SIMPÓSIO DOUTORAL 2006

Relatório de Desenvolvimento/Resultados do Projecto de Doutoramento

Doutorando: Ana Cristina Wanzeller Guedes de Lacerda
cwanzeller@di.estv.ipv.pt

Título da Tese: Elaboração Semi-Automática de Planos de Mineração para a Integração de Ferramentas de Data Mining em Ambientes de Data Webhousing

Orientador: Professor Doutor **ORLANDO BELO**, Universidade do Minho, Departamento de Informática
obelo@di.uminho.pt

Data Início: - Oficial 2 Abril 2003 - Efectiva 1 Janeiro 2004

Data Término: Prevista: 31 Dezembro 2006

B. Resumo

B.1 Área de Investigação e Desenvolvimento

Informática – Mineração de dados de Utilização da Web (*Web Usage Mining* – WUM).

B.2 Resumo

O desenvolvimento e a aplicação de processos de mineração de dados, particularmente relacionados com a informação relativa à utilização da Web, são normalmente de grande complexidade para utilizadores sem conhecimentos profundos nesse domínio. Uma das muitas dificuldades com que os especialistas desta área de trabalho se debatem está relacionada com a selecção de métodos de mineração a aplicar sobre um problema específico de análise de dados, com o objectivo de obter resultados úteis para um determinado fim. A principal meta do presente trabalho consiste, precisamente, em proporcionar suporte ao analista nessa tarefa de selecção, considerando a natureza do problema e as particularidades do contexto em que este se coloca. A abordagem defendida para esse efeito envolve a reutilização da experiência adquirida a partir de problemas similares que proporcionaram processos de mineração de dados bem sucedidos. Uma limitação desta abordagem decorre da indisponibilidade de descrições acerca destes processos, convenientemente estruturadas, detalhadas, acessíveis e aplicáveis, de acordo com as necessidades específicas de cada organização. Esta abordagem requer ainda um mecanismo efectivo que auxilie os analistas a relacionar os novos problemas de análise de dados com os processos de mineração existentes e conducentes a soluções promissoras para resolver esses problemas.

O paradigma de Raciocínio Baseado em Casos é uma abordagem de resolução de problemas, capaz de utilizar o conhecimento específico de situações de problemas concretos, previamente experimentados e explicitamente documentados. Este trabalho descreve um sistema que explora este paradigma, concebido e implementado com o propósito de assistir os analistas em duas formas principais: (i) organizar e armazenar num repositório partilhado os exemplos de processos mineração bem sucedidos; (ii) seleccionar os planos de mineração mais adequados para a resolução de um determinado problema de análise de dados de utilização de Web, dada uma descrição de alto nível desse problema. Este sistema actua a partir de um conjunto de requisitos da análise e de características do conjunto de dados disponível, e, com base na

experiência adquirida acerca da aplicação de funções e modelos de mineração de dados consegue estabelecer uma solução: planos de mineração para esses dados.

B.3 Objectivos Estratégicos

Contribuir para a simplificação e acréscimo dos níveis de produtividade e de efectividade em iniciativas de aplicação e desenvolvimento de processos de mineração de dados de utilização da Web, assistindo o analista na selecção dos métodos apropriados de tratamento de um problema específico de análise de dados, considerando a natureza desse problema e as particularidades do seu contexto.

C. Contribuições

C.1 Principais contribuições técnico-científicas

1. Sistematização de factores com influência na selecção de métodos de mineração de dados, no âmbito específico da WUM:

- Identificação de um conjunto de características que podem descrever um conjunto de dados, em termos de propriedades relevantes para um conjunto alargado de funções de mineração de dados, considerando atributos genéricos e atributos específicos relevantes no âmbito da WUM;
- Descrição de alto nível da tarefa de WUM pretendida, baseando a sua especificação em abstracções relacionadas com os problemas reais que se pretende solucionar e adaptáveis aos requisitos particulares de cada organização;
- Flexibilidade da incorporação no sistema de atributos descritores do problema, a fim de produzir uma ferramenta que facilite o estudo experimental da influência de diferentes factores na selecção de métodos de mineração de dados.

