Jorge Manuel Esparteiro Garcia

Assistente Web para Análise Automática e Colaborativa de Dados - WebCAD

Departamento de Engenharia Electrotécnica e de Computadores
Faculdade de Engenharia da Universidade do Porto
Março de 2007

Jorge Manuel Esparteiro Garcia

Assistente Web para Análise Automática e Colaborativa de Dados - WebCAD

Tese submetida à Faculdade de Engenharia da Universidade do Porto para obtenção do grau de Mestre em Engenharia Informática sob a orientação do Professor Doutor Rui Camacho e co-orientação do Professor Doutor João Correia Lopes

Departamento de Engenharia Electrotécnica e de Computadores
Faculdade de Engenharia da Universidade do Porto
Março de 2007

Para a Ana Luísa, meus Pais e Irmãos

Agradecimentos

Desejo agradecer a colaboração e o apoio prestado de todos aqueles que contribuíram em diferentes momentos para que este trabalho chegasse a um bom termo.

Em particular, gostaria de agradecer todo o valioso trabalho de orientação do Prof. Doutor Rui Camacho e do Prof. Doutor João Correia Lopes, pela forma como me apoiaram e orientaram o trabalho, bem como pela disponibilidade e dedicação permanente que sempre demonstraram, mesmo em ocasiões de agenda muito preenchida.
Gostaria também de agradecer a todos os meus Professores e colegas que me acompanharam durante o mestrado e que em diferentes níveis contribuíram para que pudesse realizar este trabalho.
Um agradecimento também especial aos todos os meus amigos e colegas de trabalho da Escola Superior de Ciências Empresariais do Instituto Politécnico de Viana do Castelo que me apoiaram e incentivaram durante a realização deste trabalho.

Desejo agradecer à minha namorada Ana Luísa, pelo seu constante apoio e por me acompanhar em todos os momentos da realização deste trabalho.

Quero agradecer à minha família pelo apoio inequívoco que me deu, particularmente aos meus pais, à minha irmã Cláudia e ao meu irmão Luís, sem os quais este trabalho não seria possível.

Finalmente, agradeço a todos os meus colegas e amigos que sempre me ajudaram e apoiaram, e cujos nomes me é impossível aqui citar, por falta de espaço. A todos muito obrigado!

Abstract

Significant improvements in research activities may be achieved by means of: collaboration with other researchers in the same field; easy use of computational resources to solve research problems without the need to know the tool's details; joining inter-disciplinary research teams; use tools to fasten the access to relevant and valuable information; share results of our own experiments and use other researchers's experiences and experimental results.

The work presented in this thesis consisted in the development of a Collaborative and Automatic Data Analysis tool that addresses the above mentioned desirable features of the research activity. The tool allows each researcher to use an automatic data analysis tool in a user friendly fashion. It maintains records and logs of all experimental activity of each user. Each user has a private and public "space" where he can store or disclose experimental results and data. The tool also provides a "web service" that enables the search for related information in the public spaces of other users located in different web sites.

The automatic data analysis tool used is an Inductive Logic Programming (ILP) system called IndLog and we developed a *wrapper* program to hide the ILP specific details of IndLog. A variety of different kinds of data involved in the experimental activity is stored in a database and the tool is accessible from the Web. There may be several sites housing the tool and each site may accommodate several users.
The deployment of the tool showed both that IndLog´s details of use were completely hidden from the user and sharing of information was desirable and quite useful.

Resumo

Melhoramentos significativos em actividades de investigação podem ser alcançados através do trabalho desenvolvido nesta dissertação. Tais melhoramentos incluem: a colaboração com outros investigadores da mesma área de investigação; a facilidade de utilização de recursos computacionais para solucionar problemas de investigação sem necessidade de conhecer os detalhes técnicos das ferramentas; a junção de equipas de investigação inter-disciplinares; a disponibilização de resultados das nossas experiências e utilização de experiências e resultados experimentais de outros investigadores.

O trabalho teve como objectivo o desenvolvimento de uma ferramenta de Análise Automática e Colaborativa de Dados que aborda as funcionalidades da actividade de investigação acima mencionadas e desejadas. A ferramenta permite a cada investigador utilizar uma ferramenta automática de análise de dados de forma "amigável". Mantém um registo para todos os utilizadores de toda a sua actividade experimental. Cada utilizador tem um "espaço" privado e um "espaço" público onde pode armazenar ou publicar dados e resultados de experiências. A ferramenta também fornece um "serviço *Web*" que permite a procura de informação útil no espaço público de outros investigadores mesmo trabalhando em sítios *Web* diferentes.

A ferramenta automática de análise de dados utilizada é um sistema de Indução de Programas em Lógica chamado IndLog, que executa por baixo de uma aplicação *wrapper*, desenvolvida para esconder os detalhes técnicos de utilização do IndLog. Os diferentes tipos de dados envolvidos no processo experimental são armazenados numa base de dados e a ferramenta é acessível a partir da *Web*. A ferramenta pode ser alojada em diversos sítios *Web*, e cada um desses sítios permite o acesso a diversos utilizadores.

A implementação realizada e a sua utilização mostram que é possível esconder completamente os detalhes de utilização do IndLog e que a partilha de informação se revela extremamente útil.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 O Problema Abordado

É de conhecimento geral que numa vasta gama de domínios, tanto científicos como empresariais, a quantidade de dados recolhidos e processados é muitas vezes gigantesca. É portanto muito difícil, senão mesmo impossível para analistas humanos, efectuar uma análise adequada em tão elevada quantidade de dados e mesmo para um especialista é difícil identificar informação valiosa implícita em tantos dados, ou que traga novo conhecimento para o domínio.
Hoje em dia, a análise de grandes volumes de dados é feita usando ferramentas computacionais complexas que utilizam técnicas de *Data Mining*[1] e que permitem analisar e identificar informação de grande valor implícita nos dados e até mesmo descobrir novo conhecimento.
Um dos problemas destas abordagens com técnicas de *Data Mining* é que requerem um especialista ou, pelo menos, uma pessoa que tenha tido formação para a correcta utilização destas ferramentas. Para que estas técnicas possam ser usadas por uma vasta gama de utilizadores é crucial e bastante útil que qualquer pessoa seja capaz de as usar, sem que tenha de frequentar uma formação específica para a utilização da ferramenta, ou que necessite da assistência de um especialista.

Outros problemas que existem actualmente na análise de dados dizem respeito à

[1]Daqui em diante utilizar-se-á a expressão inglesa *Data Mining* para fazer referência a Extracção de Conhecimento.

partilha de resultados e mesmo de informação sobre experiências efectuadas entre os investigadores ou as pessoas que fazem análise de dados.
A análise de dados é uma actividade que pode ser muito custosa tanto em termos de tempo despendido, como por vezes também em termos financeiros. Numa organização de grande dimensão, ou numa comunidade de investigação, existe bastante conhecimento que sendo partilhado pode poupar bastante tempo de investigação, podendo-se "poupar" por vezes alguns passos de investigação e também dinheiro.

O problema abordado nesta trabalho é o de permitir a investigadores que não tenham conhecimentos específicos de utilização de ferramentas automáticas de análise de dados poderem utilizar uma classe de ferramentas eficazmente, explorando e potencializando as características destas ferramentas, sem que necessitem de qualquer formação nem a assistência de um especialista.
A possibilidade da utilização de trabalho colaborativo na análise automática de dados é também considerada: permitir que investigadores em diferentes localizações geográficas possam trabalhar dentro da mesma área de investigação e partilhem resultados, dados e diversa informação relativa a experiências que efectuem dentro da sua área de investigação.

1.2 Motivação para o Trabalho

Embora já existam actualmente, algumas ferramentas para fazer análise (semi)-automática de dados, estas exigem, na maior parte das vezes, um utilizador treinado e formado que saiba pelo menos como configurar e parametrizar de modo adequado a ferramenta para obter melhores resultados.
O desenvolvimento de uma ferramenta com um interface gráfico "amigável"[2] que permita ao utilizador abstrair-se do conhecimento específico da ferramenta é sempre uma mais valia. Uma ferramenta deste género, poderia ser utilizada por uma gama mais vasta de utilizadores. Utilizadores do meio empresarial ou investigadores apenas necessitariam de conhecimentos base de utilização de um computador, para conseguirem utilizar todas as potencialidades e características da ferramenta.

A investigação em ferramentas de análise de dados, com interface "amigável" é portanto uma uma linha de investigação com muito valor.

[2]Daqui em diante utilizar-se-á a designação "amigável" como tradução de *user-friendly*.

A análise de dados, é na maior parte das vezes, uma operação muito morosa e, por vezes, pode também ser muito dispendiosa em termos financeiros.
A partilha de informação pode, contudo, poupar bastante tempo. A informação partilhada pode ser valiosa e de importância para a execução de análises de grandes volumes de dados, como são os resultados de análises já efectuadas ou experimentações no domínio científico.
A partilha deste tipo de informação pode por exemplo, evitar a repetição de análises de dados que já foram efectuadas por utilizadores diferentes dentro de uma organização empresarial ou dentro de um instituto ou comunidade de investigação. Pode também permitir a um dado utilizador visualizar resultados de determinadas análises anteriores ou de outros que lhe permitirão convergir na determinação de certos resultados e que sem essa informação seriam mais difíceis de alcançar, ou por vezes, mesmo inalcançáveis.
A partilha de resultados, pode assim, permitir a uma organização empresarial ou uma comunidade de investigação poupar tempo e dinheiro, que podem ser factores críticos para obtenção de bons resultados.

1.3 Abordagem Adoptada

Neste trabalho são abordados dois problemas: a facilidade e de utilização de ferramentas automáticas de análise de dados e a partilha de conhecimento.
Para abordar o primeiro problema foi criada uma ferramenta que permite ao utilizador abstrair-se de qualquer conhecimento específico da ferramenta de análise de dados usada, podendo utilizar a ferramenta qualquer utilizador com apenas conhecimentos-base de informática.
Para abordar o problema de partilha de conhecimento, foram desenvolvidas bibliotecas de recursos que são muito úteis para a análise de dados, contendo dados e conhecimento de fundo. A informação partilhada inclui ainda resultados e os contextos de experiências que cada utilizador julga serem úteis. Essa partilha é feita, possibilitando a cada utilizador a disponibilização pública dos resultados, conjuntos de dados e conhecimento de fundo[3], além de fornecer serviços para cada utilizador poder pesquisar informação relevante para as suas análises e experiências.

[3]Conhecimento de fundo é a tradução para *Background Knowledge* e, como se verá no Capítulo 2, inclui informação relevante para a construção do modelo para os dados.

Neste trabalho, foi desenvolvido um *wrapper*[4] que "esconde" todo o conhecimento específico que é necessário para uma ferramenta de análise de dados.
Este *wrapper* pode ser usado através de um interface *Web*, tornando-o acessível a qualquer pessoa que o pretenda utilizar como ferramenta de análise de dados. Esta ferramenta fornece ainda alojamento para os dados e para toda a informação produzida no processo de análise de dados por cada utilizador. Permite também a definição de dados ou informação privada e de informação e dados "partilháveis". Por fim, a ferramenta oferece mecanismos de pesquisa de informação relevante, seja ela informação de experiências, conjunto de dados, conhecimento de fundo ou resultados no espaço público de outros utilizadores que tenham partilhado essa informação.

1.4 Principais Contribuições da Tese

As principais contribuições desta tese são a criação de uma aplicação que permite a análise (semi)-automática de dados numa ferramenta poderosa de análise de dados como é um sistema de Indução de Programas em Lógica, sem que o utilizador tenha um conhecimento específico dela, através de uma aplicação de simples utilização como são as aplicações baseadas em *browser*. Com o desenvolvimento desta aplicação, o utilizador, pode executar as suas experiências de análise de dados como se trabalhasse directamente com a ferramenta, mas abstraindo-se de qualquer especificação técnica dela.

Outra contribuição desta tese é o desenvolvimento de uma plataforma que permite a utilizadores que trabalhem dentro da mesma instituição ou mesmo em outras instituições que partilhem em tempo real conjuntos de dados, conhecimento de fundo e resultados de experiências entre si. Esta partilha de informação pode permitir que os utilizadores ou investigadores, poupem alguns passos de investigação através da visualização de resultados das experiências de outros utilizadores que trabalhem dentro da mesma área de investigação.

As contribuições desta tese foram materializadas no sistema WebCAD para análise automática de dados acessível através de uma página *Web*.

[4]Em português é designado por "envolvente".

1.5 Estrutura da Tese

O resto da dissertação está estruturado da forma que é descrita a seguir.
No Capítulo 2 são introduzidos conceitos básicos das tecnologias usadas de modo a se perceber melhor a proposta apresentada na dissertação. São ainda referidos trabalhos efectuados dentro da área da Aprendizagem Computacional e trabalhos realizados com portais de algum modo relacionadas com o trabalho aqui apresentando, bem como em trabalhos que permitam a partilha de conhecimento e resultados através de ferramentas colaborativas.

No Capítulo 3 são descritos os requisitos funcionais e não funcionais do utilizador para a aplicação WebCAD. São apresentados, em diagramas UML, os casos de utilização que o WebCAD deverá respeitar, bem como os respectivos diagramas de actividades.

No Capítulo 4 é apresentada a arquitecta da solução proposta para o WebCAD. É detalhada a arquitectura lógica e física da solução e são apresentadas as tecnologias que foram utilizadas para a implementação desta aplicação.

No Capítulo 5 descrevem-se os aspectos mais relevantes da implementação do WebCAD. É detalhada a forma como foi desenvolvido o WebCAD, nomeadamente o armazenamento dos dados utilizados e como é feita a comunicação e as trocas de informação entre o sistema WebCAD e o sistema de análise de dados (IndLog).

No Capítulo 6 são apresentados alguns casos de estudo que foram feitos para a aplicação WebCAD. São detalhadas as principais formas de utilização da aplicação e são mostrados exemplos para cada um dos casos de utilização apresentados.

No Capítulo 7 apresentam-se as conclusões sobre o trabalho desenvolvido nesta dissertação e perspectiva-se algum do trabalho futuro que poderá ser desenvolvido como continuação deste projecto.

No Apêndice A estão listados os acrónimos utilizados nesta dissertação.

Finalmente no Apêndice B é apresentado um dicionário com a tradução dos termos estrangeiros utilizados nesta dissertação.

Capítulo 2

Estado da Arte

2.1 Indução de Programas em Lógica

A Indução de Programas em Lógica[1] é uma sub-área da Aprendizagem Computacional focada no estudo da automatização de processos indutivos. A Indução de Programas em Lógica (ILP) tem influências de áreas como a Aprendizagem Computacional, da Programação em Lógica e da Estatística. "Herda" da Programação em Lógica a representação baseada em Lógica de Predicados de Primeira Ordem (LPPO), e utiliza-a tanto para representar os dados como os modelos construídos[2] Da Estatística utiliza essencialmente métodos para avaliar a qualidade das hipóteses geradas automaticamente.

Um sistema de ILP utiliza habitualmente como dados de entrada um conjunto de exemplos, um conjunto de informação (designado *Background Knowledge*) relevante para a construção das hipóteses e um conjunto de restrições. Os exemplos podem ser de dois tipos: positivos ($\mathbf{E}^+$) ou negativos ($\mathbf{E}^-$). Os exemplos positivos são instâncias do conceito alvo. Os negativos não são instâncias do conceito alvo e são usados para evitar sobregeneralização. O *Background Knowledge* ($\mathbf{B}$) é um conjunto de informação que é útil para a construção das hipóteses. Pode incluir tanto relações como funções (numéricas, estatísticas, geométricas etc). As restrições permitem melhorar o desempenho do sistema de ILP evitando computações

[1] Tradução de *Inductive Logic Programming* — ILP.

[2] Usaremos o termo hipótese como sinónimo de modelo. Uma hipótese é representada por um conjunto de cláusulas.

desnecessárias. Toda esta informação é codificada usando LPPO. O objectivo de um sistema de ILP é construir uma hipótese usando o *Background Knowledge* que "explica" os exemplos positivos, não explica os exemplos negativos e é consistente com as restrições.

em lógica. É pois uma tecnologia para geração e validação de hipóteses

O ILP[3] permite resolver duas limitações das técnicas "tradicionais" da Aprendizagem Computacional: a utilização de uma representação formal de conhecimento limitada; e as dificuldades em usar conhecimento de fundo substancial durante o processo de aprendizagem.

Esta maior facilidade de manipulação dos programas lógicos, ainda segundo Muggleton [Muggleton and Raedt, 1994], deve-se ao facto de numa lógica clausal pura, se poderem efectuar modificações num programa, simplesmente adicionando ou apagando cláusulas completas ou literais numa cláusula, sem que haja qualquer preocupação com os efeitos de ordenação. Devido às semânticas dos programas de lógica estarem tão ligadas à sua sintaxe, essas alterações têm um efeito e impacto bastante suave na generalidade dos programas resultantes.

Podem-se considerar como vantagens de utilização do ILP, a utilização da LPPO, uma linguagem muito poderosa, como linguagem de representação de dados, a rapidez de execução que estes sistemas oferecem aos investigadores, bem como a compreensibilidade dos resultados obtidos. Outras das vantagens do ILP, são a sua capacidade de aceitar informação adicional sobre o domínio, bem como a possibilidade combinar harmoniosamente computações numéricas com relações.
São ainda consideradas vantagens do ILP, o largo espectro de domínios de aplicação em que pode ser utilizado e o facto da maioria destes sistemas ILP estarem disponíveis na *Web* para *download*.

Os sistemas de ILP são utilizados numa vasta gama de domínios. Algumas das aplicações do ILP são por exemplo: concepção de novos fármacos, detecção de problemas de tráfego; dimensionamento de malhas em Métodos de Elementos finitos; trefilagem de aço; análise de dados de acidentes rodoviários, entre outras

[3]Daqui em diante utilizar-se-á a abreviatura inglesa ILP para as referências à Indução de Programas em Lógica.

que serão referidas mais em detalhe na Secção 2.1.3.

2.1.1 Breve Resumo do ILP

Um sistema de ILP tem como objectivo a construção automática de modelos para dados. Durante o processo de criação do modelo final gera e avalia um conjunto de hipóteses. O modelo final deve ser consistente com os exemplos.

Um sistema de ILP funciona da seguinte forma: Recebe como entrada Conhecimento de Fundo (*Background Knowledge*)[4] B e um conjunto de exemplos (positivos e negativos) $E = E^+ \wedge E^-$. Através de um processo de Indução, o sistema ILP, procura criar um programa lógico que prove todos os exemplos positivos e nenhum dos exemplos negativos. Como resultado retorna a "melhor" hipótese H (conjunto de cláusulas) satisfazendo um determinado conjunto de restrições. As restrições são necessárias para reduzir o número de hipóteses candidatas. O processo de indução é mapeado numa procura num *espaço de hipóteses* (um grafo). Um sistema de ILP percorre este espaço de hipóteses criando novos nós do grafo (hipóteses) e avaliando a qualidade das hipótese geradas.

O sistema que é usado no WebCAD é baseado no sistema PROGOL [Muggleton and Raedt, 1994] e Aleph [Srinivasan, 2001] de Srinivasan e chama-se Indlog [Camacho, 2000] de Rui Camacho.

2.1.2 O Sistema IndLog

Como foi referido o sistema de ILP utilizado neste trabalho é o IndLog [Camacho, 2000]. É um sistema de ILP empírico, de acordo com a classificação de De Raedt [De Raedt and Bruynooghe, 1992].
É um sistema que tem por base a técnica de *Mode Directed Inverse Entailment* [Muggleton, 1995], realizando uma pesquisa no espaço de hipóteses do topo para a base.
Apesar de ser um sistema semelhante a outros sistemas ILP existentes como o Aleph

[4]Daqui em diante utilizar-se-á a expressão inglesa *Background Knowledge* para as referências ao Conhecimento de Fundo.

[Srinivasan, 2001], Progol [Muggleton and Raedt, 1994] ou o SKILit [Jorge, 1998], o Indlog difere destes pois utiliza técnicas para construir uma cláusula inicial a fim de reduzir o espaço de procura [5], que conjuntamente com informações fornecidas pelo utilizador ou deduzidas a partir dos dados já disponíveis, oferece melhoramentos significativos de eficiência na procura de resultados.

O sistema IndLog incorpora ainda certas funcionalidades que não foram desenvolvidas em sistemas de ILP anteriores, como por exemplo, a tarefa de aprendizagem que neste sistema resulta numa teoria não-redundante.
Segundo Camacho [Camacho, 2000] é também um sistema adequado para processar os dados usados na construção de controladores, pois inclui características vocacionadas nesse sentido, como a avaliação retardada de literais, a possibilidade de utilizar como conhecimento de fundo um conjunto qualquer de predicados Prolog, a possibilidade de utilização como constantes dos elementos do Universo de Herbrand, a utilização de uma função de custo definida pelo utilizador para guiar a procura de hipóteses e a minimização do *Mean Square Error* (MSE).
O IndLog aprende com base em exemplos num determinado momento e executa técnicas de "saturação" e "alisamento" antes do passo final de "refinamento" dos resultados. O IndLog pode também reduzir substancialmente o tamanho da cláusula base quando a linguagem de hipóteses está restringida a "teorias conectadas".

2.1.3 Aplicações de ILP

2.1.3.1 Aquisição de Conhecimento

A aquisição de conhecimento é uma tarefa muito morosa e difícil, e susceptível a erros, como qualquer outra actividade humana. Sistemas de ILP contribuem para esta aquisição no desenvolvimento de técnicas e ferramentas que assistem e automatizam parcialmente o processo de aquisição de conhecimento.
O ILP pode assim ser usado para automaticamente construir bases de conhecimento para sistemas inteligentes em modelos qualitativos, nos quais as técnicas de aprendizagem proporcional não são suficientes. Por outro lado, é muito complicado obter aquisição de conhecimento apenas com uma técnica individual, é mais adequado

[5] Técnica existente também no sistema Golem [Muggleton and Feng, 1990] por exemplo

haver um sistema com um conjunto de técnicas base, e seleccionar uma ou mais técnicas conforme a natureza do domínio ou o tipo de conhecimento em investigação [Harmon et al., 1988].

Existem ainda duas abordagens à aquisição do conhecimento no ILP, uma empírica e outra interactiva. A abordagem empírica do ILP à aquisição do conhecimento é mais adequada a domínios onde existe um conhecimento (*Background Knowledge*) estável do domínio em estudo além de um grande conjunto de exemplos para um dado conceito. A abordagem interactiva é mais adequada para domínios onde diversos conceitos têm de ser adquiridos e existe pouco conhecimento (*Background Knowledge*) ou, possivelmente, esse conhecimento está errado.

2.1.3.2 Descoberta de Conhecimento em Bases de Dados

A descoberta de conhecimento em bases de dados consiste na extracção de informação implícita, que potencialmente é bastante útil, e que ainda não é conhecida. Esta informação, normalmente, consiste em dados presentes em bases de dados de grande dimensão onde se pode extrair informação muito valiosa, mas que está "escondida" na vasta quantidade de dados presentes na base de dados. Essa extracção feita com ferramentas convencionais, como simples aplicações de consultas a base de dados, é normalmente muito morosa e custosa para o utilizador, fazendo com que muita informação potencialmente útil, nunca seja retirada de uma base de dados. Além dessa dificuldade, existe ainda o problema da existência de informação ou dados com ruído presentes nas bases de dados, como indica [Frawley et al., 1992] e que dificulta a descoberta do conhecimento. Existem diversas técnicas de preparação dos dados no processo de descoberta de conhecimento nas bases de dados, que removem ou minimizam os efeitos destes dados ruidosos e outros dados inconsistentes ou incompletos.

