BIOBI- FERRAMENTA PARA ANÁLISE DE DADOS BIOLÓGICOS

HERBERT RAUSCH FERNANDES

BIOBI- FERRAMENTA PARA ANÁLISE DE DADOS BIOLÓGICOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Sérgio Vale Aguiar Campos
Coorientador: Alessandra Conceição Faria Aguiar Campos

Belo Horizonte
Setembro de 2012

Fernandes, Herbert Rausch

F363b **BioBI**- Ferramenta para Análise de Dados Biológicos / Herbert Rausch Fernandes. — Belo Horizonte, 2012
xx, 73 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Sérgio Vale Aguiar Campos

1. Computação - Teses. 2. Bioinformática - Teses. 3. Mineração de Dados - Teses. I. Orientador. II. Título.

CDU 519.6*93 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

BioBI - Ferramenta para análise de dados biológicos

HERBERT RAUSCH FERNANDES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. SÉRGIO VALE AGUIAR CAMPOS - Orientador
Departamento de Ciência da Computação - UFMG

DRA. ALESSANDRA C. FARIA AGUIAR CAMPOS - Co-orientadora
Bolsista Pós-Doc/ ICB - UFMG

DR. GUILHERME CORREA DE OLIVEIRA
Centro de Pesq. René Rachou Fund. Oswaldo Cruz

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 10 de outubro de 2012.

Dedico este trabalho a todos que me acompanharam e apoiaram nesta jornada.

"Não importa o quanto você bate, mas sim o quanto
aguenta apanhar e continuar lutando.
O quanto pode suportar e seguir em frente.
É assim que se ganha."
(Rocky Balboa)

Resumo

O processo moderno de pesquisa biomédica gera um grande volume de dados de diferentes tipos por diversos equipamentos. O gerenciamento desses dados e o acompanhamento da execução das atividades demandam sistemas de gerenciamento para auxiliar nessas atividades, conhecidos como *Laboratory Information Management Systems* ou LIMS. Esses sistemas computacionais são utilizados para gerenciar dados produzidos em laboratórios com ênfase na qualidade. Entretanto, essas ferramentas não são suficientes para analisar e fornecer informações aos usuários sobre os dados, sendo necessários outros sistemas e técnicas especificas para a realizar esta atividade com maior eficiência. Os chamados *Sistemas de Apoio a Decisão* são ferramentas que têm este objetivo. Estes sistemas são uma solução muito utilizada para explorar dados em busca de conhecimento que agreguem valor aos seus usuários. Esse conhecimento pode ser obtido através de ferramenta OLAP(*Online Analytical Processing* e de técnicas de mineração de dados, as quais compõem uma aplicação conhecidas como *Business Intelligence*. Contudo, essas ferramentas, principalmente a OLAP, não são muito utilizadas na pesquisa biomédica devido a natureza dos dados e a complexidade dos processos desta área de estudos. Entretanto, a necessidade de ferramentas desta natureza para análise dos dados biológicos é grande. Assim, neste trabalho, foi desenvolvido o **BioBI**, um sistema *web* de apoio a decisão que integra diferentes ferramentas de análise de dados biológicos e clínicos. Neste trabalho, ele foi utilizado para realizar a análise de dados de pacientes acometidos pela doença **Paracoccidioidomicose** através de técnicas de mineração de dados e análises OLAP. A análise inicial dos dados produziu informações novas que serão revistas pelos médicos especialistas e fornecerão novas pistas na caracterização dos dados sobre a doença. Dada a necessidade crescente de ferramentas desta natureza, esperamos contribuir com o sistema **BioBI** para melhorar os LIMS existentes e facilitar o processo de obtenção de conhecimento na áreas biomédicas.

Palavras-chave: Bioinformática, Armazém de Dados, Mineração de dados, Paracoc-

cidioidomicose.

Abstract

The modern processes on biomedical research produce a large volume of different data from a variety of equipments. The management of these data and the tracking of experiments requires management systems to assist on these activities, known as Laboratory Information Management Systems or LIMS. These computational systems are used to manage data produced in laboratories with emphasis in quality. However, these tools are not enough to analyze and provide information about the data to the user and other systems and specific techniques are necessary to perform these tasks efficiently. Business Inteligence systems are tools designed to achieve this goal. These systems are a solution used to explore data in order to obtain knowledge that presents value to the users. These knowledge may be obtained by using the tool OLAP (Online Analytical Processing) and techniques of data mining that are part of the Business Inteligence applications. These tools, specially OLAP, have not been used frequently in biomedical research, because of the nature of the data and the complexity of the processes in this research field. Nevertheless, the need of tools of this nature to analyze biomedical data is significant. Thereby, in this work, we have developed BioBI, a business inteligence web system that integrates different tools for biological and clinical data analysis. BioBI was used to perform the analysis of data from patients of the disease Paracoccidioidomicosis using data mining techniques and OLAP analysis. The initial analysis has generated new information that will be reviewed by medical specialists and will help in the characterization of the data regarding the disease. Developing the BioBI system we hope to contribute to improve the existing LIMS and to make easier the process of getting knowledge in the biomedical area.

Keywords: Bioinformatics, Data Warehouse, Data Mining, Paracoccidioidomicose.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Sistemas de Apoio a Decisão

O avanço das pesquisas nas áreas biomédicas provocou um grande aumento na geração de dados resultantes da análise clínica e de exames realizados. É um desafio gerenciar e analisar os dados produzidos devido ao seu grande volume e complexidade, tornando-se necessário a utilização de ferramentas computacionais para auxiliar nestas tarefas. Os Sistemas de Gerenciamento de Laboratório ou *Laboratory Information Management System* (LIMS) surgiram para suprir esta necessidade e auxiliar os laboratórios a organizar os dados. Os LIMS são sistemas para integração e gerenciamento de dados de laboratório com ênfase em melhoria na qualidade de dados e geração de resultados de forma consistente e eficiente[Hinton, 1995]. Além disso, o LIMS realiza o controle das atividades e armazena os dados de todo o processo experimental de um laboratório (calibração de equipamentos, amostra experimental utilizada, fornecedor de material, resultados, etc). Contudo, a utilização de um LIMS apenas não é suficiente, uma vez que estes não são especializados em análise de uma massa de dados. Existem ferramentas específicas para esse propósito, que são os sistemas de apoio a decisão ou *Business Intelligence (BI)*. Estes sistemas utilizam uma abordagem diferenciada de armazenamento de dados e técnicas de mineração de dados e análise *ad-hoc* que permitem realizar correlações entre as variáveis e encontrar novas informações úteis.

Os sistemas de apoio a decisão são bem aceitos e utilizados no mundo corporativo. Contudo, seu uso nas áreas de pesquisa, principalmente biomédicas, ainda é restrito. Enquanto que os processos de negócios nas corporações são bem definidos e relativamente simples, os processos biológicos possuem uma metodologia complexa e variada. Além disso, os dados necessários para análise estão, muitas das vezes, distribuídos em diversas fontes externas (banco de dados públicos, artigos, patentes, entre outros) e

são produzidos por diferentes equipamentos, os quais armazenam os dados de maneira particular. Assim, torna-se um grande desafio realizar a coleta e a integração desses dados biológicos e principalmente, analisar esses tipos complexos de dados.

Da mesma maneira que os sistemas de apoio a decisão auxiliam as corporações na tomada de decisões, eles também contribuem para a obtenção de novas informações relevantes na área da pesquisa biomédica. Com o uso de ferramentas computacionais é possível identificar que os médicos não estão seguindo corretamente o protocolo de anamnése (etapa inicial que o profissional de saúde coleta informações históricas do paciente para iniciar o diagnóstico de uma doença) ou estão falhando no preenchimento de prontuários. Além disso, é possível com estas ferramentas identificar variáveis clínicas críticas na reincidência de uma doença, auxiliando consideravelmente na pesquisa biomédica.

1.2 Ferramentas

Várias ferramentas estão disponíveis atualmente para uma análise de dados específica. Algumas são direcionadas para análise multidimensional, outras para mineração de dados e outras ainda para testes estatísticos. Isso faz com que os pesquisadores utilizem diferentes ferramentas, em diferentes plataformas (*web* ou *desktop*), para obter todas as informações relevantes. Neste trabalho foram utilizadas três ferramentas para exploração dos dados: Mondrian, WEKA e R.

O Mondrian[Mondrian, 2012] é uma ferramenta que permite visualizar os dados por diferentes perspectivas e níveis de detalhamento. Um exemplo da utilização do Mondrian seria avaliar a média da idade dos pacientes portadores da doença por localização geográfica e para cada região ou identificar qual a média da idade dos pacientes que tiveram raio x alterado e não alterado. A ferramenta permite que o usuário altere essa consulta de forma interativa e de maneira simples.

O WEKA - *Waikato Environment for Knowledge Analysis*[Hall et al., 2009] é um sistema que agrupa diversas técnicas para a descoberta de conhecimento, tais como seleção de atributos, agrupamentos e classificação. Com o uso das técnicas existentes no WEKA, é possivel que pesquisadores encontrem atributos clínicos relevantes que podem indicar o momento de interrupção de um tratamento médico, por exemplo, ou identifiquem padrões em variáveis clínicas que possam servir de indicadores de rescidivas da doença.

O uso de ferramentas estatísticas é fundamental para análise dos dados. O pacote R[R, 2012] é um ambiente de programação para análise estatística e criação de gráficos.

Através de *scripts* podem ser realizados diversos testes estatísticos, análise temporal, cálculos e operações em vetores e matrizes, entre outros.

Para realizar a análise dos dados é de grande interesse que as ferramentas de análise estejam agregadas em um sistema único e se possível associadas a um LIMS que possa ser utilizado para a coleta de dados e exibição dos resultados. O sistema SIGLa (Sistema Integrado de Gerência de Laboratórios) é um LIMS baseado em *workflows* que se adapta a diversos laboratórios visando a melhoria na qualidade da entrada dos dados [SIMOES et al., 2010]. Essa ferramenta facilita a implantação de um sistema de apoio a decisão para laboratórios visto que os dados estarão organizados, preenchidos com os valores corretos e concistentes com o fluxo de trabalho definido. Portanto, o SIGLa foi o LIMS de escolha para implementação dos sistema de apoio de decisão para análise de dados biológicos.

1.3 Estudo de Caso — Paracoccidioidomicose — PCM

A paracoccidioidomicose (PCM) é uma micose tipicamente brasileira, causada pelo fungo *Paracoccidioides brasiliensis*[RESTREPO et al., 2001]. A inalação do fungo faz com que o agente infeccioso atinja o epitélio pulmonar e a não eliminação deste possibilita a disseminação do fungo para outros orgãos e tecidos. A análise clínica dos pacientes portadores do fungo segue um protocolo definido pelo Centro de Treinamento e Referência de Doenças Infecto-Parasitárias - CTR-DIP, Anexo Orestes Diniz, do Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG). Esse protocolo possui muitas variáveis clínicas que são avaliadas pelos médicos em cada consulta assim como exames radiológicos e sorológicos, os quais também são realizados no acompanhamento da evolução da doença. Entretanto, na maioria das vezes, estes exames não estão disponíveis na rede pública de saúde.

Além disso, o momento da interrupção do tratamento ainda é um desafio para os médicos, visto que ainda não há um parâmetro clínico que seja confiável para a determinação desse momento ou de predição das reincidências da doença. Técnicas de análise de dados, como a mineração de dados, podem ser bastante úteis para auxiliar a identificar as variáveis críticas que determinam o momento de interrupção ou de predizer uma possível reincidência da doença.

1.4 BioBI

O sistema **BioBI** foi desenvolvido com o objetivo de criar um sistema de apoio a decisão onde pesquisadores possam ter diferentes ferramentas de análise em um único sistema acessível via *web*. Ele integra Mondrian, WEKA e R para que os pesquisadores explorem seus dados em busca de novas informações relevantes e utiliza o SIGLa para obter os dados a partir do fluxo de trabalho dos laboratórios.

O objetivo final do **BioBI** é ser uma ferramenta que integra diferentes técnicas e programas disponíveis de análise de dados em uma única plataforma *web* utilizada por pesquisadores para análise e extração de informação de dados clínicos e biológicos. Neste trabalho, ele será utilizado para caracterização dos dados de pacientes que sofrem com a doença paracoccidioidomicose. Com o **BioBI** espera-se contribuir para agilizar o processo de obtenção de conhecimento na análise de dados sobre pacientes de PCM e que o modelo possa ser extendido para outras análises de dados de outras doenças.

1.5 Resultados

Utilizou-se o **BioBI** para realizar uma caracterização da base de dados de pacientes portadores da PCM. Inicialmente, foi realizada uma analise da frequência de distribuição de valores para cada um dos atributos da base de dados. Observou-se que uma grande quantidade de atributos não são avaliados na anamnése clínica ou os médicos não estão preenchendo todas as informações na ficha do paciente.

A dimensionalidade de análise foi reduzida utilizando uma das técnicas de mineração de dados para identificar o conjunto de atributos mais relevante para uma determinada classe. Com as variáveis selecionadas, realizou-se uma análise de agrupamento para identificar alguns padrões no conjunto de dados. Foi detectada uma falha na execução do protocolo de primeira consulta: o exame de raio x é uma análise obrigatória a ser feita pelos médicos, contudo, 40% dos pacientes da base de dados não tiveram o resultado do exame mencionado no protocolo.

Outro fator evidenciado através da análise de grupos e detalhado pela análise multidimensional é a relação entre o tempo médio da evolução dos pacientes com tabagismo prévio. Pode-se perceber que os pacientes que foram tabagistas apresentaram um tempo de evolução muito maior aos não tabagistas. Essa informação, dentre outras identificadas, serão revistas pelos especialistas para que possam ser utilizadas no estudo da doença.

1.6 Organização do texto

Esta dissertação está organizada da seguinte forma: O Capítulo 1 introduz o problema e discute brevemente a motivação para a realização do trabalho. O Capítulo 2 apresenta o modelo dimensional e suas características, as técnicas de mineração de dados e descreve as operações de consulta multidimensionais. O Capítulo 3 aborda a implementação do **BioBI** e a sua arquitetura. Neste capítulo são também descritas as ferramentas de análise de dados utilizadas e a maneira que como foi feita a integração de cada uma delas. O Capítulo 4 apresenta o estudo de caso e o conjunto de dados utilizados para análise e as etapas de implantação do **BioBI**. Além disso, aborda a metodologia de análise e os resultados obtidos. Por fim, as conclusões e os trabalhos futuros são descritos no Capítulo 5.

Capítulo 2

Sistemas de Apoio a Decisão

Uma grande quantidade de dados é gerada e armazenada em sistemas computacionais. Nas empresas, muitos dos dados gerados são essenciais para o gerenciamento do dia a dia do negócio. Entretanto, estes sistemas não provêem boas ferramentas de análises de dados e a informação obtida não é detalhada e consolidada. Além disso, a rápida geração, coleta e armazenamento dos dados fazem com que as análises manuais tornam-se inviáveis sem a utilização de ferramentas específicas. Com isso, as decisões muitas vezes eram tomadas pela intuição, devido à ausência de uma informação rica e confiável, o que é inaceitável nos dias atuais.

O surgimento das ferramentas de apoio a decisão permitiu que as informações "escondidas" no grande volume de dados fossem descobertas e utilizadas para a tomada de decisões. Um exemplo disso é a identificação de padrões de compras de consumidores em um supermercado - identificar que há um grupo de consumidores que compra fraldas no período da noite e também compra cerveja; isto auxilia o departamento de marketing a direcionar uma campanha de publicidade para um determinado perfil consumidor.

A maneira como os sistemas de informação estruturam e organizam os dados armazenados não é apropriada para a realização de análises elaboradas. Além disso, os dados a serem analisados podem ser originados de diferentes fontes, sendo necessário integrá-los em uma base de dados centralizada. Essa base de dados centralizada e estruturada de maneira que otimize as consultas é chamada de armazém de dados (seção 2.1). Para realizar a integração dos dados de diferentes fontes, é necessario realizar, inicialmente, o processo de ETL - *Extraction*, *Transformation* e *Load* (seção 2.1.2), o qual seleciona os repositórios e realiza a limpeza dos dados eliminando ruídos e duplicações de registros.

O processo de ETL é uma das etapas críticas do processo de implantação de um sis-

tema de apoio a decisão. Como os dados a serem analisados podem ser originados em diferentes fontes e cada uma delas possui uma estrutura própria de armazenamento dos dados, muitos desafios surgem no momento de integrá-los em um mesmo repositório. Um dos desafios dessa fase é identificar os registros armazenados com valores distintos que possuem o mesmo valor semântico. Por exemplo, duas bases de dados distintas podem armazenar o registro de uma mesma cidade de formas diferentes, como "Belo Horizonte", "BH", "BHZTE". Na base centralizada, estes dados devem ser tratados de maneira que o valor armazenado seja apenas um. Isso mantém os dados consistentes e sem redundância.

Todas essas etapas são importantes para implantar um sistema de apoio a decisão. Elas são responsáveis pela coerência dos dados e por fazer com que estes estejam armazenados de maneira adequada e integrada para análises futuras. Além disso, a estrutura de armazenamento diferenciada permite que as consultas e processo de obtenção de novas informações úteis e relevantes tenham uma melhor performance.

Neste capítulo são apresentadas as ferramentas e técnicas que tornam viáveis a extração de informação e conhecimento para facilitar o processo de análise de dados. Um modelo de armazenamento de dados que facilita o processo de extração da informação é apresentado na seção 2.1. Nas seções 2.2 e 2.3 são apresentadas ferramentas e técnicas que auxiliam os analistas na aquisição de informações que podem ser aplicadas no seu domínio de atuação.

2.1 Armazém de Dados

Com o aumento da demanda por informações sobre uma base dados, surgiram os sistemas de apoio a decisão com ferramentas apropriadas para manipularem os dados de maneira a extrair destes informação útil. Essas ferramentas utilizam uma estrutura de armazenamento de dados, os Armazéns de Dados (*Data Warehouse*), que favorece a geração de relatórios, o processamento de análises (*Online Analytical Processing* - OLAP) e a mineração de dados em grande volume. Os Armazéns de Dados caracterizam-se por integrar e consolidar os dados de diversas fontes disponíveis (sistemas legados, arquivos textos, planilhas, páginas da *web*) de maneira a tornar os dados armazenados acessíveis para análise.

O processo de integrar dados de diferentes fontes demanda um esforço considerável na implantação de um armazém de dados visto que é necessário conhecer como os dados são armazenados e estruturados em cada repositório para que eles possam ser combinados. Também é necessário determinar como um registro de uma base pode

ser encontrado em outra base, além de integrá-los de forma consistente e não redundante. Um dos grandes desafios neste processo de integração de dados é a estratégia de normalização de nomes, ou seja, identificar registros com valores distintos mas que possuam a mesma semântica.

2.1.1 Modelo Dimensional

O esquema convencional de armazenamento de dados, o modelo relacional, evita que os dados sejam inseridos de maneira redundante. Este modelo normalizado tem benefícios para a construção de sistemas de informação que realizam muitas transações (inserções, atualizações e remoções) tais como o armazenamento consistente dos dados, a diminuição de ruídos e erros na entrada dos registros, e a facilidade em atualização das informações além de representar como é o fluxo de informação no sistema. A eliminação da redundância é realizada criando uma nova tabela no banco de dados para armazenar os dados redundantes, por exemplo: ao preencher a informação de naturalidade do paciente, existem dezenas de registros com o mesmo valor, tal como "Belo Horizonte". Para eliminar a duplicação deste dado, pode-se criar uma tabela "cidade" com o registro "Belo Horizonte" e, na tabela "paciente" no atributo "naturalidade" armazena apenas a referência para o registro "Belo Horizonte". Entretanto, ao manipular uma grande massa de dados para consulta, este modelo torna-se complexo e ineficiente para recuperar os dados pois é necessário realizar operações de junções em diferentes tabelas no banco.