2. Estudo e extensão de métodos de determinação de soluções e optimização da sua disponibilização:

- Estudo comparativo de diferentes métodos de comparação entre atributos constituídos por conjuntos de valores (itens com cardinalidade múltipla), baseado em propostas provenientes de distintas linhas de investigação (e.g. Mineração de Dados Multi-Relacional, Análise de Agrupamento e Raciocínio Baseado em Casos);

- Extensão de métodos propostos na literatura, respeitantes ao cálculo de similaridade entre atributos com cardinalidade múltipla, de forma a reflectir a semântica pretendida;
- Extensão do processo recuperar do ciclo de Raciocínio Baseado em Casos, identificando e agrupando diferentes categorias de soluções, a fim de produzir explicitamente recomendações de métodos de mineração;
- Optimização da apresentação de soluções maximizando a sua utilidade, sob o ponto de vista do analista, conjugando dois propósitos distintos: (i) diversidade das categorias de soluções, de acordo com nível de apropriação e critérios de avaliação relevantes para o analista; (ii) variedade de casos de cada solução, a fim proporcionar vários exemplos de uma solução com eventual interesse particular para o analista.

C.2 Publicações

Wanzeller, C., Belo, O., "**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, Praga, República Checa, 2005.

Abstract: *Web Usage Mining (WUM) involves the application of mining methods to data related with the interaction processes between visitors and Web sites. Usually, this data is recognized by Web explorers as clickstream data. Data Mining (DM) tools, in general, and WUM tools, in particular, can be used to support several decision problems and users' data with very different levels of knowledge in the area. Nevertheless, these tools and their underlying paradigms are too complex to be used without the aid of specialists in the area. Selecting the most suitable methods to apply on a specific data analysis problem, in order to get useful results for a particular goal, is an important and known challenge faced on the DM step of the knowledge discovery process. This challenge is the main motivation of our work, which aims at promoting a more effective, productive and simplified exploration of clickstream data analysis potentialities. The way defended to achieve this goal involves the reuse of the acquired experience from similar problems, in order to assist the development and application of these processes. The Case Based Reasoning (CBR) paradigm comes towards ours purposes, providing support to the solution of new problems reusing past solutions of similar problems. Therefore, we implemented a CBR system with the ability to propose mining plans more adjusted to clickstream data analysis problems. The case based representation models can also be able to act as exploration and sharing bases over knowledge repositories, promoting sustained learning involving clickstream data exploitation. Our system is also based on abstractions related to the problems to solve, meaning that it could serve the particular needs of less skilled analysts that wish to*

learn how to handle a concrete problem, being also useful to specialists interested in reusing successful solutions, instead of solving the problems from scratch.

Wanzeller, C., Belo, O., "**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), Granada, Espanha, 2005.

Abstract: *The development and application of data mining processes, specifically related with information concerning Web usage, are normally very complex to users without deep knowledge in this domain. One of the many difficulties faced by experts from this working area is frequently related with the selection of mining methods (and transformation operations) to apply on a specific data analysis problem, in order to get useful results for a particular goal. The defended solution to cut off or, at least, attenuate this difficulty involves the reutilization of the acquired experience from similar problems, which had provided successful data mining processes. The Case Based Reasoning paradigm allows achieving this purpose, since it provides support to the resolution of new problems, reusing previous solutions of similar problems. By this way and following a previous proposal involving a mining plans selection system seated in this paradigm, this paper describes its respective implementation.*

Sousa, A., Wanzeller, C., Duarte, P., Baptista, M., Soares, A., Fernandes, C. "**Contacts Manager: A Mobile Web Application Consumer of Web Services**", Proceedings of the IADIS WWW/Internet 2004, Madrid, Outubro 2004.

Abstract: *With the constant innovations in the mobile devices domain, the need to make available new forms of applications is growing. We believe that this kind of devices may become large consumers of Web services. Therefore, we aim to explore the development of applications for mobile devices based in Web services. Our project involves the implementation of an application which allows the WAP access to the Intranet/Extranet of an information technologies company, aiming to obtain the information about its costumers, from anywhere.*

D. Enquadramento

D.1 Enquadramento Científico

A prossecução dos objectivos de doutoramento envolve a realização de actividades de investigação, que permitam dar resposta ao problema da selecção de abordagens conducentes à resolução de problemas de WUM, no sentido de alcançar as seguintes metas:

- Viabilizar a exploração, 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;
- Identificar um conjunto de soluções possíveis, discriminando abordagens alternativas em termos de métodos e ferramentas disponíveis, bem como critérios de apoio na selecção de soluções.