Em geral, o processo de descoberta de conhecimento, que se desenrola em várias fases, inclui a gestão dos algoritmos de *Data Mining*, utilizados para extrair padrões dos dados e a interpretação dos padrões encontrados pelos mesmos. Uma das vantagens da utilização do ILP em *Data Mining* é que pode utilizar com facilidade as tabelas das bases de dados através de *Data Mining* multi-relacional.

2.1.3.3 Descoberta de Conhecimento Científico

A Descoberta Científica é uma das aplicações do ILP em que os investigadores têm desenvolvido mais trabalho.
O processo de descoberta de conhecimento pelos cientistas ou por investigadores de descoberta de conhecimento é feito normalmente com base num determinado número de observações ou exemplos, que são generalizados usando o conhecimento já existente sobre o domínio de estudo. O resultado é uma teoria que é inicialmente desconhecida e que pode ser considerada como um novo dado ou como uma nova formulação de conhecimento. Normalmente, na geração destas teorias, são feitas novas experimentações para confirmar a nova teoria formulada ou para, por outro lado, rejeitar esta nova teoria. De facto, esta teoria, por tratar-se apenas de uma hipótese, tem de ser completamente testada para que seja validada. O ILP pode também ajudar os cientistas e investigadores nos passos necessários para a construção de uma nova teoria, nomeadamente: na geração empírica ou interactiva das teorias lógicas gerais criadas a partir de observações efectuadas, na geração interactiva de experimentações cruciais na descoberta da teoria lógica e nos testes sistemáticos de uma nova teoria lógica com experimentações cruciais que podem inequivocamente falsificar a teoria.

A geração de teorias a partir de um conjunto de observações ou exemplos pode ser feita tanto em sistemas ILP empíricos como em interactivos.
Os sistemas ILP empíricos, que são os sistemas utilizados neste trabalho, desenham as experimentações durante a descoberta de conhecimento científico, ou seja, vão sendo desenhadas novas experimentações com novos parâmetros durante o processo, de forma a poder chegar a uma nova teoria lógica. Nesta tarefa é importante que todas as experimentações efectuadas sejam relevantes e que sejam coerentes com as definições explicadas pela nova teoria formulada.
A fase final do processo que envolve o teste sistemático de uma nova hipotética teoria para falsificar ou confirmar a nova teoria, pode ser feita também como sistemas ILP interactivos.

2.1.3.4 Outras Aplicações

Existem outras aplicações do ILP, nomeadamente em áreas de Engenharia, Processamento de Linguagem Natural, Ciências do Ambiente e Ciências da Vida, entre

outras.
Algumas destas aplicações em que pode ser utilizado o ILP acima mencionadas, baseiam-se na capacidade do ILP de aprendizagem a partir de dados da vida real, o que faz com que o ILP seja aplicado nos mais diversos domínios. Alguns exemplos destas aplicações são a descoberta de regras que orientam a topologia tridimensional da estrutura de uma proteína [Turcotte et al., 1998], a Detecção de problemas de tráfego, a partir de diferentes situações de tráfego [Džeroski et al., 1998], Descoberta de conhecimento a partir de dados extraídos de mapas [Popelynsky, 1998].
Uma descrição detalhada destas aplicações pode ser encontrada no sítio *Web* do projecto ILPnet2 em "http://www.cs.bris.ac.uk/ ILPnet2/".

2.2 Portais sobre ILP

Existem actualmente alguns recursos e aplicações de ILP na Internet. Os portais de ILP são especialmente importantes, enquanto agregadores de recursos relacionados com ILP, devido ao crescente interesse que existe pelos métodos e processos de ILP nomeadamente para a descoberta de conhecimento em bases de dados, para problemas de *Data Mining* e para descoberta de conhecimento científico.
Além disso, os portais, têm outro papel, que é o de ligar a comunidade de investigadores em ILP existente, com todos os utilizadores interessados em utilizar os sistemas ILP. Entre estes portais, pode-se encontrar o ILPnet2 [ILPnet2, 2006] e o portal do Grupo de *Machine Learning* da Universidade de Oxford [MLCL, 2006], da Universidade de York [UY-MLG, 2006], Imperial College [CBL, 2006], etc.

2.2.1 ILP Oxford

Neste portal, são apresentados alguns projectos dos membros do grupo envolvendo Machine Learning e especialmente ILP.
São disponibilizadas informações sobre todos os projectos, bem como explicações sobre os trabalhos e ainda estão disponíveis algumas aplicações desenvolvidas em ILP pelo grupo.
Além desta informação, são também disponibilizados alguns conjuntos de dados

(*Datasets*)[6]. Estes *Datasets*, podem ser usados por diversos sistemas ILP existentes. A informação disponibilizada é estática descrevendo as aplicações e indicando referências bibliográficas que descrevem os resultados das experiências realizadas com as aplicações. O sistema Aleph [Srinivasan, 2001] é o sistema de ILP mais utilizado.

2.2.2 ILPnet2

O ILPnet2 foi uma rede de excelência composta por 37 universidades e institutos de investigação. Esta rede teve como missão disponibilizar informação científica relacionada com ILP. Essa informação é composta por sistemas ILP existentes, informações e dados de aplicações desenvolvidas em ILP e qualquer outro tipo de projecto em ILP.
No seu portal, o ILPnet2, disponibiliza também *Datasets* para diversos sistemas e a explicação desses mesmos dados e de como podem ser utilizados em sistemas ILP. Além desta informação, o ILPnet2, tem uma base de dados com referências a artigos, livros e todo o tipo de documentos científicos sobre ILP. É o portal mais completo referenciando os avanços científicos feitos com ILP.

2.2.3 Projecto Metal

O Projecto Metal [Metal, 2002] é um projecto cujo objectivo é o desenvolvimento de métodos e ferramentas para ajudar utilizadores na escolha de algoritmos de *Machine Learning* adequados para processar os seus dados.
Neste projecto, foi disponibilizada uma aplicação *Web* que possibilita a escolha do método ou algoritmo de *Data Mining*, dentro de um conjunto de métodos disponíveis, mais ajustado para os dados que o utilizar pretende analisar, além de mostrar uma classificação dos vários algoritmos de *Data Mining* mais ajustados para os dados. Além disto, a aplicação sugere ainda alguns passos de transformação dos dados de forma a obter melhor resultados.

[6]Daqui em diante utilizar-se-á a expressão inglesa *Dataset* para fazer referência a um conjunto de dados.

2.2.4 BioGrid Toolkit

O BioGrid Toolkit [Barbosa and Monteiro, 2006] é uma aplicação que oferece um ambiente de programação paralela e distribuída que permite aos utilizadores criar, correr e visualizar aplicações paralelas e distribuídas em redes contendo recursos computacionais bastante heterogéneos, especificamente aplicações bioinformáticas de diversas áreas de investigação. O utilizador pode depois monitorizar o estado dos processos dessas aplicações.

2.3 Colaboração

2.3.1 O Projecto SolEuNet

Algum trabalho de colaboração foi feito num projecto de investigação chamado "*Data Mining and Decision Support for Business Competitiveness: A European Virtual Enterprise*" [Mladenic et al., 2003]. Este projecto envolveu doze equipas académicas e de negócio para formar uma empresa virtual cujo objectivo era o desenvolvimento de um *Data Mining* prático e soluções de apoio e suporte à decisão para clientes finais.
Foram apresentadas uma estrutura, métodos e ferramentas de *Data Mining* e de suporte de decisão, bem como uma aplicação para ajudar a solucionar problemas das empresas numa forma colaborativa.

O projecto aborda de uma forma geral as tecnologias usadas na análise dos dados, na solução de problemas de decisão e colaboração numa organização virtual de *Data Mining* e na tecnologia usada para o *Data Mining* colaborativo. Descreve também os aspectos relacionados com a integração de *Data Mining* e suporte à decisão no pré-processamento de dados e ao longo da estandardização e visualização. Finalmente são apresentadas algumas das vantagens práticas e limitações da estrutura para a resolução de problemas colaborativos numa empresa virtual formada por equipas colaborando principalmente e maioritariamente via *Web*.

2.3.2 SETI@home

O SETI@home[Anderson et al., 2002] cujo acrónimo em português significa "Procura de Inteligência Extra-Terrestre" é uma aplicação de computação distribuída criada em 1999 com base no projecto SETI, projecto que tem como objectivo a detecção da existência de possíveis comunicações entre seres extra-terrestres em planetas distantes da Terra. O projecto SETI analisa os sinais de rádio (ondas electromagnéticas) captados por radiotelescópios terrestres, pois estas representam a forma de transmissão de informação mais rápida conhecida.

O SETI@home possibilita que voluntários no mundo inteiro colaborem nesta investigação através da cedência de tempo de processador dos seus computadores para a análise dos dados captados e separados em pequenas fracções pelo projecto SETI, de forma a poderem ser analisados, na maioria dos casos em computadores pessoais.
Cada voluntário faz o *download* de uma certa quantidade de dados pela Internet, que depois de analisados automaticamente pela aplicação do SETI@home, são enviados novamente à equipa responsável pelo projecto. Cerca de 500 mil pessoas espalhadas por todo o mundo colaboram neste projecto.

Actualmente, a capacidade de processamento dos voluntários supera mesmo em mais de 5 vezes o volume de dados disponíveis para análise, apesar do projecto ainda não ter encontrado qualquer dado relevante relacionado com o objectivo base do projecto.

2.4 Porquê o WebCAD?

O WebCAD pode ser dividido em três áreas base: *Web*, Colaboração e Análise de Dados. Estas são as áreas em que o WebCAD pode ser muito útil nas actividades de investigação, nas quais este projecto pode actuar.

A nível da *Web* como foi referido, o ILP Oxford e o ILPnet2 são os dois principais sistemas existentes. Apesar de terem uma base de dados bastante completa de *Datasets* e referências para dados e informações de ILP, estes portais, apenas providenciam esta informação, são apenas repositórios de informação e de *Datasets*.

No projecto METAL é feito um trabalho semelhante ao WebCAD, na medida em que há uma análise de uma grande quantidade de dados através de técnicas de *Machine Learning* e *Data Mining*. A aplicação também é baseada na *Web*, tal como o WebCAD. Contudo, a aplicação apenas sugere qual o melhor algoritmo para os dados que o utilizador pretende analisar, não há qualquer controlo da execução desses mesmos algoritmos, nem é feita qualquer análise dos dados.
No projecto "BioGrid Toolkit", pelo contrário, o utilizador pode correr diversas aplicações de *Machine Learning* e *Data Mining*, bem como monitorizar o estado de execução dessas aplicações. Contudo, o "BioGrid Toolkit", não oferece ao utilizador a possibilidade de guardar e visualizar os resultados da execução destas aplicações.

Comparando o WebCAD com estes portais e aplicações, este diferencia-se na possibilidade de analisar os dados, por monitorizar a execução das experiências, na possibilidade de poder visualizar e guardar estes dados, bem como em permitir a alteração de parâmetros na execução da experiência. Permite também, interagir directamente com o sistema ILP, sem que para isso seja necessário conhecimentos técnicos de ILP, ou sem que para isso seja necessário um técnico especialista nestes sistemas.

No que respeita à dimensão colaborativa do WebCAD, e comparando especificamente com o projecto SolEuNet, embora este tenha sido um projecto com bastante sucesso, as equipas envolvidas trabalharam separadamente e cada equipa tinha apenas conhecimento do seu próprio projecto individual. Não foi partilhada qualquer informação, dados ou conhecimentos entre as doze equipas durante e após o desenvolvimento do projecto global.
No WebCAD, praticamente toda a informação com que o utilizador trabalha pode ser partilhada directamente na aplicação com outros investigadores dentro da mesma instituição, bem como outros investigadores noutras instituições que utilizem este sistema. Os utilizadores podem inclusivamente seleccionar qual a informação que pretendem partilhar e com que grupo de utilizadores querem partilhar essa informação. Os utilizadores podem ainda aceder a experimentações efectuadas por outros investigadores e, se desejarem, podem tentar melhorar os resultados dessas mesmas experimentações.

2.5 Conclusões

Tendo em conta, o estado da arte que aqui foi analisado, o projecto WebCAD pode colmatar alguma lacunas que existem actualmente na área do *Machine Learning* e do *Data Mining*. Principalmente no que diz respeito ao acesso a estas ferramentas por parte de pessoas que não sejam especialistas tecnicamente nestes sistemas, que até aqui necessitavam de técnicas especializados para poderem efectuar estas experimentações nestes sistemas, nomeadamente em ILP.
Além desta vantagem, os investigadores, podem também trabalhar, numa aplicação bastante amigável correndo num *Web browser*, e que lhes permite visualizar, monitorizar e executar as experimentações de uma forma fácil. Finalmente, podem também, partilhar o conhecimento adquirido com outros investigadores que trabalham na mesma área de investigação, fazendo com que não haja redundância de estudos efectuados e fazendo com que haja partilha de informação e conhecimentos de uma forma mais rápida, seja resultados, *Datasets* ou *Background Knowledge*.

Capítulo 3

Requisitos do Utilizador do WebCAD

3.1 Introdução

Neste Capítulo são descritos os requisitos do utilizador para o sistema WebCAD. Para modelar os requisitos dos utilizadores , são identificados e descritos os casos de utilização da aplicação em diagramas UML [Rumbaugh et al., 2004].
É também descrito o modelo de objectos do domínio através de um diagrama de classes UML por forma a clarificar os conceitos relevantes do domínio do problema.

3.2 Objectivos

O WebCAD tem por finalidade a inovação e aperfeiçoamento das actividades de investigação. Hoje em dia, pode-se atingir este desiderato principalmente com a ajuda de melhoramentos tecnológicos. Identificam-se assim, neste projecto, algumas áreas onde há possibilidade de evidentes melhorias para as actividades de investigação e de descoberta científica.
A colaboração com outros investigadores no mesmo campo de investigação; a facilidade de uso de recursos computacionais para executar experiências sem necessidade de especialização na ferramenta a utilizar; a junção de equipas de investigação interdisciplinares; o uso de ferramentas para poder aceder a informação importante e

valiosa; o fornecimento de resultados das experiências e utilizar outros resultados e dados de outros investigadores da mesma área. Estas são as áreas principais em que o WebCAD considera que pode contribuir para o melhoramento das actividades de investigação.
Para alcançar este objectivo, a aplicação proposta, o WebCAD, permite que os investigadores trabalhem com uma ferramenta de análise automática de dados, bem como uma ferramenta que permite a partilha de conhecimentos, tendo também uma característica colaborativa, que permite preencher os requisitos das características das actividades de investigação científica acima mencionadas.

3.3 Requisitos Gerais

Nesta Secção são identificados e descritos os requisitos gerais da solução, comuns a todos os casos de utilização. Estes requisitos estão associados ao tipo de utilizadores que irão utilizar a aplicação bem como aos dados que irão ser utilizados na solução que irá ser apresentada.

3.3.1 Aplicação Web

Os utilizadores do WebCAD são, na sua maioria, investigadores científicos e desde o início do desenvolvimento do projecto foi assumido que na sua maioria estes investigadores não são especializados nem em ciência de computadores nem sistemas de ILP. Deste modo, a aplicação deverá ser desenvolvida para ser de fácil utilização e com um aspecto gráfico "amigável", para que qualquer investigador possa facilmente utilizar a aplicação sem que tenha um conhecimento profundo de informática. O interface gráfico "amigável" permite a qualquer utilizador executar tarefas de análise de dados de forma a que não necessite saber detalhes da ferramenta de análise de dados, nomeadamente como parametrizar o sistema ILP.

O WebCAD deverá ser desenvolvido para poder responder a estas necessidades. É uma aplicação *Web*, corre portanto num browser, sendo um sítio que é acessível na *Web* para utilizadores pré-autorizados que pretendam usar o WebCAD. Os utilizadores e os investigadores estão organizados na aplicação em grupos relacionados com as instituições de investigação a que pertencem, o que faz com que possam

partilhar dados e resultados com outros investigadores dentro da mesma instituição de forma directa e fácil. Este tipo de organização dos utilizadores permite também a partilha de resultados, experiências e *Datasets* entre instituições que utilizem o WebCAD. Integrando cada sítio de WebCAD numa rede de sítios permite aos utilizadores assim, executar e trocar dados e informações no que respeita a análise de dados e *Datasets*. Esta rede permite que todos os dados envolvidos e obtidos na análise de dados sejam partilhados por todos os utilizadores e investigadores que usem a aplicação.

Esta característica ajuda os investigadores a colaborar com outros no mesmo campo de investigação e dá-lhes oportunidade de utilizar resultados de experiências de outros investigadores da mesma instituição ou de qualquer outra que use o WebCAD, desde que essa informação seja previamente seleccionada para ser partilhada ou pública.

3.3.2 *DataSets* e *Background Knowledge*

A aplicação mantém numa bases de dados o registo de todos os dados necessário para as experiências. Os dados necessários às experiência são os *Datasets* (conjuntos de dados) e o *Background Knowledge* (conhecimento de fundo). Estes dados são necessários aos investigadores para fazerem as experiências e para analisarem todos os resultados obtidos.

Os *Datasets* e o *Background Knowledge* são inseridos nas base de dados dos sítios WebCAD pelos investigadores. Alguns destes dados, os mais comuns e mais utilizados, já estão nas base de dados, dependendo da área científica do investigador. Os *Datasets* podem também ser descarregados na *Web* nos repositórios de *Datasets* de ILP, depois disso ser enviados para o servidor WebCAD num formulário de envio específico, onde o utilizador pode parametrizar e definir o *Dataset* e também o *Background Knowledge*.

3.3.3 Experiências

Um dos objectivos do projecto é o de permitir que os investigadores trabalhem com uma ferramenta de análise de dados poderosa num ambiente amigável, como

é um *browser*.
No WebCAD, a ferramenta utilizada para a análise de dados, como foi já referido anteriormente, é o ILP.
Devido à complexidade desta ferramenta, apenas utilizadores experientes e especializados nestas ferramentas de análise de dados conseguem utilizá-la de forma correcta e eficaz. Um utilizador experiente é capaz de efectuar tarefas como o afinamento dos parâmetros e tarefas que necessitem conhecimentos específicos dum sistema deste tipo.
No WebCAD, para fazer experiências de análise de dados, o utilizador necessita apenas de seleccionar os dados que pretende analisar, seleccionar algumas opções para a experiência e lançá-la. Pode ainda, efectuar outras experiências, com base em resultados de experiências, alterando apenas para isso os parâmetros desejados, como por exemplo: o tempo de execução; a precisão; o número de cláusulas; e outras.

3.4 Modelo de Casos de Utilização

No WebCAD, o utilizador consegue controlar toda a informação e dados envolvidos na aplicação. Existem casos de utilização diferentes para cada secção da aplicação. No diagrama de casos de utilização, que se apresenta na Figura 3.1 são apresentados os casos de utilização da aplicação. De seguida são descritos cada um dos casos de utilização, bem como ilustrados com diagramas de actividades em UML [Rumbaugh et al., 2004], por forma a detalhar as actividades envolvidas em cada um dos casos.

3.4.1 Diagrama de Casos de Utilização

A aplicação tem cinco casos de utilização principais: Fazer *Login*, Enviar Dados, Visualizar Dados e Resultados, Executar Experiências e Pesquisar Dados e Resultados. Além destes casos de utilização existem outros casos de utilização mais específicos, descritos mais à frente.

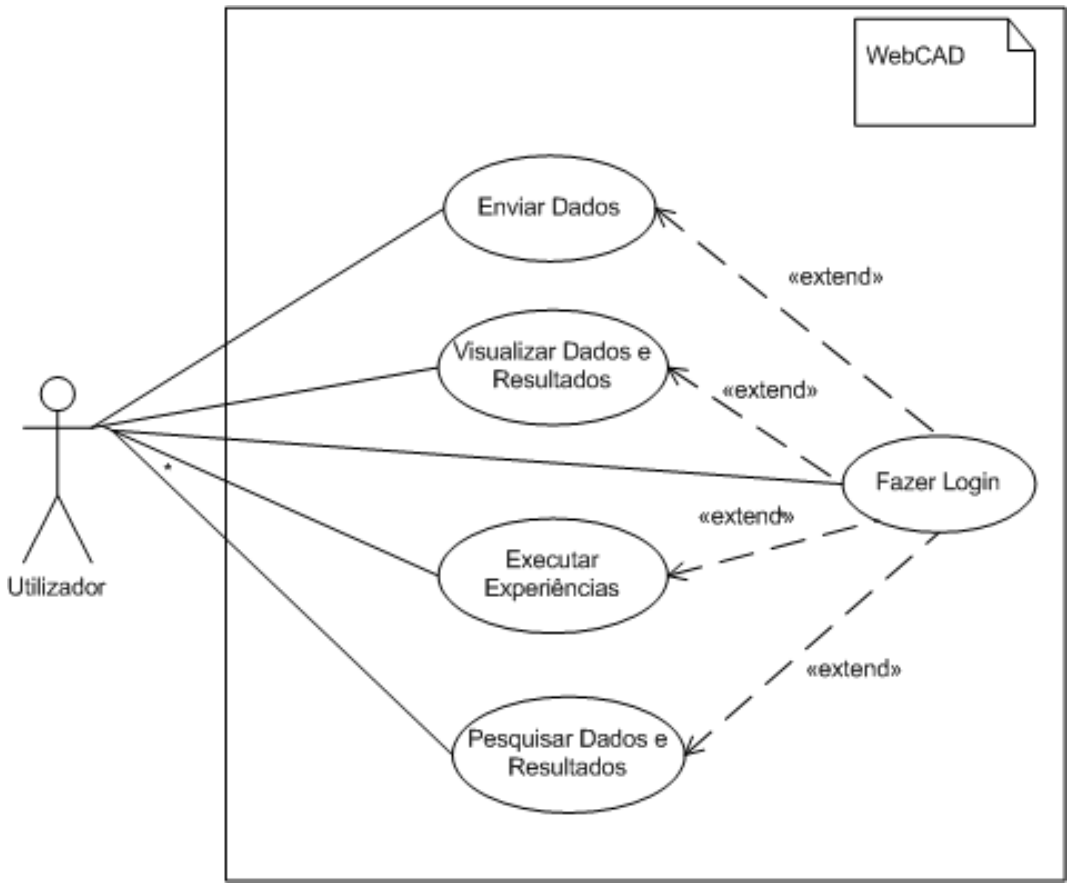

Figura 3.1: Diagrama de Casos de Utilização

3.4.2 Fazer *Login*

Para se poder efectuar alguma operação na aplicação é necessário fazer *Login*. O utilizador necessita inserir o seu nome de utilizador e *password* para ter acesso às funcionalidades do WebCAD. Isto permite ao utilizador ter privacidade nos dados das suas experiências e melhor visualização das mesmas.