O Modelo dimensional, também conhecido como esquema estrela pela forma em que as tabelas ficam dispostas no diagrama, é preparado para realizar consultas mais intuitivas e com bom desempenho ao acesso aos dados [Kimball & Ross, 2002]. Isso ocorre devido a maneira como os dados são organizados, onde há uma tabela central, chamada de tabela de fatos (seção 2.1.1.2), que se relaciona com as outras tabelas, conhecidas como tabela de dimensões (seção 2.1.1.1). Estas se conectam com a tabela de fatos através de um relacionamento, o que facilita a associação entre os dados e consequentemente a obtenção dos resultados e elaboração de consultas mais simples. As ferramentas de análise que permitem a navegação e agregação dos dados dependem dessa estrutura de armazenamento.

A figura 2.1 é um exemplo de um modelo dimensional que representa uma base de dados de pacientes de um hospital. Cada registro na tabela de fatos representa uma consulta clínica associada a um paciente, uma forma de diagnóstico, localização geográfica e data de ocorrência. Uma das medidas (métricas) que poderia existir na tabela de fatos seria o peso e a idade do paciente no momento da consulta. Este mo-

delo de dados se diferencia do modelo relacional, apresentado na figura 2.2, no qual a idade do paciente seria um atributo calculado entre a data da consulta e a data de nascimento. Contudo, essa operação tem um custo que é evitado no esquema dimensional para otimizar as consultas.

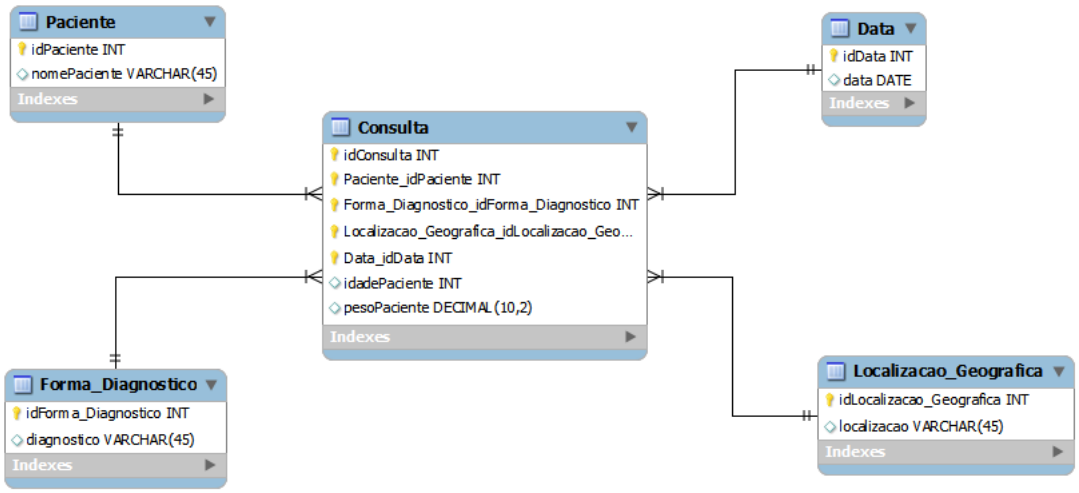


Figura 2.1. Exemplo de Modelo Dimensional.

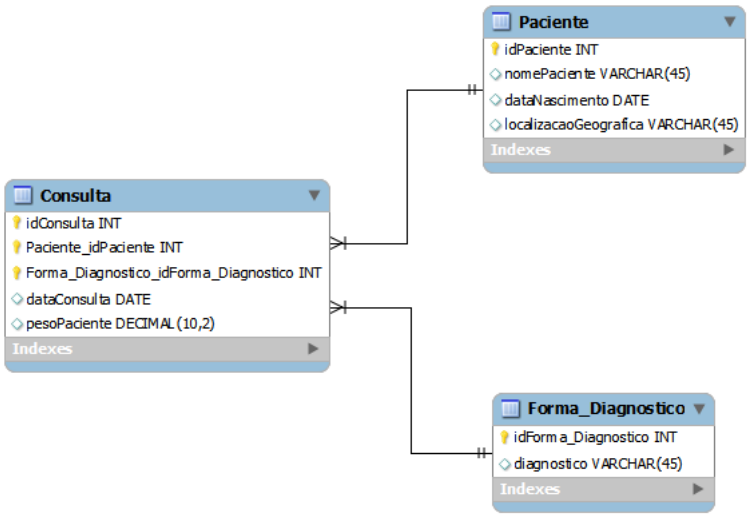


Figura 2.2. Exemplo de Modelo Relacional tradicional.

Com o modelo dimensional dos dados, pode-se analisar a contagem de consultas realizadas para cada dimensão e o cruzamento entre elas, como por exemplo, quantas consultas foram realizadas em pacientes do sexo feminino, sendo que a forma de diagnóstico foi a necropsia. Além disso, pode-se analisar a média de idade dos pacientes observados para cada forma de diagnóstico, conforme apresentado na figura 2.3.

DATASET SELECTED: PARACOCCIDIOIDOMICOSE

		Measures	
SEXO	FORMA DIAGNOSTICO	Fact Count	MEDIA IDADE 1 CONSULTA
All SEXOs	All FORMA DIAGNOSTICOs	227	36,22
feminino	All FORMA DIAGNOSTICOs	41	26,78
	biópsia	34	27,88
	clínico	2	27,
	cultura	1	16,
	exame micológico a fresco	3	20,
	necropsia	1	20,
masculino	All FORMA DIAGNOSTICOs	183	38,94
	biópsia	136	38,99
	clínico	7	29,71
	combinação	1	44,
	cultura	1	46,
	ensaio imunológico	3	40,33
	esfregaço citológico	2	26,5
	exame micológico a fresco	14	40,14
	necropsia	1	82,
	NI	18	39,28

Figura 2.3. Exemplo de uma consulta multidimensional.

2.1.1.1 Dimensões

As dimensões são tabelas independentes entre si que guardam os dados textuais que caracterizam um registro na tabela e que podem ser utilizados para as consultas. As dimensões são compostas de níveis hierárquicos e podem ter informações mais detalhadas ou sumarizadas através das funções OLAP de *Drill-Down* e *Roll-Up* a serem discutidos posteriormente. Por exemplo, em uma dimensão "Localidade" pode-se ter a representação dos níveis "País", "Estado" e "Cidade" sendo que, o usuário pode definir o nível de detalhe que deseja através das consultas dinâmicas[Kimball & Ross, 2002]. A figura 2.4 representa o conceito de hierarquia utilizando como exemplo a localidade do paciente. Ela demonstra como os dados podem ser detalhados ou sumarizados através das hierarquias de uma dimensão. No nível mais sumarizado, sem detalhamento, temos o total de todos os pacientes: 230. No próximo nível de detalhe — país — tem-se o total de pacientes para "Brasil" e "Chile". Detalhando os pacientes brasileiros, é apresentada a quantidade de pacientes por estado e por último, no maior nível de detalhe, são apresentados os pacientes por municípios.

As hierarquias podem ser implementadas utilizando o esquema estrela tradicional, com as dimensões não normalizadas e com redundância dos dados. Ou seja, para cada registro armazenado na dimensão, os dados textuais de "País", "Estado" e "Cidade" são armazenados para cada registro na tabela dimensão[Kimball & Ross, 2002]. A figura 2.5 apresenta um modelo físico de um armazém de dados onde a dimensão Localidade possui essa abordagem de estrutura para representar as hierarquias.

O relacionamento entre as dimensões ocorre na tabela de fatos, onde um registro

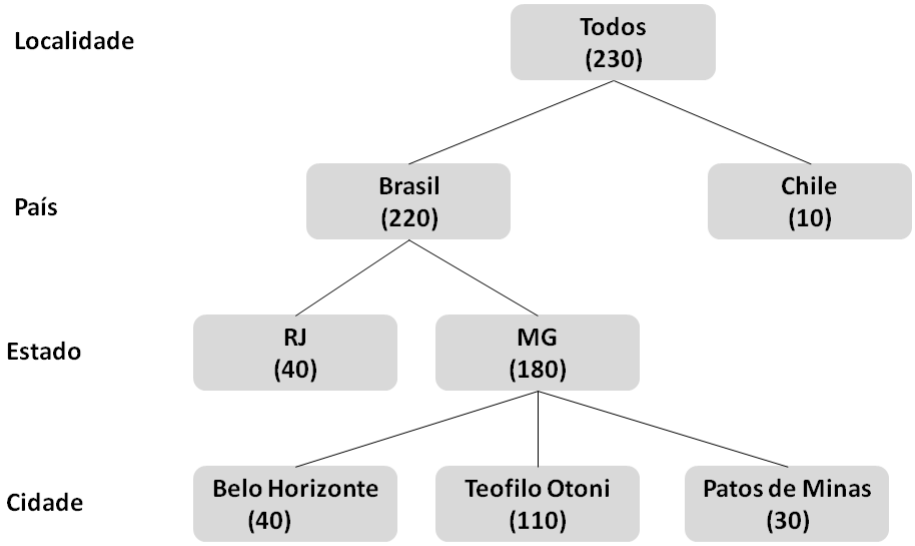


Figura 2.4. Conceito de hierarquias para a dimensão Localidade.

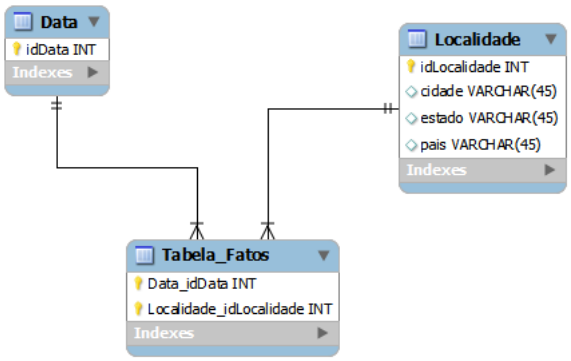


Figura 2.5. Modelo dimensional com as hierarquias não normalizadas.

nessa tabela tem referência para as dimensões no modelo. Com essa estrutura é possível que o analista realize o cruzamento desses dados.

2.1.1.2 Fatos

A tabela de fatos é composta por chaves estrangeiras, que referenciam as tabelas de dimensões, e por campos numéricos. Isso significa que cada registro na tabela de fatos possui métricas (medidas), que são valores numéricos que podem ser calculadas utilizando funções de agregação (soma, média, mediana, entre outras) de acordo com

o nível de detalhe definido[Han & Kamber, 2001]. A tabela de fatos também possui referências para as dimensões associadas a ela. O relacionamento entre as dimensões e as métricas da tabela de fatos é responsável por dar um valor semântico ao registro armazenado[Kimball & Ross, 2002]. Para isso, é importante definir o que será analisado e o seu nível de detalhe, ou seja, é necessário definir a granularidade da tabela de fatos, onde o grão descreve o que representa um item na tabela de fatos.
A granularidade diz respeito ao nível de detalhe das unidades do armazém de dados, sendo que, quanto maior o nível de detalhe, menor será a granularidade dos registros e consequentemente, um maior espaço de armazenamento será utilizado. Nas figuras a seguir temos dois exemplos de um armazém de dados para analisar as vendas de uma loja. Em - 2.6 o grão da tabela de fatos é o total da venda de uma loja para um cliente, enquanto que na figura 2.7 o grão armazenado é o produto vendido para um determinado cliente. Percebe-se que no segundo armazém de dados, onde o grão da tabela de fatos é menor (cada produto vendido por cliente), há uma possibilidade maior de análise sobre os dados em relação ao primeiro (total de uma venda por cliente). Um exemplo de análise que pode ser feita no armazém de menor grão (e não é possível na outra abordagem) é identificar quais os produtos um determinado cliente mais compra. O interessante dessa análise é que permite ao comerciante fazer um preço diferenciado em um produto para um determinado cliente utilizando o seu histórico de compras.
É necessário ter cuidado para definir a granularidade pois quanto menor for o grão maior o volume de dados que deve ser armazenado, coletado e integrado. Assim, a escolha do grão deve ser feita considerando a complexidade para se obter as informações úteis e evitar a sobrecarga de armazenamento e processamento de dados desnecessários.

2.1.2 Integração

Uma das características de um armazém de dados é centralizar e integrar dados disponíveis em diversas fontes tais como sistemas transacionais, legados, arquivos textos, planilhas e inclusive, dados disponíveis na *web*. Para tanto, é necessário conhecer a forma como os dados são armazenados nos repositórios de origem e identificar formas de integrá-los de maneira consistente. Esta é uma etapa crítica para a implantação de armazém de dados, o qual armazena todos os dados que serão analisados de forma homogênea em um repositório único e integrados. Esta etapa é conhecida como ETL, do inglês *Extraction*, *Transformation* e *Loading*.

A figura 2.8 apresenta as etapas do processo ETL, o qual possui complexidade considerável, visto que consiste na captação e estruturação dos dados que serão analisados. É nessa fase também que se conhece a qualidade e condições dos dados arma-

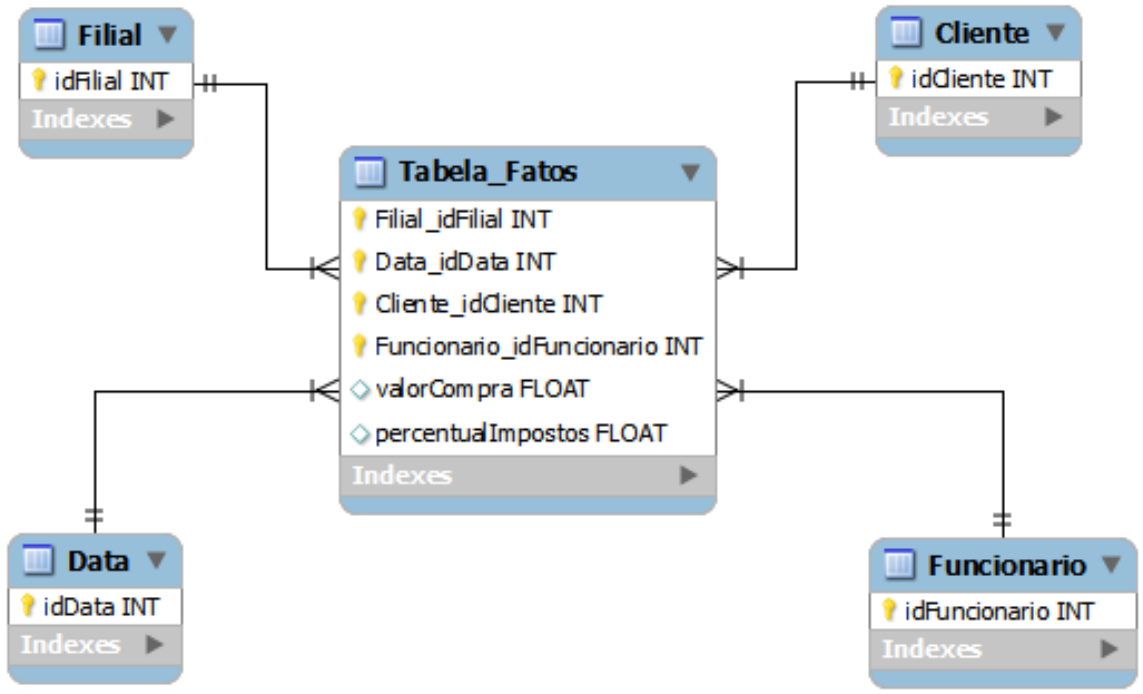


Figura 2.6. Exemplo de modelo em que o grão da tabela de fatos representa uma venda.

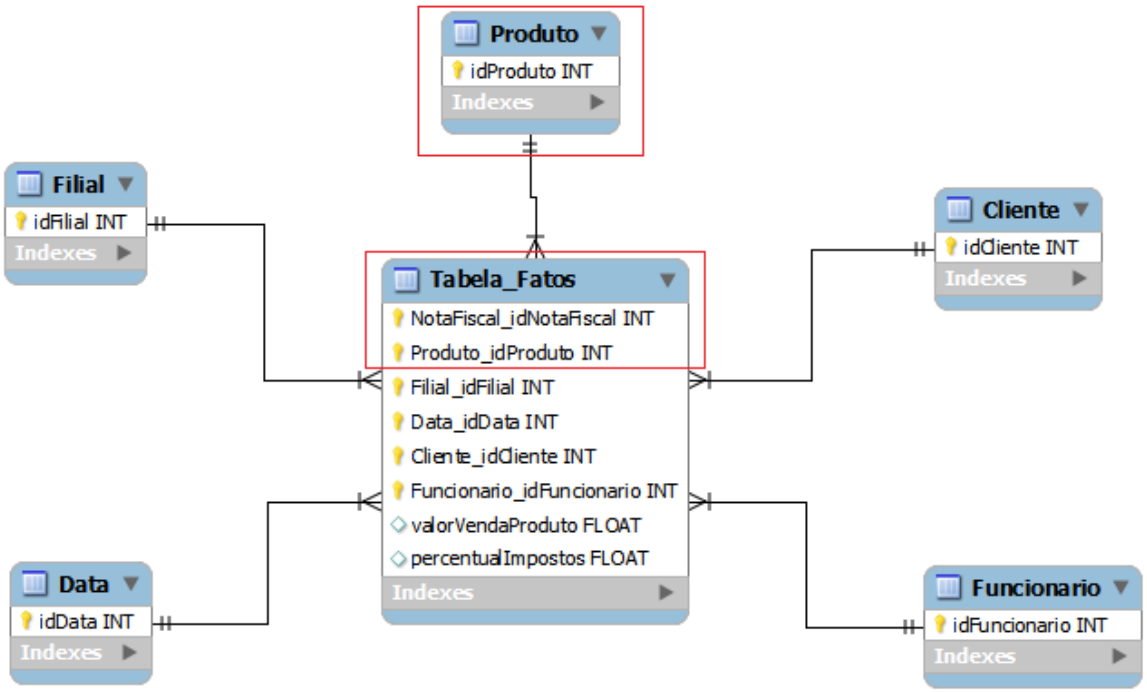


Figura 2.7. Exemplo de modelo em que o grão da tabela de fatos representa um produto vendido.

Figura 2.8. Etapas do processo ETL — Extração, Transformação e Carga dos dados.

zenados nos bancos disponíveis, e se necessário, realiza-se transformações específicas antes de armazená-los no armazém de dados. Cada etapa do ETL é descrita a seguir.

Extração

Esta é a primeira etapa do processo de movimentação de dados para um armazém de dados. A extração é o momento onde há o entendimento das fontes de dados e a seleção dos dados que serão utilizados na transformação.

