntos de dados.

A segunda fase da resolução de problemas consiste na descrição de requisitos. Esta fase apenas abrange a primeira sub-tarefa da construção de problemas, sendo a sua implementação retratada pelo diagrama de classes da Figura 27. A descrição de requisitos tem início com a obtenção e apresentação das listas de casos existentes e conjuntos de dados prévios ou provisórios, com vista a permitir que um destes aspectos seja usado para basear e auxiliar a especificação dos requisitos da análise corrente. Esta acção é despoletada pela classe *InterfaceRequisitos* que, por sua vez, recorre à classe *GerarListas* (não ilustrada na figura), para consolidar e disponibilizar estes elementos. Seguidamente, a classe *GerarAtribVals* faculta os descritores e as suas propriedades, instanciando, ainda, os valores por omissão dos descritores, com um problema ou conjunto de dados escolhido. Estas actividades são efectuadas contando com a colaboração de diversos tipos de classes, entre os quais *PropsRR*, *CaractDados*, várias classes da biblioteca de *domínio* e as respectivas classes de persistência das duas últimas. A ajuda subsequente à formulação do problema corrente é, essencialmente, concretizada por funcionalidades de interacção com o utilizador.

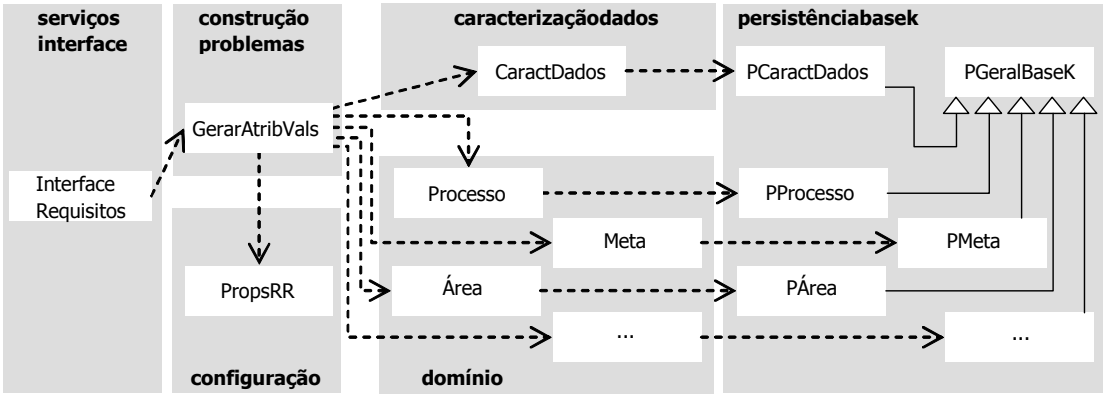


Figura 27 – Diagrama de classes da descrição de requisitos

A finalização da descrição dos requisitos permite encetar a terceira fase da resolução de problemas, implementada com base no diagrama de classes da Figura 28. Esta fase, denotada por produção da solução, é desencadeada pela classe *InterfaceResolução* e realizada pela classe *ResolverProblema*. A classe *ResolverProblema* interliga a sub-tarefa de definição do problema alvo com as actividades que procuram solucionar o problema. A classe *AlvoCbr* implementa a sub-tarefa definir alvo e simboliza o seu resultado. Esta classe utiliza, ainda, métodos da classe *PropsRR*, a fim de alterar algumas propriedades dinâmicas, com base nas indicações do utilizador, tais como, os níveis de relevância específicos atribuídos aos descritores e a ordenação de critérios de avaliação, em função da importância relativa especificada. A classe *Resultados*, por sua vez, combina e controla a persecução das tarefas recuperar e reutilizar.

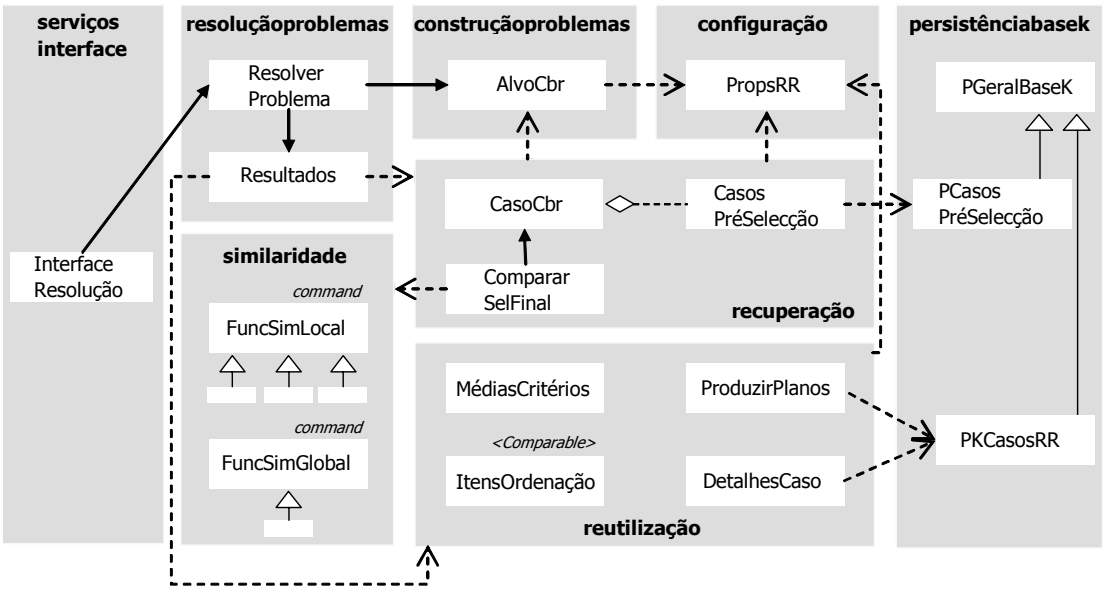


Figura 28 – Diagrama de classes da produção da solução

A classe *CasosPréSelecção* representa todos os casos candidatos plausíveis, implementando a sub-tarefa procurar casos. Para este efeito, esta classe usa as especificações do alvo, contidas num objecto do tipo *AlvoCbr*, e os serviços da respectiva subclasse de persistência *PCasosPréSelecção*. O método de procura da subclasse *PCasosPréSelecção* devolve apenas o subconjunto de casos aplicáveis e os atributos requeridos nesta fase, encapsulando a conversão de tuplos e atributos, distribuídos por várias tabelas da base de casos, em objectos candidatos do tipo *CasosPréSelecção*.

A subclasse utiliza os seus métodos específicos e também os comuns e partilhados, concedidos pela (super) classe *PGeralBaseK*. A classe *CasoCbr* simboliza cada caso candidato e o seu comportamento, possibilitando a sua manipulação ao nível individual.

A recuperação de casos é concluída pela classe *CompararSelFinal*, a qual efectiva as sub-tarefas comparar e seleccionar, recorrendo, respectivamente, a funções de similaridade e a um limiar de similitude. Este limiar pode ser definido pelo utilizador, durante a descrição de requisitos. O recurso ao padrão de desenho comando (encapsulamento de comandos como objectos) [Gamma et al. 95] assume um papel proeminente, durante a sub-tarefa comparar, permitindo que a actividade de recuperação permaneça genérica e extensível, já que, para além de já ser independente dos atributos dos casos, torna-se também independente das funções de similaridade. Nomeadamente, as classes abstractas *FuncSimLocal* e *FuncSimGlobal* definem uma interface para a execução de uma função de semelhança concreta. As diferentes funções de similitude a aplicar são obtidas sob a forma de objectos e por intermédio da classe *PropsRR*, evitando instruções condicionais complexas, com ramificações numerosas, que conduziriam a dependência do código em relação às medidas de similaridade em uso.

Segue-se o passo de reutilização dos casos mais similares ao problema alvo. As sub-tarefas avaliar e organizar são, fundamentalmente, asseguradas pelas classes *MédiasCritérios* e *ItensOrdenação*. Estas sub-tarefas são extensíveis, uma vez que os critérios de avaliação podem ser alterados (acrescentados e removidos), sem afectar o seu funcionamento. A classe *MédiasCritérios* determina valores médios dos critérios de avaliação, atendendo ao tipo de solução dos casos candidatos, e considerando um conjunto de atributos com tamanho variável, gerado via serviços da classe *PropsRR*. A classe *ItensOrdenação* implementa a interface *comparable*, proporcionando uma lógica de ordenação dinâmica, baseada no nível de similaridade e na importância relativa atribuída aos critérios, tendo em conta, novamente, uma série variável de atributos. Seguidamente, a classe *ProduzirPlanos* constrói os planos de WUM e obtém informação geral acerca dos casos que vão instanciar cada plano, por meio de serviços da classe *PKCasosRR*.

Finalmente, os planos de WUM construídos são apresentados pelo SPM ao utilizador, juntamente com ligações para aceder informação detalhada dos casos. A lógica subjacente a esta funcionalidade foi encapsulada num componente *(JavaBean)*, implementado pela classe *DetalhesCaso*. Esta classe usa, igualmente, os serviços de persistência da classe *PKCasosRR*, para obter uma descrição detalhada de um determinado caso, disponibilizando essa informação para consulta.

5.2.2 Aprendizagem

A função de aprendizagem é conduzida em múltiplas iterações de consubstanciação progressiva do conteúdo dos casos. Nesta secção abordam-se apenas as questões mais significativas da vertente inicial e mais preponderante desta função, respeitante à aquisição de informação geral e etapas de um novo processo. Por conseguinte, omitem-se os aspectos da implementação da recolha de itens de conhecimento prévio e derivado (e.g. factos e descobertas) e de informação suplementar (e.g. fontes e autores), os quais são garantidos por serviços de domínio e suas classes de persistência. Também não se descrevem nesta secção as operações auxiliares de alteração, adição e remoção de etapas, entre outras opções facultadas por intermédio desta função.

A Tabela 10 estabelece a correspondência entre as sub-tarefas da vertente mencionada da aprendizagem e as principais classes que as implementam. A descrição da implementação desta vertente foi subdividida em duas fases. A primeira fase, designada descrição de processos, é reportada no diagrama de classes da Figura 29. Esta fase cobre a sub-tarefa combinar descrições, contemplando, essencialmente, serviços de persistência e interacção. O primeiro tipo de serviços permite obter as instâncias das muitas tabelas que sustentam a descrição de processos. A classe *InterfaceDescProcesso* desencadeia a realização desses serviços, executados através da classe *DadosCasos*. Esta última classe consolida e disponibiliza os resultados das consultas à base de conhecimento. Tais consultas são, efectivamente, efectuadas por subclasses de persistência, respeitantes à classe *CaractDados* e a diversas outras da biblioteca de *domínio*. Quanto ao segundo tipo de serviços, durante a interacção com o utilizador, são requeridas capacidades sofisticadas para assistir e colectar a especificação de itens variados. Designadamente, é imprescindível gerir o povoamento de estruturas de dados, ao nível da sessão, referentes a todos os dados fornecidos, nos vários tipos de páginas Web interligadas. Adicionalmente, torna-se necessário simular alguns efeitos, como a persistência de alguns destes itens (e.g. variáveis acrescentadas em etapas de transformação), de forma a possibilitar a sua disponibilização imediata, para uso na especificação de etapas subsequentes da mesma sessão. Neste sentido, as capacidades mais complexas foram delegadas em componentes *(JavaBeans)*, representados pelas seguintes classes:

- *DescFichsPmml* – lida com a enumeração de documentos PMML.
- *InfEtapaIds* – trata a descrição de cada etapa via interacção explícita.
- *DescInfCaso* – gere a sequenciação de etapas e o registo de toda a informação acerca das mesmas, incluindo o tipo de fonte e os seus dados.

Tabela 10 – Correspondência entre sub-tarefas e classes da aprendizagem

Tarefa	Sub-tarefas	Principais classes de implementação das sub-tarefas
Conciliar descrições	**Combinar** descrições	*InterfaceDescProcesso, DadosCasos, DescInfCaso, DescFichsPmml, InfEtapaIds*
	Transformar conteúdo	*InfEtapaDgs* e classes da biblioteca acessopmml
Reter	**Integrar** elementos	*Reter, IReterEtapas, ReterEtapasPmml, ReterEtapasInteracção, InfCaso*
	Estruturar elementos	*PInfCaso, PInfCasoVars*
	Registar caso	Algumas classes de domínio e outro tipo de serviços de persistência

A classe *FonteReter* e as suas subclasses *FontePmml* e *FonteInteracção* participam na implementação do padrão de desenho comando, simplificando a posterior invocação de métodos da classe adequada, durante o tratamento de cada etapa. A interface *IReterEtapas* define a assinatura do método responsável pela execução desse tratamento.

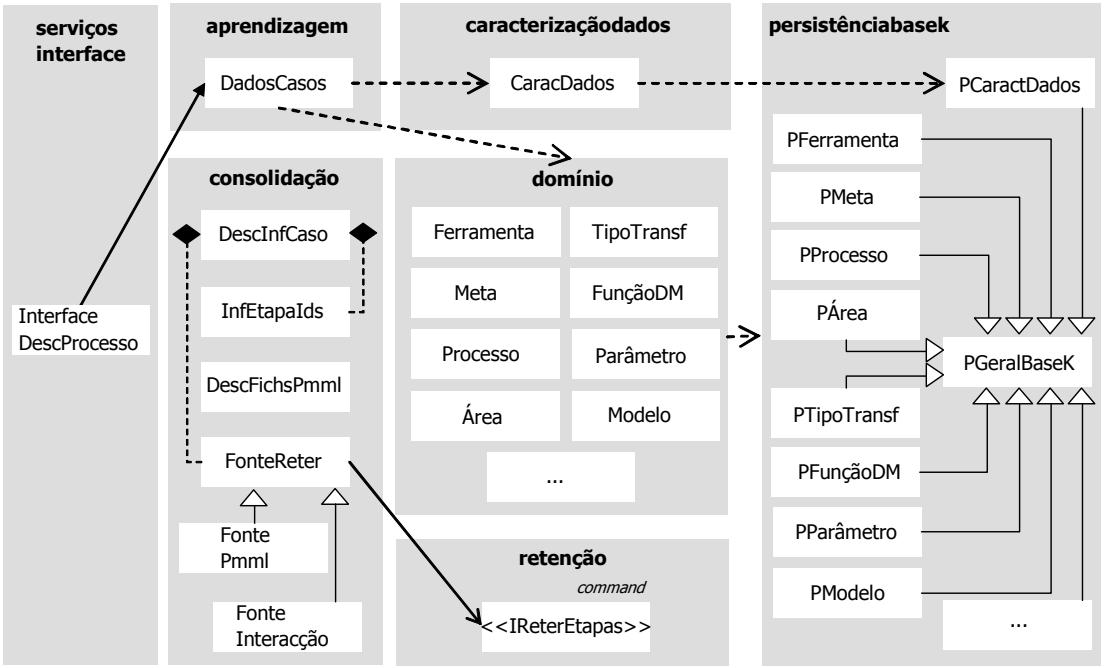


Figura 29 – Diagrama de classes da descrição de processos

Concluída a fase da descrição do processo, pode iniciar-se o seu processamento, encetando a segunda fase da aprendizagem, ilustrada no diagrama de classes da Figura 30. Antes de se passar à explicação desta fase, importa clarificar alguns dos aspectos que sustentam a sua concretização.

A descrição das etapas de modelação e transformação é facultada por um objecto do tipo *DescInfCaso*. Um objecto deste tipo é constituído por uma lista de elementos, contendo, cada elemento, um objecto do tipo *InfEtapaIds* ou uma série de referências para documentos PMML e, também, a indicação do tipo de fonte subjacente. A classe *InfEtapaIds* define as propriedades e métodos de uma etapa descrita por interacção explícita, denotando o facto do seu conteúdo se basear em identificadores ou códigos (maioritariamente valores numéricos). Pelo contrário, o conteúdo que vai ser extraído de documentos PMML, durante a persecução da sub-tarefa transformar, envolve designações (sobretudo texto), exigindo um tipo de manipulação distinto, o qual é proporcionado pela classe *InfEtapaDgs*. A manipulação deste conteúdo abrange também duas actividades acrescidas. A primeira é a validação de cada etapa, enquadrada no âmbito de uma sequência, a qual só poderá ser efectivada após a extracção do conteúdo PMML. A segunda prende-se com a eventualidade de estas designações poderem originar novas instâncias nas respectivas tabelas relacionais, obrigando à verificação exaustiva da sua existência.

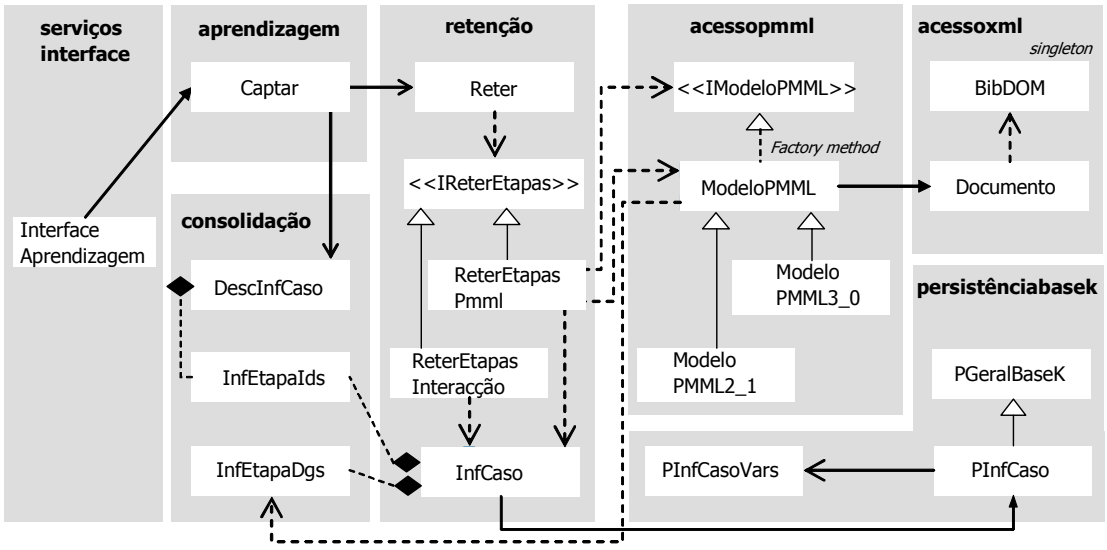


Figura 30 – Diagrama de classes do processamento da descrição de processos

A segunda fase da aprendizagem é desencadeada pela classe *InterfaceAprendizagem* e conduzida pela classe *Captar*. A primeira classe invoca uma das operações disponíveis, enquanto a segunda prepara e encapsula essas operações, interligando as tarefas de consolidação e retenção. Entre as operações disponíveis, incluem-se duas com tratamento afim: criação de um novo processo e

adição de etapas a um processo já existente. O tratamento destas duas operações é, efectivamente, gerido pela classe *Reter*, a qual itera a sequência de etapas, atendendo a várias das condicionantes já debatidas. Conforme já foi referido, a interface *IReterEtapas* concede um método genérico de tratamento de etapas. Este método é implementado, de forma específica, pelas subclasses *ReterEtapasPmml* e *ReterEtapasInteracção*, em consonância com o tipo de fonte. A classe *InfCaso* é usada para registar e representar o resultado de todo o processamento integrador da descrição submetida.

O tratamento de documentos PMML é, necessariamente, mais complicado, sendo executado intercaladamente com a sub-tarefa de integração. Na realidade, apesar das sub-tarefas combinar descrições e transformar conteúdo continuarem a contribuir, conjuntamente, para um desígnio convergente de conciliação, a concretização da sub-tarefa de transformação nestes moldes é, substancialmente, mais eficiente e conveniente, principalmente no que concerne à validação dos seus resultados, no âmbito de uma série de etapas consecutivas. Cada etapa deste tipo é extraída para um novo objecto do tipo *InfEtapaDgs*, integrado e mantido num objecto *InfCaso*. Um objecto *InfCaso*, representa, portanto, todo o conteúdo da descrição de um processo de WUM.

Em termos práticos, a sub-tarefa de transformação de conteúdo é levada a cabo pelas classes da biblioteca *acessopmml*, com o apoio das classes da biblioteca *acessoxml*, sob a orientação da classe *ReterEtapasPmml*, cuja actuação é, ainda, controlada pela classe *Reter*. Para efeitos de interpretação do conteúdo PMML, considerou-se apenas duas versões da especificação (2.1 e 3.0), nomeadamente, as mais recentes, no momento da implementação (actualmente já existe a versão 3.1). Pretendia-se, essencialmente, comprovar a possibilidade de estender facilmente a biblioteca, para incorporar versões futuras do padrão. Com este propósito em vista, recorreu-se à tradicional interface, para estabelecer a assinatura de um método genérico de extracção e, ainda, a outras soluções de desenho comuns. No presente contexto, a solução essencial é o encapsulamento da funcionalidade requerida para seleccionar e instanciar a classe apropriada, dentro de um método, referido por método de fabrico (*factory method*) [Gamma et al. 95], dado que, tal decisão, depende da versão do documento PMML. A classe *ReterEtapasPmml* instancia um objecto *ModeloPMML*. A (super) classe *ModeloPMML*, por sua vez, cria um objecto do tipo *Documento*, para cada documento PMML submetido, e, após a extracção da versão do mesmo, instancia a subclasse apropriada no método de fabrico, devolvendo um objecto de um tipo genérico. Por conseguinte, podem ser acrescentadas novas versões da especificação à biblioteca, através de subclasses adicionais, sem afectar as restantes classes. A classe *ModeloPMML* centraliza, ainda, serviços comuns, partilhados pelas suas subclasses *ModeloPMML_2_1* e *ModeloPMML_3_0*.

O processamento subjacente de XML é assegurado pela biblioteca *acessoxml*, com base na especificação e API DOM, a qual define uma forma normalizada para aceder e manipular documentos estruturados. Neste modelo de processamento, o documento XML inteiro é armazenado em memória, numa estrutura em árvore, permitindo a navegação e alteração dos seus nós constituintes. Esta abordagem de processamento possibilita um tratamento flexível e versátil, apesar das desvantagens que lhe são apontadas, no que respeita a complexidade e consumo de recursos. Recorre-se, também, à API JAXP, incluída como parte da versão da plataforma J2SE usada (biblioteca *javax.xml.parsers*). Esta API veio uniformizar a criação de instâncias de processadores e analisadores sintácticos (*parsers*) de XML em aplicações Java, sendo conforme com várias API padrão (e.g. DOM) e garantindo a independência dessas aplicações, relativamente a implementações concretas dos mesmos, como, por exemplo, o *parser Xerces* (*Xerces-J* versão 2.6.2 [WWW11]) utilizado.

O tratamento de documentos XML foi repartido por duas classes fundamentais. A classe *BibDOM* gere e encapsula o acesso a ficheiros XML e a construção de árvores DOM, facultando os resultados dessas actividades, designadamente, um objecto (do tipo *Document*) que, conceptualmente, corresponde à raiz da árvore DOM, montada a partir de cada documento PMML processado. A classe cria um objecto *BibDOM*, único na aplicação, que controla a sincronização no acesso a ficheiros, implementando o padrão de desenho *singleton* (garantia de instância única de uma classe e fornecimento de uma forma de acesso global e único à mesma). Além disso, usa as classes da biblioteca *javax.xml.parsers* para gerar uma instância da classe DOM *Document*, obtendo, desta forma, um objecto que implementa a interface *Document,* definida na biblioteca *org.w3c.dom*. Já a classe *Documento* desenvolvida acrescenta outro nível de encapsulamento, implementando métodos genéricos de procura e extracção de partes de um documento. Este documento é obtido via serviços de *BibDOM*, estando internamente disponível sob a forma de um objecto DOM *Document*. Estes métodos oferecem muitas formas de pesquisa de partes do documento (elementos, seu texto e atributos), restituindo os itens encontrados via tipos de dados básicos ou colecções. Deste modo, a biblioteca *acessopmml* foi isolada dos detalhes do processamento e acesso a partes de documentos XML. As suas classes somente lidam com objectos mais intuitivos, do tipo *Documento*, estabelecendo os critérios de procura e extracção de itens, a partir dos mesmos, com base nos métodos públicos da classe *Documento*.

Seguem-se as sub-tarefas estruturar elementos e registar caso. As classes *PInfCaso* e *PInfCasoVars* levam a efeito a estruturação do conteúdo de *InfCaso*, de acordo com o esquema interno de representação de casos. A classe *PInfCasoVars* lida, particularmente, com os esquemas

de mineração e transformação, verificando e tratando o surgimento de novas variáveis. Esta estruturação permite preparar a sub-tarefa de registo efectivo das novas instâncias nas várias tabelas relacionais. Idealmente, estas actividades deveriam ser conduzidas, em função das diversas classes de domínio e respectivas classes de persistência. Esta abordagem chegou a ser parcialmente explorada, mas revelou-se inadequada. Ao invés, uma parte substancial das operações foi delegada em procedimentos armazenados, apesar da desvantagem da falta de portabilidade dos mesmos. A biblioteca de *domínio* mantém o seu papel de isolamento do modelo de domínio, relativamente aos métodos responsáveis pelos passos de raciocínio, contemplando cerca de uma quinzena de classes distintas, as quais realizam variadas operações. No entanto, as operações mais exigentes (i.e. mais intensivas ou envolvendo a manipulação de um número considerável de tabelas) são asseguradas por intermédio de procedimentos armazenados. Esta situação verifica-se, especialmente na criação de um novo processo e, também, na adição de etapas a um processo existente e, consequentemente, a sub-tarefa registar caso é concretizada por uma série de procedimentos armazenados, sendo estes executados no âmbito de uma transacção.

5.3 Utilização do Sistema

Os recursos usados na implementação do sistema SPM oferecem capacidades para sustentar o desenvolvimento de uma ferramenta de assistência bastante funcional. Porém, as potencialidades de tais recursos não foram plenamente exploradas. Por um lado, as actividades de desenvolvimento levadas a cabo nortearam-se pela finalidade de proporcionar um protótipo, correntemente, sem ambições de uso real, visando, sobretudo, materializar as ideias defendidas que fundamentam o sistema proposto. Por outro lado, privilegiaram-se as vertentes de funcionalidade e robustez do funcionamento de algumas operações, focando a eficácia do desenho na interacção com o utilizador e a acessibilidade ao longo da organização, em detrimento de uma apresentação mais cuidada e apelativa. Procurou-se, ainda, não sobrecarregar demasiado as páginas Web, interligando séries de páginas para viabilizar a persecução de algumas operações, nomeadamente, aquelas que abrangem uma gama significativa de elementos. Mesmo assim, num cenário de utilização real do sistema, o desenho das páginas teria de ser melhorado, pois algumas continuam a comportar conteúdo excessivo.

A Figura 31 apresenta a barra de navegação, incluída em todas as páginas Web do protótipo, cobrindo as funcionalidades essenciais do sistema SPM. Estas funcionalidades foram organizadas

em dois grupos. O primeiro grupo (Itens Auxiliares) contém as várias categorias dos elementos base do domínio, facultando acesso a operações de manipulação dos mesmos, através das respectivas ligações. O segundo grupo (Operações Principais) engloba as funcionalidades mais relevantes do sistema. A descrição de contextos e de conjuntos de dados foram implementadas como operações autónomas, correspondendo a última ao módulo de caracterização de dados. Apesar da caracterização de dados pertencer à resolução de problemas, na prática, a segunda pode basear-se num problema ou conjunto de dados já existente, sendo mais conveniente oferecer a primeira como uma operação independente. As duas funções mais proeminentes do sistema, a resolução de problemas, propriamente dita, e aprendizagem, encontram-se disponíveis, respectivamente, por intermédio de duas ligações:

- Resolver problema, representando os módulos de construção de problemas, recuperação e reutilização, nos moldes que se acabou de referir.
- Reter processo, simbolizando os módulos de consolidação de descrições e retenção.

