UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

SISTEMA DE AQUISIÇÃO DE CONHECIMENTO PARA APOIO À CONSULTA DE SUBVISÃO

Cláudia Martins Antunes
(Licenciada)

Dissertação para a obtenção do grau de Mestre
em Engenharia Electrotécnica e de Computadores

Orientador Científico: Doutor João Emílio Segurado Pavão Martins

Júri: Doutor João Emílio Segurado Pavão Martins
Doutor Ernesto José Fernandes Costa
Doutor Arlindo Manuel Limede de Oliveira
Doutor Nuno João Neves Mamede

Abril de 2001

RESUMO

Alguns domínios médicos requerem o acompanhamento de doentes que perderam algumas capacidades, por alterações do funcionamento de um ou mais órgãos, devido a lesões nesses mesmos órgãos. Um caso particular deste tipo de domínio é a Subvisão, que se traduz numa perda parcial da visão. Em domínios onde a população é reduzida (como é o caso da Subvisão), os avanços científicos são conseguidos à custa de estudos de condução difícil, e que não são estatisticamente representativos, na sua maioria das vezes.

Este trabalho contribui com a criação de um sistema de informação integrado, que permite o suporte de todas as actividades da consulta de subvisão, e comporta duas componentes fundamentais: um sistema transaccional e um sistema de aquisição de conhecimento. Com o sistema transaccional consegue-se fazer o registo/gestão dos dados dos doentes de modo distribuído, possibilitando o alargamento de utilização do mesmo sistema por várias consultas da mesma especialidade. O objectivo do sistema de aquisição de conhecimento é auxiliar na descoberta das relações entre as alterações ao nível do órgão com as alterações ao nível das capacidades do indivíduo. Este sistema é baseado em técnicas de extracção de conhecimento em bases de dados (KDD) e é independente do desenho da base de dados.

Para além da descrição do sistema implementado, apresenta-se uma descrição de cada uma das etapas do processo de aquisição de conhecimento, referindo os mecanismos existentes mais importantes, assim como a sua adequação a cada tipo de situação.

Palavras Chave: Data Mining, Sistemas de Informação, Aquisição de Conhecimento em Bases de Dados, Regras de Associação, Medicina, Subvisão.

ABSTRACT

Some medical domains require patient monitoring when the patient has lost some abilities due functional changes at the organ level, caused by a lesion in that organ. Low vision is a particular case of this type of medical domain, and is expressed as a partial vision loss.

Most of the time, in emerging domains where the population is reduced (such as Low vision), traditional studies are very difficult to conduct and not statistically representative.

This work contributes with the creation of an integrated information system, that supports the entire low vision consultation, and is based on two fundamental components: a transactional system and a knowledge acquisition system. With the transactional system, patient's data recording/management is done in a distributed way, making it possible to expand the system's utilization to every consultation of the same type in the different hospitals of a country. The main goal of the knowledge acquisition system is to help on the discovery of relations between the changes at the organ level and the changes at the individual abilities level. This system is based on knowledge discovery from database techniques, and is independent of the database design.

Besides the system description, is presented a description of each step of the knowledge acquisition process, referring the most relevant data mining mechanisms, and their compliance to each situation.

Key Words: Data Mining, Information System, Knowledge Discovery in Databases, Association Rules, Medicine, Low vision

AGRADECIMENTOS

Ao meu orientador Professor João Pavão Martins, os meus agradecimentos, por ter acreditado num projecto que abrange não só o mundo académico, mas também um mundo um pouco desconhecido para nós: o mundo da medicina. O meu obrigada pela liberdade concedida e pela confiança depositada, por ter possibilitado a minha iniciação num domínio novo.

À Dr.ª Conceição Neves, médica responsável pela Consulta de Subvisão do Hospital de Santa Maria, um muito obrigada pela cedência do seu tempo e paciência na introdução a uma área nova e complexa como é a subvisão. Um muito obrigada pela forma como correspondeu às minhas dúvidas e expectativas.

Uma palavra de agradecimento à Professora Teresa Vazão pela simpatia com que me apresentou várias críticas e sugestões quanto à redacção deste texto.

Um muito obrigado aos meus pais e irmã pelo apoio e permanência ao longo de todos estes anos. Aos meus alunos, à Inês Lynce e à Dona Teresinha pela simpatia e amizade demonstrada.

Por fim, um obrigada muito especial ao Pedro Sinogas, com quem discuti todos os problemas e possíveis abordagens, assim como pelas leituras e críticas efectuadas em cada uma das fases do trabalho. Um obrigada muito especial por me apoiar nos momentos mais difíceis e menos optimistas, por me acompanhar nestes momentos de transição do mundo de estudante para o mundo da docência.

A todos um muito obrigada!

Lisboa, Abril de 2001

Cláudia Martins Antunes

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Capítulo I - INTRODUÇÃO

Actualmente existe já alguma experiência na criação de sistemas de apoio à actividade médica, indo desde sistemas de gestão dos dados dos doentes, até sistemas de apoio à fase de diagnóstico. Porém, a integração destes sistemas, de forma a suportar todo o processo conduzido pelo médico na sua actividade diária, é ainda incipiente.

O projecto de informatização da Consulta de Subvisão do Hospital de Santa Maria teve início há cerca de quatro anos com a criação de uma base de dados e um sistema de gestão desses dados. Com esta primeira aplicação pretendeu-se suportar a fase de gestão dos dados dos doentes, de modo a facilitar o seu acompanhamento pelos profissionais da consulta. No ano seguinte as preocupações voltaram-se para a criação de mecanismos de avaliação e reabilitação de crianças com problemas de subvisão. Estas aplicações, apesar de semelhantes, tinham objectivos claramente diferenciados. Se por um lado a aplicação de avaliação tendia a aproximar-se de um sistema de apoio ao diagnóstico, a segunda dirigia-se fundamentalmente ao apoio ao doente, ao fornecer alguns mecanismos básicos de reabilitação.

Criados os mecanismos básicos de apoio à consulta de subvisão, o passo natural seguinte traduzia-se pela integração dos vários mecanismos.

1 - ENQUADRAMENTO

1.1 - Subvisão

A subvisão é uma área terapêutica recente, e traduz-se numa perda parcial da visão. Quando se fala de perda de visão existem alguns aspectos a ter em consideração, dos quais se destacam as

alterações funcionais ao nível do olho – funções visuais – e as alterações ao nível das capacidades do indivíduo que dependem directa ou indirectamente do órgão afectado – visão funcional. Repare-se que se as funções visuais podem ser avaliadas quantitativamente e objectivamente, o mesmo não se passa com a visão funcional, avaliada apenas de forma qualitativa e subjectiva. [Colenbrander 1999]

De tudo isto se compreende que o diagnóstico é determinado pela ponderação dos dados registados sobre as funções visuais do doente e a avaliação das suas capacidades – visão funcional. É pois esta ponderação o cerne da questão. Porém, não é só, pois para além da avaliação da visão funcional ser subjectiva, as medidas das várias funções visuais ainda não se encontram padronizadas. Na verdade, os avanços neste domínio continuam a ser feitos com base em estudos tradicionais de avaliação de conjuntos restritos (uma ou poucas dezenas) de indivíduos com subvisão.

É nesta área que as tecnologias da informação podem dar um contributo significativo, e é a isso que se propõe este trabalho, com a construção de um sistema de informação capaz de suportar todas as vertentes da consulta de subvisão.

1.2 - Sistemas de Informação

"Um Sistema de Informação pode ser definido como um conjunto de componentes interligados, funcionando juntos para recolher, processar, armazenar, e disseminar a informação, de modo a suportar a tomada, coordenação, controlo, análise e visualização das decisões numa organização." [Laudon 1998] Um tipo particular de sistemas de informação são os sistemas transaccionais, que executam e registam as transacções diárias efectuadas pela e na organização. Dadas as cada vez maiores quantidades de dados registados pelos sistemas transaccionais, a análise destes torna-se demasiado complexa para ser efectuada pelos métodos tradicionais. É neste contexto que surgem os modernos sistemas de aquisição de conhecimento.

Um sistema de aquisição de conhecimento é um sistema que permite aos utilizadores extraírem conhecimento útil a partir de grandes quantidades de dados, "extracção esta não trivial de conhecimento implícito, previamente desconhecido e potencialmente útil, feita a partir dos dados registados". [Frawley 1992] O objectivo primordial dos sistemas de aquisição de conhecimento é portanto, a partir dos dados registados em bases de dados, obter descrições compactas da informação ali registada, como por exemplo as relações entre atributos ou a classificação de instâncias de entidades. O processo de aquisição de conhecimento é composto por três etapas fundamentais: o pré-processamento, o *data mining* e o pós-processamento, sendo cada uma destas etapas constituídas por várias tarefas.

A etapa de pré-processamento constitui uma espécie de engenharia dos dados, ao reconstruir os dados de entrada de forma a melhorar a performance dos esquemas de *data mining* usados para a descoberta da informação. Por seu lado, os esquemas de *data mining*, através da análise sistemática dos dados efectuam a extracção de informação, propriamente dita. Tal como a etapa de pré-processamento, o pós-processamento constitui uma espécie de engenharia dos dados, desta vez processando os dados de saída, de modo a tornar os resultados dos esquemas de *data mining* mais compreensíveis para os seus utilizadores finais. [Addrians 1996]

2 - OBJECTIVOS

O objectivo da presente tese é definir e implementar um sistema de informação, capaz de dar resposta às necessidades de alguns domínios médicos, onde a avaliação da situação dos doentes é feita com base na ponderação entre a avaliação das funcionalidades dos seus orgãos e a avaliação das suas capacidades efectivas. Um caso particular deste tipo de domínio é a Subvisão.

A abordagem seguida define uma solução que para além de registar/gerir os dados dos doentes com subvisão – sistema transaccional –, consegue analisar esses dados quer de modo a contribuir para estabelecer aquelas medidas padronizadas, quer a descobrir as relações entre as alterações ao nível do órgão e das alterações nas capacidades do indivíduo – sistema de aquisição de conhecimento.

De acordo com o que foi descrito, a criação do sistema integrado de apoio à área da subvisão deverá permitir:

- suportar o registo/gestão dos dados dos doentes da consulta, e de outras semelhantes;
- descobrir relações entre os dados registados.

A abordagem proposta garante que o sistema pode ser adoptado por outros domínios da medicina, desde que seja definida uma nova base de dados dedicada a esse domínio. Uma vez que o mecanismo de aquisição de conhecimento proposto é independente da estrutura da base de dados, a análise dos dados registados nessas novas bases de dados é perfeitamente viável.

3 - RESULTADOS

A abordagem proposta baseia-se pois na criação de um sistema realmente integrado que permite ao pessoal médico a utilização de um único sistema capaz de prestar auxílio em todas as fases da consulta.

Este sistema é composto por dois módulos fundamentais: um sistema transaccional e um sistema de aquisição de conhecimento. O sistema transaccional tem uma arquitectura do tipo cliente/servidor, o que permite a centralização dos dados numa única máquina (servidor) e a acessibilidade à informação através de outras (clientes). Para além disso, a base de dados regista todos os dados referentes aos doentes, englobando tanto os aspectos ao nível das alterações ao nível do órgão, como ao nível do indivíduo.

Por fim o sistema de aquisição de conhecimento. Este sistema recorre a um conjunto de técnicas de aquisição de conhecimento em bases de dados (KDD – Knowledge Discovery in Databases), na perspectiva de descobrir relações entre os dados registados. Das técnicas existentes, foram usadas as que permitem uma representação compreensível das relações descobertas – informação, de modo a permitir a sua análise e utilização para tomada de decisões futuras. Estas representações capturam a estrutura das decisões numa forma explícita, ou, por outras palavras, ajudam a explicar algo acerca dos dados. No presente trabalho esta informação é representada sob a forma de regras, tais como "Todos os doentes com glaucoma têm uma perda de visão periférica significativa". Desta forma, espera-se vir a descobrir regras ainda desconhecidas entre a comunidade científica, e/ou validar as já conhecidas.

4 - ESTRUTURA DO TEXTO

A tese está estruturada em sete capítulos. Para um melhor enquadramento do projecto no seu contexto real e concreto, no Capítulo II - apresentam-se os elementos necessários à compreensão do domínio, assim como a descrição da sua situação actual. A par das noções de oftalmologia, surge uma descrição das necessidades dos doentes de subvisão, assim como do processo a que são submetidos durante a sua passagem pela consulta de subvisão.

No Capítulo III - descreve-se sucintamente a abordagem seguida para dar resposta às necessidades dos profissionais de subvisão, apresentando-se uma perspectiva sobre a tecnologia dos sistemas de informação, dando especial destaque aos sistemas transaccionais e de apoio à decisão. Faz-se ainda uma breve referência ao papel destes sistemas no domínio da medicina e propõe-se a arquitectura de um sistema capaz de dar respostas às necessidades da Consulta de Subvisão.

No Capítulo IV - descreve-se detalhadamente o sistema criado, dando especial relevo à sua arquitectura. Também os mecanismos de comunicação usados entre os vários módulos são descritos, assim como as políticas de segurança implementadas para essas comunicações.

No Capítulo V -, é apresentada uma descrição da Tecnologia de Sistemas de Aquisição de

Conhecimento, focando cada uma das suas etapas. Em seguida é explicitado o problema a tratar, fazendo-se uma descrição da estrutura e tipo dos dados, assim como a quantidade e a qualidade dos dados registados. Neste contexto é também feita a apresentação do objectivo a atingir – tipos de regras a descobrir. A descrição da Arquitectura do Sistema de Aquisição de Conhecimento criado é feita na última secção deste capítulo.

No Capítulo VI - efectua-se uma apresentação sumária das regras descobertas, acompanhadas por algumas apreciações relativas ao seu significado e à sua significância em geral. Por último no Capítulo VII -, termina-se com a apreciação global do trabalho desenvolvido e algumas linhas orientadoras para trabalho a desenvolver no futuro.

Capítulo II - ENQUADRAMENTO DO DOMÍNIO

Neste capítulo apresentam-se os elementos necessários à compreensão do domínio da Subvisão, assim como a descrição da situação actual nesse domínio. A par das noções de oftalmologia, surge uma descrição das necessidades dos doentes de subvisão, assim como do processo a que são submetidos durante a sua passagem pela consulta de subvisão do Hospital de Santa Maria. Este processo é descrito de forma pormenorizada, de modo a evidenciar cada uma das suas fases, dos seus objectivos e dos seus possíveis resultados.

1 - DEFINIÇÃO DE SUBVISÃO

A *subvisão* é uma perda parcial da visão que não pode ser corrigida com os óculos ou lentes habituais. No entanto, apesar desta perda ser bastante acentuada, o indivíduo continua a possuir alguma capacidade visual. A sua visão residual não é suficiente para realizar a maioria das actividades da sua vida diária, mas não atinge os níveis legalmente aceites para determinar o seu estado como sendo de cegueira.

Quando se fala de perda de visão existem quatro aspectos distintos a ter em consideração.

Tabela 1- Aspectos da Perda de Visão (adaptado de [Colenbrander 1999])

	ÓRGÃO		INDIVÍDUO	
Aspectos:	Alterações estruturais ou anatómicas	Alterações funcionais	Capacidades	Consequências sócio-económicas
Termos Neutros:	Condição de saúde	Funcionamento do órgão	Capacidades	Participação social
Perda:	Lesão	Debilidade	Incapacidade	Deficiência
Em VISÃO:		**Funções Visuais**	**Visão Funcional**	

Os dois primeiros aspectos dizem respeito ao órgão (no caso da visão: o olho), sendo o primeiro referente a alterações de índole anatómica ou estrutural, designadas por lesões, e o segundo referente a alterações funcionais ao nível do órgão – *funções visuais*, designadas por debilidades.

Os dois últimos aspectos estão relacionados com o indivíduo e não com o órgão. O primeiro dos dois descreve as capacidades do indivíduo e é designado por *visão funcional*, enquanto o segundo descreve as consequências socio-económicas provocadas pelas capacidades perdidas e são designadas por deficiências.

É possível afirmar-se que um indivíduo com *subvisão* sofre de algumas *incapacidades* provocadas por *debilitação* das suas funções visuais, devido a alguma *lesão* do olho. [Colenbrander 1999]

Como lesões que causam a subvisão encontra-se em primeiro lugar a *Degenerescência Macular da Retina*, resultado do envelhecimento, e que afecta primordialmente a visão central e consequentemente a *acuidade visual*. No caso específico das crianças, a principal causa é a prematuridade, traduzindo-se na maioria dos casos pela *Diminuição da Visão Cortical* e pela *Retinopatia da Prematuridade*. Para além destas, doenças como o *Albinismo*, as *Cataratas*, a *Retinopatia*, o *Glaucoma*, a *Histoplasmosis*, o *Nístagma*, o *Descolamento da Retina* ou a *Retinite Pigmentosa* são também causadoras das deficiências visuais, quer ao nível central e/ou periférico do campo visual, quer ao nível da acuidade visual, quer nas dificuldades perante a luminosidade ou contraste, quer simplesmente na visão da cor. A natureza destas enfermidades pode ser de ordem hereditária, congénita ou adquirida.

A subvisão não é portanto uma doença particular, mas sim a situação resultante das debilidades provocadas por uma ou mais lesões visuais, e respectivas consequências para um indivíduo enquanto entidade social.

Definida que está a doença, é necessário compreender as duas fases fundamentais do acompanhamento do doente: como se avalia a gravidade da situação (Avaliação) e como se melhora a sua qualidade de vida (Reabilitação).

1.1 - Avaliação

A área terapêutica da subvisão é recente e encontra-se inserida nos tradicionais serviços de oftalmologia. Apesar da sua curta idade, as evoluções nesta área têm sido significativas nos últimos anos. Inicialmente, a subvisão era distinguida das outras situações visuais, pela constatação da ineficácia dos tratamentos e cirurgias conhecidos. Quando um doente chegava à consulta de subvisão,

tipicamente tinha sido sujeito a um conjunto de exames e observações oftalmológicas que permitiam ao oftalmologista determinar a sua situação como doente de subvisão.

Com os avanços levados a cabo neste domínio, hoje essa determinação é feita tendo em conta dois dos aspectos apresentados na Tabela 1, nomeadamente: as funções visuais e a visão funcional.

Funções Visuais

As funções visuais (por exemplo: a acuidade visual ou o campo visual), podem ser avaliadas quantitativamente e expressas como quantidades relativas às medidas padrão. Isto permite caracterizar objectivamente o funcionamento dos órgãos visuais do indivíduo, e pode ser repetido ao longo do tempo, de modo a acompanhar a evolução da situação.

No entanto, apesar destas técnicas de medição terem vindo a evoluir nos últimos anos, apenas existem padrões estabelecidos e aceites pela comunidade, para duas das funções visuais: a acuidade visual e o campo visual. Funções como a sensibilidade ao contraste, a sensibilidade à luz ou a visão da cor, continuam sem poder ser avaliadas inequivocamente.

Visão Funcional

Pelo contrário, a visão funcional, que engloba a avaliação das capacidades/incapacidades do indivíduo (tais como a leitura, mobilidade ou orientação) podem apenas ser descritas qualitativamente, não existindo ainda escalas globalmente aceites. Estas avaliações podem ser efectuadas seguindo uma de três abordagens:

- **Estimação de capacidades**: a avaliação é baseada na medição das funções visuais, e a partir dessas deriva-se o valor das capacidades; este tipo de avaliação ignora todos os factores individuais.
- **Descrição directa das capacidades**: a avaliação é baseada na descrição da actuação do indivíduo, tendo em conta os factores individuais. Porém, é dependente da interpretação do observador e aquela descrição é algo subjectiva.
- **Abordagem híbrida**: usam-se as estimativas das capacidades como ponto de partida e fazem-se ajustes baseados na informação individual (se necessário). Repare-se que caso os ajustes existam, as observações que lhes dão origem devem ser bem documentadas e os argumentos ajustados ao caso. A documentação deve ser tal, que os ajustes efectuados sejam passíveis de serem reproduzidos.

É interessante referir que enquanto um profissional de visão tipicamente descreve a severidade

de um caso em termos da função visual afectada (“*a acuidade visual diminuiu duas vezes*”), um paciente apresenta as suas queixas em termos da perda de capacidades (“*já não sou capaz de ler o jornal”*).

É ainda de notar que, apesar dos avanços no domínio, a avaliação da gravidade da situação continua a ser estabelecida com base nos exames oftalmológicos efectuados e na experiência do especialista que acompanha o doente. Porém, este é apenas o último passo de um longo percurso. A escolha dos exames a efectuar é uma tarefa delicada, pois para além do seu custo, é necessário ter em conta o estado psicológico do doente. Na verdade, é habitual e compreensível a exaustão e desconfiança destes doentes face a novos exames e/ou tratamentos, uma vez que vão sendo sucessivamente submetidos a tratamentos sem qualquer resultado concreto. A escolha dos exames deve, naturalmente, ser realizada com base nas características do doente (dados pessoais, seus antecedentes e expectativas), pois, como se referiu, as doenças causadoras de subvisão podem ser diferenciadas, por exemplo, em função da idade do doente.

Em resumo, a avaliação é efectuada pela ponderação dos resultados registados durante o acompanhamento do doente e a avaliação do seu desempenho.

1.2 - Reabilitação

Como já foi dito, a situação oftalmológica do doente é inalterável ao nível físico e actualmente existem apenas duas formas de prosseguir o objectivo de melhorar a sua qualidade de vida: por um lado, o “aumento” da sua capacidade visual, por outro, a aquisição de novas estratégias para executar as mesmas tarefas.

O “aumento” da capacidade visual do doente é apenas um eufemismo, pois o dito aumento é conseguido artificialmente, quer através da utilização de alguns dispositivos visuais menos tradicionais (como é o caso de algumas combinações de lentes ou microtelescópios para efectuar ampliações - dispositivos ópticos), quer pela substituição dos utensílios usuais por utensílios adaptados às capacidades visuais dos doentes (como por exemplo livros/revistas com letras aumentadas, designados por dispositivos não-ópticos).

Na verdade, os dispositivos ópticos e não ópticos são somente objectos que minoram as dificuldades do doente, ou seja, melhoram a sua qualidade de vida sem melhorarem as suas funções visuais. Ao contrário da utilização destes dispositivos, o desenvolvimento de novas estratégias pretende fornecer um novo modo de superar as dificuldades, baseando-se no reaproveitamento da visão residual e no desenvolvimento das restantes capacidades sensitivas. De facto, e dado o

arrastamento progressivo da situação do doente ao longo do tempo, é habitual que ele tenha até certo ponto conseguido adquirir as suas próprias estratégias, de modo a minorar as suas dificuldades.

À combinação do desenvolvimento de estratégias e da utilização daqueles dispositivos dá-se o nome de *reabilitação em subvisão*.

Dois aspectos são ainda relevantes para a melhor compreensão desta fase: o acompanhamento do doente e a satisfação das suas expectativas. Habitualmente, nesta fase o doente é acompanhado por um profissional de reabilitação, que o estimula a adquirir o conjunto de estratégias que o ajudará a realizar as suas actividades diárias, começando por tentar satisfazer as suas necessidades mais prementes, indo deste modo de encontro às expectativas do doente.

Por exemplo, as crianças em idade escolar que sofrem de subvisão têm normalmente grandes dificuldades em acompanhar as actividades lectivas; acompanhadas por um professor do ensino especial, aprendem a usar materiais que os auxiliam com as suas tarefas escolares, tais como a utilização de guias durante a leitura ou papéis menos brilhantes de forma a melhorar a legibilidade dos textos. Por outro lado, treinam o movimento dos olhos de maneira a compensar as suas deficiências, por exemplo, com jogos semelhantes aos descritos em [Antunes 2000].