3.4.3 Enviar Dados

A aplicação permite o envio de todos os dados necessários à execução de experiências. No caso de utilização ilustrado na Figura 3.3, o utilizador tem de seleccionar o domínio a que o *Dataset* pertence, descrevê-lo com toda a informação relacionada, preenchendo todos os campos com as definições do *Dataset*.

3.4.3.1 Enviar *Dataset*

Ao inserir a informação relacionada com o *Dataset*, o utilizar necessita de enviar três ficheiros. Esses ficheiros são os ficheiros necessários para a realização de uma experiência de análise de dados no sistema ILP. São um ficheiro de exemplos negativos, um ficheiro de exemplos positivos e o um *Background Knowledge* por omissão. Sem estes três ficheiros, o *Dataset* não é inserido, pois são ficheiros imprescindíveis para a execução destas experiências em ILP.

3.4.3.2 Enviar *Background Knowledge*

Neste caso de utilização, o utilizador pode também inserir *Background Knowledge* na base de dados. Para isso, tem de preencher também um formulário, de forma semelhante à inserção do *Dataset* e escolher o tipo de definição do *Background Knowledge*.

No acto de inserção de ambos estes tipos de dados, estes podem ser classificados como públicos ou privados. Existe ainda a possibilidade de seleccionar estes dados como públicos ou privados dentro da instituição em que o utilizador está integrado ou para outros sítios WebCAD de outras instituições.

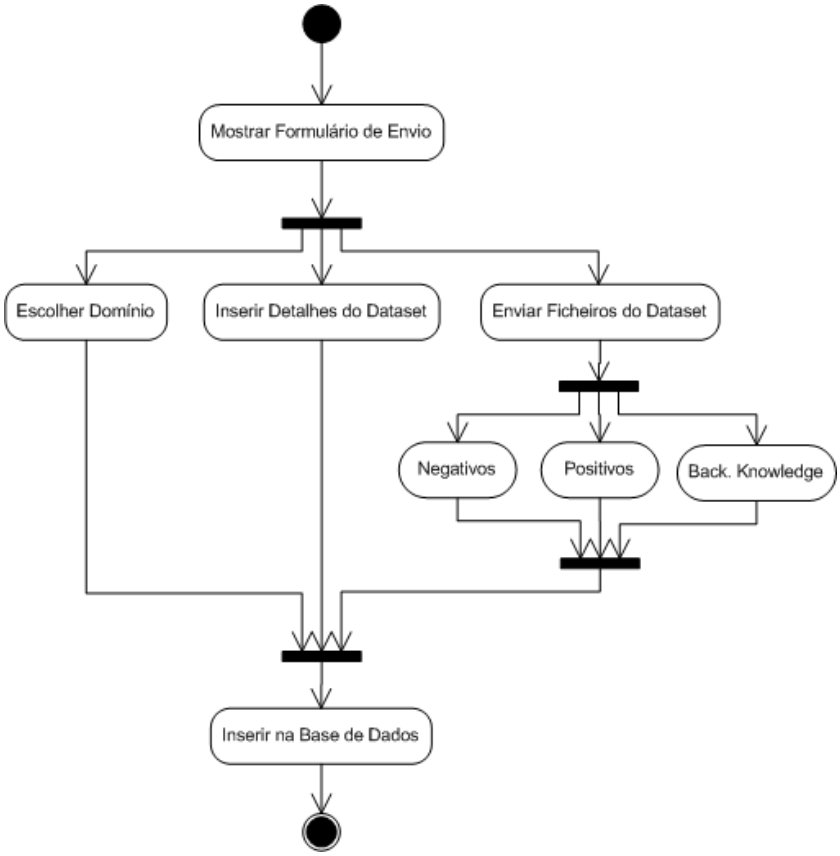

Figura 3.2: Enviar *Dataset*

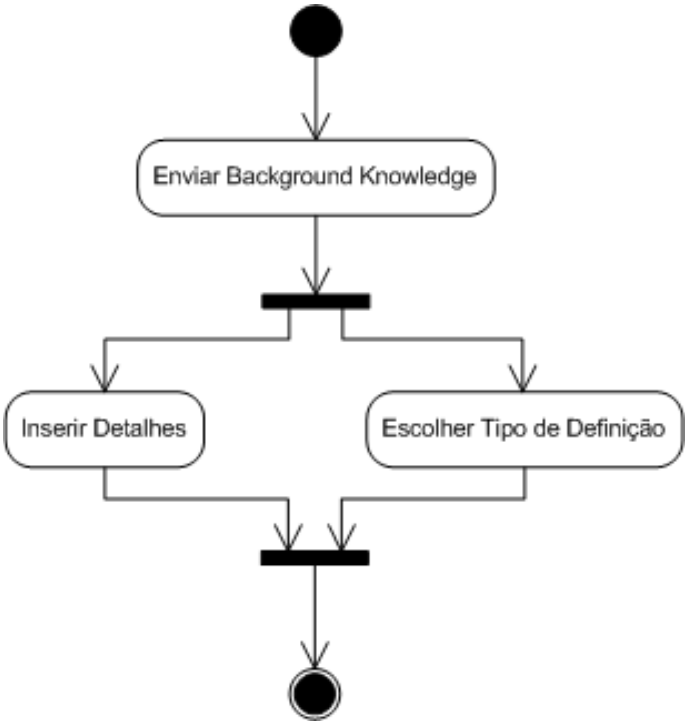

Figura 3.3: Enviar *Background Knowledge*

3.4.4 Visualizar Dados e Resultados

Este é um dos casos de utilização mais importantes de toda a aplicação. Aqui, o utilizar pode visualizar todos os resultados das experiências, os dados e as informações existentes na base de dados para a análise de dados.
A visualização dos dados está separada em: visualização dos dados para análise, dados que irão ser analisados no sistema ILP e os dados obtidos com as todas as experiências efectuadas no WebCAD.

No que respeita aos dados a ser analisados, a aplicação permite que o utilizador visualize todos os *Datasets* e *Background Knowledge* que existem no sistema. Esta visualização dos dados pode ser mais eficaz, pois o utilizador pode filtrá-la pelo domínio em que o *Dataset* e o *Background Knowledge* se insere, o que permite visualizar e encontrar a informação desejada de forma mais rápida.

A visualização dos resultados das experiências está também organizada de forma a poder ser feita de forma mais fácil e mais rápida. Está portanto, dividida, em

duas partes: informação relacionada com experiências que já foram consideradas pelo utilizador/investigador como terminadas e informação relacionada com experiências que ainda não foram terminadas e que podem inclusivamente ter dados ainda em análise pelo sistema ILP.
Em ambas, o utilizador pode visualizar toda a informação relacionada com a experiência. Nomeadamente, data de execução, parâmetros escolhidos, *Dataset* e *Background Knowledge* utilizados, resultados obtidos, entre outras informações devolvidas pelo sistema ILP. Pode também visualizar sub-experiências que foram efectuadas dentro duma experiência, usando parâmetros diferentes, depois de uma primeira experiência.

É neste caso de utilização que o utilizador selecciona quais são os resultados das experiências que pretende partilhar, ou pelo contrário, manter privados, com outros utilizadores da mesma instituição ou de outras instituições com sítios WebCAD.

Finalmente, para ajudar à visualização de um grande conjunto de dados, está disponível um conjunto de filtros para mostrar apenas os resultados desejados. O utilizador pode também ordenar os resultados da visualização por cada um dos campos presentes na Tabela.

3.4.5 Executar Experiências

Neste caso de utilização, a aplicação controla a execução de experiências de análise de dados com a ajuda de um sistema ILP.

Para executar uma experiência, o utilizador tem de escolher primeiro os parâmetros necessários. Tem de escolher o *Dataset*, o *Background Knowledge* e o método de avaliação pretendido. Depois de efectuar esta selecção, o utilizador lança a experiência.
Após esta acção, a aplicação mostra ao utilizador o estado da experiência. Ou seja, se foi iniciada correctamente, ou se pelo contrário não foi possível iniciá-la. Além desta informação, é mostrada toda a informação relacionada com a experiência: data de execução, *Dataset* e *Background Knowledge* escolhido, utilizador, entre outras.
Ao lançar a experiência, o utilizador selecciona se a pretende partilhar ou não,

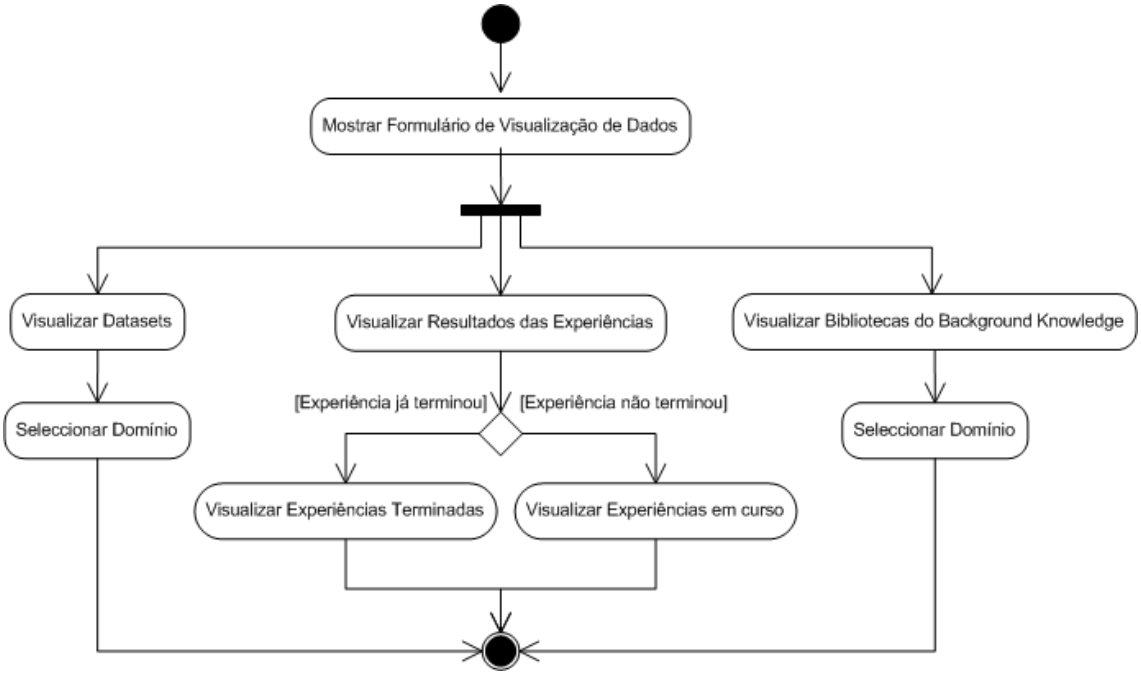

Figura 3.4: "Visualização de Dados e Resultados"

segundo as regras já definidas anteriormente no caso de utilização "Enviar Dados".

Esta experiência, pode ser uma sub-experiência de outra, detalhada no caso de utilização "Gerir Resultados de uma Experiência".

3.4.6 Pesquisar Dados e Resultados

Neste caso de utilização, a aplicação permite a pesquisa de dados e resultados partilhados por outros utilizadores/investigadores de outras investigações que utilizam o WebCAD. E colaborar na investigação de experiências desses utilizadores, que é uma das principais propostas e objectivos do WebCAD.

Neste caso de utilização é disponibilizada toda a informação partilhada pelos utilizadores de todos os sítios WebCAD que existem dentro da rede. Esta informação é disponibilizada através de duas secções: localmente, no servidor local, e externamente, em todos os outros sítios de WebCAD.

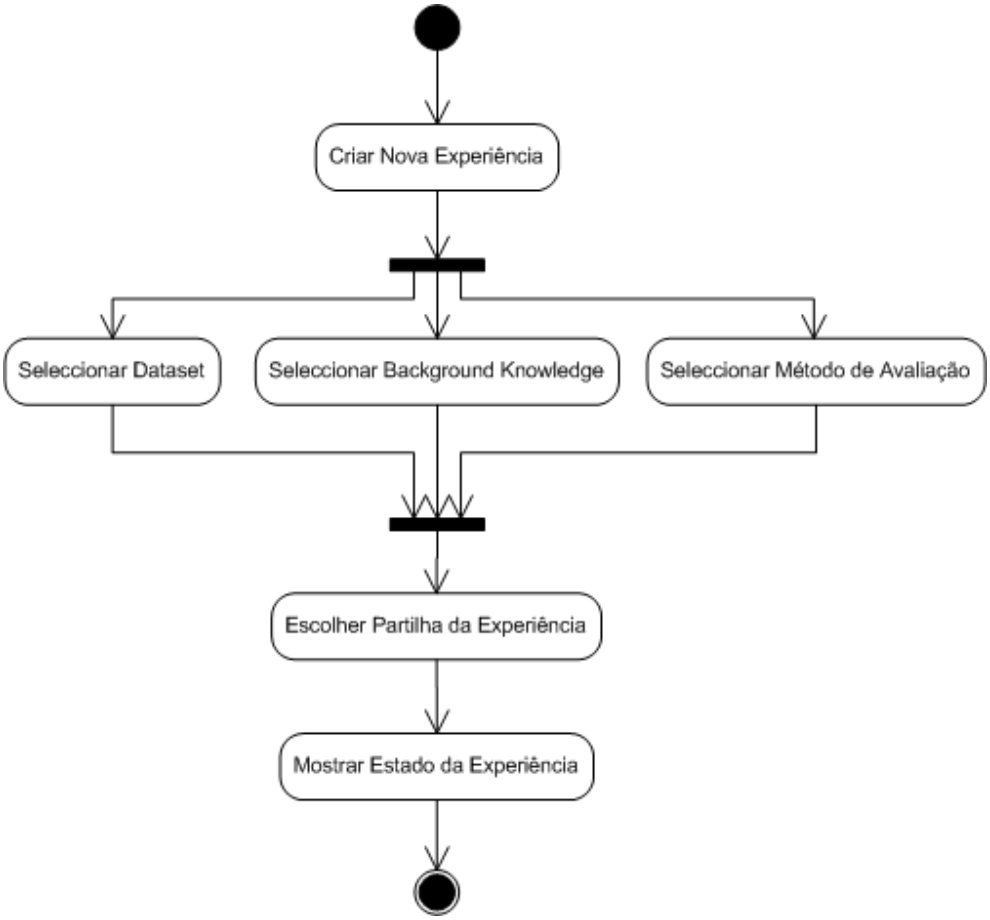

Figura 3.5: "Executar Experiências"

No sítio local, o utilizador visualiza toda a informação partilhada pelos utilizadores da instituição que utilizam a aplicação. Essa informação pode ser um *Dataset*, *Background Knowledge* ou um resultado de uma experiência. Toda a informação que é partilhada, seja de que tipo for, pode ser visualizada nesta secção. O utilizador pode apenas visualizar e utilizar essa informação, em caso algum, o utilizador, tem a possibilidade de alterar essa informação.

Nos outros sítios WebCAD, o utiliza visualiza a informação que foi seleccionada especificamente para ser partilhada com outros sítios WebCAD. Aqui também, a informação partilhada pode ser um *Dataset*, *Background Knowledge*, o resultado de uma experiência, etc.

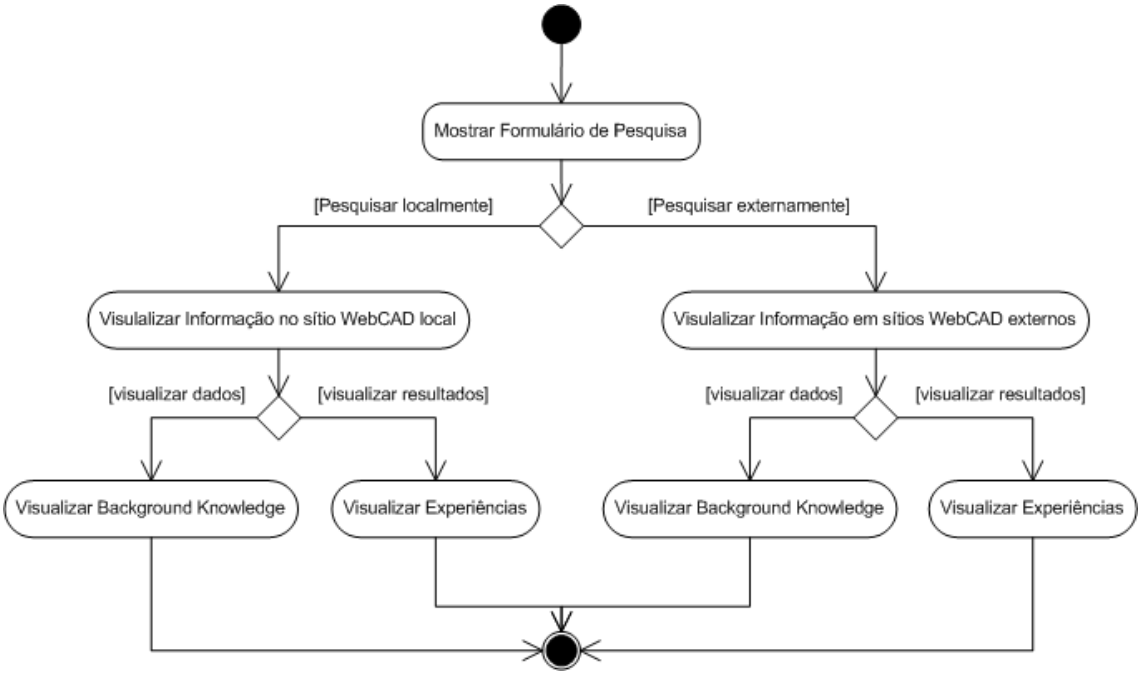

Figura 3.6: "Pesquisar Dados e Resultados"

3.4.7 Gerir Resultados de uma Experiência

Este é um caso de utilização específico, associado à visualização de resultados de uma experiência.

Neste caso o utilizador tem duas actividades possíveis: visualizar os resultados de uma experiência ou executar uma experiência com base nos resultados de uma experiência anterior, refazendo essa experiência através da alteração de parâmetros.
No primeiro caso de utilização, o utilizador visualiza toda a informação relacionada com a experiência: data de execução, parâmetros escolhidos, *Dataset* e *Background Knowledge* utilizados, resultados obtidos, entre outras informações devolvidas pelo sistema ILP. Pode também visualizar as outras sub-experiências, já detalhadas anteriormente.

No segundo caso de utilização, a aplicação oferece ao utilizador, a possibilidade de executar uma experiência com base nos resultados de outra experiência previamente efectuada através da alteração dos parâmetros associados à experiência.
O utilizador visualiza os resultados e os parâmetros utilizados na experiência, e caso entenda, pode alterar os parâmetros desta experiência para obter novos resultados. Para isso, selecciona e altera os parâmetros que deseja modificar.
Depois disso, verifica se a selecção dos parâmetros é válida. Ou seja, a aplicação verifica se a experiência com os novos parâmetros seleccionados irá produzir novos resultados. Caso isso aconteça, a experiência será executada já com esses novos parâmetros. Caso contrário, a aplicação informa o utilizador que a selecção actual dos parâmetros não produzirá novos resultados. O utilizador pode alterar a selecção dos parâmetros as vezes que desejar até obter uma resposta válida.
Caso o utilizador verifique que não irá produzir novos resultados nesta experiência pode sair, sem lançar a experiência.

3.5 Modelo de Objectos do Domínio

Para se identificar as entidades do domínio do problema e descrever os relacionamentos entre elas, apresenta-se o Diagrama de Objectos do Domínio através de um diagrama de classes UML.

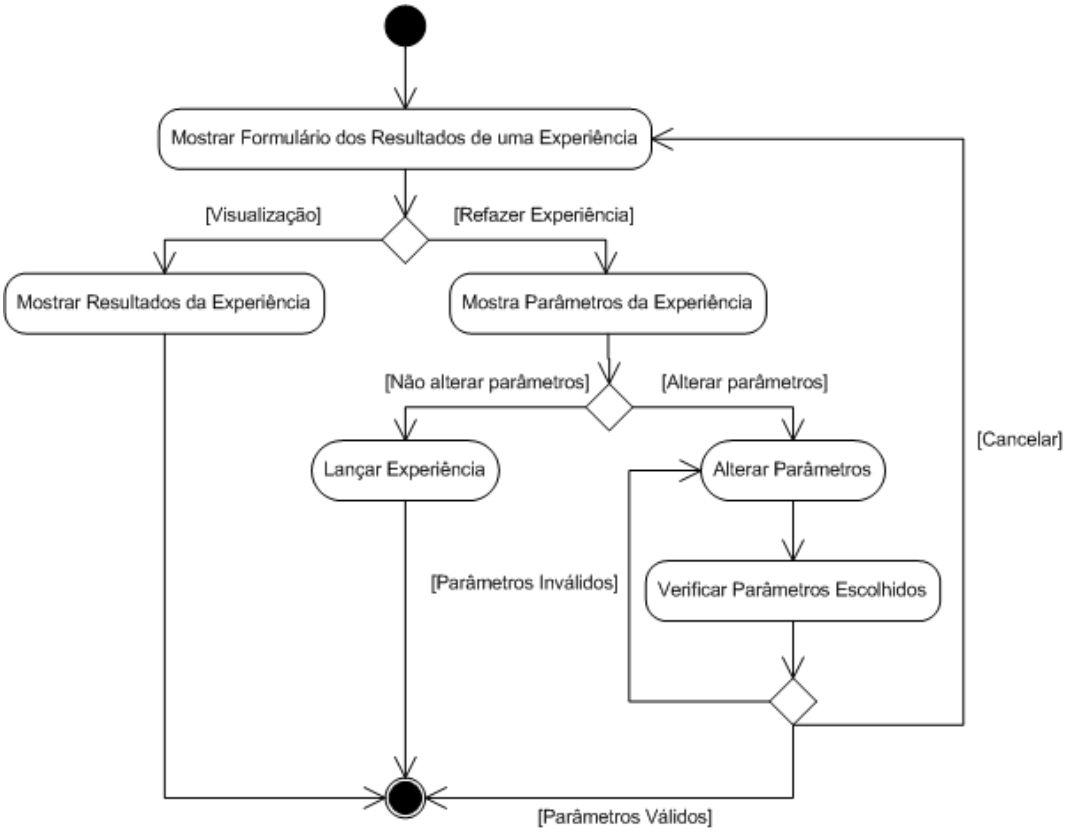

Figura 3.7: "Gerir Resultados de uma Experiência"