A funcionalidade de administração de casos é devotada ao administrador do sistema e, presentemente, concede acesso a operações mais restritivas, tais como (algumas das) alterações de estado (e.g. protegido e obsoleto) e remoção de casos. As operações adicionais de manipulação de contextos, conjuntos de dados e processos (casos) são fornecidas por meio de ligações incorporadas nas páginas subsequentes.

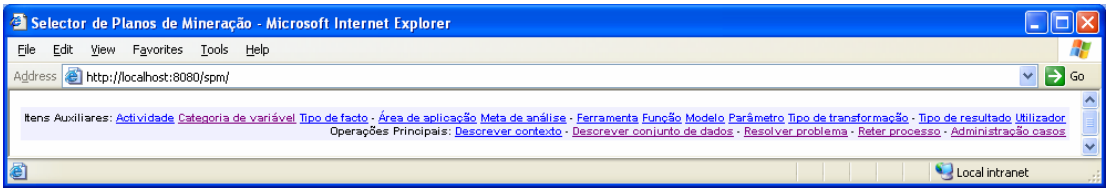


Figura 31 – Barra de navegação do protótipo do sistema

A utilização do sistema SPM foi prevista com base em três cenários primordiais: exploratório; assistência; administração. O cenário exploratório prende-se com o uso do sistema para obter indicações acerca de aspectos de interesse, tipicamente, através de uma descrição incompleta do problema. Este cenário enquadra-se num estado preliminar do processo, em que a informação existente pode ser usada para guiar o utilizador. No âmbito da WUM, a formulação de um problema, mesmo em termos de um conjunto de dados e requisitos da análise, possui dificuldade

acrescida. A referência a experiências passadas é esclarecedora quando a situação em causa não é simples nem está completamente compreendida. Os casos, ao retractarem experiências bem sucedidas de integração de diversas abstracções, podem apoiar decisões, pela apresentação de exemplos de perspectivas adoptadas por especialistas no domínio. Este cenário de utilização é iniciado pela navegação ao longo das opções de itens auxiliares, para consultar os seus elementos constituintes. Segue-se a execução da operação resolver problema, tendo uma ideia, usualmente pouco definida, do problema actual, a fim de obter resposta para tipos de questões como os seguintes:

- Características do conjunto de dados necessário, tais como as categorias de dados a usar (*clickstream*, transaccionais ou combinação de ambos), os atributos adequados e disponíveis a eleger e os critérios de selecção de registos a adoptar.
- Operações de transformação envolvidas, passíveis de reaplicação.
- Critérios de avaliação importantes para um determinado problema, contemplando a interpretação do significado dos mesmos e a escolha ajustada daqueles que devem ser aplicados para definir o problema actual.
- Metas de análise e áreas de aplicação, com o intuito de identificar o contexto, intenções e objectivos subjacentes na caracterização de análises relacionadas, para assim melhorar a especificação subsequente do problema corrente.

Em contraste, o cenário de assistência pressupõe o conhecimento do problema actual, assim como dos dados relevantes para a análise, das metas, áreas de aplicação e critérios de avaliação que devem orientar o processo, e a sua submissão ao sistema usando uma descrição focada, no sentido de obter respostas mais selectivas. A noção mais precisa da situação em causa e, também, do seu relacionamento com as abstracções em uso, permite discriminar melhor os descritores de problemas mais significativos e, mesmo, os níveis de preponderância das suas contribuições individuais. A flexibilidade no suporte à formulação do problema e na procura de soluções, com base no grau de semelhança, relativamente à situação actual, é proeminente, em particular, quando não foi encontrado nenhum caso coincidente. Não obstante, a combinação da aplicação de critérios de filtragem exacta também pode ser de grande utilidade. Este tipo de filtragem permite excluir do resultado os casos referentes a circunstâncias denotadas, à partida, como irrelevantes. Este cenário decorre de acordo com o ciclo CBR adoptado, englobando todos os seus passos, apesar de só se ter mencionado os aspectos directamente associados à resolução de problemas.

Os dois cenários anteriores de utilização do sistema são especialmente devotados a utilizadores sem experiência no domínio da WUM. O modo de administração, pelo contrário, requer indivíduos

com conhecimentos e experiência nesta área, detentores de uma visão global da base de conhecimento e dotados de capacidades para discernir quais casos são pertinentes, entre os vários registados, fruto das contribuições individuais dos colaboradores da organização. O perfil de administração do sistema tem sob a sua alçada a gestão da base de conhecimento, de forma a esta se manter actualizada, coerente e relevante, em consonância com a perspectiva específica do organismo. Para o cumprimento desta missão, torna-se conveniente, antes de mais, adoptar uma orientação fundamentada no estatuto alcançado pelos casos, reflectido pelo seu estado. Para além da consulta de casos, previamente organizados pelo seu estado, e da manipulação dos diversos itens auxiliares, as operações mais significativas neste cenário, de acordo com a implementação corrente, são a alteração do estado de casos e a remoção de casos.

5.3.1 Exemplificação da Resolução de Problemas

Para demonstrar a vertente de resolução de problemas do sistema SPM considerou-se um exemplo de um problema típico de WUM. Em termos genéricos, este problema centra-se em conhecer a forma através da qual os visitantes estão a usar o sítio Web, com o propósito de melhorar a conveniência da navegação ao longo deste. Mais propriamente, a acção planeada reside em acrescentar ligações entre algumas páginas Web, para agilizar o acesso às mesmas. O analista pretende determinar quais são as melhores páginas Web para incluir nas tais ligações e em que páginas essas ligações devem ser adicionadas. Os critérios a ter em conta nessa decisão são o nível de utilidade e importância das páginas, do ponto de vista da generalidade dos visitantes do sítio Web.

Adoptando a perspectiva do cenário exploratório, uma das atitudes possíveis é a consulta de metas de análise e de áreas de aplicação registadas no sistema. De momento, a base de casos contém um número limitado de instâncias. Este facto e a corrente intenção da sua generalização restringem a diversidade das metas de análise e áreas de aplicação existentes. Nomeadamente, a hierarquia de áreas de aplicação possui, presentemente, apenas dois níveis, com três áreas de topo, englobando, cada uma, duas sub-áreas, conforme se ilustra na Figura 32. Por exemplo, a área de topo de qualidade de serviço foca acções de melhoramento de sistemas e do sítio Web, envolvendo expectativas básicas e afectando todos os visitantes. O nível de afinidade entre as instâncias desta hierarquia é traduzido numa matriz de similaridade assimétrica, para que possam ser definidos valores mais adequados entre itens de níveis diferentes e em função do sentido da comparação.

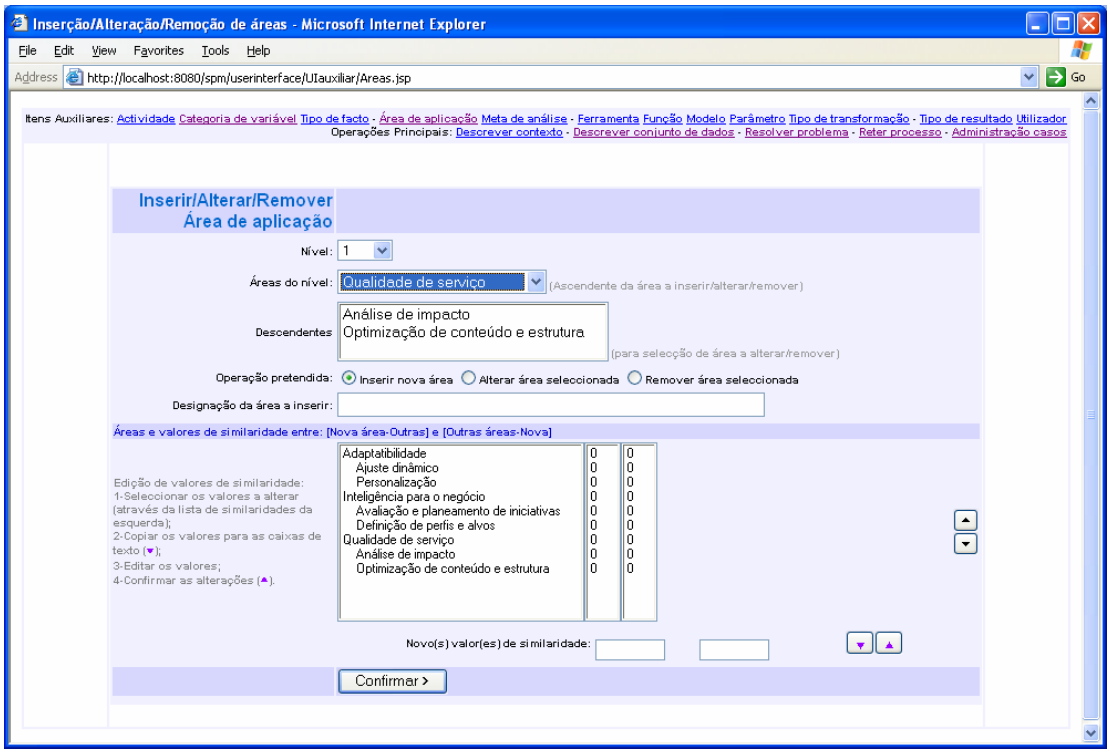


Figura 32 – Operações de manipulação de áreas de aplicação

A consulta de metas de análise e de áreas de aplicação poderá dar lugar ao uso exploratório da operação de resolução de problemas, com vista a aceder aos pormenores dos casos associados a um subconjunto plausível destas abstracções. O primeiro passo neste sentido consiste em definir o ponto de partida da descrição de requisitos, de acordo com o conteúdo da página apresentada na Figura 33. Escolheu-se a primeira opção, pois esta é a mais indicada para conduzir uma pesquisa baseada somente num subconjunto de metas de análise e de áreas de aplicação. Alternativamente, a utilização exploratória do sistema poderia ser orientada com base num conjunto de dados alvo e, ainda, conjugando aspectos relativos a restrições explícitas ou mesmo recorrendo a um problema de um caso existente, para apoiar a especificação de todos os requisitos. Nestas circunstâncias as restantes opções da definição do ponto de partida da descrição de requisitos seriam mais convenientes.

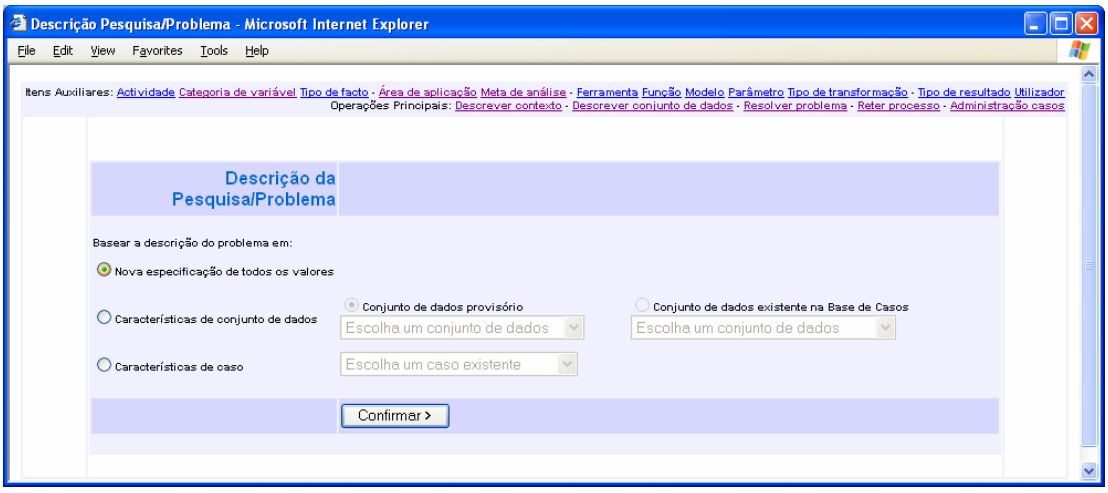


Figura 33 – Especificação do ponto de partida da descrição de requisitos

A Figura 34 exemplifica uma descrição de requisitos, atendendo aos elementos disponíveis, até ao momento, acerca do problema alvo exemplo. Seleccionou-se as duas sub-areas da qualidade de serviço, uma vez que são ambas relevantes e as mais próximas das acções intencionadas. Quanto a metas de análise, usou-se todas as que podem proporcionar informação sobre relacionamentos entre páginas. Outras metas existentes são:

- identificar e caracterizar diferentes tipos de visitas e visitantes;
- distinguir visitas em função de eventos alvo;
- determinar relações entre eventos de sessões do mesmo visitante.

Aplicou-se ainda a funcionalidade de filtragem exacta aos dois descritores, a fim de restringir a procura às abstracções mencionadas e, também, de exemplificar o uso desta funcionalidade.

O conteúdo da página de descrição de requisitos é, em grande parte, construído dinamicamente. Os atributos descritores de problemas foram organizados em várias categorias mais estáveis. Não se espera alterar estas categorias, mas sim, eventualmente, os seus elementos constituintes ou propriedades, maioritariamente facultados através do preenchimento dinâmico de controlos de interacção. O único destes elementos cujos requisitos são completamente especificados nesta página é a data do processo. Todos os restantes elementos são definidos em janelas próprias, acessíveis a partir das hiperligações localizadas na parte esquerda da página, conforme se verá mais adiante.

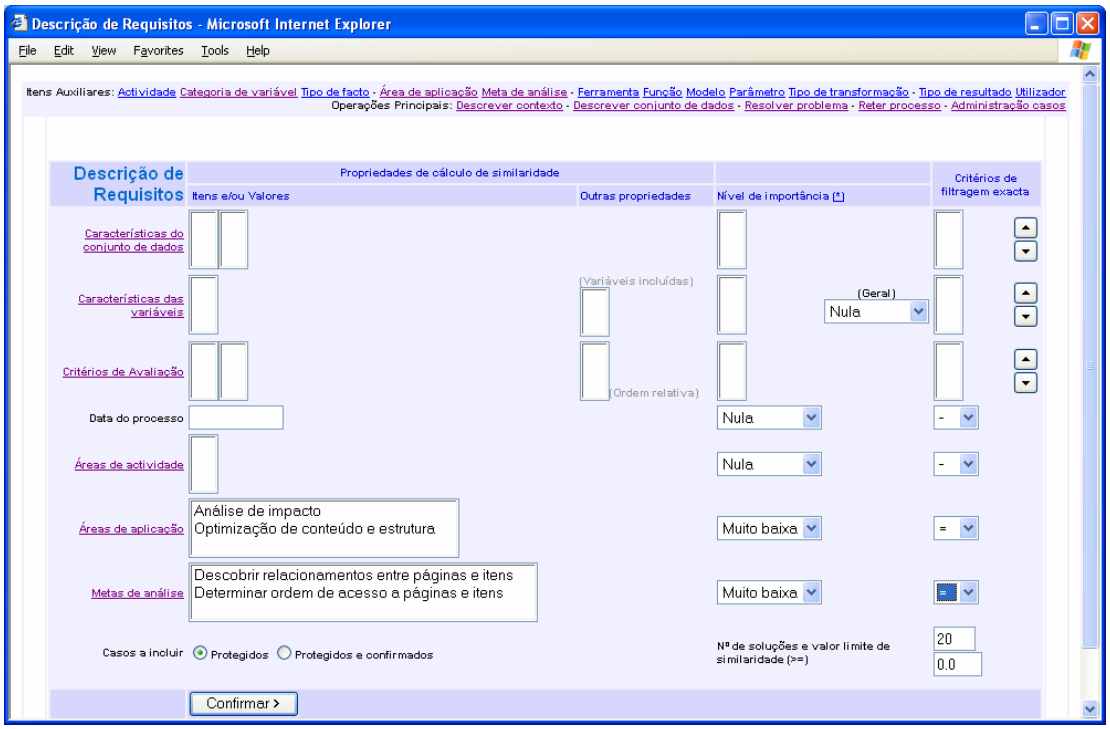


Figura 34 – Exemplo de descrição de requisitos no cenário de utilização exploratório

A descrição de requisitos anterior foi usada para gerar os planos de WUM relacionados com as metas de análise e áreas de aplicação pretendidas. Os três planos produzidos, apresentados na coluna D da Figura 35, são simples, uma vez que apenas contemplam casos com uma única etapa. Estes planos permitem obter uma ideia dos métodos de DM aplicáveis ao problema alvo, inclusive dos valores médios dos critérios de avaliação (coluna C), respeitantes a todos os casos recuperados. No entanto, os aspectos mais informativos da solução prendem-se com os sucessivos detalhes que podem ser consultados por intermédio deste resultado, em particular, acedendo:

- aos casos seleccionados (mais similares) que instanciam os planos, seguindo as ligações da coluna A, cuja similitude com o alvo é mostrada na coluna B;
- aos restantes casos recuperados (menos similares) de cada categoria de modelo, elegendo um caso entre as opções das caixas de combinação da coluna A, as quais reportam o número do caso e o nível de semelhança com o alvo.

Uma constatação a assinalar, observando especialmente a caixa de combinação expandida (em E), é o nível de similaridade, que tende a ser idêntico para todos os casos da mesma categoria de

modelo, e mesmo para casos de diferentes categorias. Esta situação deve-se ao facto de o nível de semelhança ter sido aferido apenas em função das metas de análise e áreas de aplicação. A resolução de problemas pode ser melhorada fornecendo mais requisitos, tal como se explica seguidamente, no âmbito do cenário de assistência.

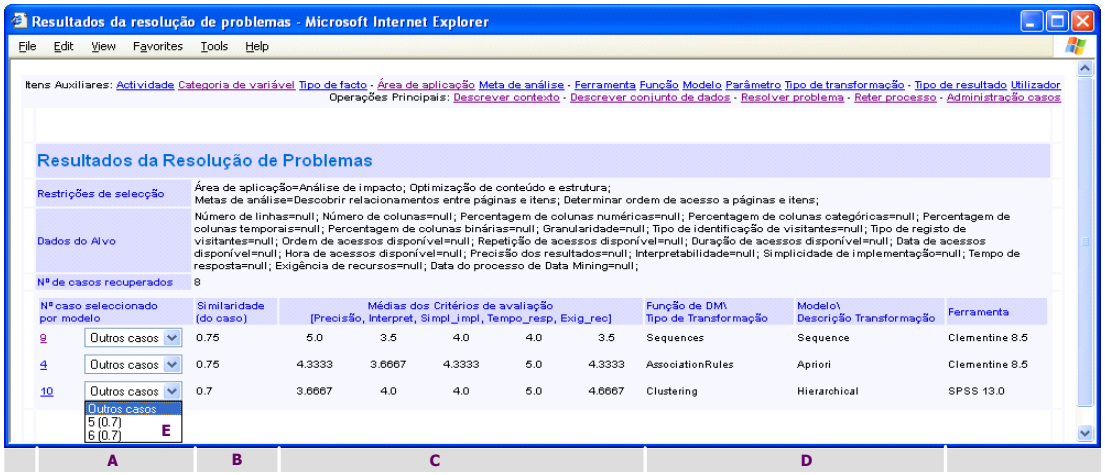


Figura 35 – Exemplo do resultado da resolução de problemas no cenário exploratório

A primeira fase do cenário de assistência corresponde à caracterização de dados, pressupondo a situação mais típica da existência de um conjunto de dados alvo, no qual se pode basear a descrição de requisitos. O conjunto de dados deste exemplo consiste num ficheiro de histórico de um servidor Web, contendo informação (8 variáveis) ao nível do acesso a páginas (granularidade). O conteúdo deste ficheiro foi, previamente, tratado de forma básica, praticamente apenas para individualizar os campos dos seus registos e para agrupar as páginas visitadas em categorias de páginas (variável URL). A Figura 36 reporta o resultado desta fase, após a edição das categorias de algumas variáveis, a saber, IP e URL, respectivamente como *IDsessão* e *Pageview*. O único aspecto que ainda não foi referido neste documento prende-se com a coluna do número de variáveis representadas (*NºVars repres*). Existem conjuntos de dados com variáveis da ordem das centenas, entre as quais, muitas com propriedades idênticas. Os exemplos mais flagrantes são matrizes binárias de paginas X sessões. Com o intuito de optimizar o tratamento destas situações, optou-se por agrupar variáveis com propriedades idênticas em variáveis representativas, sendo necessário preservar o número de elementos de cada grupo para não se perder informação.

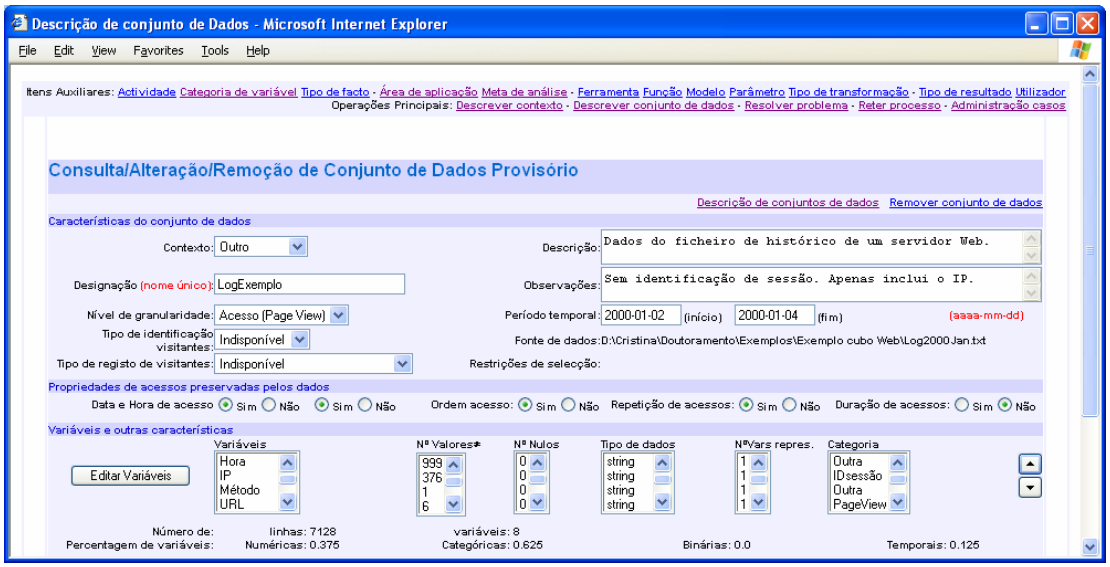


Figura 36 – Exemplo da caracterização de dados

Concluída a caracterização de dados, pode encetar-se a resolução de problemas propriamente dita, recorrendo agora à opção de basear a descrição de requisitos nas características de um conjunto de dados (segunda opção da Figura 33). A Figura 37 ilustra a nova descrição de requisitos. Começou-se por alterar a importância do descritor granularidade, de muito baixa para média. Pretende-se exemplificar o uso desta funcionalidade, dando ênfase a um aspecto pertinente, mas sem introduzir efeitos muito significativos no resultado deste exemplo concreto. Ainda em termos de características de dados, seleccionou-se as variáveis IP e URL, como as mais relevantes, significando que apenas estas duas variáveis serão incluídas na comparação com as variáveis usadas em etapas de modelação dos casos prévios. Quando a critérios de avaliação do sucesso do processo e dos seus resultados, atribuiu-se o valor máximo (5) a todos os critérios e a maior importância relativa à interpretabilidade, seguida pelos restantes critérios (Figura 38). No que concerne a metas de análise e a áreas de aplicação, usou-se as mesmas abstracções do anterior cenário exploratório, mas sem aplicar critérios de filtragem exacta. Por último, alterou-se a valor do limiar de similaridade para 0.5, com o intuito de excluir do resultado casos com um nível de semelhança baixo.

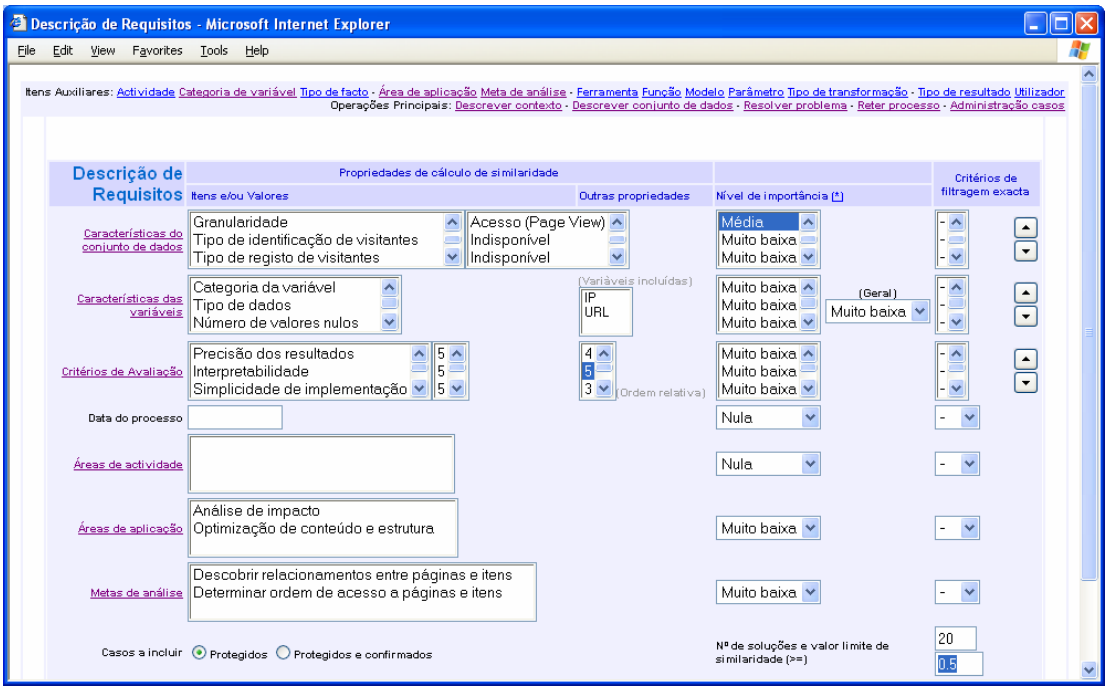


Figura 37 – Exemplo da descrição de requisitos no cenário de utilização de assistência

A Figura 38 mostra a edição de requisitos de critérios de avaliação, incluindo a especificação de valores, ordem de relevância relativa, níveis de importância e restrições de filtragem exacta.

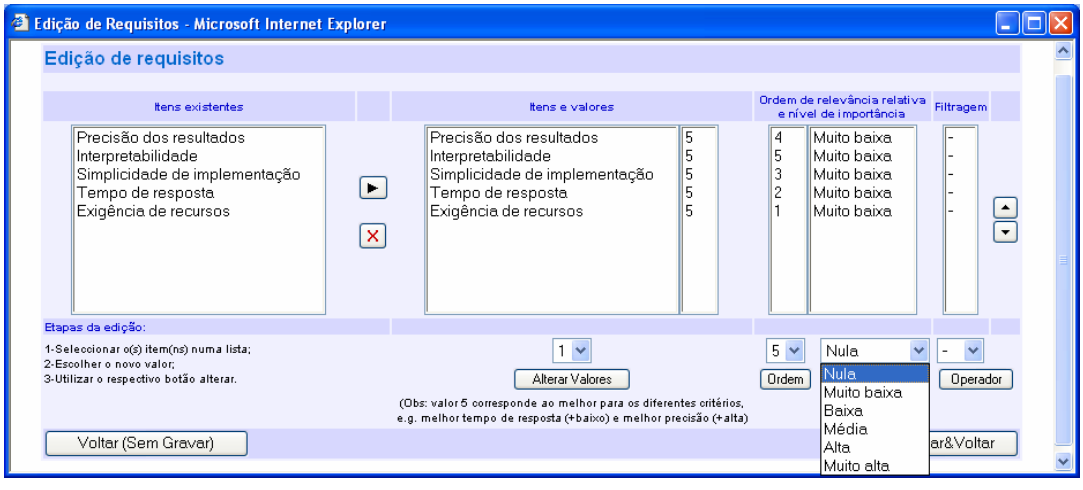


Figura 38 – Exemplo da edição de requisitos para critérios de avaliação

A Figura 39 apresenta a descrição de uma solução possível para o problema exemplo. Agora os planos já são verdadeiramente instanciados nos casos mais similares ao alvo e os níveis de similitude já distinguem os casos e, portanto, os seus pormenores tendem a ser mais ajustados ao problema em questão. Tal como anteriormente, todas as funções e modelos sugeridos são adequados para abordar o problema exemplo, mas a ordem de recomendação dos planos alterou-se. O modelo *Apriori* passou a ser recomendado como o mais apropriado. Esta sugestão pode ser considerada plausível. Na realidade, as regras de associação são um bom compromisso entre precisão e cobertura. Por uma lado, são mais informativas e precisas do que o método de agrupamento, pois facultam aspectos adicionais, como o antecedente e consequente das regras e os respectivos níveis de suporte e confiança. Por outro lado, as regras de associação, usualmente, concedem uma cobertura geral maior do que as técnicas de mineração de sequências, apesar de serem menos informativas do que tais técnicas, as quais podem proporcionar informação mais detalhada, como a ordem do acesso às páginas.