2 - DESCRIÇÃO DA SITUAÇÃO ACTUAL

A subvisão é uma área de investigação em franco desenvolvimento, existindo vários especialistas debruçados sobre este campo, não só da oftalmologia, como também da reabilitação pessoal e ciências sociais. Para além da investigação científica no campo da oftalmologia, tem-se desenvolvido um esforço significativo ao nível da reabilitação, tanto na sua componente estritamente pessoal como ao nível social, nomeadamente através da reinserção social dos doentes no mundo do trabalho. Também em Portugal, a Associação Portuguesa de Cegos e Amblíopes (ACAPO) tem vindo a desenvolver esforços significativos nesta área.

À data do início do projecto (1998), enfrentavam-se dois desafios principais: o de criar novas formas de avaliação e reabilitação, e o de avaliar e comparar os métodos usados.

Hoje, já existem algumas técnicas de avaliação padronizadas e aceites como mais correctas e adequadas, como as descritas em [Colenbrander 1999], mas continuam a existir várias instituições neste domínio que não registam digitalmente os dados dos seus doentes. Por esta razão, os avanços continuam a ser feitos com base em estudos tradicionais de avaliação de um conjunto restrito de indivíduos com subvisão. Visto que se trata (felizmente) de uma população reduzida, os estudos

resumem-se a analisar o comportamento de poucas dezenas de pessoas.

No âmbito da reabilitação quase tudo está por fazer. Por um lado, a elaboração dos planos de reabilitação é algo ainda difícil de explicar e efectuado exclusivamente por um grupo de especialistas em reabilitação ou ensino especial; por outro lado, a aquisição de novas estratégias, um dos passos mais importantes nesta fase, é quase inteiramente da responsabilidade do doente. Dada esta quase completa ausência de conhecimento na área, as dificuldades inerentes à formalização da elaboração e aconselhamento de planos de reabilitação aumenta profundamente.

Em resumo, dois aspectos se destacam pela sua extrema importância: a necessidade de estudar as implicações efectivas da debilitação das funções visuais nas capacidades efectivas dos doentes e a investigação ao nível da elaboração de planos de reabilitação adequados.

Acompanhamento na Consulta de Subvisão do Hospital de Santa Maria

A Consulta de Subvisão do Hospital de Santa Maria teve início em 1996, e habitualmente quando um doente chega à consulta já foi submetido a uma série de exames e cirurgias oftalmológicas, que se revelaram infrutíferas. É portanto bastante provável que o doente venha acompanhado de um diagnóstico prévio, composto essencialmente pelo conjunto de resultados dos exames oftalmológicos realizados, assim como alguns dados pessoais, tais como o seu historial médico.

Estes dados referentes às funções visuais, assim como as respostas dadas a um conjunto de questionários, que têm em vista a avaliação da visão funcional do doente, são registados numa base de dados criada em 1997 ([Pina 1997]). Os dados de cada doente são, evidentemente, confidenciais, não podendo ser acedidos excepto pelo médico, ou outro profissional responsável pelo seu acompanhamento. Porém, a introdução dos seus dados pessoais, assim como os resultados obtidos nos exames clínicos realizados, pode ser efectuada por funcionários do hospital.

Com as evoluções no domínio da avaliação surgidas nos últimos anos, o conjunto de atributos seleccionados em 1997 (como sendo as características essenciais e determinantes da situação do doente) têm evidenciado algumas insuficiências. A juntar a estas, deficiências quer ao nível da utilização, quer ao nível da extensão da própria base de dados, têm trazido dificuldades acrescidas ao projecto. Uma das deficiências mais significativas é a falta de flexibilidade para registar dados diferentes para doentes substancialmente diferentes, por exemplo, crianças *versus* adultos, analfabetos *versus* leitores. Por outro lado, não permite um acesso ao sistema, simultâneo, a vários utilizadores.

Apesar do esforço de informatização levado a cabo nos últimos anos, a fase de avaliação continua a resumir-se à recolha, registo e apreciação dos dados feita pelo médico, sendo baseada

exclusivamente na sua experiência enquanto profissional de subvisão. Assim, tanto a adopção de novos exames, como a apreciação do desempenho do doente face a um conjunto de tarefas é da responsabilidade exclusiva do médico. Repare-se no entanto, que para além da conhecida falta de médicos, maior é a falta de pessoal especialista nesta área.

Ao nível da reabilitação, as coisas não são melhores. Para além da selecção dos métodos de reabilitação baseada exclusivamente na experiência dos profissionais de reabilitação, quer a aplicação destes métodos, quer os resultados daí obtidos não são registados. A este nível, o essencial é a definição de um plano de reabilitação, constituído por uma sequência de passos, em que cada um deve atingir um sub-objectivo previamente identificado; cada plano deve ter em conta não só as características específicas de cada doente, mas também as suas expectativas. Repare-se no entanto, que os planos usados ainda são bastante rudimentares, e para além da sua escassez, muito pouco formalizados.

Por outro lado, actualmente e como atrás foi referido, a reabilitação é sobretudo ao nível psicossocial, e não ao nível da oftalmologia.

Capítulo III - SISTEMA DE INFORMAÇÃO

Neste capítulo descreve-se sucintamente a abordagem seguida para dar resposta às necessidades dos profissionais de subvisão, apresentando-se uma perspectiva sobre a tecnologia dos sistemas de informação, dando especial destaque aos sistemas transaccionais e de apoio à decisão.

Faz-se ainda uma breve referência ao papel destes sistemas no domínio da medicina e propõe-se a arquitectura de um sistema capaz de dar respostas às necessidades da Consulta de Subvisão.

1 - ABORDAGEM SEGUIDA

A necessidade de modernizar o serviço de acompanhamento dos doentes, quer na sua fase de avaliação quer na sua fase de reabilitação, torna-se clara.

Em primeiro lugar, é inadiável a melhoria do sistema informático existente, de modo a ser capaz de dar resposta às especificidades do domínio da subvisão, contemplando as suas duas fases, e abrangendo os dois aspectos relevantes já referidos: as funções visuais e a visão funcional. Em segundo lugar, a possibilidade de analisar os dados registados de forma sistemática e rigorosa, de modo a facilitar e melhorar os estudos realizados sobre as relações entre as funções visuais e a visão funcional.

Um meio de ultrapassar estas dificuldades seria desenvolver um sistema de suporte ao registo dos dados dos doentes, que pudesse ser usado não só numa instituição, mas por exemplo por várias das instituições semelhantes do mesmo país. Só desta forma serão registados dados suficientes que permitam uma análise estatística representativa, que poderá viabilizar e facilitar os estudos referidos.

Em termos práticos, um tal sistema deveria ter um conjunto de características fundamentais, nomeadamente:

- uma arquitectura definida por forma a permitir o acesso simultâneo aos dados, por parte de diferentes utilizadores (profissionais do serviço), de maneira a alargar a utilização do sistema e consequentemente a abranger mais doentes;
- uma base de dados com uma estrutura flexível, capaz de registar quer a avaliação das funções visuais consideradas relevantes, quer as capacidades individuais levadas em consideração para a determinação da visão funcional;
- um mecanismo de conjugação da avaliação quantitativa das funções visuais com a avaliação qualitativa da visão funcional.

Se o primeiro ponto é de fácil resolução pelo uso das novas tecnologias da informação, nomeadamente a tecnologia dos *sistemas distribuídos*, o segundo ponto aborda algumas questões não triviais de resolver, tais como a escolha da representação para o registo ao nível da visão funcional. Não menos problemático é o terceiro ponto, onde existem apenas duas formas adequadas à resolução do problema: a utilização intensiva do conhecimento dos especialistas da área (através da tecnologia dos *sistemas periciais*) ou a descoberta desse e de novo conhecimento, através do estudo das relações existentes entre os dados registados para cada doente (recorrendo às tecnologias dos *sistemas de aquisição de conhecimento*).

Uma vez que, como já foi mencionado, o conhecimento especializado neste domínio é ainda reduzido, questões relacionadas com o estudo do comportamento de doentes com subvisão conduzem-nos impreterivelmente à análise dos dados registados. Através desta análise espera-se não só obter uma melhor aproximação ao problema, como também contribuir significativamente para o estudo das implicações das lesões visuais nas capacidades efectivas dos doentes.

Neste contexto, a utilização da tecnologia dos *sistemas de informação***,** em particular dos *sistemas de aquisição de conhecimento,* é fundamental para modernizar o sistema actualmente em uso.

2 - TECNOLOGIA DOS SISTEMAS DE INFORMAÇÃO

"Um Sistema de Informação pode ser definido como um conjunto de componentes interligados, funcionando juntos para recolher, processar, armazenar, e disseminar a informação, de modo a suportar a tomada, coordenação, controlo, análise e visualização das decisões numa organização." [Laudon 1998]

Perante tais funcionalidades e utilizações, torna-se clara a importância de tais sistemas no funcionamento de qualquer organização. Porém, a diversidade de interesses, especialidades e níveis aí

existentes, é de tal ordem que um único sistema de informação é insuficiente para dar resposta a todas as suas necessidades. Dada esta diversidade, existem vários tipos de sistemas de informação, dos quais se destacam os sistemas transaccionais ao nível operacional e os sistemas de suporte à decisão ao nível da gestão.

Um *sistema transaccional* é o sistema básico de uma organização, que funcionando ao nível operacional, executa e regista as transacções diárias efectuadas pela e na organização. A este nível as tarefas, recursos e objectivos são pré-definidos e altamente estruturados, ou seja, bem conhecidos. Os sistemas transaccionais são os sistemas fundamentais para o funcionamento diário de uma organização: qualquer falha num destes sistemas pode inviabilizar o normal funcionamento da actividade da organização.

Quanto aos *sistemas de apoio à decisão* são desenhados especialmente para melhorar o processo de tomada de decisões. "Um sistema de apoio à decisão é um sistema computacional ao nível da gestão de uma organização, que combina dados, ferramentas analíticas e modelos para apoiar a tomada de decisões semi-estruturadas ou não estruturadas" [Laudon 1998]. Por problema estruturado entenda-se um problema frequente e conhecido pela organização, para o qual é conhecida a solução adequada; por problema não-estruturado entenda-se um problema novo e não usual, para o qual não é conhecida uma solução algorítmica.

Um sistema de apoio à decisão tem normalmente um maior poder analítico, e é criado explicitamente com uma variedade de modelos para analisar os dados. Um dos tipos de sistemas de apoio à decisão, são os *sistemas de apoio à decisão orientados aos dados*, que apoiam a tomada de decisão ao permitir aos utilizadores extraírem informação útil a partir das grandes quantidades de dados recolhidas pela organização, habitualmente através dos seus sistemas transaccionais. Esta extracção de informação é normalmente feita pela utilização de técnicas de *data mining* ou de aquisição de conhecimento.

Data mining é uma tecnologia com raízes na área da Inteligência Artificial, para descobrir padrões não explícitos e relações entre os dados existentes em grandes bases de dados, de modo a inferir regras para predizer comportamentos futuros.

Papel dos sistemas de apoio à decisão no domínio da medicina

Actualmente, existe já uma longa tradição no desenvolvimento de sistemas de apoio às actividades clínicas que se centram na assistência aos médicos durante a fase de diagnóstico. Porém, por um lado, o diagnóstico é apenas uma das muitas fases da medicina que necessita de apoio, por

outro lado, nem sempre os sistemas existentes foram convenientemente usados e criados de forma a trazerem mais valias ao acompanhamento dos doentes.

De acordo com Coiera [Coiera 1994] existem dois aspectos fundamentais a ter em conta no desenvolvimento de sistemas de apoio à prática clínica, nomeadamente: suportar os requisitos dos médicos em vez de lhes ditar novas práticas, e focar o desenvolvimento dos sistemas no suporte do processo de gestão do doente, no seu todo, em vez de abordar apenas a fase de diagnóstico.

Ainda de acordo com Coiera, existe um conjunto alargado de áreas da prática clínica em que a utilização de sistemas de apoio à decisão trazem benefícios. De entre estas aplicações destacam-se: ferramentas para desenhar, construir e manter protocolos clínicos; sistemas de aconselhamento para dosagem de medicamentos; sistemas para reconhecimento e interpretação de imagens ou sinais; sistemas que assegurem a qualidade da informação armazenada na base de conhecimento; sistemas de apoio à educação / reabilitação de doentes; e finalmente, sistemas que suportem a gestão da informação clínica dos doentes.

Seguindo de perto este estudo de Coiera, e comparando as suas indicações com a situação actual da Consulta de Subvisão do Hospital de Santa Maria, evidencia-se a necessidade de expansão do sistema em uso, de forma a cobrir todas as fases do processo e não somente a da gestão da informação dos doentes.

Esta abordagem está perfeitamente de acordo com o que tem sido feito nos últimos anos: primeiro com a criação da aplicação da gestão da informação dos doentes [Pina 1997], e depois com a criação de alguns testes de avaliação para crianças [Antunes 1998] e de alguns exercícios de reabilitação conhecidos [Antunes 2000], têm-se portanto criado pequenas aplicações que mais tarde podem ser integradas num sistema completo.

Porém, e como também foi apresentado, o conhecimento existente na área da subvisão é ainda bastante insuficiente para o sucesso de um sistema com aquelas características. Portanto, resta apenas uma solução: tentar adquirir mais conhecimento nesta área, de forma automática ou não, de modo a possibilitar a criação de tal sistema.

O objectivo da presente tese é desenvolver um tal sistema. Em primeiro lugar, substituindo o actual sistema de gestão dos dados dos doentes (*sistema transaccional*), de modo a suprimir as actuais faltas, e em segundo lugar, o desenvolvimento de um *sistema de aquisição de conhecimento* que possibilite a descoberta das relações existentes entre os dados registados de modo a facilitar a futura criação do sistema integrado.

Em seguida analisa-se que funcionalidades deve ter o sistema proposto de forma a conciliar estas características.

3 - FUNCIONALIDADES E ESPECIFICAÇÃO

De modo a integrar as várias etapas da consulta de subvisão com um mecanismo de aquisição de conhecimento que ajude na condução de novas investigações nesta área (e na própria melhoria do sistema), o sistema proposto, de uma forma genérica, deve permitir gerir os dados do doente e extrair informação a partir desses dados.

Mais concretamente, o sistema será usado pelo pessoal do hospital no registo dos dados dos doentes da Consulta de Subvisão, caracterizando-se esses dados por um conjunto de atributos identificativos, atributos relativos a diagnósticos e a exames já realizados, e finalmente atributos relativos ao desempenho do doente na sua vida diária. Uma vez que estes dados são confidenciais, o acesso ao sistema deve ser condicionado ao papel desempenhado pelo pessoal do hospital no acompanhamento dos doentes. Assim, apenas o médico responsável pelos doentes deverá ter livre acesso a todos os dados dos seus doentes. Porém, para além do médico responsável, outros funcionários deverão ter a possibilidade de introduzir resultados de exames e outros. Por fim, deverá existir a possibilidade de outros colaboradores da consulta terem acesso a uma apreciação global do doente. Para além da gestão dos dados dos doentes, o sistema deve permitir ao médico responsável e/ou à sua equipa de investigação, a formulação de pedidos de análise sobre os dados registados.

Com base na descrição das funcionalidades apresentada, pode-se desenhar o seguinte diagrama de fluxos de dados.

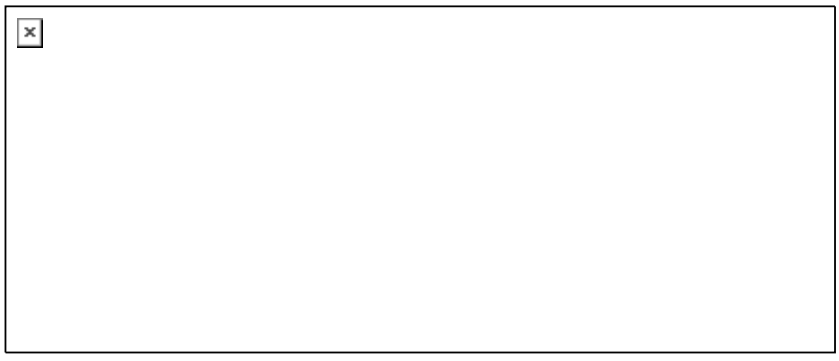

Figura 1 – Diagrama de Fluxo de Dados de nível 0 (DFD)

Conforme se mostra na **Figura 1**, o sistema é constituído por cinco elementos distintos: um processo responsável por gerir a consulta propriamente dita (‘Gestão da Consulta’); o processo

responsável pela aquisição de conhecimento (‘Aquisição de Conhecimento’); a base de dados onde estão armazenados os dados relativos aos doentes (‘Doentes’), e que responde indistintamente aos pedidos dos dois processos; e por fim a existência de duas classes distintas de utilizadores (entidades externas), uma referente à gestão da consulta (‘Profissionais da Consulta’) e outra de manipulação do processo de aquisição de conhecimento (‘Investigadores’).

Com base no diagrama de fluxo de dados apresentado na **Figura 1**, consegue-se identificar as componentes básicas, que conceptualmente constituirão a arquitectura do sistema. Assim,

- Em primeiro lugar, a existência de uma *base de dados* torna-se estritamente necessária. Repositório de todos os dados mantidos em computador, a base de dados é responsável por possibilitar uma consulta eficiente aos dados referentes aos doentes da consulta de subvisão, aí armazenados.

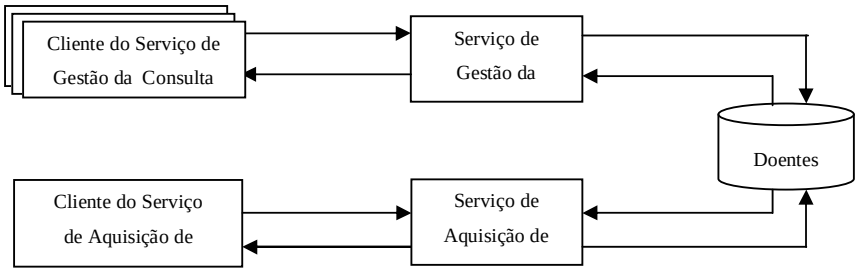

Figura 2 – Arquitectura geral do sistema

- De forma a manipular dados armazenados na base de dados, são necessários mecanismos capazes de interagir com eles. Com este intuito criam-se as aplicações que se designam por *serviço de gestão da consulta* e *serviço de aquisição de conhecimento*, de forma a servirem os pedidos efectuados pelos utilizadores (entidades externas atrás designadas por “profissionais da consulta” e “investigadores”). Estes serviços acedem à base de dados para dar resposta àqueles pedidos, sendo simultaneamente responsáveis pela garantia da coerência e correcção dos dados veiculados entre os utilizadores e a base de dados.

- No fim da cadeia não podem deixar de se encontrar as aplicações que tornam possível um acesso generalizado aos dados armazenados e aos serviços responsáveis pelo seu processamento. Aplicações desta natureza são normalmente designadas por aplicações cliente, e as criadas no sistema designadas por *cliente do serviço de gestão da consulta* e *cliente do serviço de aquisição de conhecimento*. Estas têm como principal função fornecer a interacção entre os utilizadores e os sistemas de processamento e armazenamento dos dados atrás descritos. Por interacção entenda-se a formulação de pedidos ao sistema e a aplicação

do tratamento adequado às respostas enviadas pelo serviço respectivo.

Em seguida apresentam-se, de forma detalhada, as funcionalidades que cada um destes componentes deve apresentar.

Gestão da Consulta

Os clientes são os únicos componentes a interagir directamente com os utilizadores, e em nome destes, junto do serviço de gestão da consulta. Como tal, devem apresentar uma interface gráfica, o mais adequada possível aos utilizadores considerados.

No caso do *Cliente do Serviço de Gestão da Consulta*, os utilizadores finais são os profissionais da equipa médica e de reabilitação, de onde se destacam os médicos especialistas, psicólogos, professores de ensino especial, entre outros funcionários do hospital. Desta forma, estamos provavelmente perante utilizadores com baixa motivação e pouca experiência na utilização de sistemas informáticos. Recorrendo às técnicas de Interfaces Pessoa Máquina para apoiar a escolha do tipo de interface a usar, a decisão recaiu numa interface baseada em *formulários* e *menus*, uma vez que se adequa perfeitamente ao tipo de utilizador descrito, assim como à natureza da tarefa que vão desempenhar no sistema: uma tarefa bem estruturada e pouco complexa.[Mayhew 1992]

Ainda a referir, há a existência de diferentes classes de acesso ao sistema, de acordo com as responsabilidades de cada utilizador no funcionamento da consulta.

O *Serviço de Gestão da Consulta* é a componente do sistema responsável por gerir quer os acessos aos dados, quer a sua manipulação. Identificam-se assim dois aspectos principais a gerir: os dados relativos aos doentes e os utilizadores que a podem manipular. Por um lado, há a garantir o registo dos utilizadores e a atribuição dos seus direitos de acesso; por outro, é necessário manter a coerência dos dados armazenados na base de dados. De acordo com estas necessidades, o serviço deve suportar as seguintes funções:

Tabela 2 – Funções do Serviço de Gestão da Consulta

FUNÇÕES	DESCRIÇÃO
Registar Utilizador	Regista um novo utilizador no sistema, guardando a sua *password* e atribuindo-lhe o nível de acesso adequado
Login / Logout	Permite a entrada e a saída do utilizador do sistema
Registar Doente	Introduzir um novo doente no sistema
Consulta	Consultar os dados armazenados
Alteração	Alterar os dados armazenados

Aquisição de Conhecimento

O *Serviço de Aquisição de Conhecimento* é a componente do sistema que visa a análise

sistemática e rigorosa dos dados, ajudando assim a realização de estudos sobre o impacto da debilitação dos órgãos nas capacidades efectivas dos doentes. Esta análise é designada por descoberta de padrões nos dados, nos domínios do *data mining* e da aprendizagem artificial. Estes padrões são capazes de tornar explícitas as relações implícitas existentes entre os vários dados registados, e para além disso deve possibilitar a consulta dos dados a um nível estatístico.

Quanto ao *Cliente do Serviço de Aquisição de Conhecimento* este deve apresentar uma interface o mais simples e flexível possível, de modo a possibilitar a análise dos dados de forma eficiente por parte quer de profissionais de informática quer de investigadores na área da saúde.

As funcionalidades deste serviço, assim como as suas características serão apresentadas detalhadamente mais adiante no Capítulo V -1 -.

Base de Dados

A *base de dados* é entendida normalmente como o suporte físico da informação armazenada (num computador). Porém o que lhe confere o estatuto que actualmente possui é a separação entre os componentes físicos usados para o armazenamento da informação (*base de dados física*) e a representação abstracta da mesma (*base de dados lógica*).

Fazendo uso destas vantagens, a base de dados criada armazena todos os dados manipulados pelo sistema, quer na sua vertente da consulta, quer da aquisição de conhecimento. Incluem-se nestes dados, tanto os referentes aos doentes, como aos utilizadores.

Capítulo IV - ARQUITECTURA E DECISÕES DE IMPLEMENTAÇÃO

No presente capítulo descreve-se detalhadamente o sistema criado, dando especial destaque à sua arquitectura. Também os mecanismos de comunicação usados entre os vários módulos são descritos, assim como as políticas de segurança implementadas para essas comunicações.

1 - ARQUITECTURA GERAL DO SISTEMA DE INFORMAÇÃO

Definidas as várias funcionalidades que o sistema deve incorporar, segue-se uma descrição da abordagem seguida para o concretizar. Dado que os principais objectivos de índole tecnológica são a portabilidade, a segurança e a escalabilidade, e que para além destas propriedades existe a necessidade de criar um sistema de aquisição de conhecimento, a *centralização* da informação num sistema único torna-se vantajosa. Porém, o acesso ao sistema deve ser *distribuído*, de forma a possibilitar a interacção simultânea de vários utilizadores com o sistema.

