UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ISTAR: UM ESQUEMA ESTRELA OTIMIZADO PARA IMAGE DATA WAREHOUSES BASEADO EM SIMILARIDADE

LUANA PEIXOTO ANNIBAL

ORIENTADOR: PROF. DR. RICARDO RODRIGUES CIFERRI

CO-ORIENTADOR: PROF. DR. JOAQUIM CEZAR FELIPE

São Carlos - SP
Agosto/2011

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ISTAR: UM ESQUEMA ESTRELA OTIMIZADO PARA IMAGE DATA WAREHOUSES BASEADO EM SIMILARIDADE

LUANA PEIXOTO ANNIBAL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Engenharia de Software e Banco de Dados.
Orientador: Dr. Ricardo Rodrigues Ciferri
Co-Orientador: Joaquim Cezar Felipe

São Carlos - SP
Agosto/2011

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

A615ie

Annibal, Luana Peixoto.
Istar : um esquema estrela otimizado para *Image Data Warehouses* baseado em similaridade / Luana Peixoto Annibal. -- São Carlos : UFSCar, 2011.
145 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2011.

1. Ciência da computação. 2. Processamento de imagens. 3. ETL para imagens. 4. Estrutura de indexação de imagens. 5. OLAP. I. Título.

CDD: 004 (20ª)

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

"iStar: Um Esquema Estrela Otimizado para Image Data Warehouses baseado em Similaridade"

LUANA PEIXOTO ANNIBAL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação

Membros da Banca:

Ricardo Rodrigues Ciferri

Prof. Dr. Ricardo Rodrigues Ciferri
(Orientador - DC/UFSCar)

Prof. Dr. Joaquim Cezar Felipe
(Co-orientador - USP/RP)

Marilde Santos

Profa. Dra. Marilde Terezinha Prado Santos
(DC/UFSCar)

Profa. Dra. Valéria Cesário Times
(UFPE)

São Carlos
Agosto/2011

Dedico esta dissertação a Deus e à minha família

AGRADECIMENTO

Agradeço humildemente a Deus, ao Espírito Santo, à Santa Catarina e à nossa Senhora por toda a iluminação e proteção.

Aos meus pais, Italo e Celia, por todo amor, carinho e apoio que estes tem me dado em toda a minha vida.

Ao meu amado, Gustavo, pelo apoio e dedicação, que fizeram que esta etapa fosse vencida com mais leveza e tranquilidade.

À minha avó, ao Micico, à Renata e às minhas queridas crianças, Luisa, Henrique, Lucas e Bruninho, e aos demais familiares pela alegria e auxílio.

Aos meus amigos, Dú, Vanessa, Walter, Debora, Alê, Vínicius, Renata, Felipe, Jaqueline, Arthur, Ives, Thiago, Bruno e demais colegas por me acolherem em São Carlos e me ajudarem em todos esses dois anos e meio.

Às eternas amigas, Gisele, Daniane, Mariane e Lucimara, por, mesmo distantes, estarem sempre ao meu lado.

Agradeço especialmente ao Prof. Ricardo Rodrigues Ciferri, por todo ensinamento, orientação, auxílio e principalmente por ser palmeirense.

Agradeço aos professores Marilde Terezinha Prado Santos, Renato Bueno e Caetano Traina Jr., e especialmente aos professores Cristina Dutra Aguiar Ciferri e Joaquim Cezar Felipe por toda ajuda prestada, atenção e conselhos na realização deste trabalho.

A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original

Albert Einstein

RESUMO

Um ambiente de *data warehousing (DWing)* auxilia seus usuários a tomarem decisões a partir de investigações e análises dos dados de maneira organizada e ágil. Entretanto, os atuais recursos de DWing não possibilitam que o processo de tomada de decisão seja realizado com base em comparações do conteúdo intrínseco de imagens. Esta análise não pode ser realizada por aplicações de DW convencionais porque essa utiliza, como base, imagens digitais e necessita realizar operações baseadas em similaridade, para as quais um DW convencional não oferece suporte. Neste trabalho, é proposto um ambiente de *data warehouse* chamado iCube que provê suporte ao processamento de consultas IOLAP (*Image On-Line Analytical Processing*) baseadas em diversas percepções de similaridade entre as imagens. O iCube realiza adaptações nas três principais fases de um ambiente de *data warehousing* convencional para permitir o uso de imagens como dados de um *data warehouse* (DW). Para a fase de integração, ou fase ETL (*Extract, Trasnform and Load*), nós propomos um processo para representar as imagens a partir de seu conteúdo intrínseco (i.e., por exemplo por meio de descritores numéricos que representam cor ou textura dessas imagens) e integrar esse conteúdo intrínseco a dados convencionais em um DW. Neste trabalho, nós também propomos um esquema estrela otimizado para o iCube, denominado iStar, que armazena tanto dados convencionais quanto dados de representação do conteúdo intrínseco das imagens. Ademais, nesta fase, o iStar foi projetado para representar e prover suporte ao uso de diferentes camadas perceptuais definidas pelo usuário. Para a fase de análise de dados, o iCube permite que processos OLAP sejam executados com o uso de comparações de similaridade como predicado de consultas e com o uso de mecanismos de filtragem para acelerar o processamento de consultas OLAP. O iCube foi validado a partir de testes de desempenho para a construção da estrutura e para o processamento de consultas IOLAP. Os resultados demonstraram que o iCube melhora significativamente o desempenho no processamento de consultas IOLAP quando comparado aos atuais recursos de IDWing. Os ganhos de desempenho do iCube contra o melhor trabalho correlato (i.e. SingleOnion) foram de até 98,21%.

Palavras-chave: *Image Data Warehouse*, ETL para imagens, *Image On-Line Analytical Processing*, Consulta baseada em similaridade, iStar, Camadas Perceptuais, iCube, Onion-tree.

ABSTRACT

A data warehousing environment supports the decision-making process through the investigation and analysis of data in an organized and agile way. However, the current data warehousing technologies do not allow that the decision-making processe be carried out based on images pictorial (intrinsic) features. This analysis can not be carried out in a conventional data warehousing because it requires the management of data related to the intrinsic features of the images to perform similarity comparisons. In this work, we propose a new data warehousing environment called iCube to enable the processing of OLAP perceptual similarity queries over images, based on their pictorial (intrinsic) features. Our approach deals with and extends the three main phases of the traditional data warehousing process to allow the use of images as data. For the data integration phase, or ETL phase, we propose a process to represent the image by its intrinsic content (such as color or texture numerical descriptors) and integrate this data with conventional data in the DW. For the dimensional modeling phase, we propose a star schema, called iStar, that stores both the intrinsic and the conventional image data. Moreover, at this stage, our approach models the schema to represent and support the use of different user-defined perceptual layers. For the data analysis phase, we propose an environment in which the OLAP engine uses the image similarity as a query predicate. This environment employs a filter mechanism to speed-up the query execution. The iStar was validated through performance tests for evaluating both the building cost and the cost to process IOLAP queries. The results showed that our approach provided an impressive performance improvement in IOLAP query processing. The performance gain of the iCube over the best related work (i.e. SingleOnion) was up to 98,21%.

Keywords: *Image Data Warehouse*, ETL for images, *Image On-Line Analytical Processing*, Similarity-Based Query, iStar, Perceptual Layers, iCube, Onion-tree.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CBIR – *Content-Based Image Retrieval*

DW – *Data Warehouse*

DWing – *Data Warehousing*

DM – *Data Mart*

DICOM – *Digital Imaging and Communications in Medicine*

EMIAT – *European Myocardial Infarct Amiodarone Trial*

ETL – *Extract, Transform and Load*

ECG - Eletrocardiograma

EEG – Eletroencefalograma

iCube – *Image Cube*

IDW – *Image Data Warehouse*

IDWing – *Image Data Warehousing*

IOLAP – *Image OLAP*

Id – Identificador

iStar – *image Star schema*

mbOr – *minimum-bounding-Omni-region*

MOLAP – *OLAP multidimensional*

NIDS – *Neuroinformatics Database System*

OLAP – *On-Line Analytical Processing*

OLTP – *On-Line Transaction Processing*

PACS – *Picture Archiving and Communication System*

PET - Tomografia por Emissão de Pósitrons

ROLAP – *OLAP Relacional*

ROT - Relação de Ordem Total

SSD – Sistemas de Suporte à Decisão

SGBD – Sistema Gerenciador de Banco de Dados

SUMÁRIO

Capítulo 1

INTRODUÇÃO

Este capítulo apresenta o contexto, a motivação e os desafios que deram origem ao desenvolvimento desse projeto de pesquisa em nível de mestrado. Os objetivos são discutidos e as principais contribuições são destacadas, finalizando com a apresentação da organização da dissertação.

1.1 Considerações Iniciais

Em ambientes empresariais e de pesquisa, um conjunto específico de tecnologias, denominado *Data Warehousing* (DWing), tem sido desenvolvido com o objetivo de prover suporte ao processo de tomada de decisão estratégica (CHAUDHURI; DAYAL, 1997; INMON, 2005). Por meio de ferramentas de ETL (*Extract, Translate, Load*), uma aplicação de *data warehousing* permite a integração de dados oriundos de fontes de dados heterogêneas em uma única base de dados especializada, chamada Data Warehouse (DW).

Ambientes de *data warehousing (DWing)* promovem o suporte à tomada de decisão, pois proporcionam um meio apropriado para o armazenamento e a análise multidimensional de dados (CHAUDHURI; DAYAL, 1997; CIFERRI, 2002; KIMBALL; ROSS, 2002; SONG, 2009). Nestes ambientes, o próprio usuário de Sistema de Suporte a Decisão (SSD) pode realizar análises sobre os dados de maneira rápida e consistente a partir da confecção de relatórios, do uso de técnicas de mineração de dados e submissão *on-line* de consultas complexas e multidimensionais (INMON, 2005; RIZZI, 2006; SONG, 2009).

Devido à maturidade de ambientes de DWing e devido a sua característica de fácil adaptação a modelos de negócio, ambientes de DWing estão sendo desenvolvidos e utilizados em diversas áreas, tais como a área médica, agropecuária e ambiental. O desenvolvimento de ambiente de DWing para esses setores é mais complexo do que para o setor clássico empresarial devido à complexidade dos tipos de dados armazenados. Além de dados convencionais como dados numéricos, datas e textos curtos, armazena-se também dados complexos como texto livre, imagem, áudio e vídeo.

Esta dissertação enfocou especificamente em dados de imagens médicas, devido ao fato do armazenamento e a posterior recuperação de informações sobre imagens ser um requisito fundamental na área médica. No entanto, as propostas e conclusões apresentadas por este trabalho também podem ser aplicadas em outros setores que realizem a tomada de decisão utilizando imagens digitais, como o setor agropecuário, ambiental, dentro outros.

Em geral, as imagens médicas são armazenadas tanto em seu formato original, como exemplo *Digital Imaging and Communications in Medicine* (DICOM) e JPEG, quanto são armazenadas as características dessas imagens que facilitam o seu tratamento e posteriormente recuperação (FELIPE; JR.; TRAINA, 2009; TRAINA et al., 2009). Nesse sentido, um DW voltado para o armazenamento de imagens deve conter um conjunto de características relevantes das imagens que permita a sua recuperação eficiente.

Este trabalho de pesquisa em nível de mestrado propõe um ambiente de DWing que provê suporte ao uso de imagens médicas no processo de tomada de decisão estratégica. Para tanto, este trabalho propõe um ambiente de DWing que mantém as suas características de SSD e continua a ser ágil e flexível às intenções de consulta do usuário.

1.2 Motivação

A ampliação da qualidade do atendimento na área médica está diretamente relacionada a práticas de pesquisa e de otimização da gestão de recursos da instituição (XÉXEO; SANTOS, 2000; WANG et al., 2006). Estas atividades são

realizadas com base na análise de grandes volumes de dados e, assim, necessitam de infraestrutura computacional robusta e eficiente.

Assim como as práticas do mundo empresarial, aplicações de DWing também são indicadas para a área médica como uma ferramenta para aperfeiçoar o armazenamento dos dados e a geração de maneira ágil de informação e conhecimento voltados para o processo de tomada de decisão. Neste contexto, estudos apresentados por Murphy (MURPHY et al., 1999) avaliam o desempenho de aplicações de data warehousing para a área médica. Nesse trabalho, Murphy declara que consultas analíticas OLAP (*On-Line Analytical Processing*) realizadas por usuários da área médica foram executadas de 3 a 4 ordens de magnitude mais rápidas com aplicações de *data warehousing* do que em ambientes OLTP.

Um ambiente de *data warehousing* auxilia seus usuários a tomarem decisões a partir de investigações e análises dos dados. Por exemplo, pesquisas estatísticas sobre a eficiência de um tratamento pode ser realizada facilmente com o uso de um DWing; assim como tendências e hipóteses podem ser descobertas e validadas sobre uma grande quantidade de dados; políticas públicas, como a aquisição de novos medicamentos ou a contratação de novos profissionais da saúde, também podem ser investigadas com base em dados históricos armazenados no DW.

Entretanto, existe uma nova gama de consultas OLAP que atualmente não podem ser respondidas por aplicações de DWing convencionais. Uma equipe médica pode estar interessada em analisar a quantidade de imagens que são parecidas com uma determinada imagem de câncer de mama, para avaliar, por exemplo, a evolução da prevalência de formas patológicas segundo critérios de interesse, como exemplo a distribuição geográfica. Esta análise não pode ser respondida por aplicações de DW convencionais, porque essa utiliza, como base, imagens médicas e necessita realizar operações baseadas em similaridade, para as quais um DW convencional não oferece suporte.

A utilização de um *data warehouse* de imagens médicas também é justificada na aplicação de políticas públicas com relação ao gerenciamento de recursos da instituição hospitalar. Por exemplo, a partir de análises comparativas baseadas em imagens (e.g., consulta *"Qual a quantidade de imagens que são similares a uma determinada imagem de câncer de mama para os últimos três anos?"*), os administradores possuem mais subsídios para determinar se será necessário

contratar mais oncologistas ou mesmo avaliar a compra de novos materiais e medicamentos utilizados no tratamento de câncer de mama.