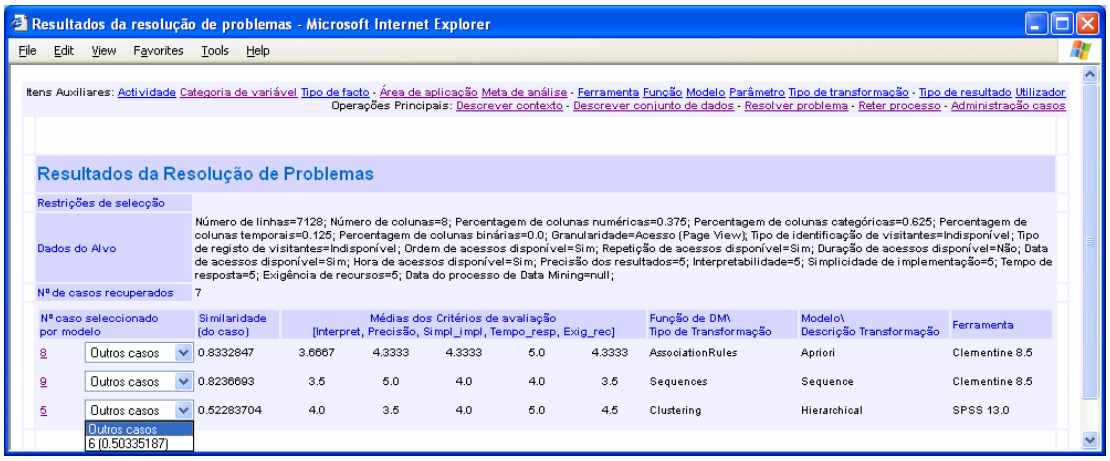


Figura 39 – Exemplo do resultado da resolução de problemas no cenário de assistência

A diferença no nível de similaridade entre os dois primeiros planos recomendados é pequena, significando que o modelo sequencial também deve ser explorado. Já a similitude dos casos do modelo de agrupamento hierárquico é substancialmente inferior. O sistema dá ênfase às características dos dados e os respectivos descritores estão em maioria, em relação a outros tipos de descritores. De facto, o modelo hierárquico foi aplicado a conjuntos de dados com propriedades

muito distintas, designadamente, matrizes binárias de páginas X sessões, que diferem na própria granularidade dos dados. A alteração do nível de importância do descritor de granularidade, durante a especificação de requisitos, também acentuou esta diferença. Sem esta alteração os valores de semelhança dos casos seleccionados seriam aproximadamente os seguintes: caso 8 - 0.803; caso 9 – 0.792; caso 5 – 0.621. Em contrapartida, as análises dos casos 8 e 9 foram realizadas usando conjuntos de dados mais similares ao alvo. De qualquer forma, entende-se que a integração do modelo hierárquico na solução final pode ser útil, pois é possível transformar os dados alvo no formato vulgarmente usado para aplicar este modelo.

Outro aspecto a assinalar é a disposição dos critérios de avaliação e dos seus valores médios para os sete casos recuperados. O critério de interpretabilidade passou a ser o primeiro, pois foi-lhe atribuída a maior importância relativa na descrição de requisitos. Além disso, usando a organização do resultado ilustrada o analista pode avaliar e consultar diversos tipos de soluções (planos) para o mesmo problema e, também, várias instâncias de um determinado tipo de solução. Por conseguinte, maximiza-se a utilidade das soluções encontradas para dois propósitos distintos: diversidade de tipos de soluções alternativas para um dado problema e variedade de instâncias para cada tipo de solução, com eventual interesse particular para o analista.

No que respeita a informação detalhada, seguindo, por exemplo, a ligação do caso 8 é possível aceder, sucessivamente, a todo o seu conteúdo, englobando a descrição dos seguintes aspectos, disponíveis via ligações, tal como se mostra na Figura 40:

- dados gerais do caso;
- etapas de modelação e transformação;
- resultados derivados no processo de WUM;
- propriedades do conjunto de dados alvo.

A pagina de informação detalhada integra ainda os principais elementos descritivos do exercício de WUM, entre os quais:

- a meta de análise e as áreas de aplicação;
- as funções de DM ou os tipos de transformação usados;
- os modelos de DM ou as descrições das operações de transformação aplicados;
- os parâmetros de configuração e o esquema de mineração de etapas de modelação;
- explicações e justificações das actividades e decisões envolvidas, sob a forma de observações.

Por último, conforme já foi dito, o Anexo A exemplifica o cálculo de similaridade entre dois problemas, reportando-se ao caso 8 e ao problema alvo discutido nesta secção.

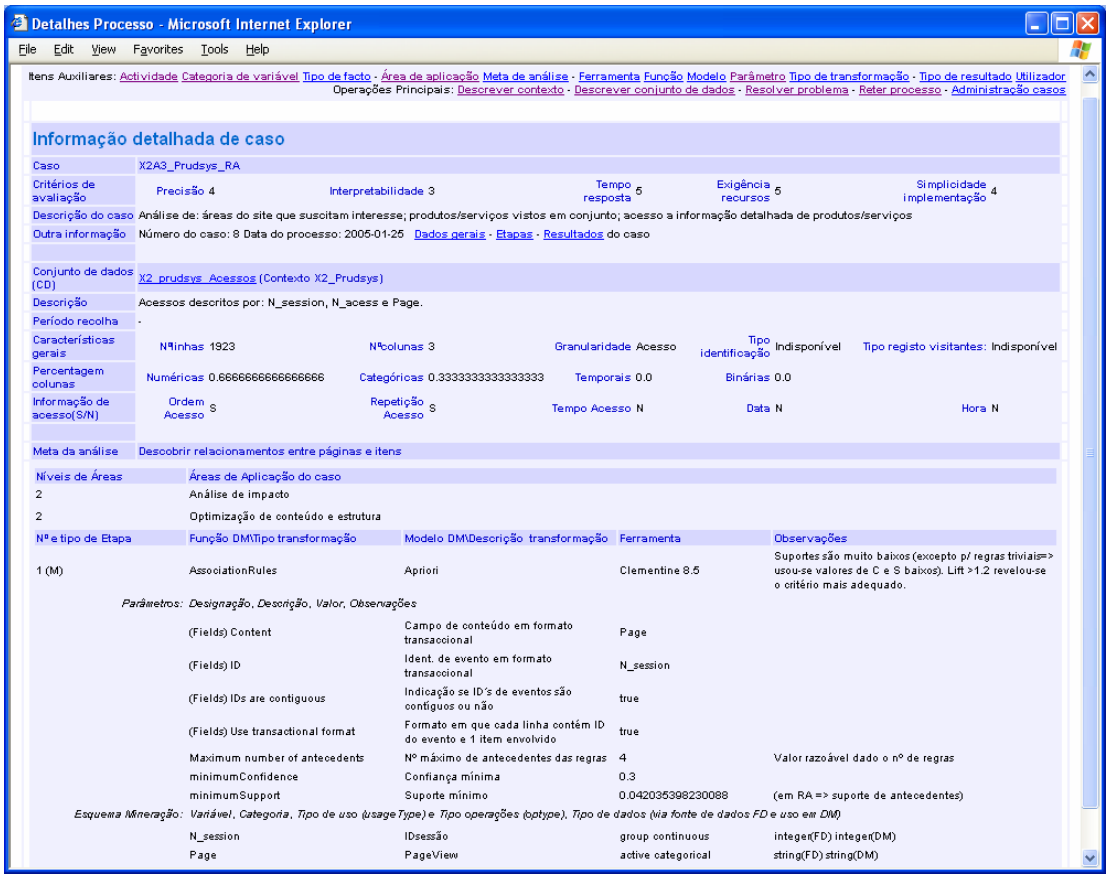


Figura 40 – Exemplo de acesso aos detalhes de um caso recuperado

5.3.2 Exemplificação da Aprendizagem

A vertente de aprendizagem do SPM permite registar a descrição de qualquer exercício de WUM entendido como útil. O seu único requisito é a caracterização prévia dos dados usados nesse estudo, reforçando, mais uma vez, a pertinência de proporcionar a descrição de conjuntos de dados como uma operação autónoma. Contudo, esta vertente do SPM enquadra-se, tipicamente, no cenário de assistência, dando continuidade à fase de resolução de problemas, após a revisão de um processo de WUM, como o discutido na secção anterior. Não obstante, nesta secção optou-se por recorrer a um exemplo baseado numa situação distinta, não só para conferir independência à secção, como também para aumentar a diversidade dos exemplos e dos aspectos abordados.

A exemplificação da funcionalidade de aprendizagem fundamenta-se num processo de WUM devotado ao estudo do fenómeno do abandono de "cartões" (ou "carrinhos") de compra, envolvendo, portanto, dados de *clickstream* relativos a um sítio Web dedicado à actividade de comércio electrónico. O conjunto de dados alvo já foi explorado noutras análises e consiste numa matriz binária, indicando, basicamente, se os visitantes acederam ou não às secções ou funcionalidades primordiais do sítio Web (8 variáveis), ao longo de várias sessões (185). A Figura 41 reporta a descrição do processo exemplo, no momento da sua submissão ao sistema, contemplando, entre outros, a enumeração dos seguintes tipos de aspectos:

- o conjunto de dados alvo, existente na base de casos, cuja selecção não pode ser alterada nesta sessão, em virtude de já terem sido descritos outros elementos com base neste;
- a meta de análise, uma área de aplicação e os valores dos critérios de avaliação atribuídos ao exercício de WUM;
- duas etapas, anteriormente descritas em janelas auxiliares (disponíveis através de botões de comando), e uma fonte do processo, a qual foi obtida e associada automaticamente, conforme se explica seguidamente.

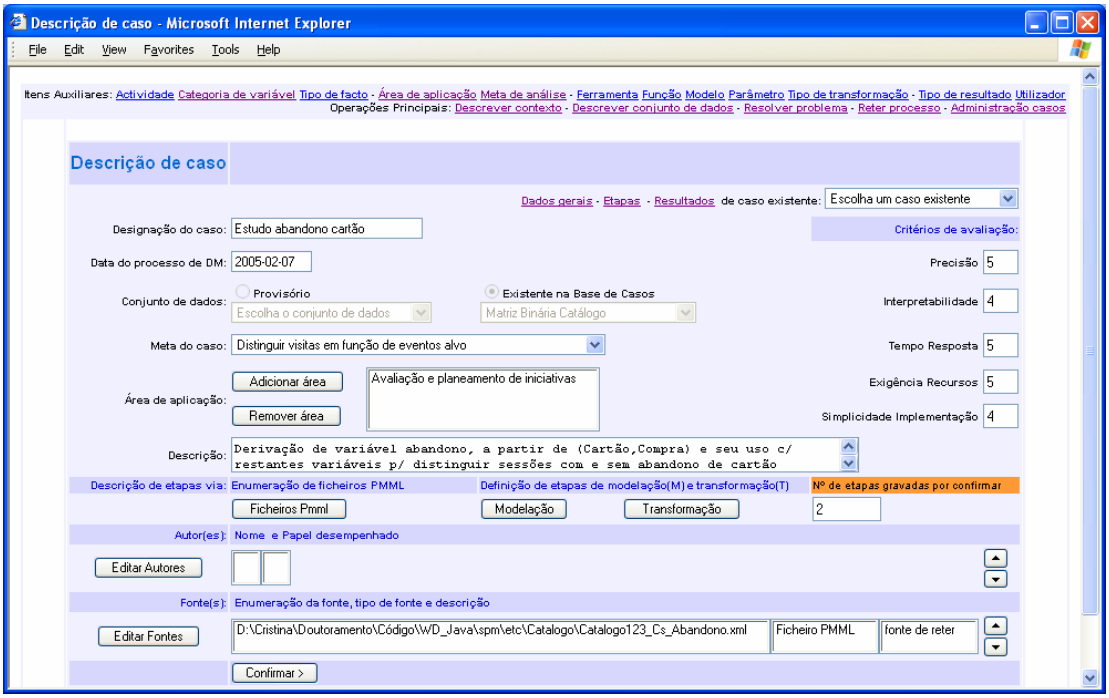


Figura 41 – Exemplo da descrição de um processo de WUM

A primeira etapa do processo foi descrita por interacção (explícita) com o utilizador, correspondendo a uma operação de transformação, para derivar uma nova variável denominada abandono. Este tipo de acção é concretizado através do botão de comando "Transformação" (Figura 41). Conforme se constata através da Figura 42, esta variável foi gerada a partir de duas variáveis iniciais: Cartão e Compra. A primeira variável assinala a existência (ou inexistência) de itens no cartão de compra, enquanto a segunda indica se a sessão do visitante incluiu ou não acções de compra. A situação de abandono pode, por conseguinte, ser identificada pela ocorrência de um cartão com itens (cartão=1 ou verdadeiro), cuja compra não se efectivou (Compra=0 ou falso). No seguimento da definição desta etapa, a variável abandono fica disponível para ser usada na especificação de outras etapas da mesma sessão, apesar de o seu registo só se efectivar após a submissão da descrição do processo ao sistema.

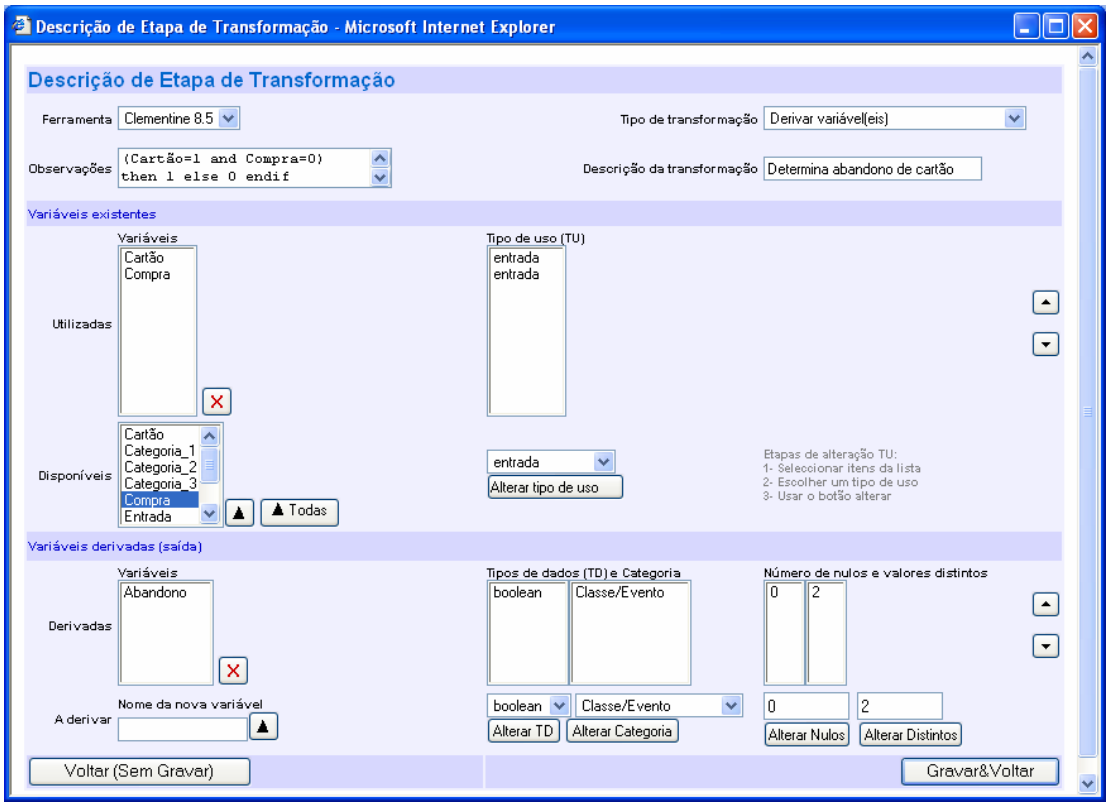


Figura 42 – Exemplo da especificação de uma etapa de transformação de dados

Prosseguiu-se com a definição de uma etapa de modelação, com vista a distinguir as sessões com e sem abandono de cartão, para tentar identificar os factores que influenciam ou determinam este tipo de comportamento. Para este efeito, recorreu-se maioritariamente a modelos da função de DM de classificação. Esta etapa foi descrita por intermédio de um documento PMML, cuja acção de enumeração se ilustra na Figura 43 e se encontra disponível através do botão de comando "Ficheiros PMML" (Figura 41). A enumeração de um documento PMML resulta, por omissão, na criação automática de uma fonte do processo, tal como se mencionou e mostrou anteriormente (Figura 41). Alternativamente, uma etapa de modelação pode também ser especificada por meio de interacção explícita com o utilizador, recorrendo, neste caso, ao botão de comando "Modelação" (Figura 41). As funcionalidades deste tipo de especificação são similares às reportadas na Figura 46, no âmbito da edição de uma etapa de modelação existente. As diferentes formas de especificação de etapas podem ser intercaladas e combinadas, durante a mesma sessão, para dar resposta às diferentes circunstâncias que se colocam neste contexto, tais como, indisponibilidade de todos os documentos PMML necessários para representar o exercício de WUM

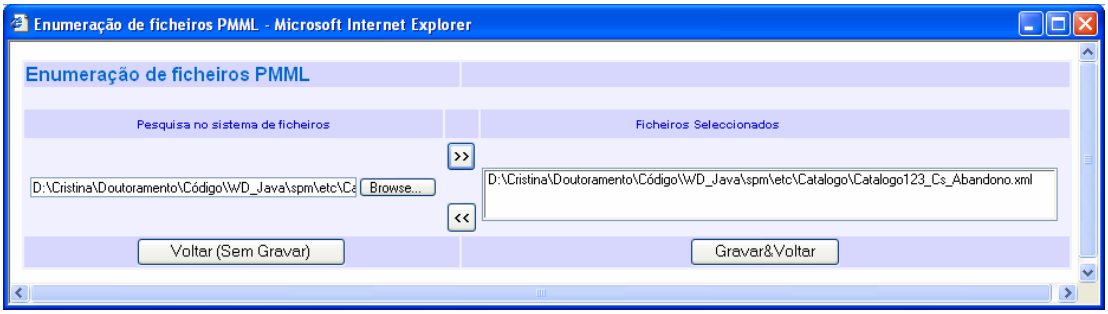


Figura 43 – Exemplo da enumeração de documentos PMML

A Figura 44 apresenta um extracto do conteúdo do documento PMML submetido ao sistema, salientando a negrito os principais elementos e atributos extraídos a partir deste, de acordo com as correspondências estipuladas na Tabela 8 e discutidas na respectiva secção 4.8.2. O primeiro item a ser obtido consiste na versão da especificação PMML, na qual o documento se baseia, uma vez que esta condiciona todo o processamento subsequente. Segue-se a extracção dos vários itens do documento, em parte dependente do tipo de modelo aplicado. Os elementos e atributos do dicionário de dados (*DataDictionary*) e do esquema de mineração (*MiningSchema*) facultam

informação adicional integrada no componente Esquema_Variável da base de casos do sistema SPM. Já alguns itens do cabeçalho (elemento *Application* de *Header*) e os atributos do modelo de mineração *("mining model")* permitem povoar automaticamente outros componentes da base de casos, tais como Ferramenta, Função, Modelo e Parâmetro.

A análise foi realizada recorrendo ao modelo de Rede Neuronal, disponível na ferramenta *SPSS Clementine* utilizada. Este modelo foi eleito, entre os vários modelos testados, apesar de não poder ser considerado apropriado para a análise corrente, em virtude de ter sido o que produziu melhores resultados. O modelo de rede neuronal não devolve uma explicação para a classificação produzida, a qual é necessária para entender os factores que influenciam o comportamento de abandono. Porém, os restantes modelos aplicados não proporcionam resultados razoáveis (e.g. precisão e selecção de alguma variável relevante). Além disso, o número de variáveis do conjunto de dados em causa é, excepcionalmente, muito reduzido, facilitando a fase de pós-processamento de descobertas, com base na análise sensitiva complementar, a qual indica a importância relativa das variáveis de entrada da rede. Na realidade, os resultados deste processo foram interessantes e inclusive consistentes com os gerados através de uma análise de agrupamento.

A nova variável abandono assumiu o papel de alvo (*usageType predicted*) no estudo. As variáveis Cartão e Compra foram omitidas nesta etapa do processo. As (8) variáveis que constam (parcialmente) no dicionário de dados e (exaustivamente) no esquema de mineração da Figura 44 reportam-se:

- às restantes (6) variáveis iniciais do conjunto de dados original, com papel activo na construção do modelo (*usageType active)*, as quais são indicativas dos diferentes recursos acedidos:
    - página de entrada do sítio Web (Entrada);
    - página que contém uma lista das várias categorias de produtos comercializados (Lista);
    - páginas individuais das diferentes categorias de produtos (Categoria_1, Categoria_2 e Categoria_3);
    - funcionalidade de pesquisa de produtos específicos (Procura);
- à variável alvo Abandono ($N-Abandono);
- à variável adicional correspondente ao alvo, automaticamente gerada ($NC-Abandono) pela ferramenta de DM neste modelo e com papel suplementar (*usageType supplementary)*.

```
<?xml version="1.0" encoding="UTF-8"?>

<PMML version="2.1" xmlns="http://www.dmg.org/PMML-2_1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

    <Header copyright="Copyright (c) Integral Solutions Ltd., 1994 - 2004. All rights reserved.">
        <Application name="Clementine" version="8.5"/>
        <Annotation>Exported with PMML extensions for use with SPSS SmartScore</Annotation>
    </Header>

    <DataDictionary numberOfFields="8">
        <DataField name="Entrada" displayName="" optype="categorical" dataType="integer">
          <Extension name="x-storageType" value="numeric" extender="spss"/>
          <Extension name="x-measure" value="flag" extender="spss"/>
          <Value value="0" property="valid"/>
          <Value value="1" property="valid"/>
        </DataField>
        <DataField name="Categoria_1" displayName="" optype="categorical" dataType="integer">
          <Extension name="x-storageType" value="numeric" extender="spss"/>
          <Extension name="x-measure" value="flag" extender="spss"/>
          <Value value="0" property="valid"/>
          <Value value="1" property="valid"/>
        </DataField>
        ...
        <DataField name="$N-Abandono" displayName="" optype="categorical" dataType="integer">
          <Extension name="x-storageType" value="numeric" extender="spss"/>
          <Extension name="x-measure" value="flag" extender="spss"/>
          <Value value="0" property="valid"/>
          <Value value="1" property="valid"/>
        </DataField>
        <DataField name="$NC-Abandono" displayName="" optype="continuous" dataType="double">
          <Extension name="x-storageType" value="numeric" extender="spss"/>
          <Extension name="x-measure" value="range" extender="spss"/>
        </DataField>
    </DataDictionary>

    <! -- elemento mining model -->
    <NeuralNetwork
        modelName="AbandonoCartao" functionName="classification" algorithmName="Neural Net"
        activationFunction="logistic" x-normalizationMethod="limitedDifference">

        <! – esquema de mineração -->
        <MiningSchema>
            <MiningField name="Entrada" usageType="active"/>
            <MiningField name="Categoria_1" usageType="active"/>
            <MiningField name="Categoria_2" usageType="active"/>
            <MiningField name="Categoria_3" usageType="active"/>
            <MiningField name="Lista" usageType="active"/>
            <MiningField name="Procura" usageType="active"/>
            <MiningField name="$N-Abandono" usageType="predicted"/>
            <MiningField name="$NC-Abandono" usageType="supplementary"/>
        </MiningSchema>

        <! – elementos descritivos da rede neuronal -->
        <NeuralInputs>  …  </NeuralInputs>
        <NeuralLayer>  …  </NeuralLayer>
        …
        <NeuralOutputs> … </NeuralOutputs>
    </NeuralNetwork>
</PMML>
```

Figura 44 – Excerto do documento PMML exemplificativo de uma etapa de modelação

A definição de uma etapa de transformação não é imprescindível para a prossecução da aquisição deste caso de aplicação de WUM. Se esta etapa não tivesse sido criada, a variável Abandono poderia ser obtida, por intermédio do documento PMML submetido ao SPM, e acrescentada, como derivada, ao conjunto de dados original, tal como se explicou na secção 4.8.1. No entanto, perdia-se a descrição mais clara e detalhada desta variável derivada. De facto, o sistema não cobre a especificação de transformações de dados mais complexas e, como estas também não podem ser capturadas automaticamente, optou-se por possibilitar a omissão da sua descrição, quando são fornecidos meios alternativos para recolher novas variáveis, através de documentos PMML.

Todas as variáveis usadas para gerar o modelo encontram-se representadas no documento PMML exportado. Para outros métodos, como as Árvores de Decisão, algumas variáveis irrelevantes para a execução do modelo resultante podem ser excluídas. Todavia, o documento omite um dos parâmetros usados na aplicação do modelo de Rede Neuronal e também não contém aspectos descritivos e justificativos que o analista pode conceder e o sistema está preparado para preservar. Neste sentido, procedeu-se à edição das etapas do processo, suportada pela página mostrada na Figura 45. Esta página disponibiliza uma série de operações, desde a remoção, mudança de ordem e alteração do conteúdo de uma etapa, passando pela adição de novas etapas e pela remoção de um caso (somente acessível para o administrador do sistema).

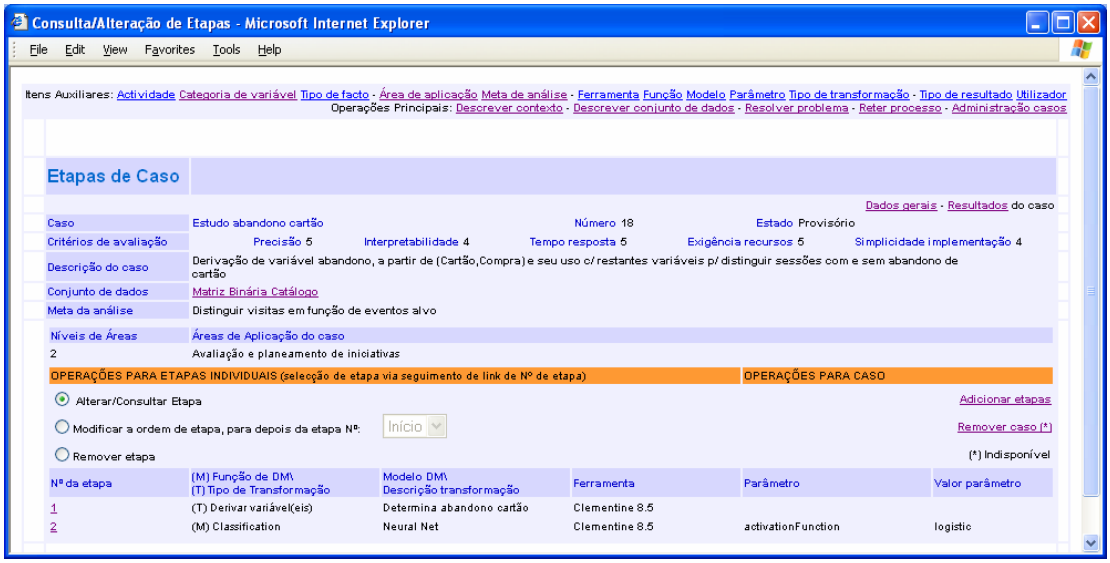


Figura 45 – Exemplo da edição de etapas de um processo

A alteração de uma etapa é realizada seguindo a ligação do seu número (Figura 45). A Figura 46 mostra o resultado da edição da etapa de modelação do processo exemplo. No âmbito desta edição salienta-se a facilidade da adição do parâmetro em falta, uma vez que os parâmetros já aplicados em cada tipo de modelo são apresentados separadamente, numa caixa de listagem dedicada a este efeito. A possibilidade de seleccionar simultaneamente todas as variáveis, como utilizadas na modelação, também se revela muito prática, designadamente quando estão em causa centenas de variáveis.