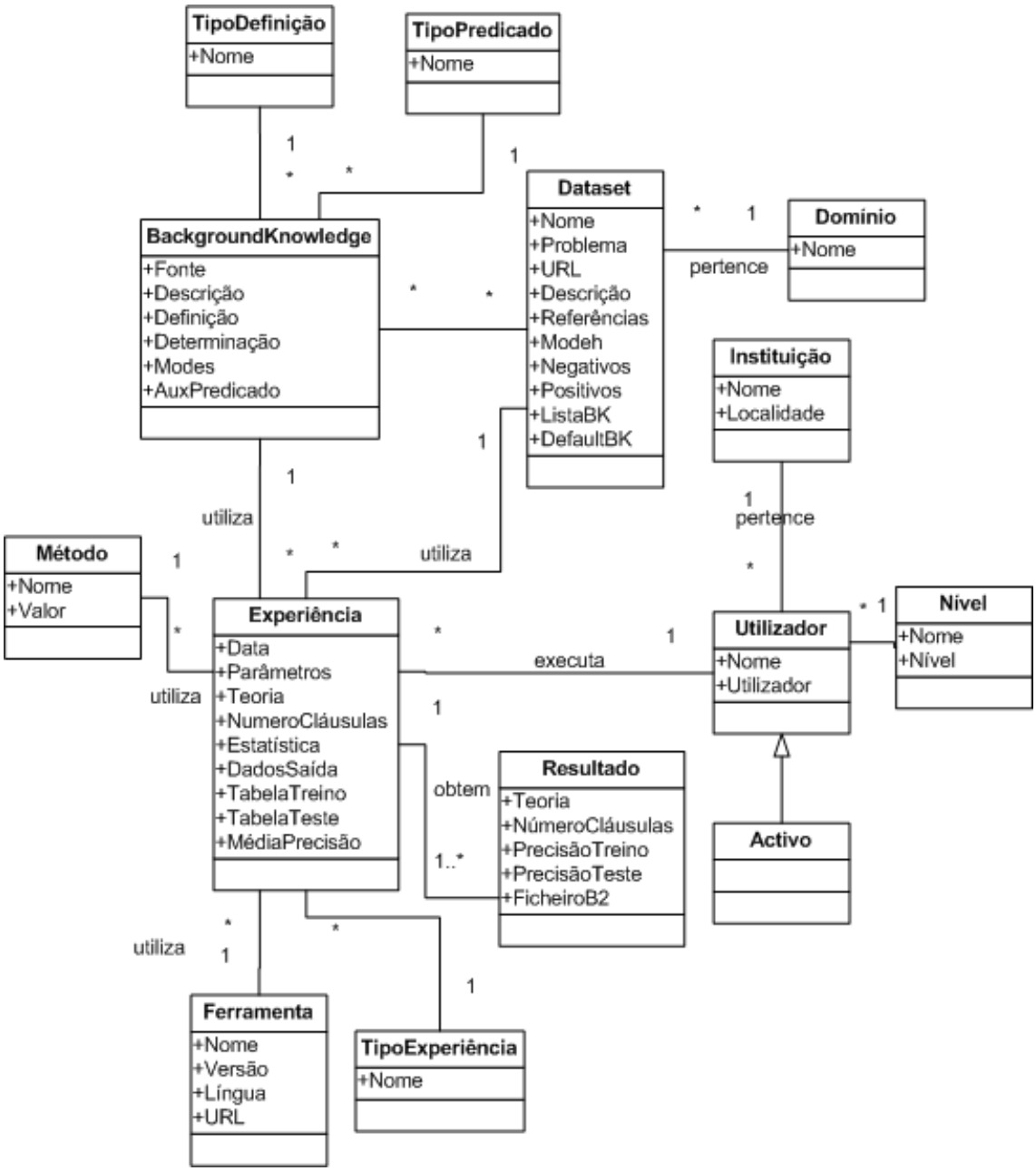

Figura 3.8: Modelo de Objectos do Domínio

Capítulo 4

Arquitectura do WebCAD

4.1 Introdução

Neste Capítulo é apresentada a arquitectura da solução proposta para o WebCAD, pensando e respeitando os requisitos funcionais e não funcionais contidos na Especificação de Requisitos.
Com a apresentação da arquitectura, pretende-se descrever a solução proposta, bem como as tecnologias que irão ser utilizadas para na implementação da aplicação a ser desenvolvida.

É apresentada a Arquitectura Lógica da solução através de um Diagrama de Pacotes de Classes em UML [Rumbaugh et al., 2004]. São também apresentados cada um dos pacotes de classes necessários no WebCAD.

É também apresentada a Arquitectura Física através de um Diagrama de Componentes, que permite perceber a estrutura física do projecto, e.g. onde ficam as classes alojadas e um Diagrama de Distribuição que permite identificar a topologia do sistema, os nós de *hardware* e as ligações entre eles.

Finalmente, é apresentada a Arquitectura Tecnológica, onde são apresentadas as tecnologias que irão ser utilizadas no WebCAD e é descrita a função dentro da arquitectura da solução.

4.2 Arquitectura Lógica

Nesta Secção é apresentada a Arquitectura Lógica da solução proposta, através de um Diagrama de Pacotes de Classes em UML. O Diagrama apresentado na Figura 4.1 está divido em cinco camadas horizontais e decomposto em diversos pacotes de classes para melhor organização. Horizontalmente, o diagrama está dividido nas camadas de *Graphic User Interface*; Páginas *Web* dinâmicas; Lógica de negócio no servidor *Web*; Camada de acesso a dados e Sistema IndLog.

O WebCAD sendo uma aplicação no *browser* e disponível na *Internet*, possuirá um servidor *Web* para responder aos pedidos HTTP do utilizadores que utilizem a aplicação. O sistema possuirá também um servidor IndLog onde será feita a análise (semi)-automática dos dados.
O pacote de classes da *Graphic User Interface*, englobará todas as classes e funções do interface gráfico do *browser* utilizado para aceder à aplicação.
No pacote de classes das Páginas *Web* dinâmicas estarão as classes necessárias para o utilizador ou investigador poder visualizar os dados e resultados utilizados na aplicação. Estas página serão geradas dinamicamente consoante as escolhas do utilizador, e consoante os resultados a apresentar. Estas páginas serão geradas dinamicamente pelo Servidor *Web*.

Dentro do pacote do Servidor *Web* existirão diversos pacotes necessários para a aplicação correr num *browser*. Num desses pacotes, o Servidor HTTP, estarão as classes necessárias para receber os pedidos HTTP do *browser*. Esses pedidos serão criados com base na informação enviada pela lógica de negócios do investigador. Dentro do pacote Servidor *Web* estará também o pacote com a Lógica de negócio do Servidor. Neste pacote estarão as classes que possibilitarão a publicação do WebCAD na *Web* e que permitirão a ligação com o pacote de classes da base de dados e com o pacote de classes da lógica de comunicação entre o Servidor *Web* e o Servidor IndLog.
O pacote de classes da base de dados permitirá guardar todos os dados necessários para a análise de dados, bem como resultados e informações das experiência e outra informação relacionada com o WebCAD.
No pacote Lógica de Comunicação estarão todas as classes, processos e protocolos utilizados na comunicação entre o Servidor *Web* e o Servidor IndLog.

No Servidor IndLog, existirão dois pacotes de classes, o pacote de Lógica de Comunicação e o pacote Sistema IndLog. No pacote Lógica de Comunicação presente no Servidor Indlog irão estar todas as classes, processos e protocolos utilizados na comunicação entre o Servidor IndLog e o Servidor *Web*.
No pacote de classes do Sistema IndLog estarão todos os processos e classes utilizadas pelo Sistema IndLog para fazer a análise (semi)-automática de dados.

4.3 Arquitectura Física

Pretende-se agora mostrar a estrutura física do sistema que será implementado. Para isso são apresentados o Diagrama de Componentes e o Diagrama de Distribuição. Estes diagramas vão permitir perceber a implementação que será feita e de como será feita a distribuição dos diversos componentes de software no hardware a utilizar.

4.3.1 Componentes do Sistema

Para poder-se visualizar quais os componentes presentes, apresenta-se um Diagrama de Componentes na Figura 4.2. Este diagrama contém quais os componentes e interfaces que o sistema irá conter, bem como as relações existentes entre os componentes.

O Pacote de Páginas Dinâmicas do WebCAD serão a parte visível do sistema. Ou seja, serão o nível superior de todo o sistema, pois são através das páginas dinâmicas que o utilizador poderá interagir com a aplicação, visualizar os dados e resultados além de executar experiências.
Nestas páginas existirá um componente principal que é o Menu Principal, index.php. Existirão páginas para diversas funções como a autenticação e apresentação do projecto. Mas os componentes principais, que no diagrama estão agrupadas em pacotes serão as secções de Realização de Experiências; Enviar Dados; Visualizar Resultados e Pesquisa de Dados e Resultados. Todos estes conjuntos de páginas serão apresentadas no *browser* do utilizador, ou seja, numa máquina cliente.
Estas páginas são geradas pelos Componentes Lógica de Negócio. Estes com-

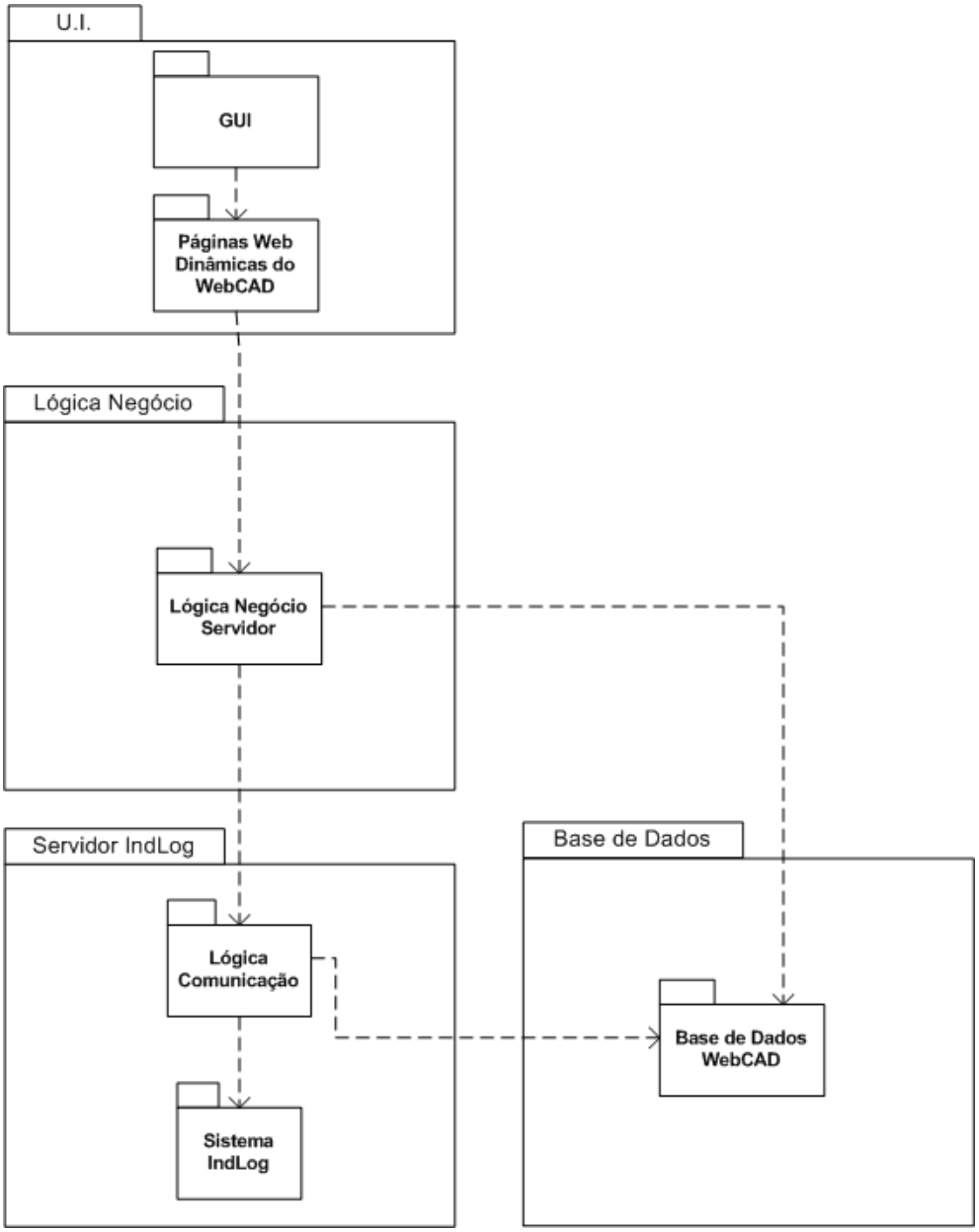

Figura 4.1: Diagrama de Pacote de Classes do WebCAD

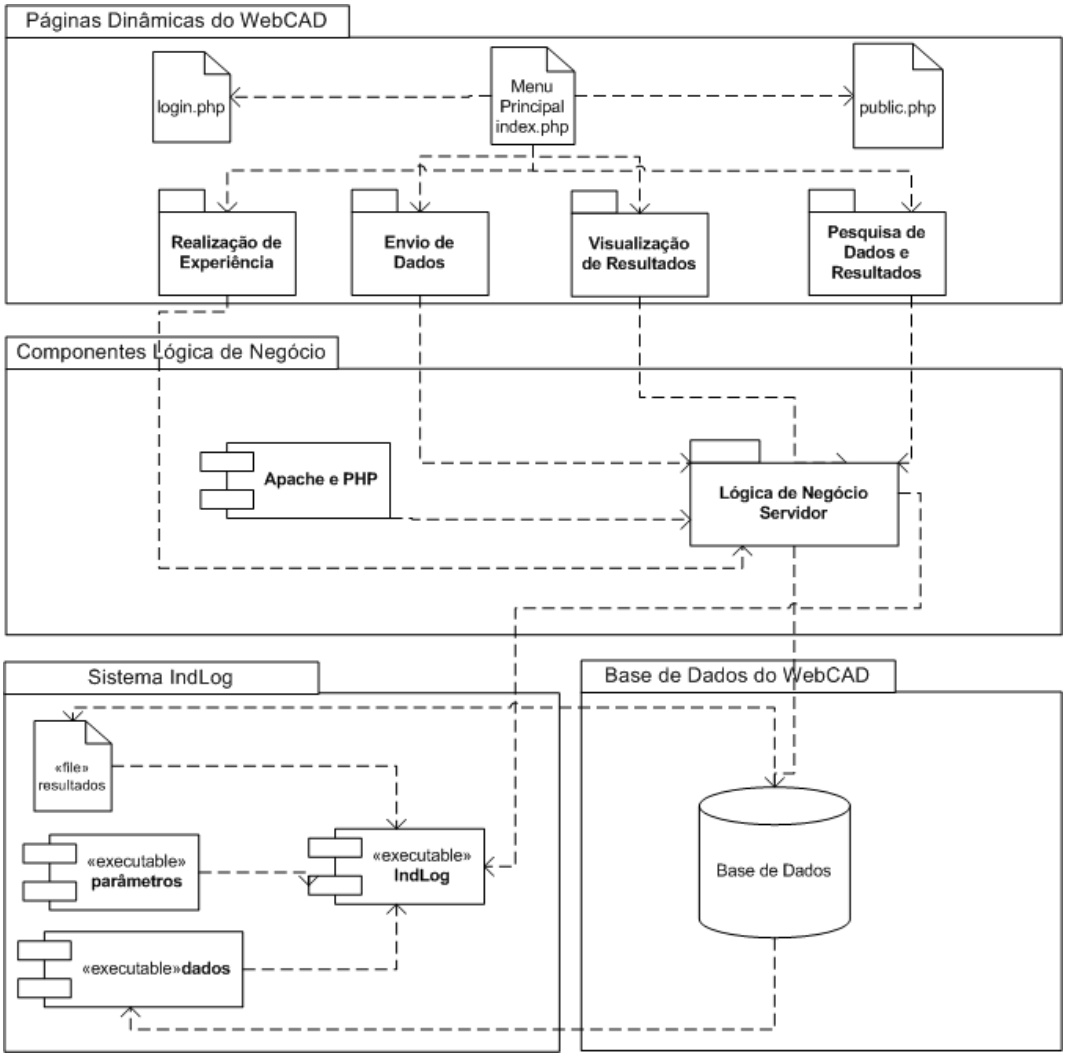

Figura 4.2: Diagrama de Componentes do WebCAD

ponentes são constituídos por diversos ficheiros de código em PHP, a tecnologia utilizada no servidor *Web*, que será visto em mais detalhe na Secção Arquitectura Tecnológica 4.4.
Nas Componentes Lógica de Negócios existirá também um componente chamado Ligação à Base de Dados que fará a ligação entre o Servidor *Web*, os ficheiros de código escritos em PHP e a Base de Dados, e que permitirá ler da base de dados para depois esses dados serem mostrados nas Páginas Dinâmicas do WebCAD, geradas através do PHP.

A Base de Dados será criada e manipulada no MySQL, também detalhada na Secção 4.4. Na Base de Dados irão estar armazenados os dados necessários para as experiências, informação relacionada com utilizadores e resultados de experiências efectuadas.
No Servidor Indlog irão estar ficheiros que possibilitarão a ligação ao Sistema IndLog que irá fazer a análise dos dados. O Sistema IndLog irá posteriormente guardar os resultados das experiências, que serão depois inseridos na Base de Dados.

4.3.2 Distribuição Física

Nesta Secção pretende-se mostrar a conceptualização das topologias locais de hardware dos sistemas sobre os quais a aplicação de software irá ser implementada. A Figura 4.3 ilustra um Diagrama de Distribuição (*Deployment*) que mostra o ambiente de hardware do sistema WebCAD ao qual são executados os componentes de software.

Visto ser uma aplicação em browser, o WebCAD deve funcionar em qualquer sistema operativo que tenha um *browser* instalado. Na máquina cliente, o utilizador necessitará apenas do *browser*, não sendo necessário a instalação de qualquer *applet* ou qualquer *plugin* para que a aplicação seja executada no *browser*.

No servidor, estará instalado um servidor *Web* Apache e um servidor de base de dados MySQL. O servidor *Web* irá interpretar os pedidos HTTP do *browser* do utilizador e possibilitará a ligação à Componente Lógica de Negócio.
A Componente Lógica de Negócio do WebCAD irá fazer a geração das páginas

dinâmicas que irão ser mostradas ao utilizador. Para poder mostrar essas páginas, a componente irá ler informação da base de dados.

No Servidor IndLog, estará disponível o Sistema IndLog que possibilitará a análise dos dados que será requerida pelo utilizador.
Para possibilitar a ligação à Componente Lógica de Negócio do WebCAD, no servidor IndLog estarão ficheiros <<*batch*>> que executarão o IndLog com os parâmetros e configurações seleccionadas pelo utilizador.
O IndLog fará a análise de dados e escreverá os resultados dessas experiências em ficheiros que servirão de suporte aos ficheiros <<*batch*>> para inserir esses resultados e informações na Base de Dados.

4.4 Arquitectura Tecnológica

As tecnologias que serão utilizadas no desenvolvimento do projecto serão um servidor *Web*, um Sistema de Gestão de base de dados e uma linguagem interpretada para possibilitar a criação da lógica de negócio do WebCAD e da lógica de apresentação.

O servidor *Web* utilizado será o Apache Web Server [Apache, 1995], um dos servidores *Web* mais utilizados na *Internet*. O Apache Server é um software livre e é capaz de executar código em diversas linguagens como o PHP, Perl, Shell Script e funciona como servidor HTTP, essencial para o funcionamento do WebCAD.

A linguagem interpretada que será utilizada é o PHP [PHP, 1994]. É uma das linguagens de programação suportadas directamente no servidor Apache. A Figura 4.4 ilustra a interacção entre o PHP, o servidor *Web* e o *browser* do utilizador. É um software livre, à semelhança de todas as tecnologias que irão ser utilizadas neste projecto. É uma linguagem bastante poderosa, orientada a objectos apesar de ser bastante simples. Caracteriza-se também por poder ser executada em vários sistemas operativos como o Windows, Unix, Linux, entre outros.
Permite uma conexão directa com bases de dados relacionais ao contrário de outras ferramentas que necessitam de drivers ODBC para poder conectar-se a uma base de dados. Entre as bases de dados que o PHP permite a conexão directa estão o

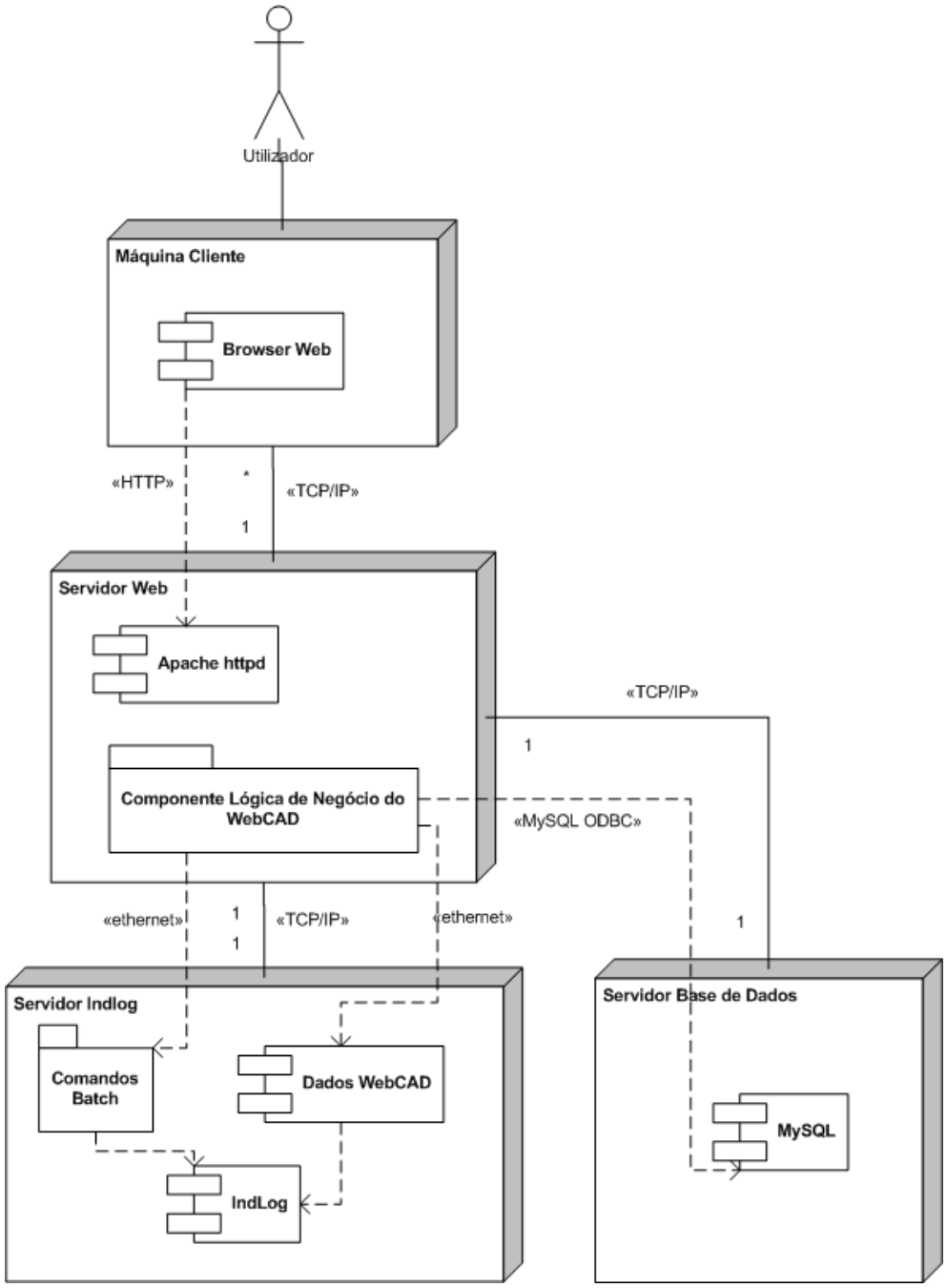

Figura 4.3: Diagrama de Distribuição do WebCAD

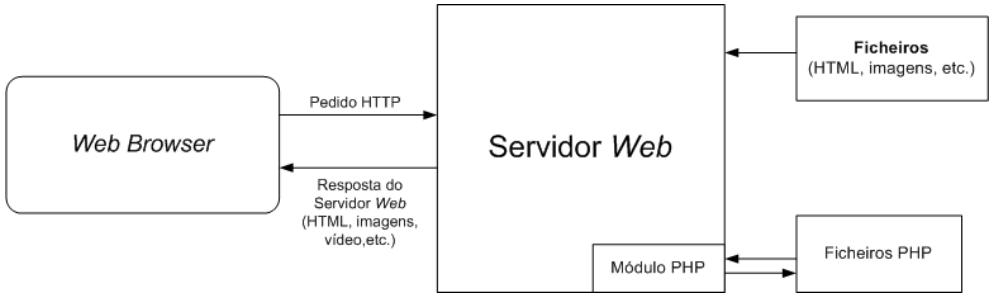

Figura 4.4: Interacção entre o PHP, o servidor *Web* e o *browser* do utilizador

Oracle, Sybase, Postgresql e o MySQL, o Sistema de Gestão de Base de Dados que será utilizado neste projecto. O PHP tem ainda suporte aos principais protocolos de *Internet* IMAP, SNMP, NNTP, POP3, HTTP, LDAP, XML-RPC, SOAP, sendo ainda possível abrir sockets e interagir com outros protocolos.