Transformação

Como os dados extraídos podem ser originados em diferentes fontes, algumas operações de transformações devem ser realizadas com intuito de adequar os dados para a inserção no armazém. A limpeza do dados é uma transformação essencial que deve ser realizada sobre os dados. Ela consiste em corrigir valores ausentes, duplicados ou errados e na formatação de valores. O objetivo é padronizar os valores inseridos de maneira não duplicada, correta e consistente. A tradução de valores é outra transformação necessária para a padronização dos registros armazenados. Por exemplo, para o atributo "sexo", esta etapa transforma o valor "1", ou "M" para um valor padrão, "Masculino", que será inserido no armazém de dados.
A operação de tradução de valores pode ser mais complexa, sendo necessária a utilização de técnicas mais elaboradas para realizar a padronização de nomes. Um exemplo é a identificação de uma mesma cidade escrita com grafias diferentes. Em um sistema fonte a cidade de "Belo Horizonte" é registrada como "Belo Horizonte" enquanto que em outro sistema possui apenas a abreviação "BH" ou "BHZTE". No armazém de dados esses dois registros precisam ser identificados e armazenados em apenas um. Outro exemplo é identificar que a pessoa "Maria José Esteves" de uma base é a mesma que "Mªjosé Esteves" de outro banco de dados. Estes dois exemplos mostram como é desafiadora e complexa a fase de transformação dos dados.

Carga

Esta é a fase onde os dados são inseridos no armazém de dados, na qual se deve atentar para aspectos importantes de integridade, tipo de carga realizada e outras tarefas importantes da manutenção dos dados. A integridade referencial dos dados deve ser respeitada no momento da carga dos dados no armazém. Ou seja, é necessário que esta atividade verifique se as chaves estrangeiras existem nas respectivas tabelas em que são chaves primárias.
A carga dos dados pode ser realizada de forma incremental ou total. A primeira mantém os dados antigos e apenas adiciona ao armazém de dados os novos itens não existentes desde a última carga. A segunda estratégia exclui os dados armazenados e os inclui novamente em seguida. A escolha da estratégia é uma decisão de projeto que deve ser analisada e definida para cada armazém de dados. Além disso, o aspecto temporal da carga — de quanto em quanto tempo ela é realizada — é definido de acordo com a necessidade de atualização da análise dos dados. Como exemplo, existem negócios em que a carga precisa ser realizada a cada 10 minutos enquanto em outros esta pode ser realizado apenas uma vez por dia.

2.2 Consultas OLAP

No modelo dimensional os dados são organizados em dimensões que podem conter diversos níveis de hierarquias dos dados. Essa estrutura de armazenamento permite que sejam realizadas consultas flexíveis chamadas de OLAP (*Online Analytical Processing*), sendo que os resultados destas são apresentados em formatos gráficos e tabelas. Os sistemas OLAP permitem que os analistas tenham uma visão multidimensional dos dados. Ou seja, pode-se analisar os dados em diferentes perspectivas de maneira mais fácil e intuitiva. A figura 2.9 é o resultado de uma consulta multidimensional que analisa a média da idade dos pacientes discriminada por sexo e por forma de diagnóstico.

Uma das características das ferramentas OLAP é permitir que os analistas "naveguem" sobre os dados sintetizando informações dos mesmos, realizando comparações e análises históricas. Para isso, elas oferecem algumas funções específicas que podem ser utilizadas pelos usuários (Figura 2.10):

- *Pivot*: é a rotação dos eixos para apresentar uma alternativa de visualização dos resultados.Por exemplo, em uma apresentação tabular a operação trocaria as linhas por colunas, ou mover uma das dimensões da linha para coluna.

		Measures	
SEXO	FORMA DIAGNOSTICO	Fact Count	MEDIA IDADE 1 CONSULTA
All SEXOs	All FORMA DIAGNOSTICOs	227	36,22
feminino	All FORMA DIAGNOSTICOs	41	26,78
	biópsia	34	27,88
	clínico	2	27,
	cultura	1	16,
	exame micológico a fresco	3	20,
	necropsia	1	20,
masculino	All FORMA DIAGNOSTICOs	183	38,94
	biópsia	136	38,99
	clínico	7	29,71
	combinação	1	44,
	cultura	1	46,
	ensaio imunológico	3	40,33
	esfregaço citológico	2	26,5
	exame micológico a fresco	14	40,14
	necropsia	1	82,
	NI	18	39,28

Figura 2.9. Exemplo de uma consulta multidimensional.

- *Slice* e *Dice*: *Slice* é a operação de selecionar uma dimensão do cubo, originando, assim, um subcubo para análise. Um exemplo desta operação é definir um subcubo de todos os eventos clínicos que ocorreram no ano de 2011. A operação *dice* é similar a *slice* porém, com a seleção de duas ou mais dimensões.

- *Roll-up / Drill-up*: A operação de *roll-up* realiza agregação dos dados do cubo. A sumarização dos dados podem ser realizados por subir o nível de hierarquia nas dimensões ou por redução de dimensão no cubo. Um exemplo para o primeiro caso é apresentar os dados de venda por estado, ao invés por cidade, a qual sumariza as informações da hierarquia mais baixa. Para o o segundo caso, considera o cubo que apresenta as vendas pela dimensão tempo e localidade. Ao remover a dimensão tempo do cubo, agregam-se os dados de total de vendas apenas pela localidade, ao invés de tempo e localidade.

- *Drill-down*: Esta operação realiza o contrário da operação *Roll-up*. Ela realiza uma navegação para o nível mais baixo de hierarquia ou adicionando novas dimensões no cubo, apresentando a informação de forma mais detalhada e menos sumarizada, como por exemplo, analisar os dados de venda no nível de cidade ao invés de estado.

- *Drill-across* é o processo de unir duas ou mais tabelas de fatos de mesmo nível de detalhes. A operação *drill-across* executa consultas envolvendo mais de um cubo. Esse pode ser o caso, por exemplo, se quisermos comparar dados de venda

de um cubo com dados de aquisições de outro cubo. Essa operação requer que os dois cubos tenham pelo menos uma dimensão em comum.

As ferramentas OLAPs podem usar diferentes estratégias de implementação, sendo que as duas principais abordagens são a ROLAP (OLAP Relacional) e a MOLAP (OLAP Multidimensional). A primeira estratégia utiliza banco de dados relacionais para o armazenamento do armazém de dados (utilizando o esquema estrela ou floco de neve), sendo que, o servidor OLAP fica em uma camada intermediária entre o banco de dados físico e o cliente. Já as ferramentas MOLAP utilizam estruturas proprietárias de armazenamento do cubo de dados com abordagens específicas de indexação e a sumarização de dados. Essa estratégia torna o processamento mais rápido além de permitir cálculos mais complexos[Han & Kamber, 2001].

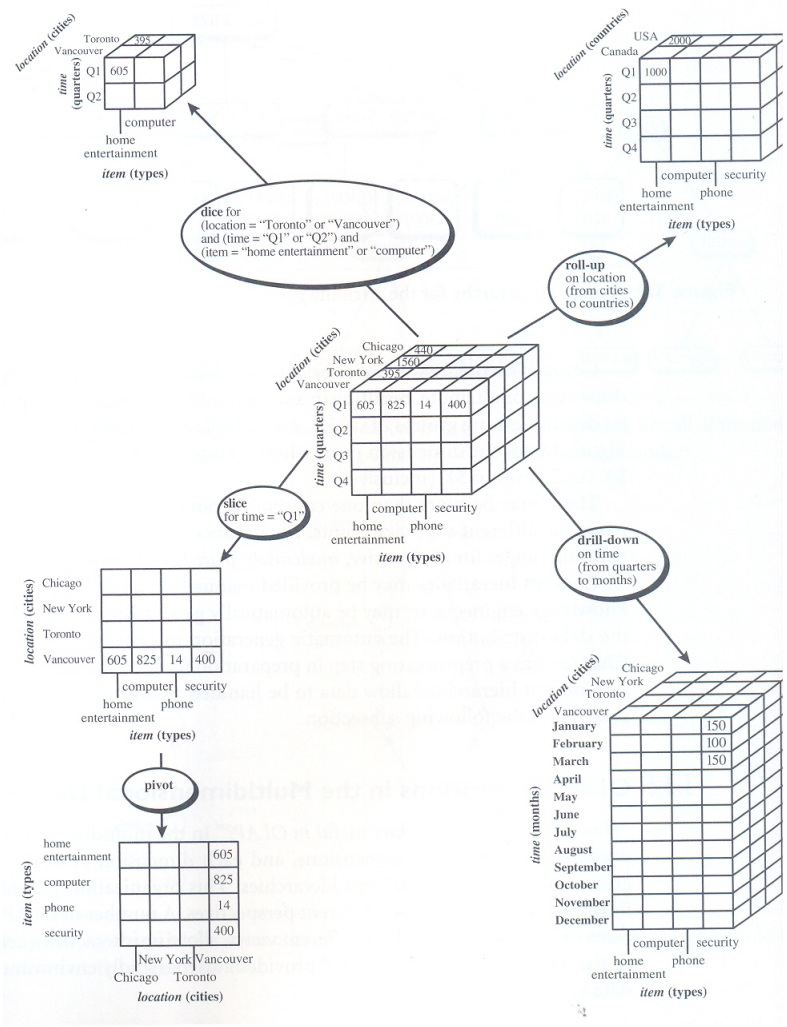


Figura 2.10. Exemplo de operações OLAP típicas em dados multidimensionais [Han & Kamber, 2001].

2.3 Mineração de Dados

Tem havido atualmente um aumento siginficativo na geração de dados seja pelos sistemas computacionais, através da internet ou por diversos equipamentos de análise. Contudo, esse grande volume de dados não tem sido sempre utilizado apropriadamente, de maneira a transformá-lo em conhecimento que possa ser aplicado no domínio de análise. A descoberta de padrões e/ou tendências pode ser de grande valia no processo de negócio de uma empresa, em experimentos científicos ou auxiliando médicos no entendimento de alguns fatores no tratamento e caracterização de uma doença.

O processo de descobrir padrões e correlações entre variáveis em uma massa de dados produzindo informações novas e úteis é chamado de mineração de dados. A mineração de dados utiliza algoritmos de classificação diversos para explorar os dados e encontrar padrões auxiliando na descoberta de conhecimento[Han & Kamber, 2001]. No mundo corporativo, um exemplo de aplicação da mineração de dados é identificar o perfil dos consumidores, revelando, por exemplo, que os consumidores noturnos que compram fraldas a noite também compram cerveja. Essa informação pode auxiliar o departamento de *marketing* de um supermercado a dirigir uma promoção especificamente para esse grupo. Na área da pesquisa biomédica, a mineração de dados pode também gerar informações interessantes, como por exemplo, identificar grupos de sintomas que caracterizam uma doença rara.

A mineração de dados é uma etapa do processo de extração de conhecimento. Para realizá-la, é necessário que os dados estejam organizados e estruturados de maneira tal que os algoritmos para executá-la tenham uma maior eficácia. Ou seja, é necessário que os dados passem por uma etapa de pré-processamento, como descrito na seção 2.1.2. Uma das técnicas para isto é a seleção de atributos que auxilia na redução das dimensões do modelo, a qual é muito utilizada antes da busca por padrões nos dados (seção 2.3.1). As técnicas de agrupamentos são igualmente importantes para que os analistas tenham uma visão do comportamento dos dados (seção 2.3.2).

2.3.1 Seleção de Atributos Relevantes

A grande quantidade de atributos dos dados afeta a performance dos algoritmos de mineração de dados além de dificultar a leitura dos resultados obtidos. Contudo, é possível remover alguns atributos destes, de maneira que os resultados obtidos sejam bem próximos daquele obtidos com o modelo que utiliza todos os atributos do conjunto de dados[Han & Kamber, 2001]. Uma das maneiras de realizar essa seleção seria um especialista avaliar os atributos e identificar quais destes são redundantes ou irrelevantes para mineração de dados. Entretanto, essa abordagem demandaria uma grande quantidade de tempo para determinar quais são os atributos importantes e um grande conhecimento prévio da base de dados. Existem algoritmos de seleção de atributos que realizam essa atividade de seleção dos atributos com maior ganho de informação.
A determinação de quais dimensões serão consideradas para a mineração de dados não é uma tarefa trivial. Para isso, há métodos para determinar os atributos irrelevantes ou fracamente relevantes [Han & Kamber, 2001; Guyon & Elisseeff, 2003; Yvan et al., 2007].

2.3.2 Agrupamentos - Clustering

A clusterização é uma boa ferramenta para separar objetos em grupos através de sua similaridade. Essa técnica agrupa objetos mais parecidos e coloca os demais em outros grupos. Essa segregação dos dados permite que os analistas tenham uma visão dos padrões existentes em uma base de dados desconhecida. Esta técnica é muito utilizada em diversas áreas, como por exemplo em biologia para definir qual é a taxonomia animal ou categorizar genes de funções similares. Na medicina ela pode ser aplicada para agrupar pacientes que possuem um conjunto de sintomas para uma determinada doença. A análise dos *clusters* ou grupos produzidos permite que os analistas tenham *insights* sobre a distribuição do dados e observar o comportamento de cada *cluster*. A partir disso, pode-se realizar uma pesquisa mais aprofundada sobre o grupo utilizando as análises OLAP e testes estatísticos.
O objetivo da clusterização é agrupar os dados de maneira a maximizar a similaridade intraclasse e minimizar a similaridade interclasse. Ou seja, instâncias de um mesmo grupo possuem grande semelhança entre elas enquanto instâncias de grupos distintos possuem pouca semelhança.

Uma das técnicas conhecidas para a obtenção dos grupos é o *k-means*. O *k-means* é um método de particionamento que define, iterativamente, qual o grupo mais apropriado para um determinado registro. Isso é feito calculando a similaridade do objeto em relação ao centróide dos grupos sendo que o *cluster* que tiver maior similaridade

será selecionado. O centróide de cada grupo é obtido através das médias dos pontos dos registros que pertencem ao grupo. Assim, para cada inserção de um item em um grupo, o centróide é recalculado. Este processo de movimentação dos registros entre os grupos é executado até que não haja mais mudanças[Han & Kamber, 2001; Witten.I et al., 2011].

A figura 2.11 apresenta as etapas do processo de agrupamento utilizando o *k-means*. A primeira etapa do algoritmo (Figura 2.11(a)) define aleatoriamente o centróide dos grupos. Utilizando a função de similaridade entre o registro e o centróide, define-se então a qual grupo ele pertence. O centróide do grupo é recalculado (Figura 2.11(c)). Este último processo se repete até que os itens do grupo atinjam a estabilidade (Figura 2.11(d)).

(a) Seleção inicial dos centróides

(b) Seleciona as instâncias mais próximas de cada centróide

(c) Redefine o centróide de cada grupo

(d) Finaliza as iterações quando atingir a estabilidade

Figura 2.11. Etapas do k-means

2.4 Comparação entre Ferramentas de Sistemas de Apoio a Decisão

Existem algumas ferramentas de Sistemas de Apoio a Decisão disponíveis gratuitamente para uso. Contudo, algumas delas são apenas ferramentas para geração de relatórios que se classificam como uma ferramenta de análise de dados, como por exemplo o BIRT[1]. Neste trabalho, foi realizado um estudo preliminar sobre as ferramentas disponíveis com intuito de selecionar uma para uso no estudo de caso. Após a avaliação preliminar foram selecionadas as ferrramentas Pentaho BI Server(Seção 2.4.1) e SpagoBI (2.4.2) por atender, em parte, as principais demandas de análises de dados deste projeto. Ainda assim, foi necessária a implementação da ferramenta desenvolvida neste projeto denominada **BioBI**, para atender a todas as necessidades como discutido na seção 2.4.3.

[1]http://www.eclipse.org/birt/phoenix/

2.4.1 Pentaho BI Server

O Pentaho Community Edition[Pentaho, 2012](versão gratuita da ferramente) é uma suíte de ferramentas que incluem análises, relatórios e *dashboards*. A ferramenta permite que os analistas realizam análises OLAP através do Mondrian (servidor OLAP que é apresentado na seção 3.3.1) e JPivot.

As empresas que mantém a suíte de ferramentas, disponibilizam outras ferramentas *desktops* independentes do Pentaho BI Server para tarefas como o ETL - através da ferramenta Pentaho Data Integration e elaboração de *templates* de relatórios, utilizando o Report Designer. Outra ferramenta adotada pela suíte de análise de dados é o Weka(detalhado na versão 3.3.2). Contudo, assim como essas últimas ferramentas, o analista precisa ter o programa instalado em seu computador, além da base de dados estar disponível em um formato de arquivo específico do programa, ou no formato CSV (Valores separados por virgula).

2.4.2 Spago BI

O SpagoBI[BI, 2012] é um sistema gratuito de Apoio a Decisão que integra diferentes ferramentas de análises de dados que possibilitam aos usuários escolher qual plataforma utilizar. A ferramenta é mantida por uma companhia italiana que adota a política "*Free and Open Source Software*" (FOSS) e disponibiliza apenas uma versão do produto, totalmente gratuita sob a licença GNU LGPL[LGPL, 2012]. Ou seja, não há uma versão paga que tem funcionalidades.

A ferramenta disponibiliza diversas opções para análises: análise multidimensional, utilizando o Mondrian/JPivot; mineração de dados, através da integração com o WEKA. Utilizando uma única ferramenta que pode ser acessível tmabém via *web*. Também são disponibilizadas outras facilidades, como ferramentas de ETL, elaboração de relatórios e *dashboards*.

2.4.3 Conclusão

Ambas as ferramentas apresentam uma série de funcionalidades que auxiliam na tomada de decisões, como por exemplo, análises multidimensionais e elaboração de relatórios. Além disso, ambas possuem controle de acesso e funcionalidades extras, como por exemplo o compartilhamento de análises elaboradas com outros usuários do sistema. O Pentaho possue uma interface mais elaborada e fácil de utilizar. O SpagoBI apresenta um acervo maior de funcionalidades integradas em um único sistema *web*.

Uma delas é a sua integração com o WEKA, que, por sua vez, ainda é limitada - estão disponíveis apenas dois algoritmos, ambos de agrupamentos (Tabela 2.1).

Neste cenário, seria necessário a customização de uma das ferramentas para que fosse incorporada a ela análises estatísticas e de mineração de dados. A customização para atender a essas necessidades, demandaria um grande período de tempo para entendimento da arquitetura desses programas. No caso do Pentaho, seria necessário incorporar o WEKA e o pacote R (*software estatístico*) enquanto que no caso do SpagoBI seria preciso customizar o módulo de mineração — para ter disponível pelo menos um algoritmo de seleção de atributos — e a incorporação do WEKA. Porém, a comunidade do SpagoBI não é muita ativa, está concentrada na França e Itália e não há muita documentação sobre o produto dificultando esta customização. Assim, decidiu-se pela implementação de uma ferramenta que possui as funcionalidades demandadas para análise de dados que são a análise multidimensional, técnicas de mineração de dados e integração com o pacote R.

Característica	**Pentaho**	**SpagoBI**
Análise Multidimensional	Mondrian - JPivot	Mondrian/JPivot
Mineração de dados	Não possui	2 algoritmos de agrupamento
Análise Estatística	Não possui	Não possui
Comunidade	Ativa	Pouco ativa e concentrada na França e Itália
Documentação	Sim	Pouco material encontrado

Tabela 2.1. Tabela comparativa entre as características e funcionalidades demandadas disponibilizadas pelo Pentaho BI Server e SpagoBI.

Capítulo 3

BioBI

BioBI é uma ferramenta desenvolvida para apoiar pesquisadores da área biomédica na análise de seus dados utilizando técnicas e ferramentas usualmente utilizada no mundo corporativo. Ela possibilita a realização de análises multidimensionais de forma iterativa, mineração de dados e testes estatísticos com a utilização de uma única ferramenta.