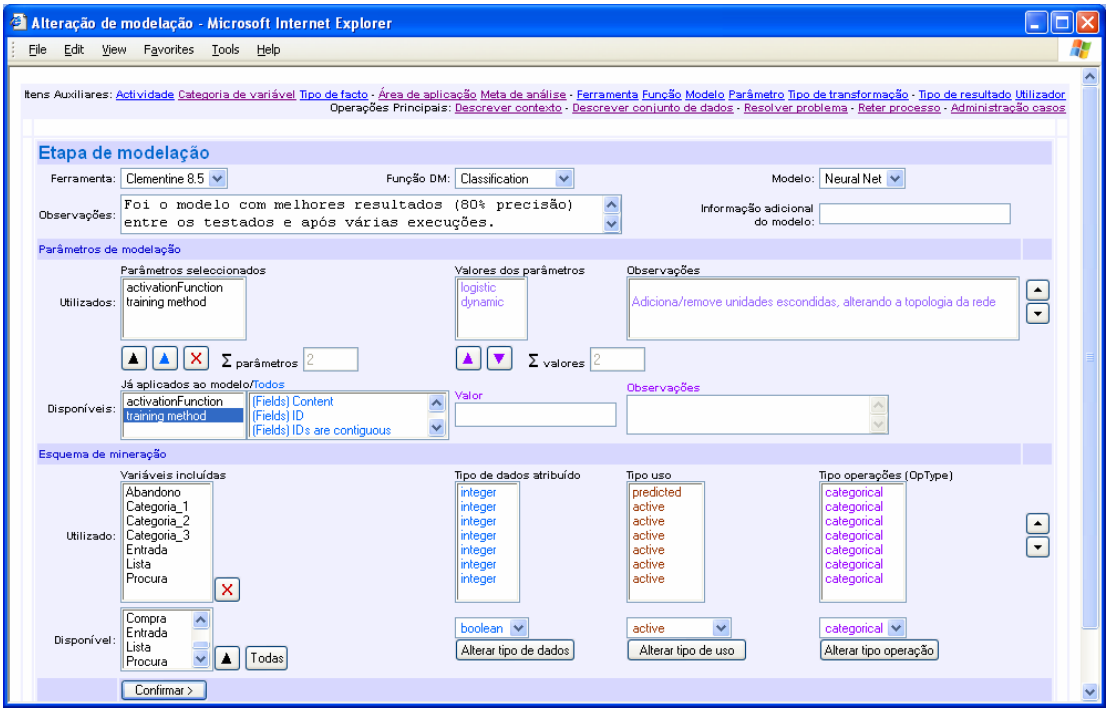


Figura 46 – Exemplo de edição de uma etapa de modelação

Terminada a edição do processo, o analista pode indicar que a descrição do processo já foi concluída, acedendo aos dados gerais do processo (ligações Dados gerais na Figura 41 ou Figura 45) e alterando o estado do caso de provisório para confirmado. Em contrapartida, o analista também pode classificar um caso provisório como inútil, como forma de comunicar que este deve ser removido. Em qualquer das situações, a partir do momento em que um caso deixa de estar no estado provisório, a gestão do seu ciclo de vida passa a ser da responsabilidade do administrador

do sistema e o cenário de utilização do sistema também muda para o âmbito da administração. Os casos inúteis e obsoletos podem ser removidos, os protegidos só podem ser consultados e o administrador pode alterar o estado de qualquer tipo de caso. Em termos gerais, a operação de administração de casos organiza os casos existentes pelo estado que estes detêm e confere um ponto de acesso a operações mais restritivas, após a validação com sucesso da informação de identificação do administrador do sistema. A manipulação de casos individuais é efectuada utilizando as mesmas páginas Web, mas disponibilizando operações de alteração de estado, para qualquer tipo de caso, e de remoção de casos inúteis ou obsoletos.

5.4 Principais Características do Protótipo do Sistema SPM

O sistema SPM foi implementado através de uma aplicação Web protótipo, a ser explorada num âmbito organizacional, de forma a promover a aquisição, partilha e reutilização de conhecimento acerca de processos de WUM ao longo da organização. As duas operações primordiais do sistema são a resolução de problemas e a aprendizagem a partir de exemplos concretos de processos de WUM. A resolução de problemas integra as tarefas caracterizar dados, construir problema, recuperar e reutilizar, enquanto a aprendizagem agrega as tarefas conciliar descrições e reter. Na implementação destas seis tarefas evidenciam-se uma série de aspectos que se passa a referir. A caracterização de dados assegura a abstracção e independência do acesso e extracção de metadados, face aos tipos de fontes de dados previstos, podendo estes ser facilmente extensíveis para outros tipos. A construção de problemas evita a dependência do sistema quanto a atributos descritores de problemas e outros tipos de propriedades, tratados dinamicamente. Esta característica, aliada à separação da implementação de funções de semelhança, permite que a recuperação seja disponibilizada como uma funcionalidade genérica e extensível. Já a reutilização é extensível, relativamente aos critérios de avaliação que baseiam a construção de planos, sendo ainda configurável, no que respeita à importância relativa atribuída a esses critérios pelo utilizador. A consolidação de descrições permite definir múltiplas etapas na mesma sessão, quer via interacção explícita, como por intermédio de documentos PMML. A obtenção de conteúdo a partir destes documentos é extensível em termos da versão PMML e acrescenta dois níveis adicionais de abstracção, correspondentes ao acesso e interpretação de documentos XML e aos mecanismos de extracção de partes destes documentos. A retenção, por sua vez, garante o registo eficiente de novos casos, abrangendo todas as operações inerentes e encapsulando os detalhes do armazenamento e estruturação da base de casos.

As duas operações mencionadas interligam os passos do ciclo CBR adoptado, fazendo parte de um dos cenários de exploração do sistema, designado assistência. Este cenário pressupõe a execução da operação de resolução de problemas, com conhecimento do problema em causa e usando uma descrição focada do mesmo, com o propósito de obter respostas ou soluções mais selectivas. Previu-se, ainda, mais dois cenários de utilização do SPM: exploratório e administração. O cenário exploratório advém de uma perspectiva mais voltada para a obtenção de indicações, recorrendo a uma descrição, usualmente, menos precisa do problema actual. No cenário de administração a finalidade reside na gestão da base de conhecimento, no sentido de manter os seus níveis de actualização, coerência e relevância. Em contraste com os outros dois cenários, o último é devotado a utilizadores com conhecimentos e capacidades para levar a cabo as operações de manutenção da base de conhecimento.

A exemplificação das duas operações fundamentais do SPM permite assinalar alguns aspectos, relativamente à forma de actuação do sistema. A operação de resolução de problemas é suficientemente versátil para comportar os requisitos dos cenários exploratório e de assistência. A procura de soluções pode ser orientada pelos requisitos da análise ou pelas propriedades dos dados em causa e, ainda, conjugando parte ou a totalidade de ambas as vertentes. A possibilidade de considerar apenas os descritores relevantes e de lhes atribuir níveis de importância específicos, bem como de estabelecer critérios de filtragem exacta, também confere flexibilidade a esta operação. A funcionalidade de aprendizagem requer o fornecimento explícito de alguns elementos, muitos dos quais não são obrigatórios, mas a descrição da maior parte das actividades de modelação pode ser automaticamente extraída, sem obrigar a esforços por parte dos utilizadores, e muitos dos componentes da representação de casos são povoados, com base em conceitos universais e, igualmente, de forma automática.

Capítulo 6

Conclusões e Trabalho Futuro

6.1 Do Problema Investigado à Solução Proposta

Actualmente, estabelecer uma presença efectiva na Web é um anseio desafiante, porém imperativo para a maior parte das organizações. A Web atingiu um elevado nível de maturidade e, naturalmente, os utilizadores depositam nela expectativas exigentes e diversas. Além disso, o uso de sítios Web, vulgarmente, não decorre conforme se esperava ou se desejava. Face a este cenário, torna-se vital avaliar o impacto e a eficácia de tais sítios e, sobretudo, identificar formas para aproveitar melhor as oportunidades emergentes da Web, actuando de modo mais responsivo e, mesmo, pró-activo, para alcançar as metas desses sítios. A descoberta de conhecimento aplicada aos dados referentes à interacção de indivíduos com sítios Web, conhecida por WUM, é um instrumento oportuno e fulcral para ambos os propósitos. Antes de mais, estes dados são registados nos ficheiros de histórico de servidores Web, sendo também, frequentemente, mantidos em bases de dados transaccionais ou consolidados em bases de dados analíticas, de modo a optimizar a sua exploração. A WUM pode então ajudar as organizações a transformar esta fonte de dados imensa, muito rica e abrangente, em conhecimento de valor e accionável, para orientar melhoramentos do sítio, conducentes ao retorno dos avultados recursos financeiros e humanos investidos. Os padrões de utilização da Web, conferidos pela WUM, são úteis para uma ampla variedade de finalidades e no apoio a vários tipos de decisões, as quais podem ser sistematizadas

em áreas de aplicação como o melhoramento de sistemas, alteração de sítios, personalização da Web e "inteligência" para o negócio.

Apesar do grande interesse e empenho na exploração de ferramentas de DM, a extracção de conhecimento permanece uma actividade bastante complicada. A curva de aprendizagem do desenvolvimento e aplicação de processos de KDD é um sério obstáculo, especialmente para utilizadores sem conhecimentos profundos e experiência no domínio. O ambiente Web, os dados de *clickstream* e, portanto, a WUM, levantam contratempos acrescidos e exigem maior eficácia e celeridade à KDD, por variados motivos:

- estes dados são normalmente muito vastos, contêm um grande número de variáveis, são intrinsecamente complexos e registam aspectos comportamentais, tornando-os susceptíveis de interpretações subtis;
- as descobertas são facilmente accionáveis;
- as restrições temporais na disponibilização dessas descobertas são inflexíveis;
- é necessário conjugar informação acerca da actividade do sítio, dos perfis de visitantes e do conteúdo e estrutura do sítio, e este encontra-se em constante mutação.

Uma questão crucial e geral da KDD é a selecção dos métodos de DM certos a aplicar, consoante a natureza do problema sob análise, com o intuito de obter resultados úteis para um determinado fim. As dificuldades críticas advêm da sobreposição desses métodos, com respeito aos tipos de problemas a que podem dar resposta, e, também, da variabilidade da sua adequação, em função de vários factores influentes, os quais, muitas vezes, são complicados, subjectivos e, por vezes, contraditórios. A apresentação das técnicas de DM mais exploradas em WUM é um reflexo destas dificuldades, evidenciando as suas diversas aplicações e sobreposições. O nível de adequação de cada técnica varia consideravelmente, introduzindo várias questões subtis, passíveis de influenciar substancialmente a selecção de métodos e abordagens de DM a adoptar. Apesar de se tratarem de questões gerais da KDD, na realidade, estas dificuldades agravam-se no âmbito da WUM pois, geralmente, os aspectos peculiares e complexos são mais abundantes e toda a estratégia de escolha de métodos encontra-se menos estudada e estruturada, em termos das categorias de problemas, tipos de actividades de mineração apropriadas, aplicações práticas relacionadas das descobertas e elementos de dados chave. Todos estes impedimentos à ampla exploração da WUM assumem, cada vez mais, maior preponderância, perante o seu alcance e utilidade, para diferentes tipos de colaboradores da organização, em virtude dos níveis variáveis de conhecimento e experiência na área dos potenciais analistas.

O reconhecimento dos desafios da KDD não é um facto novo, nem tão pouco recente, estando na origem de muitas iniciativas de assistência à persecução destes projectos. O objectivo destas iniciativas é promover a eficácia e produtividade destes processos, atenuando as dificuldades e esforços envolvidos na extracção de conhecimento útil, para o problema de análise em causa. Uma das dimensões que se pode usar para classificar tais iniciativas é o âmbito do apoio, sendo as suas vertentes primordiais a selecção de modelos, na fase de modelação, e a planificação geral de processos. A última vertente é, nitidamente, mais abrangente, pois pode contemplar diversos níveis de abstracção e fases da KDD, bem como estudos que requerem a aplicação de vários métodos de DM e outras operações. Outra dimensão significativa consiste na orientação fundamental da abordagem de assistência, no que concerne a factores influentes. As orientações predominantes centram-se nas características dos dados e nos requisitos explícitos do analista. A estratégia de recomendação é outro dos aspectos que contribuem para a descriminação de iniciativas com desígnios afins. Alguns sistemas procuram gerar a melhor solução para o problema actual ou construir uma lista de soluções plausíveis. Outros sistemas incentivam a reutilização de solução prévias e excelentes, para basear a resolução de novos problemas similares.

A abordagem de assistência adoptada é, antes de mais, especificamente devotada a projectos de WUM. Dados os contratempos da sua exploração prática, optou-se pelo âmbito de apoio de planificação de processos. Contudo, apesar de se considerar o desenvolvimento da totalidade do processo de WUM, deu-se prioridade à fase de modelação e supõe-se a existência de dados pré-processados e com qualidade, significando, principalmente, que o tratamento de etapas de transformação foi simplificado. Esta decisão deve-se à ênfase colocada em torno da recomendação de métodos de DM que, obviamente, está mais associada ao passo de modelação de exercícios de KDD. Quanto a orientações, entendeu-se que os dois factores influentes, atrás mencionados, são ambos imprescindíveis e determinantes para basear a abordagem de assistência. Os dados alvo de análise necessitam de ser convertidos para metadados relevantes e a descrição de requisitos tem de ser facilitada, recorrendo-se a abstracções relacionadas com os problemas reais a solucionar. Já a estratégia de recomendação de soluções baseou-se na reutilização de processos úteis e bem sucedidos, ao nível da organização, seleccionando a lista de planos de mineração mais adequados, para resolver um determinado problema, conforme se justifica seguidamente.

A maior parte do sucesso atingido por especialistas, quando lidam com problemas de WUM, deve-se, sobretudo, à experiência e conhecimento adquirido que estes detêm. De facto, os próprios especialistas não conseguem proporcionar regras gerais e consistentes para sustentar a resolução de problemas. Por conseguinte, a ideia defendida, para combater muitos dos desafios da WUM,

reside na gestão do conhecimento obtido a partir da experiência na resolução de problemas concretos, no âmbito de um ambiente organizacional, para, assim, constituir a base para a partilha e reutilização de tal conhecimento nesse âmbito. Esta ideia pressupõe a construção de um repositório de conhecimento sobre processos anteriores, a aquisição de conhecimento a partir de novas experiências de WUM e a disponibilização de meios para ajudar os utilizadores a relacionar os problemas enfrentados com as estratégias plausíveis de resolução de problemas registadas. O âmbito do sistema permite repercutir as especificidades da organização no conhecimento recolhido, viabilizando também a criação de uma memória centralizada, passível de ser explorada por uma audiência alargada de membros do organismo. Esta ideia foi consolidada recorrendo ao paradigma CBR, já que o mesmo atende, intrinsecamente, a vários dos requisitos essenciais a satisfazer e vai ao encontro do tipo de abordagem de assistência que se pretende facultar.

O paradigma CBR é uma abordagem de aprendizagem e resolução de problemas, sediada na utilização do conhecimento específico de experiências passadas, registadas e mantidas numa base de casos. Um novo problema é solucionado recuperando um caso prévio similar e reutilizando-o nas circunstâncias actuais. Os problemas resolvidos podem, igualmente, ser retidos, ficando disponíveis para serem aplicados a problemas futuros. Consequentemente, os modelos de representação baseados em casos podem actuar como bases para a estruturação e exploração de conhecimento, acerca de processos úteis de WUM da organização. Os mecanismos CBR de aprendizagem e resolução de problemas também favorecem, por inerência, a extensibilidade incremental sustentada do repositório de conhecimento e a reutilização ampliada do mesmo em novas situações, face à flexibilidade da comparação baseada em similaridade e à possibilidade intrínseca de compactuar com descrições incompletas e imprecisas de problemas. Entre os muitos argumentos que consubstanciam a adopção do paradigma CBR neste contexto, salientam-se os seguintes:

- a experiência no domínio possui uma importância proeminente em WUM;
- os exemplos concretos de problemas resolvidos são um meio muito útil e convincente de auxílio, pois podem simplificar a complexidade subjacente, oferecendo, ao mesmo tempo, informação detalhada, explicativa e justificativa de toda a envolvente;
- problemas recorrentes e o uso repetido dos mesmos métodos são circunstancias comuns em KDD;
- a extensibilidade é uma capacidade muito atractiva, dada a constante evolução da WUM.

Em acréscimo, a exploração do paradigma CBR presta-se à conjugação de variadas técnicas e tecnologias, requeridas para responder aos requisitos adicionais da abordagem defendida.

Um dos requisitos estipulados prende-se com o recurso a padrões em uso no domínio, no sentido de colmatar a necessidade de vocabulário aceite a adoptar e de indicações quanto à estruturação de processos de KDD. O recurso a padrões para basear a modelação de conhecimento tem a vantagem estratégica de reduzir a necessidade da intervenção de especialistas na matéria. A especificação PMML corresponde a estas expectativas, pois trata-se de uma linguagem normalizada e baseada na XML, para descrever e partilhar modelos estatísticos e de DM. O padrão é largamente suportado por ferramentas de KDD e, também, por outros tipos de aplicações, como um mecanismo de transferência e operacionalização dos resultados obtidos, independentemente da ferramenta que os gerou. Outro requisito relevante consiste na integração de tecnologias de gestão e exploração de dados, vulgarmente encontradas nas organizações, de forma a retirar partido das suas capacidades e a maximizar o seu potencial. Esta integração pode ser realizada a dois níveis: para aceder a fontes de dados da organização e para gerir o repositório de conhecimento de processos de WUM.

Por último, evidencia-se o requisito de optimização do povoamento de tal repositório, automatizando, na medida do possível, a aquisição de conhecimento, a partir de volumes consideráveis de dados e informação, provenientes de fontes diversas. Esta optimização pode ser conseguida em três modalidades:

- extracção automática e consistente de alguns metadados, ao longo de diferentes tipos de fontes de dados alvo de análises;
- captura de grande parte da descrição das actividades de mineração, por intermédio de documentos PMML;
- disponibilização de facilidades para encorajar o fornecimento de elementos complementares, por parte das fontes humanas.

O sistema *Selector de Planos de Mineração* (SPM) concretiza a abordagem de assistência defendida, estabelecendo os componentes funcionais requeridos, para atender a todos os seus requisitos e ideias subjacentes debatidas. A principal missão do sistema é assistir os analistas, no desenvolvimento e aplicação de processos de WUM, auxiliando-os a identificar as estratégias mais adequadas para solucionar problemas de análise de dados de *clickstream*, sustentado por casos respeitantes a experiências úteis de WUM, concluídas com sucesso no passado. O SPM fundamenta-se no paradigma CBR, comportando-se como uma ferramenta de gestão e reutilização de casos de aplicação de WUM, ao nível da organização. Na assistência à resolução de problemas, o sistema actua, tipicamente, a partir das características dos dados alvo e de uma série de requisitos da análise e, com base no conhecimento detido acerca da aplicação de métodos de DM,

produz uma solução – planos de WUM consentâneos. O SPM também suporta os analistas na descrição de novos processos de WUM, com vista a facilitar esta tarefa árdua e a adquirir conhecimento relativo a novas experiências. Tal como qualquer outro sistema CBR, a recolha de um caso é iniciada quando surge um novo problema, durante o uso do SPM para solucionar problemas, e prossegue, com maior intensidade, ao longo da fase de aprendizagem, culminando com o seu registo na base de conhecimento do sistema. No que respeita aos componentes fundamentais do SPM, os quais contribuem para que o sistema cumpra a sua missão, apresenta-se seguidamente um resumo das suas características primordiais.

A base de conhecimento do SPM foi organizada à volta de dois componentes nucleares. O primeiro corresponde à tradicional base de casos, contendo exemplos detalhados de processos de WUM, descritos em termos de um problema de domínio e da respectiva solução aplicada. Um problema de domínio é, basicamente, retractado por características do conjunto de dados alvo, categorizações do tipo de problema de WUM e critérios de avaliação do sucesso do processo e dos seus resultados. Já a solução aplicada reporta as actividades de mineração e transformação executadas, conhecimento prévio e derivado e informação geral do processo. O segundo componente da base de conhecimento engloba, por sua vez, quatro contentores de conhecimento, referentes a aspectos particulares do corrente âmbito de aplicação de CBR, tais como, vocabulário descritivo do domínio, propriedades usadas no cálculo de similitude entre casos, metadados provisórios de conjuntos de dados alvo de análises e configurações do comportamento do sistema. A concepção da base de conhecimento foi influenciada por princípios reconhecidos de estruturação de conhecimento em sistemas CBR, quer ao nível geral dos diferentes tipos de conhecimento, como da própria representação dos casos. A representação de casos e o vocabulário adoptado basearam-se, ainda, na especificação PMML.

Para além da base de conhecimento, o SPM inclui seis módulos funcionais, responsáveis pela persecução das tarefas essenciais da resolução de problemas e aprendizagem a partir da experiência prévia. A resolução de problemas é assegurada pelos módulos de caracterização de dados, construção de problemas, recuperação e reutilização, enquanto a aprendizagem é garantida pelos módulos de conciliação de descrições e retenção. A criação destes módulos inspirou-se nas tarefas preconizadas pelo célebre modelo do ciclo CBR, mas adaptando esse ciclo aos requisitos correntes. Como consequência, existem diferenças entre o ciclo original e o adoptado. Por um lado, o passo rever (*Revise*) não figura em nenhum módulo do SPM. A revisão e a própria adaptação (no passo de reutilização) dos planos de WUM decorrem fora do sistema. O analista leva a cabo estas actividades, contando com a ajuda de descrições e explicações pormenorizadas,

associadas ao plano de WUM que escolheu. Na realidade, o analista faz parte do processo de raciocínio, uma vez que não se pretende substituí-lo, mas sim apoiá-lo. Por outro lado, acrescentou-se três novas tarefas ou passos, individualizando sub-tarefas afins de alguns passos do ciclo CBR original, dado que as mesmas assumem papéis preponderantes no SPM e são, em certa medida, autónomas.

Uma das tarefas acrescentadas ao ciclo CBR original é realizada pelo módulo de caracterização de dados, cuja missão é converter o conjunto de dados alvo para uma meta-representação sistemática e consistente, ao longo de diferentes tipos de fontes de dados. Esta meta-representação traduz propriedades significativas, para o propósito da selecção de métodos e abordagens de KDD, e substitui os dados originais, em comparações entre diferentes conjuntos de dados. O resultado concedido por este módulo simboliza requisitos inerentes, contemplando propriedades específicas de dados de *clickstream*, especificadas pelo analista, e metadados genéricos, extraídos automaticamente pelo sistema a partir da fonte de dados. No entanto, a caracterização de dados não cobre restrições explícitas da análise, quer no que concerne às próprias propriedades dos dados, como a outras categorias de preferências ou expectativas, as quais podem melhorar a descrição do problema. Assim sendo, adicionou-se também um módulo de construção de problemas, que tem a seu cargo as funções de auxílio, recolha e organização da especificação de requisitos explícitos, para facultar ao sistema um problema alvo, convenientemente detalhado e organizado. O objectivo estratégico deste módulo é contribuir para a redução da dependência do SPM, relativamente a descritores de problemas e suas propriedades.

Os módulos de recuperação e reutilização concretizam tarefas típicas, mas com alguns aspectos particulares no SPM. O módulo de recuperação tem por incumbência encontrar os casos mais semelhantes ao problema alvo e, portanto, os que são mais promissores para solucionar esse problema. Uma actividade fulcral para assegurar a funcionalidade de recuperação é a medição do nível de similaridade do alvo, em relação a problemas prévios, registados na base de casos. Uma vez que existem relacionamentos de um para muitos, entre várias partes da descrição de um caso, o recurso a uma formulação proposicional usual não é viável e, como tal, a determinação de similitude entre problemas comporta desafios acrescidos no SPM. Neste sentido, foi necessário especificar um procedimento de cálculo e medidas de similitude ajustados, face aos requisitos subjacentes e à semântica de semelhança pretendida, para tratar as situações de confronto entre conjuntos de elementos finitos, cuja cardinalidade pode ser diferente e variável. Esta situação ocorre, por exemplo, na comparação entre as variáveis de dois conjuntos de dados. Recorreu-se também a medidas de semelhança comuns, nas situações de comparação de atributos usuais. Já o

papel desempenhado pelo módulo de reutilização prende-se com a avaliação, organização e consubstanciação dos casos candidatos recuperados, a fim de produzir uma série de planos de WUM alternativos a apresentar ao analista, acompanhados por indicações para apoiar a sua escolha e por referências para detalhes de casos exemplificativos. Os planos apresentados destacam as operações e métodos que devem ser aplicados aos dados alvo, preparando a reutilização prática das recomendações, na resolução do problema em causa. No entanto, conforme já se referiu, a adaptação dos planos às circunstâncias específicas do problema actual é efectuada pelo analista.

A capacidade de resolução de problemas de um sistema CBR depende directamente do conteúdo da sua memória de casos. Daí surge a necessidade de evolução contínua de tal memória. Os novos processos de WUM bem sucedidos e úteis da organização são a fonte essencial de aprendizagem do sistema. Todavia, o conhecimento vital sobre estes processos encontra-se disperso por vários tipos de fontes, entre os quais as humanas, e a sua recolha requer esforços avultados por parte dos analistas. Diante deste cenário, colocou-se grande empenho na concepção de uma abordagem semi-automática de aquisição de conhecimento, capaz de sistematizar e agilizar a captura do mesmo, ao longo da organização e de acordo com os preceitos já mencionados. O objectivo da terceira tarefa aditada e levada a cabo pelo módulo de conciliação de descrições é, justamente, dar suporte a parte dessa abordagem, aceitando e compatibilizando uma descrição de processos heterogénea e do ponto de vista da conveniência para o analista, baseada, simultaneamente, em documentos PMML e em interacção explícita, para colmatar as omissões desses documentos. Por fim, o módulo de retenção, o qual é responsável por uma tarefa tradicional, conclui o ciclo CBR adoptado e completa a aquisição de conhecimento, conjugando resultados dos módulos de caracterização de dados e de conciliação de descrições com as suas próprias capacidades, para permitir que novos casos de aplicação de WUM sejam acrescentados à base de casos. Todos os itens disponíveis são integrados e estruturados, formando um novo caso, que é catalogado, com o intuito de simplificar a sua reutilização no futuro, e, em seguida, é registado na memória de casos.

O sistema SPM foi implementado através de uma aplicação Web protótipo, materializando os componentes constituintes enumerados e tendo presentes as ideias discutidas, com influência na sua concepção. A escolha de uma aplicação Web deve-se, sobretudo, às vantagens de acessibilidade, flexibilidade da interacção com o utilizador e independência de plataforma desta categoria de aplicações. A aplicação recorre, fundamentalmente, a tecnologias de base de dados, PMML, Java e Web. A eleição de recursos na implementação norteou-se pelas capacidades

requeridas, dando prioridade a *software* livre, com código aberto e multi-plataforma, bem como a padrões e interfaces de programação com grande aceitação.