De forma a criar um sistema com tais características, a decisão sobre o tipo de arquitectura a usar recaiu sobre a *arquitectura do tipo cliente/servidor*.

"A designação de arquitectura Cliente/Servidor estabelece a distinção entre dois tipos de processos com comportamentos diferentes e, portanto, assimétricos na sua estrutura, que podemos sinteticamente caracterizar da seguinte forma: os servidores implementam um conjunto de funções de interesse geral para outros processos que remotamente lhes podem aceder. [...] Os clientes efectuam a interface com os utilizadores, podem executar parte das aplicações localmente e acedem remotamente a processos servidores" [Marques 1998].

Seguindo de perto este modelo, o servidor fornece um conjunto de funções que permitem a manipulação da informação existente na base de dados, e os clientes possibilitam a interacção entre os

utilizadores e o sistema.

Para além da definição do modelo de arquitectura a seguir, e dada a escolha efectuada, torna-se necessário definir o modo de comunicação a estabelecer entre os clientes e o servidor.

De acordo com a existência dos dois clientes distintos, também se nota a existência de dois modelos de comunicação diferentes. Assim, enquanto que para o *serviço de gestão da consulta* a transferência dos dados se faz através do protocolo *HTTP*, para o *serviço de aquisição de conhecimento* a comunicação efectua-se através da *Invocação de Métodos Remotos* (*RMI*), que se justifica mais adiante.

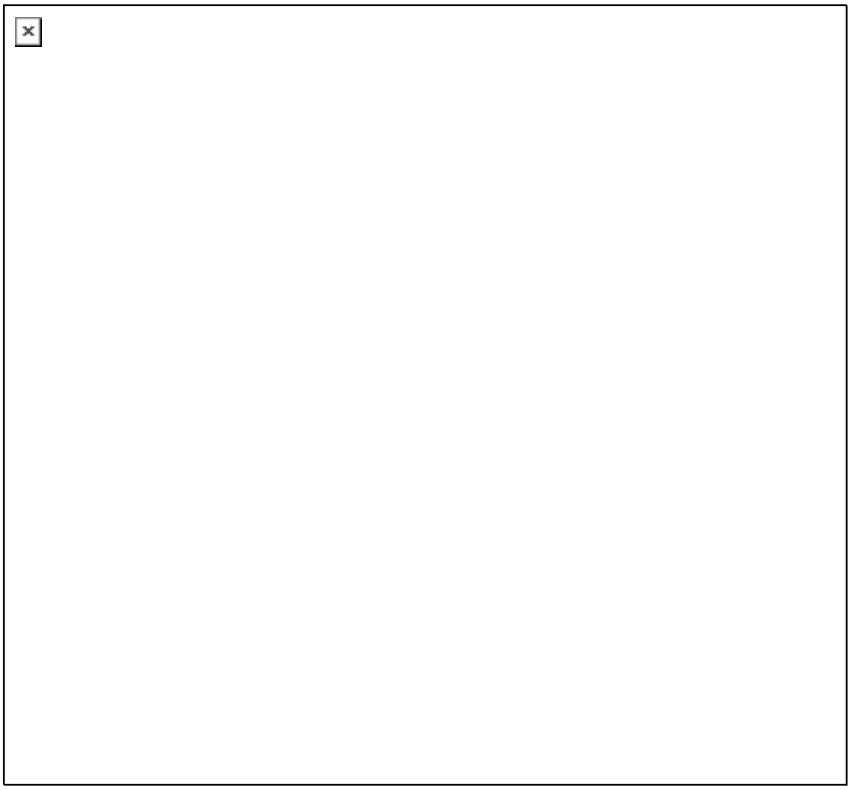

Figura 3 – Arquitectura do sistema por módulos

Para além das determinações da arquitectura e mecanismos de comunicação a usar, foi necessário a tomada de outras decisões para garantir a satisfação das propriedades referidas (portabilidade, segurança e escalabilidade). De entre estas decisões destacam-se duas: a criação de interfaces suportadas por *browsers* (*cliente da consulta*) e a utilização do Java como linguagem de programação que garantem a portabilidade destas aplicações.

Em seguida apresenta-se cada um dos componentes do sistema detalhadamente, de modo a justificar as opções tomadas.

2 - BASE DE DADOS

Do ponto de vista do armazenamento, uma base de dados não é mais que um mero repositório de dados, neste caso, dos utentes da consulta de subvisão e dos seus utilizadores. No entanto, uma base de dados por si só não é suficiente para dar resposta às necessidades de uma aplicação com as características pretendidas, é por esta razão que se engloba, como é habitual, o servidor de gestão de base de dados (SGBD) nesta componente. É pois o SGBD quem controla o acesso aos dados e garante a sua coerência.

O SGBD escolhido para a implementação do sistema foi o postgresql [Lockhart 1998], servidor que funciona em ambiente Unix, e em particular no sistema operativo Linux, escolhido para suportar os serviços fundamentais do sistema.

Modelo da Base de Dados

Como foi dito, a aplicação de manipulação dos dados registados na base de dados, em uso desde 1997, tem vindo a evidenciar algumas dificuldades. Algumas delas já apontadas no Capítulo II -, como por exemplo a impossibilidade de acesso generalizado aos dados, o registo de dados distintos para doentes substancialmente diferentes e, por último, a impossibilidade de registar os resultados de vários diagnósticos ou tratamentos obtidos em datas diferentes, em número ilimitado.

Se o primeiro dos problemas apontados não está directamente relacionado com a estrutura da base de dados (base de dados lógica) propriamente dita, e sim com a aplicação de manipulação, o mesmo não é verdade quanto aos outros dois problemas.

A questão do registo de dados adequada ao doente relaciona-se principalmente com as questões colocadas ao doente durante a avaliação da visão funcional, que como já foi apresentado, consiste na recolha de dados sobre as capacidades efectivas do doente com base na análise das suas actividades quotidianas. Este problema foi já abordado no âmbito de um trabalho final de curso realizado em 1998 ([Antunes 1998] e [Neves 2000]) tendo sido sugerido que estas questões fossem escolhidas de entre um leque variado, de acordo com algumas das características mais discriminantes, tais como a idade, o sexo, a escolaridade e o local de residência.

Quanto ao registo de vários diagnósticos e/ou tratamentos, evidencia-se a necessidade de remodelar a base de dados existente de forma a tornar trivial esta questão. Um outro aspecto a relatar é a existência de um elevado número de atributos das entidades relacionadas com os exames oftalmológicos, que na prática se traduz pela quase inexistência de dados registados.

Dado tudo isto, e sobretudo o *feedback* da equipa da consulta de subvisão após a utilização da aplicação nos últimos anos, torna-se inadiável o redesenho da base de dados de modo a resolver estes problemas. Deste modo o desenho da base de dados tem vindo a ser modificado tendo aqueles objectivos em mente. Recorrendo à modelação ER (*entity/relationship*) pode-se apresentar o seguinte modelo, como base de reconstrução da base de dados. (No Anexo A apresentam-se os atributos de cada uma das entidades, assim como o seu tipo de dados).

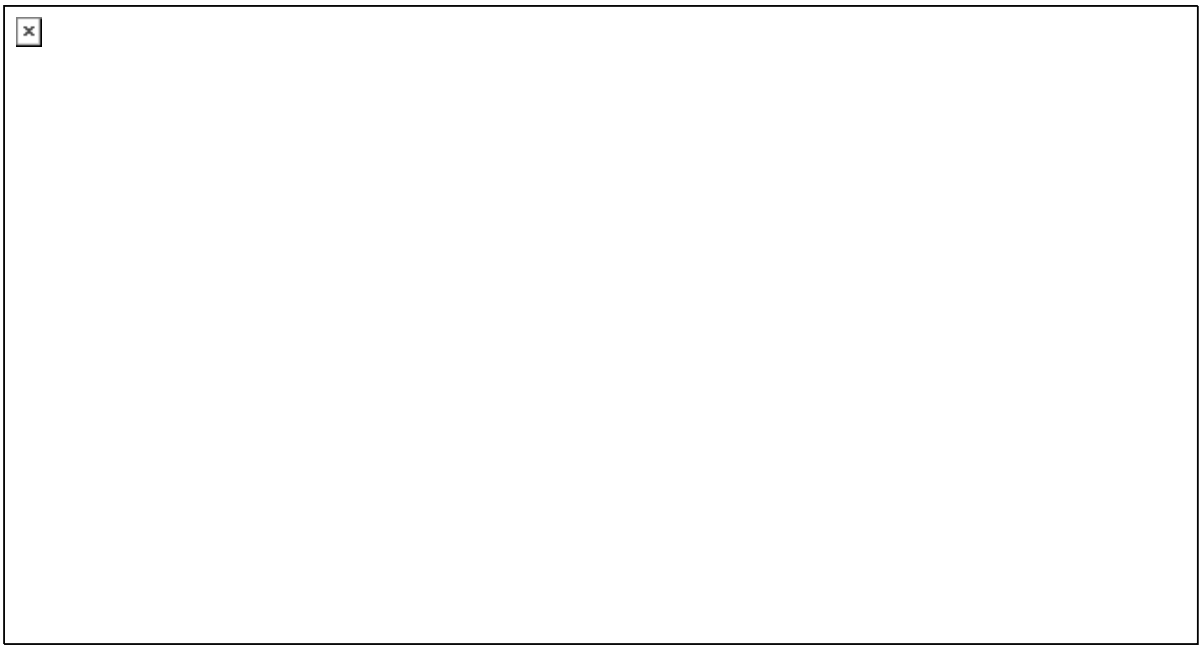

Figura 4 – Modelo ER da base de dados

A base de dados operacional está modelada na terceira forma normal (3NF) por forma a garantir um nível de redundância de dados mínimo. De modo a facilitar a implementação do módulo de aquisição de conhecimento, as várias entidades foram organizadas em agrupamentos lógicos designados esquemas. Para maior facilidade de explicação, todas as referências à estrutura da base de dados serão feitas de acordo com esta abstracção.

Tabela 3 - Distribuição das Entidades da Base de Dados por Esquemas

	ESQUEMAS			
	Identificativos	**Diagnóstico**	**Funções Visuais**	**Visão Funcional**
ENTIDADES	Doente	Diagnóstico	Funções Visuais	Visão Funcional
	Agregado Familiar	Incapacidade	Exame	Questão
		Tratamento		
		Historial Clínico		

3 - MÓDULO DE LIGAÇÃO À BASE DE DADOS

Para além da utilização do SGBD referido, foi criado um módulo intermédio, de forma a garantir a modularidade do sistema, e que se dedica exclusivamente a estabelecer a interface entre a base de dados e os serviços. Este módulo ao acumular em si todos os comandos SQL, permite aumentar a modularidade do sistema, pelo que qualquer alteração à base de dados afectará exclusivamente este módulo, não se propagando qualquer alteração aos níveis superiores.

Como já foi referido, o sistema foi programado usando a linguagem Java, que se adequa perfeitamente à natureza da aplicação a criar. Em particular este módulo recorre ao *package* `java.sql`, que fornece o *package* JDBC (um mapeamento do protocolo ODBC para Java), onde são fornecidos os meios para aceder e usar bases de dados relacionais (como é o caso presente). Assim, esta componente implementa-se a partir de uma classe - `JdbcConsulta` -, que contém em si um objecto do tipo `Connection`. Este objecto não é mais do que uma instância da classe de Java responsável por estabelecer a ligação à base de dados em causa, pelo que é através de objectos deste tipo que se efectua toda a manipulação de dados ali existente.

Ao concentrar neste módulo todas as *queries* a efectuar à base de dados, acumulam-se em si as responsabilidades de receber os pedidos de informação e de dar as respectivas respostas aos clientes.

4 - MÓDULO DA GESTÃO DA CONSULTA

A parte do sistema que aqui se designa por módulo da consulta é na verdade a soma do *cliente* e respectivo *serviço de gestão da consulta de subvisão***,** e que implementam o sistema transaccional propriamente dito.

Uma vez que um dos principais objectivos do projecto é garantir que o acesso ao sistema é possível, independentemente da máquina e sistema operativo usados, a decisão da implementação dos *clientes da consulta* seguindo a abordagem dos *clientes magros* (*thin clients*) revelou-se fundamental para garantir não só a portabilidade e escalabilidade do sistema, mas também a sua fácil manutenção. Com esta abordagem, torna-se possível uma evolução contínua do sistema, uma vez que a produção de sucessivos melhoramentos não implica a substituição do software existente nas máquinas de cada um dos utilizadores finais, mas apenas no lado do servidor.

Concretamente, o *cliente da consulta* é implementado sobre um *browser* (decisão que foi determinante para a arquitectura deste módulo), e é representada na Figura 5 – Módulo de gestão da consulta.

De acordo com a decisão de usar *browsers* para suportar a aplicação cliente, optou-se pela utilização do HTTP como protocolo de comunicação entre o cliente e o servidor.

Por seu lado, o serviço de gestão da consulta é implementado através da conjugação de um *servlet* (o `ServletConsulta`) e de um objecto do tipo `JdbcConsulta` (já atrás descrito).

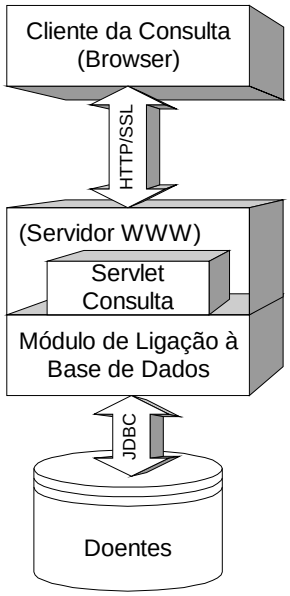

Figura 5 – Módulo de gestão da consulta

ServletConsulta

"Um servlet não é mais que um pedaço de código, normalmente (mas não obrigatoriamente) programado em Java, que se encontra associado a um servidor, aumentando assim a sua funcionalidade. Este servidor é normalmente um servidor WWW embora isso não seja imperativo quando os servlets são programados em Java" [Ferreira 1999]

Do ponto de vista da implementação, a classe `ServletConsulta` estende a classe `HttpServlet` - uma classe do Java que fornece os mecanismos básicos para criar um *servlet* dedicado ao protocolo escolhido. No presente caso, o *servlet* é programado em Java e encontra-se associado a um servidor de WWW - o *Apache* [Apache Group 1999]. Deste modo, e através do *servlet* consegue-se suportar serviços adicionais para além do simples acesso a páginas HTML, tais como aceder à base de dados de forma controlada, e criar páginas dinâmicas que apresentem os dados aí registados, ou tão somente enviar os dados introduzidos para os armazenar convenientemente.

"O funcionamento básico de um servlet consiste em receber pedidos de processos cliente, indirectamente através de um servidor […], e servi-los adequadamente enviando a resposta

correspondente. (...)Quando um *servlet* serve um pedido vindo de um cliente, existem duas interfaces Java a considerar: ServletRequest e ServletResponse que encapsulam todos os aspectos relativos, respectivamente, ao pedido efectuado pelo cliente e à resposta a ele enviada.[…] Quando o servlet em causa é para ser associado a um servidor WWW, existem duas interfaces denominadas HttpServletRequest e HttpServletResponse, que estão vocacionadas para o protocolo HTTP. Os seus métodos permitem não só aceder de uma forma simples a toda a informação vinda do cliente segundo o protocolo HTTP, mas também facilitam a construção da respectiva resposta." [Ferreira 1999]

A opção de utilizar um *servlet* adequa-se perfeitamente às necessidades da aplicação, pois este fornece um mecanismo simples e seguro de efectuar a comunicação entre os clientes e o servidor.

Sincronização

Como já foi referido, o sistema deve permitir a interacção simultânea de vários utilizadores. Ora, para que isto seja possível é necessário garantir a integridade e coerência dos dados, apesar dos vários acessos à base de dados.

Também como já foi referido, a manipulação dos dados existentes na base de dados é feita exclusivamente pelos objectos do tipo `JdbcConsulta`. É pois através da conjugação deste objecto com o *servlet* que se torna possível a manipulação dos dados. Quanto à garantia da integridade e coerência da informação, esta é responsabilidade do *servlet*, pois os métodos que interagem com o cliente da aplicação (o `doPost` e o `doGet`) são métodos síncronos - *synchronized*, o que significa que "quando uma tarefa invoca um método síncrono sobre um objecto, o trinco lógico que protege o objecto é fechado. Qualquer outra tarefa que invoque um método síncrono sobre o mesmo objecto fica bloqueada até que o trinco seja aberto." [Arnold 1998]

Deste modo e uma vez que apenas se acede à informação através destes dois métodos, o problema da integridade da informação está resolvido. Repare-se ainda que estes métodos têm como argumentos um objecto do tipo `HttpServletRequest` e outro do tipo `HttpServletResponse`, e é através destes dois objectos que recebem e enviam aos clientes os pedidos e respostas, respectivamente.

Há ainda a destacar um aspecto relevante no objecto `HttpServletRequest`. Este objecto contém um objecto do tipo `HttpSession`, que identifica a sessão (informação de contexto) e consequentemente o utilizador, responsável pelo envio do pedido, e que facilmente garante a correcção do encaminhamento das respostas ao utilizador correcto.

5 - MÓDULO DE AQUISIÇÃO DE CONHECIMENTO

Por último, o módulo de aquisição de conhecimento. Este módulo tem como objectivo a implementação do *cliente* e respectivo *serviço de aquisição de conhecimento*. Para a criação deste módulo foi necessário decidir sobre a forma de implementação do ***cliente:*** implementá-lo na mesma máquina que o servidor, ou numa outra qualquer máquina. Se por um lado a primeira abordagem simplificaria a implementação do módulo, inviabilizaria a sua manipulação fora da instituição onde o servidor estará instalado (o local habitual da consulta). Ao seguir a segunda abordagem torna-se possível o acesso à informação gerada pelo processamento do *data mining* remotamente. Por esta razão, foi necessário proceder ao estabelecimento de um mecanismo de comunicação entre os dois componentes.

Pelo lado do servidor executam-se todos os acessos à base de dados (desta vez exclusivamente para operações de leitura) assim como o processamento dos pedidos efectuados pelo cliente. Pertence ao servidor a responsabilidade de processar o *data mining* sobre aqueles dados. Uma vez que não existe a necessidade de generalizar a utilização deste serviço, não se utilizou a abordagem dos clientes magros (*thin clients*), de modo que foi necessário usar um outro mecanismo de comunicação entre as duas entidades a Invocação de Métodos Remotos (RMI).

Esta escolha deve-se fundamentalmente às facilidades que este mecanismo oferece face às alternativas existentes. Considerem-se os casos da utilização de *sockets* e de *Chamada a Procedimentos Remotos (RPC)*. Por um lado, a utilização de *sockets* obrigaria a um esforço significativo no desenho de um protocolo de codificação e descodificação das mensagens a trocar; por outro, e apesar do mecanismo de *Chamada a Procedimentos Remotos* já não exigir esse esforço, as dificuldades ao nível da comunicação entre objectos residentes em espaços de endereçamento distintos ainda são significativas. [Marques 1998]

Mecanismo de Invocação a Métodos Remotos (RMI)

O mecanismo de Invocação a Métodos Remotos supera estas dificuldades, ao ir de encontro à semântica da invocação de métodos dos objectos. Este mecanismo permite assim suportar indistintamente a invocação de métodos quer de objectos locais quer de objectos remotos. Por objectos remotos entenda-se os objectos cujos métodos podem ser invocados numa outra máquina virtual Java, potencialmente num computador diferente. Um objecto deste tipo é descrito por uma ou mais interfaces remotas, interfaces estas que declaram os métodos do objecto remoto. A Invocação a Métodos Remotos é a acção de invocar um método da interface remota num objecto remoto, e é

executada usando a mesma sintaxe da invocação de um método de um objecto local.

Fundamentalmente, uma aplicação RMI é constituída por dois programas: um servidor e um cliente. Tipicamente, a aplicação servidor cria um conjunto de objectos acessíveis aos seus clientes e aguarda que algum deles invoque um método de um daqueles objectos remotos.

Existem portanto dois aspectos relevantes: em primeiro lugar, a necessidade dos clientes localizarem os objectos remotos; e em segundo, a necessidade de esconder os detalhes da comunicação entre os objectos remotos.

Ora, com o mecanismo de RMI do Java, aqueles detalhes da comunicação são completamente escondidos, uma vez que a comunicação remota é, aos olhos do programador, uma invocação de métodos tradicional.

Quanto à dificuldade de localização dos objectos, esta é resolvida através da associação de uma referência a cada um dos objectos remotos. Deste modo, para que um cliente possa invocar um método de um dos objectos basta que conheça essa referência. Para que seja possível obtê-la é necessário proceder de uma de duas maneiras: através do registo (pela parte do servidor) dessa referência usando o mecanismo de nomeação do RMI - `rmiregistry`, ou simplesmente pela passagem das referências dos objectos como parâmetros ou resultados das operações tradicionais.

Para poder usufruir destas facilidades basta portanto criar um servidor com capacidades de criar objectos remotos e torná-los acessíveis, o que é conseguido através dos métodos fornecidos pela classe `UnicastRemoteObject`.

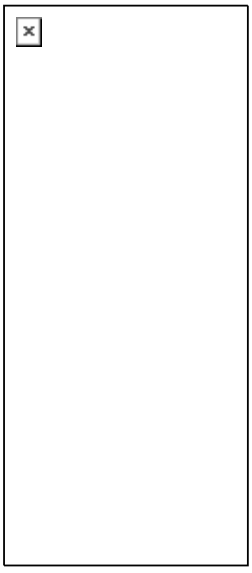

Figura 6 - Arquitectura geral do módulo de aquisição de conhecimento

No presente caso, foi pois necessário criar duas classes, uma do lado do servidor (`RmiServidor` que estende a classe `UnicastRemoteObject`) e outra do lado do cliente (`RmiCliente`), como se mostra na Figura 6.

Para além destas foi necessário definir uma interface (`RmiInterface`) que estende a classe `Remote`. Repare-se que cada um dos métodos desta interface lança um tipo específico de excepções – `java.rmi.RemoteException` – de forma a alertar a ocorrência de falhas na invocação, que se poderão dever, entre outras razões, a dificuldades de comunicação, problemas na passagem de parâmetros, ou simplesmente erros no protocolo.

Repare-se ainda que o servidor criado é uma implementação desta interface, e que no seu construtor é efectuado o seu registo. Por seu lado, o cliente possui um objecto do tipo `RmiInterface`, e através deste acede ao servidor.

6 - SEGURANÇA

"O estabelecimento de uma política de segurança advém da necessidade de proteger um bem que, no caso específico dos sistemas computacionais, corresponde à possibilidade de aceder ou modificar informação mantida pelo sistema. A segurança de um sistema informático baseia-se na definição global de uma estratégia ou política de segurança, que se apoia em diversos mecanismos de protecção à informação em causa e que terá de ter em conta o valor potencial dessa informação, as ameaças esperadas e os custos associados à implementação." [Marques 1998]

Existe uma necessidade de equilibrar os custos de implementar uma política de segurança (tal como se define), com os riscos ou ameaças que a informação manipulada pelo sistema corre. É essa tentativa de equilíbrio que se pretende atingir com a decisão de usar um mecanismo de segurança em apenas um dos módulos do sistema. Esta decisão fundamenta-se em dois aspectos principais: o número de acessos à base de dados e a não confidencialidade dos dados transmitidos.