D.2 Motivação

1. Relevância:
   - Constante aumento da importância e exploração de ferramentas de Mineração de Dados (DM – *Data Mining*) e de WUM;
   - Utilidade das ferramentas de DM/WUM para utilizadores com necessidades diversas e com diferentes níveis de conhecimentos na área;
2. Desafios:
   - Complexidade das ferramentas de DM/WUM e dos paradigmas subjacentes;
   - Influência de múltiplos factores no desenvolvimento de processos de DM/WUM, entre os quais, alguns subjectivos;
   - Sobreposição dos métodos de DM, em termos dos tipos de problemas que podem resolver;
   - Inexistência de critérios simples, gerais e consistentes que possam fundamentar a selecção de métodos de mineração de dados mais adequados a um problema particular;
3. Natureza:
   - Problemas recorrentes e o uso repetitivo dos mesmos métodos são comuns;
   - A experiência e *know-how* adquirido possuem um valor proeminente;

4. Especificidade:
   - A WUM reveste-se de características específicas que requerem um tratamento dirigido;
   - Existem problemas típicos que poderão ser adequadamente organizados e abstraídos, considerando os requisitos particulares de uma organização.

D.3 Objectivos Detalhados

- Identificar os aspectos que melhor caracterizam os problemas de WUM;
- Identificar as categorias de elementos que melhor descrevem e explicam a aplicação de métodos de mineração e a condução de outras actividades inseridas no processo de extracção de conhecimento;
- Identificar os diferentes tipos de factores, e os respectivos elementos constituintes, com maior influência na selecção de métodos de DM, em termos gerais e no âmbito específico da WUM;
- Conceber um modelo conceptual de dados, capaz de suportar a representação adequada e detalhada de processos de WUM, em termos de problemas de análise de dados e respectivas soluções;
- Implementar um repositório de casos de aplicação e desenvolvimento de processos de WUM, com base no modelo conceptual de dados concebido;
- Desenhar e implementar um sistema de selecção de planos de mineração de dados, fundamentado na exploração do paradigma CBR e em várias camadas de serviços e incluindo módulos capazes de realizar as seguintes funções:
  - caracterização de dados, com base na especificação da fonte do conjunto de dados e informação complementar;
  - recolha dos requisitos da análise, envolvendo restrições explícitas, conducentes ao refinamento da descrição do problema de análise de dados;
  - recuperação dos casos mais promissores, convenientemente organizados sob o ponto de vista de sua utilidade para o analista, a partir de uma caracterização representativa do problema de análise de dados;
  - transformação de dados produzidos por ferramentas de DM/WUM, no sentido de integrar formas expeditas de aquisição de dados acerca de processos de DM desenvolvidos;
  - retenção de novos casos, suportando as formas alternativas de submissão de dados necessárias para garantir a aquisição completa de todos os dados requeridos e a sua organização e registo;
- Optimizar os serviços de interface no sentido de suportar as funcionalidades essenciais, com a robustez requerida, entre as quais se salienta: (i) a especificação do problema de análise de dados, o mais simplificada e fidedigna possível; (ii) a descrição

de processos de DM desenvolvidos, viabilizando a introdução facilitada de dados, particularmente, quando meios mais expeditos para esse efeito não estão disponíveis.
- Preparação de uma amostra de casos, baseada em dados, necessidades e processos de WUM reais, com vista a aferir a efectividade e pertinência do sistema.

D.4 Trabalhos Alternativos

Em [3] apresenta-se um sistema que guia o analista na construção de planos válidos de processos de extracção de conhecimento. O sistema baseia-se na decomposição sistemática de tarefas (em subtarefas mais simples), ao longo de várias etapas de refinamento de um processo de DM de alto nível, ajudando o analista a construir o melhor plano, com base num conjunto de operações. Os planos concluídos são compilados, podendo ser executados numa ferramenta de DM comercial. Cada (sub)tarefa e método de resolução de problemas é descrito em termos de pré e pós condições, definindo a sua aplicabilidade e viabilizando o mapeamento entre estes. Outro trabalho, em certa medida, semelhante consiste no sistema *Intelligent Discovery Electronic Assistant* (IDEA) [1] Ao contrário do trabalho anterior, baseia–se no argumento de que é muito complicado discernir qual é o melhor plano de DM, posto que os resultados de DM são imprevisíveis e os problemas são difíceis de formular. Neste sentido, o sistema usa uma ontologia para construir uma lista de processos que são apropriados para uma dada tarefa, acrescentando operações de transformação para gerar todas as sequências válidas de operações, e mostra ao analista a lista de planos válidos, ajudando-o a escolher entre estes (usando heurísticas de ordenação).