A implementação do SPM enquadra-se numa arquitectura cliente servidor típica e foi organizada em três camadas de serviços: interacção, negócio e dados. A camada de interacção gere o acesso às operações disponíveis, sendo responsável pela recepção e encaminhamento de pedidos, recolha de dados e apresentação de resultados. Os serviços de negócio encapsulam as funcionalidades do sistema, tendo sido estruturados em bibliotecas dedicadas aos seguintes tipos de finalidades:

- concretização dos módulos do SPM;
- interligação e coordenação da actuação desses módulos, para formar operações, como a resolução de problemas e a aprendizagem;
- individualização de algumas funcionalidades ou vectores do sistema, designadamente, a gestão de objectos constituintes do modelo de domínio, as funções que implementam as medidas de similaridade em uso e o tratamento de propriedades mantidas nos contentores de vocabulário, similaridade e configuração.

A camada de dados proporciona serviços de persistência e transformação, incluindo:

- bibliotecas, mais genéricas, que lidam, directamente, com o acesso e manipulação de diferentes tipos de fontes de dados e conhecimento;
- bibliotecas, mais específicas, que realizam um tratamento das fontes, de acordo com as necessidades particulares do SPM, ou que abstraem determinadas funcionalidades, como a gestão da base de conhecimento e o acesso a fontes de dados alvo de analises.

Uma das características do protótipo é a separação das perspectivas mais significativas do sistema a diversos níveis. Em termos genéricos, esta separação dá-se nas três camadas de serviços prestados. Em termos específicos, a subdivisão surge, principalmente, na individualização das tarefas dos módulos, na separação do modelo de domínio dos mecanismos de raciocínio e no isolamento da implementação e gestão da base de conhecimento. Envidaram-se ainda muitos esforços no sentido de alcançar uma implementação, tanto quanto possível, genérica e extensível de um sistema CBR de assistência à WUM. A extensibilidade é garantida para os vectores críticos do sistema, designadamente, origens de dados e conhecimento, atributos descritores de problemas, critérios de avaliação de soluções e funções de similaridade. Como resultado, é possível assinalar uma série de aspectos significativos, indicada, seguidamente, em função das tarefas nucleares do SPM:

- Caracterização de dados – assegura a independência e extensibilidade do acesso e extracção de metadados, face aos tipos de fontes de dados alvo, contemplando,

correntemente, dois dos tipos de fontes mais usuais de suporte à exploração de dados de *clickstream* (i.e. sistemas de base de dados e ficheiros convencionais).

- Construção de problemas – evita a dependência dos mecanismos de recuperação e reutilização, relativamente a atributos descritores de problemas e a critérios de avaliação de soluções.
- Recuperação – é, ainda, independente das funções de similitude, funcionando como um mecanismo genérico.
- Reutilização – para além de ser extensível em relação aos critérios de avaliação, é configurada dinamicamente, atendendo à importância relativa atribuída pelo analista a esses critérios.
- Consolidação de descrições – admite duas formas de descrição de processos, baseadas, respectivamente, em interacção explícita e documentos PMML, garantindo a extensibilidade do sistema, quer em termos de (novas) formas de descrição de processos, quer das versões da especificação PMML.
- Retenção – integra várias origens de conhecimento, nomeadamente, fontes de metadados, documentos PMML e fontes humanas, procedendo ao registo de novos casos e encapsulando os pormenores da estruturação da base de casos.

A utilização do sistema foi prevista com base em três cenários primordiais: assistência, exploratório e administração. O primeiro é o mais relevante e foi demonstrado, com maior ênfase, à luz de exemplos concretos. Este cenário decorre de acordo com os passos do ciclo CBR adoptado, englobando a assistência na resolução de problemas e na descrição de novos processos de WUM. No que concerne à resolução de problemas, este cenário pressupõe o conhecimento do problema actual e das abstracções de especificação das orientações da análise, bem como a submissão desse problema ao sistema, usando uma descrição focada, com vista a obter respostas mais selectivas. Já no que diz respeito à descrição do processo, subentende-se que o analista conhece, em acréscimo, a solução do problema e os aspectos cruciais dessa solução, requeridos para sustentar a construção de um novo caso de aplicação de WUM. O cenário exploratório advém de uma perspectiva de uso do sistema, mais voltada para a obtenção de indicações, usando, tipicamente, uma descrição imprecisa do problema corrente. Em contraste com os outros cenários, o último é destinado a utilizadores com perfil de administrador do sistema e, portanto, detentores de conhecimento e experiência no domínio da WUM. Neste contexto, a finalidade é a gestão da base de conhecimento, cobrindo, entre outras actividades, a manutenção da base de casos. A administração de casos é uma das funcionalidades adicionais do SPM, contendo operações mais restritivas, como certas alterações de estado dos casos e a remoção de casos. Apesar da

importância de tal actividade, o SPM, presentemente, ainda não abrange meios verdadeiramente expeditos de apoio à sua persecução. Assim sendo, cabe ao administrador do sistema a incumbência de discernir os casos que são úteis, gerindo o seu ciclo de vida e toda a base de conhecimento, para que esta se mantenha actualizada, coerente e relevante.

6.2 Avaliação de Resultados

A análise dos resultados alcançados reporta-se a uma base de casos preliminar, contendo exemplos de aplicação de WUM e construída com o propósito de testar a efectividade do protótipo implementado. Todos esses exemplos são baseados em dados e análises reais, divulgados na Internet. Algumas dessas análises reproduzem processos de WUM desenvolvidos com esses dados, publicados junto com as respectivas fontes e mesmo em artigos científicos. Optou-se, quase desde o início, por não recorrer a exemplos simulados, pois a avaliação de resultados de processos de KDD é relativa, tornando-se muito difícil de aferir nesses moldes. Esta decisão mostrou-se acertada, não só, por atenuar o problema da natureza subjectiva do sucesso das descobertas, mas também devido à maior diversidade de circunstâncias com as quais se contactou. Porém, tal decisão também introduziu efeitos negativos. O sistema não foi avaliado a partir da situação idealizada e previamente estabelecida: fontes de dados pré-processados e com qualidade; problemas de análise colocados no âmbito de uma só organização.

A preparação de casos, nas novas condições, é uma tarefa mais complicada e morosa. Cada série de análises estende-se por períodos temporais longos, já que requer esforços avultados de transformação de dados e, principalmente, a compreensão desses dados e de todo o contexto que os rodeia. Na sequência de tal decisão, o perfil da base de casos também se modificou ligeiramente, em virtude de não se referir exclusivamente a uma organização, mas sim a diferentes instituições, enquadradas em contextos distintos. A diversidade de situações e problemas de WUM é benéfica para basear a concepção e o próprio povoamento de um repositório desta natureza, mas a heterogeneidade dos conjuntos de dados também conduz a diferenças mais acentuadas em termos das suas propriedades, as quais acabam por se reflectir no processo de recuperação. Já as vantagens marginais da direcção seguida foram uma sensibilidade acrescida para os desafios reais destes processos e as capacidades desenvolvidas, ao longo das numerosas actividades de tratamento e mineração destes dados.

Diante do exposto, o número de casos existentes não é grande e as actividades de preparação de casos adicionais continuam a decorrer. Por esta razão, os resultados conseguidos, até ao

momento, ainda são limitados. As próximas secções descrevem e avaliam esses resultados, em termos das duas operações mais importantes do sistema.

6.2.1 Resolução de Problemas

Um aspecto fundamental da resolução de problemas no sistema SPM é a recuperação de casos similares e, no seu seguimento, a comparação de conjuntos de dados, até porque os descritores da sua caracterização predominam, relativamente a outros tipos de atributos da descrição de problemas de WUM, conforme se pode constatar a partir da Tabela 4 (dimensão D). Os conjuntos de dados do repositório de casos construído possuem duas facetas primordiais. A primeira prende-se com o contexto em que cada conjunto de dados se enquadra. No cenário de teste do SPM, que se acabou de explicar na secção anterior, cada contexto integra os conjuntos de dados relacionados com uma determinada envolvente, correspondendo esta, simultaneamente, a uma organização, sítio Web e caso de estudo. A outra faceta resulta das configurações que os conjuntos de dados assumem, as quais são essencialmente de dois tipos:

- conjunto de dados inicial, obtido a partir de uma fonte e submetido ao SPM para caracterização, englobando todas ou parte das variáveis originais e também, tipicamente, novas variáveis derivadas, sendo utilizado em um ou mais processos de WUM;
- esquemas de mineração das análise, correspondendo a séries de variáveis efectivamente usadas em etapas de modelação de exercícios de WUM, as quais podem incluir um subconjunto ou a totalidade das variáveis iniciais e, ainda, outras variáveis adicionais, derivadas após a submissão do conjunto de dados inicial.

A ideia subjacente da coexistência dos dois tipos de configurações é captar e preservar, tanto quando possível, aspectos de várias fases do processo de KDD, de forma a criar condições para viabilizar a submissão ao sistema de problemas baseados em diferentes categorias de factores influentes. Os metadados ao nível do conjunto de dados retêm propriedades gerais, inclusive alguns aspectos respeitantes a todas as variáveis iniciais, por intermédio de atributos como as percentagens de colunas numéricas, categóricas, temporais e binárias. Já as características mais específicas das variáveis, como a categoria semântica e o tipo de dados, são mantidas nos metadados ao nível das variáveis individuais. O esquema de mineração (classe Esquema_Variável), por sua vez, regista as variáveis mais importantes para os diversos processos de WUM (i.e. as usadas em etapas de modelação). As considerações anteriores justificam e permitem distinguir os três tipos de testes mais pertinentes, no âmbito da comparação de conjuntos de dados:

- entre variáveis de conjuntos de dados;
- entre conjuntos de dados e em função de todas as variáveis iniciais;
- entre conjuntos de dados e em função das variáveis mais relevantes.

O teste específico de comparação entre variáveis de vários conjuntos de dados foi realizado separadamente, numa fase preliminar, com o intuito de fundamentar a escolha de medidas de similaridade para atributos de cardinalidade múltipla. Este tipo de teste foi estendido a todos os descritores abrangidos nestas circunstâncias e os seus resultados já foram brevemente discutidos na secção 4.5.3. Como estas comparações fazem parte do cálculo global de similaridade, são abordadas, de forma implícita, juntamente com os demais testes reportados nesta secção. Nomeadamente, a comparação de variáveis é englobada, por omissão, na comparação de conjuntos de dados, mas envolvendo apenas as variáveis que constam nos esquemas de mineração (no que respeita aos casos prévios). Pretende-se, desta forma, facultar um meio para dar ênfase às variáveis mais significativas dos processos, uma vez que as propriedades gerais das variáveis iniciais já são retractadas nos metadados ao nível do conjunto de dados.

Quanto aos restantes testes enumerados, correspondem a duas das formas mais típicas de submissão de problemas ao SPM, sendo úteis para avaliar o sistema. O primeiro destes testes (T1) representa a submissão de um problema alvo, orientado pelos requisitos inerentes de um conjunto de dados alvo, na sua forma inicial, sem assinalar variáveis relevantes, nem outros tipos de requisitos explícitos. O utilizador não selecciona as variáveis mais relevantes porque desconhece quais são ou porque são todas proeminentes. Em qualquer das possibilidades, é conveniente comparar todas as variáveis iniciais com as mais importantes dos casos existentes. Já o segundo teste (T2) traduz a submissão de um problema alvo baseado num conjunto de dados e indicando as variáveis mais relevantes, mas novamente sem especificar outros tipos de requisitos explícitos. Na sequência destes testes, torna-se necessário avaliar a submissão de problemas alvo, cobrindo também requisitos explícitos, através da adição de mais um teste (T3).

Os três tipos de testes referidos (T1,T2 e T3) foram levados a cabo recorrendo ao subconjunto de casos e conjuntos de dados mais representativos e já registados no SPM. Este subconjunto contempla 5 contextos, 16 conjuntos de dados e 23 casos (e esquemas de mineração). O teste T1 pode ser efectuado submetendo sucessivamente problemas baseados nos 16 conjuntos de dados, enquanto os testes T2 e T3 podem ser conduzidos em função dos 23 casos, usando também os esquemas de mineração desses casos, para retractar as variáveis mais relevantes, e os valores dos outros atributos de descrição de problemas, no papel de requisitos explícitos. Estes testes foram

elaborados sem aplicar as facilidades de filtragem exacta e de atribuição de níveis de importância específicos aos descritores. A Tabela 11 resume as principais características de tais testes.

Tabela 11 – Principais características dos testes realizados

Teste	**Descrição de requisitos**	**Dimensões e Nº descritores**		**Comparações**
T1	Características do conjunto de dados Características de variáveis (todas as variáveis iniciais)	D	**14** **+1** (4 atributos internos)	16 X 23
T2	Características do conjunto de dados Características de variáveis (só as variáveis relevantes)	D	**14** **+1** (4 atributos internos)	23 X 23
T3	Características do conjunto de dados Características de variáveis (só as variáveis relevantes) Restantes requisitos explícitos	D T P	14 +1 (4 atributos internos) **+9**	23 X 23

No que concerne às principais características dos casos e conjuntos de dados considerados nos testes, destaca-se a diversidade entre os dados originais, a partir dos quais se criaram os conjuntos de dados iniciais. Os dados originais variam desde históricos de servidores Web, na sua forma bruta, a dados já pré-processados, nalguns casos com séries de dados distintas, disponibilizadas para o tratamento de problemas diferentes. Por exemplo, o caso de estudo, eventualmente, com maior notoriedade nesta área, advém da competição de DM *KDD Cup 2000* [Kohavi et al. 00]. Usou-se três conjuntos de dados baseados nesta fonte, inseridos no contexto KddCup, e reproduziu-se quatro análises baseadas nestes dados. As séries de dados foram fornecidas para lidar especificamente com as questões colocadas no âmbito da competição. Neste sentido, o tratamento prévio dos dados originais acabou por ser substancialmente diferente. Para os históricos em estado bruto concretizaram-se diversas tarefas de pré-processamento e geraram-se vários conjuntos de dados iniciais, tais como, por exemplo, registos ao nível do acesso, agregações ao nível da sessão, matrizes binárias de páginas acedidas em sessões (sessões X paginas) e matrizes binárias transpostas (páginas X sessões). Já no que respeita aos restantes dados originais, essencialmente, derivaram-se novas variáveis e filtraram-se outras, tendo em conta critérios como a qualidade dos dados e sua pertinência para os problemas em causa. A maioria destas actividades não foi registada no SPM, pois estas são numerosas e possuem descrições demasiado trabalhosas, para as quais não existem meios de automatização da sua aquisição. Além disso, também não existe cobertura, por parte do sistema, para descrever convenientemente algumas dessas actividades. Como resultado, os conjuntos de dados existentes

reflectem apenas parte da diversidade original, omitindo as actividades de pré-processamento e algumas operações de transformação, mas preservam as propriedades entendidas como mais significativas para a selecção de métodos e abordagens de DM, ao longo dos vários descritores, como, por exemplo, a granularidade dos dados (e.g. acesso, sessão e visitante), o número de colunas ou variáveis (presentemente de 3 a 989819, cujo valor máximo, muito elevado, advém de matrizes transpostas) e o número de linhas ou registos (correntemente de 100 a 4179752).

Os 23 casos utilizados nos testes incluem desde análises simples, contemplando uma só etapa e a aplicação de um único método de DM, a processos mais elaborados, envolvendo várias etapas e a aplicação combinada de diversos métodos de DM. Os 23 casos abrangem ainda a totalidade de metas de análise e de áreas de aplicação definidas no sistema.

Seguidamente, reportam-se os resultados dos três testes realizados. Começando pelo teste T1, cujas condições primordiais de elaboração são ilustradas na Figura 47, só será atingido o valor de similitude 1 (nas comparações) se existirem esquemas de mineração que usam todas as variáveis iniciais do conjunto de dados. Os casos mais similares são, naturalmente, aqueles que se baseiam nesse conjunto de dados alvo e o nível de similaridade é maior quando o número de variáveis do esquema de mineração se aproxima do número de variáveis iniciais. Este resultado só tem interesse para efeitos de verificação, na medida em que se pretende simular a submissão de um conjunto de dados não existente no sistema. Consequentemente, as comparações de cada conjunto de dados com os casos baseados nestes são excluídas na restante exposição.

16	**X**	**23**
Problemas alvo definidos pelos 16 conjuntos de dados, com todas as variáveis iniciais		Casos que usam os 16 conjuntos de dados, através de 23 esquemas de mineração
e.g. Variáveis iniciais do conjunto de dados cd5: N_session, N_access, Page		*e.g. Esquema de mineração do caso 8 que usa cd5: N_session, Page*

Figura 47 – Condições de realização do teste T1

Em termos gerais, os resultados do teste T1 vão ao encontro das expectativas e da noção intuitiva de similaridade entre conjuntos de dados, no que concerne às propriedades constituintes da caracterização de dados e à ideia geral formada de cada conjunto de dados. Entre as principais tendências confirmadas, assinalam-se as seguintes:

- Quando o conjunto de dados alvo é uma matriz binária:

  - os casos (nomeadamente, os conjuntos de dados subjacentes) mais semelhantes são os baseados também em matrizes binárias. Verifica-se uma ligeira excepção para duas séries de comparações, uma vez que outro tipo de conjunto de dados (com 98% variáveis categóricas ou 56% variáveis binárias) é mais similar do que uma matriz binária;
  - os casos (conjuntos de dados) mais dissimilares são geralmente comuns, para alvos do mesmo tipo, sendo frequentemente análises com granularidade acesso.
- Quando o conjunto de dados alvo possui granularidade (ao nível do) acesso:
  - os casos (conjuntos de dados) mais semelhantes têm também granularidade acesso;
  - os casos mais dissimilares são geralmente os mesmos, para alvos do mesmo tipo.
- Quando o conjunto de dados alvo possui granularidade sessão ou outra (visitante) e não é uma matriz binária:
  - não existe um padrão de similitude claro, pois os atributos recolhidos a esses níveis são muito variados. Verifica-se alguma propensão de aumento do nível de similaridade para casos (conjuntos de dados) inseridos no mesmo contexto, apesar de a comparação não conter nenhum atributo a este respeito, mas sim em virtude do surgimento de mais características afins (e.g. dimensionalidade). Esta é precisamente a situação dos casos (conjuntos de dados) do contexto KddCup, que se evidenciam dos restantes, por serem mais similares entre si e, também, por fazerem parte, quase sempre, dos 8 a 10 casos mais dissimilares dos restantes;
  - os casos mais dissimilares possuem sobretudo a granularidade acesso.

As tendências assinaladas evidenciam padrões gerais úteis de discriminação de conjuntos de dados, com referência a factores determinantes para as abordagens de DM que podem ser levadas a cabo.

Prosseguindo com o teste T2, cujas condições de persecução são resumidas na Figura 48, agora a comparação entre variáveis decorre para as pertencentes aos esquemas de mineração, atingindo o valor de similitude 1 apenas para conjuntos de dados e esquemas de mineração comuns. Novamente, este resultado apenas é útil para fins de validação e, portanto, tal como anteriormente e no âmbito das circunstâncias que se pretende simular, as instâncias referentes à comparação de um conjunto de dados com o próprio foram ignoradas na análise seguinte.

Após a execução do teste T2 e considerando os resultados do teste anterior (T1), verificam-se as seguintes tendências:

- Os valores extremos de similitude intensificam-se muito ligeiramente, sendo maiores para os casos mais semelhantes e menores para os casos menos similares.
- As tendências identificadas no teste T1 mantêm-se, designadamente para matrizes binárias e por tipos de granularidade. Adicionalmente, em 23 casos, apenas se regista uma alteração no caso mais similar (ou o segundo mais semelhante, incluindo as instâncias ignoradas).

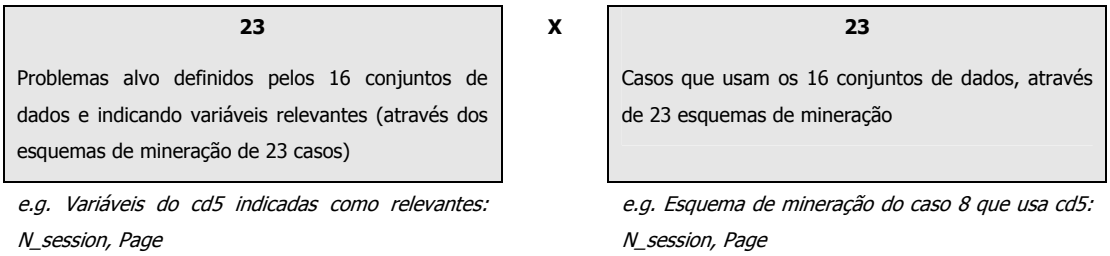


Figura 48 – Condições de realização do teste T2

Em suma, tal como se previa, as diferenças entre os dois testes realizados (T1 e T2) não são acentuadas. Esta constatação deve-se ao facto de o resultado da comparação entre as variáveis de pares de conjuntos de dados ser agregado num descritor único que, só por si, não pode produzir efeitos muito significativos. Porém, este comportamento por omissão, nomeadamente o nível de proeminência deste factor pode ser ajustado, alterando a importância relativa de tal descritor. De facto, este ajuste deveria ser executado automaticamente, pelo menos, quando o utilizador selecciona um subconjunto de variáveis relevantes, com vista a tornar este requisito significativo.

O teste T2 é ainda de grande utilidade, quando confrontado com os resultados do teste que se segue (T3), cujas condições de execução são mostradas na Figura 49. Tal comparação permite aferir a influência relativa dos vários tipos de descritores de problemas, designadamente dos (14) descritores de caracterização de dados, cobrindo também a especificação de variáveis relevantes (1 descritor), e outras formas de requisitos explícitos (9 descritores das dimensões P e T). Ao contrário da caracterização de dados, tipicamente os requisitos explícitos somente são especificados parcialmente. No entanto, dada a magnitude das variantes de elaboração deste teste, optou-se pela utilização das descrições de problemas existentes, em torno de todos os descritores com valores não nulos. Para os 23 casos em causa, tal significa a especificação de valores para todos os (24) atributos. Procurou-se deste modo diversificar os alvos, mas esta

simulação é mais irrealista do que nos restantes testes, os quais representam situações mais plausíveis de submissão de problemas alvo.

23	**X**	**23**
Problemas alvo definidos pelos requisitos implícitos e explícitos dos casos seleccionados		Casos seleccionados

Figura 49 – Condições de realização do teste T3

No teste T3 o caso mais semelhante é sempre o próprio e com valor de similitude 1, comprovando a correcção da estimativa de similaridade, de acordo com o critério de reflexividade que se impôs na selecção de medidas de semelhança para atributos de cardinalidade múltipla. Tal como anteriormente, apenas se vai considerar os restantes valores de similitude, em particular o segundo caso mais próximo. A existência de casos baseados no mesmo conjunto de dados passa a ser plausível no âmbito deste teste, uma vez que já não se está a simular a submissão de um conjunto de dados.

Quando a tendências gerais verificadas no teste T3, relativamente aos testes anteriores, evidenciam-se as seguintes:

- Os níveis de similaridade, inclusive os valores extremos, atenuam-se, com a introdução dos novos descritores, especialmente em relação aos resultados do teste T2.
- Os níveis de similitude entre casos baseados em matrizes binárias ou com granularidade acesso continuam a ser expressivamente maiores, reflectindo a importância relativa da caracterização de dados. Contudo, os restantes descritores também afectam os níveis de semelhança. Por exemplo, já são cinco os casos baseados em matrizes de binárias em que os casos mais similares não são apenas matrizes binárias (enquanto antes eram dois).

No que respeita a resultados concretos, para o caso mais próximo (excluindo o próprio), dos 23 casos e comparativamente com o teste T2 (para este efeito, ignorando apenas o caso alvo): em 9 casos (39,1%), o caso mais próximo altera-se, retractando a influência dos demais descritores; em 14 casos (60,8%) o caso mais próximo mantém-se. Entre os 14 casos para os quais o caso mais próximo não se altera, 8 correspondem a quatro pares de processos baseados num conjunto de dados comum e, por conseguinte, com (14) valores de descritores idênticos. No entanto, também existem outros casos nas mesmas circunstâncias (i.e. baseados no mesmo conjunto de dados) que possuem casos baseados em conjuntos de dados distintos como os mais semelhantes. Constata-se

ainda que os descritores acrescentados no teste T3 também contribuem, nalgumas comparações, com mais aspectos afins, justificando algumas das situações para as quais não decorrem mudanças do caso mais similar.

Relativamente à análise dos resultados concretos do teste T3, em função da sua correspondência com as expectativas, na maior parte das comparações os casos mais similares fazem todo o sentido, relacionando casos baseados em conjuntos de dados com propriedades semelhantes e análises com propósitos e aplicações afins, mas não necessariamente coincidentes. Os exemplos neste âmbito incluem os processos que recorrem aos mesmos métodos ou a métodos e abordagens de DM com resultados e aplicações, de certo modo, equivalentes, tais como mineração de padrões sequências, regras de associação e agrupamento de páginas, em conformidade com a exemplificação da secção 5.3.1.

Uma situação adversa a salientar advém das circunstâncias que basearam a criação da base de casos preliminar e surge particularmente em torno dos casos do contexto KddCup. Os conjuntos de dados deste contexto são diferentes dos restantes, possuindo muitas variáveis peculiares, algumas das quais são comuns aos seus conjuntos de dados. Questiona-se se a heterogeneidade das fontes poderá prevalecer, em detrimento das propriedades dos dados, em termos da aplicação de diferentes métodos e abordagens de DM, as quais poderão constituir uma espécie de assinatura que interessa verdadeiramente captar. No âmbito da exploração e teste do sistema nas condições idealizadas, acredita-se que seria possível atenuar muitas das diferenças, que agora se manifestam na comparação de conjuntos de dados, uma vez que estas podem dever-se aos seus contextos distintos, nomeadamente ao tipo de abstracções que é preservado ao nível de cada organização ou que depende do tipo de sítio Web. No seguimento da atenuação deste efeito, outros tipos de diferenças mais relevantes podem ser evidenciadas, tais como a designada assinatura dos métodos e abordagens de mineração. Outra contribuição neste sentido seria o aumento da importância relativa da comparação com variáveis envolvidas nas etapas de modelação.

Os testes reportados possuem ainda duas limitações proeminentes. A base de casos ainda é diminuta, não representando o espaço do domínio. Estes testes também não reproduzem com fidelidade o funcionamento do sistema SPM, dado que omitem o passo de reutilização, no qual os casos são combinados para construir planos de mineração, de forma a viabilizar a recomendação de métodos, ao invés de casos concretos.

Apesar de limitações como as mencionadas, os testes conduzidos até ao momento apontam no sentido da efectividade do SPM. O sistema consegue discriminar os conjuntos de dados mais similares e dissimilares, em função dos descritores propostos e representativos de características

relevantes para a escolha de métodos e abordagens de DM. Os restantes descritores também traduzem factores influentes e os seus efeitos são retractados na recuperação de casos, contribuindo para estabelecer a ponte entre requisitos de análises e métodos e abordagens de DM apropriados. Como os metadados de caracterização de dados estão em maioria, em relação a outros atributos de descrição problema, por omissão, a importância relativa destes descritores é maior. Este comportamento por omissão vai ao encontro do pretendido, sendo considerado um bom resultado. Na realidade, as características dos dados são sempre um factor crucial e preditivo da adequação dos métodos, uma vez que as propriedades e os pressupostos dos modelos de DM, e, mesmo, outros factores, como, por exemplo, as metas de análise, frequentemente, exigem dados com características específicas. Em acréscimo, proporcionam-se vários mecanismos para alterar o comportamento por omissão do sistema e para refinar a descrição do problema, mais concretamente, critérios de filtragem exacta ou compulsiva, especificação de níveis de importância específicos e a exclusão de descritores irrelevantes, por omissão do seu preenchimento. No entanto, conforme já foi referido, o sistema tem vindo a ser testado recorrendo a uma amostra preliminar e limitada de casos de aplicação de WUM.