Qualquer sistema que aplique técnicas de *data mining* efectua inúmeros acessos aos dados na perspectiva de descobrir as tais relações implícitas. Estes acessos devem ser o menos demorados possível, de forma a minimizar o tempo gasto pelos algoritmos. Ao cifrar os dados, aumentar-se-ia significativamente esse tempo. Para além disso, os dados transmitidos entre a base de dados e o *módulo de aquisição de conhecimento* não são personalizados, pelo que deixam de ser confidenciais. Por tudo isto, a implementação de uma política de segurança neste módulo, provocaria atrasos desnecessários no processamento dos dados, sem trazer benefícios relevantes.

Contudo, e quanto ao *módulo de gestão da consulta*, tanto as ameaças como o valor dos dados são maiores do que no módulo anterior: as *ameaças*, porque o módulo da consulta estará disponível para um número alargado de utilizadores, e os dados porque são confidenciais neste módulo.

Por tudo isto, e no seguimento das decisões de implementação tomadas, recorreu-se à atribuição de direitos de acesso e à utilização de mecanismos de autenticação e cifra, para implementar a estratégia de protecção dos dados manipulados pelo sistema.

Direitos de Acesso

O acesso ao módulo da consulta é feito por diversos funcionários do hospital, pelo que nem todos devem poder aceder a todos os dados registados. Por esta razão, foram criados três níveis de acesso: *acesso total* destinado aos médicos responsáveis pelo acompanhamento do doente em causa, *acesso técnico* para introdução de resultados dos exames ou registo/alteração de um doente, e *acesso para colaboradores* para tornar possível a transferência da informação entre a consulta propriamente dita e outros colaboradores (por exemplo outros profissionais de outros serviços do hospital).

Assim, quando é registado um novo utilizador, o serviço de gestão da consulta atribui-lhe os direitos de acesso que mais se adequam à sua função no hospital. Repare-se que apenas utilizadores com direito de acesso total, podem alterar os direitos de acesso de outro utilizador. Para sistematizar o que foi dito, apresenta-se a matriz de direitos de acesso usada:

Tabela 4 – Matriz de direitos de acesso

	AGENTES		
OBJECTOS	**Médicos**	**Técnicos**	**Colaboradores**
Dados Identificativos	Leitura / Escrita	Leitura / Escrita	Leitura
Historial Clínico	Leitura / Escrita	X	X
Observações Oftalmológicas	Leitura / Escrita	X	X
Avaliação Funcional	Leitura / Escrita	Leitura / Escrita	X
Síntese	Leitura / Escrita	Leitura / Escrita	Leitura
Utilizadores	Leitura / Escrita	X	X

A forma de armazenamento utilizada foi a associação dos direitos de acesso a cada um dos objectos, ou seja, a utilização de uma **Lista de Controlo de Acessos** (**ACL**).

"Para implementar a política de controlo de acessos é necessário dispor de duas operações de base: Autenticação – operação de validação da identidade do agente (...). Autorização – operação que valida os direitos do agente sobre o objecto antes da execução da operação." [Marques 1998]

Enquanto que a **autenticação** dos clientes é concretizada através da verificação da *password* do utilizador, a **autorização** é realizada pelo servidor com base nos direitos de acesso atribuídos ao

utilizador, aquando da sua entrada.

Segurança da Comunicação

Como já foi referido, a comunicação entre o cliente e o serviço de gestão da consulta efectua-se recorrendo ao protocolo HTTP sobre SSL (*Secure Socket Layer*), um protocolo criado pela Netscape para autenticações e cifras gerais sobre redes TCP/IP, o que permite garantir a segurança das comunicações. [Freier 1996]

O *SSL* é um nível de protocolo que deve ser colocado no patamar entre o nível da rede (por exemplo, TCP/IP) e o nível da aplicação (por exemplo HTTP). O *SSL* fornece comunicação segura entre o cliente e o servidor ao permitir a utilização de autenticação mútua, assinaturas digitais e cifra. As sessões de *SSL* entre o cliente e o servidor são estabelecidas seguindo a sequência definida pelo protocolo *handshaking*, podendo esta sequência variar dependendo de o servidor estar configurado para fornecer e/ou exigir um certificado.

De um modo geral o protocolo de *handshaking* segue os passos apresentados na Figura 7:

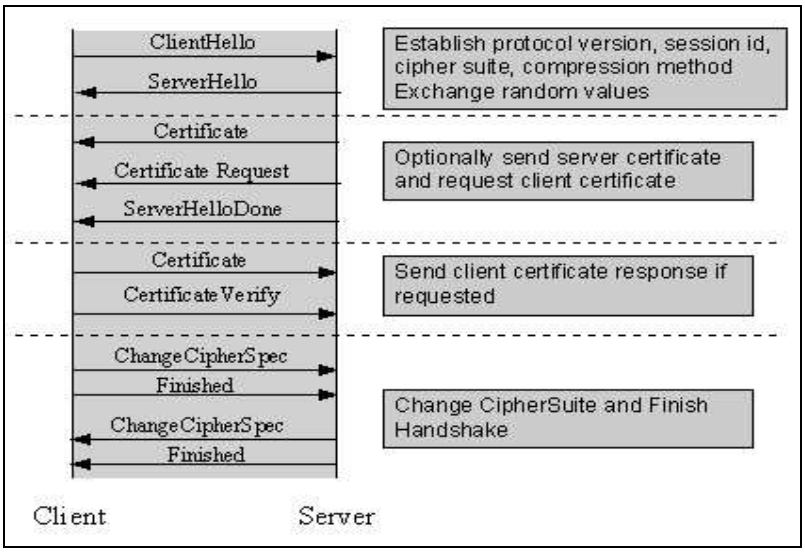

Figura 7 - Esquema simplificado do protocolo *handshaking* (retirado de [MOS_SSL 1999])

No sistema implementado, apenas o servidor é portador de um certificado, pelo que a 4ª, 6ª e 7ª mensagem não são trocadas, ou seja, não é enviado o pedido do certificado ao cliente, e consequentemente este não é enviado nem verificado.

Repare-se que o primeiro passo do protocolo, a negociação do tipo de cifra a usar na comunicação, permite escolher um tipo de cifra suportado tanto pelo servidor como pelo cliente. A versão utilizada, SSL3.0, usa o algoritmo convencional de criptografia – criptografia simétrica – e

permite 9 opções. De entre estas foi usada a técnica de *RC4-40* (chave secreta de 40 bits). É ainda de referir que foi necessário recorrer à criação de uma autoridade de certificação para produzir o **Certificado do Servidor**, produzindo-se desta forma dois certificados. É a partir deste último que se efectua a cifra de todas as transmissões.

Capítulo V - SISTEMA DE AQUISIÇÃO DE CONHECIMENTO

Com o presente capítulo termina-se a descrição do sistema desenvolvido. A primeira secção – Tecnologia de Sistemas de Aquisição de Conhecimento – apresenta um pequeno resumo sobre esta tecnologia, focando cada uma das suas etapas. Na segunda secção – Definição do Problema – faz-se uma descrição detalhada do problema a tratar, começando-se por uma descrição da estrutura e tipo dos dados, assim como a quantidade e a qualidade da informação registada. Termina-se com a apresentação do objectivo a atingir – tipos de regras que se pretende descobrir. Por último, na secção Arquitectura do Sistema de Aquisição de Conhecimento descreve-se o sistema desenvolvido, destacando-se cada um dos seus componentes.

1 - TECNOLOGIA DE SISTEMAS DE AQUISIÇÃO DE CONHECIMENTO

O sucesso dos *sistemas de apoio à decisão* deve-se essencialmente à sua capacidade de extracção de informação a partir do aglomerado de dados armazenados pelos sistemas transaccionais da organização. A componente cuja responsabilidade exclusiva é a extracção de informação a partir dos dados, será designada por *sistema de aquisição de conhecimento.*

Empregam-se os termos aquisição ou extracção de conhecimento para descrever todo o processo de extracção de conhecimento a partir dos dados registados numa base de dados (*Knowledge Discovery from Databases - KDD*), "extracção esta não trivial de conhecimento implícito, previamente desconhecido e potencialmente útil, feita a partir dos dados registados" [Frawley 1992]

A este nível, existe uma distinção clara entre os termos *dados***,** *informação* e *conhecimento***.** "Se os dados são identificados como os factos registados, então a informação é o conjunto de padrões ou expectativas, que advêm desses dados. Pode-se definir conhecimento como a acumulação do conjunto de expectativas, e a sabedoria como o valor associado ao conhecimento." [Witten 2000]

A solução para a criação de sistemas de suporte à decisão passa portanto por acumular um conjunto de padrões que descrevam os dados conhecidos. Como proceder então para adquirir o conhecimento necessário?

No final da década de 70, descobriu-se que o conhecimento dos peritos em determinados domínios bem restritos, podia ser descrito por regras simples do tipo "Se ... então...", bastando portanto criar um sistema capaz de armazenar essas regras e seleccioná-las convenientemente para obter um sistema de verdadeiro apoio à decisão. Contudo, a aquisição destas regras dependia quase exclusivamente da capacidade do perito em exprimi-las, o que na maioria das situações se revelou extremamente difícil (este problema da aquisição de conhecimento ficou conhecido como o *knowledge bottleneck problem*).

A aquisição de conhecimento nos humanos é feita através do processo de aprendizagem. Do ponto de vista de um sistema, esta pode ser definida como as alterações do sistema, que lhe permitem refazer as mesmas tarefas mais eficaz e eficientemente no futuro. [Simon 1990] Por outro lado, do ponto de vista da matemática, a aprendizagem pode ser vista como a compressão de conjuntos de dados. [Adriaans 1996]

De acordo com as definições apresentadas, a aprendizagem torna-se possível ao basear-se na experiência e tendo em vista a compressão de conjuntos de dados. Ora, com o desenvolvimento das técnicas de bases de dados, a proliferação do registo de grandes quantidades de dados tornou-se uma realidade, e esses dados não são mais do que o conjunto de resultados registados sobre as experiências passadas. Deste modo, tentando comprimir as experiências registadas nas bases de dados, a aprendizagem artificial começou a produzir alguns resultados significativos.

O objectivo primordial dos sistemas de aquisição de conhecimento é portanto, a partir dos dados registados em bases de dados, obter descrições compactas da informação ali registada, como por exemplo as relações entre atributos ou a classificação de instâncias de entidades.

As aplicações deste tipo de sistema são extremamente variadas. Frases como

> *"Os homens de negócios / empresários querem descobrir se os dados acumulados podem ajudá-los a tomar melhores decisões; os engenheiros querem que os dados indiquem o que há a melhorar; e os cientistas pretendem fazer descobertas com base nos dados que têm."*
>
> *[Lu 1997, p. vii]*

caracterizam o estado actual da aplicação destes sistemas.

O processo da aquisição de conhecimento é composto por três etapas fundamentais: o pré-processamento, o *data mining* e o pós-processamento, sendo cada uma destas etapas constituídas por várias tarefas, como se pode ver na Figura 8.

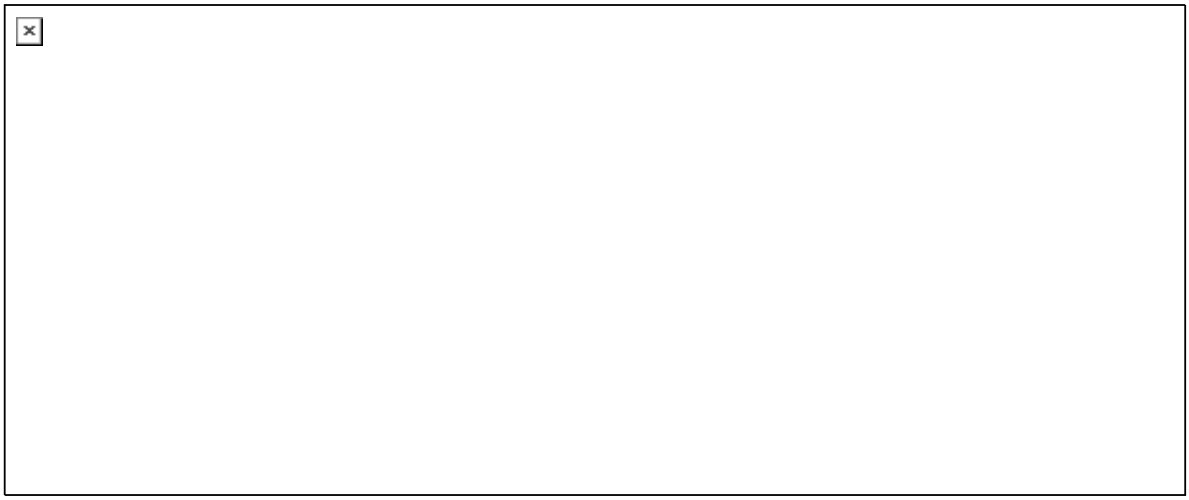

Figura 8 - Processo de descoberta de conhecimento a partir de bases de dados (adaptado de [Adriaans 1996, p.38])

Repare-se que nenhuma destas etapas é definitiva, sendo sempre possível refazer uma ou mais novamente. Segue-se agora uma descrição de cada uma destas etapas.

1.1 - Etapa de Pré-processamento

A etapa de pré-processamento consiste na aplicação de um conjunto de processos que melhoram o desempenho dos esquemas de *data mining* usados para a descoberta da informação. Estes processos constituem uma espécie de engenharia dos dados, ao reconstruir os dados de entrada de forma a adequarem-se ao esquema de aprendizagem escolhido.

Definição do Objectivo

O primeiro passo do longo processo de aquisição de conhecimento é a formulação dos requisitos de informação, associado a uma acção específica, isto é, o que se quer saber e o que se vai fazer com essa informação ou conhecimento.

Para além disso, é necessário estabelecer os critérios de avaliação a usar para identificar o sucesso do processo. E finalmente, é indispensável identificar a origem, ou seja a(s) base(s) de dados, onde os dados que irão suportar o processo de aquisição de conhecimento estão armazenados.

Selecção

Normalmente, os dados existem em bases de dados operacionais e é necessário recolhê-lhos para uma só base de dados centralizada e dedicada exclusivamente ao processo. Este tipo de base de dados é habitualmente designada por data warehouse.

Esta recolha pode tornar-se um processo complexo devido à situação actualmente existente nas organizações ao nível dos sistemas de informação. Muitas vezes os dados que se encontram nas diferentes bases de dados operacionais não são trivialmente relacionáveis, uma vez que é frequente a existência de diferentes formas de identificar a mesma entidade nas várias bases de dados operacionais. Outro aspecto relevante é a actualidade dos dados: nem todas as bases de dados contêm dados referentes à mesma janela temporal.

Depuração

O passo seguinte consiste na apreciação dos dados recolhidos, de modo a formar uma ideia / imagem sobre as suas potencialidades. Esta apreciação torna-se ainda mais difícil quando se trata de uma base de dados de grande dimensão. Nestes casos é habitual fazer esta apreciação em várias amostras aleatórias, e cruzar as impressões obtidas em cada uma delas.

Os aspectos a apreciar nos dados são essencialmente ao nível da sua qualidade, nomeadamente consistência e existência de valores omissos.

Apesar desta fase poder ser em parte processada automaticamente, não se deve esperar que a depuração dos dados seja total. De facto, a poluição das grandes bases de dados reveste-se de formas extremamente subtis e difíceis de identificar. Porém, de modo a identificar as potencialidades dos dados, é conveniente que parte desta fase seja efectuada manualmente.

Um aspecto importante a referenciar é o facto desta fase permitir identificar potenciais problemas na forma como os dados estão a ser recolhidos pelos sistemas transaccionais, isto porque muitas vezes a poluição deve-se essencialmente ao método pelo qual é efectuada a recolha. A correcção deste tipo de erros envolve futuras reengenharias do processo.

Enriquecimento

A fase de enriquecimento apresenta-se sob duas formas: na adição de novos atributos derivados de atributos já existentes e na introdução de dados de bases de dados exteriores à organização.

Se a primeira abordagem é trivialmente executada, o mesmo não se passa com a segunda. De facto os problemas enfrentados são semelhantes aos da fase de selecção, na recolha de dados das várias bases de dados operacionais.

Há no entanto um aspecto a ter em conta, e que se relaciona com o contexto jurídico envolvente. Em muitos países a venda / utilização de dados exteriores à organização não é permitida, pelo que esta abordagem não pode ser seguida.

Codificação

Por fim a fase de codificação. Esta consiste na formatação final dos dados para entrada do mecanismo de data mining. Nesta fase é necessário ter em conta toda a informação recolhida nas fases anteriores, em especial a estrutura dos dados e a existência de valores omissos.

Estrutura dos Dados

Em termos da estrutura de dados há dois aspectos a ter em atenção. Em primeiro lugar é necessário verificar se os tipos de atributos existentes são processados pelos mecanismos de *data mining* a usar na próxima etapa. Se isto não se verificar, existe a possibilidade de os pré-processar de forma a que este problema seja ultrapassado. Por exemplo, atributos do tipo numérico podem ser transformados de modo a tornarem-se nominais, através de um processo de *discretização*.

Um outro problema surge na presença de atributos textuais, em que a pessoa que introduz os dados tem oportunidade de escrever livremente. Este tipo de atributo não é processado por nenhum dos tradicionais esquemas de *data mining*. Uma solução possível passa pelo pré-processamento destes valores de modo a criar novos atributos nominais, capazes de traduzir o atributo textual.

Por outro lado, existem atributos extremamente específicos (como a data de nascimento ou a morada) que não trazem grandes vantagens, pelo que uma outra operação possível é a *tradução* destes atributos em atributos mais significativos para o requisito de informação previamente definido (um exemplo desta operação é a tradução da data de nascimento para a idade).

Outro aspecto importante é a estrutura das entidades registadas na base de dados. Se existir mais que um registo para uma só instância da entidade, estes registos devem poder ser processados conjuntamente de modo a descobrir as possíveis relações existentes entre esses dados. Este tipo de situação é encontrada quando se identificam dependências funcionais entre os atributos de duas ou mais entidades. De forma a permitir esta operação é necessário transformar os vários registos da mesma instância (normalmente representados como linhas de uma tabela) num só registo que englobe todos os dados (numa só linha de uma nova tabela). Esta operação é designada por *desnormalização* (tirar da forma normal). Uma outra manifestação desta operação é na junção de registos de tabelas diferentes, relacionadas entre si [Witten 2000].

Substituição de Valores Omissos

Um problema frequente nas bases de dados é a existência de grandes quantidades de atributos sem valores para cada uma das instâncias. Contudo, este é um factor extremamente importante no

sucesso da descoberta de informação.

Existem algumas abordagens a este problema: a exclusão dos registos com valores omissos, a substituição do valor omisso por um valor novo, designado *valor desconhecido*, e a substituição do valor omisso pelo *valor mais provável* (Quinlan sugere a construção de regras que prevejam o valor do atributo omisso, baseado nos valores dos restantes atributos do registo e na informação de classe [Quinlan 1986]).

Repare-se que qualquer uma destas abordagens tem inconvenientes e vantagens face às restantes. Por exemplo o facto de excluir alguns registos está a diminuir a amostra e portanto as potencialidades dos dados. É ainda de notar que muitas das vezes a existência de valores omissos evidenciam por si só alguns sintomas de problemas na recolha dos dados.

1.2 - Etapa de Data Mining

O termo data mining está associado à etapa de descoberta do conhecimento, propriamente dita. Existem dois critérios que diferenciam o *data mining* da aprendizagem artificial (*machine learning*), o primeiro diz respeito ao número e qualidade dos dados alvo e o segundo aos algoritmos usados. De facto o *data mining* aplica-se à extracção de conhecimento em bases de dados, que se traduzem na maioria das vezes em enormes quantidades de dados com alguns problemas de falta de qualidade, como descritos anteriormente. Pelo contrário, os dados usados pelos investigadores em aprendizagem artificial, agrupados em conjuntos de treino, são escolhidos de modo a evidenciarem apenas algumas determinadas características, que o algoritmo a testar deve tratar [Holsheimer 1994]. Quanto aos algoritmos, o que diferencia o *data mining* da aprendizagem é o facto dos algoritmos usados pelo primeiro deverem gozar da propriedade da escalabilidade, isto é, conhecidos os recursos do sistema (memória, velocidade de processamento, ...) o tempo de execução do algoritmo deve crescer linearmente com o tamanho do conjunto de dados. Grande parte dos algoritmos usados no domínio do *data mining* tiveram a sua origem na aprendizagem e sofreram (ou não) algumas alterações de modo a se tornarem escaláveis. [Ramakrishnan 2000]

A etapa de *data mining* pode ser resumida na Figura 9

Figura 9 – Etapa de *Data Mining* (adaptado de [Wu 1995, p. 21])

O mecanismo de aprendizagem ou de aquisição de conhecimento consiste num algoritmo ou combinação de algoritmos, que acede aos dados (já tratados) registados na base de dados e os compactam ou traduzem em informação. Ao longo das últimas décadas foram desenvolvidos vários tipos de algoritmos de aprendizagem – designados por *esquemas de aprendizagem*. Estes diferem sobretudo na forma como traduzem a informação descoberta e no processo como é feita essa descoberta. Apesar de algumas dificuldades na escolha do esquema a usar em cada problema específico, alguns esquemas são mais adequados a determinados tipos de problemas que outros.

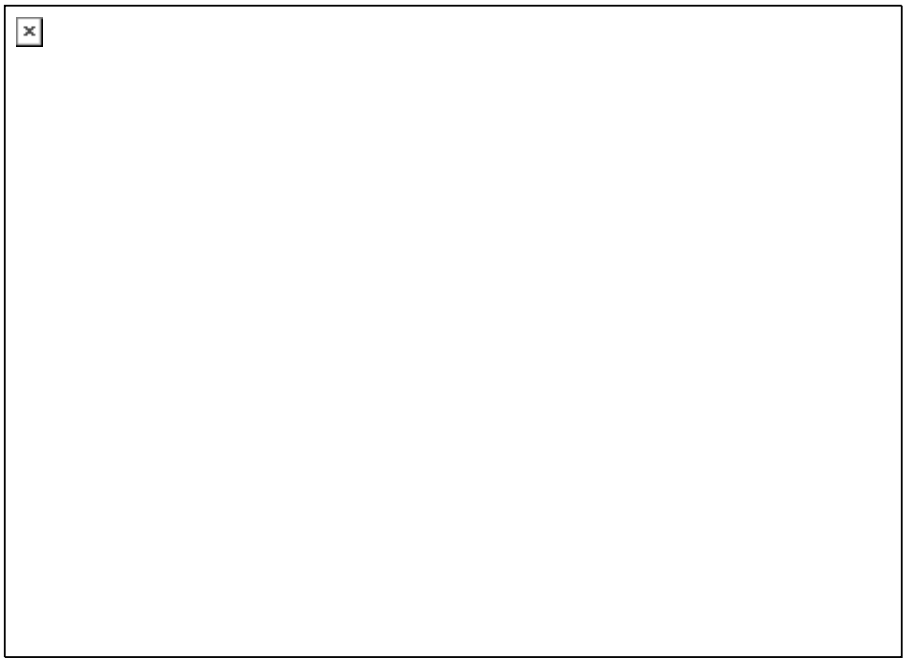

Figura 10 – Comparação entre esquemas de aprendizagem (adaptado de [Adriaans 1996, p. 87])

Podem identificar-se três tipos distintos de tarefas nos problemas de aquisição de conhecimento: as tradicionais tarefas de classificação, as tarefas de resolução de problemas e as tarefas de engenharia do conhecimento. Enquanto que as tarefas de classificação se baseiam na determinação de funções ou regras com base nos dados do conjunto de treino, a resolução de problemas envolve a descoberta de métodos de aplicação daquelas funções ou regras para resolver novos problemas [Shavlik 1990]. Por fim, a engenharia do conhecimento diz respeito ao processo de encontrar a representação formal de uma determinada porção de conhecimento, de modo a representá-lo num sistema baseado em conhecimento [Adriaans 1996]. A Figura 10 apresenta uma possível classificação para alguns dos esquemas de aprendizagem mais usados, em função destas tarefas.