3.1 Visão Geral

BioBI é uma aplicação *web*, complementar ao Sistema Integrado de Gerência de Laboratórios (SIGLa), para análise de dados biológicos e clínicos. Ela integra um conjunto de programas de código aberto, tais como o Mondrian, WEKA e o R, para permitir aos pesquisadores o acesso a essas tecnologias de análise de dados através da *internet*. Ela foi desenvolvida utilizando a linguagem Java e o Apache Tomcat [Tomcat, 2012] como servidor *web* tornando possível a sua instalação em múltiplas plataformas e o acesso utilizando múltiplos navegadores. A ferramenta utiliza como sistema gerenciador de banco de dados (SGBD) o MySQL que é um SGBD gratuito [MySQL, 2012]. Entretanto, a **BioBI** está preparada para trabalhar com qualquer outro SGBD relacional, uma vez que ela utiliza o *framework* Hibernate para realizar as operações no banco de dados ou realiza estas através de conexão JDBC - Java Database Connectivity [Hibernate, 2012].

A **BioBI** possui um banco de dados próprio que possui uma estrutura mínima de tabelas para executar atividades de gerenciamento da aplicação. Essas atividades são as de controle de acesso de usuários e de gestão do armazém de dados. Essa última é responsável por armazenar as informações de conexão da ferramenta com o armazém de dados. Esta estrutura permite que a **BioBI** possa ter acesso a diferentes armazéns de dados para análise.

3.2 SIGLa

O Sistema Integrado de Gerência de Laboratórios (SIGLa) é um LIMS de código aberto que se adapta a diversos laboratórios visando a melhoria na qualidade da entrada dos dados[SIMOES et al., 2010]. A flexibilidade do SIGLa deve-se ao fato deste possuir uma estrutura de gerenciamento de fluxo de trabalho (*workflow*). Essa estrutura permite que os usuários definam as atividades, transações e regras de associações. Por exemplo, na modelagem do *workflow* de um laboratório define-se o atributo de cada atividade, o seu tipo (numérico, textual, data, etc), seu formato, o conjunto de valores possíveis e outras características. Além disso, pode-se definir um conjunto de dados de saídas de uma atividade que será utilizada na atividade seguinte como entrada. A figura 3.1 mostra um exemplo simplificado de um fluxo de trabalho de anamnese de um paciente.

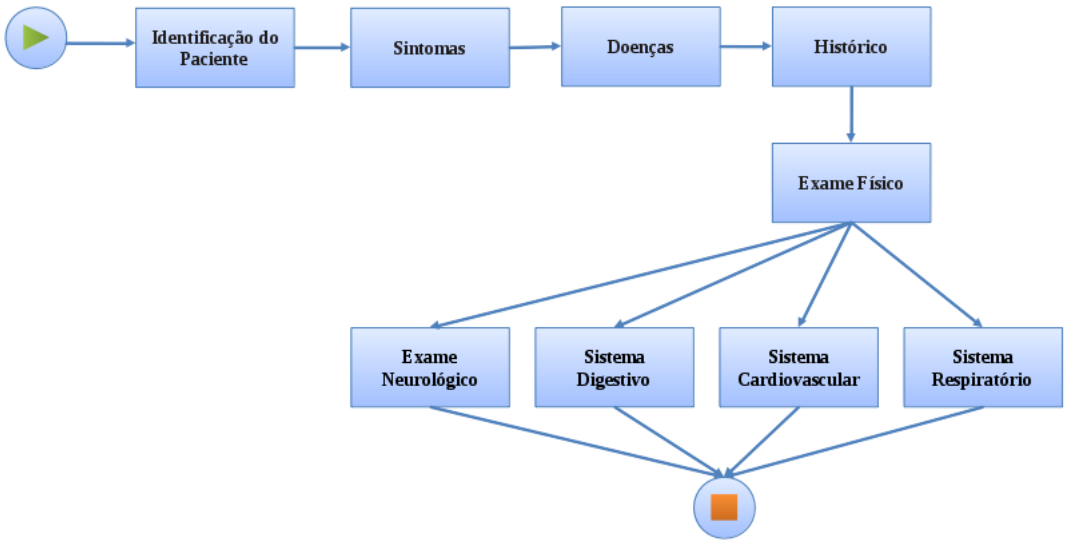


Figura 3.1. Exemplo de um fluxo de trabalho de anamnese de um paciente.

A partir do *workflow* modelado, o SIGLa é responsável por gerenciar a execução das atividades definidas. Através da *web*, o SIGLa auxilia os usuários na execução de cada atividade, sendo que os dados preenchidos de cada atributo são armazenados no banco de dados. A ferramenta também permite que eles tenham o conhecimento das atividades futuras que podem ser executadas. A figura 3.2 mostra como uma atividade do *workflow* é executado no sistema.

A utilização do SIGLa traz benefícios para o laboratório no armazenamento dos seus dados, uma vez que ele auxilia na melhoria da qualidade dos dados através da manutenção da execução correta das atividades, da obrigatoriedade de preenchimento

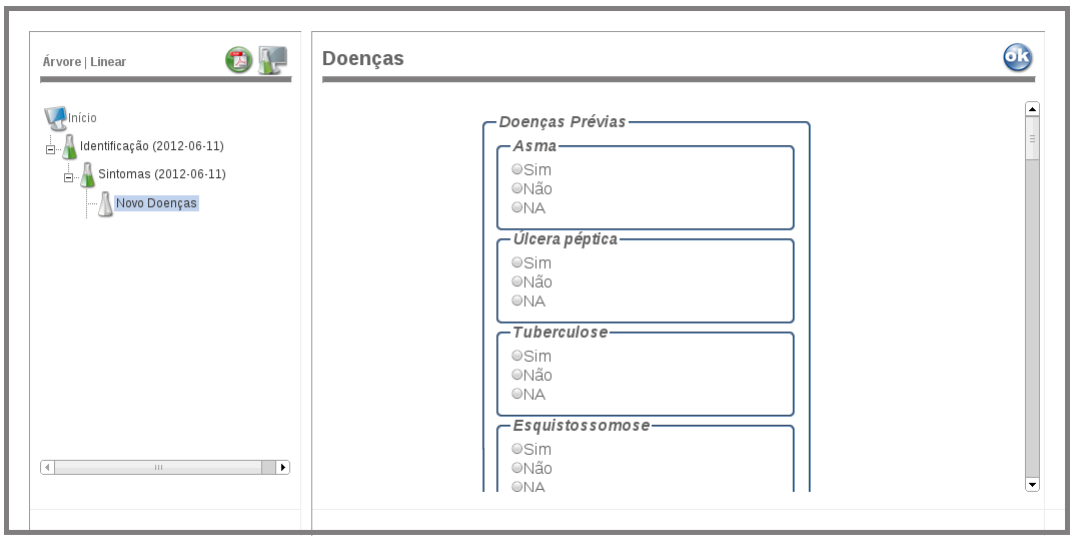


Figura 3.2. Tela de execução de um *workflow* no SIGLa.

dos campos e da corretude deste preenchimento validando a forma como um campo foi preenchido. Este comportamento do SIGLa faz que o tempo para realizar a etapa de ETL dos dados para o armazém de dados seja menor e os processos menos complexos.

3.3 Integração das Ferramentas de Apoio a Decisão

A **BioBI** é uma ferramenta de apoio a decisão que possibilita aos pesquisadores ter acesso a tecnologias de análise de dados através de um navegador *web*. Ela integra três diferentes ferramentas de análise de dados: Mondrian, WEKA e o pacote estatístico R.

A primeira etapa para a concepção do projeto da BioBI foi a análise e o estudo das ferramentas a serem integradas e das maneiras como estas poderiam ser integradas em uma única ferramenta. O Mondrian e o WEKA são ferramentas desenvolvidas em Java e que podem ser incluídas como bibliotecas em outras aplicações. Para o pacote R, foi necessária a instalação de um servidor deste e a utilização de uma biblioteca que realizasse a comunicação entre a **BioBI** e o servidor.

A estrutura comum de integração com as três ferramentas é que a **BioBI** possui páginas *web* desenvolvidas em JSP e essas se comunicam com um conjunto de classes que são os *servlets*. Essas classes são classes especiais em Java, que recebem as requisições dos usuários através das páginas *web*, processam as informações recebidas e retornam o resultado para os usuários. São nessas classes que a **BioBI** utiliza as bibliotecas do WEKA e do pacote R para realizar a análise dos dados. No caso específico

do Mondrian não é necessário a implementação dos *servlets* visto que ele já possui uma estrutura similar. O detalhamento de cada ferramenta e como ocorre a integração com a **BioBI** é apresentado a seguir.

3.3.1 Mondrian

O Mondrian é um servidor OLAP gratuito que possibilita aos usuários explorarem grande volume de dados, analisando-os em diferentes perspectivas e cruzando informações [Mondrian, 2012]. Os dados podem estar armazenados fisicamente em um SGBD relacional. Dessa forma, é necessário que o Mondrian interprete as consultas e operações multidimensionais para consultas SQL tradicionais. Isso é realizado através de um mapeamento do modelo dimensional para o modelo físico relacional.

Esse mapeamento entre os modelos ocorre utilizando um arquivo XML (Extensible Markup Language) que descreve o modelo dimensional (as dimensões, as hierárquias, a tabela de fatos e as métricas) e como ele se encontra fisicamente. Ou seja, este arquivo contém a definição da dimensão: qual tabela no banco de dados a representa, qual a chave primária, qual campo da tabela representa o primeiro nível da hierarquia, entre outros. O mesmo ocorre para a definição da tabela de fatos, sendo que o XML descreve qual a tabela no banco em que ela está armazenada, quais são as métricas e as funções de agregações (média, soma, mediana, etc) associadas, quais são as chaves estrangeiras e a dimensões relativas, entre outras propriedades. É importante salientar que o esquema multidimensional não precisa ser idêntico ao modelo físico relacional. Por exemplo, existem situações em que várias dimensões do esquema estrela são mapeadas para apenas uma tabela física no banco de dados. A figura 3.3 é um exemplo de como o modelo dimensional pode ser mapeado fisicamente em um SGBD.

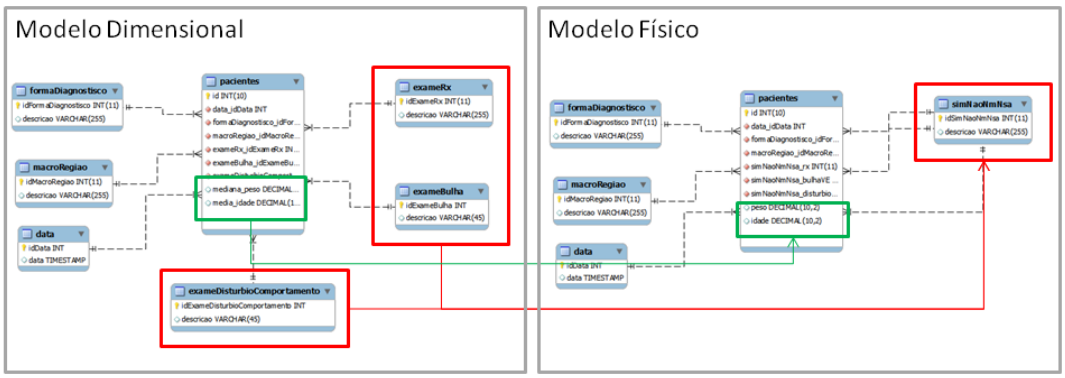


Figura 3.3. Exemplo do mapeamento do modelo dimensional para o modelo físico.

As consultas multidimensionais possuem uma linguagem própria, que é o MDX (*MultiDimensional eXpression*). O Mondrian é responsável por processar essas consultas em um banco de dados relacional tradicional utilizando a linguagem SQL. Isso é possível através do arquivo XML que contêm a definição do cubo e como ele está armazenado fisicamente no banco.

Existem vários clientes com diferentes paradigmas (interface *web* ou *desktop*) e tecnologias que comunicam com o Mondrian. Essas aplicações realizam as consultas no armazém de dados através da linguagem MDX, que são traduzidas pelo Mondrian em SQL. O retorno da consulta no banco de dados é exibido para o usuário de maneira que ele consiga visualizar o resultado de forma tabular.

Um cliente *web* muito utilizado e adotado pelo **BioBI** é o JPivot[JPivot, 2012]. Ele é responsável por permitir que os usuários explorem os dados fazendo as operações típicas de análises multidimensionais (seção 2.2) e por exibir os resultados das consultas de forma tabular. As operações de consultas podem ser realizadas através da interface sem a necessidade de que os analistas tenham conhecimento da linguagem MDX. Além da tabela multidimensional, o JPivot fornece para os usuários a opção de visualizar o resultado de forma gráfica, sendo que o tipo de gráfico (gráfico de barras, colunas ou pizza) e suas propriedades (legenda, tamanho, título, entre outros) podem ser alterados facilmente.

A **BioBI** utiliza o Mondrian em conjunto com o JPivot para as análises multidimensionais. Para tanto, foi necessário a implementação de uma página JSP que utiliza as funcionalidades providas pelas ferramentas para a execução das operações OLAP. O Mondrian manipula a consulta requisitada pelo usuário e a transforma em uma consulta nativa SQL. Posteriormente ele utiliza o arquivo XML, que define o esquema multidimensional e como ele está armazenado fisicamente no banco de dados, para realizar o mapeamento entre a consulta MDX para SQL. A consulta é realizada no banco de dados, sendo que o resultado é recebido pelo Mondrian que, por sua vez repassa para o JPivot. Ao receber o conjunto de dados de retorno, o JPivot exibe os resultado de maneira tabular e gráfica para os analistas. Esse fluxo de interação entre a página JSP do **BioBI**, JPivot e Mondrian, é representado na figura 3.4.

A figura 3.5 mostra como os analistas podem elaborar as consultas para a exploração dos dados, de maneira iterativa sem a necessidade de conhecimento da linguagem MDX. O resultado é exibido de forma tabular podendo ser visualizado de forma gráfica (figura 3.6). A consulta MDX correspondente ao resultado é apresentada na figura 3.7.

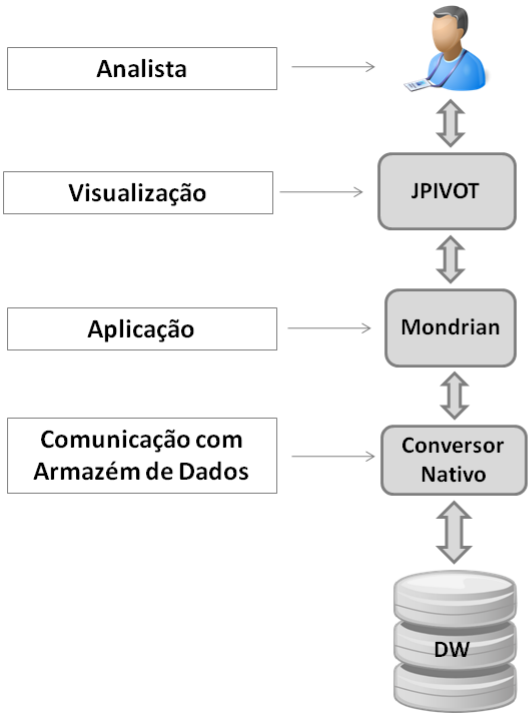


Figura 3.4. Representação da arquitetura de comunicação da **BioBI** com o JPivot e Mondrian

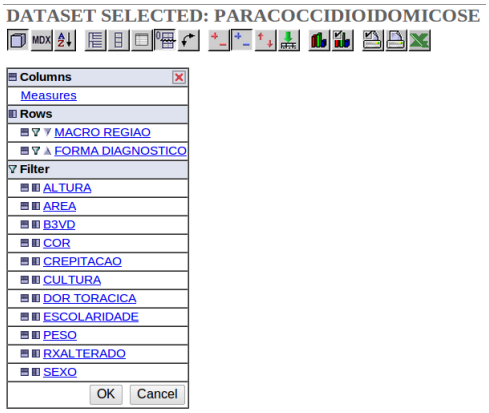


Figura 3.5. Interface para elaboração da consulta.

3.3.2 WEKA

O WEKA - *Waikato Environment for Knowledge Analysis* é uma ferramenta que unifica diversas técnicas de análise de dados, tais como os algoritmos de classificação, agrupamentos, regras de associação e seleção de atributos [Hall et al., 2009]. Ele também possibilita aos analistas realizar o pré-processamento e visualização dos dados que permite uma análise prévia da informação. O projeto se originou em 1992 com intuito

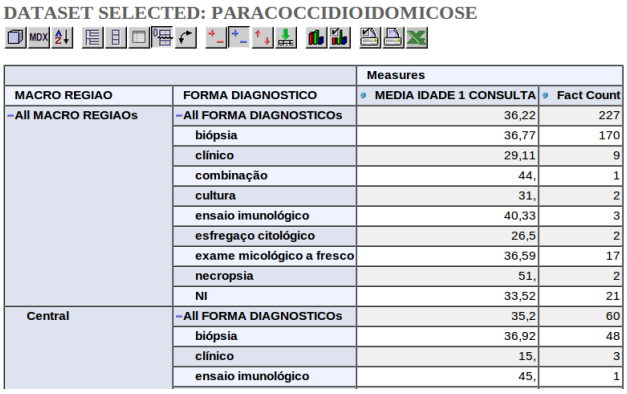


Figura 3.6. Interface de exibição do resultado da consulta.

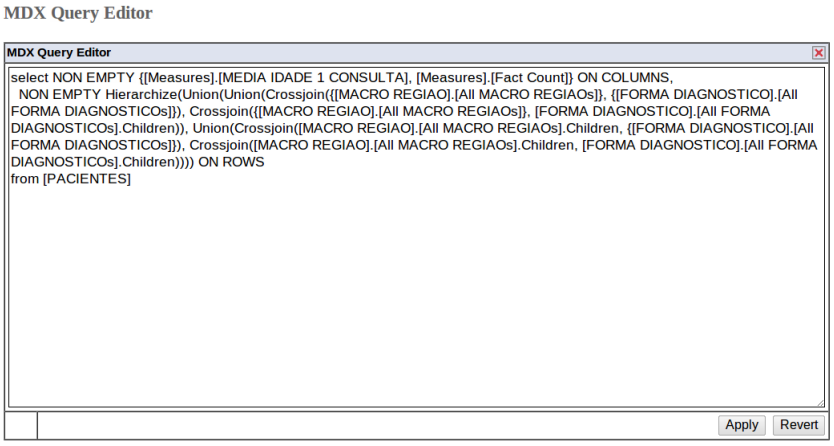


Figura 3.7. MDX correspondente a consulta.

de integrar em uma plataforma padronizada às técnicas de extração de conhecimento existentes em diferentes linguagens. Atualmente, ele também é um *framework*, possibilitando que aos pesquisadores desenvolver seus próprios algoritmos de mineração de dados sem se preocupar com a manipulação dos dados. Esta é feita utilizando um arquivo texto no formato ARFF (Attribute Relationship File Format) que define os atributos, suas relações, e as instâncias do conjunto de dados. O arquivo é separado em duas partes, sendo que a primeira é um cabeçalho que descreve os atributos e seu tipo e a segunda apresenta cada instância do banco de dados e os valores para cada atributo.