6.2.2 Aprendizagem

A abordagem semi-automática de aquisição de casos de aplicação de WUM permite capturar grande parte da descrição de um caso e das abstracções subjacentes, sem o esforço do utilizador, proporcionando um suporte abrangente para documentar processos de WUM, cobrindo a especificação de aspectos como:

- Sequências de actividades, abarcando etapas de transformação e de modelação, juntamente com os modelos aplicados, variáveis implicadas, configurações de parâmetros, ferramentas usadas, explicações e justificações.
- Contexto, fontes, autores, conhecimento prévio e derivado de processos, assim como classificações de processos e resultados, obtidas a partir de situações pragmáticas e sob critérios de avaliação, vulgarmente usados na área.
- Problemas definidos através dos aspectos mais influentes, contemplando caracterizações de dados, ao nível geral e das variáveis individuais, e abstracções relacionadas com os problemas a resolver, as quais podem ser estabelecidas, de acordo com as necessidades da organização.

O esquema da base de casos reflecte a diversidade de situações práticas com as quais se lidou e que influenciaram a sua concepção, com o intuito de captar mais informação acerca da envolvente dos casos que povoaram a sua versão preliminar. Mais uma vez, os esforços envidados no tratamento de casos de estudo reais foram recompensados, especialmente no que respeita às condições de exploração idealizadas, uma vez que muitas organizações e os seus sítios Web possuem diversas facetas. Este esquema viabilizou o registo dos principais aspectos desses casos. Porém, no contexto hipotético da exploração do sistema, no âmbito de diversas organizações, este esquema seria redutor, dado que não preserva características específicas dos sítios Web.

A aquisição automática de conhecimento foi usada com sucesso para todos os casos para os quais foi possível gerar ficheiros PMML, tendo sido complementada por interacção explícita, mas com um esforço francamente reduzido. Esta abordagem revelou-se eficaz, não existindo limitações a assinalar, para além das que já se referiu, quanto aos itens que o padrão PMML abrange. As actividades de transformação de dados são uma dessas limitações. Em acréscimo, o SPM apenas suporta a descrição de operações de transformação simples. Na realidade, tal como já se referiu, a ideia era partir de fontes de dados pré-processados e com qualidade, abarcando apenas a definição de operações simples. De facto, esta visão é a que faz mais sentido, pois não se pode esperar que os utilizadores desenvolvam individualmente actividades de pré-processamento ou de transformação de grande envergadura, sob pena de comprometer ou inviabilizar o estudo dos dados. Estas actividades têm de ser centralizadas e unificadas. De qualquer forma abre-se espaço para a descrição das operações que permitem mapear os dados existentes nas fontes para os efectivamente usados, nem que seja sob a forma de uma descrição textual. Uma descrição exaustiva e precisa destas operações, em qualquer dos cenários, encontra-se, correntemente, fora do âmbito deste trabalho. O melhoramento do suporte destas descrições não deixa, no entanto, de ser oportuno, quando encarado na perspectiva apontada.

Por último, refere-se o carácter generalista das metas de análise e áreas de aplicação definidas correntemente. Tal situação impede que o sistema seja mais específico na recuperação de casos, apesar de ser adequada nas circunstâncias actuais de um repositório de casos preliminar. Esta limitação pode ser facilmente ultrapassada, acrescentando metas de análise e áreas de aplicação igualmente mais específicas. Deste modo, seria possível atingir um nível de abstracção maior e, eventualmente, alcançar o verdadeiro potencial do SPM, na sua vertente preponderante de simplificação de complexidade, especialmente no que concerne a utilizadores sem conhecimentos profundos na área.

6.3 Contribuições do Trabalho

O carácter do problema investigado suscitou muitas ideias e despoletou vários tipos de iniciativas, conjugando resultados de pesquisas ao longo de diversas áreas e o recurso a diferentes tipos de tecnologias. O objectivo inicial de simplificação de exercícios de WUM não é fácil de alcançar. Esta pretensão é dificultada, sobretudo, pela subjectividade e modo iterativo e interactivo, pelo qual estes exercícios são levados a cabo, para não falar das características adversas dos variados factores que influenciam a selecção de métodos, bem como da sua sobreposição, quanto aos problemas que podem solucionar, igualmente, verificada no sentido inverso. A adopção de uma orientação mais voltada para retirar benefícios da perspectiva prática da exploração de WUM, focando também os aspectos humanos e organizacionais, foi preponderante para o trabalho concretizado. O objectivo inicial evoluiu, ganhou novos contornos e materializou-se num sistema de assistência a processos de WUM, fundamentado na gestão e reutilização de conhecimento ao nível organizacional, com base no paradigma CBR e, ainda, implementado por um protótipo, com uma arquitectura cliente servidor e baseado no ambiente Web.

A relevância do sistema SPM não se coloca em termos da sua viabilidade comercial, mas sim como um veículo para comunicar e testar ideias conceptuais. Trata-se efectivamente, de um protótipo e não de um sistema CBR em funcionamento. Nunca foi meta deste projecto de investigação produzir uma ferramenta comercial viável, nem mesmo introduzir avanços no domínio CBR. A ambição deste trabalho sempre foi resolver ou, pelo menos, atenuar um problema concreto e o CBR surgiu como um rumo promissor para ir ao encontro desse ensejo. A meta primária de investigação traduziu-se em desenvolver uma solução para combater o desafio da selecção de métodos de DM, especificamente em processos de WUM, e, no seu seguimento, demonstrar como tal solução poderia ser implementada, como uma ferramenta de suporte à decisão e acessível num ambiente organizacional.

Acredita-se que o trabalho desenvolvido nesta dissertação inclui contributos válidos e a diversos níveis, pois representa uma abordagem original de apoio à decisão em WUM e fornece uma solução para o problema estudado, englobando resultados de variados tipos de actividades, desde a investigação em torno de várias áreas, à condução de estudos experimentais e execução de muitas análises de dados, passando pela idealização de uma estratégia para abordar o problema em questão e pela concepção de um sistema para o tratar, culminando com a implementação de um protótipo demonstrativo de tal sistema. Entre estas contribuições, evidenciam-se as que se passa a enumerar, dispostas segundo a ordem mais conveniente da sua exposição:

- **Sistematização de factores úteis no apoio à selecção de métodos em WUM** – Apesar dos vectores essenciais e de alguns dos elementos constituintes de tais factores se inspirarem em trabalhos relacionados, incorporou-se ideias novas na determinação e estruturação desses elementos. A caracterização de dados é simples, mas abrange um conjunto alargado de funções de DM e considera, não só, atributos genéricos, como também, atributos específicos, significativos no âmbito da WUM. Também se acrescentou uma dimensão de categorização de tipos de problemas de WUM, útil para abstrair a descrição da categoria de análise pretendida e para classificar processos. Já a experiência acerca de actividades de KDD foi estendida, para interligar elementos descritivos da sua aplicação, envolvente, avaliação e utilidade.
- **Desenvolvimento de modelos de representação de conhecimento** – A representação de casos é uma das questões cruciais num sistema CBR. O meta-modelo proposto, para esse efeito, faculta uma série de componentes, a fim de permitir que os diferentes aspectos envolvidos em exercícios de WUM sejam descritos, segundo múltiplas perspectivas, complementares entre si. Os modelos desenvolvidos no estudo, para os restantes contentores de conhecimento, organizam outras formas de conhecimento típicas ou mais específicas deste domínio. Neste ponto, a contribuição primordial do trabalho reside na estruturação da representação de conhecimento, quer ao nível geral da base de conhecimento, como dos seus contentores individuais, em particular, da base de casos. Um dos aspectos a destacar é a conciliação de três tipos de orientações: princípios oriundos do domínio CBR; ideias consolidadas acerca dos tipos de contentores e elementos constituintes imprescindíveis; e, especialmente, vocabulário e indicações, retiradas a partir do padrão PMML, estrategicamente adoptados para uniformizar a diversidade verificada no domínio geral da KDD ou DM e para fundamentar a modelação de conhecimento. Outro aspecto relevante é a separação do registo dos atributos pertencentes aos descritores de problemas, assim como das suas propriedades e papéis que estes desempenham, num único local do sistema, no sentido de viabilizar a parametrização dos mecanismos de recuperação e reutilização.
- **Concepção de um sistema para aplicar CBR na assistência a WUM** – O sistema concebido durante o estudo modela a abordagem de assistência defendida, integrando os blocos de construção ajustados a cenários de WUM, de acordo com os requisitos estabelecidos, e associando esses blocos a componentes e passos CBR típicos. O desenho resultante proporciona o contexto, especifica a funcionalidade geral e levanta as questões que têm de ser tratadas na implementação do sistema. Neste tópico a contribuição

fundamental surge na forma como se dividem e interagem os módulos do sistema, adaptando o ciclo CBR convencional ao problema investigado.

- **Simplificação da formulação de problemas de WUM** – O trabalho desenvolvido pautou-se pela grande ênfase depositada nesta simplificação e distingue-se, particularmente, pela viabilização de uma descrição de alto nível da tarefa de mineração pretendida. Este efeito é conseguido evitando a tradução de requisitos para alvos técnicos e expressando-os em termos de abstracções próximas do problema real. Para além da congregação de vários meios para apoiar a especificação de problemas, recorrendo a tais abstracções, possibilita-se a definição personalizada do seu conteúdo, consoante os requisitos da organização. A contribuição a assinalar possuiu maior importância ao nível conceptual. Pretende-se transcender descrições prescritas e técnicas do que a actividade de WUM é, para começar a pensar mais em função do que esta oferece e pode fazer na prática. Sugere-se, portanto, que a utilidade pragmática e as características gerais dos resultados da aplicação de WUM se tornem a fundação, não só para entender, mas também para mediar o relacionamento entre os requisitos de análises e os métodos de DM.
- **Recuperação genérica e extensível para uma formulação multi-relacional** – A recuperação é uma funcionalidade nevrálgica de qualquer sistema CBR, sendo representativa dos principais contributos desta dissertação. A primeira contribuição consiste no modelo de similaridade e algoritmos de recuperação concebidos, para lidar com uma representação de casos multi-relacional. A segunda contribuição corresponde às medidas de similitude desenvolvidas, nomeadamente, aquelas que confrontam atributos de cardinalidade múltipla, de forma a reflectir a semântica de semelhança almejada. Esta contribuição é original, pois tais medidas foram derivadas, na sequência de estudos experimentais e através da conjugação de ideias provenientes de medidas propostas em distintas linhas de investigação, tendo gerado resultados consideravelmente melhores, em termos dos requisitos estipulados e da semântica pretendida. A última contribuição neste contexto prende-se com o facto de se ter implementado um mecanismo de recuperação genérico e extensível, inicialmente baseado numa formulação proposicional, mas que evoluiu para suportar uma estrutura multi-relacional, mantendo as mesmas características. Este mecanismo é genérico e extensível porque pode ser utilizado em qualquer área de aplicação e ajustado em termos dos principais vectores de variabilidade, essencialmente, em virtude da sua independência relativamente ao modelo de domínio, tipo de armazenamento dos casos, atributos descritores de problemas e funções de similaridade

em uso. Uma das vantagens decorrentes, com maior importância no âmbito do presente trabalho e da evolução que se antevê para o mesmo, é a facilidade da incorporação de novos atributos descritores de problemas e funções de similaridade, para optimização futura do protótipo.

- **Reutilização de casos para recomendação de planos de WUM** – A reutilização foi abordada neste trabalho sem considerar a adaptação, propriamente dita e em toda a sua extensão, ao problema corrente, mas tentando ir mais além da mera recuperação de casos, para contemplar, explicitamente, a recomendação de planos de WUM. A recomendação dos métodos que solucionaram os problemas reveste-se de maior utilidade, em virtude de preparar o processo de adaptação. Adicionalmente, a avaliação que acompanha os planos reporta-se aos métodos de DM, calculando os valores médios de critérios, para os casos agrupados, com vista a atenuar efeitos adversos, tais como, a sua subjectividade. Os valores individuais de avaliação de cada caso encontram-se, igualmente, acessíveis através dos seus detalhes. Outra contribuição neste âmbito advém da optimização da apresentação de soluções, com o objectivo de maximizar a sua utilidade, sob o ponto de vista do analista, simultaneamente, para dois propósitos diferentes: diversidade de categorias de soluções, de acordo com o nível de apropriação e critérios de avaliação importantes para o analista; variedade de casos de cada solução, com o intuito de conceder exemplos variados de uma solução, com eventual interesse para o analista.
- **Abordagem semi-automática de aquisição de conhecimento** – A fase de aprendizagem permite que um sistema CBR melhore a sua competência, especialmente, por intermédio da recolha de novos casos e, preferencialmente, de forma automatizada, para minimizar os esforços subjacentes. No sistema SPM, um dos aspectos inovadores desta fase é o tratamento de documentos PMML, como uma fonte de conhecimento, capaz de viabilizar a aquisição automática de parte da descrição de processos de WUM, para derivar conhecimento e povoar os componentes dos respectivos contentores. Para além da questão crucial da eficiência, decorrente da redução drástica de esforços humanos, as instâncias adicionadas a tais componentes baseiam-se em conceitos normalizados e aceites. Outra vertente desta fase relaciona-se com as acções que promovem a reutilização futura de tal conhecimento. Nesta vertente, salienta-se a catalogação dos casos, em função das soluções técnicas que estes contêm. A identificação dos diferentes tipos de soluções é automática, para evitar a necessidade da intervenção de especialistas. O catálogo de tipos de soluções é construído dinamicamente, atendendo aos métodos de

DM aplicados em cada processo e na determinação das distintas combinações de métodos observadas. Quanto à vertente associada de manutenção, estabeleceu-se a fundação para aplicar políticas de acesso e, também, de auxílio à actuação do administrador do sistema, na selecção do conteúdo da base de casos, através de vários estados possíveis, que retractam o estatuto atingido pelos casos.

- **Implementação de um protótipo em ambiente Web** – Neste contexto, a arquitectura cliente servidor desenvolvida e a organização dos serviços implementados são os aspectos mais proeminentes, uma vez que materializam a solução proposta para o problema investigado. Entre as características mais significativas do protótipo, evidenciam-se, a separação do modelo de domínio dos mecanismos de raciocínio, a integração e encapsulamento de fontes heterogéneas de dados e conhecimento, a extensibilidade em função dos vectores fulcrais de variabilidade e as potencialidades acrescidas da combinação de CBR com tecnologias Web.
- **Criação de um repositório de casos de aplicação de WUM** – A preparação de uma amostra de casos, baseada em dados, necessidades e análises reais, reproduzindo processos de WUM desenvolvidos em competições de KDD e outros publicados em artigos científicos, permite criar um repositório diversificado e útil para vários fins. Para além de sustentar a avaliação do sistema, tal repositório pode ser usado como uma base de casos preliminar, na fase experimental do SPM, no seio de uma organização concreta. O crescimento e evolução previstos, no curto prazo, para este repositório permitirão, ainda, que o mesmo assuma valor e possa constituir uma fonte consolidada de exemplos variados da área, passível de ser disponibilizada de diferentes formas. A inexistência de um repositório de tal natureza e com o nível de detalhe que este possui justifica a pertinência desta contribuição.

As contribuições referidas e as principais ideias discutidas ao longo desta dissertação foram publicadas em alguns artigos. Nesta dissertação aborda-se, com maior profundidade, as questões tratadas nesses artigos, em conjunto com outros aspectos, não abrangidos em tais publicações, e com referência ao estado corrente do sistema e do correspondente protótipo que, entretanto, têm evoluído e sido alvo de melhoramentos. A lista seguinte contém os artigos que esta dissertação originou, ordenados pela data de publicação, bem como os respectivos assuntos neles debatidos:

- *C. Wanzeller e O. Belo. "Selecting Clickstream Data Mining Plans", Proceedings of the 2nd Workshop Data Gadgets 2005, integrada nas X Jornadas sobre Ingeniería del Software y Bases de Datos (JISBD' 2005), 27-41. Granada, Espanha. Setembro, 2005.* – Apresenta e justifica as ideias defendidas para abordar o problema estudado, as estratégias adoptadas

na concepção e implementação do sistema proposto e os desafios enfrentados na realização dessas actividades, reflectindo uma fase preliminar do trabalho desenvolvido e do estado da implementação do protótipo.

- *C. Wanzeller e O. Belo. "Towards a More Suitable System for Web Usage Mining", Proceedings of the International Conference on Knowledge Management in Organizations (KMO' 2006), 103-108. Maribor, Eslovénia. Junho, 2006.* – Foca os aspectos da recolha e organização da aquisição de conhecimento, no contexto de um sistema CBR de assistência a processos de WUM, em termos do desenvolvimento e actuação do sistema. Para além de definir o modelo de representação de casos e de salientar as diferentes vertentes da exploração da especificação PMML, descreve e exemplifica a abordagem semi-automática de captura de casos de aplicação de WUM.
- *C. Wanzeller e O. Belo. "Recomendação de Planos de Mineração de Dados para Clickstreams", Actas da 1ª Conferência Ibérica de Sistemas e Tecnologias da Informação (CISTI' 2006), 2, 249-264. Esposende, Portugal. Junho, 2006.* – Estabelece os módulos fundamentais do SPM e as suas funcionalidades e interacções, facultando uma visão geral do sistema sugerido, da sua finalidade e do seu modo de actuação.
- *C. Wanzeller e O. Belo. "Selecting Clickstream Data Mining Plans Using a Case-Based Reasoning Application", Proceedings of the 7th International Conference on Data, Text and Web Mining and their Business Applications and Management Information Engineering (DMIE' 2006), 223-232. Praga, República Checa. Julho, 2006.* – Proporciona uma perspectiva geral do sistema proposto, dando ênfase às questões da aplicação de CBR e da adaptação do respectivo ciclo no âmbito de processos de WUM. Adicionalmente, representa uma evolução do sistema, na medida em que acrescenta um novo passo ao ciclo CBR adoptado, determinando os blocos de construção existentes correntemente.
- *C. Wanzeller e O. Belo. "Improving Effectiveness on Clickstream Data Mining", Proceedings of the 6th Industrial Conference on Data Mining (ICDM' 2006), 161-175. Leipzig, Alemanha. Julho, 2006.* – Demonstra a exploração do sistema proposto, no apoio a decisões envolvidas em problemas de WUM, concedendo, particularmente, uma visão da vertente de resolução de problemas do SPM. Para esse efeito, aponta dois cenários de uso do SPM, indica os passos do cenário mais relevante e exemplifica a forma como o sistema actua, especialmente em termos da descrição de problemas e da recomendação de soluções, à luz de uma situação concreta.
- *C. Wanzeller e O. Belo. "Clickstream Data Mining Assistance: A Case-Based Reasoning Task Model", Proceedings of the 1st International Conference on Software and Data*

Technologies (ICSOFT' 2006). Setúbal, Portugal. Setembro, 2006. – Trata as questões da modelação e estruturação da implementação do sistema SPM, no que concerne à série de (sub-) tarefas que cada módulo tem de desempenhar, a fim de cumprir a sua missão.

Por último, produziu-se um relatório técnico devotado ao problema do cálculo de similaridade entre casos de aplicação de WUM, no contexto de uma representação de casos multi-relacional. Este trabalho reporta os resultados de estudos experimentais conduzidos, definindo as medidas desenvolvidas para a comparação de descritores com cardinalidade múltipla e justificando a necessidade de medidas ajustadas. Pretende-se submeter este trabalho a conferências internacionais a breve trecho.

6.4 Considerações Finais

Actualmente, as organizações procuram, com premência crescente, meios capazes de lhes facultar vantagens competitivas. Um desses meios reside na maximização do ganho que pode ser obtido a partir da exploração dos seus recursos de informação, para suporte do processo estratégico de tomada de decisão. A KDD vem corresponder a estes anseios, dada a sua capacidade para descobrir conhecimento implícito nos dados, o qual não é possível antever, nem identificar, por meio de análises mais superficiais. De facto, as capacidades de DM tendem a tornar-se parte integrante de qualquer tecnologia de informação ao nível organizacional, conforme se pode constatar facilmente, examinando as funcionalidades das ferramentas que vão sendo disponibilizadas no mercado. Também se verifica uma franca expansão no número de indivíduos, dentro das organizações, que necessitam de efectuar análises, cada vez, mais sofisticadas. Os dados mantidos em sistemas operacionais e analíticos são, hoje em dia, acedidos por uma gama muito diversificada e alargada de utilizadores, sem experiência na concretização destas actividades, graças ao uso de variados tipos de ferramentas e interfaces. Esses indivíduos enfrentam problemas de decisão reais e são os mais aptos para reconhecer os factos ou padrões que os podem ajudar a combater esses problemas. Tais competências têm, porém, de ser transferidas, para que possam ser os próprios a levar a cabo os seus estudos.

Condições envolventes, em alteração súbita, e a competição, ao nível global, introduzem uma pressão tremenda no dia a dia das organizações, exigindo uma presença efectiva na Web e atitudes mais elaboradas, para tentar alcançar o seu potencial. A KDD aplicada a dados de *clickstream*, denotada por WUM, assume, mais uma vez, um papel preponderante. A WUM tem sido o instrumento mais importante e efectivo, usado pelas organizações para traçar padrões de

comportamento dos visitantes, para os seus sítios na Internet. Trata-se, indubitavelmente, de um instrumento, verdadeiramente, capaz de ajudar as organizações a estabelecer a ponte entre dados vastíssimos de *clickstream* e conhecimento accionável, a fim de apoiar o planeamento de melhoramentos oportunos de sítios Web. No entanto, a curva de aprendizagem da WUM é um sério obstáculo, sendo importante poder contar com alguma estratégia indicativa da forma pela qual se deve proceder. É sempre pertinente poder recorrer a alguém familiarizado com a indústria e as formas, pelas quais, dados e problemas com características similares foram tratados, com sucesso, no passado. Mais do que isso, não existem razões que justifiquem o desperdício de esforços na resolução de problemas que já foram solucionados no passado.

O trabalho que se propôs e desenvolveu visa contribuir para uma exploração simplificada, mais produtiva e efectiva das potencialidades da WUM, assim como promover um uso mais sinergético de recursos organizacionais, reduzindo o tempo e o esforço, requeridos para derivar conhecimento útil, e congregando as contribuições válidas, dos vários membros da organização. Para atingir estes objectivos concebeu-se um sistema CBR, especificamente devotado à assistência ao desenvolvimento e aplicação de processos de WUM, a ser explorado em ambiente organizacional, consolidando os exemplos bem sucedidos da organização, numa memória comum. Um sistema desta natureza e com este âmbito permite defrontar os desafios no contexto de cada organização, dado que é mais difícil transpor outros exemplos de WUM para os problemas concretos no seio de um organismo.

O apoio à decisão, envolvendo exercícios de extracção de conhecimento, é um rumo de trabalho muito fértil, em virtude do contraste entre a utilidade e a facilidade da exploração da KDD, sendo, por este motivo, contemplado em muitos projectos de investigação. O paradigma CBR também é adoptado em iniciativas com finalidades afins. O trabalho realizado pode ser distinguido, do trabalho directamente relacionado, pela conjugação de vários aspectos, designadamente, a planificação ao nível de processos e incluindo diferentes tipos de etapas, a ajuda abrangendo diversas funções de DM e, sobretudo, a intenção de atingir um nível elevado de abstracção na descrição de tarefas de DM, que pode ainda ser ajustado aos requisitos particulares da organização. Além disso, o sistema proposto é dedicado, especialmente, ao âmbito específico da WUM, procurando dar uma resposta dirigida aos problemas que se levantam neste contexto. As dificuldades acrescidas da WUM são plenamente reconhecidas e, por conseguinte, o presente trabalho simboliza uma direcção de investigação mais específica. Adicionalmente, a combinação das áreas de WUM e CBR, para efeitos de assistir a resolução de problemas de análise de dados de *clickstream*, pode mesmo ser considerada uma estratégia inovadora. Tal coligação proporciona

novas formas para abordar problemas de WUM, delineadas no sistema SPM concebido e comprovadas pelo respectivo protótipo.

A implementação do protótipo interliga, essencialmente, tecnologias Web, PMML, base de dados e Java. Estas opções revelaram-se frutíferas para o cumprimento dos requisitos estipulados, tendo permitido:

- ganhar flexibilidade, com respeito à interacção com o utilizador e ao uso e acessibilidade do sistema;
- retirar vantagens de influências benéficas de padrões do domínio da DM, em termos de vocabulário da área, indicações para a estruturação da base de casos e automatização de parte da aquisição de conhecimento;
- garantir a interoperacionalidade com tecnologias de gestão de dados ao nível da organização, de maneira a aproveitar as capacidades, vulgarmente, disponíveis e a maximizar o seu potencial;
- usufruir da portabilidade, características de orientação por objectos, flexibilidade e vantagens no desenvolvimento de aplicações Web do ambiente Java.

Em acréscimo, a implementação norteou-se pelo desígnio de assegurar a extensibilidade do sistema e o tratamento, tanto quanto possível, independente das vertentes de armazenamento dos casos, modelo de domínio e mecanismos de raciocínio. Estas orientações foram proeminentes, tendo viabilizado e simplificado o melhoramento progressivo das sucessivas versões do protótipo do sistema.

O protótipo implementado demonstra a viabilidade da abordagem conceptual formulada, para suportar a aquisição e disseminação de conhecimento relativo a iniciativas de WUM. Ao longo do desenvolvimento e implementação do sistema considerou-se fontes humanas da instituição e outros recursos, designadamente, fontes de dados organizacionais e documentos PMML, representativos de conhecimento já extraído e baseados no padrão do domínio da DM mais aceite e suportado. Estes recursos são usados pelo sistema para recolher novos casos, de acordo com uma abordagem semi-automática de aquisição de conhecimento. Esta abordagem mostrou-se efectiva, uma vez que consegue captar uma parte significativa do conhecimento acerca de experiências de WUM, em consonância com noções universais e evitando esforços humanos. É um facto inegável que uma base de casos que não seja constantemente alimentada, com novo conhecimento, tende, a breve trecho, a ficar desactualizada e a comprometer a competência do sistema para resolver problemas. Contar, exclusivamente, com os utilizadores para a condução manual desta tarefa, tipicamente mais trabalhosa no contexto da WUM, também é, nitidamente,

uma situação insustentável. Neste sentido, a abordagem semi-automática referida afigura-se preponderante, mesmo sendo apenas uma solução parcial, já que a automatização exaustiva desta actividade não é exequível.

A abordagem defendida também parece ser viável e útil para o propósito da transferência de conhecimento de projectos prévios para novos. Apesar do estado corrente do protótipo ainda estar longe de ser um sistema completamente funcional, já presta ajuda com valor. O SPM pode ser explorado como uma ferramenta tradicional de pesquisa num repositório, em função de critérios de filtragem exacta, bem como para comparar experiências retidas nesse repositório e, principalmente, para tirar proveito dos mecanismos CBR de recuperação baseada em similaridade, para obter soluções plausíveis para novos problemas, não necessariamente coincidentes com aqueles que se encontram registados. É justamente esta capacidade, de dar resposta a problemas não antecipados, que reveste os sistemas CBR de grandes potencialidades. O impacto deste potencial também é notório, quando o paradigma é usado como um meio ao serviço do apoio à decisão, indo, precisamente, ao encontro do tipo de assistência que mais se adequa em KDD e WUM. A capacidade de recomendar planos de mineração plausíveis e justificados, explicados e detalhados, com referência a exemplos concretos, torna-se muito conveniente neste âmbito, dado que os analistas necessitam de ser auxiliados e não substituídos. De facto, a produção de um processo de KDD completo e bem sucedido, de forma automática, é uma situação irrealista e a omissão de pormenores demonstrativos é demasiado redutora. A conjugação de CBR com tecnologias Web incrementa, ainda mais, este potencial, permitindo conceder uma ferramenta de recuperação inteligente e apontando o caminho para a transferência mais alargada de *know-how*.

O SPM baseia-se em abstracções relacionadas com os problemas a resolver, significando que pode servir as necessidades de utilizadores inexperientes, que tencionam lidar com problemas de WUM, em cooperação com o sistema, ou, simplesmente, reconhecer aplicações promissoras da WUM e obter acesso a descobertas e a indicações de formas de exploração da WUM, a partir dos contributos e da experiência de outros membros da organização. Os utilizadores terão as suas tarefas de desenvolvimento de processos de WUM simplificadas e alcançarão, eventualmente, uma série mais ampla e não antevista de resultados úteis de exploração. O sistema poderá reduzir drasticamente a taxa de esforço de especialistas nesta área, tão específica, podendo, ainda, ser útil, mesmo para estes peritos, caso estejam interessados em relembrar e reutilizar, de forma mais consistente, soluções prévias efectivas, ao invés de resolverem os problemas do início.