O sistema de agentes GLS [8], tal como os anteriores, suporta o planeamento de processos de extracção de conhecimento e baseia-se em condições de pré e pós condições de aplicabilidade. Neste sistema é requerida interacção com o analista a fim de optimizar as sequências válidas geradas automaticamente. Os dados e as operações são descritas ao meta-nível, mas não existe a noção de um processo completo a este nível.

O sistema Mining Mart [7] possui um âmbito diferente e mais limitado do que os trabalhos anteriores, centrando-se em processos de pré-processamento e não de DM, apesar de contemplar a aplicação de algoritmos de aprendizagem automática que necessitam de ser utilizados neste contexto. A principal meta deste projecto consiste na reutilização de processos bem sucedidos de operações de pré-processamento de dados, que foram desenvolvidas 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 utilizados em diferentes aplicações, são organizados em ontologias, a fim de suportar um nível de generalização que facilite e promova a sua reutilização. A exploração destes processos envolve a procura na base de metadados daqueles que parecem ser mais apropriados para o

problema actual. 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, que podem ser executados automaticamente. Este sistema possui ainda um nível de integração elevado com SGBD, focando-se, por conseguinte em conjuntos de dados reais e volumosos.

O objectivo do projecto METAL[1] prende-se com o desenvolvimento de métodos e ferramentas para proporcionar suporte a utilizadores de tecnologias de aprendizagem automática (Machine Learning) e DM. Para este efeito as principais metas de investigação estabelecidas incluíram o desenvolvimento de caracterizações genéricas de metaobjectos (e.g. conjuntos de dados e algoritmos), a recolha e formalização de metaconhecimento, recorrendo a diferentes formas de representação (e.g. regras e casos) e a integração da base de metaconhecimento com metaaprendizagem para atingir adaptabilidade. Um dos resultados deste projecto consistiu num protótipo de um sistema de assistência – *Data Mining Assistent* (DMA) –, que auxilia os utilizadores na selecção de algoritmos, em função da sua adequação esperada para uma dada tarefa. Esta ferramenta proporciona recomendações sob a forma de uma lista ordenada de algoritmos candidatos, de acordo com uma combinação ponderada de critérios de desempenho (precisão e tempo de aprendizagem). O âmbito deste projecto é claramente distinto do dos trabalhos anteriores, posto que a sua atenção incidiu sobre a selecção de algoritmos e, particularmente, em problemas de regressão e classificação.

O projecto METAL envolveu muitos esforços de investigação e desenvolvimento, alguns dos quais recorrendo à utilização do paradigma CBR. Um desses esforços consistiu na concepção de uma infra-estrutura, que inclui um sistema de selecção de algoritmos *Algorithm Selection Tool* (AST) [6] baseado neste paradigma. De acordo com a perspectiva defendida neste trabalho, um caso é descrito pelas restrições de aplicação, metadados do conjunto de dados e conhecimento acerca da aplicação de algoritmos, materializado sob a forma de metadados relativos a algoritmos e experiência da sua utilização. A inclusão de características de algoritmos e restrições de aplicações representa um contributo, em relação a outros trabalhos, baseados apenas em características dos dados. Adicionalmente, este trabalho antevê a necessidade de refinar o modelo de metadados e de incluir na estrutura de casos metadados relativos à configuração de algoritmos. No entanto, este trabalho apenas aborda o problema da selecção de algoritmos, não de processos, e não contempla múltiplas funções de DM.

Outro trabalho baseado no paradigma CBR e no âmbito de projecto METAL consiste no desenvolvimento de um assistente para selecção de algoritmos, que integra numa infra-estrutura unificada três características principais. A primeira consiste na combinação de

conhecimento acerca de algoritmos de aprendizagem com características de conjuntos de dados. Esta característica visa a simplificação da selecção de modelos, focando-se nas ferramentas mais promissoras, com base em restrições e preferências definidas pelo utilizador. A segunda característica corresponde à incorporação de um mecanismo para distinguir entre diferentes parâmetros de ferramentas, estendendo a selecção de modelos para contemplar a escolha de algoritmos e as configurações específicas de parâmetros. A terceira característica reside na integração de metadados obtidos a partir de diferentes experiências com metaconhecimento, não só de algoritmos como também da sua modelação, métricas de desempenho e estruturas de avaliação. Apesar dos seus contributos, tal como no trabalho anterior, este sistema não considera operações de transformação nem a aplicação composta de múltiplos algoritmos envolvendo diferentes funções de mineração.