Quando se encara a aquisição de conhecimento a partir de fontes externas, o que se pretende é analisar o conjunto de dados fornecido para produzir regras ou procedimentos gerais. Este tipo de

aprendizagem é designada por aprendizagem indutiva ou empírica, e pode ser realizado seguindo uma de duas estratégias, escolhida de acordo com a natureza dos dados em análise. Estas estratégias designam-se por aprendizagem supervisionada e não supervisionada, e distinguem-se fundamentalmente pela existência ou ausência da informação acerca da classe a que o registo pertence. Mais concretamente, na *aprendizagem supervisionada* os exemplos fornecidos ao mecanismo de aprendizagem (algoritmo) são da forma (x_i, y_i), e o objectivo deste mecanismo é descobrir a função *f* tal que $f(x_i)=y_i$. Mais ainda, a função *f* deve capturar os "padrões gerais" do conjunto de treino, de modo a que *f* possa ser aplicado para predizer o valor de *y* para elementos *x* ainda não analisados. [Shavlik 1990] Exemplos deste tipo de aprendizagem são a classificação e a predição numérica. Na primeira, o esquema de aprendizagem recebe um conjunto de exemplos, a partir dos quais deve ser capaz de aprender uma forma de classificar exemplos não conhecidos. Na predição numérica, o resultado a prever não é uma classe discreta mas sim uma quantidade numérica. [Witten 2000] Na *aprendizagem não supervisionada* os dados de treino são fornecidos e a tarefa do mecanismo de aprendizagem consiste na procura de alguma regularidade nesses dados, sem ter qualquer informação referente à existência de classes. [Shavlik 1990]

A seguir apresentam-se sucintamente os esquemas de aprendizagem mais relevantes, que podem compor o mecanismo de aquisição de conhecimento da etapa de *data mining*.

Aprendizagem Simbólica

Árvores de Decisão

As árvores de decisão são um dos métodos mais simples e bem sucedidos da aprendizagem. Uma Árvore de Decisão é uma representação de um conjunto de regras de classificação, e classificam as instâncias ordenando-as de acordo com a árvore, desde o nó raiz até a algum dos nós terminais (folhas), que fornece a classificação para a instância. Cada nó da árvore, especifica um teste para algum dos atributos da instância (variáveis), e cada ramo descendente desse nó corresponde a um dos valores possíveis para esse atributo. Uma instância é classificada começando por testar o atributo especificado pelo nó raiz, e depois seguindo o ramo correspondente ao valor do atributo na instância. Este processo é então repetido para a sub-árvore com raiz no novo nó. De um modo geral, uma árvore de decisão representa uma disjunção de conjunções de restrições no valor dos atributos. Cada caminho desde o nó raiz até uma folha corresponde a uma conjunção de testes de atributos e a própria árvore a uma disjunção de conjunções. [Mitchell 1997]

O objectivo é, portanto, gerar a árvore que melhor se adeque ao problema, ou seja, que melhor

classifique as instâncias do domínio considerado. Para tal, têm sido estudados vários algoritmos para seleccionar os atributos para os nós da árvore. Em *data mining* é habitual construir a árvore a partir de um subconjunto dos dados existentes na base de dados, e avaliar o desempenho da árvore no conjunto de todos os dados. Caso a árvore não classifique correctamente todos os casos, as excepções são adicionadas ao conjunto de treino e o processo é repetido. [Chen 1996]

As árvores de decisão são um modelo adequado para problemas de n-dimensões, tratando-se normalmente do método preditivo não linear mais rápido. Outra vantagem deste método é a expressividade do modelo de representação de conhecimento usado. [Weiss 1998]

Aprendizagem Relacional ou Sistemas de Programação Lógica Indutiva

Enquanto que as árvores de decisão usam uma representação restrita à lógica proposicional, a aprendizagem relacional usa a lógica de primeira ordem. Isto confere-lhe uma maior flexibilidade, permitindo a representação de objectos com estruturas complexas, assim como a representação de relações entre esses objectos ou entre as suas componentes. Com base num menor número de exemplos, conseguem-se especificar algumas relações, que são então generalizadas para induzir uma definição lógica para as relações. Como estas definições podem ser recursivas, podem ser expressas através do formalismo das bases de dados dedutivas.

Os sistemas de programação lógica indutiva podem ser classificados segundo várias dimensões de entre as quais se destacam as possibilidades de: aprender apenas um ou mais conceitos, verificar a validade das generalizações efectuadas durante o processo e integrar uma hipótese inicial, que é então revista.

A tarefa destes sistemas é aprender as definições lógicas completas e consistentes de vários predicados, usando a linguagem de 1ª ordem, partindo de um conjunto de exemplos de treino e do conhecimento do domínio existente. Uma outra aplicação deste tipo de sistemas é a reestruturação interactiva de bases de dados por meio de indução. A ideia principal é usar a indução para identificar as restrições de integridade de uma base de dados válidas, e depois usar essas restrições para reestruturar a base de dados.

Os sistemas de programação lógica indutiva preocupam-se com a indução de regras de primeira ordem, sob a forma de teorias clausais ou programas em lógica. Este processo ocorre no contexto das bases de dados dedutivas, e pode ser usado para descobrir padrões que envolvam várias relações. [Dzeroski 1996]

Na medida em que estes sistemas permitem a identificação e manipulação da própria estrutura

dos dados, são um mecanismo adequado às tarefas de engenharia do conhecimento.

Derivação de Dependências

A modelação de dependências consiste na procura de um modelo que descreva as dependências relevantes entre as variáveis. Existem modelos de dependências a dois níveis: ao nível estrutural e ao nível quantitativo. Ao nível estrutural, os modelos especificam (normalmente através de uma forma gráfica) quais as variáveis que são localmente dependentes entre si. Ao nível quantitativo, os modelos especificam a intensidade das dependências, fazendo uso de uma escala numérica. [Fayyad 1996]

Na derivação de dependências, o objectivo é portanto identificar as relações entre os vários atributos de uma entidade (variáveis), pelo que a informação da classe a que pertence cada entidade não é relevante para o objectivo. Neste contexto, está-se perante uma aprendizagem do tipo não supervisionado.

Regras de Associação

As regras de associação são uma forma de representar as dependências entre os atributos, ao nível quantitativo. Sejam $\mathbf{I} = \{i_1, i_2, ..., i_m\}$ um conjunto de literais (chamados itens), **D** um conjunto de transações, em que cada transacção **T** é um conjunto de itens, tal que $\mathbf{T} \subseteq \mathbf{I}$. Diz-se que **T** contém X, um conjunto de itens de **I**, se $X \subseteq \mathbf{I}$. Uma Regra de Associação é uma implicação da forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$ e $X \cap Y = \emptyset$. A regra $X \Rightarrow Y$ é válida no conjunto de transacções **D**, com confiança ***c***, se ***c***% das transações de **D** que contêm X, também contêm Y. A mesma regra tem suporte ***s*** no conjunto de transacções **D**, se ***s***% das transacções de **D** contêm $X \cup Y$. [Agrawal 1994]

O objectivo da determinação das regras de associação é portanto gerar todas as regras de associação que tenham suporte e confiança maior que dois valores pré-determinados (suporte e confiança mínimos).

Repare-se que os literais ou itens correspondem a pares atributo–valor, característicos das entidades registadas na base de dados. As transações correspondem às relações existentes entre as várias entidades. Um caso particular das regras de associação é permitir apenas que o consequente da implicação seja apenas um literal, o que pode ser encarado como uma forma de classificação. Ao descobrir alguns padrões de regularidades entre os dados registados, as regras de associação são um mecanismo adequado às tarefas de engenharia do conhecimento.

Redes de Bayes

As Redes de Bayes, são um exemplo de mecanismos de modelação de dependências ao nível estrutural. Estas redes consistem em grafos que descrevem as relações causais entre as variáveis, representando as variáveis através de nós, e por arcos as relações causais entre elas. Geralmente uma rede de Bayes representa a função de distribuição de probabilidade conjunta do domínio, à custa da especificação de um conjunto de asserções de independência condicional entre as variáveis e das probabilidades condicionadas locais. [Mitchell 1997]

O primeiro passo na construção de uma rede de Bayes é decidir quais as variáveis a modelar; em segundo lugar constrói-se o grafo acíclico capaz de codificar as asserções de independência condicional entre as variáveis, e por fim, determinam-se as distribuições de probabilidades locais.

Uma das vantagens das Redes de Bayes é a facilidade de introdução de conhecimento sobre o domínio, normalmente expresso pelos peritos, para melhorar o processo de aquisição de conhecimento. Para tal, codifica-se esse conhecimento tal como se faz na construção de sistemas periciais probabilísticos, e em seguida usam-se os dados registados na base de dados para actualizar aquele conhecimento, criando uma ou mais novas redes. O resultado traduz-se, assim, por um refinamento do conhecimento inicial expresso pelo perito, e por vezes a identificação de novas distinções e relações. Para além desta vantagem, outra característica que justifica a utilização das redes de Bayes, é que a interpretação do conhecimento representado (grafo) é realizada com bastante facilidade. [Heckerman 1996]

Clustering

Os termos *clustering* e aprendizagem não supervisionada são muitas vezes confundidos, porém o *clustering* é apenas um caso daquele tipo de aprendizagem. A tarefa de *clustering* é baseada numa metodologia do tipo “dividir para conquistar”, e tem como linha orientadora a identificação de um número finito de partições de um conjunto de dados. Ao decompor um grande sistema em componentes mais pequenas, simplifica a sua modelação e implementação.

O objectivo dos algoritmos de *clustering* é criar uma partição do conjunto de dados, em que os dados semelhantes pertencem ao mesmo subconjunto, e os diferentes pertencem a subconjuntos diferentes – cada um destes subconjuntos designa-se *cluster*. A semelhança entre os dados é medida computacionalmente através de uma *função de distância*. Esta função recebe dois elementos e devolve o valor correspondente à “semelhança” entre os dois elementos. Cada aplicação tem uma noção diferente de semelhança, pelo que não existe uma função de distância adequada a todos os domínios.

[Ramakrishnan 2000]

Encarada como uma tarefa de *data mining*, o *clustering* pode ser visto como uma forma de simplificar o registo e comunicação: em vez de memorizar todos os elementos, é mais fácil memorizar apenas o conceito envolvido, e em seguida nomear as excepções. Repare-se que se trata de uma forma de compressão dos dados, com perda de alguma informação mas com ganho na economia da representação. [Chen 1996]

Existem várias abordagens para representar os *clusters*. Uma das formas mais básicas de o fazer é representar um *cluster* através do registo de todos os elementos desse *cluster*. Esta abordagem torna a admissão de novos elementos a um dos *clusters* uma tarefa pouco eficiente, pois é necessário medir a semelhança do novo elemento com todos os elementos de cada um dos *clusters*. Uma abordagem mais económica consiste em representar o *cluster* por um elemento representativo, que pode ou não pertencer ao conjunto de treino. Deste modo, a admissão de novos elementos requer apenas a comparação com o representante de cada um dos *clusters*, sendo possível prever o valor dos atributos dos elementos de um *cluster* com base no valor dos atributos do representante desse *cluster*. Uma terceira abordagem consiste em representar o *cluster* como uma função de distribuição de probabilidades sobre o espaço dos valores possíveis dos atributos. Desta forma os novos exemplos são admitidos no *cluster* onde assumem maior probabilidade (determinada de acordo com a regra de Bayes). Por último, uma quarta abordagem consiste na definição das condições necessárias e suficientes para admitir elementos num *cluster*. O objectivo é, portanto, determinar aquelas condições da forma mais simples possível. Novos exemplos são admitidos num *cluster* se e só se satisfizerem todas as condições referidas, e os valores dos atributos dos elementos de um *cluster* são determinados directamente pelas condições.

A abordagem de *clustering* ao problema da extracção de conhecimento em bases de dados grandes, traduz-se muitas vezes na procura de *clusters* numa amostra dos dados da base de dados. A ideia base é que se a amostra for estatisticamente representativa, então representa correctamente todo o conjunto de dados, e consequentemente os representantes de cada *cluster* serão semelhantes aos que seriam seleccionados com base em todos os dados da base de dados. Usando técnicas como esta ou recorrendo a representações económicas dos *clusters*, este mecanismo de aprendizagem torna-se num poderoso mecanismo para a extracção de conhecimento em grandes bases de dados.

Ao separar os dados em vários *clusters*, este mecanismo executa uma classificação dos dados, pelo que se torna numa ferramenta adequada para tarefas de classificação. Contudo, esta separação é feita à custa da análise e abstracção das várias características dos dados, pelo que se aproxima das

soluções para problemas de engenharia do conhecimento.

Redes Neuronais

As Redes Neuronais artificiais, também conhecidas como modelos conexionistas, são redes densamente interligadas de unidades computacionais simples, designadas neurónios. As entradas da rede consistem num conjunto de N valores numéricos, assim como as suas saídas. Um neurónio calcula a diferença da soma ponderada das suas entradas com um *treshold*, e determina o valor de uma função não linear (por exemplo a sigmóide) para esse valor. As redes neuronais são criadas a partir da ligação da saída de um neurónio à entrada de um ou mais outros. Uma das possíveis topologias das redes neuronais é o perceptrão multi-camada, em que a saída de cada neurónio apenas é usada como entrada de unidades que se encontrem numa camada posterior. Este tipo de redes são especialmente úteis para tarefas de classificação, atribuindo um valor alto à saída que representa a classe referente aos valores de entrada, e um valor mais baixo às saídas referentes às restantes classes.

O objectivo é determinar os pesos associados às entradas de cada unidade computacional, e o treino é efectuado fornecendo um exemplo de cada vez, e alterando os valores dos pesos de acordo com a classificação do exemplo, pelo que se trata de um mecanismo de aprendizagem supervisionada.

Existem algumas desvantagens na utilização de redes neuronais em *data mining*, a primeira das quais diz respeito à baixa velocidade do processo de aprendizagem quando comparado com os mecanismos de aprendizagem simbólica. Uma segunda desvantagem é que a informação gerada pelas redes é representada sob a forma de uma rede, traduzindo-se na sua topologia e nos valores de cada um dos seus pesos, em vez de explicitamente representada sob a forma de regras ou padrões conceptuais. Por último, é difícil, embora possível, incorporar conhecimento do domínio ou interacção do utilizador durante o processo de aprendizagem. [Holsheimer 1994]

Algoritmos Genéticos

Os algoritmos genéticos pertencem a uma família de modelos computacionais inspirados na teoria da evolução. Estes algoritmos codificam uma potencial solução para um problema específico sob a forma de uma única estrutura de dados semelhante a um cromossoma. Geralmente os algoritmos genéticos são encarados como um método para optimizar funções.

A abordagem consiste, essencialmente, em usar um conjunto de descrições candidatas (tipicamente aleatória) – designado população inicial – e melhorar a sua qualidade gradualmente, ao gerar novas descrições a partir das melhores componentes das descrições iniciais. Estas recombinações de membros de uma população terminam quando as descrições têm uma qualidade suficiente, ou não

se verifiquem melhorias significativas no mérito da população.

A primeira suposição que se faz é que os parâmetros do problema são passíveis de ser representados através de *cadeias* de *bits*. Estas *cadeias* contêm uma *sub-cadeia* para cada um dos atributos, e nesta *sub-cadeia* cada elemento representa um valor do domínio do atributo, que assume o valor 1 caso seja esse o valor actual para o atributo, e 0 caso contrário.

Para além da codificação da população é necessário determinar a função de avaliação, o que envolve muitas vezes a criação de uma simulação. Para cada elemento da população, a função de avaliação é calculada com base na sua acuidade de classificação (isto é, no rácio de exemplos positivos e negativos cobertos) e na sua generalidade (isto é, o número de exemplos cobertos pela descrição em causa relativamente ao número de exemplos cobertos pela população total).

Uma das utilizações mais frequentes dos algoritmos genéticos é no processamento de problemas não lineares, o que implica normalmente que não é possível tratar cada parâmetro como uma variável independente, de forma isolada das restantes. Uma das principais vantagens dos modelos paralelos, como é o caso dos algoritmos genéticos, é a riqueza do seu método de computação: pequenas e simples alterações nos algoritmos resultam muitas vezes em surpreendentes comportamentos emergentes.

Tal como nas redes neuronais é difícil a incorporação de conhecimento externo durante o processo de aprendizagem. Por outro lado, também o número de avaliações necessárias, para atingir uma população com boas soluções para o problema, é elevado. [Holsheimer 1994]

Assim, a utilização de algoritmos genéticos é mais adequada para aquisição de conhecimento em aplicações onde existe pouco conhecimento do domínio, e na presença de bases de dados de menores dimensões (bases de dados com poucos atributos). Repare-se que a função de avaliação da população necessita da informação referente às classes a que pertencem os exemplos de treino, pelo que estamos uma vez mais na presença de aprendizagem supervisionada.

1.3 - Etapa de Pós-processamento

Tal como a etapa de pré-processamento, esta etapa constitui uma espécie de engenharia dos dados, desta vez processando os dados de saída em vez dos de entrada. O objectivo é, portanto, construir um modelo de saída para tornar os resultados dos esquemas de *data mining* mais eficientes, ou seja, para tornar os resultados mais compreensíveis para os seus utilizadores finais.

Existem dois aspectos que merecem atenção nesta fase: a tradução da informação encontrada para uma forma perceptível ao utilizador do sistema, e a verificação da qualidade dessa informação.

Esta verificação pode ser encarada como um passo da etapa de *data mining,* mais concretamente a fase de Avaliação do Desempenho do Mecanismo de Aprendizagem. Inclui-se esta fase na etapa de pós-processamento uma vez que a determinação dos valores de confiança e suporte são dados extremamente importantes para a validação das regras encontradas e portanto devem ser fornecidos ao utilizador final.

2 - DEFINIÇÃO DO PROBLEMA

2.1 - Análise dos Dados

Os dados objecto do presente estudo, caracterizam-se por estarem distribuídos por quatro esquemas distintos: *esquema de identificação, esquema de diagnóstico, esquema de dados das funções visuais* e *esquema de dados da visão funcional***,** como descrito anteriormente.

Por esquema de identificação entende-se o conjunto de dados referentes à identificação do doente, assim como do seu agregado familiar; por esquema de diagnóstico o conjunto de dados que definem o historial clínico do doente; por esquema de dados das funções visuais classificam-se todos os dados referentes às observações oftalmológicas (exames oftalmológicos); finalmente por dados da visão funcional designam-se todos os dados relacionados com as capacidades efectivas do doente, recolhidas sob a forma de questionários. Tal como apresentado na Tabela 3.

Várias considerações há ainda a fazer acerca da base de dados. Em primeiro lugar, há a destacar um número extremamente baixo de registos (o número de doentes da consulta é à data 170), de entre os quais se identificaram cerca de 5 registos "fantasma" (registos que não têm quaisquer dados à excepção do nome e números de identificação do doente). Assim, o número de registos para qualquer outra entidade não vai além de 165.

Um outro aspecto a registar é a quase inexistência de registos ao nível de dados das funções visuais. Quanto aos dados da visão funcional, foram detectados alguns erros no registo dos dados, essencialmente relacionados com atributos do tipo lógico, em que um valor omisso é codificado como o valor falso. Para além destas particularidades há ainda a registar o grande número de atributos por cada entidade, alguns dos quais do tipo texto. Por fim, é de destacar a grande quantidade de valores omissos existentes entre os dados.

No Anexo B faz-se uma pequena análise do número e qualidade dos dados disponíveis para o processo de aquisição de conhecimento.

2.2 - Descrição dos Objectivos

Do ponto de vista da aquisição de conhecimento, pode-se identificar o principal objectivo como a descoberta das relações existentes entre os dados registados na base de dados.

Porém, e dada a distribuição dos dados em quatro esquemas distintos, é trivial a classificação das relações que podem ser detectadas entre os dados. Naturalmente, podem ser relações entre atributos do mesmo esquema, ou entre atributos de esquemas diferentes. Uma outra possibilidade passa pela descoberta das relações existentes não dentro de um esquema, mas apenas numa tabela. Neste caso surgem duas possibilidades: relações numa tabela tal como existe na base de dados, ou numa tabela sujeita à operação de desnormalização, como descrita na secção 1.1 - do presente capítulo.

Pode-se, portanto classificar as potenciais regras em duas classes genéricas: relações intra esquemas e relações inter esquemas. De entre as relações intra esquemas tem-se: relações numa tabela, relações numa tabela desnormalizada, ou relações entre várias tabelas de um mesmo esquema.

Esta distinção é feita porque ao nível da informação, estas diferentes relações emitem diferentes contribuições.

Tabela 5 - Contribuição das classes de relações

Contribuição de cada classe de relações			
→	**De Diagnóstico**	**Funções Visuais**	**Visão Funcional**
De Diagnóstico	Relacionar patologias e tratamentos	Seleccionar exames oftalmológicos Diagnosticar patologias	Prever capacidades efectivas
Funções Visuais		Prever resultados dos exames Estabelecer equivalências entre exames	Estabelecer impacto das funções visuais na visão funcional
Visão Funcional			Optimizar questionários

Note-se, que algumas das relações são principalmente de interesse médico e que outras são essencialmente de interesse para a optimização do sistema.

3 - Arquitectura do Sistema de Aquisição de Conhecimento

De um modo geral o sistema de aquisição de conhecimento, propriamente dito, é composto por três componentes distintas, nomeadamente: o cliente onde são efectuados os requisitos de informação, o servidor que dá resposta a estes requisitos, e a base de dados onde estão armazenados os dados.

Como se viu a comunicação entre estes componentes é efectuada através da ligação por JDBC para aceder à base de dados, e por invocação de métodos remotos (RMI) para estabelecer a ligação

entre o servidor e o cliente.

Ao contrário do que foi apresentado como solução habitualmente aceite, não foi criada nenhuma *data warehouse* para o sistema desenvolvido. Isto deve-se essencialmente ao reduzido número de dados registados e sobretudo à sua previsível evolução (é perfeitamente expectável que em condições normais, a população da consulta não cresça substancialmente). Assim, a base de dados usada pelo sistema transaccional desenvolvido, é o repositório dos dados a submeter ao *data mining*.

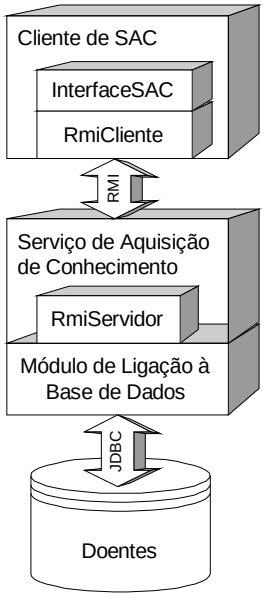

Figura 11 – Arquitectura geral do módulo de Aquisição de Conhecimento

Quanto ao cliente, como já vimos é apenas responsável pela interacção entre o serviço e o utilizador final deste módulo. Assim, por agora, limita-se quase exclusivamente à recepção de requisitos de informação e a apresentação das respostas obtidas. Com o futuro desenvolvimento de um sistema de apoio à decisão na sua plenitude, este componente poderá vir a ser alargado de forma a acumular em si as responsabilidades de seleccionar o conjunto de regras que se adeque ao problema formulado no pedido do utilizador, e não apenas na sua apresentação.

Finalmente é responsabilidade do servidor dar resposta aos pedidos efectuados, através do recurso aos mecanismos de aquisição de conhecimento propriamente ditos.