O WEKA é bem aceito nas universidades e nas companhias privadas, e se tornou largamente utilizável como ferramenta de mineração de dados. Um dos motivos dessa adoção é que ela é implementada em Java, tornando-a portável para diversas plataformas computacionais. Além disso, é uma ferramenta gratuita que pode ser utilizada por uma interface gráfica *desktop*, através de linha de comando, e também pode ser

incorporada por outros sistemas adicionando-a como uma biblioteca do novo sistema.

BioBI incorpora a biblioteca do WEKA no projeto e, atualmente, disponibiliza alguns de seus recursos para os analistas através de uma interface *web*. Estas são implementadas utilizando JSP e tem como objetivo permitir que os usuários definam os parâmetros dos algoritmos a serem executados e visualizem os resultados obtidos. As páginas implementadas interagem com os *servlets* responsáveis por instanciar as classes necessárias para realizar a atividade de mineração de dados. O fluxo de comunicação pode ser observado na figura 3.8.

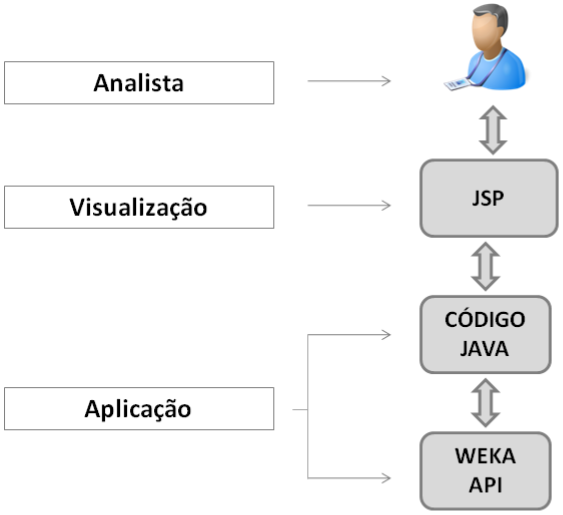


Figura 3.8. Representação da arquitetura de comunicação do **BioBI** com o WEKA

O WEKA disponibiliza um leque de opções de técnicas e algoritmos que auxiliam no processo de descoberta do conhecimento. Atualmente, o **BioBI** utiliza duas das técnicas de mineração de dados disponíveis: seleção de atributos relevantes e agrupamentos. Essas duas técnicas foram priorizadas devido ao domínio de estudo, detalhado na seção 4.1, ter centenas de atributos e por não termos o conhecimento prévio do comportamento da base de dados.

Seleção de Atributos

A técnica de seleção de atributos tem como objetivo encontrar um conjunto mínimo de dimensões capaz de obter um resultado mais próximo possível do que se estivesse

processando todos os atributos. Com a redução da dimensionalidade dos dados obtém-se um ganho no processamento das informações além de facilitar na legibilidade dos resultados. **BioBI** possui uma interface *web* (Figura 3.9) para que os usuários definam os parâmetros dos algoritmos, que são repassados para um *servlet*. A figura 3.10 mostra o trecho de código desse *servlet* que recebe os parâmetros definidos pelo usuário para a execução do algoritmo.

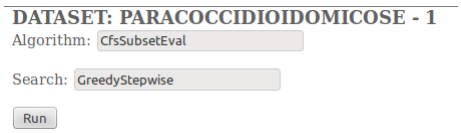


Figura 3.9. Interface do **BioBI** para a seleção de atributos relevantes.

```
AttributeSelection attsel = new AttributeSelection();

CfsSubsetEval eval = new CfsSubsetEval();

GreedyStepwise search = new GreedyStepwise();
search.setSearchBackwards(true);

attsel.setEvaluator(eval);
attsel.setSearch(search);
attsel.SelectAttributes(data);
int[] indices = attsel.selectedAttributes();
```

Figura 3.10. Trecho de código do **BioBI** utilizando o a biblioteca do WEKA para seleção de atributos.

Na implementação atual, o **BioBI** utiliza os algoritmos *CfsSubsetEval*[Witten.I et al., 2011] e o *InfoGainAttributeEval*[Witten.I et al., 2011] para identificar os atributos mais relevantes da base de dados. O primeiro algoritmo é responsável por identificar o melhor subconjunto de atributos associados a uma determinada variável, e o segundo apresenta qual o grau de correlação de cada atributo com a classe selecionada. Estes algoritmos foram inicialmente selecionados para serem incorporados no **BioBI** pois eles seus resultados são mais simples para serem interpretados e por possuir poucos parâmetros de configuração para execução. As estratégias de buscas do conjunto de atributos são o *GreedyStepwise*[Witten.I et al., 2011] e o *Ranker*[Witten.I et al., 2011].

Agrupamentos

Como a grande parte dos atributos da base de dados do estudo de caso são do tipo nominal, o algoritmo *Simple k-Means*[Witten.I et al., 2011] foi selecionado para ser utilizado no **BioBI**Ṫambém, o resultado da execução desse algoritmo é de fácil leitura

para os analistas. Além disso, **BioBI** utiliza a estratégia da distância euclidiana para calcular a similaridade dos registros, ou seja, identificar qual o melhor grupo a ser atribuído para um registro. A tela para parametrização do algoritmo é apresentada na figura 3.11 onde o analista deve definir o número de *clusters* e quais atributos devem ser considerados para a identificação dos grupos.

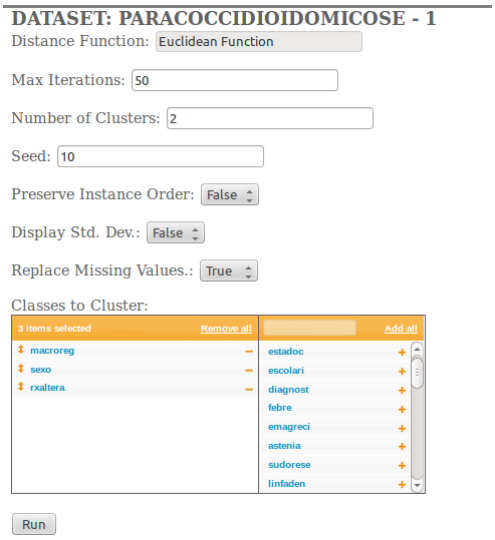


Figura 3.11. Interface para parametrização do algoritmo de agrupamento

O trecho de código do *servlet* que realiza a comunicação entre os parâmetros definidos pelos usuários com a API do WEKA é apresentado na figura 3.12. Nele ocorre a instanciação da classe do algoritmo de agrupamentos, existente na API do WEKA, e a definição dos valores das variáveis de execução (número de *clusters* e de iterações, atributos a serem considerados, e outros) e a função de distância (distância euclidiana).

3.3.3 R

O pacote R [R, 2012] é um ambiente de programação gratuito para análise estatística de dados e geração de gráficos. Através uma linguagem de programação, similar a da ferramenta de análise de dados "S"(*software* proprietário), os usuários podem programar diversos *scripts* para analisar e visualizar os dados [Ricci, 2004]. Ele provê uma grande variedade de comandos que facilita na realização de testes estatísticos, análise temporal, cálculos e operações em vetores e matrizes, entre outros.

O pacote R permite que desenvolvedores implementem pacotes que tem como objetivo adicionar novos comandos que facilitam uma determinada análise. Por exemplo, pode-se criar um pacote que possui um comando que encapsula um trecho de código R

```
SimpleKMeans skm = new SimpleKMeans();

skm.setDistanceFunction(new EuclideanDistance());
skm.setDisplayStdDevs(false);
skm.setMaxIterations(maxIteration);
skm.setNumClusters(numCluster);
skm.setSeed(seed);
skm.setDontReplaceMissingValues(replaceMissingValues);
skm.setPreserveInstancesOrder(preserverInstanceOrder);
skm.setDisplayStdDevs(displayStdDev);
skm.buildClusterer(data2Cluster);

ClusterEvaluation eval;
eval = new ClusterEvaluation();
eval.setClusterer(skm);
eval.evaluateClusterer(data2Cluster);
```

Figura 3.12. Trecho de código do **BioBI** utilizando o a biblioteca do WEKA para agrupamentos.

que realiza um complexo teste estatístico. Apesar de toda sua capacidade de processamento e análise de dados, o R é utilizado apenas por grupos restritos de pessoas que possuam conhecimento de programação, do ambiente e da linguagem. Isso se deve ao fato de todos os testes e análises precisarem ser implementados utilizando a linguagem R através de comandos da linguagem. O objetivo do **BioBI** ao integrar o R a sua plataforma é permitir que usuários tenham acesso a análise estatística dos seus dados através de uma interface *web* sem a necessidade de programação de código.

A integração do R com o **BioBI** é mais complexa que a das outras aplicações descritas anteriormente. Para a utilização do R em outra aplicação é necessário que ele esteja instalado em uma máquina, e que essa seja o servidor do programa R. Isso é feito através da instalação de um pacote R chamado RServe. Esse pacote é responsável por permitir que a máquina que tenha a instalação do R receba e execute *scripts* de uma outra aplicação através de uma conexão TCP/IP [Urbanek, 2003].

Para que a aplicação se comunique com o servidor R e possa enviar os *scripts* e receber o resultado do processamento, ela precisa utilizar uma biblioteca responsável por isso. Essa biblioteca é disponibilizada pelo projeto que desenvolveu e mantém o pacote RServe e é utilizada pelo **BioBI** para realizar as análises estatísticas e geração de gráficos através do R.

Para uma melhor modularização da ferramenta, o **BioBI** possui um módulo próprio, BIO R, que encapsula determinados *scripts* R em métodos Java. É nesse módulo que ocorre a comunicação entre o **BioBI** e o R utilizando a biblioteca do RServe.

A figura 3.13 mostra como ocorre a integração do **BioBI** com o R. Os analistas realizam uma requisição através das páginas JSP que repassam a solicitação para os

servlets. Esses recuperam os dados necessários no armazém de dados, os quais são repassados para o BIO R. Ele é responsável por enviar a requisição da execução de um *script* e receber seu resultado através da API do RServe disponibilizada. Dessa forma, o resultado retornado é visualizado pelos analistas através do navegador.

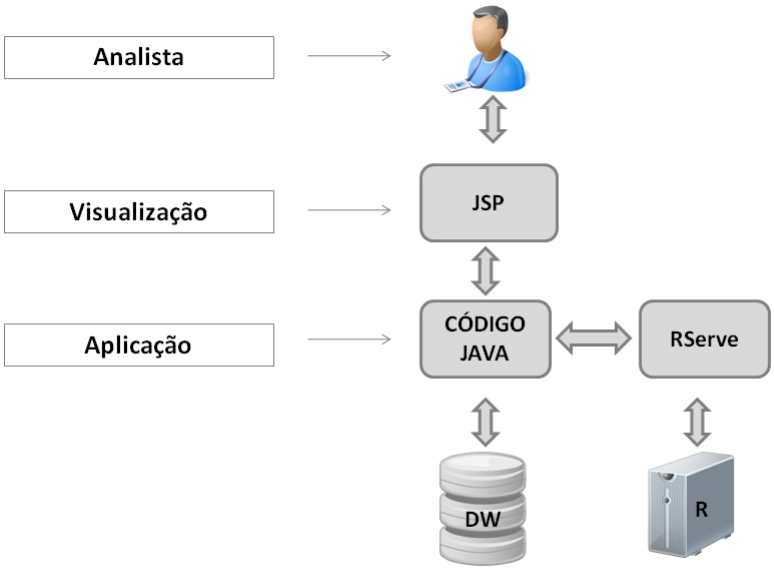


Figura 3.13. Representação da arquitetura de comunicação do **BioBI** com o R

A integração com o R possibilita que o **BioBI** crie gráficos analíticos de forma mais rápida. Essa funcionalidade é muito importante para que os analistas realizem uma pré-análise dos dados armazenados de maneira que eles consigam ter uma visão macro de como está a distribuição dos registros para cada atributo. A figura 3.14 mostra um histograma criado pelo R que apresenta a frequência de ocorrências de cada valor distinto na base de dados. Ao lado do gráfico também são apresentadas algumas análises estatísticas descritivas do atributo, como o número distinto de valores, média, desvio padrão, valor máximo e valor mínimo.

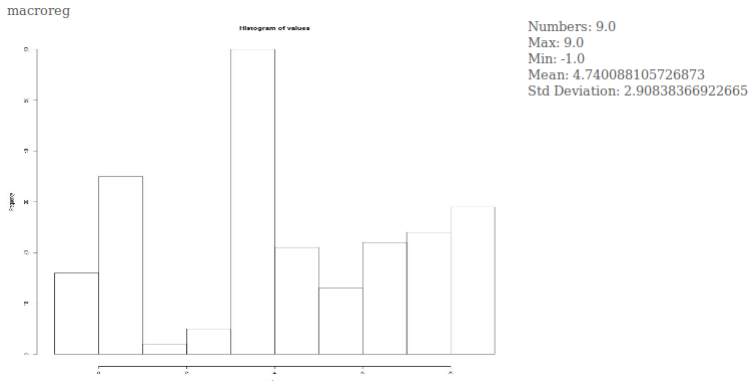


Figura 3.14. Histogramas criados pelo R

Capítulo 4

Estudo de Caso

O **BioBI** é uma ferramenta de apoio aos pesquisadores para análise de dados biológicos e clínicos. Este trabalho utiliza os dados de pacientes acometidos por uma doença tropical que são analisados utilizando a ferramenta. Para isso, é necessária a implantação de um armazém de dados com a estrutura dos dados organizada de maneira apropriada para as análises OLAP e mineração de dados.

Este capítulo está estruturado da seguinte maneira: A seção 4.1 apresenta uma breve descrição da doença em estudo: a paracoccidioidomicose(PCM). A seção 4.2 descreve o conjunto de dados de pacientes portadores da doença que foram disponibilizados. O armazém de dados desenvolvido e o processo de ETL necessário para implantação é detalhado nas seções 4.3 e 4.4. Finalmente na seção 4.5, são apresentados os resultados das análises realizadas para caracterização da base de dados dos pacientes acometidos pela PCM.

4.1 Estudo de Caso — Paracoccidioidomicose — PCM

A paracoccidioidomicose (PCM) é uma micose causada pelo fungo *Paracoccidioides brasiliensis*, uma espécie tipicamente brasileira[RESTREPO et al., 2001]. A infecção ocorre, habitualmente, através da inalação atingindo o epitélio alveolar pulmonar. Se não houver a eliminação do fungo, ele pode se disseminar para outros órgãos e tecidos caracterizando a doença.

A análise clínica dos pacientes segue um protocolo definido pelo Centro de Treinamento e Referência de Doenças Infecto-Parasitárias - CTR-DIP, Anexo Orestes Diniz, do Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG). Esse proto-

colo possui diversas variáveis clínicas que são avaliadas pelos médicos em cada consulta. O momento adequado para a interrupção do tratamento dos pacientes acometidos pela doença é uma preocupação e um desafio para os médicos. Atualmente, os critérios utilizados são clínicos, radiológicos e sorológicos que, na maioria das vezes, não estão disponíveis nos postos de atendimento público de saúde. Além disso, ainda não há um parâmetro clínico que seja confiável para determinar o momento de interrupção ou que permita prever as reicidências da doença.

4.2 Conjunto de Dados

Nesta trabalho, os dados utilizados para análise contêm as informações clínicas da primeira consulta que os pacientes afetados pela PCM realizaram, conforme o protocolo padrão (anexo B). Eles foram disponibilizados pelo CTR-DIP do Hospital das Clínicas. A base de dados consiste no exame clínico de 227 pacientes onde são analisados 314 parâmetros. Sendo que, 09 atributos são do tipo numérico (idade, tempo de evolução, tamanho do RCD, tamanho do AX, tamanho boyd, tamanho da lesão, área da lesão, tempo para inativação e tempo de tratamento até a suspensão) e os outros 305 atributos são do tipo nominal.

4.3 Modelo Dimensional

significado do grão é dado pela junção das dimensões na tabela de fatos, sendo que, essa estrutura contribui para a mineração de dados e consultas OLAP.

O armazém de dados para análise dos pacientes com PCM foi modelado a partir dos dados disponibilizados. O grão da tabela de fatos representa uma consulta clínica com as suas diversas variáveis. Ou seja, cada registro na tabela de fatos representa um consulta clínica. O exame clínico contempla diversos parâmetros, sendo que alguns foram definidos como métricas e outros como dimensões. As variáveis quantitativas existentes na base de dados de origem (idade, tempo de evolução, tamanho do RCD, tamanho do AX, tamanho boyd, tamanho da lesão, área da lesão, tempo para inativação e tempo de tratamento até a suspensão) foram definidas como métricas no cubo OLAP, enquanto que as variáveis nominais (que representam uma categoria/característica sem um valor quantitativo) foram definidas com as dimensões do armazém de dados.

O esquema simplificado do armazém de dados deste estudo de caso é apresentado na figura 4.1. A tabela central, "consulta", é a tabela de fatos e as tabelas em seu entorno são as dimensões. Contudo, conforme descrito na seção 2.1, o cubo pode ser

armazenado fisicamente de maneira diferente do modelo dimensional. As dimensões "frequenciaCardiaca" e "frequenciaPulso" são armazenadas em uma única tabela no banco: "frequencia". As dimensões "rxAlterado", "urinaAlterada" e "placas" são armazenadas na tabela "simNaoNmNsa". Isso ocorre, porque essas dimensões possuem os mesmos valores de classificação. As métricas obtidas através de agrupamentos de um conjunto de dados, tais como a média, soma e mediana, também não são armazenadas fisicamente no banco de dados. O Mondrian é responsável por realizar o cálculo dessas métricas de acordo com o conjunto de dados selecionados na consulta.