Além disso, aproveitando o recurso de análise de imagens (i.e., a imagem sendo um eixo de análise), o gestor pode tomar decisões especificando o nível de gravidade de uma patologia. Por exemplo, o gestor pode analisar a incidência de uma patologia com alto nível de gravidade para avaliar a necessidade de aumentar o estoque de um determinado medicamento ou mesmo aumentar o valor de cobrança dos tratamentos devido à complexidade da patologia e dos gastos dos procedimentos para tratá-la.

Em especial, este projeto de pesquisa tem o objetivo de ampliar a capacidade de processamento e armazenamento de ambientes de DWing, ao viabilizar que consultas OLAP baseadas em imagens médicas sejam realizadas em um DWing. Consequentemente espera-se ampliar o escopo das atividades de tomada de decisão a partir do suporte a imagens médicas.

O desafio em tornar esta nova gama de consultas executáveis por aplicações de DWing está relacionado à maneira segundo a qual às imagens deverão ser representados no cubo de dados (e.g., representados por descritores baseados em conteúdo) e como operações OLAP serão processadas no cubo de dados de imagens.

1.3 Objetivo

Este trabalho tem por objetivo ampliar as possibilidades de processamento e de armazenamento de dados em aplicações de *data warehousing* ao viabilizar que consultas OLAP baseadas em características intrínsecas de imagens sejam realizadas em um DWing. Consequentemente, o iCube visa ampliar a gama de atividades de tomada de decisão permitindo, para isso, o uso de imagens médicas.

São três os grandes desafios para permitir que esta nova gama de consultas seja respondida por aplicações de DWing. O primeiro destes desafios refere-se ao desenvolvimento e implementação de ferramentas ETL (*Extract, Transform and Load data tools*) para integrar as imagens, representadas por seu conteúdo intrínseco a dados convencionais (i.e., dados numéricos, datas e textos curtos) em um DW. O

segundo desafio refere-se ao DW de modo a permitir o armazenamento destes dados. Por fim, o terceiro desafio consiste em como consultas OLAP baseadas na similaridade de imagens podem ser processadas em ambiente DWing com baixo tempo de resposta.

Neste trabalho, abordaremos estes três desafios através da proposta de iCube, um ambiente de *data warehousing* estendido que engloba:

- Uma abordagem de transformação ETL adaptada para o tratamento de imagens (Seção 5.2);
- Um esquema estrela que possui uma tabela dedicada a representação e referenciamento de imagens (Seção 5.3);
- Uma abordagem ágil de processamento de consultas OLAP baseada em similaridade de imagens. Esta abordagem possui mecanismos de filtragem com o propósito de atender eficientemente às consultas por similaridade dos usuários de sistemas de suporte à decisão (Seção 5.4).

Este trabalho também estendeu a proposta de esquema estrela do iCube (Seção 5.2) de modo a investigar o esquema ideal de um IDW composto por múltiplas descrições do conteúdo intrínseco da imagem. Este estudo foi necessário, pois um dos maiores desafios de sistemas de recuperação de imagens baseada em seu conteúdo intrínseco (CBIR) é a redução do *gap* semântico. Como apresentado em (PONCIANO-SILVA et al., 2009), o *gap* semântico pode ser reduzido com o uso de diferentes mecanismos (i.e., descritores) para descrever o conteúdo intrínseco das imagens (Seção 2.2). Neste contexto, este trabalho propõe o iStar, um esquema estrela otimizado para *image data warehouses* baseados em similaridade (Capítulo 6 -). O iStar é um esquema estrela pertencente ao ambiente iCube, sendo caracterizado por possuir um conjunto de tabelas de dimensão especialmente projetadas para representar as percepções dos usuários sobre a similaridade entre as imagens, e, portanto, provê suporte ao processamento de consultas OLAP baseadas em diversos critérios de similaridade entre imagens.

1.4 Organização do Trabalho

O conteúdo desta monografia está estruturado da seguinte maneira:

- O Capítulo 2 descreve a fundamentação teórica usada no desenvolvimento desse trabalho;
- O Capítulo 3 descreve os trabalhos correlatos a este projeto;
- O Capítulo 4 apresenta a adaptação do método de acesso métrico Onion-Tree para o seu uso em ambientes de IDWing;
- O Capítulo 5 detalha o trabalho desta pesquisa em nível de mestrado, descrevendo o ambiente proposto iCube;
- O Capítulo 6 apresenta propostas de esquema estrela para IDWing e propõe o iStar como o esquema estrela com múltiplos descritores adaptado para *image data warehouses* baseado em similaridade;
- O Capítulo 7 sumariza as principais contribuições e conclusões obtidas a partir do desenvolvimento das propostas do iCube e do iStar e indica trabalhos futuros;
- Por fim, são apresentadas as referências bibliográficas e os apêndices.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

As propostas e discussões apresentadas neste trabalho foram fundamentadas pelos conceitos apresentados neste capítulo, que abrange conceitos sobre a atual tecnologia de data warehousing, consultas por similaridade, image data warehousing e estruturas de indexação para ambientes de IDWing. Analisando tal contextualização, foi possível identificar e discutir os desafios e algumas das principais questões ainda em aberto para novas pesquisas.

2.1 Tecnologia de Data Warehousing

A obtenção de informações estratégicas é uma etapa essencial no processo de tomada de decisão. Inicialmente, as informações estratégicas eram extraídas diretamente por aplicações OLTP (*On-line Transaction Processing*), as quais automatizam operações cotidianas de processamento de dados. Tradicionalmente, ambientes OLTP alimentam e acessam banco de dados operacionais, realizando operações caracterizadas por serem simples, repetitivas e volumosas (CHAUDHURI; DAYAL, 1997; KIMBALL; ROSS, 2002; INMON, 2005).

Mesmo com o uso de "*extratores de informação*", a obtenção de informações estratégicas por aplicações OLTP é apontada como pouco aconselhada. Os principais motivos apresentados por Inmon (INMON, 2005) são: Desempenho inaceitável no processamento de consultas complexas; Inconsistência e baixa quantidade de informações obtida por relatórios e análises de tendências, uma vez que os bancos de dados operacionais armazenam apenas dados atuais e não o histórico dos dados; Comprometimento no desempenho do processamento de

operações transacionais; Enorme esforço e tempo gasto para o tratamento de dados de fontes heterogêneas; e Diminuição da produtividade no processo de análise.

Neste contexto, foram propostos e implantados ambientes específicos para sistemas de suporte à decisão estratégica, que são denominados ambientes informacionais e são caracterizados pelo processamento de consultas OLAP (*On-Line Analytical Processing*). Os requisitos funcionais e de desempenho de um ambiente informacional, como o ambiente de *data warehousing*, são fundamentalmente diferentes de um ambiente operacional. Ademais, as diferenças entre os tipos de dados, tecnologias, usuários e processamento de ambos os ambientes são justificativas para manter um ambiente informacional separado de bancos de dados operacionais. Na Tabela 2.1, são apresentadas as principais diferenças existentes entre os ambientes operacional e informacional.

Ambientes de *data warehousing* (DWing) surgiram em meados dos anos 90 como uma solução para auxiliar, de maneira rápida e consistente, às organizações a enfrentarem desafios cada vez mais complexos em ambientes globais (SONG, 2009). Um ambiente de DWing consiste em coleções de tecnologias que possibilitam que dados originários de provedores de informação autônomos, heterogêneos e distribuídos sejam integrados em um único repositório conhecido como *data warehouse* (DW), e mantidos por um longo período de tempo (CHAUDHURI; DAYAL, 1997; KIMBALL; ROSS, 2002; INMON, 2005). No contexto empresarial, usuários de SSD utilizam ambientes de *data warehousing* para investigar o rumo de sua empresa, analisando milhares de dados agrupados e levantando diversas questões sobre esses dados. Exemplos dessas investigações são: avaliação de produtos, processos e profissionais; análises para obter novos clientes; assim como manter sua clientela. Enfim, as investigações realizadas têm como objetivo atingir ou manter a empresa em patamares competitivos no mercado e ampliar seus lucros.

No domínio médico-hospitalar, ambientes de DWing são implantados, principalmente, para garantir uma boa qualidade clínica, agilizar as pesquisas médicas e permitir que análises visando a redução de custos tanto para o paciente quanto para a instituição sejam realizadas em menor tempo e de forma organizada (PEDERSEN; JENSEN, 1998; BANEK; TJOA; STOLBA, 2006). Ademais, um DWing médico é a base para a descoberta de novas relações de causa e efeito das doenças por meio de mineração de dados, assim como na identificação de pacientes específicos qualificados para pesquisas ou tratamentos (MURPHY, 2009).

Tabela 2.1: Ambiente operacional versus ambiente informacional (Adaptado de: (CIFERRI, 2002; SONG, 2009)).

Aspecto	**Ambiente Operacional**	**Ambiente Informacional**
Ambiente	Voltado ao processamento de transações OLTP	Voltado ao processamento de transações OLAP
Operação mais frequente	Atualização, remoção e inserção	Leitura
Volume de transações	Relativamente alto	Relativamente baixo
Características das transações	Pequenas e simples. Acessam poucos registros por vez	Longas e complexas. Acessam muitos registros por vez e realizam várias varreduras e junções de tabelas
Tipo de usuários	Administradores do sistema, projetistas do sistema e usuários de entrada dos dados	Usuários de SSD (e.g. executivos, gestores e analistas)
Número de usuários	Grande (geralmente milhares)	Relativamente pequeno (geralmente centenas)
Predominantes interações com os usuários	Pré-determinadas e estáticas	Ad-hoc e dinâmicas
Tipo de dados	Dados primitivos (essencialmente atuais)	Dados derivados (geralmente dados históricos)
Volume de dados	Gigabytes e terabytes	Terabytes e pentabytes
Projeto de banco de dados	Normalizado para suporte às propriedades de atomicidade, consistência, isolação e durabilidade	Multidimensional, refletindo as necessidades de análise dos usuários de SSD
Granularidade dos dados	Detalhado	Agregado e detalhado
Principal questão de desempenho	Produtividade da transação	Produtividade da consulta
Tempo de resposta	Geralmente poucos segundos	De minutos a horas
Exemplo de aplicação	Transações bancárias, contas a pagar e empréstimos de livros	Planejamento de marketing, análise financeira

2.1.1 Data Warehouse Convencional

Como já especificado na Seção 2.1, em ambientes de *data warehousing*, dados de fontes distribuídas e heterogêneas são integrados em um banco de dados especializado denominado *data warehouse* (DW). Inmon define um DW como sendo uma “coleção de dados orientados a assunto, integrados, não voláteis e históricos, utilizado em processos de tomada de decisão” (INMON, 2005).

Dados armazenados em DW são caracterizados como orientados a assunto por serem relacionados a um tema de negócio de interesse da organização. Esta propriedade permite que seus usuários realizem análises multidimensionais sobre esse assunto para tomar decisões estratégicas.

Os dados em um *data warehouse* são compreendidos como integrados, pois mesmo sendo dados derivados de diversas fontes, esses são adaptados de forma a compor um único repositório. Vale ressaltar que um DW além de integrar dados de diferentes sistemas de banco de dados operacionais, também integra dados de outras fontes externas, tais com arquivos textos e planilhas. Para tanto, diversas fontes de dados não integradas e espalhadas são pesquisadas em busca de dados relativos ao tema de interesse da organização, e caso haja dados de interesse, esses são extraídos e integrados no DW. Todavia, antes de realizar o armazenamento desses dados, ocorre a eliminação de inconsistências dos dados devido à heterogeneidade das fontes (e.g. problemas de forma de representação, em uma fonte de dados o sexo é representado por f e m, já em outra é representado por 0 e 1). Esses procedimentos tornam o DW um repositório centralizado com todos os dados de negócio contendo a mesma semântica e formato.

A propriedade não volátil refere-se ao fato dos dados em um DW serem atualizados com pouca frequência, consequentemente, seus dados permanecem estáveis por longos períodos de tempo.

Por fim, os dados de um DW são ditos históricos, pois, diferentemente de dados de ambientes operacionais que sofrem diversas alterações em seu valor, estes possuem o mesmo valor em um longo período do tempo. Assim que os dados são armazenados no DW, esses não são removidos e toda alteração é armazenada como uma nova entrada associada a um componente de tempo, como snapshots, resultando em uma base histórica.

2.1.2 Arquitetura de Data Warehousing

Ambientes de *data warehousing* são compostos por diversas tecnologias (CHAUDHURI; DAYAL, 1997), as quais são tipicamente organizadas em cinco camadas: primeira camada consiste nos sistemas de fonte de dados; segunda camada refere-se aos sistema de gerenciamento de ETL (*Extract, Transform and Load*); terceira camada é composta por DW, *data mart* (DM) e repositório de

metadados; quarta camada consiste nos servidores OLAP; e quinta camada é composta pelas ferramentas de consulta e análise dos usuários finais. Esta arquitetura é ilustrada na Figura 2.1.