Em contraste com os trabalhos referidos anteriormente, [2],[4] e outros usam metaregras derivadas de estudos experimentais, para ajudar a prever que algoritmos são mais adequados. As regras consideram características mensuráveis dos dados, tais como número de registos e número de atributos.

D.5 Bibliografia Principal

[1] Bernstein A., Provost F. (2001). An Intelligent Assistant for the Knowledge Discovery Process. In Proceedings of the IJCAI-01 Workshop on Wrappers for Performance Enhancement in KDD.

[2] Brazdil, P. , Gama, J. And Henery, B. (1994). Characterizing the Applicability of Classification Algorithms Using Meta-Level Learning. In Proceeedings of the European Conference on Machine Learning (ECML-94), 83-102.

[3] Engels, R., Lindner, G., and Studer., R. (1997). A Guided Tour through the Data Mining Jungle. In Proceedings of the 3nd International Conference on Knowledge Discovery in Databases (KDD-97), 14–17.

[4] Gama J. and Brazdil P.(1995). Characterization of Classification Algorithms. 7th Portuguese Conference on Artificial Intelligence, EPIA '95, 189-200.

[5] Hilario, M., and Kalousis, A. (2001). 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. Springer.

[6] Lindner, C., Studer, R. (1999). AST: Support for algorithm selection with a CBR approach. Proceedings of the 16th International Conference on Machine Learning, Workshop on Recent Advances in Meta-Learning and Future Work.

[1]http://www.metal-kdd.org/; https://www.cs.bris.ac.uk/Research/MachineLearning/METAL/index.html

[7] Morik K. And Scholz M. (2003). The MiningMart Approach to Knowledge Discovery in Databases. Intelligent Technologies for Information Analysis.

[8] Zhong, N., Liu, C. and Ohsuga S. (1997). A Way of Increasing both Autonomy and Versatility of a KDD System. In Z.W. Ras and A. Skowron, editors, Foundations of Intelligent Systems, pages 94–105. Springer.

E. Desenvolvimento

E.1 Macro-planeamento das Actividades

Plano de trabalhos inicial:

Fase	Actividades
F1	Estudo aprofundado de área e tecnologias envolvidas. Análise de sistemas disponíveis que incorporam estes tipos de tecnologias.
F2	Recolha de características e estudo comparativo de ferramentas de WUM disponíveis. Definição e implementação de componentes do sistema. Estudo da adequação das componentes desenvolvidas sobre cenários de aplicação real.
F3	Análise, fundamentação e planeamento de um modelo para uma arquitectura funcional do sistema . Implementação de um sistema piloto para o modelo projectado.
F4	Análise e preparação de uma amostra de dados real para teste e validação do sistema desenvolvido.
F5	Análise do desempenho do sistema e da qualidade dos resultados obtidos na fase anterior. Optimização de resultados.
F6	Escrita da tese de doutoramento.

A figura 1 apresenta o diagrama temporal do plano previsto. A data de início (efectiva) consiste em a de Janeiro de 2004.