Em suma, apesar das imensas potencialidades da WUM, subsiste o problema reconhecido da dificuldade na exploração de dados de *clickstream*, o qual também se sentiu na prática, ao longo

das inúmeras análises realizadas, e que tende a abranger um número, cada vez, maior de analistas informais. Constatou-se, por experiência própria, que o acesso a descrições de processos de WUM bem sucedidos, envolvendo propósitos ou dados afins, foi a orientação mais frutífera para lidar com os desafios enfrentados. O sistema SPM materializa, justamente, as estratégias de auxílio à resolução de problemas de WUM que se revelaram mais efectivas para ultrapassar tais desafios, reproduzindo o procedimento que os especialistas seguem, por inerência, graças ao *know-how* que detêm, e que todos os utilizadores também podem adoptar, de forma mais sistemática e contando com uma memória de conhecimento adquirido ampliada. Concluindo, acredita-se que tal sistema pode ser um instrumento verdadeiramente profícuo para a criação, partilha e reutilização de conhecimento, ao nível da organização, não só acerca da experiência de desenvolvimento e aplicação de processos de WUM, como também das respectivas descobertas.

6.5 Trabalho Futuro

O trabalho descrito neste documento é um passo no sentido de fornecer uma solução para assistir o desenvolvimento e aplicação de processos de WUM, mas não esgota todas as ideias e actividades que se vislumbrou, nem as novas questões que foram surgindo no seu decurso. Envidaram-se muitos esforços na implementação do sistema, mas o resultado prático destes esforços é, apenas, um protótipo. Continuam a existir muitas possibilidades de optimização, a considerar em implementações futuras, para ser viável suportar decisões reais e colocar o sistema em funcionamento efectivo. Nesta secção indicam-se actividades que já estão em curso ou vão ser realizadas muito brevemente, enumeram-se sugestões de uma série de melhoramentos primordiais e delineiam-se as direcções mais promissoras de investigação futura. De imediato, identificam-se três actividades prementes, cuja concretização complementa o trabalho corrente e se enquadra numa etapa inicial do planeamento relativo ao futuro:

- **Construção de novos casos de aplicação de WUM e consolidação da base de casos preliminar** – Correntemente, já se está a elaborar novos casos de aplicação de WUM. Estas actividades vão prosseguir, com grande ênfase, englobando processos, cada vez, mais complexos, contando, igualmente, com dados reais, disponíveis publicamente na Internet, e com as descrições de análises bem sucedidas, executadas sobre os mesmos. Conforme já foi dito, esta opção tem a desvantagem de ser mais morosa, mas permite superar o problema da subjectividade do sucesso dos projectos e adquirir exemplos mais diversos, sustentando a continuação do povoamento da base de casos preliminar.

- **Avaliação mais exaustiva e sistemática do protótipo** – O protótipo do sistema SPM tem vindo a ser testado com referência a uma base de casos com representatividade e dimensionalidade limitadas. Na sequência da actividade precedente, torna-se possível e desejável proceder a uma avaliação mais exaustiva e sistemática da efectividade do sistema. Esta avaliação também pode decorrer em condições de maior exigência, resultante do alargamento da base de casos.
- **Preparação de um caso de estudo acerca de uma organização alvo** – O recurso a exemplos de WUM reais e públicos permite povoar a base de casos preliminar, satisfazendo parte dos pressupostos do sistema SPM. Porém, alguns pressupostos não são abrangidos e não se explora plenamente as potencialidades de abstracção de complexidade do sistema. Para este efeito torna-se necessário recolher e catalogar os problemas mais específicos que surgem em organizações particulares. Assim sendo, seria oportuno construir um caso de estudo baseado em dados e necessidades de uma organização alvo concreta, recorrendo à base de casos preliminar para viabilizar o funcionamento inicial do sistema. No seguimento desta actividade, torna-se viável conduzir testes experimentais com utilizadores finais, para avaliar a aplicação do sistema e validar, mais objectivamente, a abordagem subjacente.

No âmbito do melhoramento do protótipo são múltiplas as actividades que podem ser reportadas. Entre essas actividades, elegeu-se um subconjunto, dando prioridade às que mais foram discutidas, implícita ou explicitamente, nos vários capítulos desta dissertação:

- **Suporte de bases de dados analíticas como fontes de dados** – Ao longo deste documento evidenciou-se o papel e importância de sistemas de DW e, mais especificamente, DhW, na exploração de dados de *clickstream*. Estes sistemas foram indicados como uma das fontes mais relevantes, senão a mais preponderante, da WUM. Mencionou-se ainda, por diversas vezes, as vantagens das capacidades OLAP. Não se quis reduzir o destaque de tais fontes e capacidades, meramente com base no facto de o seu suporte ter sido relegado para segundo plano. Presentemente, o sistema integra fontes de dados baseadas em ficheiros simples e bases de dados convencionais. Começou-se por estas fontes, seguindo a ordem mais lógica e por uma questão de conveniência da sua instanciação, para as séries de dados disponíveis, mas garantiu-se a extensibilidade do sistema, comprovada pelo tratamento de, pelo menos, dois tipos de fontes distintos e da extracção consistente de metadados a partir das mesmas. A implementação exaustiva para as diversas modalidades dos diferentes eixos de variabilidade do sistema não é uma meta desta dissertação, dada a dimensão do trabalho, as inevitáveis restrições temporais e,

como tal, a exigência de estabelecer prioridades. No entanto, é necessário admitir que esta vertente do trabalho fazia parte do plano inicial e continua a ser interessante e mais desafiante que as restantes. Por esta razão, cumpre apresentar a mesma como uma actividade a conduzir no futuro.

- **Extensão da descrição de actividades de transformação** – O carácter superficial do tratamento de operações de transformação foi mencionado, várias vezes, neste documento. Se por uma lado, a indispensabilidade da aplicação deste tipo de operações foi incentivadora de acções conducentes a um suporte consentâneo para as documentar, por outro lado, a inexistência de meios para automatizar a sua aquisição é dissuadora destas acções, nas circunstância actuais, pelo menos no que concerne à especificação PMML. Por conseguinte, esta actividade foi remetida para segundo plano e acabou por ser adiada.
- **Optimização da interacção com o utilizador** – Apesar das limitações já apontadas ao ambiente de interacção, o tempo e esforços envidados no seu desenvolvimento foram consideráveis, uma vez que a sua funcionalidade e robustez se revelaram essenciais para atender a vários requisitos básicos de funcionamento. Todavia, para além da sobrecarga do conteúdo das páginas Web concebidas, assume-se que os termos usados nem sempre são os mais intuitivos para utilizadores novatos, os quais formam o público alvo primordial do sistema. O melhoramento a diversos níveis do ambiente de interacção é uma tarefa que se antevê no futuro, particularmente, no âmbito da utilização do sistema, no seio de alguma organização.
- **Disponibilização de funcionalidades de configuração do sistema** – No contexto desta actividade, pretende-se, por um lado, elaborar mais o contentor de configuração da base de conhecimento, de modo a gerir, centralizadamente, mais variantes do comportamento do sistema, como, por exemplo, os critérios de determinação de diferentes tipos de soluções técnicas. Por outro lado, torna-se oportuno acrescentar um ponto de acesso a operações de configuração do sistema, para permitir que se actue, de forma mais flexível, sobre propriedades já parametrizadas e outras que venham a ser alvo do mesmo tipo de tratamento.

Para além das actividades abordadas, existem muitas questões em aberto, definindo tópicos a explorar em trabalho subsequente, seguindo a mesma linha condutora. De seguida destacam-se algumas possibilidades de pesquisa que se pretende investigar num futuro próximo:

- **Refinamento do modelo conceptual de representação de casos** – A extensão do modelo conceptual de representação de casos é vista como uma actividade constante, sempre que tal se justifique. A preparação de casos mais complexos e a realização de

testes mais exaustivos poderão proporcionar indicações a este respeito, nomeadamente, relativamente a atributos descritores de problemas. Esta possibilidade foi contemplada durante a implementação, assegurando a extensibilidade do sistema em termos deste vector, propiciando, portanto, atitudes voltadas para a identificação de novos atributos com interesse e a sua incorporação no sistema. Outra forma de extensão deste modelo, a ponderar, reside em captar novos elementos, capazes de retractar mais características do sítio Web.

- **Indexação de casos e optimização da atribuição de ponderações** – A indexação de casos é um factor importante para o desempenho de um sistema CBR, principalmente quando a base de casos é volumosa e em condições de operação do sistema, nas quais é exigida eficiência. Conforme já foi referido, ainda não se implementou nenhum mecanismo de indexação, em parte, em virtude de o sistema não se encontrar nas circunstâncias anteriores. Um dos primeiros passos neste sentido seria a mineração dos dados dos casos registados, para tentar identificar os atributos descritores mais relevantes e com maior poder discriminatório, e, consequentemente, aqueles que devem ser considerados no esquema de indexação. Contudo, não se dispõe, no momento, de um número suficiente de casos, capaz de representar o espaço do domínio. A indexação de uma base de casos também não é uma incumbência simples, pelo contrário. Existem muitas iniciativas de investigação especialmente devotadas a esta questão, sendo necessário proceder a um estudo aprofundado da mesma. Pelo exposto, uma das direcções de trabalho futuro coloca-se em torno desta questão desafiante e envolve a optimização do módulo de recuperação e, por inerência, a correspondente actualização do módulo de retenção. Outra actividade relacionada e planeada consiste em melhorar a atribuição de valores por omissão às ponderações dos atributos. Em primeira instância, esta atribuição poderia basear-se nos resultados da aplicação de DM aos casos existentes. Porém, novamente, abre-se espaço para iniciativas de investigação, pois este tópico também não é trivial.
- **Apoio à manutenção da base de casos** – A manutenção da base de casos é outro aspecto complicado que ainda não foi objecto de um tratamento verdadeiramente consentâneo, sendo, também, mais significativo no contexto do uso real do sistema e do crescimento contínuo da base de casos. Para lidar com esta problemática, está previsto, no curto prazo, atender a factores como o nível de utilidade e representatividade dos casos, com base em estatísticas de utilização e grau de relevância dos casos. O nível de relevância dos casos poderá, eventualmente, ser aferido em função de aspectos, como os diferentes tipos de soluções que estes contêm. Os factores citados podem ser usados para

fundamentar a concepção de mecanismos mais ágeis, capazes de ajudar o administrador do sistema, na realização da tarefa de manutenção da base de conhecimento. De qualquer forma, ainda é necessário estudar detalhadamente esta questão, dado que a mesma dá corpo a uma linha activa de investigação.

- **Estudo de exequibilidade de mecanismos de adaptação** – O papel do sistema SPM é coadjuvar o utilizador, continuando a decisão humana a ser preponderante. No entanto, é pertinente equacionar o estudo de meios que possam oferecer sugestões adicionais, para um determinado plano escolhido, em consonância com a situação específica do problema actual. Este estudo constitui outra direcção de trabalho futuro plausível.
- **Integração com outras ferramentas** – A interligação do sistema SPM com outros tipos de ferramentas é uma ideia interessante, designadamente, no que concerne a ferramentas de KDD. A possibilidade de executar processos de WUM registados no sistema, aplicando-os aos mesmos ou a novos dados, seria didáctica e útil. Esta possibilidade permitiria, ainda, esbater a distância, em relação ao ambiente de operação, de recursos imprescindíveis neste âmbito e conferir uma nova dimensão de utilidade prática ao SPM.

As actividades enumeradas reforçam nitidamente o facto de que o trabalho desenvolvido, até ao momento, é apenas um passo de uma solução possível, para resolver ou minorar o problema da complexidade do desenvolvimento e aplicação de processos de WUM. Também se torna patente que a continuação deste trabalho e a evolução do sistema de assistência à WUM proposto requerem novos esforços de conjugação de resultados de diversas linhas activas de investigação e, certamente, darão lugar a novas ideias e questões em aberto. No entanto, este passo inicial poderá abrir novas perspectivas à condução de iniciativas de WUM, pois dá ênfase a factores preponderantes, como a abstracção de complexidade, acessibilidade a conhecimento e minimização de esforços humanos, os quais contribuem para colocar as potencialidades da WUM ao serviço das organizações, incluindo, particularmente, os seus membros sem experiência na área. A concretização das actividades indicadas e das demais que venham a surgir, permitirá encetar outros passos, de um novo rumo que se afigura promissor, no sentido de uma exploração mais eficaz e expedita da WUM, para fazer face a solicitações de apoio sustentado e atempado dos agentes de decisão das organizações.

Bibliografia

[Aamodt e Plaza 94] A. Aamodt e E. Plaza. "Case-Based Reasoning: Foundational Issues, Methodological Variations and Systems Approaches". Artificial Intelligence Communications, IOS Press, 7(1), 39-59. Março, 1994.

[Aamodt 95] A. Aamodt. "Knowledge Acquisition and Learning by Experience - The Role of Case Specific Knowledge". Chapter in Machine Learning and Knowledge Acquisition - Integrated Approaches, Academic Press, 197-245. 1995.

[Aha 92] D. W. Aha. "Generalizing from Case Studies: A Case Study". Proceedings of the 9th International Workshop on Machine Learning (ML'92), 1-10. Aberdeen, Scotland, UK. Julho, 1992.

[Althoff 01] K.-D. Althoff. "Case-Based Reasoning". Chapter in Handbook on Software Engineering and Knowledge Engineering, World Scientific, 549-588. 2001.

[Ansari et al. 01] S. Ansari, R. Kohavi, L. Mason e Z. Zheng. "Integrating E-Commerce and Data Mining: Architecture and Challenges". Proceedings of the IEEE International Conference on Data Mining (ICDM'01), 27-34. San Jose, CA, USA. Novembro-Dezembro, 2001.

[Berendt et al. 02] B. Berendt, B. Mobasher e M. Spiliopoulou. "Web Usage Mining for E-business Applications". Tutorial at European Conference on Machine Learning/Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'02). Helsinki, Finland. Agosto, 2002.

[Bergmann e Stahl 98] R. Bergmann e A. Stahl. "Similarity Measures for Object-Oriented Case Representations". Proceedings of the 4th European Workshop on Case-Based Reasoning (EWCBR'98), 1488, 25-36. Dublin, Ireland. Setembro, 1998.

[Bergmann 01] R. Bergmann. "Highlights of the European INRECA Projects". Proceedings of the 4th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR'01), 1-15. Vancouver, Canada. Julho-Agosto, 2001.

[Bernstein e Provost 01] A. Bernstein e F. Provost. "An Intelligent Assistant for the Knowledge Discovery Process". Proceedings of the IJCAI-01 Workshop on Wrappers for Performance Enhancement in KDD. Seattle, WA, USA. Agosto, 2001.

[Bohnebeck et al. 98] U. Bohnebeck, T. Horváth e S. Wrobel. "Term Comparisons in First-Order Similarity Measures". Proceedings of the 8th International Workshop on Inductive Logic Programming (ILP'98), 1446, 65-79. Madison, Wisconsin, USA. Julho, 1998.

[Borges e Levene 99] José Borges e Mark Levene. "Data Mining of User Navigation Patterns". Proceedings of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), 31-36. San Diego, CA, USA. Agosto, 1999.

[Borges 00] J. Borges. "A Data Mining Model to Capture User Web Navigation Patterns". Ph. D. thesis, Department of Computer Science, University College London, London University. Julho, 2000.

[Brazdil et al. 94] P. Brazdil, J. Gama e B. Henery. "Characterizing the Applicability of Classification Algorithms Using Meta-Level Learning". Proceedings of the European Conference on Machine Learning (ECML'94), 83-102. Catania, Italy. Abril, 1994.

[Büchner et al. 99] A. Büchner, M. Mulvenna, S. Anand e J. Hughes. "An Internet-enabled Knowledge Discovery Process". Proceedings of the 9th International Database Conference, 13-27. Hong Kong. Julho, 1999.

[Chaudhuri e Dayal 97] S. Chaudhuri e U. Dayal. "An Overview of Data Warehouse and Olap Technology". ACM SIGMOD Record, 65-74. Março, 1997.

[Codd et al. 93] E. F. Codd, S. Codd e C. Salley. "Providing OLAP (On-Line Analytical Processing) to User-Analyst: An IT Mandate". Technical Report, E. F. Codd and Associates, IBM. San Jose, CA, USA. 1993.

[Cooley et al. 97] R. Cooley, B. Mobasher e J. Srivastava. "Web Mining: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 558-567. Newport Beach, CA, USA. Novembro, 1997.

[Cooley et al. 99] R. Cooley, B. Mobasher e J. Srivastava. "Data Preparation for Mining World Wide Web Browsing Patterns". Journal of Knowledge and Information Systems, Springer, 1(1), 5-32. Fevereiro, 1999.

[Cooley 00] R. Cooley. "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data". Ph. D. thesis, Department of Computer Science, Faculty of the Graduate School, University of Minnesota, Minneapolis, USA. Maio, 2000.

[Craw et al. 92] S. Craw, D. Sleeman, N. Graner, M. Rissakis e S. Sharma. "Consultant: Providing Advice for the Machine Learning Toolbox". Proceedings of the Research and Development in Expert Systems IX Cup, 5-23. Cambridge, UK. 1992.

[Duda et al. 01] R. Duda, P. Hart e D. Stork. "Pattern Classification", Chapter 10 - Unsupervised Learning and Clustering, 2nd ed., John Willey and Sons. 2001.

[Eiter e Mannila 97] T. Eiter e H. Mannila. "Distance Measures for Point Sets and their Computation". Acta Informatica, 34(2), 109–133. 1997.

[Emde e Wettschereck 96] W. Emde e D. Wettschereck. "Relational Instance-based Learning". Proceedings of the 13th International Conference on Machine Learning (ICML'96), 122-130. Bari, Italy. Julho, 1996.

[Engels 96] R. Engels. "Planning Tasks for Knowledge Discovery in Databases; Performing Task–Oriented User–Guidance". Proceedings of the 2nd International Conference on Knowledge Discovery in Databases (KDD'96), 170-175. Portland, Oregon, USA. 1996.

[Engels et al. 97] R. Engels, G. Lindner e R. Studer. "A Guided Tour Through the Data Mining Jungle". Proceedings of the 3rd International Conference on Knowledge Discovery in Databases (KDD'97), 14–17. Newport Beach, CA, USA. Agosto, 1997.

[Engels et al. 98] R. Engels, G. Lindner e R. Studer. "Providing User Support for Developing Knowledge Discovery Applications: A midterm report". The German AI Journal Künstliche Intelligenz, 12(1), 40–45. 1998.

[Engels e Theusinger 98] R. Engels e C. Theusinger. "Using a Data Metric for Offering Preprocessing Advice in Data Mining Applications". Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), 430-434. Brighton, UK. Agosto, 1998.

[Fayyad et al. 96] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth, "From Data Mining to Knowledge Discovery: An Overview". Advances in Knowledge Discovery and Data Mining. AAAI Press/MIT Press, 1-34. 1996.

[Flach et al. 98] P. Flach, C. Giraud-Carrier e J. Lloyd. "Strongly Typed Inductive Concept Learning". Proceedings of the 8th International Workshop on Inductive Logic Programming (ILP'98), 1446, 185-194. Madison, Wisconsin, USA. Julho, 1998.

[Gama e Brazdil 95] J. Gama e P. Brazdil. "Characterization of Classification Algorithms". Proceedings of the 7th Portuguese Conference on Artificial Intelligence (EPIA'95), 189-200. Funchal, Portugal. Outubro, 1995.

[Gamma et al. 95] E. Gamma, R. Helm, R. Jonhson e J. Vlissides. "Design Patterns: Elements of Reusable Object Oriented Software". Addison Wesley. 1995.

[Gomory et al. 99] S. Gomory, R. Hoch, J. Lee, M. Podlaseck e E. Schonberg. "E-Commerce Intelligence: Measuring, Analyzing, and Reporting on Merchandising Effectiveness of Online Stores". IBM T. J. Watson Research Center. Julho, 1999.

[Gregori et al. 05] V. Gregori, C. Ramírez, J. Orallo e M. Quintana. "A Survey of (pseudo-distance) Functions for Structured-Data". Proceedings of the III Taller Nacional de Minería de Datos y Aprendizaje (TAMIDA'2005), Congreso Español de Informática (CEDI'2005), 233-242. Granada, Espanha. Setembro, 2005.

[Grossman et al. 02] L. R. Grossman, M. F. Hornick e G. Meyer. "Data Mining Standards Initiatives". Communications of the ACM, 45(8), 59-61. Agosto, 2002.

[Han 98] J. Han. "Towards On-Line Analytical Mining in Large Databases", SIGMOD Record, 27(1), 97-107. Março, 1998.

[Hilario e Kalousis 01] M. Hilario e A. Kalousis. "Fusion of Meta-Knowledge and Meta-Data for Case-Based Model Selection". Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2001), 180-191. Freiburg, Germany. Setembro, 2001.

[Hilario e Kalousis 03] M. Hilario e A. Kalousis. "Representational Issues in Meta-Learning". Proceedings of the 20th International Conference on Machine Learning (ICML'03), 313-320. Washington, DC, USA. Agosto, 2003.

[Hu e Cercone 04] X. Hu e N. Cercone. "A Data Warehouse/OLAP Framework for Web Usage Mining and Business Intelligence Reporting". International Journal of Intelligence Systems, 19(7), 567-584. Julho, 2004.

[Jespersen et al. 02] S. Jespersen, J. Thorhauge e T. Pedersen. "A Hybrid Approach To Web Usage Mining". Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, 4(6), 73-82. Aix-en-Provence, France. Setembro, 2002.

[Kimball e Merz 00] R. Kimball e R. Merz. "The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse". Wiley Computer Publishing, John Wiley & Sons, Inc. 2000.

[Kirsten e Wrobel 98] M. Kirsten e S. Wrobel. "Relational Distance Based Clustering". Proceedings of the 8th International Workshop on Inductive Logic Programming (ILP'98), 1446, 261-270. Madison, Wisconsin, USA. Julho, 1998.

[Kirsten et al 01] M. Kirsten, S. Wrobel e T. Horvath. "Relational Data Mining". Chapter 9 - Distance Based Approaches to Relational Learning and Clustering, Springer, Berlin, 212-232. 2001.

[Kodratoff et al. 92] Y. Kodratoff, D. Sleeman, M. Uszynski, K. Causse and S. Craw. "Building a Machine Learning Toolbox". Enhancing the Knowledge Engineering Process, North-Holland, Elsevier Science Publishers, 81-108. 1992.

[Kohavi et al. 00] R. Kohavi, C. Brodley, B. Frasca, L. Mason e Z. Zheng. "KDD-Cup 2000 Organizers' Report: Peeling the Onion". SIGKDD Explorations, 2(2), 86-98. Dezembro, 2000.

[Kohavi 01] R. Kohavi. "Mining E-Commerce Data: The Good, the Bad, and the Ugly". Invited paper at Industrial Track, 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01). San Francisco, USA. Agosto, 2001.

[Kohavi e Provost 01] R. Kohavi e F. Provost. "Applications of Data Mining to Electronic Commerce". International Journal of Data Mining and Knowledge Discovery, Special Issue on E-Commerce and Data Mining, Kluwer Academic Publishers, 5(1-2), 5-10. Janeiro-Abril, 2001.

[Kolodner 91] J. Kolodner. "Improving Human Decision Making Through Case-Based Decision Aiding". AI Magazine, American Association for Artificial Intelligence, 12(2), 52-68. Menlo Park, CA, USA. 1991.

[Kolodner 93] J. Kolodner. "Case-Based Reasoning". Morgan Kaufman Publishers Inc., San Mateo, CA, USA, 1993.

[Kotásek e Zendulka 02] P. Kotásek e J. Zendulka. "Describing the Data Mining Process with DMSL". Proceedings of the 6th East European Conference on Advances in Database and Information Systems (ADBIS'02), 2 Research Communications, 131-140. Bratislava, Slovakia. Setembro, 2002.

[Koutri et al. 05] M. Koutri, N. Avouris e S. Daskalaki. "A Survey on Web Usage Mining Techniques for Web-Based Adaptive Hypermedia Systems". Chapter in Adaptable and Adaptive Hypermedia Systems, Idea Publishing Inc., Hershey. 2005.

[Kosala e Blockeel 00] R. Kosala e H. Blockeel. "Web Mining Research: A Survey". SIGKDD Explorations, 2(1), 1-15. Julho, 2000.

[Laer 02] W. Laer. "From Propositional to First Order Logic in Machine Learning and Data Mining". Ph. D thesis, Department of Computer Science, Katholieke Universiteit Leuven, Belgium. Junho, 2002.

[Lindner e Studer 99] C. Lindner e R. Studer. "AST: Support for Algorithm Selection with a CBR Approach". Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'99), 418-423. Prague, Czech Republic. Setembro, 1999.

[Mantaras et al. 05] R. Mantaras, D. McSherry, D. Bridge, D. Leake, B. Smyth, S. Craw, B. Faltings, M. Maher, M. Cox, K. Forbus, M. Keane, A. Aamodt e I. Watson. "Retrieval, Reuse, Revision, and Retention in Case-Based Reasoning". The Knowledge Engineering Review, Cambridge University Press, 20(3), 215-240. Setembro, 2005.

[Michie et al. 94] D. Michie, D. Spiegelhalter e C. Taylor. "Machine Learning, Neural and Statistical Classification". Series in Artificial Intelligence, Ellis Horwood. 1994.

[Mobasher et al. 96] B. Mobasher, N. Jain, E.-H. Han e J. Srivastava. "Web Mining: Pattern Discovery from World Wide Web Transactions". Technical Report TR-96050, Department of Computer Science, University of Minnesota. Setembro, 1996.

[Mobasher et al. 00] B. Mobasher, R. Cooley e J. Srivastava. "Automatic Personalization Based On Web Usage Mining". Communication of ACM, 43(8), 142-151. Agosto, 2000.

[Mobasher et al. 01] B. Mobasher, B. Berendt, B. e M. Spiliopoulou. "KDD for Personalization". Tutorial at European Conference on Machine Learning/Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'01). Freiburg, Germany. Setembro, 2001.

[Mobasher et al. 02] B. Mobasher, H. Dai, T. Luo e M. Nakagawa. "Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks". Proceedings of the IEEE International Conference on Data Mining (ICDM'02), 669-672. Maebashi City, Japan. Dezembro, 2002.

[Mobasher 04] B. Mobasher. "Web Usage Mining and Personalization". Chapter in Practical Handbook of Internet Computing. CRC Press. 2004.

[Morik e Scholz 04] K. Morik e M. Scholz. "The MiningMart Approach to Knowledge Discovery in Databases". Chapter in Intelligent Technologies for Information Analysis. Springer, 47-65. 2004.

[Pei et al. 00] J. Pei, J. Han, B. Mortazavi-Asl e H. Zhu. "Mining Access Patterns Efficiently from Web Logs". Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'00), 396-407. Kyoto, Japan. Abril, 2000.

[Ramon 02] J. Ramon. "Clustering and Instance Based Learning in First Order Logic". Ph. D thesis, Department of Computer Science, Katholieke Universiteit Leuven, Belgium. Outubro, 2002.

[Richter 95] M. Richter. "The Knowledge Contained in Similarity Measures". Invited Talk at the 1st International Conference on Case-Based Reasoning (ICCBR'95), Lecture Notes in Artificial Intelligence, Springer Verlag. Sesimbra, Portugal. Outubro, 1995.

[Spiliopoulou e Faulstich 98] M. Spiliopoulou e L. Faulstich. "WUM: A Tool for Web Utilization Analysis". Proceedings of the EDBT Workshop on the Web and Databases (WebDB'98), 184–203. Valencia, Spain. Março, 1998.

[Srivastava et al. 00] J. Srivastava, R. Cooley, M. Deshpande e P.-N. Tan. "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". SIGKDD Explorations, 1(2), 12-23. Janeiro, 2000.