A criação deste serviço foi feita tendo em conta as tarefas do processo de aquisição de conhecimento a partir de bases de dados. Para tal, disponibilizaram-se algumas daquelas tarefas através deste serviço, nomeadamente a selecção dos dados e a sua codificação. As restantes tarefas foram implementadas de forma a serem feitas automaticamente sem intervenção do utilizador. Entre elas

encontra-se a tarefa de enriquecimento pela criação de novos atributos. Assim, ao utilizador é dada a possibilidade de seleccionar os dados a analisar, pré-processá-los e "miná-los". Para que isto fosse possível optou-se por estruturar o serviço de forma a facilitar quer a sua implementação, quer a sua utilização, através da sua modelação recorrendo à metodologia orientada a objectos.

O funcionamento do serviço pode ser descrito como o ciclo composto pelos passos de selecção dos dados a analisar, seguido da aplicação de um mecanismo de pré-processamento (opcional), e a subsequente aplicação de um mecanismo de *data mining*, na tentativa de descoberta das regras que lhe estão na base.

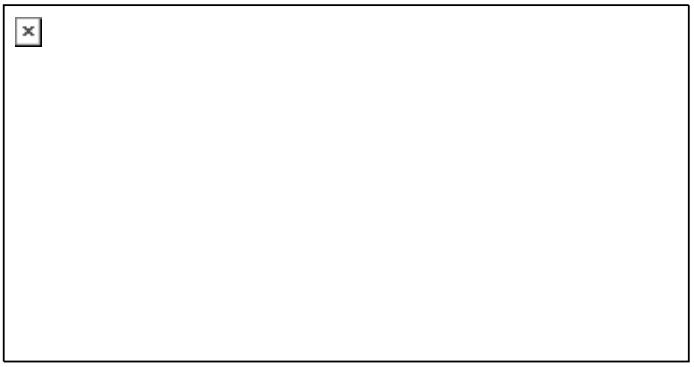

Figura 12 – Ciclo de Funcionamento do Serviço de Aquisição de Conhecimento

Um aspecto fundamental da criação deste módulo foi a utilização do pacote de software *Waikato Environment Knowledge Analysis* (WEKA), programado em `Java` e desenvolvido por **Ian H. Witten** e **Eibe Frank**. Este pacote de software contém implementações de vários métodos de aprendizagem, assim como um conjunto de ferramentas que permitem a sua fácil utilização [Witten 2000].

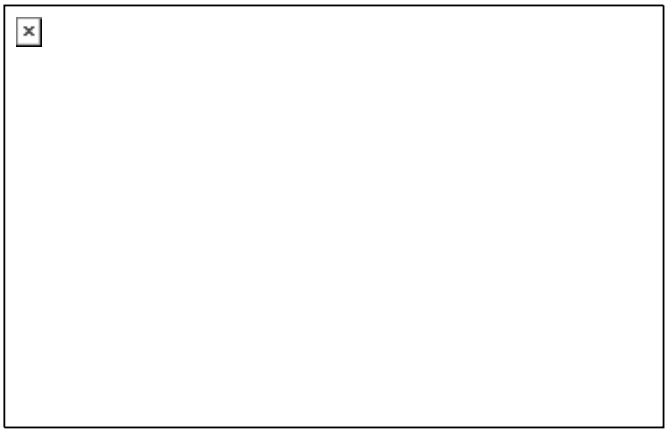

Figura 13 - Estrutura do Serviço de Aquisição de Conhecimento

Assim o serviço é implementado através de uma classe que permite a comunicação por RMI

(`RmiServidor`), uma classe capaz de armazenar os dados seleccionados (`Tabela`), um método de aprendizagem (`Apriori`) e algumas instâncias de classes derivadas da classe `Filter`.

3.1 - Acesso ao repositório de dados

O acesso ao repositório dos dados é feito, como já foi dito, através de uma ligação à base de dados, através da qual se recolhem todas as instâncias que respeitam os requisitos inicialmente especificados pelo utilizador.

Estes dados são então guardados em memória numa instância da classe `Tabela`, que entre os atributos possui um nome, um vector de cadeias de caracteres que representam a(s) sua(s) chave(s), assim como uma instância da classe `Instances`, existente no *package* `weka.core`. Esta classe, por sua vez, contém um conjunto de instâncias da classe `Instance`, também pertencente ao *package* `weka.core`. Uma `Instance` possui um conjunto de pares atributo – valor, entre outras coisas. Estas classes fornecem os meios de registar em memória os dados lidos directamente da base de dados, consistindo um objecto `Instances` no conjunto de dados lido, e um objecto `Instance` numa instância particular de uma qualquer entidade.

Ao manter em memória os dados seleccionados torna-se possível a sua manipulação sem ser necessário voltar a aceder à base de dados, o que é viável neste caso, dado a reduzida quantidade de dados registados.

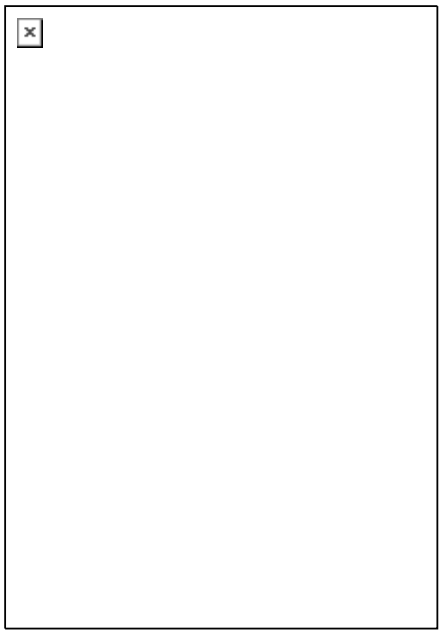

Figura 14 - Classe Tabela

3.2 - Mecanismos de pré-processamento

Das várias possibilidades de mecanismos de pré-processamento descritos na tarefa de codificação, foram implementados apenas três deles: a desnormalização de tabelas, a substituição de valores omissos pelo valor "desconhecido" e o tratamento dos valores numéricos, pela sua discretização. Mecanismos como o de substituição de valores omissos pelo valor mais provável não se aplicam aos dados a analisar, uma vez que se está perante um processo de aprendizagem não supervisionada, pelo que o valor mais provável não pode ser estimado com base na classificação da instância, como sugerido em [Quinlan 1986]. Ao ser aplicado qualquer um destes mecanismos, cria-se um novo objecto Tabela, em tudo idêntico ao descrito anteriormente, mas cujo processo de criação tem algumas particularidades. Uma vez que tudo o que pode ser feito numa tabela pode ser feito numa tabela pré-processada é possível fazer-se uso de um dos mais ricos conceitos da modelação orientada a objectos: a herança.

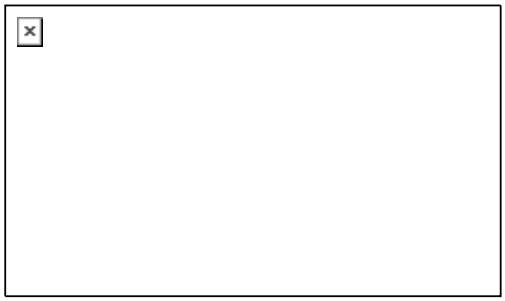

Figura 15 - Classes Derivadas da Classe Tabela

Tabela Pré-processada

Designa-se por `Tabela Pré-processada` uma tabela criada a partir da aplicação de um filtro a outra tabela já existente. Um Filtro é uma classe do *package* `weka.filters`, e tem como objectivo transformar um conjunto de instâncias (objecto `Instances`) noutro pela alteração/selecção de alguns dos seus atributos.

Foram usados dois filtros diferentes de modo a substituir valores omissos e tratar valores numéricos. No caso dos valores numéricos foi usado o filtro `DiscretizeFilter` existente no package `weka.filters`, e que possibilita a discretização de valores numéricos segundo o método de discretização não supervisionada através da partição do espectro de valores assumidos pelo atributo. Esta partição é inicialmente estabelecida automaticamente e optimizada através de um processo de validação cruzada (*cross-validation*) que estima a máxima verosimilhança dos dados. Quanto ao filtro de substituição (`ReplaceMissingValuesWithUnknownFilter`) foi criado a partir da classe abstracta

`Filter`, e baseia o seu funcionamento na procura e substituição dos valores omissos pelo novo valor "`desconhecido`".

Tabela Desnormalizada

Uma `Tabela Desnormalizada` é também originada pelo pré-processamento de uma tabela já existente, mas não pela aplicação de um filtro que transforma as suas instâncias.

Foram desenvolvidos dois processos diferentes de desnormalização, um aplicado a uma tabela, e o outro aplicável a duas ou mais tabelas.

O processo de desnormalização de uma tabela consiste na descoberta do número máximo de registos da mesma entidade na tabela (por exemplo o número máximo de tratamentos a que um doente foi submetido) e a replicação dos seus atributos tantas vezes quantas as ditadas por esse número.

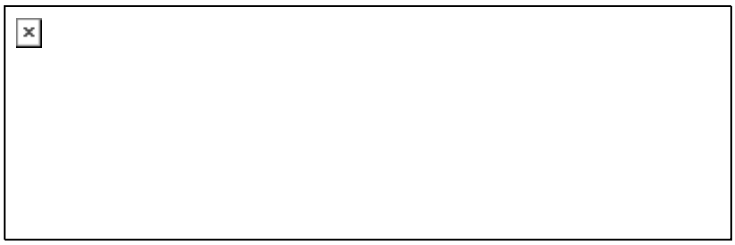

Figura 16 – Exemplo da Desnormalização de uma Tabela

A desnormalização de várias tabelas consiste na aglomeração dos atributos de mais do que uma tabela, numa espécie de super-tabela. Mais concretamente, este tipo de desnormalização só pode ser realizado quando existe uma relação entre elas, e consiste em encontrar as instâncias de cada uma das entidades (referentes ao mesmo doente, por exemplo) e juntá-las numa nova linha de uma nova tabela.

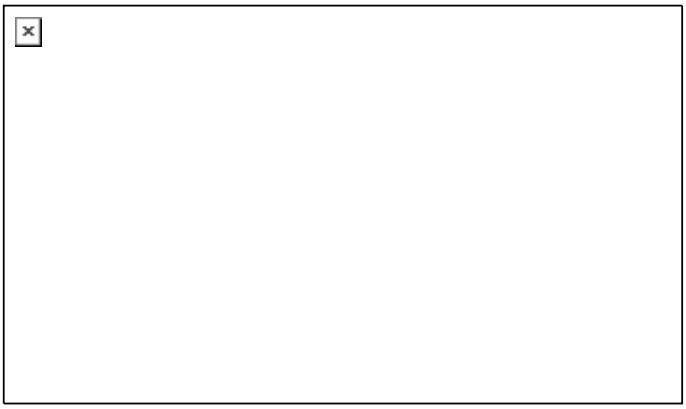

Figura 17 – Exemplo da Desnormalização de duas Tabelas

3.3 - Mecanismos de *data mining*

Estando na presença de um conjunto de dados com as características enunciadas na secção 2.1 - e dos objectivos estabelecidos na secção 2.2 -, a escolha do mecanismo de aquisição de conhecimento a usar ficou limitada. De facto, e como foi apresentado, uma abordagem ao problema com Redes Neuronais ou Algoritmos Genéticos não seria adequada, dada a opacidade do tipo de representação de conhecimento que usam e que muito dificilmente poderia ser traduzida de forma a tornar-se perceptível para os utilizadores finais. Por outro lado, a inexistência de uma classificação dos dados registados, tornam mais complexa a aplicação de mecanismos de aprendizagem supervisionada.

Dos tipos de aprendizagem não supervisionada apresentados restam a Derivação de Dependências – Regras de Associação – e o *Clustering*. Com o primeiro mecanismo consegue-se descobrir as relações "escondidas" entre os dados registados – o que vai ao encontro dos objectivos estabelecidos. Com o *clustering*, é possível agrupar os doentes registados em vários aglomerados, de acordo com a semelhança das suas características, o que está para além dos objectivos actuais.

Apriori

O algoritmo usado para descobrir as Regras de Associação foi o ***Apriori***, desenvolvido por Rakesh Agrawal e Ramakrishnan Srikant [Agrawal 1994]. Este algoritmo encara a descoberta de relações como um processo com duas etapas: a geração dos *conjuntos de itens* (*item sets*) com suporte acima do mínimo especificado, e a determinação das *regras* com grau de confiança acima do mínimo estabelecido, a partir de cada um destes conjuntos.

O passo de geração dos conjuntos de itens consiste na criação dos conjuntos com apenas um item com suporte acima do mínimo especificado, e a partir destes gerar conjuntos com dois elementos, a partir do emparelhamento de conjuntos de apenas um item. Repare-se que os conjuntos de dois itens devem relacionar dois atributos diferentes, uma vez que se relacionarem o mesmo atributo não são consistentes, pois em cada registo um atributo pode apenas ter um valor associado. A geração destes segundos conjuntos envolve nova passagem pelos dados de modo a determinar o suporte de cada um dos novos conjuntos de itens. Depois desta passagem os conjuntos de itens com suporte inferior ao estabelecido são excluídos, e os sobreviventes guardados.

O segundo passo do algoritmo pega em cada conjunto de itens e gera as regras a partir dele, verificando se estas regras têm grau de confiança acima do mínimo estabelecido.

Tabela 6 – Notação (adaptado de [Agrawal 1994])

k-itemset.	Um itemset com k items.
L_k	Conjunto de k-itemsets com suporte mínimo (large k-itemset). Cada membro deste conjunto tem dois campos: i) conjunto de itens e ii) contador de suporte.
C_k	Conjunto de k-itemsets candidates. Cada membro deste conjunto tem dois campos: i) conjunto de itens e ii) contador de suporte.
Apriori-gen	Função que recebe o conjunto de k-itemsets com suporte mínimo como argumento, e devolve o super-conjunto de conjuntos de todos os k-itemsets com suporte mínimo.

```
1)  L1 = {large 1-itemsets};
2)  for (k=2; Lk-1≠0; k++) do begin
3)   Ck = apriori-gen(Lk-1);
4)   forall transactions t∈D do begin
5)    Ct = subset(Ck, t);
6)    forall candidates c∈Ct do
7)      c.count++;
8)   end
9)   Lk = {c∈Ck | c.count ≥ minsup}
10) end
11) Answer = ∪k Lk ;
```

Figura 18 – Algoritmo Apriori [Agrawal 1994]

A implementação do algoritmo *Apriori* usado foi a fornecida no *package* `weka.associations`.

3.4 - Mecanismo de Pós-processamento

O mecanismo de pós-processamento limita-se à tradução das regras encontradas no passo anterior, e ao seu armazenamento em instâncias da classe `Regra`. Esta tradução é feita recorrendo às tabelas que contêm todos os valores possíveis para os atributos nominais.

Uma regra contém um antecedente, um consequente, um grau de confiança e um suporte. Tanto o antecedente como o consequente são apenas cadeias de caracteres que englobam conjunções de literais.

Considere-se a regra $X \Rightarrow Y$. Por suporte entende-se o número de instâncias encontradas em que X e Y se verificam, e por grau de confiança entende-se o rácio entre o suporte de X e o suporte de Y, isto é, o número de instâncias correctamente previstas como proporção do número de instâncias a que a regra se aplica.

Capítulo VI - ANÁLISE CRÍTICA DOS RESULTADOS OBTIDOS

No presente capítulo faz-se uma apresentação sumária das regras descobertas, fazendo algumas apreciações ao seu significado e à sua credibilidade em geral.

1 - DESCRIÇÃO DOS RESULTADOS OBTIDOS

Devido aos factores atrás descritos na secção de Análise dos Dados, apenas foi possível utilizar os dados referentes aos atributos dos esquemas de identificação e de diagnóstico.

Por tabela

Os resultados sem qualquer pré-processamento nas tabelas referentes aos atributos de diagnóstico foram os seguintes:

```
Tabela Historial Clínico
➔ familiares = Irrelevantes ==> prematuridade = false (conf->0.825, sup->0.196)
Tabela INCAPACIDADE
➔ evolução = Fásica ==> instalação = Progressiva (conf->0.818, sup->0.134)
➔ instalação = Súbita ==> evolução = Estável (conf->0.762, sup->0.238)
Tabela TRATAMENTO
```

Não foram encontradas regras para a tabela Tratamento

```
Tabela DIAGNÓSTICO
➔ descrição Patologia = DMRI ==> patologia = Adquirida (conf->1.0, sup->0.111)
```

Repare-se que a tabela Tratamento possui uma vasta quantidade de valores omissos especialmente no seu atributo descrição do Tratamento.

Por tabela desnormalizada

Ao aplicar a operação de desnormalização os resultados obtidos foram substancialmente melhores.

Apresentam-se em seguida algumas das regras encontradas a partir da desnormalização da tabela

diagnóstico.

```
➔ patologia#1 = Congénita ⇨ patologia = Congénita (conf->0.867, sup->0.158)
➔ patologia#1 = Adquirida ⇨ patologia = Adquirida (conf->0.765, sup->0.108)
```

Estas duas regras traduzem o facto de que na maioria dos casos, um doente com mais de uma patologia diagnosticada, estas patologias são do mesmo tipo (ambas congénitas ou ambas adquiridas).

Outro tipo de regras que se pode encontrar é um par de regras semelhante às seguintes

```
➔ descrição Patologia = DMRI ⇨ patologia = Adquirida (conf->1.0, sup->0.158)
➔ patologia = Adquirida ⇨ descrição Patologia = DMRI (conf->0.514, sup->0.158)
```

À primeira vista estas regras podem parecer idênticas, no entanto têm significados distintos e possibilitam a verificação de que as regras traduzem possíveis implicações, mas não equivalências. De facto pelos valores de confiança, verifica-se que todos os doentes que tem DMRI tem uma patologia do tipo Adquirida, mas que o contrário se verifica em apenas metade das situações.

Um outro caso de interesse é o facto de não terem sido detectadas relações entre os dados registados na tabela Tratamento, mas terem sido descobertas regras para a Tabela Tratamento desnormalizada.

```
➔ tratamento#1 = Cirúrgico ⇨ tratamento = Médico (conf->0.818, sup->0.134)
```

Este tipo de regra traduz a realização de vários tratamentos espaçados no tempo, em que na maioria das vezes quando existe um tratamento do tipo Cirúrgico, já foi usado um tratamento do tipo Médico.

Entre várias tabelas do mesmo esquema

Ao juntar as várias tabelas de um mesmo esquema obtiveram-se também alguns resultados, dos quais se destaca a redescoberta das relações anteriormente encontradas e a descoberta de novas relações entre diferentes tabelas.

Casos de redescoberta de regras:

```
➔ Evolução = Fásica ==> instalação = Progressiva (conf->0.818, sup->0.097)
➔ Instalação = Súbita ==> evolução = Estável (conf->0.762, sup->0.171)
```

Casos de descoberta de novas regras:

```
➔ Instalação = Progressiva evolução = Fásica ==> prematuridade = false (conf->1.0, sup-
 >0.107)
➔ Evolução = Linear ==> prematuridade = false (conf->1.0, sup->0.101)
➔ Familiares = Irrelevantes evolução = Estável ==> prematuridade = false (conf->0.826,
 sup->0.113)
```

Estas regras pouca informação trazem. Porém dão ideia de algumas relações fora de uma só tabela. Por outro lado, é necessário referir que o atributo “familiares” é um atributo textual, pelo que traz dificuldades acrescidas à descoberta de regras como já foi referido.

Entre várias tabelas de diferentes esquemas

Uma vez mais a redescoberta de algumas regras já encontradas tais como:

```
➔ descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup->0.101)
```

Porém consegue-se uma vez mais a descoberta de algumas novas regras tais como:

```
➔ patologia = Adquirida sexo = Masculino ==> raça = Caucasiana (conf->1.0, sup->0.107)
➔ patologia = Adquirida descpatologia = DMRI ==> raça = Caucasiana (conf->1.0, sup->0.101)
➔ descpatologia = DMRI ==> raça = Caucasiana (conf->1.0, sup->)
➔ patologia = Congénita sexo = Masculino ==> raça = Caucasiana (conf->0.969, sup->0.19)
➔ descpatologia = Disfunção mista da retina ==> raça = Caucasiana (conf->0.954, sup-
 >0.125)
➔ patologia = Congénita ==> raça = Caucasiana (conf->0.948, sup->0.327)
```

Também desta vez a informação descoberta é de pouco valor, uma vez que apenas traduz alguma caracterização da população que é acompanhada na consulta.

Substituição de valores omissos

Feita que está a apreciação dos resultados não utilizando nenhuma espécie de pré-processamento relacionada com o tipo de dados, verifique-se agora os resultados ao substituir os valores omissos pelo novo valor “Desconhecido”.

Algumas das regras encontradas trazem alguma informação

```
➔ Pessoais = Desconhecido ==> familiares = Desconhecido (conf->0.742, sup->0.154)
➔ Instalação = Desconhecido ==> evolução = Estável (conf->0.648, sup->0.179)
```

Todavia quase a totalidade destas regras traduz apenas a existência de um número demasiado elevado de valores omissos.

```
➔ Tratamento = Laser ==> descrição tratamento = Desconhecido (conf->1.0, sup->0.149)
➔ Tratamento = Medico ==> descrição tratamento = Desconhecido (conf->0.967, sup->0.333)
➔ Tratamento = Cirúrgico ==> descrição tratamento = Desconhecido (conf->0.75, sup->0.379)
```

De facto as regras encontradas cobrem todos os valores possíveis para o atributo Tratamento, e não diferenciam a sua descrição.

Pré-processamento de valores numéricos

Os resultados obtidos com o pré-processamento dos valores numéricos, através da discretização, foram sem dúvida melhor sucedidos. De facto foram, pela primeira vez encontradas relações envolvendo este tipo de valor em conjunção com os valores nominais.

```
➔ Idade incapacidade ='(-inf-8.9]' ==> idade doença ='(-inf-8.5]' (conf->1.0, sup->0.111)
➔ Idade doença ='(-inf-8.5]' ==> patologia = Congénita (conf->0.621, sup->0.179)
```

As regras apresentadas traduzem a relação esperada entre as durações da doença e da incapacidade, no caso de patologias congénitas e adquiridas.

Outras regras, relacionam estes atributos discretizados com os restantes

```
➔ Idade doença ='(-inf-8.5]' instalação = Súbita ==> evolução = Estável (conf->0.827, sup-
 >0.179)
➔ Idade incapacidade ='(53.4-inf)' ==> instalação = Progressiva (conf->0.695, sup->0.119)
```

Outras regras de interesse são as do tipo

```
➔ Id ='(-inf-0.2]' ==> tratamento = Médico (conf->0.626, sup->0.149)
➔ Id ='(0.2-inf)' ==> tratamento = Cirúrgico (conf->0.65, sup->0.482)
```

em que se verifica uma vez mais que um tratamento Cirúrgico é normalmente antecedido por um tratamento do tipo Médico, o que pode ser generalizado para todos as entidades com mais que um registo.

No Anexo C, apresentam-se todas as regras descobertas.

2 - COMPARAÇÃO DOS RESULTADOS

Pela descrição dos resultados obtidos, são visíveis vantagens na utilização de pré-processamento, sobretudo ao nível da desnormalização de tabelas e do tratamento de valores numéricos.

Dois aspectos há ainda a ser referenciados: os graus de confiança obtidos e a comparação dos resultados perante a variação do suporte.

Todas as regras encontradas têm um suporte mínimo de 10% e um grau de confiança superior ou igual a 50%, existindo uma forte percentagem com graus de confiança superiores a 75%. Estes valores são bastante razoáveis, e permitem atribuir alguma credibilidade às regras encontradas. Porém se se olhar para os valores de suporte, esta credibilidade é um pouco afectada, uma vez que o facto de existir uma regra com 100% de confiança e que foi descoberta com base em poucos registos (baixo suporte), não pode ser generalizada a outras situações com grande segurança.