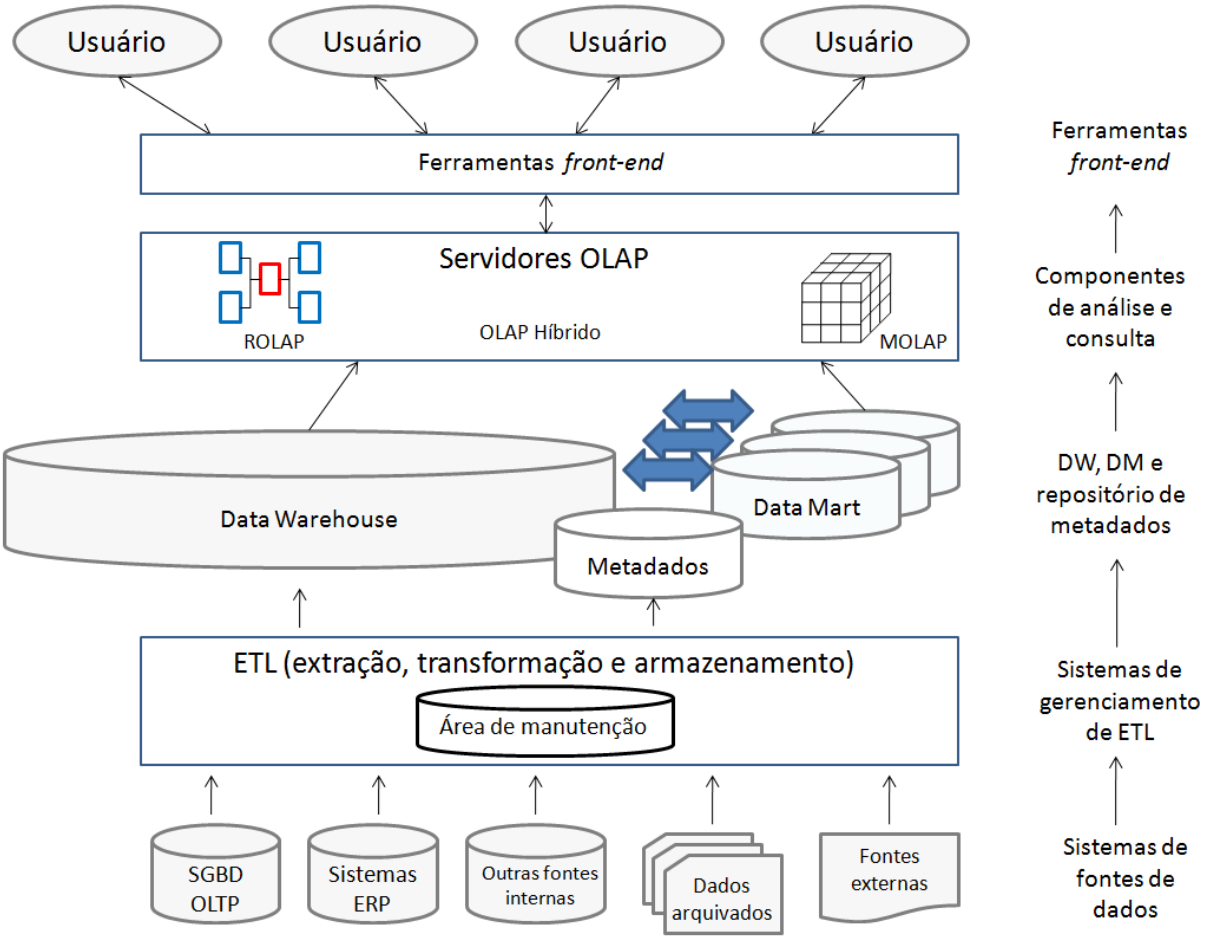

Figura 2.1: Arquitetura típica de um ambiente de data warehousing (adaptado de (SONG, 2009)).

A camada de sistemas de fontes de dados tem a função de referenciar todas as origens de dados armazenados no DW. Como descrito na seção anterior, um *data warehouse* integra dados importantes na tomada de decisão estratégica, oriundos de fontes heterogêneas. Estas fontes de dados são geralmente bases de dados operacionais (OLTP), mas outras fontes internas, tais como, banco de dados legados, planilhas e arquivos, e fontes externas também podem ser acessadas. Como o foco desta integração é compor um DW rico de dados referentes ao assunto de interesse do usuário, não há restrições quanto ao formato dos dados. Dessa maneira, vários tipos de dados podem ser extraídos, desde dados estruturados até dados semi-estruturados e não estruturados.

A segunda camada de uma arquitetura de data warehousing é tipicamente composta pela ferramenta ETL. Esta ferramenta trata a diversidade de formatos e estrutura dos dados, garantindo assim que o DW ofereça uma “versão única da verdade”. Para tanto desempenha três processos fundamentais: extração, transformação e armazenamento.

Durante o processo de extração, a ferramenta ETL é utilizada para indicar quais são as fontes de dados que contêm dados de interesse e para identificar quais são esses dados. Uma vez identificados os dados, esses são copiados e tratados no processo de transformação. A extração de dados tem um papel fundamental em um ambiente de *data warehousing*, pois o torna independente das fontes de dados, além de garantir que alterações indesejadas ocorram nas bases de dados. O intervalo de acesso às fontes pode ser mensal, semanal, diária ou até mesmo em tempo real, estando diretamente relacionado às atualizações da fonte e, principalmente, relacionado à necessidade do usuário.

O segundo processo realizado pela ferramenta de ETL é o processo de transformação, que elimina as inconsistências e a diversidade de formatos dos dados devido à heterogeneidade das fontes. Os exemplos de erros e inconsistências tratadas nesta etapa são: inconsistência no comprimento de campo, dados incompletos ou em branco, duplicações, descrições inconsistentes, associações de valores inconsistentes, abreviações não padronizadas, valores inapropriados (e.g. campo idade é preenchido com o valor “670” ou “1a2”) e violações de restrição de integridade. Os dados, que durante essa etapa não puderam ser transformados, não são armazenados no repositório.

O processo de transformação também é responsável pela uniformidade dos dados. Por exemplo, um dos dados de interesse do DW é o sexo dos clientes e esses dados são oriundos de duas fontes de dados, em uma fonte de dados o sexo é representado por f ou m, já em outra é representado por 0 ou 1. Neste caso, a ferramenta ETL irá copiar esses dados, mas antes de armazená-los esses serão transformados em um único formato. Dessa maneira, o resultado desta etapa consiste na padronização dos dados do DW em relação ao tipo de dado, formato, tamanho, unidade, codificação de valores e semântica.

Terminada a transformação dos dados (i.e., transformação e limpeza), inicia o processo de armazenamento. O armazenamento dos dados, geralmente, é realizado após a execução de vários processamentos adicionais (CIFERRI et al., 2007), tais

como, verificação de restrição de integridade, ordenação dos dados, geração de agregações, além de construção de índices.

Medidas de segurança referentes à falha no processo de armazenamento devem ser tomadas. A recuperação de falhas evita gasto computacional desnecessário, pois o carregamento é reiniciado a partir dos dados que ainda não foram corretamente calculados ou armazenados antes de uma eventual falha.

O processo ETL é considerado por vários autores como a fase de maior consumo de tempo no desenvolvimento de ambientes de DWing, consumindo de 60% a 80% do esforço de todo o desenvolvimento (INMON, 2005). Portanto, é altamente recomendável a automatização desta etapa utilizando ferramentas ETL.

A terceira camada de uma arquitetura típica de *data warehousing* é composta pelo DW, pelos *data marts* e pelo repositório de metadados. Todos os dados enviados pela ferramenta ETL são mantidos no *data warehouse* em seu nível atômico e primitivo. Como o volume de dados armazenados em um DW é muito grande, de terabytes a pentabytes, torna-se necessário a elaboração de rotinas de remoção de dados. Usualmente, são definidos limites de tempo em relação a retenção dos dados e após esse período, os dados são excluídos ou arquivados em outro repositório, como exemplo em memória terciária (e.g. *juke box* com fitas magnéticas ou DVD). Os dados de um DW podem ser agregados de forma a materializar respostas para consultas OLAP. Desta forma, um DW pode ser composto de vários níveis de agregação, construídos a partir do nível anterior, os quais formam o conceito de cubo de dados.

Um *data mart* é visto por (SONG, 2009) como um DW de pequeno porte ou departamental, pois os dados de ambos os repositórios compartilham as mesmas características, isto é, os dados são orientados a assunto, integrados, não voláteis e históricos, ademais seus dados são organizados em diferentes níveis de agregação. Um *data mart* é caracterizado como de pequeno porte, pois seu volume é limitado aos dados de interesse a um departamento, ao invés de atender às necessidades de toda empresa. Outra característica de um *data mart* refere-se ao nível de agregação, de forma a agregar os dados a um nível consistente com as necessidades de seus usuários. Armazenar dados agregados, mesmo que seja em um nível de pequena granularidade, reduz o tempo de resposta no processamento de consultas OLAP, simplifica o entendimento de seu projeto e a sua manutenção. A criação de um *data*

mart também resulta em uma descentralização de acesso aos dados do *data warehouse*.

Existem duas abordagens de representação lógica de um DW: MOLAP (OLAP multidimensional) e ROLAP (OLAP relacional). Abordagens MOLAP organizam e armazenam os dados em estruturas especiais, isto é, matrizes multidimensionais. Esta abordagem é caracterizada por operações OLAP serem executadas diretamente sobre sua estrutura. Já as abordagens ROLAP organizam e armazenam os dados de um DW em SGBD relacional, o qual é estendido para realizar as consultas de tomada de decisão. As estruturas ROLAP mais conhecidas são esquema estrela e esquema floco de neve. Esta dissertação de mestrado enfocará estruturas ROLAP, mais especificamente a esquemas estrela.

O esquema estrela é o modelo de dados lógico mais utilizado, composto por apenas dois tipos de tabela: uma tabela de fatos, geralmente posicionada no centro do esquema, e várias dimensões ligadas a essa tabela central (SONG, 2009). Uma tabela de fatos armazena dados conhecidos como medidas, sendo identificada por uma chave primária composta por chaves estrangeiras para todas as dimensões do esquema, mantendo assim um relacionamento com cada uma das dimensões. Uma dimensão armazena atributos que servem como um eixo de análise dos dados, que podem estar organizados em uma hierarquia de atributos. A Figura 2.2 exibe um esquema estrela referente à incidência de câncer de mama.

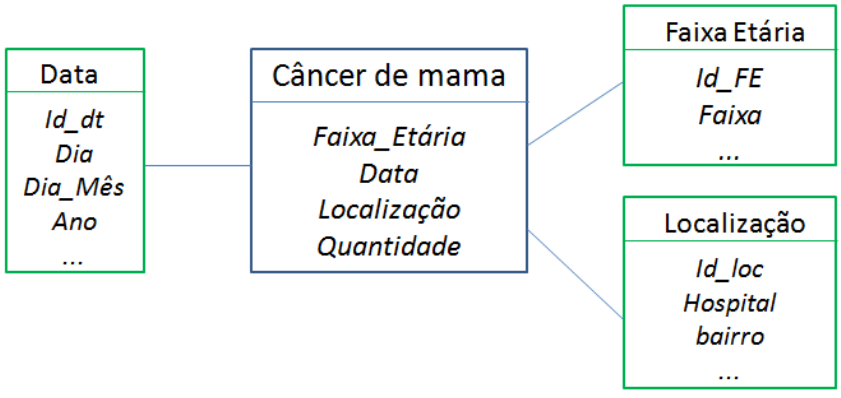

Figura 2.2: Esquema estrela para a incidência de câncer de mama.

Diferentemente do esquema estrela, em um esquema floco de neve, as hierarquias de atributos são representadas explicitamente, por meio da normalização das tabelas de dimensões. Geralmente, o desempenho no processamento de

consultas OLAP em um DW convencional representado pelo esquema floco de neve é inferior por exigir a junção de várias tabelas, tornando o esquema estrela preferível ao esquema floco-de-neve (KIMBALL; ROSS, 2002).

Outro componente importante da terceira camada de um ambiente de DWing consiste no repositório de metadados. Este tem por objetivo armazenar dados e informações referentes às estruturas, operações e conteúdo do DW e *data marts*, permitindo que usuários administradores organizem, compreendam e gerenciem o ambiente de DWing. Existem três classes de metadados (SONG, 2009): (i) Metadados de Negócio, relacionado ao cumprimento das regras de negócio da organização; (ii) Metadados Técnicos, referente aos elementos do *data warehouse* (e.g. tabelas, índices, tipos de dados) e aos elementos de fontes de dados (e.g. informações sobre o processo de ETL, frequência de extração); (iii) Metadados de Processo, relativos aos eventos ocorridos em qualquer processo (e.g. estatísticas sobre o tempo de execução de consultas, segundos de CPU, acessos a disco).

Servidores OLAP representam a quarta camada de uma arquitetura de *data warehousing*. Análises e consultas sobre os dados de um DW ou de um *data mart*, geralmente, são executadas por aplicações OLAP. Na Seção 2.1.3 são detalhados os conceitos referentes à estrutura multidimensional. Dentre os servidores OLAP atualmente disponíveis, pode-se destacar o Mondrian (http://mondrian.pentaho.com/) e Microsoft SQL (http://www.microsoft.com/sqlserver/en/us/default.aspx).

A quinta camada de uma arquitetura de *data warehousing* é composta por ferramentas de interface para os usuários finais (i.e., usuários de SSD), utilizadas na exploração do conteúdo do DW e *data marts* por meio de relatórios, consultas ad-hoc, análises OLAP, mineração de dados, painéis e outros aplicativos de BI (*business intelligence)*.

2.1.3 Consultas OLAP

Aplicações OLAP permitem uma visão multidimensional de medidas sobre um conjunto de dimensões (KIMBALL; ROSS, 2002; INMON, 2005; RIZZI, 2006), em que as medidas são os objetos de análise relevantes ao negócio, enquanto as dimensões determinam o contexto e as perspectivas para essas medidas.

Uma medida pode ser definida em função de suas dimensões correspondentes, representando, desta forma, um valor no espaço multidimensional.

Por exemplo, em uma aplicação de DW para a área de saúde, a medida numérica "incidência de câncer de mama" pode ser determinada em função das dimensões hospital (H), Idade (F) e data (D). O cubo de dados para esta aplicação é ilustrado na Figura 2.3, na qual cada eixo denota estas dimensões. Neste cubo, os valores em cada célula do cubo de dados representam a medida numérica de interesse denominada "incidência de câncer de mama".