Como já foi referido, o Sistema de Gestão de Base de Dados que irá ser utilizado é o MySQL [MySQL, 2006]. É um sistema de gestão de base de dados livre e utiliza o SQL, como linguagem de consulta e controle de dados. Caracteriza-se pela sua simplicidade e eficiência e está optimizado para aplicações baseadas na *Web*. Devido à sua fácil integração com o PHP, foi o servidor de base de dados escolhido para o desenvolvimento do WebCAD. À semelhança do PHP, funciona também em todos os principais sistemas operativos existentes actualmente.

Capítulo 5

Implementação

5.1 Introdução

Neste Capítulo descrevem-se os aspectos mais relevantes relativos à implementação do sistema WebCAD. A implementação foi feita com base na arquitectura proposta para este projecto e tendo em conta os seus requisitos identificados.

Primeiro será apresentada a forma como foi implementada a estrutura e configuração do sistema, ao nível dos servidores e configuração das aplicações utilizadas. De seguida será apresentado como está estruturado cada um dos sítios WebCAD. Depois será detalhada a implementação da aplicação WebCAD. Será apresentada a base de dados criada para armazenamento dos dados do sistema, bem como a forma como foi implementada a transferência dos ficheiros dos *Datasets* para o servidor WebCAD.

Serão depois apresentados alguns detalhes de implementação a nível da visualização dos dados e resultados e a execução das experiências de análise de dados. Finalmente será apresentada a implementação da pesquisa de dados e resultados noutros sítios WebCAD e a implementação que foi feita de forma a ser possível sequenciar experiências, utilizando resultados de experiências anteriores.

5.2 Instalação e Configuração do Sistema

A estrutura física do sistema criado possui um servidor *Web* onde está alojado o sítio WebCAD. O servidor *Web* que foi utilizado é o Apache e nesta implementação, este é executado numa máquina Linux.
A estrutura possui ainda um servidor de base de dados MySQL e um interpretador PHP utilizado para a lógica de negócio. Nesta implementação optou-se por configurar tanto a base de dados MySQL como o interpretador PHP no mesmo servidor, pelo que cada sítio WebCAD que foi desenvolvido foi instalado e executa a partir de uma só máquina.
O IndLog, que realiza a análise dos dados, é executado na mesma máquina do servidor *Web*. Apesar de grande parte das experiências efectuadas demorarem bastante tempo a ser executadas e as várias experiências executadas em paralelo requererem um esforço grande por parte do processador, na solução implementada utilizou-se apenas uma máquina para ambos os servidores.
Nos outros sítios WebCAD desenvolvidos, o servidor WebCAD, que contém o servidor *Web*, a base de dados e o PHP, estão também na mesma máquina que o servidor do IndLog, apesar de em termos de eficiência ser melhor separar os dois servidores.

Na Figura 5.1 é apresentada uma representação esquemática de como foi implementado o WebCAD.

Dado que o sistema WebCAD, possui também uma componente colaborativa, este pressupõe a instalação de vários sítios WebCAD. Cada um dos sistemas WebCAD pode ser instalado de forma diferente desde que esteja acessível via *Web*, e que respeite a arquitectura definida.

5.3 O Sítio WebCAD

O sistema WebCAD sendo uma aplicação Web, foi desenvolvido para estar disponível para qualquer utilizador que o queira utilizar. Cada sítio WebCAD, tem uma secção pública e uma secção privada. Na secção pública onde estão disponíveis diversas informações sobre o projecto, o utilizador pode requerer o acesso ao WebCAD para si ou para a sua Instituição, como se mostra na Figura 5.2.

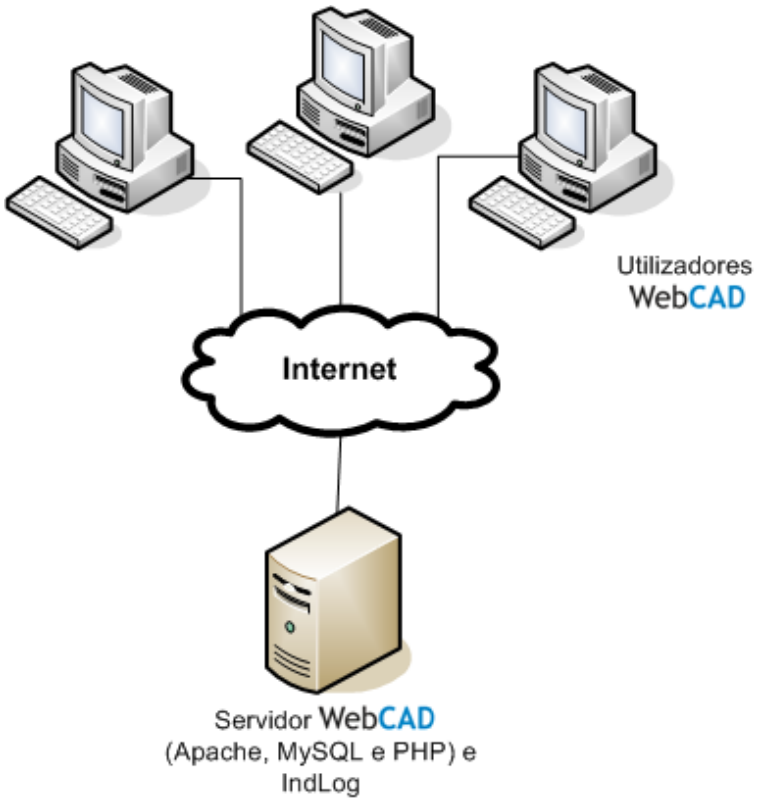

Figura 5.1: Representação Esquemática da Disposição Física do Sistema WebCAD

Figura 5.2: Página inicial do WebCAD

Ao requerer o acesso ao sistema WebCAD, este enviará um e-mail aos administradores do sistema que validarão e disponibilizarão os acessos requeridos, se assim o desejarem.

5.4 A Aplicação WebCAD

Dentro da secção privada do sítio WebCAD, depois de devidamente autenticado, o utilizador poderá aceder à aplicação em si, onde poderá fazer as suas experiências de análise de dados.
A linguagem PHP é utilizada para criar a estrutura do sítio, mas também para criar a lógica de negócio da aplicação. É utilizado ainda para a criação dos formulários de interacção com o utilizador e para o *upload* dos dados a serem analisados para o servidor. Para fazer a ligação com o servidor do IndLog, são utilizado *batch files* armazenados no servidor IndLog que vão executar a realização da análise dos dados evocando o sistema IndLog.

5.4.1 Armazenamento de Dados

Para se poder armazenar toda a informação relacionada com as experiências, como os *datasets*, *Background Knowledge*, resultados, além de dados necessários à aplicação, criou-se uma base de dados MySQL. Esta base de dados tem a mesma estrutura para todos os sítios que forem instalados.

A Figura 5.3 ilustra o esquema da base de dados relacional que foi criado de forma a poder armazenar convenientemente esta informação. O esquema da base de dados está na forma normal de BCNF [Ramakrishnan and Gehrke, 2002], por isso é garantido que não existe redundância da informação devido às Dependências Funcionais.

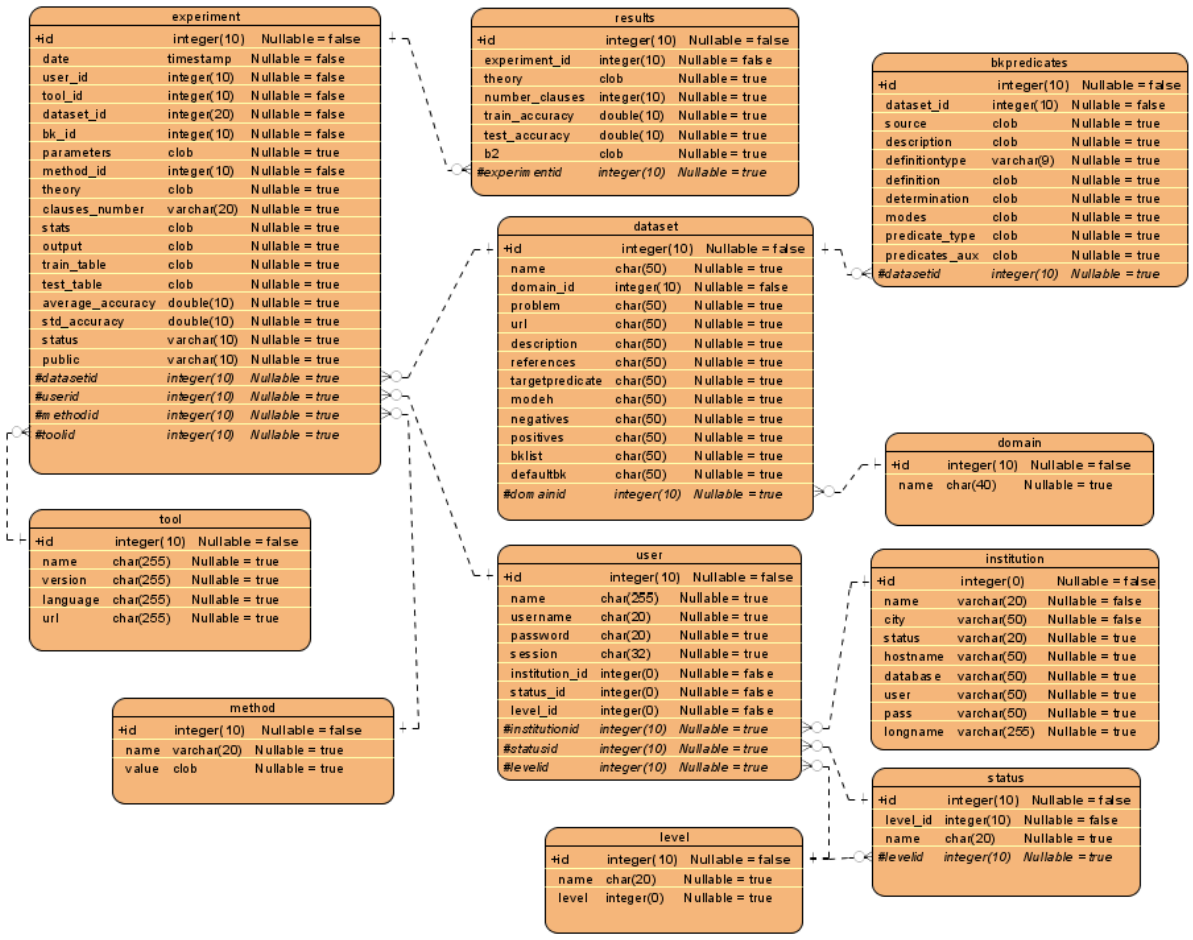

Figura 5.3: Esquema da Base de Dados do WebCAD

5.4.2 Exemplos e *Background Knowledge*

Os Exemplo e o *Background Knowledge* que são os dadoss[1] necessários para a realização das experiências de análise de dados são ambos guardados na base de dados. Toda a informação relacionada com o *Background Knowledge* é inserida na base de dados, enquanto que no caso dos *Datasets*, só é armazenada informação relacionada com o *Dataset* (*Meta data*). Os dados para análise são guardados em ficheiros num directório onde estão armazenados os dados de todos os *Datasets* existentes no sistema WebCAD.

Nome da Tabela	Dados armazenados
Experiment	Experiências realizadas
Results	Resultados das experiências realizadas
Bkpredicates	Dados relacionadas com o *Background Knowledge*
Tool	Ferramentas utilizadas para aquisição dos dados das experiências
Method	Métodos de Avaliação utilizados na realização das análises de dados
Dataset	Conjuntos de dados a ser analisados
User	Utilizadores registados do sistema WebCAD
Domain	Domínios científicos dos *Datasets*
Institution	Empresa ou Organização registada no sistema WebCAD
Level	Nível de utilizador
Status	Estado do utilizador

Tabela 5.1: Tabelas Utilizadas na Base de Dados

5.4.3 Dados e Resultados das Experiências

Ao ser realizada uma nova experiência, é armazenada na base de dados a informação relacionada com ela, bem como os parâmetros e configurações que foram utilizados para a sua realização.
Para monitorizar a execução da experiência, é utilizado um *batch file*, cujo processo será detalhado na Secção 5.4.6.

[1] O *Dataset* em ILP refere-se habitualmente aos exemplos positivos (E^+), exemplos negativos (E^-) e ao *Background Knowledge* B.

Posteriormente à execução da experiência, são armazenados na base de dados os resultados e estatísticas da experiência de análise de dados efectuadas.

5.4.4 Envio de Dados no WebCAD

Para possibilitar o envio dos ficheiros necessários para cada *Dataset* é utilizado um formulário de *upload* em PHP. Neste formulário, o utilizador selecciona na sua máquina os três ficheiros necessários para inserir um *dataset*. Estes três ficheiros enviados são os exemplos negativos e positivos e o *Background Knowledge*.

Programa 1 Excerto de código utilizado para envio dos dados de um *Dataset*.

```
for($x=1;$x<4;$x++)
{
$file_name = $_FILES['uploadFile'. $x]['name'];

$file_name = stripslashes($file_name);
$file_name = str_replace(‘‘’’’,’’’’,$file_name);
$copy = copy($_FILES['uploadFile'. $x]['tmp_name'],''datasets/''.$file_name);
$nomes[$x]=$file_name;

 if($copy)
 {
 $mensagem=''File ‘‘.$file_name.'' uploaded to WebCAD Server''.''<br>'';
 }
 else
 {
 $i++;
 }
}
```

No que respeita ao *Background Knowledge*, este é inserido na base de dado, através da inserção dos predicados em Prolog que são utilizados e outros parâmetros que lhe estão associados, não sendo necessário o envio de qualquer ficheiro.

5.4.5 Visualização dos Dados e Resultados

Para visualizar os dados existentes no sistema, foram criados diversos ecrãs que possibilitam, de uma forma intuitiva, a visualização da informação.

Nestes ecrãs é possível a visualização de Exemplos, *Background Knowledge* e os resultados das experiências, divididos por "Experiências Terminadas" e "Experiências em Execução". O Utilizador pode definir a experiência como terminada quando assim o desejar. Existe também um *batch file* no servidor IndLog, que verifica se a experiência já terminou ou está em execução.

Programa 2 *Batch file* "indlogstatus" que permite verificar o estado de execução de uma determinada experiência

```
#!/bin/csh

set resultsFile=‘echo "/tmp/experiencia"$2‘

echo ‘ps -f -u $1 |grep yap |grep $resultsFile|awk ’{print $7}’‘
```

Este *batch file* verifica o processo em execução no servidor IndLog e escreve num ficheiro o estado de execução da experiência.

Na visualização dos dados, o utilizador pode também filtrar os resultados e pode aplicar regras de ordenação em todos os campos das grelhas de visualização.
No caso da visualização dos resultados das experiências, o utilizador, pode optar por visualizar apenas as suas experiências ou visualizar todas as experiências dentro da sua experimentação. Nesta grelha o utilizador pode também seleccionar se uma experiência é privada ou pública.

5.4.5.1 Organização dos resultados

Para melhor visualização dos resultados de uma experiência, pois uma experiência pode conter diversos resultados, como é detalhado em 5.4.8, são utilizadas camadas de HTML.

Estas camadas permitem ao utilizador mostrar ou esconder os resultados de uma experiência. As Figuras 5.4 e 5.5 ilustram a selecção dos resultados de uma experiência para visualização.

A expansão e a contracção dos resultados permite uma melhor organização dos dados, especialmente quando a quantidade de dados é considerável.

Programa 3 Excerto de código *javascript* que permite ocultar ou mostrar dados relacionados com uma experiência

```
function show(planoid) {
mostra(planoid);
changeimage('seta' +planoid,'images/close.gif');
document.getElementById('seta' + planoid).parentNode.href =
"javascript:hide('" + planoid + "');";
}
function hide(planoid) {
esconde(planoid);
changeimage('seta'  +planoid,'images/open.gif');
document.getElementById('seta' + planoid).parentNode.href =
"javascript:show('" + planoid + "');";
}
```

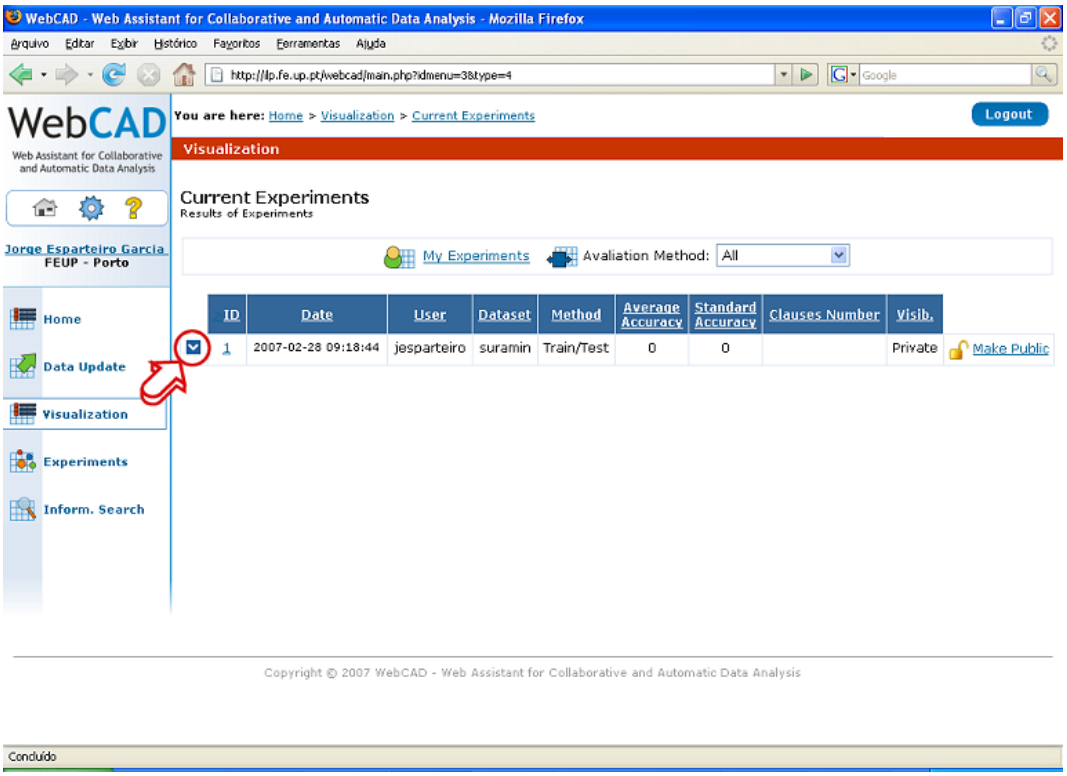

Figura 5.4: Selecção dos Resultados de uma Experiência para Visualização

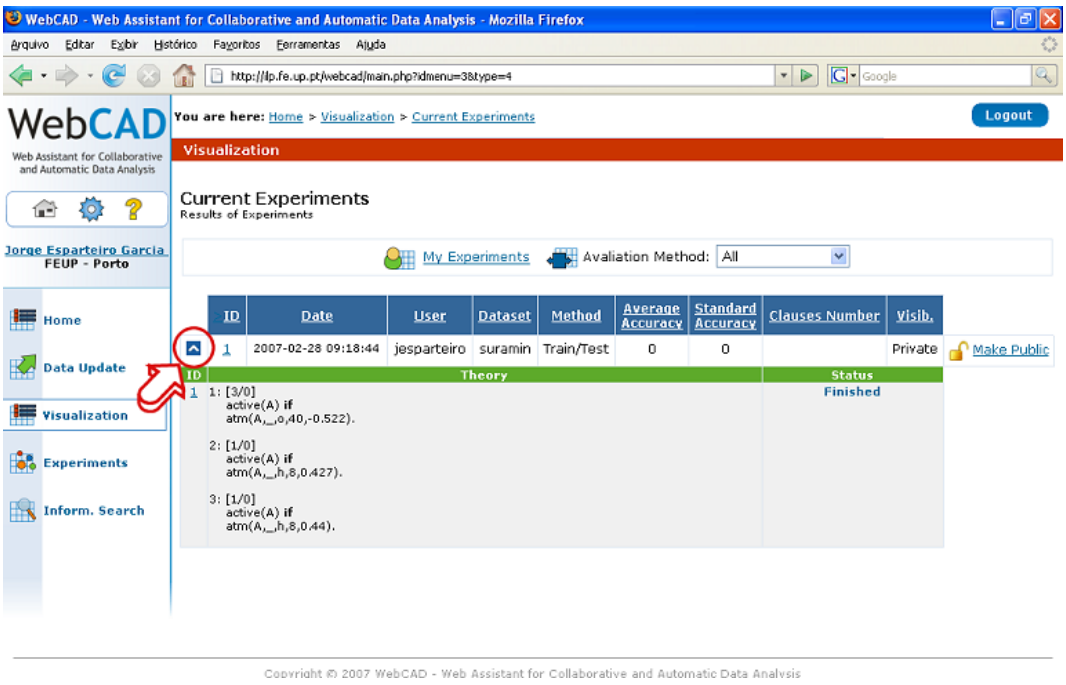

Figura 5.5: Expansão dos Resultados de uma Experiência

5.4.6 Execução das Experiências de Análise de Dados

Para a análise de dados pelo IndLog foi necessário criar *batch files* que permitissem que o utilizador, a partir de um *browser*, fosse capaz de executar experiências de análises de dados.