Esta situação deve-se essencialmente, às grandes diferenças entre os doentesregistados.

Capítulo VII - CONCLUSÃO

> *"Pode bem ser o caso, que a maioria dos clínicos não sinta dificuldades em fazer diagnósticos, mas sim a estabelecer a imagem do estado do mundo, e a clarificar os seus objectivos e suposições, antes de fazer o diagnóstico. Se for este o caso, a avaliação daquela situação é o verdadeiro problema da tomada de decisão(...) Sistemas que apoiem os clínicos a fazer a avaliação dos dados em análise podem ser de maior valia do que os sistemas que tentam fabricar um diagnóstico."*
>
> *[Coiera 1994, pg. 3]*

1 - RESUMO DO TRABALHO DESENVOLVIDO

O desenvolvimento de sistemas de apoio à decisão para o domínio da medicina é uma desafio que requer sensibilidade às características únicas do ambiente deste domínio. Até à data, grande parte dos projectos nesta área têm sido motivados pela percepção da necessidade de apoio à tomada de decisões, sem qualquer ponto de referência para a sua natureza particular. O que se pretende com este trabalho é continuar a concertação de esforços com a equipa da Consulta de Subvisão do Hospital de Santa Maria, de modo a tornar possível a compreensão da natureza da prática clínica, e em particular das tarefas a que se pretende dar apoio.

Dada a já informatização desta consulta, o primeiro dos objectivos propostos, relacionava-se com a melhoria do processo de recolha de informação dos doentes. Esta melhoria foi conseguida com a distribuição da aplicação de gestão da consulta, assim como pela remodelação da base de dados existente (tentando resolver os problemas detectados durante a análise dos dados).

O segundo dos objectivos propostos relacionava-se com a descoberta de relações entre os dados registados. Apesar dos problemas de quantidade e qualidade dos dados registados ao longo dos últimos quatro anos, o sistema de aquisição de conhecimento criado conseguiu descobrir algumas regras. Essas regras, que apesar de pouco numerosas são válidas, anunciam já boas perspectivas face a um maior

número de dados. Porém, pequenos objectivos, tais como a melhoria dos questionários usados para a avaliação da visão funcional, ficaram uma vez mais adiados, essencialmente devido à escassez e incorrecção dos dados existentes.

Por outro lado, é de realçar que os sistemas de informação e de aquisição de conhecimento não são exclusivamente utilizáveis pela Consulta de Subvisão, e que para adopção destes sistemas por outros domínios de acompanhamento de doentes é perfeitamente viável, desde que seja criada uma nova base de dados semelhante à criada para a consulta de subvisão.

2 - TRABALHO FUTURO

De modo a conseguir a melhoria dos questionários é necessário identificar um conjunto de questões equivalentes dirigidas aos vários perfis de doentes. Dois aspectos há aqui a realçar: a classificação de doentes em perfis e a equivalência entre questões.

Para determinar o perfil de um doente, é necessário, em primeiro lugar, descobrir as várias classes. Esta descoberta pode ser feita recorrendo à metodologia da aprendizagem baseada em instâncias (mecanismos de *Clustering*), com base quer nos dados identificativos (como idade e sexo), quer nos dados de diagnóstico. Em relação à equivalência entre questões é necessário atingir dois objectivos: identificar os aspectos avaliados por uma questão, e definir novas questões que consigam avaliar o mesmo aspecto, mas dirigidas a cada um dos perfis. Facilmente se verifica que este problema só pode ser resolvido com um forte empenho do pessoal médico. Por outro lado, a descoberta de perfis usando clustering só é possível mediante o estabelecimento de "funções de distância" para cada um dos atributos de cada uma das entidades. Esta é mais uma das tarefas que não pode ser concretizada sem um apoio claro do pessoal médico.

Outra abordagem interessante seria a aplicação de mecanismo de Programação Lógica Indutiva sobre as regras encontradas pelo sistema desenvolvido, de modo a generalizá-las e a torná-las mais facilmente aplicáveis num sistema de apoio à decisão.

Por fim, um dos principais problemas da consulta não foi sequer abordado. Este problema diz respeito ao registo dos planos de reabilitação e sua interligação com a fase de avaliação. Este problema não foi abordado devido à grande inexperiência existente nesta área e às enormes dificuldades de formalização dos planos de reabilitação. Este é o verdadeiro ponto que necessita de um sistema de apoio à decisão: face ao diagnóstico efectuado, propor um plano de reabilitação adequado a cada doente. É possivelmente por aí que se poderá melhorar o acompanhamento dos doentes.

BIBLIOGRAFIA

[Adriaans 1996] Adriaans, P. e D. Zantinge. *Data Mining*. Addison-Wesley, 1996.

[Agrawal 1994] Agrawal, R. e R. Srikant. "Fast algorithms for mining association rules in large databases". *Proceedings of International Conference on Very Large Databases*, pp. 478-499. Santiago, Chile: Morgan Kaufmann, Los Altos, CA. 1994.

[Antunes 2000] Antunes, C. e I. Lynce, J.M. Pereira, J.P. Martins. "Realidade Virtual em Subvisão". *9º Encontro Nacional de Computação Gráfica*. 2000.

[Antunes 1998] Antunes, C. e I. Lynce. "Subvisão: avaliação e reabilitação em crianças", trabalho final de curso, IST. 1998.

[Apache 1999] Apache Group. *Apache 1.3 User's Guide*. 1999. (http://www.apache.org/)

[Apache Project 1999] *The Java Apache Project*. Apache JServ Servlet Engine Documentation. 1999. (http://java.apache.org/)

[Arnold 1998] Arnold, K. e J. Gosling. *The Java Programming Language*. Addison Wesley, 2ª edição. 1998.

[Chen 1996] Chen, Ming-Syan e Jiawei Han e Philip Yu. "Data Mining: An Overview from Database Perspective", in *IEEE Transaction on Knowledge and Data Engineering*, vol. 8, nº. 6, (pg. 866-883). 1996.

[Coiera 1994] Coiera, E. "The Role of Knowledge Based Systems in Clinical Practice", in P.Barahona, J.P. Christensen(eds.), *Knowledge and Decisions in Health Telematics – The Next Decada* (pg. 199-203). IOS Press. Amsterdam. 1994.

[Fayyad 1996] Fayyad, U.M., G. Shapiro, P. Smyth. "From Data Mining to Knowledge

Discovery: An Overview", in Fayyad, U.M., G. Shapiro, P. Smyth, R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining* (pg. 1-36). AAAI Press. 1996.

[Ferreira 1999] Ferreira, P. *Suporte de Aplicações Distribuídas*, folhas da cadeira de Aplicações em Redes de Grande Escala, do IST.1999.

[Frawley 1992] Frawley, W.J., G. Piatetsky-Shapiro e C.J. Matheus. "Knowledge discovery in databases: An overview". *AI Magazine*, vol. 13 nº. 3 (pg. 57-70). 1992.

[Freier 1996] Freier, A.O. e P. Karlton e P.C. Kocher. *The SSL Protocol Version 3.0*. 1996. (http://www.netscape.com/eng/ssl3/draft302.txt)

[Giarratano 1994] Giarratano, J. e G. Riley. *Expert Systems: Principles and Programming*, 2ª Edição. Publishing Company Boston. 1994.

[Heckerman 1996] Heckerman, D. "Bayesian Networks for Knowledge Discovery", in Fayyad, U.M., G. Shapiro, P. Smyth, R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining* (pg. 273-306). AAAI Press. 1996.

[Holsheimer 1994] Holsheimer, M. e A. Siebes. *Data Mining: the search for knowledge in databases*. Report CS-R9406, CWI. Amsterdam.1994.

[Holsheimer 1995] Holsheimer, M. e M.L. Kersten e H. Mannila e H. Toivonen. "A perspective on databases and data mining". *First International Conference on Knowledge Discovery and Data Mining*– KDD'95 (pg. 150 – 155). AAAI Press. 1995.

[Kryszkiewicz 1997] Kryszkiewicz, M. "Generation of Rules from Incomplete Information Systems", in *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence, vol. 1263 (pg. 156-166). Springer. 1997.

[Kryszkiewicz 1998] Kryszkiewicz, M. "Representative Association Rules", in *Research and Development in Knowledge Discovery and Data Mining*, Lecture Notes in Artificial Intelligence, vol. 1394 (pg. 198-211). Springer. 1998.

[Lockhart 1998] Lockhart, T. *PostgreSQL Documentation*. 1998. (http://www.postgresql.org/)

[Lu 1997] Lu, H. e H. Motoda e H. Liu. *KDD: Techniques and Applications*. World Scientific. 1997.

[Kononenko 1997] Kononenko, I. e I. Bratko e M. Kukar. "Application of Machine Learning to Medical Diagnosis". In R. Michalsky. e I. Bratko e M. Kubat. *Machine Learning and Data Mining – Methods and Aplications* (pg. 389-408). John Wiley and Sons, Lda. 1998.

[Marques 1998] Marques, J.A. e P. Guedes. *Tecnologia de Sistemas Distribuídos*. FCA. 1998.

[Mayhew 1992] Mayhew, D.J. *Principles and Guidelines in Software User Interface De*sign. Englewood Cliffs, NJ: Prenctice Hall. 1992.

[Minsky 1985] Minsky, M. *The Society of Mind*. Simon and Schuster. 1985.

[Mitchell 1997] Mitchell, T.M. *Machine Learning*. MacGraw-Hill International Editions. 1997.

[Mitchell 1981] Mitchell, T.M. "Generalization as Search". In Shavlik e Dietterich, *Readings in Machine Learning*, pág. 96-107. Morgan Kaufmann 1990.

[Nakhaeizadeh 1997] Nakhaeizadeh, G. e C. Taylor. *Machine Learning and Statistics, The Interface*. Wiley InterScience. 1997.

[Neves 2000] Neves, C. e C.Antunes e I. Lynce e J.P. Martins e A.C. Dinis."A Computer-Based Assessment and Rehabilitation of Visual Function in Low vision Children", in Stuen, C., A. Arditi, A. Horowitz, M.A. Lang, B. Rosenthal and K. Seidman, *Vision Rehabilitation, Assessment, Intervention, and Outcomes* (pg. 376-379).Swets & Zeitlinger Publishers. 2000.

[OpenSSL 1999] *The OpenSSL Project*. OpenSSL 0.9.4 README file. 1999. (http://www.openssl.org/)

[Pina 1997] Pina, L. e V. Norte e C. Neves e J.P. Martins e C. Dinis. *Informatização da Consulta de Subvisão*, trabalho final de curso, IST. 1997.

[Pressman 1992] Pressman, R.S. *Software Engineering: a Practitioner's Approach*. European Edition, Berkshire, Inglaterra: McGraw-Hill Publishing Company. 1992.

[Quinlan 1990] Quinlan, J.R. "Learning Logical Definitions from Relations", in *Machine Learning*, vol. 5 (pg 239-266). Kluwer Academic Publishers. 1990.

[Quinlan 1989] Quinlan, J.R. *Inferring Decision Trees Using Minimum Description Length Principle*. Academic Press. 1989.

[Quinlan 1986] Quinlan, J.R. "Induction of Decision Trees", in *Machine Learning*, vol. 1 (pg 81-106). Kluwer Academic Publishers.1986.

[Simon 1990] Simon, H.A. e G. Lea. "Problem Solving and Rule Induction: a Unified View". Shavlik, J. e Dietterich, T. *Readings in Machine Learning* (pg. 26-37). Morgan Kaufmann. 1990.

[Ragel 1998] Ragel, B. e Crémilleux. "Treatment of Missing Values for Association Rules", in *Research and Development in Knowledge Discovery and Data Mining*, Lecture Notes in Artificial

Intelligence, vol. 1394 (pg. 258-270). Springer. 1998.

[Ralf 1999] Ralf, S. *MOD_SSL version 2.4. The Apache Interface to OpenSSL User Manual*. 1999. (http://www.modssl.org/)

[Ramakrishnan 2000] Ramakrishnan, R. e J. Gehrke. Database Management Systems. McGraw Hill. 2000.

[Russel 1995] Russel, S. e P. Norvig. *Artificial Intelligence – a modern approach*. Prentice Hall International Editions. 1995.

[Shavlik 1990] Shavlik, J. e Dietterich, T. *Readings in Machine Learning* (pg. 1-10). Morgan Kaufmann. 1990.

[Silverstone 1999] Silverstone, B. "Coverage for Vision Rehabilitation Services: a National Imperative", in *LightHouse News*. LightHouse International. 1999.

(http://www.lighthouse.org/lighthouse_news/lighthouse_news_fall99_rehab.htm)

[Stefanowski 1997] Stefanowski, J. e K. Slowinski. "Rough Set Theory and Rule Induction Techniques for Discovery of Attribute Dependencies in Medical Information Systems", in *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence, vol. 1263 (pg. 36-46). Springer. 1997.

[Viisola 1998] Viisola, M. *Statistics on Children and Visual Impairments.* Lighthouse International. 1998.

[Weiss 1998] Weiss, S. e N. Indurkhya. *Predictive Data Mining – a practical guide*. Morgan Kaufmann Publishers, Inc. 1998.

[Witten 2000] Witten, I.H. e E. Frank. *Data Mining Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann. 2000.

[Wu 1995] Wu, X. *Knowledge Acquisition from Databases.* Ablex Publishing Corporation. 1995.

ANEXO A – BASE DE DADOS

Tabelas do Esquema Dados

```
CREATE TABLE dados_Doente
(
     numSV              int,
     dtNascimento       date,
     dt1aConsulta       date,
     raca               text,
     sexo               text,
     enviadoPor         text,
     grauEscolaridade   text,
     sitProfissional    text,
     sisSegSocial       text,
     localResidencia    text,
     agregadoFamiliar   text,

     PRIMARY KEY (numSV)
);

CREATE TABLE dados_HistorialClinico
(
     numSV              int,
     prematuridade      bool,
     pessoais           text,
     familiares         text,

     PRIMARY KEY (numSV)
);

CREATE TABLE dados_Diagnostico
(
     numSV              int,
     id                 int,
     patologia          text,
     descPatologia      text,
     orgaoAfectado      text,
     PRIMARY KEY (numSV, id)
);
CREATE TABLE dados_Incapacidade
(
```

```
    numSV        int,
    idadeDoe     int,
    idadeIncap   int,
    instalacao   text,
    evolucao     text,
    PRIMARY KEY (numSV)
);

CREATE TABLE dados_Tratamento
(
    numSV              int,
    id                 int,
    idade              int,
    tratamento         text,
    descrTratamento    text,
    PRIMARY KEY (numSV, id)
);
```

Tabelas do Esquema Confidenciais

```
CREATE TABLE conf_Doente
(
    numSV              int,
    numHospital        text,
    nome               text,
    telefone           text,
    numSegSocial       text,

    PRIMARY KEY (numSV)
);

CREATE TABLE conf_Morada
(
    numSV              int,
    rua                text,
    codPostalLocal     text,
    codPostalNum       text,

    PRIMARY KEY (numSV)
);
```

Tabelas do Esquema Valores

```
CREATE TABLE tp_Raca
(
    id           int,
    raca         text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_Sexo
(
    id           int,
    sexo         text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_EnviadoPor
(
    id           int,
    origem       text,
```

```
    PRIMARY KEY (id)
);

CREATE TABLE tp_Escolaridade
(
    id           int,
    escolaridade text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_SitProfissional
(
    id                int,
    sitProfissional   text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_LocalResidencia
(
    id                int,
    localResidencia   text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_AgregadoFamiliar
(
    id                int,
    agregadoFamiliar  text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_SisSegSocial
(
    id           int,
    sisSegSocial text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_Tratamento
(
    id           int,
    tratamento   text,
    PRIMARY KEY (id)
);

CREATE TABLE tp_Patologia
(
    id           int,
    patologia    text,
    PRIMARY KEY  (id)
);

CREATE TABLE tp_DescPatologia
(
    id           int,
    descPatologia text,
    PRIMARY KEY (id)
);
```

```
CREATE TABLE tp_Incapacidade
(
     id           int,
     incapacidade text,
     PRIMARY KEY  (id)
);

CREATE TABLE tp_Instalacao
(
     id           int,
     instalacao   text,
     PRIMARY KEY  (id)
);

CREATE TABLE tp_Evolucao
(
     id           int,
     evolucao     text,
     PRIMARY KEY  (id)
);
```

ANEXO B – ANÁLISE DOS DADOS

Apresentam-se alguns gráficos que dão uma ideia geral sobre os dados registados na base de dados, e usados como fonte para o sistema de aquisição de conhecimento.

O número de registos efectuados para cada um dos exames, que avaliam as Funções Visuais, é extremamente diminuto, quantificando-se na maioria dos casos por apenas 10 – 15 registos.

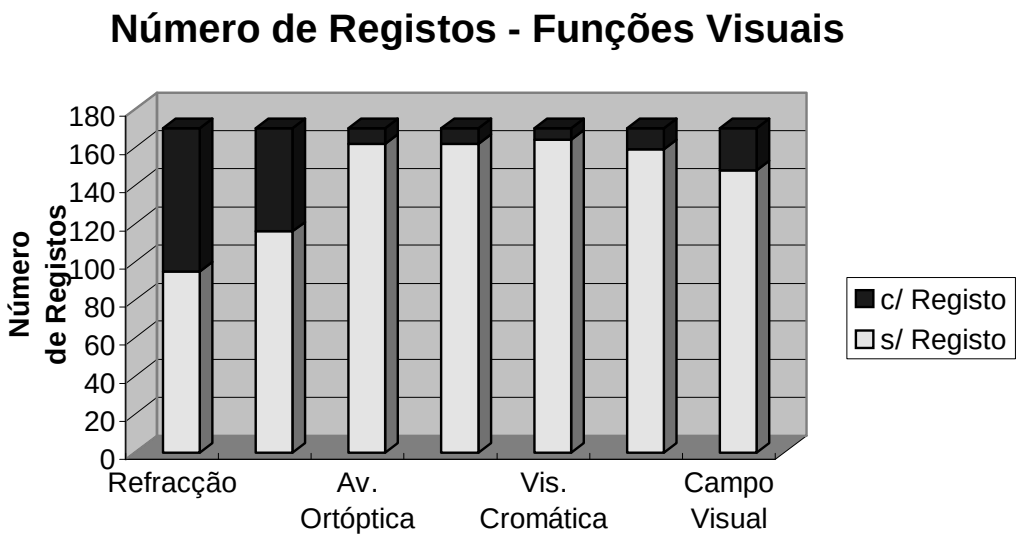

Figura 19 - Número de registos - Funções Visuais

Ao nível dos registos da avaliação da Visão Funcional, os números são significativamente maiores. Contudo foram detectados alguns problemas na representação dos valores FALSO e não preenchido, o que levou à impossibilidade da análise dos dados.

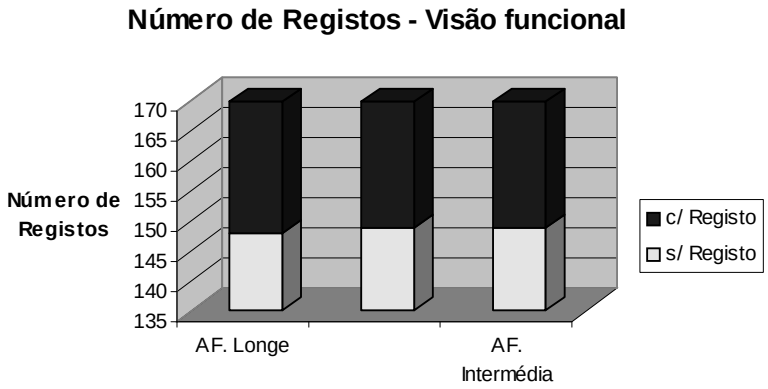

Figura 20 - Número de registos - Visão Funcional

Como se pode perceber pelo gráfico seguinte, a população distribui-se por cinco faixas etárias de forma quase uniforme.

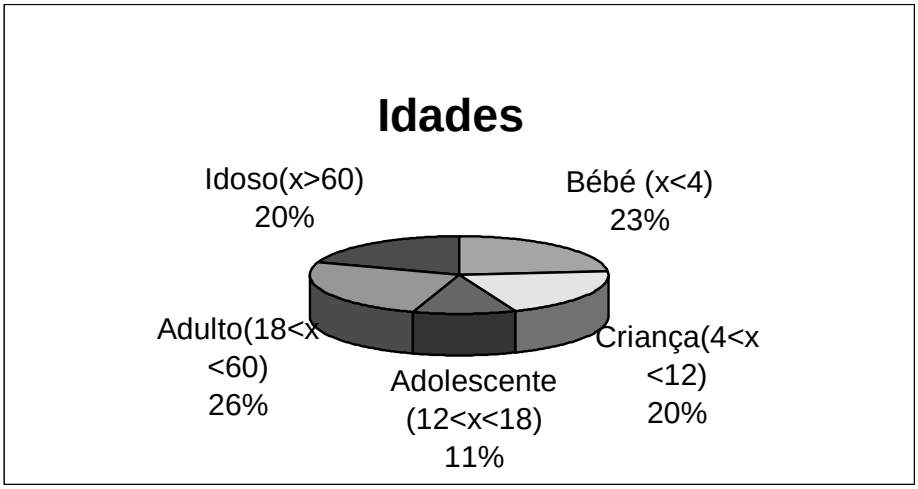

Figura 21 - Distribuição dos doentes por idade

Relativamente aos Diagósticos, Tratamentos e Incapacidades verifica-se a existência de alguns valores omissos, e uma distribuição pouco equilibrada, entre os valores possíveis para cada um dos atributos.

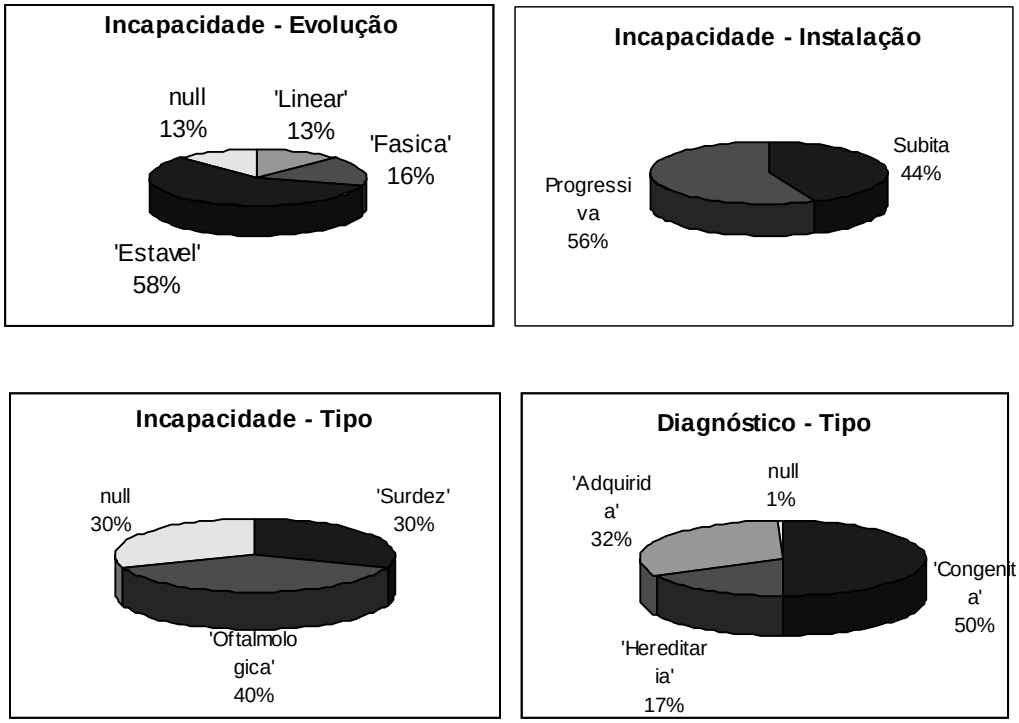

Figura 22 - Valores assumidos por alguns dos atributos de algumas entidades

Repare-se ainda que o número de registos de diagnósticos e tratamentos por doente é, na maioria dos casos, igual a 1. Isto evidencia alguns problemas na recolha de dados: possivelmente o pessoal médico eliminou dados anteriores de modo a registar dados mais recentes para o mesmo doente.