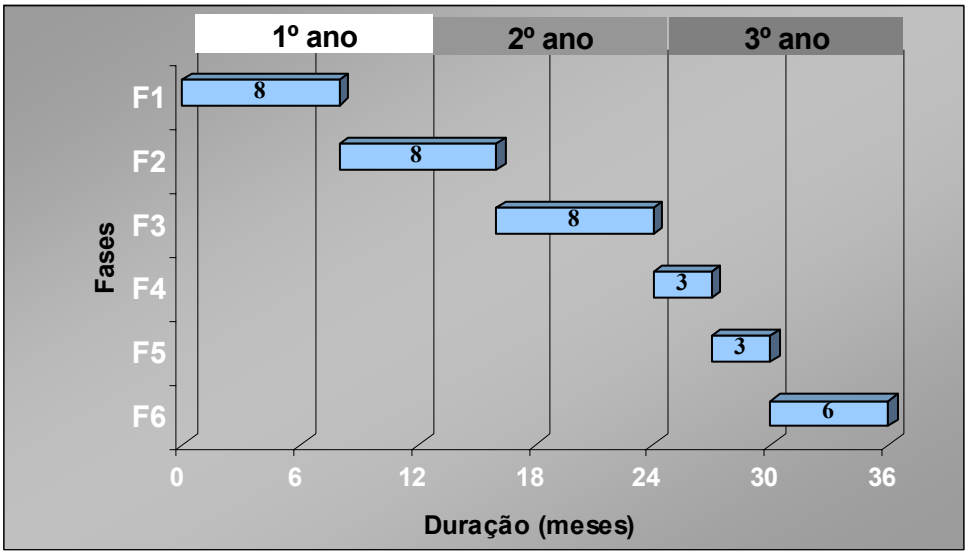

Figura 1 – Cronograma de Actividades

Situação actual

A calendarização prevista na proposta de doutoramento sofreu alterações, pelo facto de a dispensa formal de serviço docente só ter sido concedida em Janeiro de 2004 e, ainda, devido à impossibilidade de interrupção abrupta da prestação desse serviço, por motivos pedagógicos. Adicionalmente, em virtude da descoberta de um novo rumo de trabalho, que se processou no mês de Novembro 2004, a proposta de trabalho inicial também sofreu alterações de conteúdo. No entanto, procurou-se seguir as linhas orientadoras do plano de trabalhos inicial, efectuando as devidas correspondências.

Com base no exposto, alternou-se actividades de diferentes fases por dois motivos principais, que se passa a referir. Em primeiro lugar, foi necessário aprofundar competências consentâneas com o novo rumo de trabalho, realizando actividades enquadradas em fases iniciais. Em segundo lugar, as actividades de implementação de um sistema piloto (F3), de análise do desempenho do sistema e da qualidade dos resultados obtidos (F5) e a própria preparação de uma amostra de dados real (F4), decorreram progressivamente em várias etapas e, em certa medida, de forma intercalada. No momento, estão a ser realizadas actividades respeitantes às fases F4 e F5, uma vez que as fases anteriores já foram concluídas. Nomeadamente, a conclusão da implementação do sistema piloto ocorreu em Novembro de 2005. Paralelamente, estão a ser envidados esforços no sentido de publicar os resultados que vão sendo obtidos.

E.2 Recursos Necessários

Tratando-se de um doutorando externo ao DI, este ponto não é aplicável.

E.3 Recursos Disponibilizados

Tratando-se de um doutorando externo ao DI, este ponto não é aplicável.

F. Avaliação

F.1 Análise Comparativa

O âmbito dos trabalhos referidos na secção D.4 são diferentes do âmbito do trabalho proposto. Por um lado, não se considera todo o processo de descoberta de conhecimento. Por outro lado, também não se estudar apenas a questão da selecção de diferentes algoritmos. A ênfase do nosso trabalho reside na selecção de métodos de mineração de diferentes funções, uma vez que as soluções são distinguidas em termos dos métodos que aplicam. No entanto, prevê-se suporte ao nível de processos, entendidos como sequências de aplicações de métodos/operações, incluindo indicações e descrições acerca de operações de transformação e envolvendo a escolha das próprias funções de mineração de dados. Por conseguinte, os trabalhos que contemplam todo o processo de extracção de conhecimento proporcionam um suporte mais alargado, apesar de mais generalista do que o nosso.

Com a excepção do projecto *MiningMart*, não parece existir nos trabalhos referidos uma preocupação (ou necessidade, no caso dos de âmbito mais estreito) em abstrair a descrição do tarefa de análise de dados. Esta é tipicamente definida indicando uma função de DM. De acordo, com a nossa perspectiva, esta indicação faz parte da solução, dada a conhecida sobreposição de determinados métodos e funções de DM, em termos dos problemas que podem resolver.

No que respeita às principais abordagens em que os sistemas abordados se baseiam, podemos considerar a existência de três tipos fundamentais: planificação de processos válidos, representação explícita de conhecimento e representação do conhecimento específico baseado em casos. Quanto à primeira, a principal desvantagem reside no facto de processos válidos não serem necessariamente processos bem sucedidos, entendidos como soluções excelentes e com resultados comprovados. Acredita-se que as potencialidades de gerar um processo válido através da adaptação de um processo bem sucedido são maiores do que o inverso. Adicionalmente, as condições de validade sofrem alterações substanciais perante a

aplicação de operações de transformação. Por conseguinte, esta abordagem parece-nos ser mais indicada no apoio ao analista, após a identificação de uma solução bem sucedida e apropriada para o problema corrente.