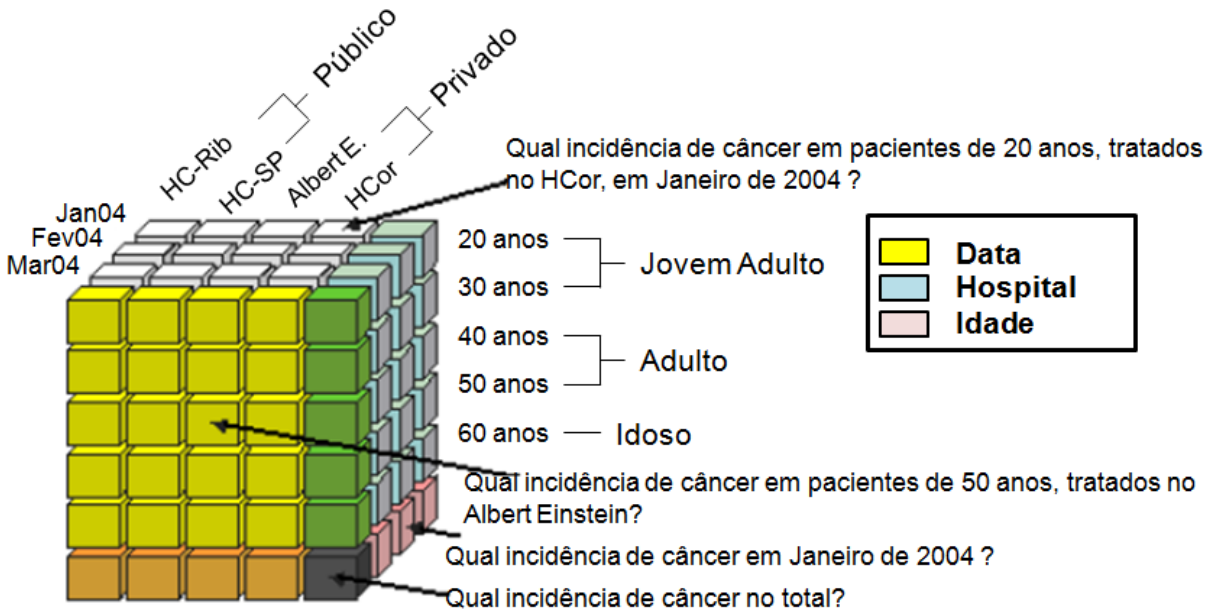

Figura 2.3: Cubo sobre a incidência de câncer de mama pelas perspectivas hospital, data e idade (adaptado de (SIQUEIRA, 2009)).

Além de serem modelados multidimensionalmente, os dados em um *data warehouse* também são organizados em diferentes níveis de agregação, desde um nível inferior que possui dados detalhados, até um nível superior que possui dados muito resumidos. Também podem existir vários níveis intermediários entre estes dois níveis que representam graus de agregação crescentes. Neste sentido, a semântica subjacente ao cubo de dados multidimensional permite não somente a visualização dos valores de coordenadas específicas, mas também a identificação das várias agregações que podem ser originadas ao longo de todas as dimensões.

No exemplo corrente, o cubo de dados multidimensional permite que a medida numérica "incidência de câncer de mama" seja examinada segundo diferentes perspectivas, tais como incidência de câncer de mama por faixa etária, incidência de câncer de mama por hospital, incidência de câncer de mama por dia por hospital e incidência de câncer de mama por faixa etária por hospital. Em

especial, a medida numérica “incidência de câncer de mama” é aditiva, uma vez que por meio da combinação de suas dimensões ela pode ser aritmeticamente somada. Assim, pode-se determinar a incidência de câncer de mama por mês por faixa etária por hospital somando-se os valores das incidências de câncer de mama para cada um dos dias que formam o mês. Tipicamente, outras medidas estatísticas, como média e variância, podem ser obtidas e, em cada caso específico, as diferentes projeções são calculadas por algoritmos específicos (NEUMUTH et al., 2008).

A forma de organização dos dados do DW em diferentes níveis de agregação garante uma melhoria no desempenho do processamento de consultas OLAP (BECKERA; RUIZ, 2004; FIDALGO et al., 2004; FORLANI; CIFERRI; CIFERRI, 2006; CIFERRI et al., 2007). Operações OLAP típicas incluem consultas *drill-down*, *roll-up*, *slice-and-dice*, *pivot* e *drill-across* (CHAUDHURI; DAYAL, 1997; KIMBALL; ROSS, 2002; INMON, 2005; RIZZI, 2006). Enquanto consultas *drill-down* analisam os dados em níveis de agregação progressivamente mais detalhados, consultas *roll-up* inversamente investigam os dados em níveis de agregação progressivamente menos detalhados. Já a operação *slice-and-dice* permite que os usuários de SSD restrinjam os dados sendo analisados a um subconjunto destes dados. Diferentes perspectivas dos mesmos dados podem ser obtidas pela operação *pivot*, a qual reorienta a visão multidimensional dos dados. Esta operação altera a ordem das dimensões, modificando, consequentemente, a orientação sob a qual os dados são visualizados. Por fim, consultas *drill-across* são consultas que comparam medidas numéricas de cubos de dados diferentes que são relacionados entre si por uma ou mais dimensões em comum.

2.2 Consultas por Similaridade

Uma imagem digital é determinada por uma matriz bidimensional M(x, y), em que cada coordenada (x, y) corresponde a um pixel da imagem. Mesmo para imagens com dimensões reduzidas, uma comparação pixel-a-pixel entre imagens é um processo impraticável devido ao seu alto custo computacional. Atualmente, as análises computacionais para determinar a similaridade entre duas imagens são

realizadas a partir da comparação sobre o conteúdo intrínseco dessas imagens (FELIPE, 2005).

Algoritmos de processamento de imagem, denominados descritores, são ferramentas que representam o conteúdo intrínseco de uma imagem através de assinaturas matemáticas, conhecidas como vetores de características (GONZALEZ; WOODS, 2008). A similaridade entre duas imagens é, normalmente, determinada utilizando uma função de distância, para mensurar a similaridade, e um algoritmo de consulta para estabelecer o critério na consulta por similaridade. Dessa maneira, uma função de distância é aplicada sobre os vetores de características dessas duas imagens a fim de obter o valor de distância (i.e., dissimilaridade) entre esses dois vetores. Esse valor de distância mensura a similaridade entre duas imagens em uma relação inversamente proporcional, ou seja, quanto menor o valor de distância entre os vetores de características, maior é a similaridade dessas imagens.

Usualmente, consultas por similaridade utilizam os algoritmos de consulta por abrangência e dos k-vizinhos mais próximos (i.e., kNN) (BÖHM; BERCHTOLD; KEIM, 2001; CHÁVEZ et al., 2001; FELIPE, 2005; FELIPE; TRAINA; TRAINA, 2005).

O algoritmo de consultas por abrangência recupera os elementos definindo um limite de distância, em que fornecido um conjunto de elementos $S = \{s_1, s_2, \dots, s_n\}$ e uma função de distância $ds_q(s_i)$, onde $s_i \in S$, a função $ds_q(s_i)$ determina a distância do elemento s_i ao elemento consulta s_q; o conjunto resultante da consulta por abrangência é determinado por:

$$R(s_q, r) = \{s_i \mid ds_q(s_i) \leq r, \forall s_i \in S\} \qquad (2.1)$$

Dessa maneira, todos os elementos s_i que tenham sua distância ao elemento de consulta s_q menor ou igual a um limite *r* são considerados similares a s_q.

O algoritmo de consulta aos k-vizinhos mais próximos recupera todas as k imagens com menor distância em relação ao elemento de referência. Logo, fornecido um conjunto de elementos $S = \{s_1, s_2, \dots, s_n\}$ e uma função de distância $ds_q(s_i)$, onde $s_i \in S$, a função $ds_q(s_i)$ determina a distância do elemento s_i ao elemento consulta s_q; o conjunto *Q* resultante da consulta por k-vizinhos mais próximos é determinado por:

$$kNN(s_q,k) = \{s_i | s_i \in Q, |Q| = k, Q \subseteq S, \forall s_i \in Q, s_j \in S - Q, ds_q(s_i) \leq ds_q(s_j)\} \quad (2.2)$$

Na Figura 2.4, a seleção pelos algoritmos de consulta por abrangência e por k-vizinhos mais próximos é ilustrada em um plano de duas dimensões. Ambos os algoritmos são centrados no mesmo elemento de consulta, representado pelo circulo preto, no entanto, o critério de similaridade do algoritmo de consulta por abrangência refere-se ao raio de abrangência *r,* enquanto para o kNN refere-se ao valor de k.

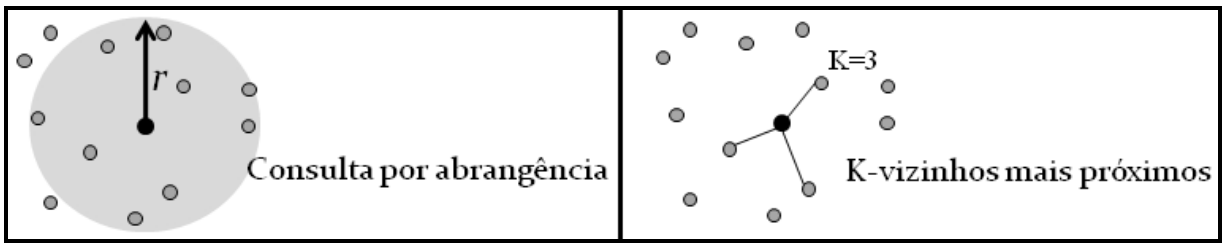

Figura 2.4: Exemplo de consultas por abrangência e por k-vizinhos mais próximos.

Um dos maiores desafios em consultas por similaridade é a redução do *gap* semântico, em outras palavras, este desafio está relacionado com a diferença entre o resultado gerado pelo consulta computacional por similaridade e o resultado esperado pelo usuário. O *gap* semântico ocorre, pois diferentes usuários, ou o mesmo usuário sob diferentes intenções de consulta, podem escolher diferentes condições para determinar a similaridade entre um conjunto de imagens. Portanto, a precisão de uma consulta é uma medida subjetiva à percepção do usuário. Nesse contexto, um Sistema de Recuperação de Informação Baseado em Conteúdo (CBIR) deve ser capaz de executar consultas com diferentes percepções de usuários ao aplicar extratores bem adequados à representação dessas percepções e ao aplicar funções de distância que melhor discrimine a similaridade entre as imagens (PONCIANO-SILVA et al., 2009).

2.2.1 Modelo de Espaço Métrico e Espaço Vetorial

Na literatura, existem duas abordagens típicas de modelagem dos vetores de características: o modelo de espaço vetorial e o modelo de espaço métrico. O modelo de espaço vetorial descreve o conteúdo intrínseco da imagem em um vetor

composto por *n* atributos (i.e., n = número fixo de características). Dessa maneira, cada vetor de características pode ser visto como um ponto em um espaço *n*-dimensional. Em um modelo de espaço vetorial, a similaridade entre duas imagens pode ser calculada com o uso de funções de distância, como as funções de distância de Minkowski: Euclidiana (L_2), Manhatan (L_1) ou Infinity(L_∞) (TRAINA-JR. et al., 2007; GONZALEZ; WOODS, 2008).

Em contrapartida, o modelo de espaço métrico não representa o conteúdo intrínseco em um conjunto fixo de características. Um exemplo é a representação de impressões digitais, as quais são caracterizadas pelo número de arcos, laços e espirais (TRAINA-JR. et al., 2007). Como cada digital é única e o número de características pode variar, não é possível representar este objeto como um ponto em um espaço *n*-dimensional. Mesmo não possuindo um número fixo de características, é possível calcular a similaridade entre dois objetos pertencentes a um espaço métrico M, onde $M = \langle \mathbb{S}, d(\,) \rangle$, $\mathbb{S}$ consiste no universo de elementos válidos e $d(\,)$ é a função que determina a distância (i.e., dissimilaridade) entre os elementos de $\mathbb{S}$ e que atende as propriedade de:

- Simetria: $d(s_i, s_j) = d(s_j, s_i)$
- Não negatividade: $0 < d(s_i, s_j) < \infty$, $se\ s_i \neq s_j$, e $d(s_i, s_i) = 0$
- Desigualdade triangular: $d(s_i, s_k) \leq d(s_i, s_j) + d(s_j, s_k)$, $onde\ s_i, s_j, s_k \in \mathbb{S}$

Nesta pesquisa em nível de mestrado, vamos utilizar o modelo de espaço métrico por este ser menos restritivo, pelo fato do conjunto $\mathbb{S}$ incluir elementos representados em espaço vetorial e pelo fato das funções de Minkowski atenderem as três propriedades e serem casos especiais de modelos de espaço métrico.

2.2.2 Funções de Distância

Existem muitas funções de distância que podem ser aplicadas para medir a similaridade entre dois objetos. Exemplos de funções de distância para os domínios de objetos complexos em espaços métricos incluem a função de distância LEdit (LEVENSHTEIN, 1966) e a função de distância MH() (TRAINA et al., 2002). Uma das funções mais amplamente utilizadas é a função de distância Minkowski, a qual

pode também ser aplicada a domínios vetoriais como casos especiais de modelos de espaço métrico, sendo baseada nas normas Lp (FELIPE, 2005). Considerando-se um conjunto de imagens, cada qual possuindo um vetor de características, que consiste em um vetor de valores de suas n características, e tomando duas imagens, q e c, representadas pelos vetores de características q = (q1,q2,...,qn) e c = (c1,c2,...,cn), a distância Minkowski é calculada por:

$$d(Q,C) = \sqrt[p]{\sum_{i=1}^{n}|c_i - q_i|^p} \qquad (2.3)$$

As variações de Minkowski usualmente utilizadas são: L1 (i.e., Manhattan), L2 (i.e., Euclidiana) e L∞ (i.e., Chebychev). A distância L1 de um espaço geométrico determinado por duas coordenadas (e.g. x e y) é definida por todos os pontos que possuem o valor da soma de d(Q,C) = |cy – qy| + |cx – qx|. O valor dessa distância é compreendido como o deslocamento de um ponto a outro nas ruas de uma cidade por meios de segmentos lineares paralelos ou perpendiculares entre si (Figura 2.5).

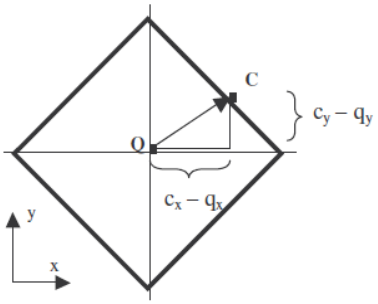

Figura 2.5: Lugar geométrico definido pela distância L1 (FELIPE, 2005).

A distância L2 de um espaço geométrico determinado por duas coordenadas é definida pela Equação 2.4. O valor dessa distância é compreendido como o deslocamento de um ponto em linha reta (Figura 2.6).

$$d(Q,C) = \sqrt[2]{(c_y - q_y)^2 + (c_x - q_x)^2} \qquad (2.4)$$

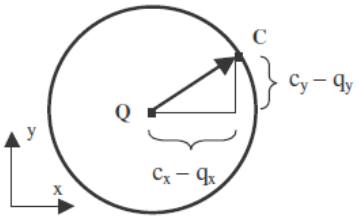

Figura 2.6: Lugar geométrico definido pela distância L2 (FELIPE, 2005).