[Todorovski e Dzeroski 99] L. Todorovski e S. Dzeroski. "Experiments in Meta-Level Learning with ILP". Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'99), 98-106. Prague, Czech Republic. Setembro, 1999.

[Verdenius e Engels 97] F. Verdenius e R. Engels. "A Process Model for Developing Inductive Applications". Proceedings of the 7th Belgian-Dutch Conference on Machine Learning (BENELEARN-97), 119-128. Tilburg, Netherlands. Outubro, 1997.

[Watson e Marir 94] I. Watson e F. Marir. "Case-Based Reasoning: A Review". The Knowledge Engineering Review, 9(4), 327-354. 1994

[Witten e Frank 05] I. Witten e E. Frank. "Data Mining: Practical Machine Learning Tools and Techniques", 2nd ed., Morgan Kaufmann, San Francisco, 2005.

[Zaïane et al. 98] O. Zaïane, M. Xin e J. Han. "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs". Proceedings of the Advances in Digital Libraries Conference (ADL'98), 12-29. Santa Barbara, CA, USA. Abril, 1998.

[Zaïane 99] O. Zaïane. "Resource and Knowledge Discovery from the Internet and Multimedia Repositories". Ph. D thesis, School of Computing Science, Simon Fraser University, Vancouver, BC, Canada. Março, 1999.

Referências WWW

[WWW1] http://www.w3.org/

O consórcio *World Wide Web* (W3C) é um órgão internacional que congrega as contribuições de múltiplas instituições, pessoas e público em geral, com a missão de conduzir a Web ao seu potencial máximo, através do desenvolvimento de tecnologias interoperáveis e da promoção de fóruns abertos, para incentivar a sua evolução sustentada. Entre as várias normas, cuja especificação é tutelada por este consórcio, evidenciam-se as seguintes, em virtude de as mesmas serem extensivamente usadas nesta dissertação: a linguagem de marcação de hipertexto HTML (*HyperText Markup Language*), utilizada na produção e apresentação de documentos na WWW; a meta-linguagem de marcação de dados XML (*eXtensible Markup Language*), um formato universal de marcação de texto, que faculta uma forma flexível e portável de representação e transferência de informação; o modelo de objectos de documentos estruturados DOM (*Document Object Model*), uma interface de programação, fundamentada na representação em árvore do documento, para permitir que as aplicações acedam e alterem dinamicamente o conteúdo, estrutura e estilo de um documento; folhas de estilo CSS (*Cascading Style Sheets*), um mecanismo para descrever e controlar separadamente regras de formatação de documentos estruturados. O sítio Web do consórcio W3C é, indubitavelmente, um local fulcral de acesso a um manancial de informação acerca destes padrões, nas suas diferentes versões, e de uma diversidade de muitas outras tecnologias relacionadas.

[WWW2] http://www.dmg.org/

A *Predictive Model Markup Language* (PMML) é uma linguagem normalizada, baseada na XML, para descrever e partilhar modelos estatísticos e de mineração de dados entre aplicações aderentes à mesma. A linguagem PMML é mantida pelo *Data Mining Group*

(DMG), um consórcio voltado para a especificação de padrões na área da mineração de dados e formado por vários organismos com um papel preponderante nesta área e noutros domínios afins. O sítio WWW da DMG disponibiliza informação relativa à norma, nas suas diversas versões, e, também outros recursos, como uma lista de produtos aderentes e exemplos de documentos PMML, juntamente com os conjuntos de dados utilizados nas respectivas análises.

[WWW3] http://www.metal-kdd.org/

O projecto *MetaL* teve como objectivo primordial o desenvolvimento de métodos e ferramentas para proporcionar suporte a utilizadores de tecnologias de aprendizagem automática e de mineração de dados. No sítio Web devotado ao projecto encontra-se a informação e os artigos científicos mais relevantes, sendo, ainda, possível executar a ferramenta de assistência a mineração de dados, desenvolvida no seu âmbito, denominada *Data Mining Assistant*. Em termos gerais, esta ferramenta sugere uma lista de algoritmos, ordenada pelo nível de desempenho e tempo de execução previstos, a partir de um conjunto de dados alvo e para tarefas de classificação ou regressão.

[WWW4] http://www.spss.com/clementine/

Este sítio concede informação respeitante à ferramenta de mineração de dados SPSS *Clementine*. Trata-se de uma ferramenta que cobre todas as fases do processo de descoberta de conhecimento, oferecendo uma série muito vasta de funcionalidades de modelação, transformação de dados e análise de resultados, bem como opções variadas para importação e exportação de dados e resultados. A interacção gráfica para programação visual, baseada na peculiar metáfora de nós e ligações, é uma das características mais atractivas deste *software*, pois permite criar aplicações bastante complexas, de forma simples, interligando diferentes tipos de operações num fluxo. Outro ponto a seu favor, com importância particular para a presente dissertação, é a possibilidade de exportar modelos para documentos PMML, complementada pela capacidade de importação no sentido inverso.

[WWW5] http://www.spss.com/spss/

O *Statistical Package for the Social Sciences* (SPSS) é uma ferramenta de tratamento e análise estatística de dados. O SPSS é especialmente vocacionado e muito utilizado na exploração de dados recolhidos em questionários, nas ciências, no marketing e nos

estudos demográficos, económicos e sociais. Para além de um ambiente gráfico intuitivo, semelhante ao de folhas de cálculo, e de operações de importação, transferência e exportação de dados, entre outras, este produto integra uma gama considerável de técnicas estatísticas e permite exportar o resultado da aplicação de algumas destas técnicas para o formato PMML. O endereço WWW indicado corresponde ao sítio oficial deste produto, contendo documentação sobre o mesmo.

[WWW6] http://www.omg.org/uml

A *Unified Modeling Language* (UML) é uma linguagem de especificação e modelação da estrutura, comportamento e arquitectura de aplicações, contemplando também outros aspectos, nomeadamente, processos de negócio e estrutura de dados. Trata-se de uma linguagem, essencialmente, visual, baseada em vários tipos de diagramas, de forma a retractar as múltiplas perspectivas sobre as quais versa. A linguagem UML possui amplo reconhecimento e consagrou-se como uma norma, ao ser adoptada pelo consórcio internacional *Object Management Group* (OMG), o qual se dedica à especificação de padrões da indústria de *software*. Esta página do sítio da OMG é uma fonte alargada de recursos sobre UML, referentes a versões prévias e correntes da especificação, manuais introdutórios, artigos e listas de ferramentas conformes, entre outros elementos.

[WWW7] http://www.java.sun.com/

O sítio *da Sun* fornece uma elevada quantidade de informação acerca da linguagem orientada a objectos Java. É possível encontrar uma vasta gama de recursos, que vão desde as versões mais actuais de Java, passando por uma extensa documentação, até grupos de discussão e investigação.

[WWW8] http://java.sun.com/javase/

A partir desta página do sítio da *Sun* é possível obter informação detalhada acerca do ambiente específico da plataforma *Java 2 Platform Standard Edition* (J2SE). O J2SE é a plataforma Java mais utilizada e abrangente, podendo mesmo dizer-se que é a principal, já que as restantes plataformas têm a sua fundação nesta.

[WWW9] http://www.javasoft.com/products/jdbc/

O JDBC, vulgarmente interpretado como o acrónimo de *Java Database Connectivity*, é um padrão industrial para acesso de dados universal, a partir da linguagem de programação

Java. O JDBC traduz-se numa interface de programação, particularmente utilizada para interagir com bases de dados, de forma independente dos SGBD concretos subjacentes, recorrendo a controladores específicos que implementam essa interface. O sítio da *Sun* reúne documentação muito exaustiva sobre JDBC, incluindo do seu uso e suporte pela indústria.

[WWW10] http://java.sun.com/webservices/jaxp/

A *Java API for XML Processing* (JAXP) veio uniformizar a forma como se criam instâncias de analisadores sintácticos (*parsers*) de XML em aplicações Java, subscrevendo e permitindo o uso de várias interfaces de programação normalizadas, designadamente DOM e SAX (*Simple API for XML)*, com o intuito de garantir a independência dessas aplicações em relação às implementações concretas de *parsers* de diferentes vendedores. O endereço WWW referido corresponde à secção do sítio da *Sun* dedicado a esta API, contendo muita documentação de apoio de grande utilidade.

[WWW11] http://xerces.apache.org/

O *Xerces* é um processador ou analisador sintáctico (*parser*) de XML, inicialmente desenvolvido no âmbito do projecto *Apache XML Project*, mas que adquiriu um estatuto próprio, tornando-se num projecto de topo da fundação *Apache*. A *Apache Software Foundation* (ASF) é uma organização representativa de uma comunidade descentralizada de desenvolvedores de *software*. Os projectos *Apache* são caracterizados por um processo colaborativo e consensual, e, sobretudo, pela distribuição dos produtos resultantes, como *software* livre, sob uma licença aberta e pragmática. A versão Java da implementação do *Xerces* é muito usada, sendo recomendada e incorporada como o *parser* de XML por omissão da API JAXP. O sítio Web deste projecto é, na sua essência, um repositório de código fonte e de documentação relativa ao projecto e ao seu produto alvo.

[WWW12] http://java.sun.com/products/jsp/

Nesta página do sítio da *Sun* encontra-se informação diversa acerca da especificação *Java Server Pages* (JSP), desde documentação de referência, a material didáctico e interligações com a comunidade de utilizadores. Do ponto de vista operacional, a finalidade das tecnologias de *Servlets* e JSP é permitir a criação dinâmica de conteúdos WWW, de forma mais simplificada e produtiva, no que diz respeito a JSP.

[WWW13] http://tomcat.apache.org/

O sítio Web indicado é devotado ao *software Tomcat*, da responsabilidade da fundação *Apache*, envolvendo, neste caso, o membro mais ilustre do projecto *Jakarta*, cujo objectivo é desenvolver soluções com código aberto e baseadas na plataforma Java. O *Tomcat* viabiliza a execução de aplicações Web, tendo, como característica fundamental, o facto de se centrar nas tecnologias *Servlet* e JSP, constituindo o contentor da implementação de referência destas tecnologias. Em termos práticos, o *Tomcat* pode ser usado isoladamente, assumindo o papel de servidor Web, ou em conjunto com outro servidor, como, por exemplo, o *Apache*, que passa, então, a atender às solicitações de páginas estáticas. Para além das sucessivas versões do *Tomcat*, este sítio fornece a respectiva documentação, informação ao nível do projecto, ligações para tópicos relacionados e ajuda, sob a forma de guias de uso e instalação.

[WWW14] http://www.eclipse.org/

O *Eclipse* é um ambiente de desenvolvimento e de integração de ferramentas, sob o auspício de uma fundação ou comunidade com nome igual. Os princípios defendidos por esta comunidade, também focados na condução dos seus projectos, permitiram dotar esta ferramenta de produtividade, desenvolvida na linguagem de programação Java, de uma série relevante de características. Nessas características, destacam-se o seu código aberto, a extensibilidade e a neutralidade, em termos, quer da plataforma, como da linguagem de programação e de vendedores. O endereço WWW apontado disponibiliza a plataforma *Eclipse* e a documentação associada.

[WWW15] http://www.sysdeo.com/eclipse/tomcatplugin

Esta página do sítio da *Sysdeo* disponibiliza o *Eclipse Tomcat Launcher Plugin*, uma extensão para a integração do servidor *Tomcat* no ambiente *Eclipse*, juntamente com informação descritiva da sua funcionalidade e de indicações de instalação.

[WWW16] http://www.microsoft.com/frontpage/

O sítio Web oficial do *Microsoft Office Frontpage* concede informação acerca desta ferramenta de edição de HTML e de administração de sítios Web.

Anexo

A. Exemplificação do Cálculo de Similaridade entre Casos de Aplicação de WUM

Neste anexo procede-se à exemplificação do procedimento adoptado na determinação do nível de similaridade entre casos de aplicação de WUM, comparando as descrições de um problema alvo e de um caso prévio, registado na base de casos do sistema SPM. Para este efeito, concilia-se aspectos abordados em diferentes secções da presente dissertação. A questão do cálculo de similitude foi tratada na secção 4.5, no âmbito da descrição do módulo de recuperação. Mais concretamente, as medidas de semelhança, aplicadas em tal procedimento e neste anexo, foram definidas na secção 4.5.3 (Medidas de Similaridade). Já as descrições de problemas e os respectivos valores, usados nesta exemplificação, reportam-se à secção 5.3.1 (Exemplificação da Resolução de Problemas). Utiliza-se as propriedades do problema exemplo introduzido nessa secção, como o alvo, e caso 8, igualmente referido nessa secção, como o caso prévio. O caso 8 foi recuperado, no seguimento da submissão do problema exemplo ao sistema SPM, tendo sido seleccionado como o caso existente mais semelhante a este alvo.

A Tabela 12 recapitula os tipos de funções de similitude a que o módulo de recuperação do SPM recorre correntemente. As abreviaturas incluídas, na primeira coluna dessa tabela, são usadas, ao longo de toda a exemplificação, para indicar explicitamente a função de similaridade que está a ser aplicada em cada passo. Não se utiliza (nem apresenta na tabela) nenhuma função, quando o nível de semelhança entre os valores de um descritor é retractado numa matriz de similitude (MS). Nestas circunstâncias o resultado é proporcionado directamente, através de uma procura numa matriz, com base nos valores e no descritor em causa. Conforme já se mencionou, estas matrizes poderão ser simétricas (e.g. para o descritor de metas de análise) ou assimétricas (e.g. para o descritor de áreas de aplicação).

Tabela 12 – Lista de tipos de funções de similaridade em uso

Abreviatura	Descrição	Função
(G)	Média ponderada (função de similaridade global)	$Sim_{Global}(a,c) = \frac{\sum_{d=1}^{n} Sim_{Local}(a.d, c.d) * w_d}{\sum_{d=1}^{n} w_d}$
(DNM)	Distância de *Manhattan* normalizada	$Sim_{Local\,simples}(a.d, c.d) = 1 - \frac{\lvert a.d - c.d \rvert}{d_{\max} - d_{\min}}$
(DNMa)	Distância de *Manhattan* normalizada alterada	$Sim_{Local\ simples}(a.d, c.d) = \begin{cases} 1 & c.d \geq a.d \\ 1 - \frac{\lvert a.d - c.d \rvert}{d_{\max} - d_{\min}} & c.d < a.d \end{cases}$
(E)	Exacta (texto ou binário)	$Sim_{Local\ simples}(a.d, c.d) = \begin{cases} 1 & c.d = a.d \\ 0 & c.d \neq a.d \end{cases}$
(MS)	Matriz de Similaridade (simétrica ou assimétrica)	
(MM)	Média dos Máximos	$Sim_{MM}(A,B) = \frac{1}{n_A + n_B}\left(\sum_{1}^{n_A} \max_{a \in A}(sim(a,b)) + \sum_{1}^{n_B} \max_{b \in B}(sim(a,b))\right)$
(MMA)	Média dos Máximos do Alvo	$Sim_{MMA}(A,B) = \frac{1}{n_A}\sum_{1}^{n_A} \max_{a \in A}(sim(a,b)) \quad \mathrm{A} \subset Alvo, \mathrm{B} \subset Caso$

a, **c** – alvo e caso ou parte destes
***Sim*$_{LOCAL}$** – uma medida de similaridade local
a.d, **c.d** – valores do alvo e caso para o descritor d
n – número de descritores usados na comparação
***w*$_d$** – ponderação atribuída ao descritor d
***d*$_{max}$**, ***d*$_{min}$** – valores máximo e mínimo observados para o descritor d
A, **B** - dois conjuntos, tais que a ∈ A e b ∈ B
sim(a,b) - similitude entre cada par de elementos de dois conjuntos
n$_A$, **n$_B$** - cardinalidade dos conjuntos A e B

A enumeração dos descritores de problemas, suas propriedades e valores, para o alvo (**a**) e caso prévio (**c**) considerados nesta exemplificação, foi estruturada numa série de tabelas, usando, em algumas das colunas, a notação estabelecida na Tabela 12. A Tabela 13 apresenta esses elementos, no que concerne a caracterizações ou metadados ao nível geral do conjunto de dados (i.e. sem integrar as variáveis individuais), assim como os resultados de similaridade local aferidos (i.e. entre descritores). Estes resultados locais são agregados mais adiante, atendendo às ponderações dos descritores (w_d), para produzir um valor global de similitude. Nesta tabela destacam-se dois aspectos:

- a aplicação da função baseada na distância normalizada de *Manhattan* (DNM), materializada para dois descritores em notas de rodapé (da tabela);
- a ponderação do atributo de granularidade que, ao contrário de todos os outros descritores, possui o valor 5.

Tabela 13 – Exemplo de cálculo de similaridade local para descritores ao nível de conjuntos de dados

Descritor	**Tipo de valor**	**Tipo de comparação**	**Função de similaridade**	**d_{max}**	**d_{min}**	**a.d**	**c.d**	**Similaridade local**	**w_d**
Caracterização de dados genérica ao nível do conjunto de dados									
Número de registos ou linhas	Contínuo (Inteiro)	1-1	(DNM)	4179752	1	7128	1923	**a)** 0.99875	1
Número de variáveis ou colunas	Contínuo (Inteiro)	1-1	(DNM)	989819	1	8	3	0.99999	1
Percentagem de variáveis numéricas	Contínuo (Real)	1-1	(DNM)	1	0	0.375	0.66666667	**b)** 0.70833	1
Percentagem de variáveis categóricas	Contínuo (Real)	1-1	(DNM)	1	0	0.625	0.33333333	0.70833	1
Percentagem de variáveis temporais	Contínuo (Real)	1-1	(DNM)	1	0	0.125	0	0.875	1
Percentagem de variáveis binárias	Contínuo (Real)	1-1	(DNM)	1	0	0	0	1	1
Caracterização específica de WUM ao nível do conjunto de dados									
Granularidade	Categórico (Simbólico)	1-1	(E)	-	-	Acessos	Acessos	1	***5***
Tipo de identificação de visitantes	Categórico (Simbólico)	1-1	(E)	-	-	Indisponível	Indisponível	1	1
Tipo de registo de informação de visitantes	Categórico (Simbólico)	1-1	(E)	-	-	Indisponível	Indisponível	1	1
Disponibilidade de ordem de acesso	Categórico (Binário)	1-1	(E)	-	-	Sim	Sim	1	1
Disponibilidade de repetições de acesso	Categórico (Binário)	1-1	(E)	-	-	Sim	Sim	1	1
Disponibilidade de duração de acesso	Categórico (Binário)	1-1	(E)	-	-	Não	Não	1	1
Disponibilidade de data de acesso	Categórico (Binário)	1-1	(E)	-	-	Sim	Não	0	1
Disponibilidade de hora de acesso	Categórico (Binário)	1-1	(E)	-	-	Sim	Não	0	1

a) $1 - \frac{|7128 - 1923|}{4179752 - 1}$ **b)** $1 - \frac{|0.375 - 0.666(6)|}{1 - 0}$

A Tabela 14 ilustra toda a informação envolvida na estimativa da semelhança local para critérios de avaliação. Estes resultados baseiam-se sempre na distância normalizada de *Manhattan* alterada (DNMa), sendo exemplificados apenas para o descritor de "simplicidade de implementação".

Tabela 14 – Exemplo de cálculo de similaridade local para critérios de avaliação

Descritor	Tipo de valor	Tipo de comparação	Função de similaridade	d_{max}	d_{min}	a.d	c.d	Similaridade local	w_d
Critérios de avaliação									
Precisão	Ordinal	1-1	(DNMa)	5	1	5	4	0.75	1
Interpretabilidade	Ordinal	1-1	(DNMa)	5	1	5	3	0.5	1
Simplicidade de implementação	Ordinal	1-1	(DNMa)	5	1	5	4	**c)** 0.75	1
Tempo de resposta	Ordinal	1-1	(DNMa)	5	1	5	5	1	1
Exigência de recursos	Ordinal	1-1	(DNMa)	5	1	5	5	1	1

c) $c.d < a.d \left(1 - \frac{|5 - 4|}{5 - 1}\right)$

Segue-se, agora, a exemplificação do mesmo procedimento no que respeita às variáveis individuais. A Tabela 15 contém apenas dados gerais e o resultado final da comparação entre as variáveis em causa dos dois conjuntos de dados (i.e. as variáveis seleccionadas do alvo e as implicadas em etapas de modelação do caso prévio). A Figura 50 mostra os detalhes deste procedimento, executado em três fases. Na primeira fase, os valores dos descritores d' (internos) de cada par de variáveis são comparados, recorrendo às funções de similitude apropriadas. Novamente, exemplifica-se o uso da função DNM, neste caso para o atributo "número de valores distintos", das variáveis *URL* e *N_session*. Na segunda fase, os resultados locais dos descritores internos são agregados, aplicando a função de similaridade global e as ponderações desses descritores ($w_{d'}$), fornecidas na Tabela 15. Finalmente, na terceira fase, usa-se uma função de semelhança para séries de valores, designadamente a Média dos Máximos (MM), derivando, assim, o valor local final que consta na Tabela 15.

Tabela 15 – Exemplo de cálculo de similaridade local para variáveis

Descritor	Tipo de valor	Tipo de comparação	Função de similaridade	d_{max}	d_{min}	a.d e c.d	$w_{d'}$	Similaridade local	w_d
Categorização ao nível das variáveis									
Variáveis de conjunto de dados:	*Variável*	N-M'	(MM)	-	-			0.87499	1
- Categoria de variável	Categórico (Simbólico)	1-1	(MS)	-	-		1		
- Tipo de dados	Categórico (Simbólico)	1-1	(MS)	-	-		1		
- Número de valores nulos	Contínuo (Inteiro)	1-1	(DNM)	..	..		1		
- Número de valores distintos	Contínuo (Inteiro)	1-1	(DNM)	989818	1		1		

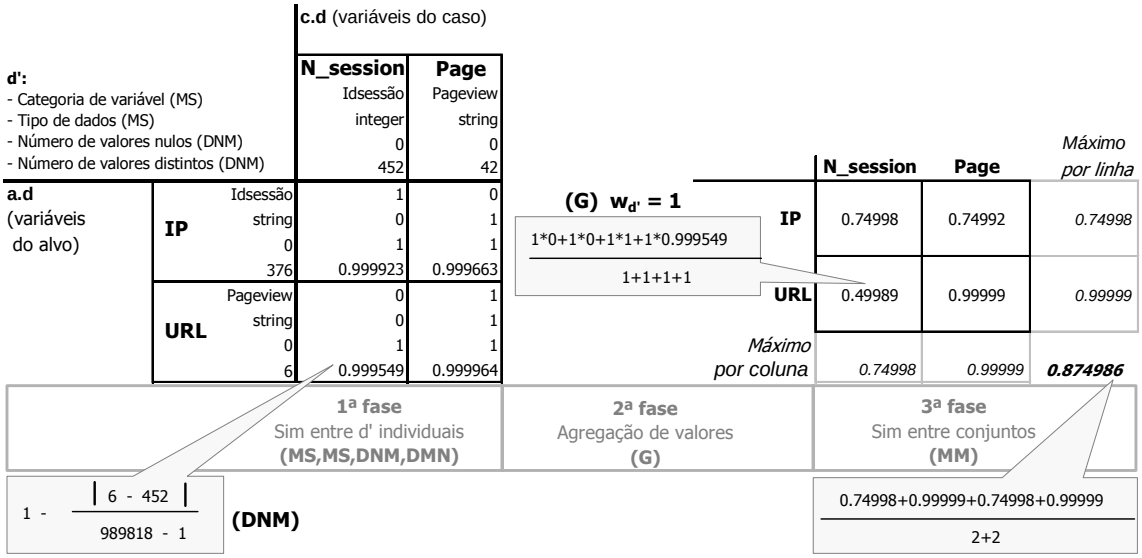


Figura 50 – Detalhes do exemplo do cálculo de similaridade local entre variáveis

A Tabela 16 e a Figura 51 reportam o procedimento de estimativa do nível de similaridade para os restantes descritores. Dois destes descritores (a "data" e "actividade") são nulos no problema alvo, não sendo, portanto, englobados nesta estimativa. Já as "metas de análise" e as "áreas de aplicação" requerem a consulta das correspondentes matrizes de similaridade, para se obter o valor de semelhança entre cada par de instâncias. Subsequentemente, estes valores são usados pela segunda medida adoptada para o confronto entre séries de valores, a Média dos Máximos do Alvo (MMA), permitindo produzir os resultados locais finais indicados na Tabela 16.

Tabela 16 – Exemplo do cálculo de similaridade para as restantes categorias de descritores

Descritor	Tipo de valor	Tipo de comparação	Função de similaridade	d_{max}	d_{min}	a.d	c.d	Similaridade local	w_d
Outras categorias									
Data do processo	Contínuo (Data)	1-1	(DNM)	..	..	nulo	..		0
Actividades:		N-1	(MMA)	-	-	nulo	..		0
- Actividade	Categórico (Simbólico)	1-1	(MS)	-	-				
Metas:		N-1	(MMA)	-	-			0.5	1
- Meta	Categórico (Simbólico)	1-1	(MS)	-	-				
Áreas de aplicação:		N-M	(MMA)	-	-			1	1
- Área de aplicação	Categórico (Simbólico)	1-1	(MS)	-	-				

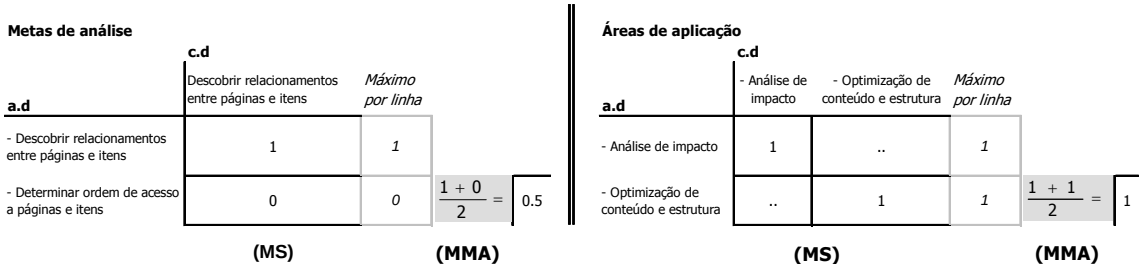


Figura 51 – Exemplo do cálculo de similaridade local entre metas e áreas de aplicação

Determinados todos os valores locais (indicados nas tabelas anteriores), pode-se proceder à sua agregação final, com base na função de similitude global (**G**), nas ponderações dos descritores (w_d) e nesses valores de similaridade local, tal como se descreve na Figura 52. Todas as ponderações possuem o valor 1, com a excepção assinalada (do atributo "granularidade"), que é 5. O resultado final coincide com o que foi gerado pelo sistema, conforme se pode constatar observando a Figura 39.

$$\frac{\begin{array}{c} 1 * \mathbf{0.99875} + 1 * \mathbf{0.99999} + 1 * \mathbf{0.70833} + 1 * \mathbf{0.70833} + 1 * \mathbf{0.875} + 1 * \mathbf{1} \\ + \mathbf{\underline{5}} * \mathbf{1} + 1 * \mathbf{1} + 1 * \mathbf{1} + 1 * \mathbf{1} + 1 * \mathbf{1} + 1 * \mathbf{1} + 1 * \mathbf{0} + 1 * \mathbf{0} \\ + 1 * \mathbf{0.75} + 1 * \mathbf{0.5} + 1 * \mathbf{0.75} + 1 * \mathbf{1} + 1 * \mathbf{1} \\ + 1 * \mathbf{0.87499} \\ + 1 * \mathbf{0.5} + 1 * \mathbf{1} \end{array}}{\begin{array}{c} 1 + 1 + 1 + 1 + 1 + 1 \\ + \mathbf{\underline{5}} + 1 + 1 + 1 + 1 + 1 + 1 + 1 \\ + 1 + 1 + 1 + 1 + 1 + 1 \\ + 1 \\ + 1 + 1 \end{array}} = 0.83328$$

Figura 52 – Exemplo da determinação do nível de similaridade global