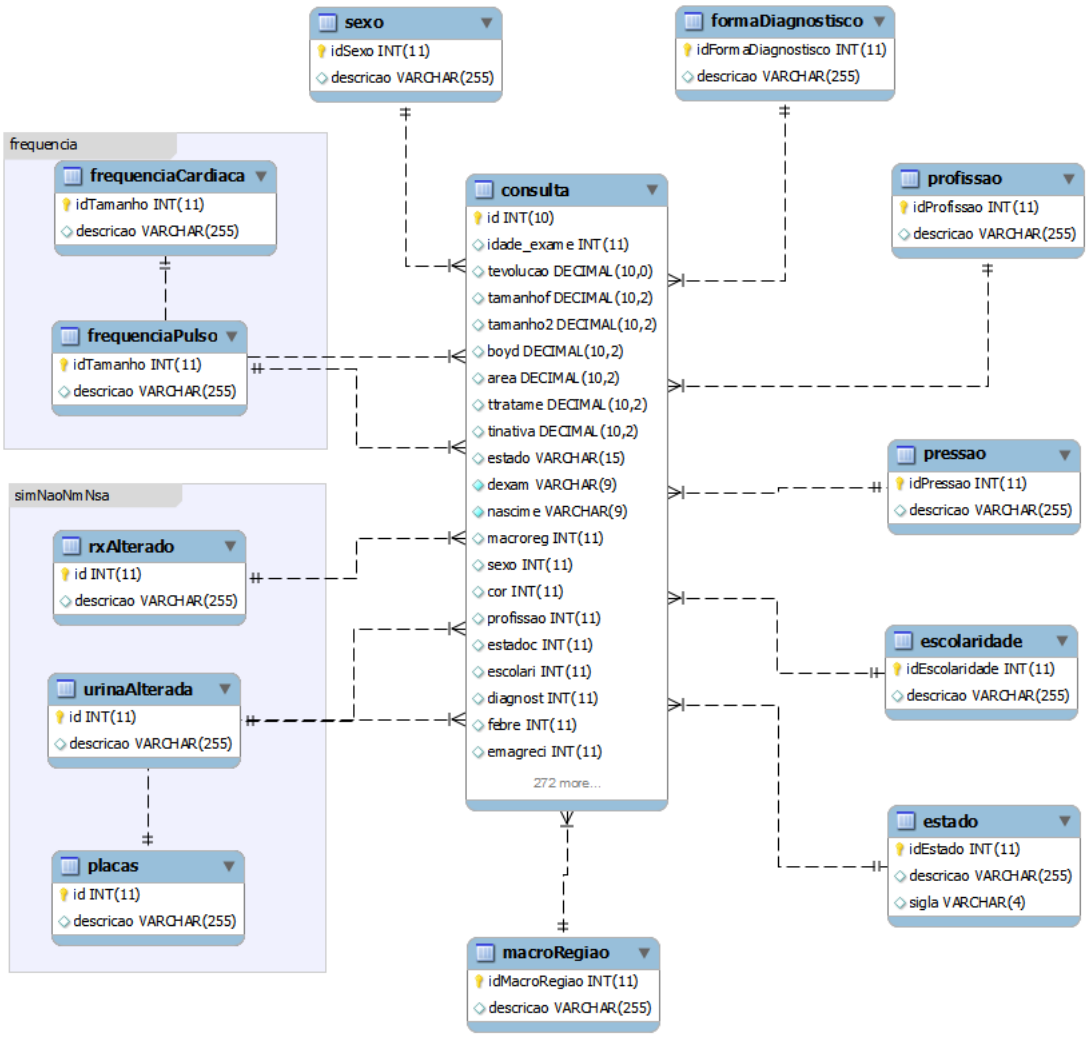


Figura 4.1. Esquema do Modelo Dimensional do Armazém de Dados.

4.4 ETL

O CTR-DIP[1] utiliza o programa estatístico proprietário SPSS[2] para armazenar os dados clínicos dos pacientes acometidos pela PCM. As informações preenchidas no protocolo de cada paciente são transcritas manualmente no SPSS. A base de dados foi disponibilizada no formato *sav*, que é a extensão padrão do SPSS. Utilizando uma versão de avaliação do programa, a base de dados foi exportada para o formato CSV (*comma-separated values*) tornando-se acessível por diversos programas. Utilizando este novo formato de armazenamento, os dados foram inseridos em uma tabela em um banco de dados MySQL.

Como os dados originais encontram-se estruturados de maneira tabular, a tabela originada pela importação está semi preparada para ser a tabela de fatos do modelo, sendo necessários alguns ajustes. Foram utilizados comandos SQLs para realizar tais transformações:

- Eliminação dos campos "same", "protocolo", "municipio", "telefone", "endereço", "naturalidade" e "nome" do modelo, visto que estes atributos do paciente não são, atualmente, relevantes para a análise dos dados;
- Padronização da ortografia do nome dos estados, pois, a base original possuía registros com ortografias diferentes para um mesmo estado;
- Atualização dos valores de registros nulos para -1 para que todos os valores dos atributos sejam do mesmo tipo: inteiro;

Apos a execução das etapas acima, foram criadas as dimensões do modelo e realizada as inserções dos valores nas tabelas do armazém de dados, detalhado no apêndice **??**. Ao final deste procedimento, o armazém de dados dos pacientes PCM está preparado para análise através do **BioBI** com objetivo de auxiliar os pesquisadores a encontrarem novas informações.

4.5 Análise dos Dados

Utilizou-se o **BioBI** para realizar a caracterização do conjunto de dados disponíveis. As análises seguiram as etapas abaixo detalhadas posteriormente.

1. Análise da distribuição de frequência das variáveis;

[1] Centro de Treinamento e Referência de Doenças Infecto-Parasitárias - CTR-DIP, Anexo Orestes Diniz, do Hospital das Clínicas da Universidade Federal de Minas Gerais - UFMG.

[2] www.ibm.com/software/analytics/spss/

2. Identificação de atributos relevantes;

3. Identificação e análise de grupos de pacientes.

4.5.1 Análise da frequência de valores de cada atributo

A distribuição da frequência tem como objetivo apresentar quantas vezes um determinado valor aparece na amostra de dados. Essa análise pode ser apresentada de forma tabular ou gráfica. Um histograma é a maneira gráfica de apresentar a distribuição de frequência, onde a base corresponde a um intervalo (ou, no caso de uma variável nominal, um valor) e a altura a respectiva frequência. A análise da distribuição da frequência é uma maneira do analista ter uma percepção da base de dados, e de como os dados estão distribuídos para cada variável.

Realizou-se uma análise da distribuição de frequência de cada atributo através dos histogramas e das tabelas de frequência gerados através do R. Pode-se perceber que existem diversos atributos no conjunto de dados que possuem uma concentração muito grande em um único valor. A tabela 4.1 apresenta os atributos onde não são preenchidos ou não são avaliados durante a anamnese do paciente. Esse comportamento dos dados pode representar que os médicos não estão avaliando os exames ou que esses não são de extrema importância na análise clínica dos pacientes com PCM.

Atributo	**Valor Prevalente**	**%(Frequência)**
Teste imunológico positivo	Não preenchido	98,69(226)
Tamanho Boyd cm	Não preenchido	85,15(200)
Área da lesão em cm2	Não preenchido	78.6(180)
Raio X de seios da face alterado	Não Avaliado	93,5(214)
Ác. Lático alterado	Não Avaliado	92,1(211)
CPK alterado	Não Avaliado	91,7(210)
CK-MB alterado	Não Avaliado	90,8(208)
HBeAg positivo	Não Avaliado	90,8(208)
Anti-HBc positivo	Não Avaliado	90,5(207)
Lipase alterada	Não Avaliado	88,2(202)
CMV positivo	Não Avaliado	87,8(201)
Anti-HIV positivo (Elisa)	Não Avaliado	87,3(200)
Chagas positivo	Não Avaliado	87,3(200)
Ác. úrico alterado	Não Avaliado	86,5(198)
Anti-HAV positivo	Não Avaliado	86,34(198)
Cultura alterada	Não Avaliado	86,4(198)
Espirometria alterada	Não Avaliado	86,5(198)
Anti-HCV positivo	Não Avaliado	86,3(197)
Amilase alterada	Não Avaliado	85,6(196)
Anti-HBs positivo	Não Avaliado	85,6(196)

Tabela 4.1. Tabela com os atributos com grande concentração na frequência no registro "Não Avaliado"e "Não preenchido".

A tabela 4.2 apresenta as variáveis que têm uma grande frequência relativa ao valor "Não alterado"no exame analisado.

Atributo	**Valor Prevalente**	**%(Frequência)**
PCM renal	"Não alterado"	95,2(218)
PCM pancreática	Não alterado	95,2(218)
PCM esplênica	Não alterado	93(217)
PCM aparelho genital	Não alterado	94,7(217)
PCM adrenal	Não alterado	93,9(215)
PCM intestinal	Não alterado	93,9(215)
PCM óssea	Não alterado	93(213)
PCM gástrica	Não alterado	93,1(213)
PCM hepática	Não alterado	92,1(211)
PCM linfática	Não alterado	91,3(209)
PCM neurológica	Não alterado	90,4(207)
Anasarca	Não alterado	87,3(200)
Bulhas arrítmicas	Não alterado	87,3(200)
MMSS	Não alterado	86 (197)
Palpebral	Não alterado	85,6(196)
PCM disseminada	Não alterado	85,6(196)

Tabela 4.2. Tabela com os atributos com grande concentração na frequência no registro de exames "Não Alterados".

Esses parâmetros que possuem uma frequência de ocorrência concentrada em um registro, normalmente são variáveis com pouca força de discriminação, ou seja, elas não são boas de serem utilizadas nos algoritmos de agrupamentos ou classificação, visto que elas não têm tanta força para separar um indivíduo de outro. Visto que o conjunto de dados possuem apenas informações da primeira consulta dos pacientes, essas variáveis podem ser decisivas para identificar uma possível interrupção no tratamento da PCM ou na recidiva da doença. Para tanto, é necessário que a base de dados possua essa informação.

É interessante notar que, dos 17 tipos de PCM avaliados no protocolo de consulta 11 tipos não são, significamente, prevalentes no conjunto de dados.

4.5.2 Seleção de Atributos Relevantes

A técnica de seleção de atributos é comumente utilizada para a redução da dimensionalidade sem afetar, de forma significativa, os resultados das análises. Além disso, a redução do espaço provoca uma otimização na execução dos algoritmos de mineração de dados e facilita a leitura do resultado, visto que o conjunto de parâmetros correlacionados é menor.

A redução da dimensionalidade é realizada de maneira que encontre o melhor

subconjunto de dados associados a uma classe(atributo) do modelo. Ou seja, a partir de uma determinada classe o algoritmo de seleção de atributos determina um subconjunto de atributos mais correlacionado com ele. Em de Moura [2008], as variáveis foram selecionadas utilizando a seleção progressiva com razão de probabilidade, em que as variáveis são selecionadas com base na significância do escore estatístico. As variáveis que apresentaram o maior escore estatístico e o valor de significância (valor-p) menor que 0,05, eram selecionados para as próximas etapas de seleção. Neste trabalho foram selecionados os 4 atributos que apresentaram o maior escore estatístico com o valor-p menor 0,05 no passo inicial de seleção de atributos apresentados por de Moura [2008]. A tabela 4.3 apresenta as classes selecionadas, sendo que, para cada uma delas, identificou-se o subconjunto de atributos mais correlacionado.

Atributo	**Escore Estatístico**	**Valor-p**
Localização da lesão nas mucosas	5,350	0,021
Localização da lesão na pele	4,040	0,044
Sexo	4,098	0,043
Vomitos	5,198	0,023

Tabela 4.3. Atributos selecionados como classes para a seleção de variáveis.

Atributos selecionados

A tabela 4.4 apresenta o subconjunto de atributos selecionados para cada classe definida. A coluna Avaliação mostra o grau de correlação entre o subconjunto de atributos[Witten.I et al., 2011] selecionados com a classe, sendo que, quanto mais próximo de 1 maior é o grau de relacionamento entre eles.

Classe	**Subconjunto de atributos**
Sexo	"idade_exame", "profissao","dorabdom","dptuberc","tabprevi","nodulo", "anasarca", "ssea", "pcmpulmo"
Vômito	"emagreci", "sudorese", "coluria", "convulsao", "obstnasa", "disfonia", "odinofag", "azia", "dorabdom", "
Localização da lesão na pele	"lesaocut", "desidrat", "tulceras", "cabeça", "mmss", "edem", "pcmcutan"
Localização da lesão na mucosa	"tevolucao", "lesaocut", "obstnasa", "odinofag", "tabprevi", "paempe", "lesmucos", "boca", "pcmmucos"

Tabela 4.4. Resultado dos algoritmos de seleção de atributos

4.5.3 Caracterização dos Dados

O agrupamento é uma técnica que realiza uma separação das instâncias em grupos. A análise do *cluster* é uma boa ferramenta para conhecer o comportamento dos dados, além de sugerir hipóteses que podem ser pesquisadas com mais detalhe através de

análises multidimensionais entre outras.

Para cada subconjunto de atributos selecionados de cada classe apresentada anteriormente, realizou-se a separação dos pacientes em grupos utilizando o algoritmo *k-means* disponível pelo **BioBI**. Este algoritmo necessita que se defina, *a priori*, um número k de grupos. Caso o número de grupos for desconhecido, umas das maneiras de encontrar o número de grupos é definir o valor de $k = 1$ e ir variando este número de forma que minimize o erro[Han & Kamber, 2001]. Neste trabalho, utilizamos o algoritmo EM[Witten.I et al., 2011] que é capaz de identificar um número de *cluster*, sendo que, este valor foi utilizado para definir o valor de k.

Após a execução do algoritmo, realizou-se uma análise prévia dos grupos com intuito de identificar atributos que poderiam ser removidos do modelo. Os atributos que não apresentaram alteração em nenhum dos *clusters* eram excluídos e, novamente, o algoritmo de agrupamento era executado.

Alguns agrupamentos evidenciaram algumas correlações entre os atributos, os quais foram investigados com mais detalhes através das análises multidimensionais. Contudo, a confirmação da interpretação dos dados só pode ser validada pelos especialistas da área.

Análise 1

A tabela 4.5 apresenta o resultado do agrupamento utilizando os atributos relevantes obtidos para a identificação da classe "Lesão mucosa". Tem-se que o cluster #0 é o grupo formado pelos pacientes com lesão cutânea e tabagismo prévio ("tabprevi"), e todos os outros exames normais. O cluster #1 já descreve as instâncias que não apresentaram lesão cutânea e obstrução nasal, sendo que as outras variáveis clínicas estavam alteradas. O agrupamento #2 são os pacientes que apresentaram ambas as lesões: cutânea e mucosa.

O cluster #4, que representa 32% das instâncias, é formado pelos pacientes que não apresentaram nenhuma alteração nos exames selecionados. O Cluster #3 diferencia do cluster #4, pois são os pacientes tabagistas e apresentaram a PCM Mucosa. O tempo médio de evolução dos pacientes deste grupo apresenta um valor destoante dos demais, o que indica que tabagismo eleva o tempo de detecção da doença.

Na tabela 4.6 apresenta um cenário com a dimensionalidade mais reduzida, considerando apenas as variáveis marcantes do resultado do agrupamento da figura 4.5: "tevolucao", "lesaocut", "tabprevi", "pcmmucos", "lesaomuc".

Atributo	Grupo Global	#0	#1	#2	#3	#4
	227(100%)	35(15%)	78(34%)	24(11%)	17(7%)	73(32%)
tevolucao	15.6167	18.9714	11.8077	17.875	65.8824	5.6301
lesaocut	ok	x	ok	x	ok	ok
obstnasa	ok	ok	ok	-	ok	ok
odinofag	ok	ok	x	ok	ok	ok
tabprevi	x	x	x	-	x	ok
boca	ok	ok	x	x	ok	ok
pcmmucos	x	ok	x	x	x	ok
lesaomuc	x	ok	x	x	x	ok

Tabela 4.5. Resultado do algoritmo de agrupamento

Atributo	Grupo Global	#0	#1	#2	#3	#4
	227(100%)	41(18%)	67(30%)	19(8%)	47(21%)	53(23%)
tevolucao	15.6167	11.8537	10.6567	17.7368	36.8298	5.2264
lesaocut	ok	x	ok	x	ok	ok
tabprevi	x	x	x	-	x	ok
pcmmucos	x	x	x	x	ok	ok
lesaomuc	x	x	x	x	ok	ok

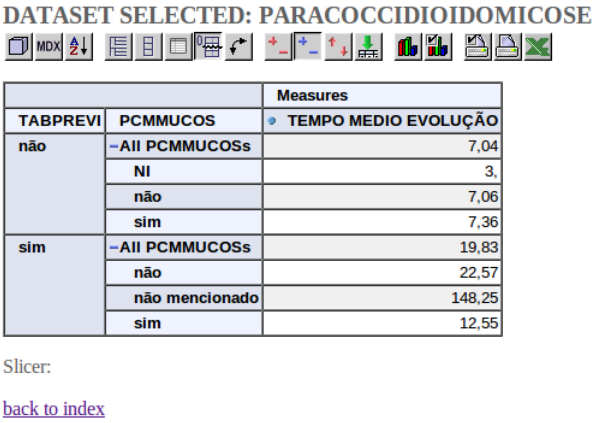
Tabela 4.6. Resultado do algoritmo de agrupamento

No cenário apresentado na tabela 4.6, os pacientes que apresentaram lesão cutânea, lesão mucosa e PCM mucosa estão no grupo #0 e #2. O que diferencia estes dois grupos é o tabagismo ("tabprevi"), que, no grupo #2 não foi avaliado esse histórico do paciente. O cluster #1 é o que se aproxima do centro global do modelo (30% das instâncias conectadas ao grupo), o qual, os pacientes com tabagismo prévio, lesão mucosa e PCM mucosa, mas não apresentaram lesão cutânea. O grupo #4 contem os pacientes que não possuíram alteração nos exames avaliados, enquanto que o cluster #3 contempla os pacientes tabagistas e que apresentaram o maior tempo médio de evolução. O mesmo ocorreu no cluster #3 apresentado na tabela 4.5 onde o tempo médio de evolução estava muito destoante dos demais grupo. A análise dos grupos provocou alguns questionamentos a serem verificados:

- Qual a relação entre tabagismo prévio ("tabprevi") com o tempo de evolução ("tevolucao") do paciente?
- Pacientes que apresentam PCM Mucosa podem não ter lesão mucosa?
- Qual a correlação entre PCM cutânea, lesão cutânea, PCM mucosa e lesão mucosa?

A figura 4.2 apresenta a análise dos pacientes com tabagismo prévio com o tempo médio de evolução da doença. Pode-se observar que os pacientes não tabagistas têm um tempo médio de evolução 7,04 meses enquanto que os pacientes tabagistas o tempo

médio foi de 19,83 meses. Ou seja, o tempo médio de evolução dos pacientes foi 2,81 vezes maior quando há relatos de tabagismo. Ao avaliar a sua correlação com pacientes que apresentam a PCM Mucosa, o tempo médio de evolução continua maior. Sendo que, os não tabagistas o tempo médio é de 7,36 meses e tabagistas 12,55 meses. Em Santos et al. [2003]; Shikanai-Yasuda et al. [2006] descrevem que há relação entre o tabagismo e a PCM.



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

		Measures
TABPREVI	PCMMUCOS	TEMPO MEDIO EVOLUÇÃO
não	-All PCMMUCOSs	7,04
	NI	3,
	não	7,06
	sim	7,36
sim	-All PCMMUCOSs	19,83
	não	22,57
	não mencionado	148,25
	sim	12,55

Slicer:

back to index

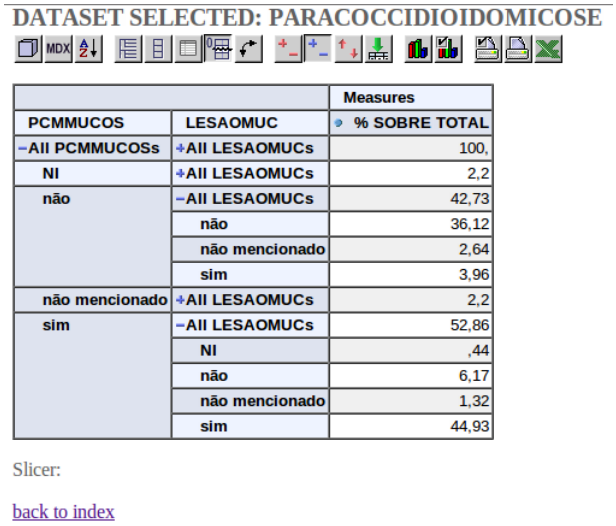
Figura 4.2. Tempo médio de evolução da doença dos pacientes tabagistas

Outro ponto apresentado pela análise do agrupamento é a existência de pacientes que apresentaram PCM Mucosa e não apresentaram lesão na mucosa, e vice-versa. A figura 4.3 mostra a análise para esse caso, e que corrobora que, na base dados atual 3,96% dos pacientes não foram diagnosticados com PCM mucosa e apresentaram lesão mucosa, e 6,17% dos pacientes diagnosticados com PCM Mucosa não apresentaram na lesão nas mucosas.