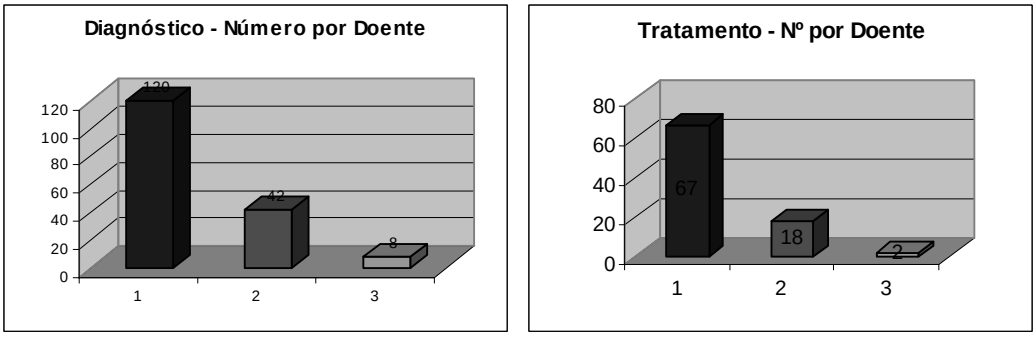

Figura 23 - Número de registos temporais para cada doente

Por fim, é interessante verificar a variedade de patologias registadas. Destas, apenas uma pequena percentagem se repete entre os doentes. As restantes ocorrem apenas num ou pouquíssimos casos.

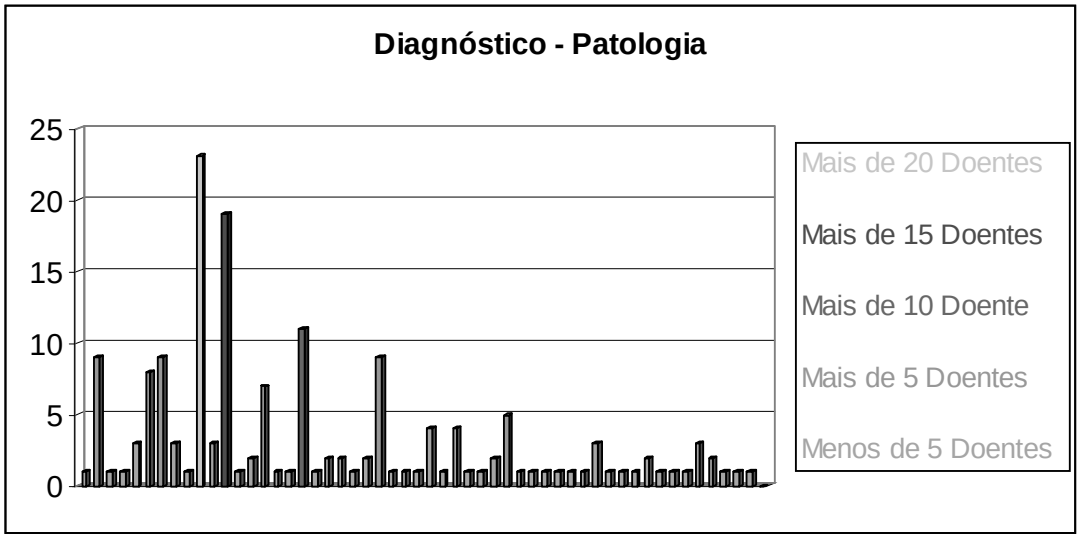

Figura 24 - Tipos de patologias

ANEXO C – REGRAS DESCOBERTAS

Por tabela

```
METODO APRIORI COM
       conf=50%, suporte=10% e numRegras=25

REGRAS TABELA dados_historialclinico
-> familiares=Irrelevantes ==> prematuridade=false (conf->0.825, sup ->0.196)

REGRAS TABELA dados_incapacidade
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup ->0.134)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.238)

REGRAS TABELA dados_tratamento
Não foram encontradas regras para a tabela dados_tratamento

REGRAS TABELA dados_diagnostico
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.111)
```

Por tabela desnormalizada

```
METODO APRIORI COM
 conf=50%, suporte=10% e numRegras=25

REGRAS TABELA tratamento desnormalizada
-> tratamento#1=Cirurgico ==> tratamento=Medico (conf->0.818, sup ->0.134)

REGRAS TABELA diagnostico desnormalizada
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.158)
-> patologia#1=Congenita ==> patologia=Congenita (conf->0.863, sup ->0.158)
-> patologia#1=Adquirida ==> patologia=Adquirida (conf->0.764, sup ->0.108)
-> descpatologia=Disfuncao mista da retina ==> patologia=Hereditaria (conf->0.636, sup -
>0.116)
-> patologia=Hereditaria ==> descpatologia=Disfuncao mista da retina (conf->0.56, sup -
>0.116)
-> patologia=Adquirida ==> descpatologia=DMRI (conf->0.513, sup ->0.158)
```

Entre várias tabelas do mesmo esquema

```
METODO APRIORI -- DADOS_DIAGNOSTICO <-> DADOS_INCAPACIDADE
       conf=50%, suporte=10% e numRegras=25

-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.097)
-> patologia#1=Congenita ==> patologia=Congenita (conf->0.857, sup ->0.134)
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.764, sup ->0.097)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.718, sup ->0.171)
-> patologia=Adquirida ==> evolucao=Estavel (conf->0.538, sup ->0.104)
-> patologia=Congenita ==> evolucao=Estavel (conf->0.509, sup ->0.194)
```

```
-> patologia=Adquirida ==> instalacao=Subita (conf->0.5, sup ->0.097)
-> patologia=Adquirida ==> descpatologia=DMRI (conf->0.5, sup ->0.097)


METODO APRIORI
       conf=50%, suporte=10% e numRegras=25

REGRAS ESQUEMA esquema—Dados
-> evolucao=Fasica ==> prematuridade=false (conf->1.0, sup ->0.130)
-> instalacao=Progressiva evolucao=Estavel ==> prematuridade=false (conf->1.0, sup ->0.119)
-> instalacao=Progressiva evolucao=Fasica ==> prematuridade=false (conf->1.0, sup ->0.107)
-> evolucao=Linear ==> prematuridade=false (conf->1.0, sup ->0.101)
-> instalacao=Progressiva ==> prematuridade=false (conf->0.981, sup ->0.309)
-> familiares=Irrelevantes evolucao=Estavel ==> prematuridade=false (conf->0.826, sup -
>0.113)
-> familiares=Irrelevantes ==> prematuridade=false (conf->0.825, sup ->0.196)
-> evolucao=Fasica ==> prematuridade=false instalacao=Progressiva (conf->0.818, sup -
>0.107)
-> prematuridade=false evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup -
>0.107)
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup ->0.107)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.190)
-> instalacao=Subita ==> prematuridade=false (conf->0.761, sup ->0.190)
-> evolucao=Estavel ==> prematuridade=false (conf->0.710, sup ->0.321)
-> prematuridade=false instalacao=Subita ==> evolucao=Estavel (conf->0.687, sup ->0.130)
-> instalacao=Subita evolucao=Estavel ==> prematuridade=false (conf->0.687, sup ->0.130)
-> prematuridade=true ==> evolucao=Estavel (conf->0.578, sup ->0.130)
-> prematuridade=false familiares=Irrelevantes ==> evolucao=Estavel (conf->0.575, sup -
>0.113)
-> familiares=Irrelevantes ==> evolucao=Estavel (conf->0.575, sup ->0.136)
-> instalacao=Subita ==> prematuridade=false evolucao=Estavel (conf->0.523, sup ->0.130)
```

Entre várias tabelas de diferentes esquemas

```
METODO APRIORI -- DADOS_DIAGNOSTICO <-> CONF_DOENTE
       conf=50%, suporte=10% e numRegras=25

-> enviadopor=ARP ==> raca=Caucasiana (conf->1.0, sup ->0.119)
-> patologia=Adquirida sexo=Masculino ==> raca=Caucasiana (conf->1.0, sup ->0.107)
-> descpatologia=DMRI ==> patologia=Adquirida raca=Caucasiana (conf->1.0, sup ->0.101)
-> patologia=Adquirida descpatologia=DMRI ==> raca=Caucasiana (conf->1.0, sup ->0.101)
-> descpatologia=DMRI raca=Caucasiana ==> patologia=Adquirida (conf->1.0, sup ->0.101)
-> descpatologia=DMRI ==> raca=Caucasiana (conf->1.0, sup ->0.101)
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.101)
-> patologia=Adquirida ==> raca=Caucasiana (conf->0.971, sup ->0.202)
-> patologia=Congenita sexo=Masculino ==> raca=Caucasiana (conf->0.969, sup ->0.190)
-> sexo=Masculino enviadopor=Consulta de Desenvolvimento ==> raca=Caucasiana (conf->0.956,
sup ->0.130)
-> descpatologia=Disfuncao mista da retina ==> raca=Caucasiana (conf->0.954, sup ->0.125)
-> patologia=Congenita ==> raca=Caucasiana (conf->0.948, sup ->0.327)
-> patologia=Congenita patologia#1=Congenita ==> raca=Caucasiana (conf->0.947, sup ->0.107)
-> patologia=Adquirida sexo=Feminino ==> raca=Caucasiana (conf->0.941, sup ->0.0952)
-> sexo=Masculino ==> raca=Caucasiana (conf->0.928, sup ->0.464)
-> patologia=Congenita sexo=Feminino ==> raca=Caucasiana (conf->0.92, sup ->0.136)
-> patologia#1=Congenita ==> raca=Caucasiana (conf->0.909, sup ->0.119)
-> patologia#1=Congenita raca=Caucasiana ==> patologia=Congenita (conf->0.9, sup ->0.107)
-> patologia=Hereditaria ==> raca=Caucasiana (conf->0.88, sup ->0.130)
-> patologia#1=Congenita ==> patologia=Congenita (conf->0.863, sup ->0.113)
-> enviadopor=Consulta de Desenvolvimento ==> raca=Caucasiana (conf->0.833, sup ->0.238)
-> patologia#1=Congenita ==> patologia=Congenita raca=Caucasiana (conf->0.818, sup ->0.107)
-> sexo=Feminino ==> raca=Caucasiana (conf->0.817, sup ->0.398)
-> sexo=Feminino enviadopor=Consulta de Desenvolvimento ==> raca=Caucasiana (conf->0.72,
sup ->0.107)
-> patologia=Congenita raca=Caucasiana ==> sexo=Masculino (conf->0.581, sup ->0.190)
```

Pré-processamento: Substituição de valores omissos – valor Desconhecido

```
METODO APRIORI EM PRE-PROCESSADAS - DESCONHECIDO
       conf=50%, suporte=10% e numRegras=25

REGRAS TABELA dados_historialclinico--preproc
-> pessoais=Desconhecido familiares=Desconhecido ==> prematuridade=false (conf->1.0, sup -
>0.154)
-> pessoais=Desconhecido ==> prematuridade=false (conf->0.971, sup ->0.202)
-> familiares=Desconhecido ==> prematuridade=false (conf->0.826, sup ->0.226)
-> familiares=Irrelevantes ==> prematuridade=false (conf->0.825, sup ->0.196)
-> prematuridade=false pessoais=Desconhecido ==> familiares=Desconhecido (conf->0.764, sup
->0.154)
-> pessoais=Desconhecido ==> prematuridade=false familiares=Desconhecido (conf->0.742, sup
->0.154)
-> pessoais=Desconhecido ==> familiares=Desconhecido (conf->0.742, sup ->0.154)
-> prematuridade=false familiares=Desconhecido ==> pessoais=Desconhecido (conf->0.684, sup
->0.154)
-> familiares=Desconhecido ==> prematuridade=false pessoais=Desconhecido (conf->0.565, sup
->0.154)
-> familiares=Desconhecido ==> pessoais=Desconhecido (conf->0.565, sup ->0.154)

REGRAS TABELA dados_incapacidade--preproc
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup ->0.134)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.238)
-> instalacao=Desconhecido ==> evolucao=Estavel (conf->0.648, sup ->0.179)

REGRAS TABELA dados_tratamento--preproc
-> tratamento=Laser ==> descrtratamento=Desconhecido (conf->1.0, sup ->0.149)
-> tratamento=Medico ==> descrtratamento=Desconhecido (conf->0.967, sup ->0.333)
-> tratamento=Cirurgico ==> descrtratamento=Desconhecido (conf->0.75, sup ->0.379)

REGRAS TABELA dados_diagnostico--preproc
-> descpatologia=Disfuncao mista da retina ==> orgaoafectado=Desconhecido (conf->1.0, sup -
>0.135)
-> descpatologia=DMRI ==> patologia=Adquirida orgaoafectado=Desconhecido (conf->1.0, sup -
>0.111)
-> patologia=Adquirida descpatologia=DMRI ==> orgaoafectado=Desconhecido (conf->1.0, sup -
>0.111)
-> descpatologia=DMRI orgaoafectado=Desconhecido ==> patologia=Adquirida (conf->1.0, sup -
>0.111)
-> descpatologia=DMRI ==> orgaoafectado=Desconhecido (conf->1.0, sup ->0.111)
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.111)
-> patologia=Hereditaria ==> orgaoafectado=Desconhecido (conf->0.965, sup ->0.164)
-> patologia=Congenita ==> orgaoafectado=Desconhecido (conf->0.811, sup ->0.405)
-> patologia=Adquirida ==> orgaoafectado=Desconhecido (conf->0.709, sup ->0.229)
-> orgaoafectado=Desconhecido ==> patologia=Congenita (conf->0.503, sup ->0.405)
```

Pré-processamento de valores numéricos

```
METODO APRIORI EM PRE-PROCESSADAS - DISCRETIZACAO
       conf=50%, suporte=10% e numRegras=25

REGRAS TABELA dados_historialclinico--preproc
-> familiares=Irrelevantes ==> prematuridade=false (conf->0.825, sup ->0.196)

REGRAS TABELA dados_incapacidade--preproc
-> idadeincap='(-inf-8.9]' ==> idadedoe='(-inf-8.5]' (conf->1.0, sup ->0.111)
-> idadedoe='(-inf-8.5]' instalacao=Subita ==> evolucao=Estavel (conf->0.827, sup ->0.179)
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup ->0.134)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.238)
-> instalacao=Subita evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.75, sup ->0.179)
-> evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.705, sup ->0.410)
-> idadeincap='(53.4-inf)' ==> instalacao=Progressiva (conf->0.695, sup ->0.119)
-> instalacao=Subita ==> idadedoe='(-inf-8.5]' (conf->0.690, sup ->0.216)
-> idadedoe='(-inf-8.5]' ==> evolucao=Estavel (conf->0.639, sup ->0.410)
```

```
-> idadeincap='(53.4-inf)' ==> evolucao=Estavel (conf->0.608, sup ->0.104)
-> instalacao=Subita ==> idadedoe='(-inf-8.5]' evolucao=Estavel (conf->0.571, sup ->0.179)

REGRAS TABELA dados_tratamento--preproc
-> idade='All' ==> tratamento=Cirurgico (conf->1.0, sup ->0.091)
-> tratamento=Cirurgico ==> id='(-inf-0.2]' (conf->0.954, sup ->0.482)
-> id='(0.2-inf)' ==> tratamento=Medico (conf->0.65, sup ->0.149)
-> id='(-inf-0.2]' ==> tratamento=Cirurgico (conf->0.626, sup ->0.482)
-> tratamento=Laser ==> id='(-inf-0.2]' (conf->0.615, sup ->0.091)
-> tratamento=Medico ==> id='(-inf-0.2]' (conf->0.566, sup ->0.195)

REGRAS TABELA dados_diagnostico--preproc
-> descpatologia=DMRI ==> id='(-inf-0.2]' patologia=Adquirida (conf->1.0, sup ->0.111)
-> id='(-inf-0.2]' descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.111)
-> patologia=Adquirida descpatologia=DMRI ==> id='(-inf-0.2]' (conf->1.0, sup ->0.111)
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.111)
-> descpatologia=DMRI ==> id='(-inf-0.2]' (conf->1.0, sup ->0.111)
-> descpatologia=Disfuncao mista da retina ==> id='(-inf-0.2]' (conf->0.956, sup ->0.129)
-> patologia=Hereditaria ==> id='(-inf-0.2]' (conf->0.862, sup ->0.147)
-> patologia=Congenita ==> id='(-inf-0.2]' (conf->0.682, sup ->0.341)
-> patologia=Adquirida ==> id='(-inf-0.2]' (conf->0.672, sup ->0.217)
-> id='(0.2-inf)' ==> patologia=Congenita (conf->0.54, sup ->0.158)
-> id='(-inf-0.2]' patologia=Adquirida ==> descpatologia=DMRI (conf->0.513, sup ->0.111)

REGRAS TABELA dados_tratamento--alisada--preproc
-> tratamento#1=Medico ==> tratamento=Cirurgico (conf->0.818, sup ->0.134)

REGRAS TABELA dados_diagnostico--alisada--preproc
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.158)
-> patologia#1=Congenita ==> patologia=Congenita (conf->0.863, sup ->0.158)
-> patologia#1=Adquirida ==> patologia=Adquirida (conf->0.764, sup ->0.108)
-> descpatologia=Disfuncao mista da retina ==> patologia=Hereditaria (conf->0.636, sup -
>0.117)
-> patologia=Hereditaria ==> descpatologia=Disfuncao mista da retina (conf->0.56, sup -
>0.117)
-> patologia=Adquirida ==> descpatologia=DMRI (conf->0.513, sup ->0.158)

METODO APRIORI EM ESQUEMA DADOS
       conf=50%, suporte=10% e numRegras=25

REGRAS TABELA esquema--Dados—preproc
-> prematuridade=true evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->1.0, sup ->0.130)
-> evolucao=Fasica ==> prematuridade=false (conf->1.0, sup ->0.130)
-> idadeincap='(53.4-inf)' ==> prematuridade=false (conf->1.0, sup ->0.130)
-> instalacao=Progressiva evolucao=Estavel ==> prematuridade=false (conf->1.0, sup ->0.119)
-> instalacao=Progressiva evolucao=Fasica ==> prematuridade=false (conf->1.0, sup ->0.107)
-> evolucao=Linear ==> prematuridade=false (conf->1.0, sup ->0.101)
-> instalacao=Progressiva ==> prematuridade=false (conf->0.981, sup ->0.309)
-> idadedoe='(-inf-8.5]' instalacao=Progressiva ==> prematuridade=false (conf->0.952, sup -
>0.119)
-> idadedoe='(-inf-8.5]' instalacao=Subita ==> evolucao=Estavel (conf->0.827, sup ->0.142)
-> familiares=Irrelevantes evolucao=Estavel ==> prematuridade=false (conf->0.826, sup -
>0.113)
-> familiares=Irrelevantes ==> prematuridade=false (conf->0.825, sup ->0.196)
-> evolucao=Fasica ==> prematuridade=false instalacao=Progressiva (conf->0.818, sup -
>0.107)
-> prematuridade=false evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup -
>0.107)
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.818, sup ->0.107)
-> prematuridade=true idadedoe='(-inf-8.5]' ==> evolucao=Estavel (conf->0.785, sup ->0.130)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.190)
-> instalacao=Subita ==> prematuridade=false (conf->0.761, sup ->0.190)
-> instalacao=Subita evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.75, sup ->0.142)
-> prematuridade=true ==> idadedoe='(-inf-8.5]' (conf->0.736, sup ->0.166)
-> evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.723, sup ->0.327)
-> evolucao=Estavel ==> prematuridade=false (conf->0.710, sup ->0.321)
-> instalacao=Subita ==> idadedoe='(-inf-8.5]' (conf->0.690, sup ->0.172)
-> prematuridade=false instalacao=Subita ==> evolucao=Estavel (conf->0.687, sup ->0.130)
```

```
-> instalacao=Subita evolucao=Estavel ==> prematuridade=false (conf->0.687, sup ->0.130)
-> idadedoe='(-inf-8.5]' ==> prematuridade=false (conf->0.674, sup ->0.345)

PRE PROCESSAMENTO -> DISCRETIZAÇÃO
-> patologia=Congenita instalacao=Subita ==> idadedoe='(-inf-8.5]' (conf->1.0, sup ->0.097)
-> idadeincap='(-inf-8.9]' ==> idadedoe='(-inf-8.5]' (conf->1.0, sup ->0.097)
-> descpatologia=DMRI ==> patologia=Adquirida (conf->1.0, sup ->0.097)
-> patologia#1=Congenita idadedoe='(-inf-8.5]' ==> patologia=Congenita (conf->0.941, sup -
>0.119)
-> idadedoe='(-inf-8.5]' instalacao=Progressiva ==> patologia=Congenita (conf->0.928, sup -
>0.097)
-> patologia=Congenita evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.923, sup -
>0.179)
-> patologia=Congenita patologia#1=Congenita ==> idadedoe='(-inf-8.5]' (conf->0.888, sup -
>0.119)
-> patologia#1=Congenita ==> patologia=Congenita (conf->0.857, sup ->0.134)
-> idadeincap='(53.4-inf)' ==> patologia=Adquirida (conf->0.842, sup ->0.119)
-> patologia#1=Congenita ==> idadedoe='(-inf-8.5]' (conf->0.809, sup ->0.126)
-> patologia=Congenita ==> idadedoe='(-inf-8.5]' (conf->0.803, sup ->0.305)
-> patologia=Congenita instalacao=Progressiva ==> idadedoe='(-inf-8.5]' (conf->0.764, sup -
>0.097)
-> evolucao=Fasica ==> instalacao=Progressiva (conf->0.764, sup ->0.097)
-> idadedoe='(-inf-8.5]' instalacao=Subita ==> evolucao=Estavel (conf->0.761, sup ->0.119)
-> patologia#1=Congenita ==> patologia=Congenita idadedoe='(-inf-8.5]' (conf->0.761, sup -
>0.119)
-> instalacao=Subita ==> evolucao=Estavel (conf->0.718, sup ->0.171)
-> evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.717, sup ->0.320)
-> instalacao=Subita evolucao=Estavel ==> idadedoe='(-inf-8.5]' (conf->0.695, sup ->0.119)
-> idadeincap='(53.4-inf)' ==> instalacao=Progressiva (conf->0.684, sup ->0.097)
-> instalacao=Subita ==> idadedoe='(-inf-8.5]' (conf->0.656, sup ->0.156)
-> idadedoe='(-inf-8.5]' ==> evolucao=Estavel (conf->0.651, sup ->0.320)
-> idadedoe='(-inf-8.5]' ==> patologia=Congenita (conf->0.621, sup ->0.305)
-> idadedoe='(-inf-8.5]' instalacao=Subita ==> patologia=Congenita (conf->0.619, sup -
>0.097)
-> patologia=Adquirida ==> idadeincap='(53.4-inf)' (conf->0.615, sup ->0.119)
-> patologia=Congenita idadedoe='(-inf-8.5]' ==> evolucao=Estavel (conf->0.585, sup -
>0.179)
```