A distância L∞ de um espaço geométrico determinado por duas coordenadas (Figura 2.7) é definida pelo valor no eixo em que a diferença entre as coordenadas dos pontos é a maior (Equação 2.5).

$$d(Q,C) = max_{i=1}^{n}|c_i - q_i| \qquad (2.5)$$

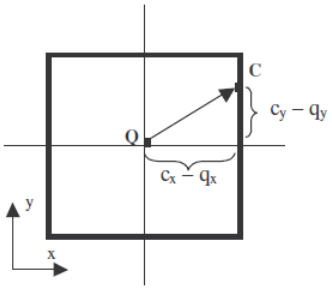

Figura 2.7: Lugar geométrico definido pela distância L∞ (FELIPE, 2005).

2.2.3 Descritores de Conteúdo Intrínseco

Os descritores para os quais as comunidades de processamento de imagens e visão computacional já desenvolveram modelos consolidados são para a representação de cor, textura e forma (DATTA et al., 2008; GONZALEZ; WOODS, 2008). A seguir são descritas as propriedades dos descritores para estes três tipos de representação da imagem (i.e., para estes três tipos de percepções).

Cor: Um exemplo de descritor de cor muito conhecido para representar a distribuição global de cores (ou de níveis de cinza) em uma imagem é o histograma de intensidade. Um histograma de intensidade de uma imagem quantifica a frequência com que um valor de luminância ocorre em pixels dessa imagem. O histograma de uma imagem com L níveis de cinza é representando em um vetor hf(L) com L elementos, também conhecido como bins, o qual hf(i) armazena a quantidade de pixels da imagem s_i em que o nível de cinza i ocorre. Usualmente,

histogramas são apresentados graficamente em que os níveis de cinza compõem a abscissa e suas frequências estabelece a escada nas ordenadas.

Uma importante característica de histograma de intensidade está relacionada ao fato de cada imagem gerar apenas um histograma, no entanto, um mesmo histograma pode ser gerado por diferentes imagens. Para tanto, os histogramas devem ter as mesmas frequências de intensidade de pixel, não importando como esses pixels estão distribuídos na imagem.

Textura: A textura é uma percepção visual intuitiva baseada no padrão de distribuição dos pixels em uma região de imagem. Uma abordagem amplamente utilizada para representar a textura de uma imagem é a abordagem estatística, que gera resultados satisfatórios com baixo custo computacional. Nessa abordagem, o conteúdo intrínseco da imagem é representado através de uma matriz de co-ocorrência P(m, n), em que P(i, j) refere à célula da linha i com a coluna j e essa célula armazena a quantidade de vezes (frequência) em que o pixel de intensidade i (e.g., intensidade de cinza) era vizinho do pixel de intensidade j em uma determinada direção e em uma distância especifica. As direções típicas são 0º, 45º, 90º e 135º, e as distâncias são escolhidas de acordo com a granularidade da imagem, geralmente variando de 1 a 5. Para cada combinação de direção e distância é construído uma matriz de co-ocorrência (FELIPE, 2005; GONZALEZ; WOODS, 2008).

Para descrever o conteúdo intrínseca da imagem em um vetor de características, Haralick em (HARALICK; SHANMUGAN; DINSTEIN, 1973) propôs um conjunto de 14 medidas obtidas através da matriz de co-ocorrência. Na Tabela 2.2, estão apresentadas as medidas de Haralick mais usadas na literatura.

Tabela 2.2: Equações das características de textura de Haralick.

Característica – Significado	Equação
Variância – nível de contraste	$\sum_i \sum_j (i-j)^2 P(i,j)$
Entropia – suavidade	$\sum_i \sum_j P(i,j)\ log\big(P(i,j)\big)$
Energia – uniformidade	$\sum_i \sum_j P^2(i,j)$

Homogeneidade	$\sum_i \sum_j \frac{P(i-j)}{(1+\lvert i-j\rvert)}$
3º momento – distorção	$\sum_i \sum_j (i-j)^3 P(i,j)$
Inversão da variância – nível inverso de contraste	$\sum_i \sum_j \frac{P(i,j)}{(i-j)^2}$

Forma: Em aplicações médicas, é usual realizar análises sobre a forma de um objeto da imagem. Por exemplo, a forma irregular de um nódulo em muitas vezes caracteriza um nódulo maligno, enquanto os nódulos de forma regular geralmente são considerados benignos. Os momentos de Zernike são usados para representar o conteúdo intrínseco das imagens através de descritores de forma. Os polinômios de Zernike são um conjunto de funções polinomiais complexas que formam uma base ortogonal dentro de um círculo de raio unitário $x^2 + y^2 \leq 1$. A representação polar dos momentos de Zernike de ordem n e repetição 1 é definida pela seguinte equação:

$$A_{nl} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^{\infty} [V_{nl}(r,\theta)] * f(r\cos\theta, r\sin\theta) r \, dr \, d\theta \qquad (2.6)$$

Para Equação 2.6, n-1 deve ser par, |1| ≤ n, *r* e θ são coordenadas polares dos pixies, $f(r\cos\theta, r\sin\theta)$ são valores de brilho do pixel representado por r e θ, V_{nl} é o modelo polinomial de Zernike, ilustrado na Equação 2.7 e $R_{nl}(r)$ é o polinômio ortogonal radial modelado na Equação 2.8.

$$V_{nl} = \sqrt{(R_{nl}\cos(l\theta))^2 + (R_{nl}\sin(l\theta))^2} \qquad (2.7)$$

$$R_{nl}(r) = \sum_{s=0}^{\frac{n-|l|}{2}} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|l|}{2}-s\right)!\left(\frac{n-|l|}{2}-s\right)!} r^{n-2s} \qquad (2.8)$$

2.3 Image Data Warehousing

Imagens são dados não estruturados, os quais podem ser representados por descritores baseados em conteúdo (ARIGON; MIQUEL; TCHOUNIKINE, 2007). Esses mecanismos confere às imagens uma representação estruturada e, consequentemente, permite que estes dados complexos sejam inseridos em um DW. Na literatura, as estruturas multidimensionais que descrevem imagens a partir de descritores em dimensões ou em medidas são denominadas como *multimedia data warehouse*, *image data warehouse* (IDW) (ZAÏANE et al., 1998; WONG et al., 2004; YOU et al., 2004; ARIGON; MIQUEL; TCHOUNIKINE, 2007). Nesta pesquisa adotamos o termo *image data warehouse*.

Ambientes de *image data warehousing* (IDWing) também são compostos por tecnologias organizadas em cinco camadas: sistemas de fonte de dados; sistema de gerenciamento de ETL; DW, *data mart* e repositório de metadados; servidores OLAP; e ferramentas de consulta e de análise (ilustrado na Figura 2.8). No entanto, essas camadas são adaptadas de forma que este ambiente dê suporte ao uso de imagens, destacando principalmente a especialização nas camadas de ETL, de DW e na camada de servidores OLAP.

Em uma arquitetura típica de IDWing, a camada de ETL além de realizar as atividades convencionais de extração, transformação e armazenamento, também realiza o processamento das imagens a fim de extrair o conteúdo intrínseco dessas imagens e armazená-lo no IDW. Dependendo do descritor definido pelo usuário, essa camada também pode conter uma estação de preparo das imagens, em que as imagens são tratadas para eliminar ruídos e, caso necessário, são segmentadas para destacar o objeto a ser analisado.

Um ambiente de IDWing também é caracterizado pela adaptação da camada de DW ao contexto de imagem. Nessa camada, o DW é especialmente projetado de forma a manter sua propriedade de multidimensionalidade ao mesmo tempo em que as imagens sejam inseridas ou tenham seu conteúdo intrínseco representado na forma de dimensões ou medidas. Um DW com essas características exige atenção especial, pois ao representar as imagens na forma de vetor de características, a definição de medidas torna-se uma tarefa não trivial devido ao fato de vetores de características não apresentarem a propriedade de relação de ordem total.

Dessa maneira, a camada de servidores OLAP de um ambiente de IDWing deve ser capaz de processar consultas OLAP baseada em algum predicado relacionado às imagens, e podem conter funções de agregação específicas capazes de agrupar as imagens. Para o contexto de representação do conteúdo intrínseco da imagem, indica-se que as funções de agregação sejam realizadas baseadas em consultas por similaridade. Esta nova gama de consultas OLAP é denominada neste projeto como consultas IOLAP (*Image On-Line Analytical Processing*), que necessariamente possuem algum critério de similaridade entre as imagens e podem também ser composta por predicados convencionais.

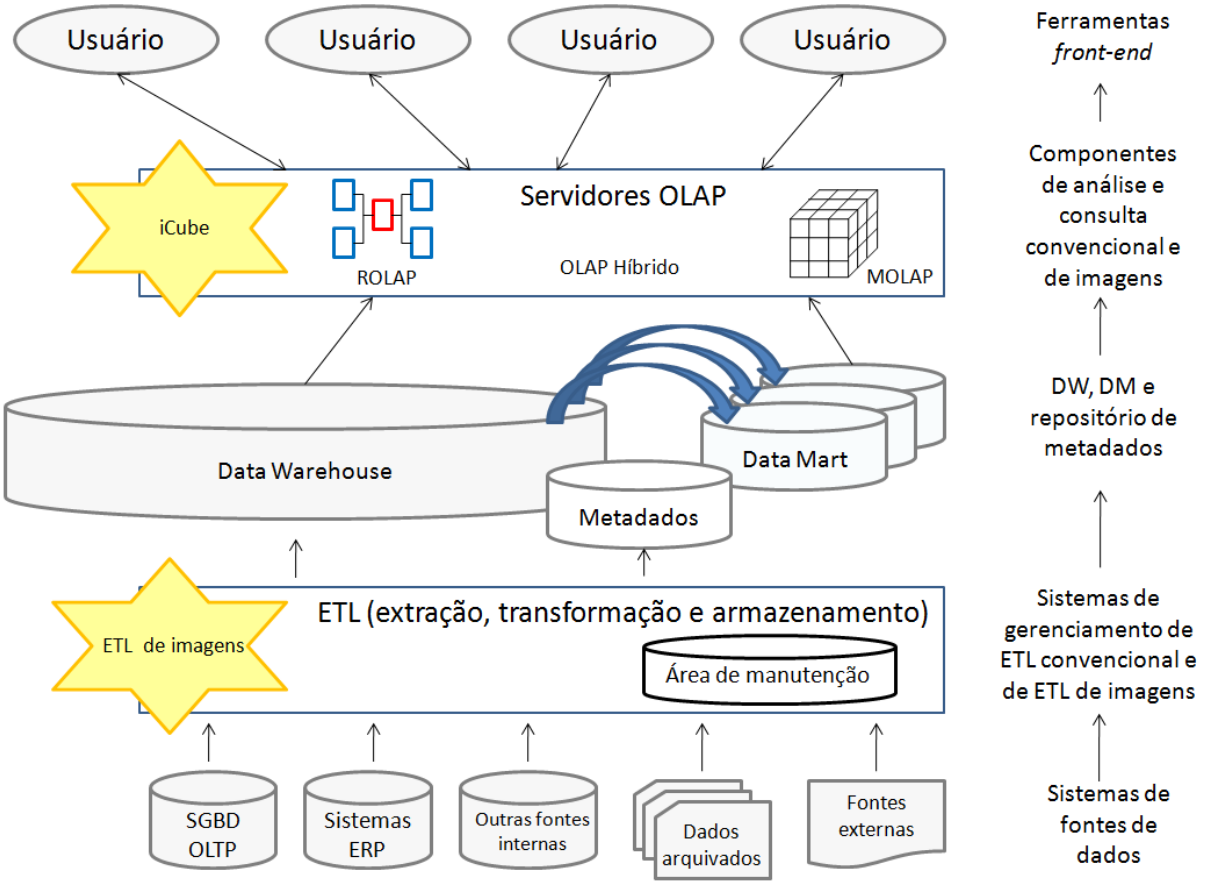

Figura 2.8: Arquitetura de um ambiente de Image Data Warehousing.

Nesta pesquisa, trabalhamos na extensão das camadas de ETL, camada de IDW e camada de servidores IOLAP. Os capítulos 5 e 6 detalham nossa proposta de extensão, assim como os resultados obtidos.

2.4 Estrutura de Indexação

Índices são estruturas de dados desenvolvidas para permitir o acesso ágil a registros em um grande banco de dados. Um índice organiza os dados segundo um ou vários predicados, sem de fato reordenar os registros, e permite encontrar um registro consultando apenas uma pequena fração dos registros do repositório (GARCIA-MOLINA; ULLMAN; WIDOW, 2000; ELMASRI; NAVATHE, 2005). Nesta pesquisa, foram investigados dois tipos de estruturas de indexação: métodos de acesso convencionais e métodos de acesso métricos.

As seções a seguir detalham o funcionamento dessas estruturas de indexação, assim como descrevem suas vantagens e desvantagens em ambientes IDWing.

2.4.1 Métodos de Acesso Métrico

O uso de índices em consultas por similaridade é importante, pois minimiza o número de cálculos de distância necessários para processar uma consulta e, por conseguinte, minimiza o seu tempo de processamento. Para consultas por similaridade em espaços métricos genéricos, os métodos de acesso métricos (MAM) são os mais adequados. MAM's são índices baseados em distância que utilizam exclusivamente funções de distância para organizar os objetos no índice. Exemplos de MAM's propostos na literatura incluem a VP-tree (*Vantage Point tree*) (YIANILOS, 1993), a MVP-tree (*Multi-Vantage Point tree*) (BOZKAYA; OZSOYOGLU, 1999), a GNAT (*Geometric Near-neighbor Access Tree*) (BRIN, 1995), a M-tree (CIACCIA; PATELLA, 2002), a Slim-tree, a DBM-tree (FILHO et al., 2001; TRAINA-JR et al., 2002; VIEIRA et al., 2004) e a Onion-tree (CARÉLO et al., 2010). Nesta pesquisa em nível de mestrado enfocamos no uso e na comparação com a Onion-tree, por esta ser um dos MAM's presente atualmente na literatura com menor tempo de resposta a consultas por similaridade (CARÉLO et al., 2010).