Estes *batch files* permitem lançar o IndLog para fazer a análise dos dados, bem como controlar e monitorizar a execução da experiência. Na Tabela estão detalhados os *batch files* criados.

Batch file	Função
filterIndLogOut	Filtrar os resultados obtidos
indlogParameterUpdate	Alteração dos Parâmetros de uma experiência
indlogstatus	Consultar o estado de execução da experiência
nextSettings	Verificação de Parâmetros válidos
runindlog	Executa a experiência e retorna os resultados
stopindlog	Pára a experiência

Tabela 5.2: *Batch Files* Utilizados no WebCAD

Os *batch files* são executados no PHP e consoante o comando a utilizar, estes podem escrever os dados directamente na base de dados ou em ficheiros auxiliares que serão posteriormente lidos e processados pela aplicação.

O *runindlog* é um dos *batch files* mais importantes criados, pois permite a execução de uma experiência. Além da execução do Indlog, actualiza os dados na base de dados relativos à experiência e escreve num ficheiro os resultados da análise dos dados. Estes dados são posteriormente lidos e processados pelo sistema WebCAD que os disponibiliza para que o utilizador visualize os resultados da experiência que executar.

Programa 4 *Batch file* "runindlog" executado pelo PHP que permite a execução do sistema IndLog

```
#!/bin/csh

set dataSet=$1
set resultsFile=‘echo "/tmp/"$2‘
set theoryFile=‘echo "/tmp/"$2".theory"‘
set outFile=‘echo "/tmp/"$2".out"‘
if ($#argv != 2) then
   echo "syntax: runindlog datasetStem outfile"
   exit
endif

/usr/local/bin/yap /usr/local/bin/savedIndlog -- $dataSet $outFile
 >& $resultsFile
/usr/local/bin/filterIndLogOut theory $2 > ${resultsFile}.theory
mysql -u webcad webcad --password=’pass’ < ${resultsFile}.theory

set nclauses=‘grep "\." ${resultsFile}.theory|wc -l‘
mysql -u webcad webcad --password=’pass’ -e "UPDATE results SET
number_clauses=$nclauses WHERE id=$2"
set trError=‘grep "Train Accuracy" ${resultsFile}|awk -F’=’ ’{print $2}’‘
mysql -u webcad webcad --password=’pass’ -e "UPDATE results SET
train_accuracy=$trError WHERE id=$2"
# rm -f $resultsFile ${resultsFile}.theory
```

institution		
+id	integer(0)	Nullable = false
name	varchar(20)	Nullable = false
city	varchar(50)	Nullable = false
status	varchar(20)	Nullable = true
hostname	varchar(50)	Nullable = true
database	varchar(50)	Nullable = true
user	varchar(50)	Nullable = true
pass	varchar(50)	Nullable = true
longname	varchar(255)	Nullable = true

Figura 5.6: Estrutura da Tabela *Institution*

5.4.7 Pesquisa de Dados noutros Sítios WebCAD

A Pesquisa de dados noutros sítios WebCAD é a funcionalidade do WebCAD que permite que haja trabalho colaborativo.
O WebCAD possibilita que sejam partilhados *Datasets*, *Background Knowledge* e resultados de experiências efectuadas entre os utilizadores dos diferentes sítios WebCAD instalados.
Todos os Exemplos e *Background Knowledge* presentes num sítio podem ser descarregados por utilizadores de outros sítios WebCAD.
Cada utilizador pode também visualizar os resultados das experiências partilhadas pelos outros utilizadores, mas não tem possibilidade de alterar essa informação, nem de efectuar outras experiências sequenciando uma experiência partilhada.

5.4.7.1 Conexão a um Sítio WebCAD

Para ser possível aceder a outros sítios WebCAD e visualizar informação partilhada, todas as base de dados dos sítios WebCAD criados possuem um utilizador padrão, com permissões limitadas, que consegue ler os dados partilhados. Os dados de cada uma das instituições que utiliza o WebCAD são inseridos em todos os sítios WebCAD existentes, como ilustra a Figura 5.6 que mostra a estrutura da Tabela *Institution* da base de dados. Aí estão armazenados para além dos dados da instituição, os dados que permitem a ligação aos diversos sítios para ler a informação partilhada por cada instituição.

Como o WebCAD é uma aplicação *Web*, pode acontecer, por vezes, que o sítio

esteja *offline*, para evitar que aconteçam erros e para informar o utilizador de quais os sítios onde poderá pesquisar informação, são feitos testes de conexão a todas as bases de dados exteriores.
Na aplicação estão representados graficamente os sítios *online* e os que estão indisponíveis, como ilustra a Figura 5.7.

Programa 5 Excerto de código para testar conexão a um sítio WebCAD

```
<?php
$link = mysql_connect('$server', '$user', '$pass');
if (!$link) {
    print 'Couldn´t connect to $server : ' . mysql_error());
    $statusimg=$OFFLINE_IMG;
}
print  'Connection established';
$statusimg=$ON_IMG;
mysql_close($link);
?>
```

O utilizador pode pesquisar por informação num sítio WebCAD específico ou pode pesquisar simultaneamente em todos os sítios WebCAD. Caso o utilizador pesquise em todos os sítios simultaneamente, irá ser mostrada toda informação partilhada existente nos sistemas WebCAD.

5.4.8 Sequenciar Experiências

Uma das funcionalidades presentes no WebCAD é a de permitir sequenciar experiências.
O utilizador pode alterar algumas características do modelo anterior criado, sem conhecer os parâmetros do IndLog. Para tal, o utilizador indica quais as características do modelo que gostaria de ver alteradas. O utilizador pode indicar se pretende mais ou menos cláusulas no modelo. Se pretende cláusulas mais curtas ou compridas, se pretende esperar mais tempo pelos resultados e se pretende mais ou menos precisão no modelo. Os parâmetros mais relevantes no IndLog podem ser alterados no WebCAD indirectamente pelo utilizador para obter diferente resultados, sem que conheça os parâmetros. Na Tabela 5.3 são apresentados os parâmetros mais relevantes do IndLog e que podem ser alterados indirectamente pelo utilizador do WebCAD.

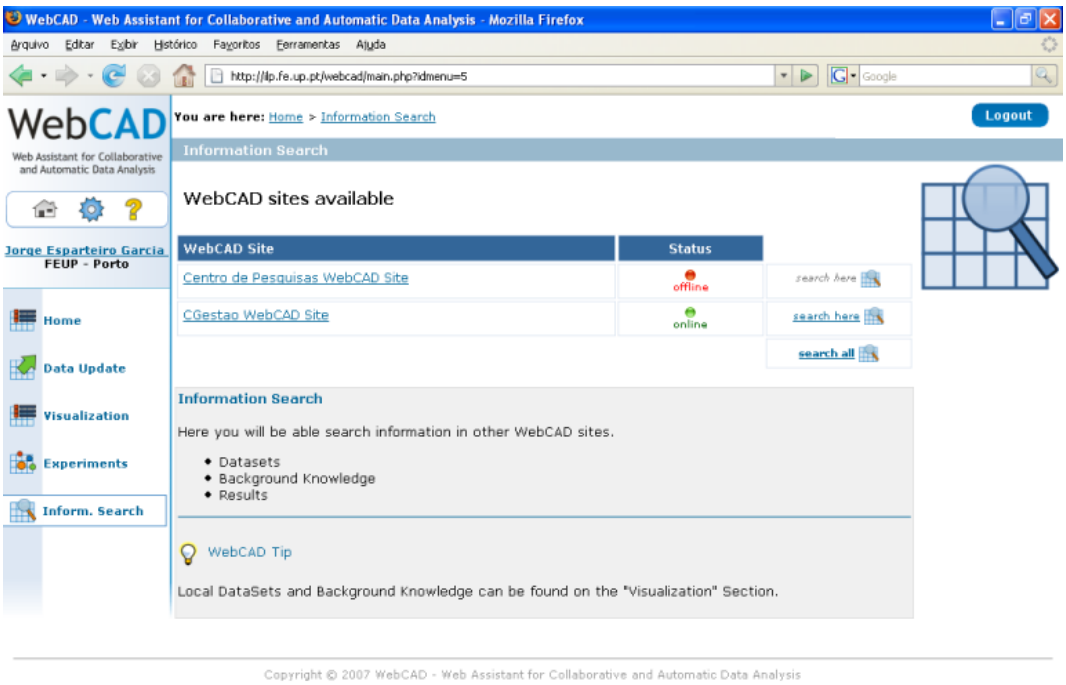

Figura 5.7: Pesquisa de Dados Partilhados

Função	Nome Parâmetro
Cobertura máxima para uma cláusula ser aceite	*mincover*
Número de nós construídos e avaliados	*nodes*
Comprimento máximo das cláusulas	*clauselength*
Exclusão de cláusulas	*forbidden* (*clausule*)
Accuracy	*noise* ou *accuracy*
Lazy evaluation	*covertest*

Tabela 5.3: Parâmetros Relevantes no IndLog

Para alterar os parâmetros desejados, o utilizador abre uma experiência anterior e aí pode visualizar o modelo constituído nessa experiência e informação sobre a experiência (cláusulas do modelo, tempo de execução, etc.).
Antes da experiência ser novamente executada, é necessário verificar se esses parâmetros irão produzir novos resultados. Esta verificação dos parâmetros será feita pelo *batch file nextSettings*. Este irá ser executado com os parâmetros que o utilizador seleccionar e verificará se a experiência com estes novos parâmetros irá produzir novos resultados.

Ficheiro 1 Exemplo de um ficheiro .next gerado

```
+
=
=
-
1
*
145
1
63.64
0
%%% Learning Bias %%%
:- set(i, 4).
:- set(language,2).
:- set(clauselength,10).
:- set(heuristic, pn).
:- set(samplesize,50).
:- set(noise,20).
:- set(mincover,10).
:- set(nodes, 100).
```

Para fazer essa verificação, irá ser gerado um ficheiro *.next*, como ilustra a Figura 5.8, com os parâmetros seleccionados pelo utilizador. Nesse ficheiro estará também todo o histórico de resultados obtidos para a experiência.

O comando *nextSettings* fará então a análise das opções do utilizador. De seguida mostrará as razões pelo qual os parâmetros seleccionados não produzirão novos resultados. Ou, caso contrário, executará uma nova experiência de análise de dados utilizando os novos parâmetros.
Para verificar se a alteração dos parâmetros produz ou não novos resultados foi criado o Algoritmo 1. Este algoritmo permite verificar se os parâmetros seleccionados pelo utilizador irão ou não produzir novos resultados se for executado. Permite também saber quais os parâmetros que não podem ser ajustados para cada experiência.

Para melhor compreensão do Algoritmo 1 são detalhadas na Tabela 5.4, as funções utilizadas no Algoritmo.

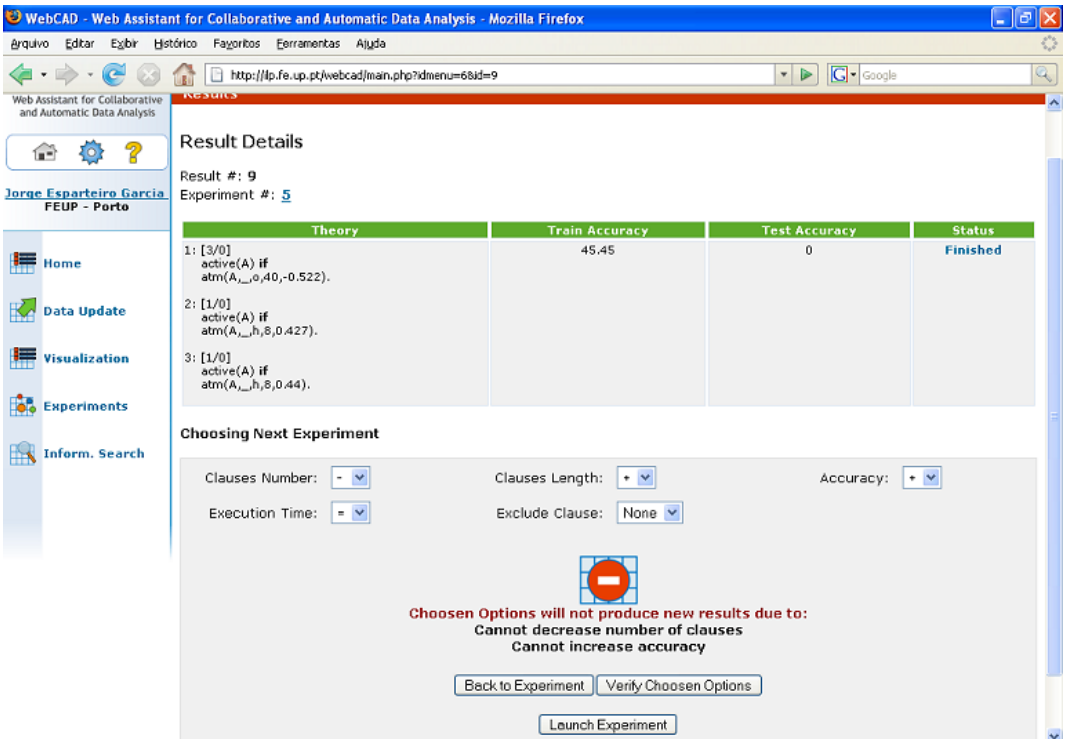

Figura 5.8: WebCAD com os Parâmetros Seleccionados

Algoritmo 1 Verificação de Parâmetros

```
Entrada:
        /* parametros do wrapper */
     ajusteNos /* incremento/decremento do numero de nós */
     ajusteMincover /* incremento/decremento do mincover */
     ajusteAccuracy /* incremento/decremento da accuracy */

   opcoesUtilizador  /* opções introduzidas pelo utilizador */

   teoria /* teoria mais recente induzida pelo IndLog */
   trace /* "trace" com os valores dos parâmetros do IndLog usados */

Saída :
      /* novos valores para os parâmetros do IndLog */
   settings  /* (nodes, mincover, accuracy, clauselength, forbidenClause) */
   mensagem /* mensagens para o utilizador */

Inicio
   mensagem = ""
   settings = valoresUltimaExecucao(trace)

   escolha = escolhaUtilizador(opcoesUtilizador, 'tempo')
   Se (escolha != '=') Então
       nodes = ajusta(trace, 'nodes', escolha, ajusteNos)
       Se nodes != valor(settings, 'nodes') Entao
          mensagem = addMsg('Não é possível alterar o tempo de execução como pretendido')
   FimSe

   escolha = escolhaUtilizador(opcoesUtilizador, 'numeroClausulas')
   Se (escolha != '=') Então
       mincover = ajusta(trace, 'mincover', escolha, ajusteMincover)
       Se mincover != valor(settings, 'mincover') Então
          mensagem = addMsg('Não é possível alterar o número de cláusulas como pretendido')
   FimSe

   escolha = escolhaUtilizador(opcoesUtilizador, 'accuracy')
   Se (escolha != '=') Então
       accuracy = ajusta(trace, 'accuracy', escolha, ajusteAccuracy)
       Se accuracy != valor(settings, 'accuracy') Então
          mensagem = addMsg('Não é possível alterar a accuracy como pretendido)
   FimSe

   escolha = escolhaUtilizador(opcoesUtilizador, 'clauseLength')
   Se (escolha != '=') Então
         clauselength = ajusta(trace, 'clauselength', escolha)
       Se clauselength != valor(settings, 'clauselength') Então
          mensagem = addMsg('Não é possível alterar o tamanho das cláusulas como pretendido)
   FimSe

   escolha = escolhaUtilizador(opcoesUtilizador, 'forbidenClause')
   indice = escolhaUtilizador(opcoesUtilizador)
   Se indice > 0 Então
        forbidenClause = teoria(indice)  /* devolve a cláusula número índice */
   FimSe

   settings = actualiza(nodes, mincover, accuracy, clauselength,forbidenClause)
Fim
```

Função	Explicação
valoresUltimaExecucao	Devolve os valores dos parâmetros do IndLog usados na última experiência
escolhaUtilizador	Devolve a opção efectuada pelo utilizador para o *pulldown menu* identificado pelo segundo parâmetro da função
ajusta	Actualiza o valor do parâmetro indicado pelo segundo parâmetro usando a opção do utilizador e os valores disponíveis no *trace*
teoria	Devolve a cláusula cuja posição na teoria esta indicada no parâmetro da função
actualiza	Actualiza o valor do parâmetro indicado como primeiro argumento
valor	Devolve o valor do parâmetro especificado na estrutura *settings*
novaMsg	Adiciona uma nova mensagem

Tabela 5.4: Funções Utilizadas no Algoritmo de Verificação de Parâmetros

Capítulo 6

Casos de Estudo

6.1 Introdução

Neste Capítulo irão ser apresentados diversos casos de estudo que foram feitos para a utilização da aplicação WebCAD desenvolvida neste projecto.
Em todos os casos de estudo apresentados, o conjunto de dados que irá ser utilizado será o *suramin* [Braddock, 1994] disponível no portal do Grupo de *Machine Learning* da Universidade de Oxford [MLCL, 2006].
Com estes casos de estudo apresentados, serão abordadas as principais formas de utilização da aplicação e serão mostrados exemplos de utilização para cada um dos casos.

O primeiro caso apresentado será a criação de um novo *dataset*, com exemplos de utilização e de envio para o servidor dos ficheiros relacionados com ele.
Será apresentado depois um caso em que se detalha a execução de uma experiência usando a aplicação. Exemplificar-se-á de seguida a alteração dos parâmetros de uma experiência para se obter novos resultados, caso seja possível.
Finalmente no último caso de estudo apresentado, mostrar-se-á uma sessão colaborativa de busca de dados, no caso *Background Knowledge* noutro sítio WebCAD.

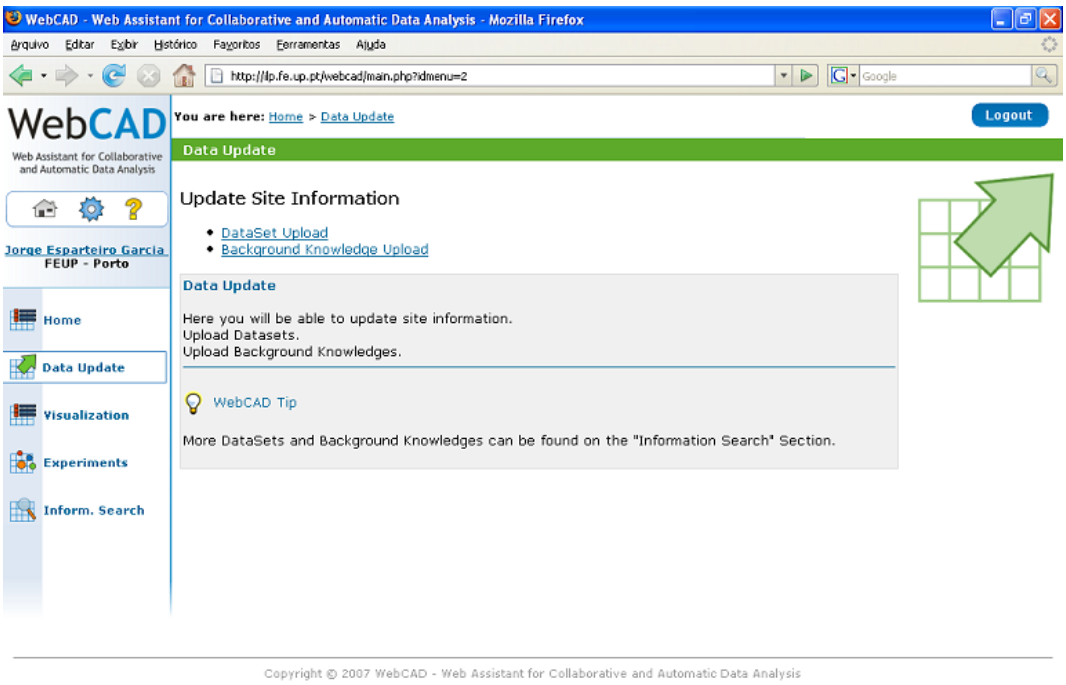

Figura 6.1: Menu de Escolha de Criação do *Dataset*

6.2 Criação de um Novo *Dataset*

Neste caso de estudo, mostra-se como se criou um novo *dataset* no WebCAD e quais os passos necessários para o envio dos ficheiros necessários para a sua criação.

No primeiro passo, o utilizador ao seleccionar o Menu *Data Update*, deve clicar em *Dataset Upload* para começar o processo de criação do *dataset*, como ilustra a Figura 6.1.

Ao entrar no menu de criação do *dataset*, é mostrado ao utilizador um ecrã de inserção de dados que completa a criação do *dataset*. Um dos dados que o utilizador necessita inserir é o domínio científico do *dataset*. Caso nenhum dos existentes se enquadre com o *dataset* que está a criar, o utilizador pode criar um novo domínio no sistema WebCAD como é detalhado na Figura 6.2.

Figura 6.2: Criação de um Novo Domínio Científico no Sistema WebCAD

6.2.1 Envio dos Ficheiros do *Dataset*

Paralelamente o utilizador vai também seleccionar os três ficheiros que pretende enviar para a criação do *dataset*.
Esses ficheiros são o ficheiro dos exemplos negativos, o ficheiro dos exemplos positivos e o ficheiro com o *Background Knowledge*. O utilizador preenche todos os dados e envia os três ficheiros paras criar correctamente o *dataset* no sistema WebCAD, caso contrário, ser-lhe-á mostrada uma mensagem de aviso ou erro consoante o caso.

Na Figura 6.3 é mostrado o interface de envio de ficheiros, no qual o utilizador envia os ficheiros e cria o *dataset*.

Finalmente, se os ficheiros foram correctamente seleccionados, e se não acontecer nenhum erro de envio ou de inserção do *dataset* na base de dados, como foi o caso, é mostrada ao utilizador uma mensagem de sucesso de envio dos ficheiros e de criação do *dataset*.