Em 4.4 temos que a PCM Mucosa (52,85%) é mais predominante que a PCM Cutânea (42,73%) e 20,26% dos pacientes apresentaram os dois tipos de PCM. A figura 4.5 apresenta a relação da localidade das lesões: mucosa *versus* cutânea. O percentual de pacientes que apresentaram ambas as lesões foi de 19,38%, que é bem próximo dos pacientes que apresentaram as duas PCM.

Análise 2

A tabela 4.7 apresenta o resultado do algoritmo de agrupamentos analisando a região de nascimento, sexo, e 4 variáveis de exames clínicos: "crepitacao", "altotorr" (alteração otorrinolaringológica), "rnialter" (alteração AP/RNI), "urinaalt" (alteração na urina) e "rxaltera" (raio x alterado). Um análise inicial mostra que o cluster #2 é o grupo das mulheres que apresentou alteração no raio x, crepitação normal e os outros exames



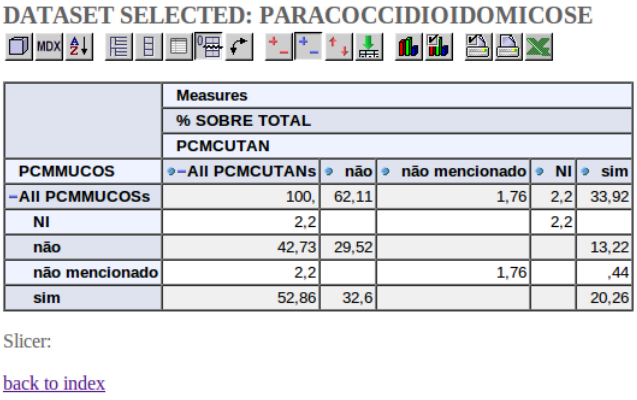
DATASET SELECTED: PARACOCCIDIOIDOMICOSE

		Measures
PCMMUCOS	LESAOMUC	% SOBRE TOTAL
−All PCMMUCOSs	+All LESAOMUCs	100,
NI	+All LESAOMUCs	2,2
não	−All LESAOMUCs	42,73
	não	36,12
	não mencionado	2,64
	sim	3,96
não mencionado	+All LESAOMUCs	2,2
sim	−All LESAOMUCs	52,86
	NI	,44
	não	6,17
	não mencionado	1,32
	sim	44,93

Slicer:

back to index

Figura 4.3. Pacientes acometidos pela PCM Mucosa mas não apresentaram lesão



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

	Measures				
	% SOBRE TOTAL				
	PCMCUTAN				
PCMMUCOS	−All PCMCUTANs	não	não mencionado	NI	sim
−All PCMMUCOSs	100,	62,11	1,76	2,2	33,92
NI	2,2			2,2	
não	42,73	29,52			13,22
não mencionado	2,2		1,76		,44
sim	52,86	32,6			20,26

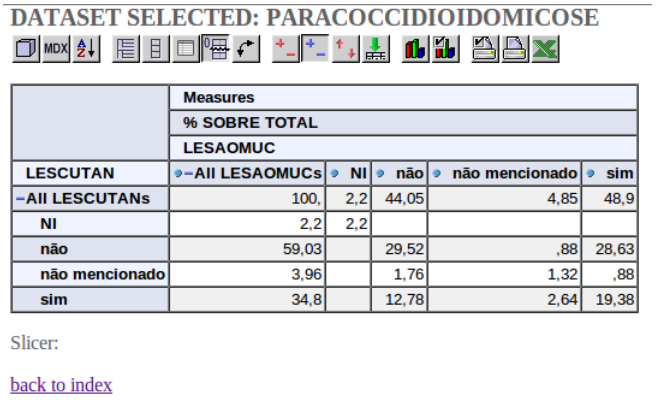
Slicer:

back to index

Figura 4.4. Relação entre a PCM Mucosa e Cutânea

não foram avaliados. Os pacientes do grupo #0 são caracterizados por serem homens da região central que apresentou alteração apenas no raio x. O cluster #3 é similar, contundo os exames "rnialter", "urinaalter" e "rxaltera" não foram avaliados. O grupo #1 é caracterizado por ser homens localizados na região do vale do rio doce que apresentaram alterações nos exames "crepitacao", "altotor" e "rxaltera"; sendo que "renialter" e "urinaalter" não foram avaliados.

O resultado deste agrupamento chamou atenção dos especialistas pelo fato de existir um grupo que o raio x não foi avaliado (cluster #3) com um grande percentual de pacientes no grupo. Esta análise foi aprofundada utilizando a análise OLAP e pode



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

	Measures				
	% SOBRE TOTAL				
	LESAOMUC				
LESCUTAN	−All LESAOMUCs	NI	não	não mencionado	sim
−All LESCUTANs	100,	2,2	44,05	4,85	48,9
NI	2,2	2,2			
não	59,03		29,52	,88	28,63
não mencionado	3,96		1,76	1,32	,88
sim	34,8		12,78	2,64	19,38

Slicer:

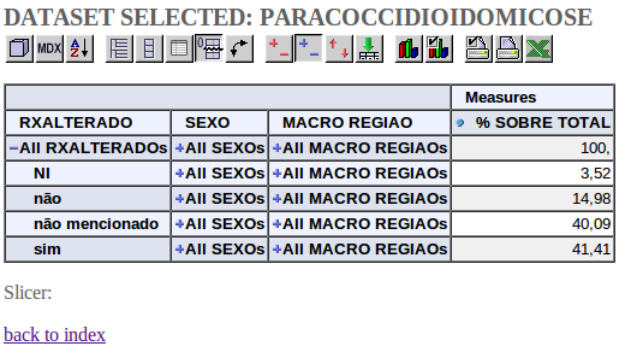
back to index

Figura 4.5. Localidade das lesões

Atributo	Grupo Global	#0	#1	#2	#3
	227(100%)	83(37%)	48(21%)	22(10%)	74(33%)
macroreg	4	4	1	1	4
sexo	M	M	M	F	M
crepitacao	ok	ok	x	ok	ok
altotor	ok	ok	x	-	ok
rnialter	-	ok	-	-	-
urinaalt	-	ok	-	-	-
rxaltera	x	x	x	x	-

Tabela 4.7. Resultado do algoritmo de agrupamento

ser comprovada que há uma boa quantidade de pacientes que não tiveram o raio x analisado ou que não apresentaram alteração no exame (figura 4.6 e 4.7).



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

			Measures
RXALTERADO	SEXO	MACRO REGIAO	% SOBRE TOTAL
−All RXALTERADOs	+All SEXOs	+All MACRO REGIAOs	100,
NI	+All SEXOs	+All MACRO REGIAOs	3,52
não	+All SEXOs	+All MACRO REGIAOs	14,98
não mencionado	+All SEXOs	+All MACRO REGIAOs	40,09
sim	+All SEXOs	+All MACRO REGIAOs	41,41

Slicer:

back to index

Figura 4.6. Análise de raio x alterado

Análise 3

Outro agrupamento realizado utilizando as variáveis de identificação da classe "vômitos" é apresentado na tabela 4.8. O cluster #2 são os pacientes que não tiveram essas variáveis mencionadas no prontuário, enquanto que no cluster #3 nenhuma

não	−All MACRO REGIAOs	14,98
	Central	5,29
	Centro-Oeste	,88
	Jequitinhonha/Mucuri	,44
	Outros Estados	2,64
	Sul/Sudoeste	,44
	Vale do Rio Doce	3,08
	Zona da Mata	1,32
não mencionado	−All MACRO REGIAOs	40,09
	Central	14,1
	Centro-Oeste	3,52
	Jequitinhonha/Mucuri	4,41
	NI	1,32
	Norte/Nordeste	,88
	Outros Estados	3,96
	Sul/Sudoeste	2,2
	Vale do Rio Doce	5,73
	Zona da Mata	3,52
sim	−All MACRO REGIAOs	41,41
	Central	6,61
	Centro-Oeste	4,85
	Jequitinhonha/Mucuri	4,85
	NI	,44
	Norte/Nordeste	1,32
	Outros Estados	6,17
	Sul/Sudoeste	3,08
	Triângulo/Alto Paranaíba	,88
	Vale do Rio Doce	5,73
	Zona da Mata	5,73

Figura 4.7. Análise de raio x alterado

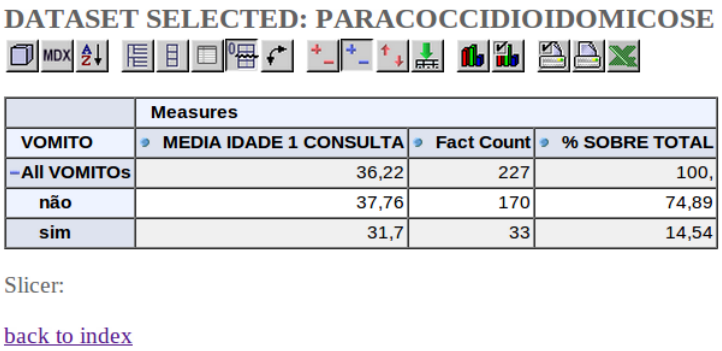
alteração foi encontrada nas análises desses atributos clínicos. Os grupos #1 e #4 apresentam os pacientes não tiveram alterações na maioria dos exames. Em #1 os pacientes apresentaram odinofagia ("odinofag") e os pacientes do *cluster* #4 foram internados. O grupo #0 é o grupo dos pacientes que apresentaram o quadro de vômitos, dor abdominal, sudorese, dispnéia e foram internados, contudo, não relataram odinofagia.

Os atributos marcantes desse agrupamento foram o "vomitos", "internacao" e "odinofag", os quais, um novo agrupamento utilizando apenas essas variáveis foi elaborado. Contudo, não houve alteração no resultado.

Atributo	**Grupo Global**	**#0**	**#1**	**#2**	**#3**	**#4**
	227(100%)	48(21%)	50(22%)	16(7%)	59(26%)	54(24%)
sudorese	ok	x	ok	-	ok	ok
odinofag	ok	ok	x	-	ok	ok
dorabdom	ok	x	ok	-	ok	ok
dispneia	ok	x	ok	-	ok	ok
internacao	x	x	ok	-	ok	ok
vomitos	ok	x	ok	-	ok	ok

Tabela 4.8. Resultado do algoritmo de agrupamento

Com as análises desses grupos, observa-se que vômito em pacientes com PCM não é muito comum, o qual 74,89% não apresentam esse quadro clínico onde pode ser observado na figura .



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

	Measures		
VOMITO	MEDIA IDADE 1 CONSULTA	Fact Count	% SOBRE TOTAL
All VOMITOs	36,22	227	100,
não	37,76	170	74,89
sim	31,7	33	14,54

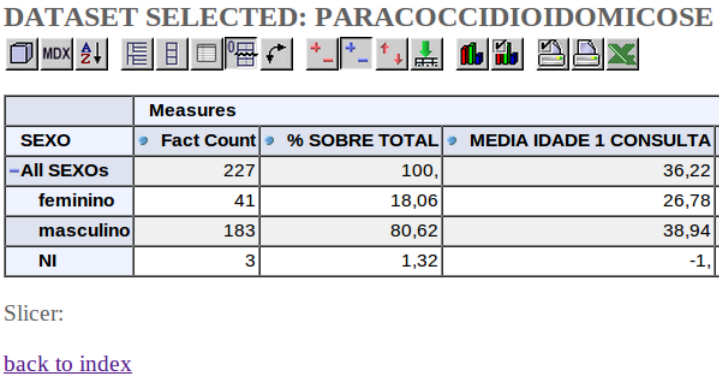
Slicer:

back to index

Figura 4.8. Relação de pacientes que apresentaram quadro de vômito

Análise 4

A grande maioria dos pacientes portadores de PCM são homens de Moura [2008], contundo duas análises foram elaboradas para um estudo sobre o comportamento da doença nas mulheres. Na figura 4.9 mostra que a média da idade das mulheres na primeira consulta (26,78 anos) é menor que a dos homens (38,94 anos). Além disso, o tempo de evolução da doença nas mulheres (9,98 meses) é quase 60% menor que a dos homens (17,15 meses). Já a figura 4.10 apresenta a relação das PCM Mucosa e Cutânea para as mulheres. De 41 pacientes, 24 (58,5%) não apresentaram a doença nessa forma, 4 (9,7%) apresentaram apenas PCM mucosa, 7(17%) a forma cutânea da doença, e 5(12,2%) ambas as formas.



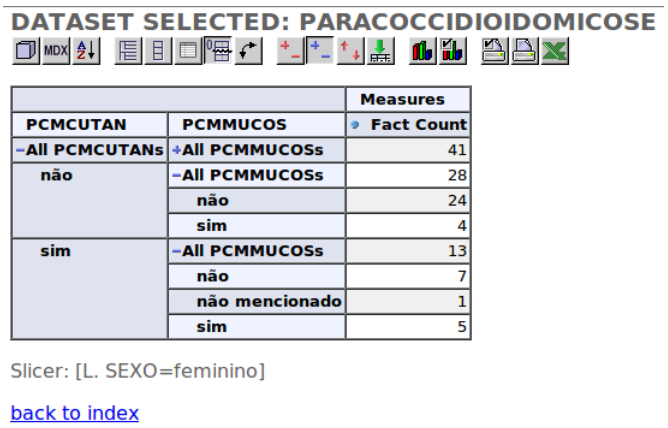
DATASET SELECTED: PARACOCCIDIOIDOMICOSE

	Measures			
SEXO	Fact Count	% SOBRE TOTAL	MEDIA IDADE 1 CONSULTA	TEMPO MEDIO EVOLUÇÃO
All SEXOs	227	100,	36,22	15,62
feminino	41	18,06	26,78	9,98
masculino	183	80,62	38,94	17,15
NI	3	1,32	-1,	-1,

Slicer:

back to index

Figura 4.9. Relção entre a idade média da primeira consulta de Homens x Mulheres



DATASET SELECTED: PARACOCCIDIOIDOMICOSE

		Measures
PCMCUTAN	PCMMUCOS	Fact Count
−All PCMCUTANs	+All PCMMUCOSs	41
não	−All PCMMUCOSs	28
	não	24
	sim	4
sim	−All PCMMUCOSs	13
	não	7
	não mencionado	1
	sim	5

Slicer: [L. SEXO=feminino]

back to index

Figura 4.10. Relação da PCM mucosa e cutânea em mulheres

Capítulo 5

Ferramentas Relacionadas

Existem ferramentas de sistema de apoio a decisão disponíveis para uso gratuitamente.

5.1 Pentaho

5.2 SpagoBI

Referências Bibliográficas

BI, S. (2012). Spago bi. `http://spagobi.eng.it/`.

de Moura, A. C. L. (2008). *Estudo Clinico e Imunologico de Controle de Cura de Paracoccidioidomicose Cronica*. PhD thesis, Universidade Federal de Minas Gerais.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157--1182.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11.

Han, J. & Kamber, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, second edi edição.

Hibernate (2012). Hibernate. `http://www.hibernate.org/`.

Hinton, M. D. (1995). *Laboratory Management Systems*. Marcel Dekker, inc. New York.

JPivot (2012). Jpivot. `http://jpivot.sourceforge.net/`.

Kimball, R. & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley Computer Publishing, second edi edição.

LGPL, G. (2012). Gnu lgpl. `http://www.gnu.org/licenses/lgpl-2.1.html`.

Mondrian (2012). Mondrian. `http://mondrian.pentaho.com/`.

MySQL (2012). Mysql. `http://www.mysql.com/`.

Pentaho (2012). Pentaho. `http://www.pentaho.com/`.

R (2012). R project. `http://www.r-project.org/`.

RESTREPO, A.; McEWEN, J. & CASTANEDA, E. (2001). The habitat of Paracoccidioides brasiliensis: how far from solving the riddle? *Med. Mycol.*, 39:233–241.

Ricci, V. (2004). R : un ambiente opensource per lánalisi statistica dei dati. *Economia e Commercio*, 1:69--82.

Santos, W. A. d.; Silva, B. M. d.; Passos, E. D.; Zandonade, E. & Falqueto, A. (2003). AssociaÃ§Ã£o entre tabagismo e paracoccidioidomicose: um estudo de caso-controle no estado do espÃrito santo, brasil. *Cadernos de SaaÃºde PÃºblica*, 19:245 – 253.

Shikanai-Yasuda, M. A.; Telles Filho, F. d. Q.; Mendes, R. P.; Colombo, A. L. & Moretti, M. L. (2006). Consenso em paracoccidioidomicose. *Revista da Sociedade Brasileira de Medicina Tropical*, 39:297 – 310.

SIMOES, A.; FARIA-CAMPOS, A. C.; DELAAT, D.; ABREU, V. & CAMPOS, S. V. A. (2010). Sigla: An adaptable lims for multiple laboratories. *BMC Genomics*.

Tomcat (2012). Apache tomcat. `http://tomcat.apache.org/`.

Urbanek, S. (2003). Rserve - a fast way to provide r functionality to applications. Em *The 3rd International Workshop on Distributed Statistical Computing*.

Witten.I; Frank.E & Hall.M (2011). *Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, third edition edição.

Yvan, S.; Inaki, I. & Pedro, L. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507â2517.

Anexo A

JPivot Tag Library

jpivot (en)

JPivot Tag Library.

Content

---

chart

Body Content JSP

Description Creates a Chart component. This component does not produce visible output directly, it must be rendered via the WCF `render` tag.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
visible	false	boolean	true	Sets the visible property of the component. The WCF Render tag will not show a component whose visible flag is false
role	false	String	true	Name of a role, optionally prefixed by '!'. Example: role='tomcat' allows members of the role tomcat to access this component, role='!tomcat' grants access to everybody except tomcat members.
query	true	OlapModel	true	Name of the Session Attribute that contains the OLAP Data (query and result) for this tag.
baseDisplayURL	false	String	true	Base link to chart display servlet. Will be parameterized with '?filename=<temp chart image file name>'
controllerURL	false	String	true	Base link to web application controller for JPivot. Useful for complicated environments like a portal, where you would use a portlet:renderURL

---

chooseQuery

Body Content JSP

Description chooses a query that has been previously created with a queryName attribute.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
queryName	false	String	true	name of the query to choose

Example

```
<jp:mondrianQuery id="query01" queryName="name1">
   SELECT ...
</jp:mondrianQuery>
<jp:mondrianQuery id="query01" queryName="name2">
   SELECT ...
</jp:mondrianQuery>
...
<jp:chooseQuery id="query01" queryName="name1"/>
```

---

clickable

Body Content EMPTY

Description Makes all members of a dimension or level clickable (i.e. generates an anchor). The generated URL contains the unique name of the member. Must be nested inside a table or query tag.

The behaviour depends on the sessionParam attribute. If its present, then the parameter value will be written into the com.tonbeller.jpivot.param.SessionParamPool before the new page is shown. If its not present, the parameter will be coded into the hyperlink and must be treated in some way by the receiving page.

This tag must be nested either inside a table tag or inside a query tag. Inside the table tag the clickable is available for all queries, inside the query tag its only available for that query.