A Onion-Tree é um Método de Acesso Métrico robusto e dinâmico, baseado em memória primária. Por ser uma extensão da MM-tree, a Onion-Tree também realiza divisões hierárquicas do espaço métrico, no entanto possui políticas novas de construção e particionamento do espaço métrico que permitem que o espaço seja

dividido em mais de quatro regiões disjuntas por nó. Testes apresentados em (CARÉLO et al., 2010), comprovam que, além da Onion-Tree ser um índice muito compacto, requerendo no máximo 4.8% da memória primária, esta estrutura produz os melhores tempos de construção do índice e os melhores resultados de desempenho no processamento de consulta quando comparada às versões baseadas em memória primária da Slim-Tree e da MM-Tree. As principais propriedades introduzidas pela Onion-Tree consistem no Processo de Expansão e na Técnica de Substituição.

O processo de expansão consiste em uma nova política de particionamento que torna a estrutura de indexação mais balanceada ao aumentar o número de regiões disjuntas definidas pelos pivôs de um nó. No particionamento da MM-Tree, em um nó N, são definidas quatro regiões disjuntas, as regiões I, II, III e IV, onde esta última é caracterizada por ser a maior região mapeada pelo nó. Evidências experimentais demonstram que a diferença no tamanho da região IV comparado ao tamanho das demais regiões resulta em um desbalanceamento na estrutura (CARÉLO et al., 2010), pois muitos elementos são atribuídos a esta. Neste contexto, o processo de expansão da Onion-Tree divide recursivamente a região mais externa do nó em quatro regiões, tornando a estrutura mais equilibrada, rasa e ampla. Ademais, esse processo não degrada o tempo de construção da árvore, uma vez que não exige novos cálculos de distância.

Dois processos de expansão são ilustrados na Figura 2.9, utilizando como exemplo um nó N que inicialmente é composto pelas regiões I, II, III e IV. Com o primeiro processo de expansão, a região IV, que é a região mais externa do nó N, é dividida em quatro regiões disjuntas, gerando o nó N’ composto por sete regiões disjuntas (i.e., I, II, III, IV, V, VI, VII), tal como ilustrado na Figura 2.9. Em seguida, a divisão é expandida à região VII, a qual é a região mais externo do nó N’, gerando as regiões VIII, IX e X ilustradas no nó N’’.

O processo de construção de uma Onion-Tree pode ser realizado de duas maneiras, pela abordagem fixa ou pela dinâmica. Na abordagem fixa, é estabelecido um número fixo de expansões para ser aplicado em todos nós da estrutura. Por outro lado, na abordagem dinâmica, o número de expansões de cada nó é definido dinamicamente, de acordo com critérios como a distância dos pivôs aos seus representantes. Outra política para construir a árvore de maneira balanceada é a técnica de substituição que assegura a melhor divisão hierárquica do subespaço

métrico de um nó folha N durante o processo de inserção de um elemento s_i a este nó. A técnica de substituição verifica se o nó N é melhor particionado ao substituir o elemento sendo inserido por um de seus pivôs, caso verifique a necessidade de substituição do pivô, o raio do nó N também é substituído pela distância entre o elemento sendo inserido e o pivô não escolhido.

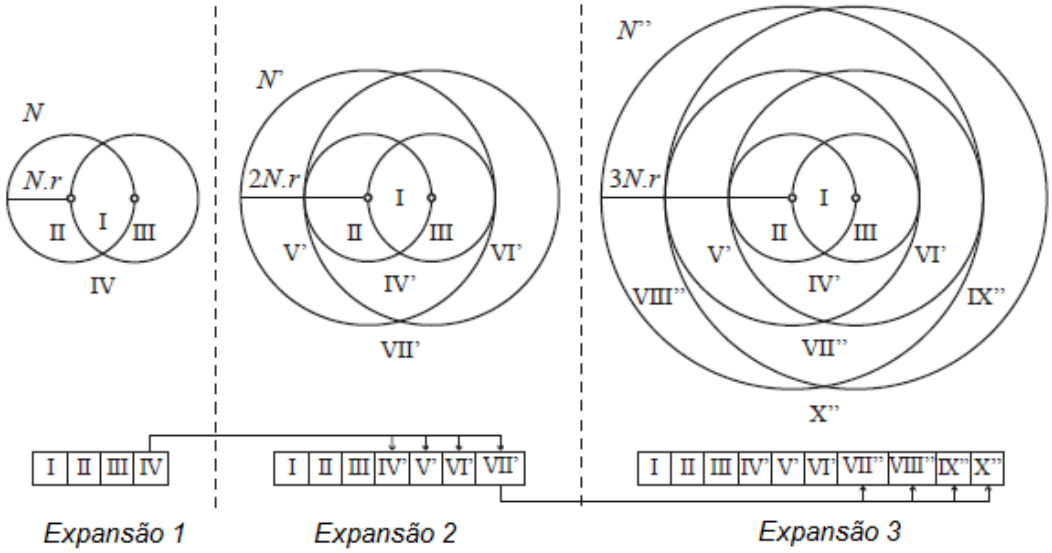

Figura 2.9: Representação das expansões dos nós N, N' e N'', e suas regiões (adaptado de (CARÉLO et al., 2009)).

Devido ao processo de expansão na Onion-Tree, os algoritmos de consulta por abrangência e k-NN foram adaptados a fim de permitir que a consulta por elementos seja realizada em todas as regiões geradas pelo processo de expansão. A adaptação no algoritmo de consulta por abrangência (Algoritmo 1) está representada na linha 8, enquanto a linha 13 do Algoritmo 2 confere a adaptação para a consulta por k-NN.

Tal como em ambientes transacionais (i.e., Online Transaction Processing - OLTP), o processamento de consultas OLAP pode ser acelerada com o uso de estruturas de indexação. No entanto, a utilização de índices convencionais (e.g. B-Tree, KD-Tree) é pouco recomendada devido à natureza multidimensional do conjunto de dados de aplicações de data warehousing. Índices Bitmaps e estruturas de aproximação, por outro lado, são excelentes estruturas de indexação para este cenário multidimensional, pois seu desempenho não degenera em grande escala, ainda que haja várias tabelas de dimensão no esquema estrela (O'NEIL, P.;

GRAEFE, 1995; SARAWAGI, 1997; O'NEIL, P.; QUASS, 1997; GOYAL; ZAVERI; SHARMA, 2006; O'NEIL, E.; O'NEIL, P.; WU, K., 2007, C. Traina Jr. Et. al. omni).

Algoritmo 1: Range (N, sq, rq)	
	Input: N (nó), s_q (centro de consulta), r_q (raio de consulta)
	Output: (ids dos elementos que satisfazem a consulta por abrangência)
1	**if** N **= null then return;**
2	**else**
3	$d_1 \leftarrow d(s_q, N.s_1)$;
4	$d_2 \leftarrow d(s_q, N.s_2)$;
5	**if** $d_1 \leq r_q$ **then** *Add* $N.s_1$*.id to the Result***;**
6	**if** $d_2 \leq r_q$ ***then*** *Add* $N.s_2$*.id to the Result***;**
7	**for** *Region* $\leftarrow 1$ **to** *N.Region* **do**
8	**if** *Query radius interesects region N.Son[Region]* **then**
9	*Range (N.Son[Region],* s_q*,* r_q*);*
10	**end**
11	**end**
12	**end**

Algoritmo 2: k-NN (N, sq, k, ra)	
	Input*: N (nó),* s_q *(centro de consulta), k (k nearest),* r_a *(raio ativo)*
	Output*: (ids dos elementos que satisfazem a consulta por k-NN)*
1	**if** *N* = **null then return;**
2	**else**
3	$d_1 \leftarrow d(s_q, N.s_1)$*;*
4	$d_2 \leftarrow d(s_q, N.s_2)$*;*
5	**if** *Result.Size() < k* **then** *ra* $\leftarrow \infty$*;*
6	**else** $r_a \leftarrow$ *Result[k].Distance;*
7	**if** $d_1 \leq r_a$ **then** *Add* $N.s_1$*.id* **to** *the Result, mantenha armazenado;*
8	**if** $d_2 \leq r_a$ **then** *Add* $N.s_2$*.id* **to** *the Result, mantenha armazenado;*
9	*Order* $\leftarrow$ *Visit order of* s_q *on N;*
10	**for** *i* $\leftarrow 1$ **to** *N.Region* **do**
11	*Region* $\leftarrow$ *Order.Next();*
12	**if** *Query radius intersects region N.Son[Region]* **then**
13	*KNN(N.Son[Region],* s_q*, k,* r_a*);*
14	**end**
15	**end**
16	**end**

2.4.2 Técnica Omni

A técnica Omni consiste em um mecanismo de filtragem baseado em representantes globais, que reduz o número de comparações necessárias para responder consultas por similaridade a partir de uma região de aproximação.

Experimentos apresentados em (FILHO et al., 2001; TRAINA-JR. et al., 2007) evidenciam que a técnica Omni proporciona melhores resultados quando comparado aos métodos de acesso métricos com relação ao tempo gasto no processamento de consultas por similaridade e com relação ao número de acessos a disco.

O processamento de consultas por abrangência é acelerada com o uso da técnica Omni, pois o número de imagens selecionadas para comparações por similaridade é reduzida com o estabelecimento de uma região de aproximação denominada mbOr (*minimum-bounding-Omni-region*, ou seja, a mínima região limitada pela Omni). Nesse contexto, a consulta por abrangência é realizada em duas etapas: filtragem e refinamento.

Na etapa de filtragem, para um espaço métrico $\mathrm{M} = \langle S, d(\,) \rangle$, a mbOr é determinada a partir dos valores de distância da imagem de consulta s_q com relação aos representantes globais e a partir do raio de abrangência r_q definido pelo usuário. Logo, a Equação (2.9) defini a $\mathrm{mbOr_F}$ (s_q,r_q) com uma consulta centrada em $s_q \in S$, um raio de abrangência r_q e para os representantes globais do conjunto S definido por $\mathrm{F} = \{f_1, f_2, \ldots, f_h \mid f_g,\ f_j \in \mathbb{S},\ f_g \neq f_j, \forall \mathrm{g} \neq \mathrm{j}\}$, onde cada f_g é um objeto de S, e h é o número de representantes para a base S.

$$mbOr_F\ (s_q, r_q) = \cap_{g=1}^{h} I_g \qquad (2.9)$$

De acordo com a Equação 2.9, I_g é um subconjunto composto por $s_i \in S$, cuja distância ao representante f_g (i.e., $\mathrm{df_g(s_i)}$) possui valor de no mínimo $I_g^{inf} = \mathrm{df_g(s_q)} - \mathrm{r_q}$ (ou zero, caso $\mathrm{df_g(s_q)} \geq \mathrm{r_q}$) e de no máximo $I_g^{sup} = \mathrm{df_g(s_q)} + \mathrm{r_q}$, onde $\mathrm{df_g(s_q)}$ consiste no valor de distância da imagem de consulta s_q ao representante f_g. Formalmente, o intervalo I_g é definido como:

$$I_g = \{s_i \mid I_g^{inf} \leq df_g(s_i) \leq I_g^{sup}, \forall s_i \in S\} \qquad (2.10)$$

Devido à propriedade de desigualdade triangular da função de distância, não há a ocorrência de falsos negativos, ou seja, imagens que são similares à imagem de consulta não são eliminadas pela etapa de filtragem. No entanto, a mbOr pode gerar falsos positivos, o que torna necessário o refinamento desse conjunto. Na etapa de refinamento, é calculada a distância de cada candidato, pertencente à $mbOr_F$ (s_q, r_q), à imagem de consulta s_q e é verificada a similaridade entre elas conforme o raio de abrangência r_q.

O ganho obtido pela técnica Omni se deve a maneira simples e concisa com que as imagens são representadas e filtradas pela distância aos representantes, uma vez que esses dados são previamente calculados e armazenados. No entanto, a escolha dos representantes globais deve ser feita cuidadosamente, pois o número de representantes e a disposição desses influenciam na seletividade da mbOr. Na Figura 2.10, é exemplificada a relação de seletividade entre a mbOr e o número de representantes globais em um espaço bidimensional.

Na Figura 2.10a, todas as imagens do conjunto *S* são submetidas à comparação de similaridade com relação à imagem de consulta s_q, pois nenhuma região de aproximação foi estabelecida a partir da mbOr (i.e., nenhum representante global foi definido). Consequentemente, este é o pior caso por ser muito custoso porquê compara todas $s_i \in S$ à imagem de consulta s_q. Por outro lado, com o estabelecimento de representantes globais, o conjunto de imagens que são submetidos à comparação de similaridade é reduzido. Como ilustrado na Figura 2.10b, um representante gera uma região no espaço em forma de anel, que restringe o conjunto de imagens que são submetidos à comparação de similaridade. Este anel consiste na representação gráfica do intervalo I_g (g = 1), em que as imagens contidas neste anel possuem um valor de distância com relação ao representante f_q entre $\mathrm{df_g}(\mathrm{s_q}) - \mathrm{r_q}$ e $\mathrm{df_g}(\mathrm{s_i}) + \mathrm{r_q}$. Na Figura 2.10b, a $mbOr_F$ (s_q,r_q) gerada pelo intervalo I_g é ilustrada pela região sombreada.

Para um conjunto de *h* representantes globais, onde h > 1, a mbOr é determinada pela intersecção dos intervalos I_g, onde g varia de 1 a *h*, conforme apresentado na Equação 2.9. Desta forma, a representação gráfica da $mbOr_F$ (s_q, r_q) é ilustrada pela região sombreada da Figura 2.10c, gerada pela intersecção dos anéis. Observa-se que a intersecção destes anéis reduz a quantidade de imagens

comparadas, o que resulta em maior agilidade no processamento de uma consulta por similaridade.