6.3 Realização de uma Sequência de Experiências

A realização de experiências de análise de dados é a principal funcionalidade desta aplicação. Neste caso de estudo, realizou-se uma sequência completa de experiências em que foi feita uma nova experiência de análise de dados do *dataset suramin* e de seguida foram feitas outras experiências de análise com base na

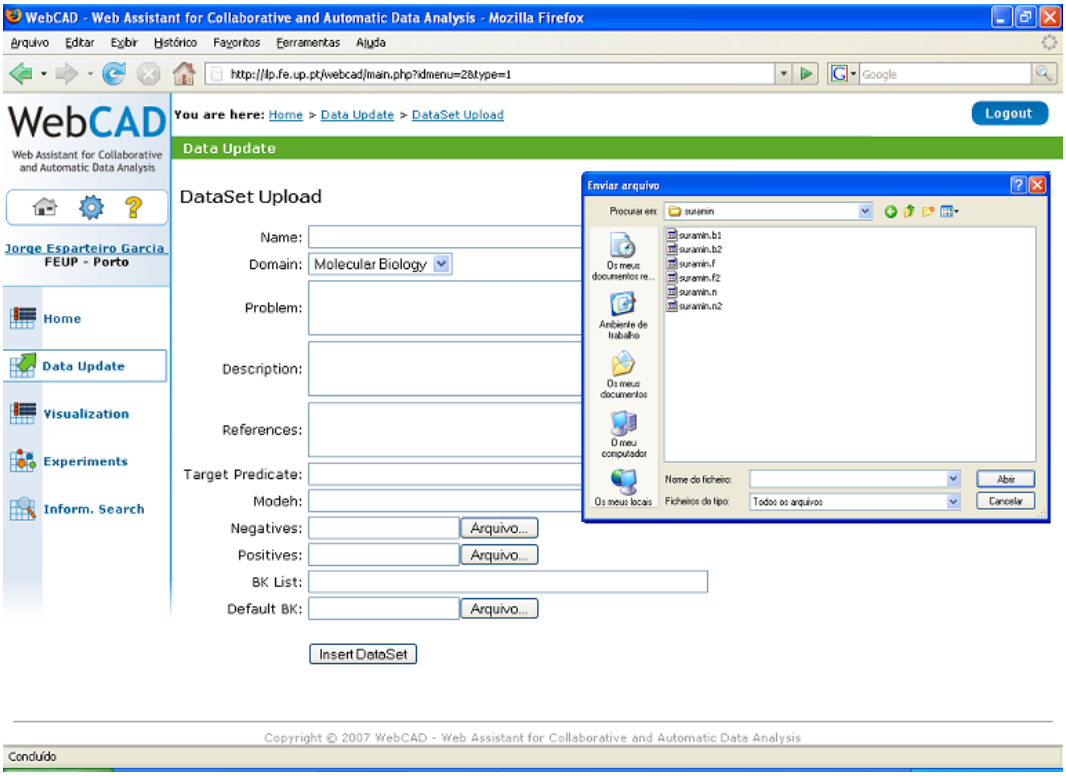

Figura 6.3: Inserção do *Dataset* com Escolha dos Ficheiros a Enviar

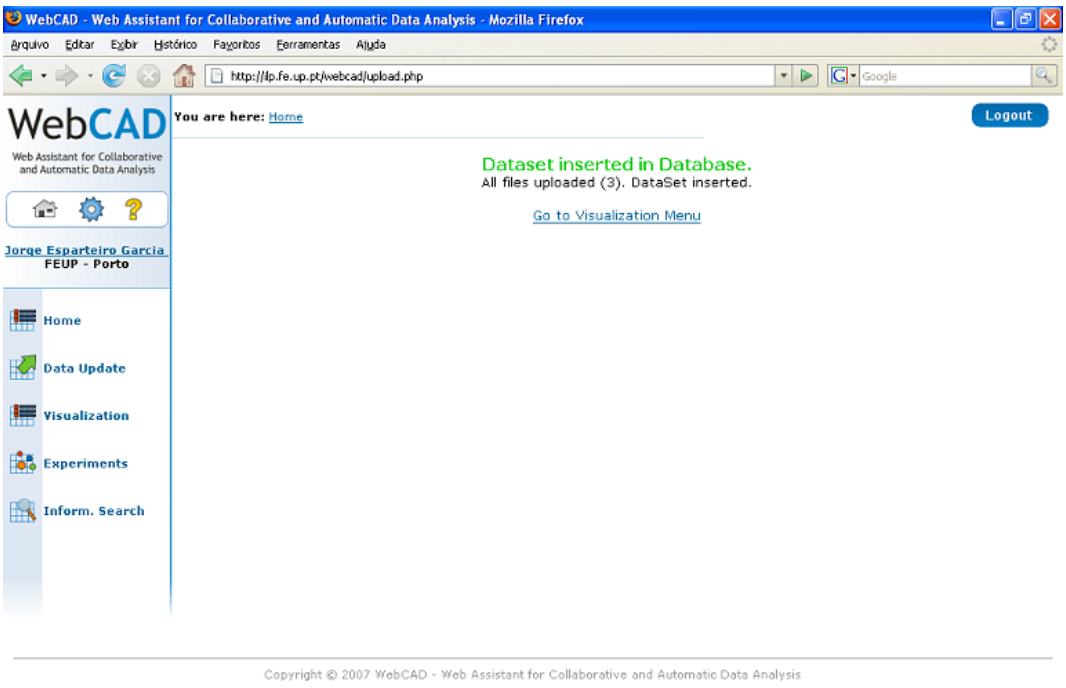

Figura 6.4: Resultado da Criação do *Dataset*

primeira experiência efectuada, tendo finalmente a experiência sido terminada depois de obtidos os resultados desejados.

Com este caso de estudo, pretende-se mostrar como um utilizador do WebCAD, pode realizar uma sequência completa de uma experiência e obter os resultados desejados, como estivesse a utilizar directamente a ferramenta de análise de dados.

6.3.1 Nova Experiência

Para se iniciar uma nova experiência, seleccionou-se a opção de lançar uma nova experiência, no menu Experiências.
Aí, foi necessário escolher o *Dataset* e o *Background Knowledge* do *suramin* bem como o método de avaliação pretendido para esta experiência de análise. Para este caso de estudo seleccionou-se o *Cross Validation*. Finalmente optou-se por partilhar esta experiência dentro da instituição e também para outros utilizadores de outras instituições que utilizem o sítio WebCAD.

Caso a experiência seja lançada, como aconteceu, a aplicação mostra um ecrã,

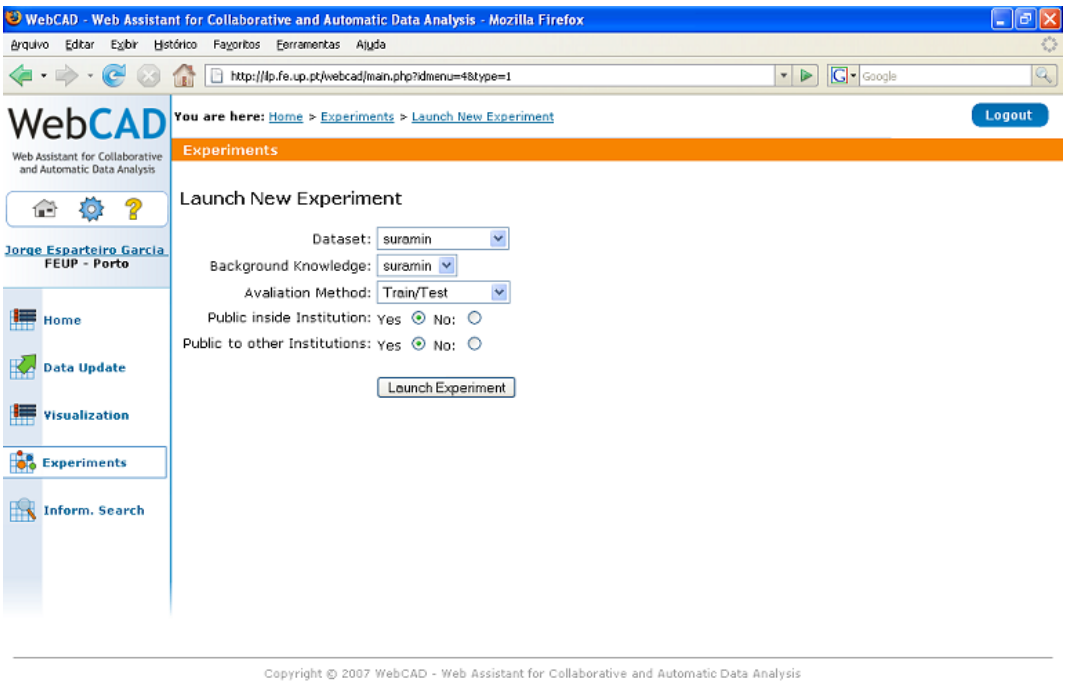

Figura 6.5: Início de uma Nova Experiência no WebCAD

ilustrado na Figura 6.6, com informação do processo de execução da experiência. Entre os dados que são mostrados, estão a informação se a aplicação conseguiu copiar o *dataset* para o servidor IndLog e a informação relacionada com a experiência iniciada como o número atribuído, *dataset* utilizado, visibilidade, hora de início, entre outros.

O tempo de execução das experiências varia consoante o *dataset* utilizado para a análise dos dados, bem como os parâmetros seleccionados. No caso de estudo que aqui é apresentado, em que o *dataset* utilizado é o *suramin*, com os parâmetros apresentados anteriormente, o tempo de execução da experiência foi inferior a 1 minuto. Contudo, consoante os *datasets* e parâmetros seleccionados, algumas experiências poderão levar algumas horas a terminarem.
Para que o utilizador saiba qual o estado da experiência que efectuou é mostrado o estado de cada uma das experiências que lançada.

Neste caso de estudo, como o tempo de execução bastante curto, a Tabela mostra a experiência efectuada já como terminada como ilustra a Figura 6.7. Ao ser expandida a linha respeitante à experiência efectuada, são mostrados todos os resultados relacionados com a experiência. Como neste momento só existia uma

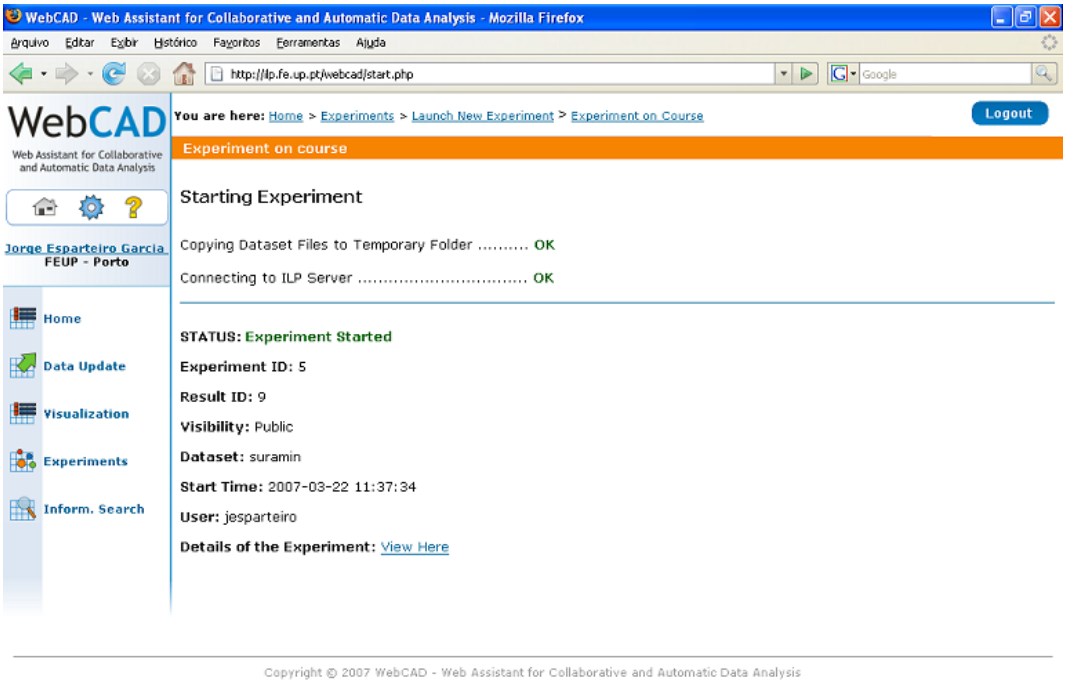

Figura 6.6: Início da Experiência

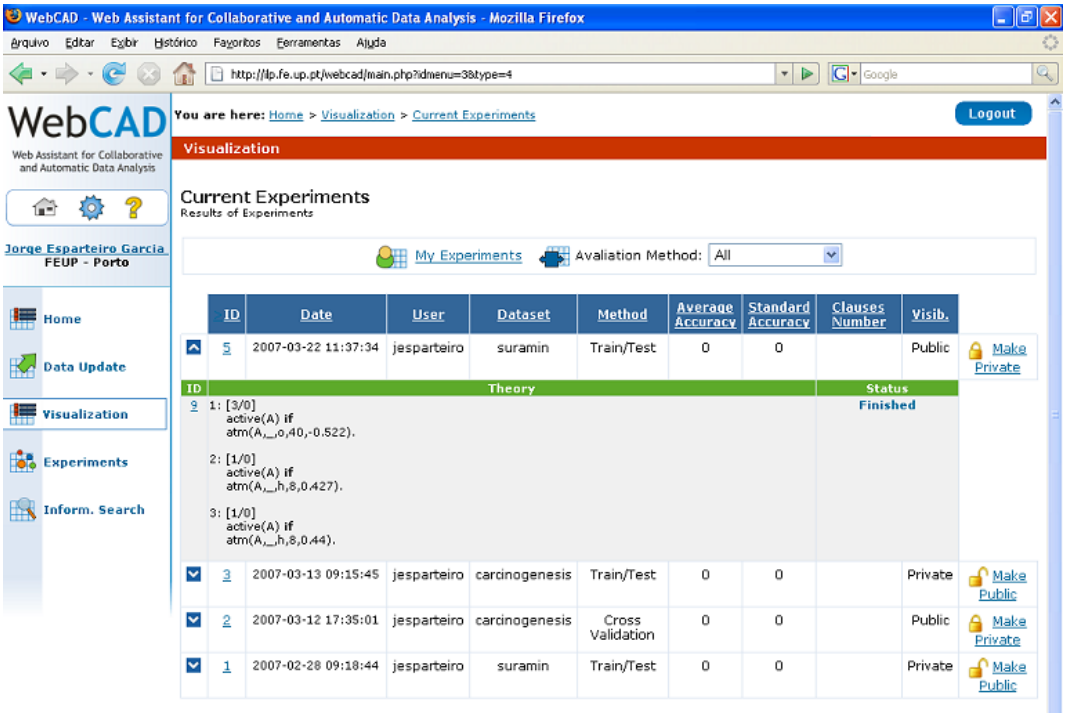

Figura 6.7: Estado da Experiência

experiência relacionada, são mostrados resultados dessa experiência, descritos no Resultado 1.

Resultado 1 Experiência Efectuada com o *Dataset* Suramin

```
1: [3/0]                  /* Positivos cobertos / Negativos cobertos */
    active(A) if
    atm(A,_,o,40,-0.522).

2: [1/0]                  /* Positivos cobertos / Negativos cobertos */
    active(A) if
    atm(A,_,h,8,0.427).

3: [1/0]                  /* Positivos cobertos / Negativos cobertos */
    active(A) if
    atm(A,_,h,8,0.44).
```

6.3.2 Refazer Experiência

Utilizando a experiência efectuada, foi refeita a experiência, tendo sido alterados os parâmetros do Número de Cláusulas e a Precisão. No Número de Cláusulas foi seleccionada a opção "-", que significa que será reduzido o número de cláusulas nas experiência, na Precisão foi seleccionada a opção "+", que significa que será aumentada a Precisão da experiência.

Depois de feita a verificação dos parâmetros, a aplicação respondeu que os parâmetros seleccionados não produziriam novos resultados, pois não era possível Diminuir o número de cláusulas e não era possível aumentar a precisão.
Foram então alterados esses parâmetros, de acordo com a verificação feita, tendo sido ainda alterado o parâmetro Tempo de Execução, para que este seja maior.

Depois da alteração feita foi feita uma nova verificação dos parâmetros seleccionados que respondeu que os parâmetros seleccionados eram válidos e iriam produzir novos resultados como se mostra na Figura 6.9.

Foi lançada então a experiência já utilizando os novos parâmetros. Esta experiência retornou resultados diferentes da anterior, confirmando o que a aplicação tinha

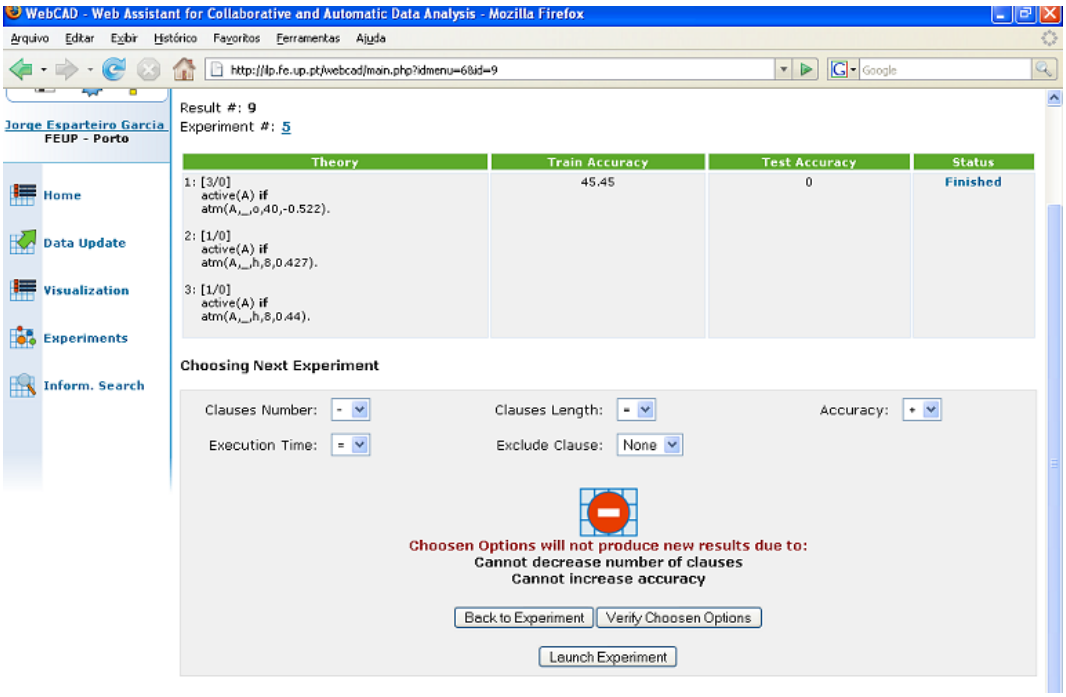

Figura 6.8: Alteração dos Parâmetros da Experiência

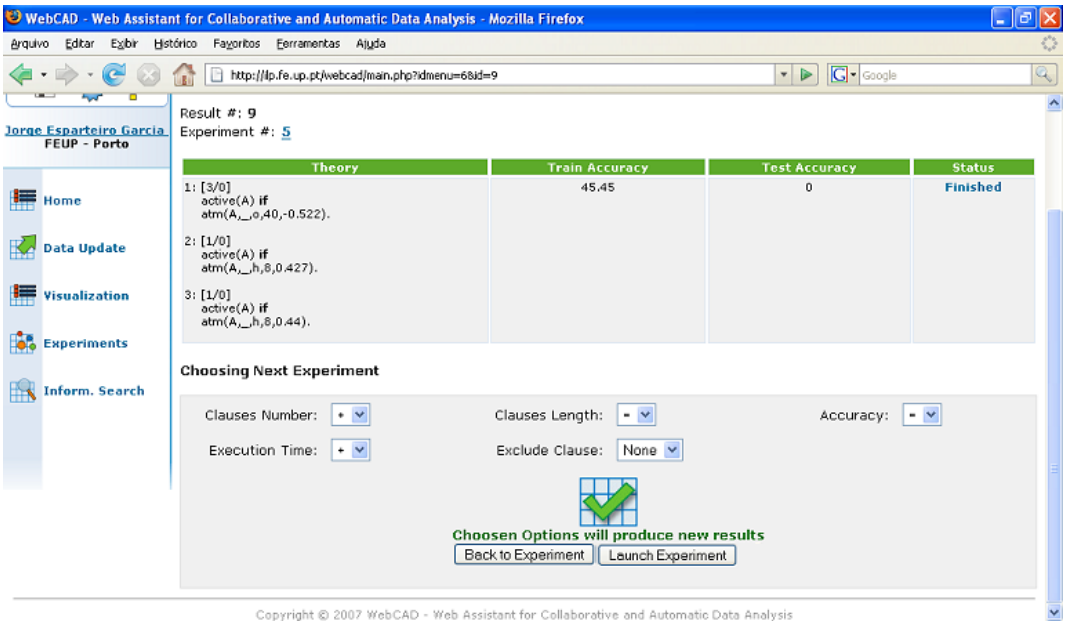

Figura 6.9: Parâmetros Alterados Válidos

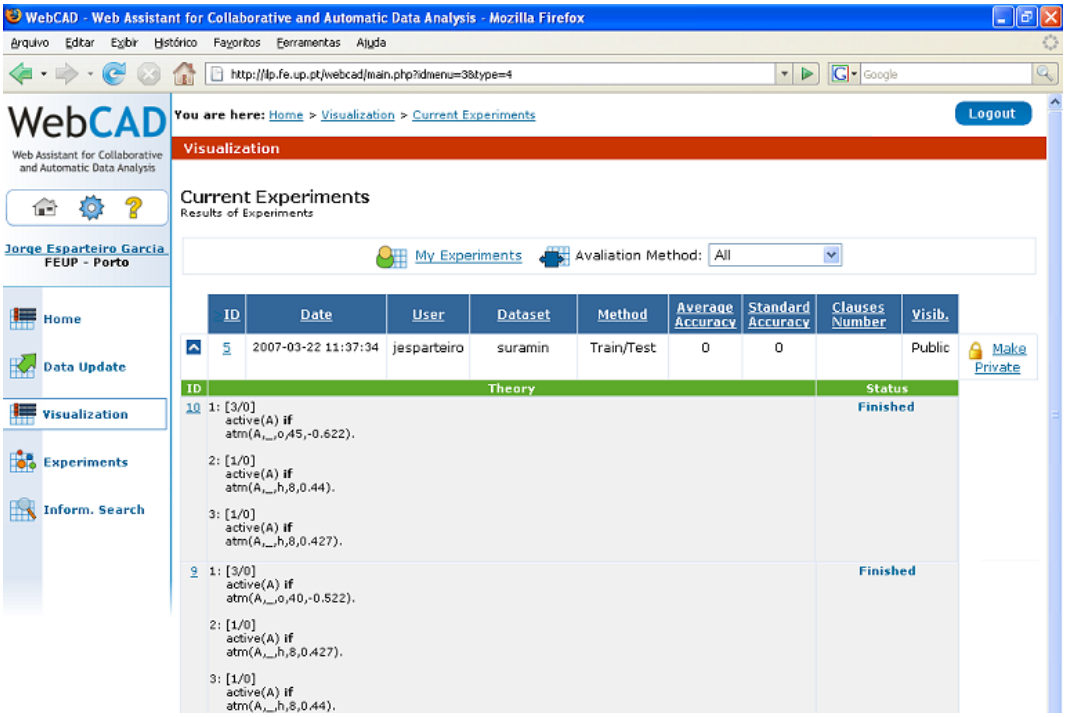

Figura 6.10: Visualização do Histórico da Experiência

respondido aquando da verificação dos novos parâmetros.
Estes novos resultados estão associados à experiência inicial em forma de histórico, e expandido a experiência desejada é possível visualizar todas as experiências efectuadas como se ilustra na Figura 6.10. Estas experiências foram efectuadas depois de terem sido alterados os parâmetros, tendo sido feita a prévia verificação dos parâmetros como foi mostrada anteriormente.

6.4 Sessão Colaborativa de Busca de Dados a outro Sítio WebCAD

Neste caso de estudo, mostrar-se uma sessão colaborativa de busca de dados, no caso *Background Knowledge* noutro sítio WebCAD.

Para se iniciar a sessão colaborativa de busca de dados, seleccionou-se no Menu a opção *Information Search* como mostra a Figura 6.11.