Quanto às restantes abordagens, aplicam-se as vantagens e desvantagens clássicas da representação explícita de conhecimento em relação ao paradigma CBR. No presente contexto salienta-se duas desvantagens da primeira abordagem: (i) a dificuldade em geral um modelo consistente e abrangente, particularmente, no que se refere a âmbitos mais alargados com um nível superior de factores subjectivos envolvidos; (ii) o problema da actualização e extensibilidade do modelo, dada a constante evolução dos domínios de DM e WUM e a consequente necessidade de acrescentar conhecimento respeitante a experiências relativas a novos modelos, algoritmos ou ferramentas de DM e suas particularidades. Dados os requisitos envolvidos, as desvantagens do paradigma CBR são ultrapassadas pelas associadas a abordagens alternativas. A este propósito, torna-se necessário referir que o sistema *MiningMart* não explora as potencialidades do metamodelo nem de funcionalidades típicas de sistemas CBR, para ajudar os utilizadores a estabelecer a correspondência entre o problema actual e os existentes.

F.2 Auto-avaliação do Trabalho Realizado

Os testes conduzidos até ao momento apontam no sentido da efectividade do sistema implementado. Nomeadamente, por defeito o sistema selecciona prioritariamente os processos que utilizam conjuntos de dados mais semelhantes, dado que os metadados do conjunto de dados estão em maioria, em relação a outros atributos descritores do problema. Este comportamento por defeito do sistema vai ao encontro do pretendido, e pode ser considerado um bom resultado, pois, por um lado, as características do conjunto de dados são sempre um factor crucial. Por outro lado, a especificação do problema pode ser refinada, recorrendo a meios como a exclusão de atributos descritores, a especificação de níveis de importância e a aplicação de critérios de filtragem exacta. No entanto, o sistema foi testado recorrendo a um conjunto reduzido de processos de WUM. As actividades de preparação de casos adicionais ainda estão a decorrer, recorrendo a conjuntos de dados disponíveis publicamente e a descrições de análises bem sucedidas que foram realizadas sobre os mesmos. Pretende-se que estes casos sustentem a criação de uma base de casos preliminar e permitam aferir o nível de efectividade do sistema em condições mais realistas. Na sequência desta actividade, pretende-se construir uma caso de estudo baseado em dados e necessidades reais de uma organização alvo.

F.3 Auto-avaliação da Documentação Produzida

Apesar de se ter envidado esforços adicionais, no sentido de atenuar os efeitos negativos da alteração dos trabalhos de doutoramento, o espaço temporal requerido para efeitos de publicação é considerável e, por conseguinte, é difícil de compensar. Maioritariamente, por esta razão, o esforço no sentido da publicação de artigos está a decorrer, no momento, com maior intensidade, esperando-se que venha a produzir resultados, passíveis de uma análise mais adequada.

G. Dificuldades

G.1 Dificuldades Técnico-Científicas

As principais dificuldades sentidas prendem-se com a investigação envolvendo áreas não conhecidas com a profundidade requerida, apesar do estudo destas áreas estar previsto na planificação do projecto. Adicionalmente, a alteração do principal objectivo do doutoramento, modificou o normal decorrer dos trabalhos. Uma das vertentes possíveis deste projecto assumiu grande importância, dando origem à descoberta de um novo rumo. O apoio à decisão na aplicação de métodos de mineração de dados e em processos de descoberta de conhecimento é uma vertente de trabalho que, por inerência, se pode considerar relacionada ou mesmo incluída na proposta de projecto de doutoramento submetida e aprovada. Apesar de não ter sido esta a vertente priorizada e alvo de ênfase na referida proposta, a nova vertente conduziu a um novo rumo a seguir, em detrimento da concretização de outras actividades previstas, pelos motivos que se passa a referir. O apoio à decisão na selecção de métodos de mineração de dados é uma área de trabalho com interesse reconhecido, que carece de um tratamento profundo e efectivamente capaz de abstrair a grande complexidade subjacente. Na altura acreditou-se ser possível prestar contributos mais pertinentes na nova área de trabalho e, por conseguinte que este novo rumo seria mais profundo, podendo assim compensar os efeitos negativos da perda de alguns esforços já envidados, uma vez que permitiria ampliar o nível de motivação na concretização deste projecto.

G.2 Outras Dificuldades

Nada a assinalar de momento.