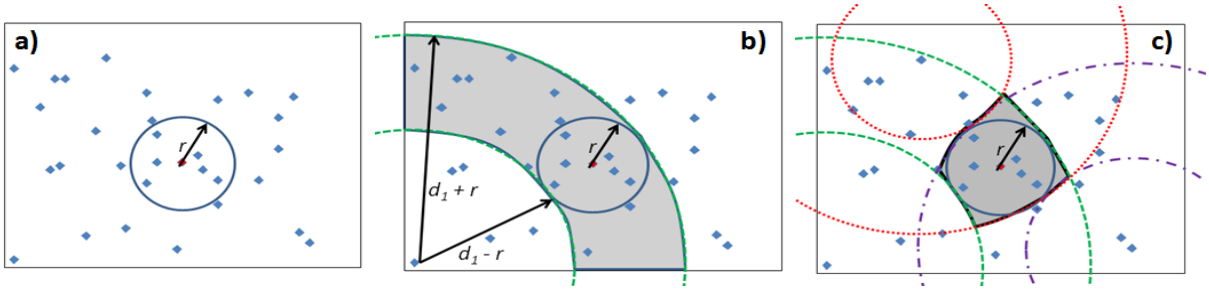

Figura 2.10: Consulta por abrangência com um raio *r* em um espaço 2D. Imagens contidas na mbOr, ilustrada pelas regiões sombreadas, são selecionadas para comparação por similaridade. a) Em um conjunto sem representantes. b) Em um conjunto com um representante. c) Em um conjunto com três representantes, a mbOr próxima da região delimitada pelo raio de abrangência (adaptado de (FILHO et al., 2001)).

Em um conjunto de dados espacial (vetorial ou métrico), (TRAINA-JR. et al., 2007) propõem que o número *h* de representantes deve ser obtido conforme a dimensionalidade intrínseca do conjunto *S*. Uma vez que a dimensionalidade intrínseca do conjunto consiste no número mínimo de atributos que são necessários para representar e diferenciar os objetos de *S*. Nessa pesquisa, foi utilizada a correlação de dimensão fractal D2 (TRAINA-JR. et al., 2000) como uma aproximação da dimensionalidade intrínseca do conjunto de dados *S*. Traina-Jr, et al. (TRAINA-JR. et al., 2007) também sugerem que o número *h* de representantes deve ser igual a $\lceil D2 \rceil$ +1, em que D2 consiste na correlação de dimensão fractal do conjunto de dados *S*, ou seja, na aproximação da dimensionalidade intrínseca desse conjunto.

A seletividade da mbOr também é influenciada pela disposição dos representantes globais. Experimentos apresentados em (FILHO et al., 2001; TRAINA-JR. et al., 2007) indicaram que os representantes globais devem ser os mais periféricos (i.e., estarem na borda do conjunto *S*) e serem mais afastados entre si. Essa relação de distribuição pode ser melhor compreendida com os exemplos ilustrados em Figura 2.11a e Figura 2.11b. Por estarem muito distantes entre si e estarem na periferia, os representantes globais definem uma mbOr bem aproximada da região determinada pelo raio de abrangência r_q (Figura 2.11a). Já a mbOr definida pelos representantes da Figura 2.11b, que estão muito próximos e pouco

distribuídos, resultou em uma região mal ajustada com relação à região definida pelo raio de abrangência, o que resulta em um aumento no número de falsos positivos.

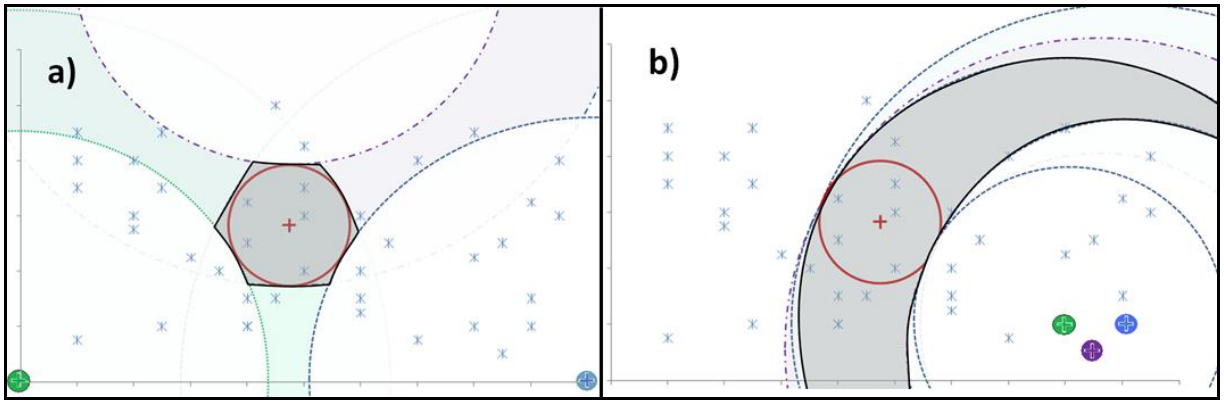

Figura 2.11: Impacto da seletividade gerada pela mbOr conforme a distribuição dos representantes globais (adaptado de (TRAINA-JR. et al., 2007)).

Neste trabalho, foi utilizado o algoritmo HF (Hull-Foci), para identificar os representantes globais de um conjunto *S*. Para maiores detalhes sobre esse algoritmo indicamos a leitura de Caetano (TRAINA-JR. et al., 2007).

2.5 Considerações Finais

A partir da fundamentação teórica e da motivação detalhada no Capítulo 1 - verificamos que existem três grandes desafios no desenvolvimento de ambientes de *image data warehousing*. Mais especificamente estes desafios se referem às camadas de ETL, IDW e servidores IOLAP.

Para a camada ETL o desafio observado consiste na integração dados convencionais e dados sobre o conteúdo intrínseco da imagem, em que dependendo do interesse dos usuários esse conteúdo deve ser representado por diferentes descritores a fim de abranger diferentes percepções visuais.

Sobre o IDW, o desafio refere-se à representação das imagens neste repositório, de modo que o ambiente dê suporte a análises multidimensionais tanto sobre dimensões convencionais quanto sobre dimensões visuais, ao mesmo tempo em que permita que o servidor IOLAP seja flexível a qualquer consulta do usuário. Para tanto, o servidor IOLAP não deve possuir imagens de consulta pré-

determinadas nem valores de raio de abrangência fixos. Além disso, o servidor IOLAP deve prover suporte a diferentes percepções de usuários, isto é, conter imagens processadas por diversos descritores de maneira a gerar camadas perceptuais.

Outro desafio interessante relacionado ao servidor IOLAP consiste na utilização de mecanismos que acelerem o tempo de processamento de consultas IOLAP, uma vez que consultas por similaridade são consideradas custosas e ambientes de DWing são caracterizados por seu grande volume de dados.

Neste contexto, este trabalho realizou uma pesquisa no estado da arte e os trabalhos correlatos estão descritos no capitulo a seguir.

Capítulo 3

TRABALHOS CORRELATOS

Neste capítulo, são descritos os trabalhos envolvendo o conceito de tecnologia de data warehousing na área médica, foco das propostas do iCube e do iStar, e trabalhos envolvendo a criação de métodos de acesso métricos, desde que as propostas do iCube e iStar são comparadas com o método de acesso Onion-tree. Os principais trabalhos correlatos a este projeto, referentes a ambientes de imagens data warehousing também são descritos e comparados.

3.1 Considerações Iniciais

No contexto de DWing, diversas técnicas têm sido propostas na literatura enfocando diferentes linhas de pesquisa relacionadas ao cubo de dados multidimensional. Considerando o enfoque deste projeto, pode-se citar os trabalhos voltados ao cálculo eficiente das agregações que compõem o cubo de dados a partir dos dados mais detalhados do DW (GRAY et al., 1995; SOUZA; SAMPAIO, 1999; HAN et al., 2001). Já no contexto da área médica, trabalhos existentes na literatura apresentam um ponto de vista mais prático, sendo voltados à utilização da tecnologia de DWing com o objetivo de tornar mais eficientes os processos informacionais relacionados à análise dos diagnósticos de saúde (MURPHY et al., 1999; LAI; NICOLLET, 2000; XÉXEO; SANTOS, 2000; CIFERRI et al., 2006). Independentemente do contexto sendo considerado, os trabalhos existentes não enfocam o armazenamento e a manipulação de imagens médicas ou o tratamento de operações baseadas em similaridade em aplicações de DWing.

Um aspecto importante a ser considerado no suporte às consultas por similaridade está relacionado à utilização de índices que possam agilizar o processamento de consultas por similaridade. Para consultas por similaridade em espaços métricos genéricos, os métodos de acesso métrico (MAM) são os mais adequados.

Embora MAM melhorem o desempenho no processamento de consultas por similaridade, MAM não enfocam características intrínsecas de aplicações de DWing, tais como a multidimensionalidade dos dados e a organização dos dados do DW em diferentes níveis de agregação, além de suporte às operações OLAP típicas. Assim, consultas tal como "*Qual a quantidade de imagens similares a uma certa imagem de consulta referentes ao ano de 2007, ao estado de SP para pacientes do sexo masculino?*", não são prontamente atendidas por um MAM.

Para tanto, visualizamos que um MAM deve ser estendido. Um MAM deve incorporar em cada entrada do índice atributos referentes às dimensões do cubo de dados. Por exemplo, associar a data da imagem, o estado de origem e o sexo do paciente. Com isto, pode-se recuperar os IDs das imagens que possuam as características requeridas pela consulta e depois contar a quantidade de imagens. Porém, a existência de hierarquias de atributos e de várias dimensões pode requerer uma grande quantidade de atributos em cada entrada do índice. Uma alternativa é a criação de um MAM distinto para cada combinação de agregações. Por exemplo, um MAM para indexar imagens por dia por estado e por paciente e outro MAM para indexar imagens por mês, por região e por tipo de paciente. Estas considerações consistem em uma primeira contribuição deste trabalho, desde que definem como um MAM pode ser estendido para prover suporte a ambientes IDWing (Capítulo 4 -).

Nas seções seguintes, serão detalhados os trabalhos identificados como correlatos ao enfoque deste trabalho de pesquisa.

3.2 Osmar Zaïane

A pesquisa de Zaïane (ZAÏANE et al., 1998) possui duas grandes motivações. A primeira motivação é relativa à crescente necessidade de desenvolver técnicas para o gerenciamento de dados multimídia, ou seja, para processos de

representação, armazenamento, indexação, recuperação de dados multimídia e, principalmente, para o desenvolvimento de pesquisas na área de mineração de dados multimídia. A segunda motivação, vista como solução para a primeira, refere-se às características e funcionalidades de um cubo de dados que permite a aplicação de métodos de mineração de dados e viabiliza análises multidimensionais dos dados. É importante salientar que em se tratando de dados multimídia essas funcionalidades são baseadas no conteúdo intrínseco e extrínseco desses dados.

Nesse contexto, Zaïane propõe o MultiMediaMiner, um sistema de mineração de imagens e vídeos disponíveis na internet, constituído pelas ferramentas DBMiner e C-BIRD. A mineração desses dados é realizada em quatro camadas: caracterização, comparação, classificação e associação.

O módulo de caracterização extrai um conjunto de características em diversos níveis de abstração de um conjunto relevante de dados multimídia. Como esta extração ocorre em diversos níveis de abstração, o sistema permite que o usuário seja capaz de analisar os dados em múltiplos níveis e realizar as operações *roll-up* e *drill-down* com base nas características da imagem.

O módulo de comparação visa descobrir características que diferenciam (i.e., representam) as classes de dados multimídia. Este módulo compara e distingue as características gerais de uma classe alvo em relação a um conjunto de dados conhecido como classe contrastante.

O módulo de associação identifica regras de associação sobre conjuntos de dados relevantes. Tendo o objetivo de apresentar os padrões que frequentemente ocorrem em um conjunto de dados. Já o módulo de classificação dos dados multimídia é baseado em classes pré-estabelecidas e tem a finalidade de classificar os dados multimídia.

Como descrito em seu artigo, para cada dado multimídia processado pelo MultiMediaMiner são coletadas informações de descrição, descritores de *layout* e características intrínsecas do dado multimídia. As características intrínsecas são armazenadas em um esquema estrela a partir da dimensão cor, em que são armazenados os vetores de 512 cores, de MFC (*Most Frequent Colour*) e de MFO (*Most Frequent Orientation*). Já para descritores de descrição são coletados o nome do arquivo, URL da imagem, formato da imagem, URLs pais, palavras-chave, *thumbnail* (i.e., versões reduzidas de imagens, utilizadas para agilizar o processo de

recuperação e reconhecimento). Por fim, os dados de descritores de *layout* são os vetores de *layout* de cor e de *layout* de borda.

O esquema estrela do MultiMediaMiner possui dez dimensões (Figura 3.1): tamanho em bytes da imagem ou vídeo; altura e largura de cada *frame* ou imagem; data da criação da dado multimídia; formato do arquivo; duração da sequência de frame em segundos como uma hierarquia numérica (e.g. zero segundos corresponde a uma imagem); domínio da página que o dado multimídia foi obtido; domínio das imagens ou vídeos da Internet; palavras-chave; dimensão cor; dimensão orientação de borda. O autor relatou que hierarquias foram construídas automaticamente para em cada dimensão, o que permite que consultas complexas de *roll-up* e *drill-down* sejam realizadas. Elementos básicos da arquitetura do cubo de dados multimídia, possíveis consultas processadas pelo MultiMediaMiner e das hierarquias geradas automaticamente não foram definidos nem apresentados no artigo.

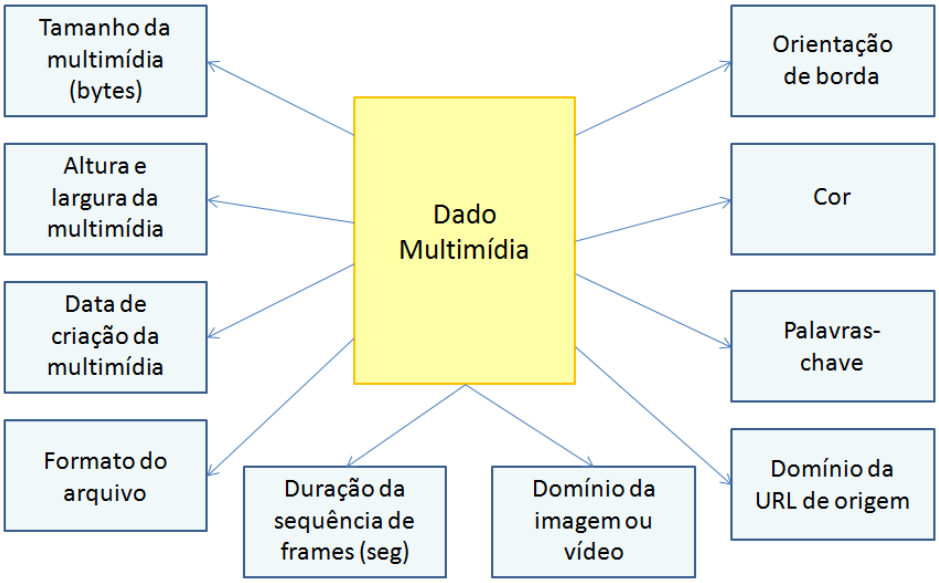

Figura 3.1: Esquema estrela do MultiMediaMiner.