Attribute	Required	Type	rtexpr	Description
urlPattern	false	String	true	DEPRECATED - do not use in conjunction with session parameters because the browser back button does not work as expected. Use `page` attribute instead. The pattern to construct the url. If sessionParam attribute is NOT present, then "{0}" will be replaced with the unique name of the member. If urlPattern starts with "/" its relative to the application context, otherwise its absolute
page	false	String	true	Name of the target page, must start with "/". If not set the current page is re-displayed.
uniqueName	true	String	true	unique name of a dimension, hierarchy or level, whose members will be clickable
menuLabel	false	String	true	if there are multiple clickables specified, they will be displayed in a popup menu, this text will be the label for the menu entry
sessionParam	false	String	true	name of the parameter, that will be placed into the com.tonbeller.jpivot.param.SessionParamPool
propertyName	false	String	true	if present, the value of that member property will be taken for the com.tonbeller.jpivot.param.Parameter sqlValue property
propertyPrefix	false	String	true	if present, multiple SessionParam will be created, one for each member property whose name starts with propertyPrefix. The name of the SessionParam will be the name of the reminder of the member properties name after the prefix. If present, the attributes sessionParam and propertyName are not allowed.
providerClass	false	String	true	if present, an instance of this class will extract the SessionParam instances from the member. The class must implement com.tonbeller.jpivot.table.navi.ClickableMember.ParameterProvider. The attributes sessionParam, propertyName, propertyPrefix are not allowed in this case.

Example

```
<jp:mondrianQuery ...>
  select .. from Sales
  <jp:clickable urlPattern="/otherpage.jsp?param={0}" uniqueName="[Customers]"/>
  <jp:clickable page="/yetotherpage.jsp" uniqueName="[Products].[Category]" sessionParam="Category"/>
</jp:mondrianQuery>
```

---

destroyQuery

Body Content JSP

Description destroys all queries stored with this id

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag

Example

```
<jp:destroyQuery id="query01"/>
```

---

mondrianQuery

Body Content JSP

Description Creates a session attribute that allows access to a Mondrian query, its result and navigations. *The attribute is always created, if the attribute already existst it will be replaced.* So one has to take care that this tag creates the attribute only if it should, e.g. by using JSTL <c:if ... > tags or similar.

Its possible to use a predefined JSBC DataSource from the application server or to specify the individual JDBC parameters with this tag.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
jdbcDriver	false	String	true	Java class name of the JDBC driver
jdbcUrl	false	String	true	Connection parameters, these are database specific
jdbcUser	false	String	true	user name to connect to the database
jdbcPassword	false	String	true	password to connect to the database
dataSource	false	String	true	JNDI name of a preconfigured JDBC DataSource, e.g. jdbc/SironTDI. Must not be used with the other JDBC attributes
catalogUri	true	String	true	Path of the Mondrian schema, relative to the application context, e.g. `/WEB-INF/FoodMart.xml`
config	false	String	true	internal use only
role	false	String	true	Role from Mondrian Schema
dynResolver	false	String	true	class resolving Mondrian Schema dynamic variables
dynLocale	false	String	true	Locale for dynamic Mondrian Schema Resolver
connectionPooling	false	String	true	"false" will prevent Mondrian from Connection Pooling
dataSourceChangeListener	false	String	true	class to detect changes in datasource
queryName	false	String	true	Allows to keep multiple queries within this session attribute. For every queryName, the last query will be stored. Use chooseQuery tag to switch between queryName's or queries.
stackMode	false	boolean	true	If set to false, all queryNames are treated equally, independent of the order they are created or shown using the chooseQuery tag. If set to true it will keep the different queryName's in a stack. Example: • A query with queryName `qn1` is created. Then this tag will display that query • A query with queryName `qn2` is created next. Now there is a stack containing `qn1` and `qn2`. `qn2` will be the visible query. • At this point, for example, if the user switches back to `qn1` using the chooseQuery tag, then `qn2` will be destroyed in stack mode. It would not be destroyed in non-stackMode. So in non-stackMode the user can arbitrary choose one of the different queries. In stackMode, if the user goes back to a previously seen query, all "younger" queries will be destroyed.

Example

```
<jp:mondrianQuery id="query01"
  jdbcDriver="com.mysql.jdbc.Driver"
  jdbcUrl="jdbc:mysql://localhost/foodmart"
  catalogUri="/WEB-INF/test/FoodMart.xml">
select
  {[Measures].[Unit Sales], [Measures].[Store Cost], [Measures].[Store Sales]} on columns,
  {[Product].[All Products]} ON rows
from Sales
where ([Time].[1997])
</jp:mondrianQuery>
```

---

navigator

Body Content JSP

Description Creates the Navigator component. This component does not produce visible output directly, it must be rendered via the WCF `render` tag.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
query	true	OlapModel	true	Name of the Session Attribute that contains the OLAP Data (query and result) for this tag.
visible	false	boolean	true	Sets the visible property of the component. The WCF Render tag will not show a component whose visible flag is false
role	false	String	true	Name of a role, optionally prefixed by '!'. Example: role='tomcat' allows members of the role tomcat to access this component, role='!tomcat' grants access to everybody except tomcat members.

Example

```
<jp:navigator id="navi01" query="#{query01}" visible="false"/>
<wcf:render ref="navi01" xslUri="/WEB-INF/jpivot/navi/navigator.xsl" xslCache="true"/>
```

---

print

Body Content JSP

Description Creates the Print component. This component does not produce visible output directly, you must create a WCF form to configure and call the print servlet to render to XLS/PDF

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag

---

scalarQuery

Body Content JSP

Description Creates a session attribute that contains an OLAP result consisting of a single cell. The values are provide through tag attributes.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
value	true	String	true	EL expression evaluating to the value (number)
formattedValue	false	String	true	EL expression evaluating to the formatted value (String)
caption	false	String	true	EL expression evaluating to the caption (String)
queryName	false	String	true	see mondrianQuery
stackMode	false	boolean	true	see mondrianQuery

Example

```
<jp:scalarQuery
  id="query01"
  value="#{some.bean.property}">
  formattedValue="#{some.bean.otherProperty}"
  caption="Some Caption" />
```

---

selectproperties

Body Content JSP

Description Creates the Select Properties component, which allows to specify which properties shall be visible and in which orderd. This component does not produce visible output directly, it must be rendered via the WCF `render` tag.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
table	true	TableComponent	true	Name of the Session Attribute that contains the Table Component for this tag.
visible	false	boolean	true	Sets the visible property of the component. The WCF Render tag will not show a component whose visible flag is false
role	false	String	true	Name of a role, optionally prefixed by '!'. Example: role='tomcat' allows members of the role tomcat to access this component, role='!tomcat' grants access to everybody except tomcat members.

Example

```
<jp:selectproperties id="selectprop01" table="#{table01}" visible="false"/>
<wcf:render ref="selectprop01" xslUri="/WEB-INF/jpivot/navi/navigator.xsl" xslCache="true"/>
```

---

setParam

Body Content JSP

Description Sets a mdx query parameter from an http parameter or from a session parameter. The body is evaluated only if the http parameter is present, so its a good place to contain a mondrian query. If you use the session Parameter, the body is never evaluated. Exactly one attribute either httpParam or sessionParam must be set.

Attribute	Required	Type	rtexpr	Description
query	true	OlapModel	true	Name of the Session Attribute that contains the OLAP Data (query and result) for this tag.
httpParam	false	String	true	Name of a http parameter. If present, its value will be parsed and set into the mdx parameter
sessionParam	false	String	true	name of the session parameter
mdxParam	true	String	true	Name of the MDX Parameter in the query to modify

Example

```
<jp:mondrianQuery id="query01"...>
  SELECT ... Parameter("Param01", ...)
  WHERE ...
</jp:mondrianQuery/>

<jp:setParam query="query01" httpParam="param" mdxParam="Param01"/>
or
<jp:setParam query="query01" sessionParam="CUSTOMER" mdxParam="Param01"/>
```

---

table

Body Content JSP

Description Creates a Pivot Table component. This component does not produce visible output directly, it must be rendered via the WCF `render` tag.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
visible	false	boolean	true	Sets the visible property of the component. The WCF Render tag will not show a component whose visible flag is false
role	false	String	true	Name of a role, optionally prefixed by '!'. Example: role='tomcat' allows members of the role tomcat to access this component, role='!tomcat' grants access to everybody except tomcat members.
query	true	OlapModel	true	Name of the Session Attribute that contains the OLAP Data (query and result) for this tag.
configXml	false	String	true	Path for a config file that allows to add customer specific code

Example

```
<jp:table id="table01" query="#{query01}" visible="true"/>
```

---

testQuery

Body Content JSP

Description Creates test data that can be displayed by the table or chart components

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
onRows	false	String	true	Whitespace separated list of dimensions to show on rows. Possible values are: Measures, Region, Products, Advertising, Material
onColumns	false	String	true	Whitespace separated list of dimensions to show on columns. Possible values are: Measures, Region, Products, Advertising, Material

Example

```
<jp:testQuery id="query01" onColumns="Measures" onRows="Products Region">
  for some reason, the body must not be empty
```

```
</jp:testQuery>
```

---

xmlaQuery

Body Content JSP

Description Creates a session attribute for an XMLA query. It will be used by components like table or navigator to display the result and navigate the cube. *The attribute is always created, if the attribute already exists it will be replaced.* So one has to take care that this tag creates the attribute only if it should, e.g. by using JSTL <c:if ... > tags or similar.

Attribute	Required	Type	rtexpr	Description
id	true	String	true	Name of the Session Attribute that is created by this tag
uri	true	String	true	The URI used to access the XMLA server
dataSource	false	String	true	DataSourceInfo specification like "Provider=MSOLAP;Data Source=local"
catalog	true	String	true	Catalog spezification like "Foodmart 2000"
config	false	String	true	Path to a config file that allows to integrat customer specific code

Example

```
<jp:xmlaQuery id="query01"
  uri="http://MYSERVER/XML4A/msxisapi.dll"
  catalog="Foodmart 2000">
select
  {[Measures].[Unit Sales], [Measures].[Store Cost], [Measures].[Store Sales]} on columns,
  {[Product].[All Products]} ON rows
from Sales
where ([Time].[1997])
</jp:xmlaQuery>
```

Anexo B

Protocolo para a primeira consulta de pacientes com paracoccidioidomicose

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE MEDICINA - HOSPITAL DAS CLÍNICAS
CENTRO DE TREINAMENTO E REFERÊNCIA EM DOENÇAS
INFECCIOSAS E PARASITÁRIAS – CTR-DIP

SAME [ ] [ ] [ ] [ ] [ ] [ ]

Protocolo [ ] [ ] [ ] [ ]

PROTOCOLO DE PARACOCCIDIOIDOMICOSE E MICOSES PROFUNDAS

PRIMEIRA CONSULTA

CODIFICAÇÃO

IDENTIFICAÇÃO:

1. Data do exame: [ ][ ][ ]
2. Nome: _____
3. Endereço: _____
4. Município: _____ 5.Estado: ____
6. Fone: [ ] _____ 7. Naturalidade: _____
8. Data de nascimento: [ ][ ][ ] 9. Idade: [ ] 10.Sexo: [ ]
11. Cor: [ ] 12. Profissão: _____13. Estado civil: [ ]
14. Escolaridade:_____15. Informante:

QUEIXA PRINCIPAL:

HISTÓRIA DA MOLÉSTIA ATUAL:

17. Início dos sintomas _____

	S	N	NA	Cód.
18. Febre				
19. Emagrecimento [ ] [ ]Kg				
20. Astenia/hipodinamia				
21. Sudorese				
22. Linfadenomegalias				
23. Lesão cutânea				
24. Lesão mucosa				
25. Prurido				
26. Icterícia				
27. Colúria				
28. Hipo ou acolia				
29. Palpitação				
30. Tonteira				
31. Cefaléia				
32. Convulsão				
33. Obstrução nasal				
34. Rouquidão				
35. Disfonia				

	S	N	NA	Cód.
36. Disfagia				
37. Odinofagia				
38. Azia/pirose				
39. Vômitos				
40. Dor abdominal				
41. Distensão abdominal				
42. Diarréia				
43. Constipação				
44. Hemorragia digestiva				
45. Artralgia ou artrite				
46. Edema				
47. Dor óssea				
48. Dor torácica				
49. Dispnéia				
50. Tosse				
51. Expectoração				
52. Chieira torácica				

SAME: [ ][ ][ ][ ][ ][ ]

HISTÓRIA PREGRESSA (incluir história epidemiológica: mudanças de localidade, tempo em cada uma, contato com culturas agrícolas, etc):

PACIENTE EM USO ATUAL DE:

TRATAMENTO PRÉVIO DE PCM?

[ ] SIM; QUAL? QUANDO?

[ ] NÃO

Doenças Prévias:	S	N	NS	Cód.
53. Asma				
54. Úlcera péptica				
55. Tuberculose				
56. Esquistossomose				
57. Sífilis				
58. SIDA				
59. Leishmaniose				
60. Neoplasia				
61. Doença do SNC				
62. Internações prévias				
63. Cirurgias prévias				
64. Pneumoconioses				
65. DST				
66. Outras				

Doenças Atuais:	S	N	NS	Cód.
67. Hipertensão arterial				
68. ICC / ICO				
69. DPOC				
70. Asma				
71. Ulcera péptica				
72. Tuberculose				
73. Esquistossomose				
74. Sífilis				
75. SIDA				
76. Leishmaniose				
77. Neoplasia				
78. Doença do SNC				
79. Doenças do Colágeno				
80. Doença de Chagas				
81. Síndrome de Addison				
82. DST				
83. Outras				

HISTÓRIA FAMILIAR:

	S	N	NS	Cód.
84. PCM				
85. Doença cardiovascular				
86. Doenças respiratórias				
87. Diabetes mellitus				
88. Neoplasias				
89. Outra doença infecciosa				
90. Doença de Chagas				
91. Esquistossomose				

HISTÓRIA SÓCIO-ECONÔMICA (descrever):

	S	N	NS	Cód.
92. Tabagismo atual				
93. Tabagismo prévio				
94. Etilismo atual				
95. Etilismo prévio				
96. Drogas ilícitas				
97. Contato com área rural				
98. Água em moradia				
99. Esgoto				
100. Trabalho com asbesto				
101. Trabalho com sílica				
102. Trabalho em minas				

SAME: [ ] [ ] [ ] [ ] [ ] [ ]

EXAME FÍSICO

103. Peso: [ ] [ ] [ ] , [ ] Kg
104. Altura: [ ] [ ] [ ] cm
105. PA deitado: [ ] X [ ] mmHg
106. PA assentado: [ ] X [ ] mmHg
107. PA em pé: [ ] X [ ] mmHg
108. Freq. Cardíaca: [ ] [ ] [ ] bpm
109. Freq. Pulso: [ ] [ ] [ ] ppm
110. Freq. Respiratória: [ ] [ ] irpm

	S	N	NA	Cód.
111. Hipotensão ortostática				
112. Palidez cutâneo-mucosa				
113. Icterícia				
114. Desidratação				
115. Cianose				
116. Hiperpigmentação				
117. Linfadenomegalias				
118. Localização: Cervical				
119. Supra / infraclavicular				
120. Axilar				
121. Inguinal				
122. Epitroclear				
123. Outros				
124. Doloroso				
125. Consistência endurecida				
126. Aderido				
127. Fistulização				
128.Tamanho > 1 cm [ ] cm				
129. Lesão mucosa.				
130. Localização: Boca				
131. Mucosa nasal/septo				
132. Faringe				
133. Laringe				
134. Genital				
135. Ânus / Reto				
136. Outra				
137.Tipo da lesão: Pápula.				
138. Nódulo				
139. Úlceras				
140. Vegetações				
141. Estomatite moriniforme				
142. Infecção secundária				

	S	N	NA	Cód.
143. Lesão cutânea.				
143. Tipo de lesão: Pápula				
144. Nódulo				
145. Úlceras				
146. Vegetações				
147. Placas				
148. Abscesso				
149. Área de lesões [ ] cm^2				
150. Localização: Cabeça				
151. Pescoço				
152. Tórax				
153. MMSS				
154. MMII				
155. Abdomen				
156. Genital				
157. Perianal / glútea				
158. Infecção secundária				
159. Pontilhado hemorrágico				
160. Edema				
161. Localização: MMII				
162. MMSS				
163. Palpebral				
164. Anasarca				
165. Tireóide anormal				
166. Nódulo palpável				
167. Bócio difuso				

Descrever linfadenomegalias por cadeias (tamanho, consistência, localização, mobilidade, etc)

APARELHO RESPIRATÓRIO:

	S	N		Cód
168. Deformidade torácica				
169. Expansibil. prejudicada				
170. Murmúrio abolido				
171. Roncos				
172. Sibilos				
173. Crepitações				
174. Atrito pleural				
175. FTV alterado				
176. Percussão alterada				

SAME: [ ] [ ] [ ] [ ] [ ] [ ]

APARELHO CARDIOVASCULAR

	S	N		Cód
177. Pulsos MMSS alterados				
178. Pulsos MMII alterados				
179. Pulsos centrais alterados				
180. Ictus cordis alterado				
181. Frêmito				
182. Bulhas alteradas				
183. Estalido protossistólico VD				
184. Estalido protossistólico VE				
Sopro cardíaco: 185.SS IM				
186. SS EA				
187. SD EM				

APARELHO DIGESTIVO

	S	N		Cód
200. Fígado palpável				
201. Tamanho: [ ] [ ] cm RCD 202. [ ] [ ] cm AX				
203. Baço palpável				
204. Tamanho: Boyd [ ]				
205. Sinais de irritação peritoneal				
206. Presença de massa palpável				

Fígado: borda, sensibilidade, consistência, superfície
Massa palpável: localização, tamanho, sensibilidade, consistência

EXAME NEUROLÓGICO

	S	N		Cód
207. Déficit focal				
208. Distúrbio do comportamento				
209. Distúrbio da consciência				
210. Distúrbio da marcha				
211. Papiledema				

HIPÓTESES DIAGNÓSTICAS:

1. _____
2. _____
3. _____
4. _____
5. _____

CONDUTA:

1. Exames solicitados:_____
_____
2. Prescrição: _____
_____
3. Orientações: _____
_____
4. Retorno em : ____ / ____ / ____

Coleta de sangue para ICB hoje: sim ( ) não( )

_____
Assinatura / Carimbo do médico e CRM

SAME: [ ] [ ] [ ] [ ] [ ] [ ]

DATA												
Hb												
Hm												
Hct												
VCM												
HCM												
CHCM												
G Leuct												
Seg %												
Bast %												
Eosin %												
Basof %												
Mono %												
Linf %												
Linf atip %												
Plaquetas												
AP / RNI												
PTT												
VHS												
PCR												
ASLO												
TGO												
TGP												
FA												
GGT												
Bb T/Bb D												
Proteinas T												
Albumina												
Glicemia												
Uréia												
Creatinina												
Na												
K												
Mg												
Cl												
P												
Amilase												
Lipase												
Colesterol T												
HDL												
LDL												
VLDL												
Triglicérides												
CPK												
CK-MB												
Ác. Úrico												
Ác. Lático												
T. tuberculin												
Fe sérico												
Ferritina												

DATA												
VDRL												
HBsAg												
Anti-HBs												
HBeAg												
Anti-HBc												
Anti-HCV												
Anti-HAV												
Anti-HIV												
Anti-EBV												
Toxoplasm												
CMV												
Chagas												

DATA				
Urina Rotina				
EPF				
Biópsia:Local /				
Achado				
Cultura:Fonte /				
Achado				
Escarro:				
Raio X de tórax				
Fibronasolaringoscopia				
Ultra som abdominal				
TC tórax				
ECG				
Espirometria				