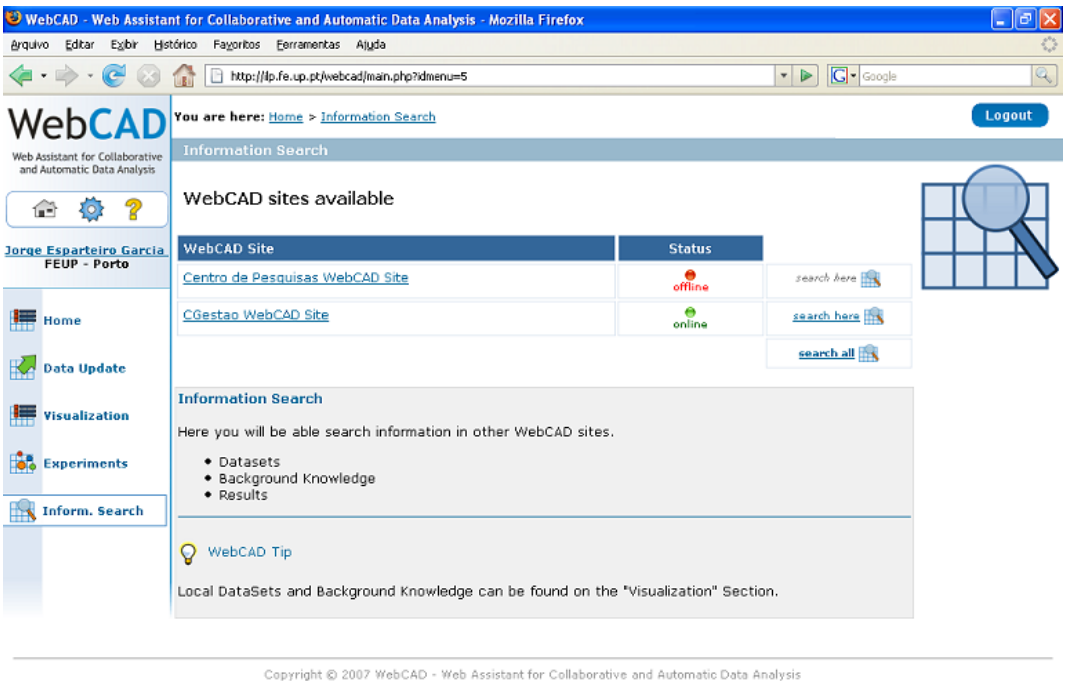

Figura 6.11: Início da Sessão Colaborativa no WebCAD

6.4.1 Visualização de Sítios WebCAD Disponíveis

Neste ecrã, foi possível visualizar quais os sítios WebCAD disponíveis e indisponíveis no momento em que foi realizado o caso de estudo. Nesse momento o sítio do "Centro de Pesquisas" estava indisponível, estando apenas disponível o sítio "CGestao WebCAD".
Neste ecrã, é possível efectuar a pesquisa apenas num sítio, ou pesquisar informação em todos os sítios WebCAD disponíveis simultaneamente.

Foi feita uma pesquisa no único sítio disponível naquele momento, o "CGestao WebCAD", como ilustra a Figura 6.12.
A pesquisa retornou os *Datasets* e *Background Knowledge* disponíveis no momento. Retornou também o número de experiências públicas que terminadas e as que ainda estão em decorrer mas que também podem ser visualizadas.

Foi seleccionado o *Dataset carcinogenesis*. Ao ser seleccionado o *Dataset*, a aplicação mostrou os seus detalhes e os ficheiros que lhe estão associados como mostra a Figura 6.13.

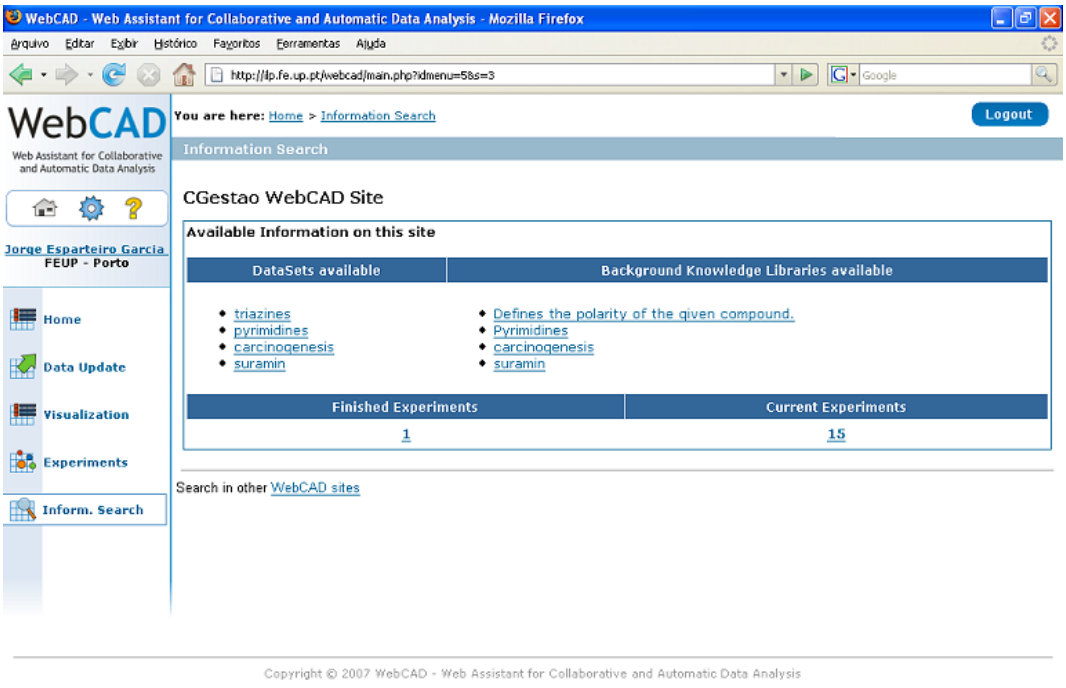

Figura 6.12: Pesquisa de Dados no Sítio "CGestao WebCAD".

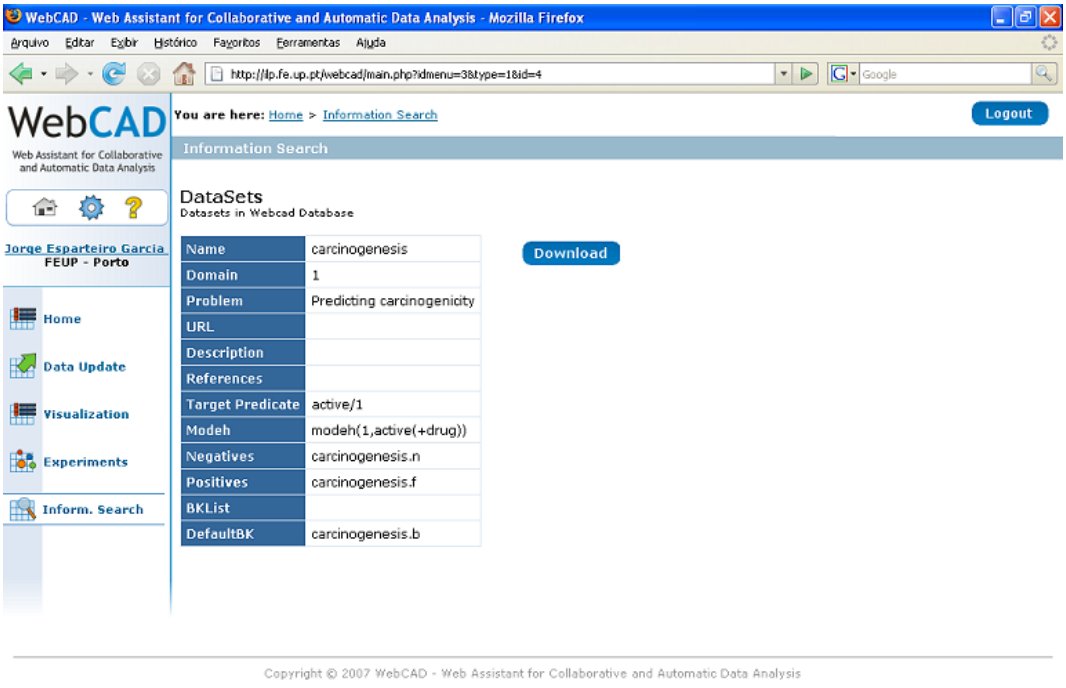

Figura 6.13: Visualização de um *Dataset* Partilhado.

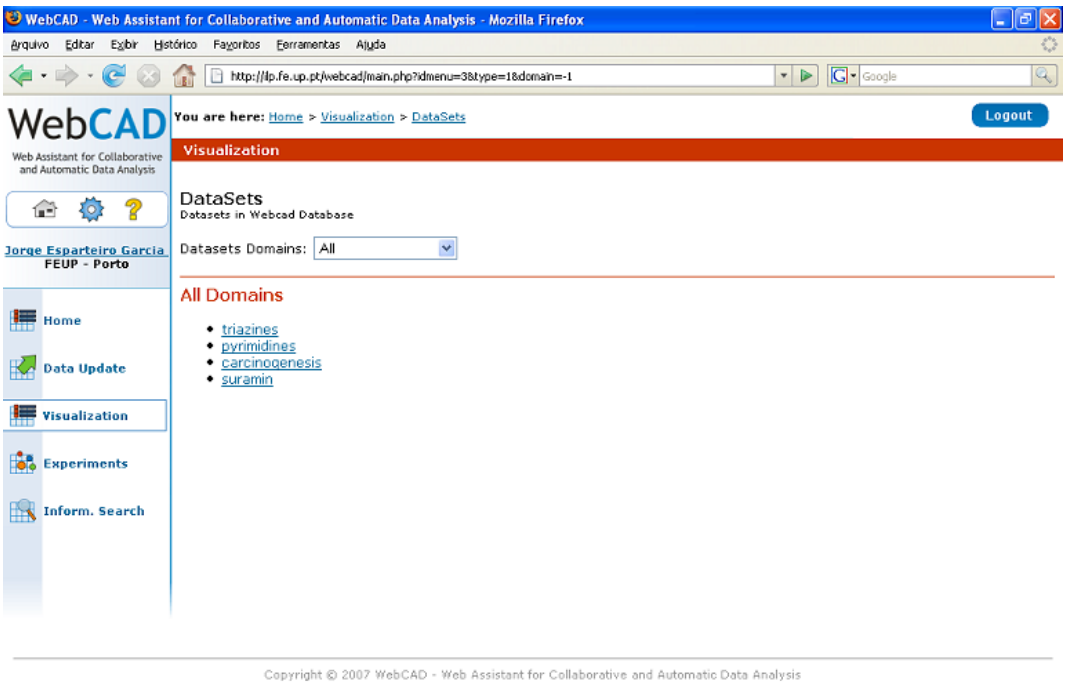

Figura 6.14: Visualização do *Dataset* Descarregado na Biblioteca dos *Datasets* Locais.

6.4.2 *Download* de *Dataset* para o WebCAD Local

Neste ecrã é mostrado também, um botão de *Download* que permite descarregar o *Dataset* seleccionado para o nosso próprio sítio WebCAD. Foi então feito o *Download* do *Dataset*, ficando este disponível na biblioteca dos *Datasets* do nosso sítio como é ilustrado na Figura 6.14.

Capítulo 7

Conclusões

7.1 Introdução

Neste capítulo é encerrada a dissertação recapitulando, na secção 7.2, os objectivos propostos no início e que foram alcançados. De seguida, na secção 7.3, são apresentadas algumas sugestões para trabalho futuro, que podem ser feitas com base no trabalho desenvolvido no âmbito desta dissertação. Finalmente, na secção 7.4, são tecidas algumas considerações finais sobre o trabalho desenvolvido e sobre a dissertação apresentada.

7.2 Objectivos do Trabalho

O sucesso do desenvolvimento do WebCAD, vai permitir que actualmente, a análise automática de dados esteja disponível para um conjunto grande de utilizadores que até agora não eram capazes de a fazer, sem a ajuda de um especialista numa ferramenta de análise de dados. Com o desenvolvimento desta ferramenta, os utilizadores poderão também analisar conjuntos complexos de dados sem que necessitem de uma máquina de grandes recursos, tornando o seu estudo e a sua investigação mais autónomo.

Os investigadores podem também utilizar esta ferramenta para partilhar conhecimento, fornecendo um conjunto de recursos para a análise de dados, permitindo

a colaboração com outros investigadores para a análise de informação de grande utilidade e importância.

O WebCAD permite também, como era requerido, aos investigadores, utilizar uma ferramenta com um interface *user-friendly* como é um *Web browser* para executar as suas experiências de análise de dados remotamente, mesmo que não tenha recursos computacionais para o fazer. Isto permite a qualquer investigador em qualquer parte do mundo fazer a análise de dados como se possuísse uma potente ferramenta de análise de dados.

O WebCAD permite também, armazenar informação relacionada com experiências de análise de dados efectuadas anteriormente, mantendo um registo de todas as experiências realizadas, e quais as configurações utilizadas. Permite também guardar informação de *Datasets* incluindo o *Background Knowledge* de diversos tipos de problemas, que podem também ser partilhados com outros investigadores.

Finalmente, foi também alcançado o objectivo de desenvolver uma aplicação de fácil acesso e utilização para uma ferramenta de análise de dados desenvolvida em Prolog, que a torna uma das primeiras aplicações que permite a possibilidade de qualquer utilizador, sem qualquer conhecimento informático aprofundado utilizar algumas das mais potentes ferramentas de análise de dados.

7.3 Trabalho Futuro

Nesta secção apresentam-se algumas ideias sobre uma possível continuação do trabalho desenvolvido, como forma de aproveitar o que já foi feito e potenciar as funcionalidades e as capacidades que o sistema WebCAD tem.

Ao longo do desenvolvimento da aplicação WebCAD, e com o cumprir dos objectivos definidos para este projecto, foram surgindo desde logo diversas ideias para um possível trabalho futuro. Entre essas ideias destacam-se as seguintes:

- **Possibilidade de escolha da ferramenta de análise de dados** — Como é descrito ao longo da dissertação, o WebCAD utiliza o IndLog para efectuar as experiências de análise de dados. A possibilidade da escolha de outras

ferramentas para efectuar as experiências de análise de dados, traria melhoramentos significativos à aplicação, pois possibilitaria aos utilizadores a utilização da ferramenta de análise que melhor se adequasse aos dados a serem analisados.
Esta funcionalidade envolveria um trabalho semelhante ao que foi feito neste trabalho para o IndLog para cada uma das ferramentas de análise de dados presentes no sistema. Seria necessário haver um estudo de cada uma das ferramentas que possibilitasse depois criar uma camada de ligação entre o WebCAD e cada uma delas. O sistema estaria depois preparado a poder fazer a análise de dados num conjunto determinado e previamente bem definido de ferramentas.

- **Independência da ferramenta de análise de dados** — Tornar o WebCAD independente da ferramenta de análise de dados que fosse utilizada permitiria que qualquer ferramenta de análise de dados que exista possa ser utilizada na aplicação. Para que a aplicação pudesse disponibilizar esta funcionalidade seria necessário criar uma nova camada lógica entre a aplicação *Web* e as ferramentas de análise de dados. Para que o WebCAD fosse completamente independente de qualquer ferramenta de análise de dados que pudesse ser utilizada, seria necessário criar um *wrapper* que envolvesse cada uma das camadas mais baixas de ligação às ferramentas. Em cada uma dessas camadas estariam publicados os comandos e as instruções necessárias para cada uma das ferramentas. O *wrapper* faria a ligação entre cada uma dessas camadas lógicas de ligação e o WebCAD.

- **Sugestão de parâmetros e de experiências de dados semelhantes** — Aproveitar as funcionalidades de colaboração presentes no sistema, pouparia alguns passos de pesquisa de resultados aos investigadores e tornaria a execução das experiências de análise de dados mais convergente e precisa. Ao efectuar uma experiência, seria sugerido e mostrado automaticamente ao utilizador, resultados já obtidos com as configurações e parâmetros seleccionados. A aplicação mostraria resultados já obtidos com as configurações seleccionadas e sugeriria possíveis configurações para obtenção de diferentes resultados.

- **Limite do número de experiências em execução** — Umas das situações que pode ocorrer no WebCAD, se utilizado em simultâneo por um largo número de utilizadores, é a sobrecarga excessiva da máquina onde corre o

IndLog. Para que não aconteça este problema, a imposição de um número limite de experiências em execução simultaneamente no IndLog preveniria essa sobrecarga. Esta funcionalidade permitiria que a execução das experiências fosse feita mais rápida e reduziria as possibilidades de sobrecarga excessiva da máquina.

7.4 Considerações Finais

A principal contribuição do trabalho realizado e descrito nesta dissertação, foi a implementação de uma aplicação que utiliza uma ferramenta de análise de dados poderosa que permite a qualquer utilizador leigo na ferramenta, utilizá-la de forma correcta e eficiente como se utilizasse a ferramenta directamente. Para isso, foi criada uma aplicação *Web* que permite a sua utilização por parte de qualquer utilizador que a queira utilizar. Destaque na aplicação, para a componente colaborativa que possui, que permite que o trabalho de análise de dados seja feito de forma colaborativo com outros utilizadores.

Os bons resultados obtidos na elaboração deste trabalho constituem uma forte motivação para que o trabalho realizado para esta tese possa ser continuado e sejam futuramente criadas novas funcionalidades.

Bibliografia

[Anderson et al., 2002] Anderson, D. P., Cobb, J., Korpela, E., Lebofsky, M., and Werthimer, D. (2002). SETI@home: an experiment in public-resource computing. *Commun. ACM*, 45(11):56–61.

[Apache, 1995] Apache (1995). Apache http server. http://httpd.apache.org/.

[Barbosa and Monteiro, 2006] Barbosa, R. N. J. and Monteiro, A. (2006). Biogrid application toolkit: a grid-based problem solving environment tool for biomedical data analysis. In *In proceedings of VECPAR'06*. Rio de Janeiro, Brasi.

[Braddock, 1994] Braddock, P. S. e. a. (1994). A structure-activity analysis of the growth factor and angiogenic activity of basic fibroblast growth factor by suramin and related polyanions. *Br. J. Cancer.*, pages 890–898.

[Camacho, 2000] Camacho, R. (2000). *Inducing Models of Human Control Skills using Machine Learning Algorithms*. PhD thesis, Departamento de Engenharia Electrónica de de Computadores, Universidade do Porto.

[CBL, 2006] CBL (2006). Imperial College Computational Bioinformatics Laboratory. http://www3.imperial.ac.uk/.

[De Raedt and Bruynooghe, 1992] De Raedt, L. and Bruynooghe, M. (1992). An overview of the interactive concept-learner and theory revisor CLINT. In Muggleton, S., editor, *ILP*, pages 163–192. AP.

[Džeroski et al., 1998] Džeroski, S., Jacobs, N., Molina, M., Moure, C., Muggleton, S., and Laer, W. V. (1998). Detecting traffic problems with ILP. In Page, C., editor, *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, LNAI 1446, pages 281–290, Berlin. Springer-Verlag.

[Frawley et al., 1992] Frawley, W. J., Shapiro, P. G., and Matheus, C. J. (1992). Knowledge discovery in databases — an overview. *Ai Magazine*, 13:57–70.

[Harmon et al., 1988] Harmon, P., Maus, R., and Morrissey, W. (1988). *Expert Systems Tools and Applications*. John Wiley & Sons, New York.

[ILPnet2, 2006] ILPnet2 (2006). Network of Excellence in Inductive Logic Programming – ILPnet2. http://www.cs.bris.ac.uk/ ILPnet2/.

[Jorge, 1998] Jorge, A. (1998). *Iterative Induction of Logic Programs: an approach to logic program synthesis form incomplete specifications*. PhD thesis, University of Porto.

[Metal, 2002] Metal (2002). A meta-learning assistant for providing user support in machine learning and data mining. http://www.metal-kdd.org/.

[Mladenic et al., 2003] Mladenic, D., Lavrac, N. A. D. A., and Bohanec, M. (2003). *Data Mining and Decision Support : Integration and Collaboration (The International Series in Engineering and Computer Science)*. Springer.

[MLCL, 2006] MLCL (2006). Oxford University Machine Learning Group. http://www.cs.york.ac.uk/.

[Muggleton, 1995] Muggleton, S. (1995). Inverse entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245–286.

[Muggleton and Feng, 1990] Muggleton, S. and Feng, C. (1990). Efficient induction of logic programs. In *Proceedings of the 1st Conference on Algorithmic Learning Theory*, pages 368–381. Ohmsma, Tokyo, Japan.

[Muggleton and Raedt, 1994] Muggleton, S. and Raedt, L. D. (1994). Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming*, 19/20:629–679.

[MySQL, 2006] MySQL (2006). Mysql — sql database management system. http://www.mysql.com.

[PHP, 1994] PHP (1994). PHP — hypertext preprocessor. http://www.php.net.

[Popelynsky, 1998] Popelynsky, L. (1998). Knowledge Discovery in Spatial Data by Means of ilp. In *PKDD '98: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 185–193, London, UK. Springer-Verlag.

[Ramakrishnan and Gehrke, 2002] Ramakrishnan, R. and Gehrke, J. (2002). *Database Management Systems*. McGraw-Hill Science/Engineering/Math.

[Rumbaugh et al., 2004] Rumbaugh, J., Jacobson, I., and Booch, G. (2004). *The Unified Modeling Language Reference Manual*. Addison-Wesley.

[Srinivasan, 2001] Srinivasan, A. (2001). *The Aleph Manual*. http://web.comlab.ox.ac.uk/oucl/research/ areas/machlearn/Aleph/.

[Turcotte et al., 1998] Turcotte, M., Muggleton, S., and Sternberg, M. (1998). Protein Fold Recognition. In Page, C., editor, *Proc. of the 8th International Workshop on Inductive Logic Programming (ILP-98)*, LNAI 1446, pages 53–64, Berlin. Springer-Verlag.

[UY-MLG, 2006] UY-MLG (2006). University of York Machine Learning Group. http://www.cs.york.ac.uk/.

Apêndice A

Acrónimos

ILP	Inductive Logic Programming
LPPO	Lógica de Predicados de Primeira Ordem
DM	Data Mining
ML	Machine Learning
IndLog	Inductive Logic
UML	Unified Modeling Language
PHP	Hypertext Preprocessor
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
BCNF	Boyce-Codd Normal Form

Apêndice B

Dicionário

Inductive logic Programming ← Indução de Programas em Lógica
Logic Programming Programação em Lógica
Inductive Logic ← Lógica Indutiva
Machine Learning ← Aprendizagem computacional
Background Knowledge ← Conhecimento de fundo
Bottom Clause ← Cláusula Base
Data Mining ← Extracção de Conhecimento
Wrapper ← Envolvente
Batch File ← Ficheiro de comandos
Unified Modeling Language ← Linguagem de Modelação Unificada
HyperText Markupo Markup Language ← Linguagem de Marcação de HiperTexto
Boyce-Codd normal form ← Forma normal de Boyce-Codd
HyperText Transfer Protocol ← Protocolo de Transferência de Hipertexto