Torna-se evidente que o presente projeto de mestrado possui foco distinto do trabalho apresentado por Zaïane, o qual propõe um cubo de dados multimídia para a mineração de dados multimídia, para o qual as principais operações realizadas são: comparação, classificação, associação e caracterização. Ademais, o próprio autor declara que esse esquema estrela de dez dimensões é limitado com relação ao

número de intervalos de cor e textura e consequentemente torna-se limitado para a execução de operações IOLAP.

3.3 Jane You

As pesquisas de Jane You em (YOU et al., 2004) foram motivadas pelo fato de sistemas de CBIR serem computacionalmente caros e dependentes de domínio, o que restringe e dificulta a sua implantação. Além do mais, a maioria dos sistemas de CBIR possui capacidades limitadas, sendo destacada no artigo a incapacidade de processar consultas resumidas e de integrar múltiplas características de imagens para medir similaridade.

Com esses argumentos, o trabalho de You tem como objetivo apresentar uma nova abordagem de integração, recuperação e armazenamento de dados multimídia, a partir da extensão e união dos conceitos de DW e CBIR. Sua proposta consiste em uma estrutura de *multimedia data warehouse* para facilitar os processos de indexação, consultas dinâmicas e hierárquicas, seleção de características baseada em estatística e, principalmente, na recuperação de informações multimídia.

Como descrito em seu artigo, os dados multimídia são armazenados em um esquema *starflake* com a finalidade de integrar múltiplos dados de multimídia em hierarquias de representação e indexação. Em um esquema *starflake*, cada campo de tabela pode ser considerado como uma tabela de fatos para o nível inferior e os detalhes deste campo são especificados nas tabelas de dimensão associadas (Figura 3.2). Assim, a tabela de fatos do primeiro nível define e armazena dados de cinco grandes classes de multimídia, as quais são especificadas nas seguintes dimensões: texto, áudio, vídeo, imagem e gráfico. Para cada classe de dados multimídia é gerado um esquema *starflake* que representa os dados básicos de seu conteúdo.

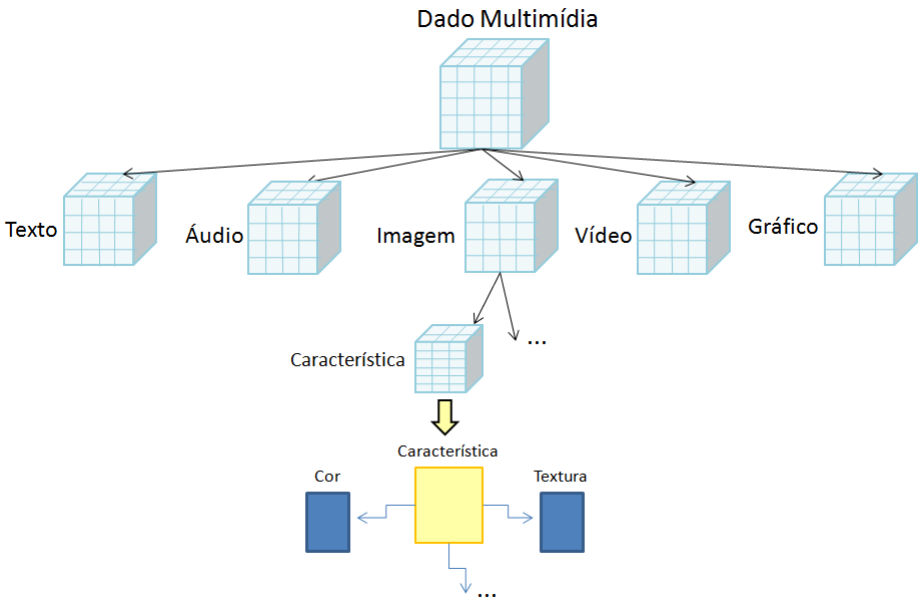

Figura 3.2: Múltiplos níveis de agregação de dados multidimensional.

As características extraídas da classe de multimídia imagem são relacionadas à cor, textura e forma. Os dados de cor são representados por um esquema estrela com sete dimensões associadas: pontos de interesse, histograma global de cor, momentos de cor, média do coeficiente de Wavelet para 4 direções, horizontal, vertical, diagonal e global. O esquema estrela que modela os dados de textura possui seis dimensões: pontos de interesse de borda, parâmetros de contorno, momentos invariantes, coeficiente de B-spline, horário e lugar. Já as características referentes a forma são armazenadas em um esquema estrela com 8 dimensões: transformada de Wavelet, o primeiro, segundo e terceiro momento central de cada camada e suas respectivas médias; e B-spline (i.e., *active contour model*). Toda essa estrutura de dados é denominada pela autora como MediaHouse e seu modelo lógico é apresentado na Figura 3.3.

As principais técnicas utilizadas para a extração das características foram: transformada de Wavelet, histograma de cor, momentos de cor e *Active Contour Model* (B-Spline) para identificar objetos de borda. O método estatístico de simetria de Tau foi estendido para realizar seleção de múltiplas características. Também é importante destacar que antes de avaliar a similaridade das imagens seus dados sofreram normalização gaussiana. Por fim, as distâncias de Hausdorff e de Cosine foram utilizadas para medir a similaridade entre as imagens.

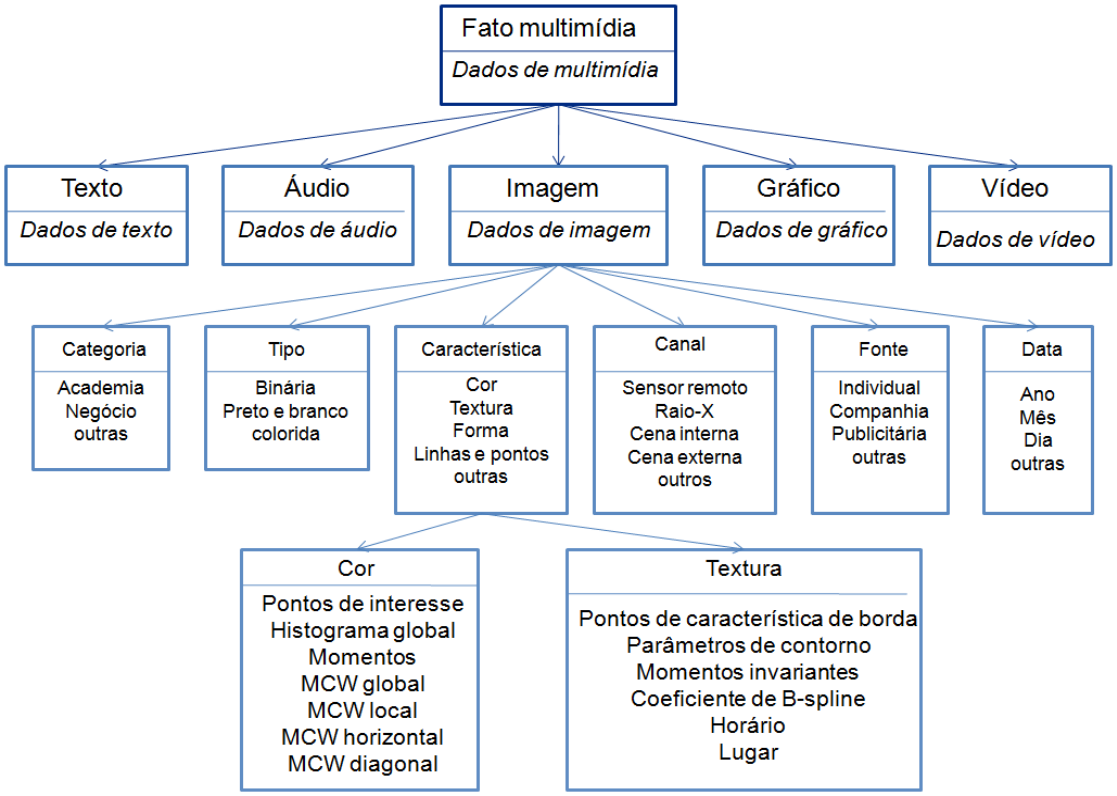

Figura 3.3: Esquema starflake do MediaHouse.

Diversos estudos de casos foram apresentados por You, os quais podem ser resumidos como método de classificação de superfície de imagens de satélite, classificação de imagens de satélite por identificação de padrões de ciclones tropicais, identificação pessoal pelo reconhecimento da palma da mão e reconhecimento de faces.

Mesmo que a proposta do MediaHouse seja desenvolver um multimedia data warehouse, o destino de utilização deste é diferente da proposta do iCube. Visto pelos experimentos realizados, o cubo de dados do MediaHouse tem a finalidade de auxiliar o usuário no processo de classificação de imagens baseados em conteúdo. Ademais, nenhuma operação OLAP ou função de agregação baseada nas características de conteúdo da imagem foram propostos nesse artigo.

Mesmo que os autores tenha declarado que o MediaHouse foi proposto com o intuito de integrar múltiplas características de imagens para medir similaridade, a forma com que o DW foi projetado é inflexível às intenções de consulta do usuário. Segundo a proposta do MediaHouse, as imagens sempre são descritas pelos mesmos extratores, o que torna essa abordagem muito restrita a estas percepções.

Ademais, o MediaHouse foi desenvolvido em um esquema *starflake*, que consiste de tabelas normalizadas e desnormalizadas. Os dados multimídia armazenados no DW são: texto, áudio, vídeo, imagem e gráfico. As características intrínsecas das imagens são armazenadas nas dimensões cor, textura e forma. Como relatado pela autora, este esquema impacta no desempenho de consultas. Dentro deste contexto, torna-se evidente que a proposta do iCube aborda estratégias e conceitos diferentes sobre IDWing.

3.4 Stephen Wong

Stephen Wong (WONG et al., 2004) relata que mesmo havendo uma ampla quantidade de dados gerados e armazenados digitalmente em centros médicos, como imagens médicas, dados de pacientes e relatórios de diagnóstico, o acesso e análise desses dados, fora do âmbito de operações clínicas diárias, é um processo longo e tedioso. Além disso, o fato dos dados estarem digitalmente armazenados não reduz o tempo do processo de diagnóstico. O diagnóstico de epilepsia, por exemplo, leva semanas para ser concluído, pois é necessário recuperar todos os dados relacionados em diversas fontes de dados e manipulá-los adequadamente. Outro agravante apontado pelo autor deve-se aos poucos esforços da comunidade científica e médica para desenvolver sistemas de informação que sejam capazes de integrar dados biomédicos originados em múltiplos laboratórios e hospitais.

Como estes agravantes estão diretamente relacionados aos principais objetivos de um hospital escola, que consistem em prover cuidados médicos, formar novos médicos e conduzir pesquisas biomédicas, Wong propõe uma nova classe de pesquisa sobre *data warehouse* utilizando dados multimídia, chamado NIDS (*Neuroinformatics Database System*). Este sistema permite que estudos clínicos e científicos sejam realizados utilizando dados de pacientes de epilepsia lobo-temporal não tratável.

O NIDS permitir realizar análises e processamento de imagens, além de consultas em texto aberto de laudos de pacientes e consultas complexas, tais como “Encontre as pacientes do sexo feminino com mais de 21 anos, com atrofia no hipocampo direito > 10% e corresponde a depleção de N-acetil-aspartato > 15%”.

Para tanto, diversos tipos de multimídias são armazenados no NIDS, dentre textos livres, laudos estruturados, imagens, sinais, gráficos, vídeos e laudos em papel digitalizados. Mais especificamente, o autor declara que os dados utilizados para alimentar o DW foram:

- Dados cadastrais do paciente e referentes a consultas, oriundos do sistema de informação hospitalar;
- Dados oriundos do sistema de *Picture Archiving and Communication System* (PACS), como dados de imagens de ressonância magnética, data do exame, etc;
- Dados de exames de tomografia por emissão de pósitrons (PET);
- Dados pré-cirúrgico e cirúrgico oriundos do banco de dados de neurocirurgia;
- Dados de testes psicométricos;
- E alguns vídeos de eletro encefalograma (EEG).

No artigo, também é destacado o fato dos dados de texto serem organizados em índices invertidos, assim o texto é representado por palavras-chave relacionadas ao seu conteúdo, identificadas em um conjunto de objetos catalogados referentes a este domínio. Esta identificação ocorre em uma abordagem a priori por intermédio da ferramenta Symphonia3. Geralmente, índices são armazenados na estrutura de dados árvore-B, devido a sua flexibilidade em inserir, reposicionar e remover índices. As consultas realizadas sobre esses índices é tipicamente aperfeiçoada com a utilização de técnicas de mineração de texto, tais como análise lexical, remoção de *stop-words, stemming*, identificação de sinônimos e atribuição de pesos.

As características de conteúdo das imagens e figuras de interesse também são armazenadas no NIDS como índices. A ferramenta Visibroker e o algoritmo desenvolvido pelo Centro de Imagens Funcionais do laboratório Lawrence Berkeley em conjunto com o Projeto de Análise e Visualização de Imagens Volumétricas da *University of Iowa Medical College*, foram utilizados para processar as imagens médicas e extrair as suas características, as quais são classificadas pelos autores como primitivas e lógicas. As características primitivas são características visuais, tais como volume, cor, textura, forma, atividade metabólica, etc. Já as características lógicas são características de alto nível semântico, tais como normalidade

volumétrica, nível de metabolismo, etc. O processo de extração de características é ilustrado na Figura 3.4.

Foram apresentados diversos estudos de caso sobre a utilização do NIDS em pesquisas de epilepsia como: análise estatística de correlação de variáveis (por exemplo, correlação de idade com volume do hipocampo); estudo clínico sobre a localização de focos epileptogênicos e para compreensão estrutural e funcional do cérebro. Igualmente foram realizadas pesquisas sobre o planejamento cirúrgico, como na determinação de pacientes que sejam bons candidatos para a cirurgia ou quais procedimentos clínicos são mais eficazes. O mesmo sistema foi utilizado para auxiliar no processo de diagnóstico.

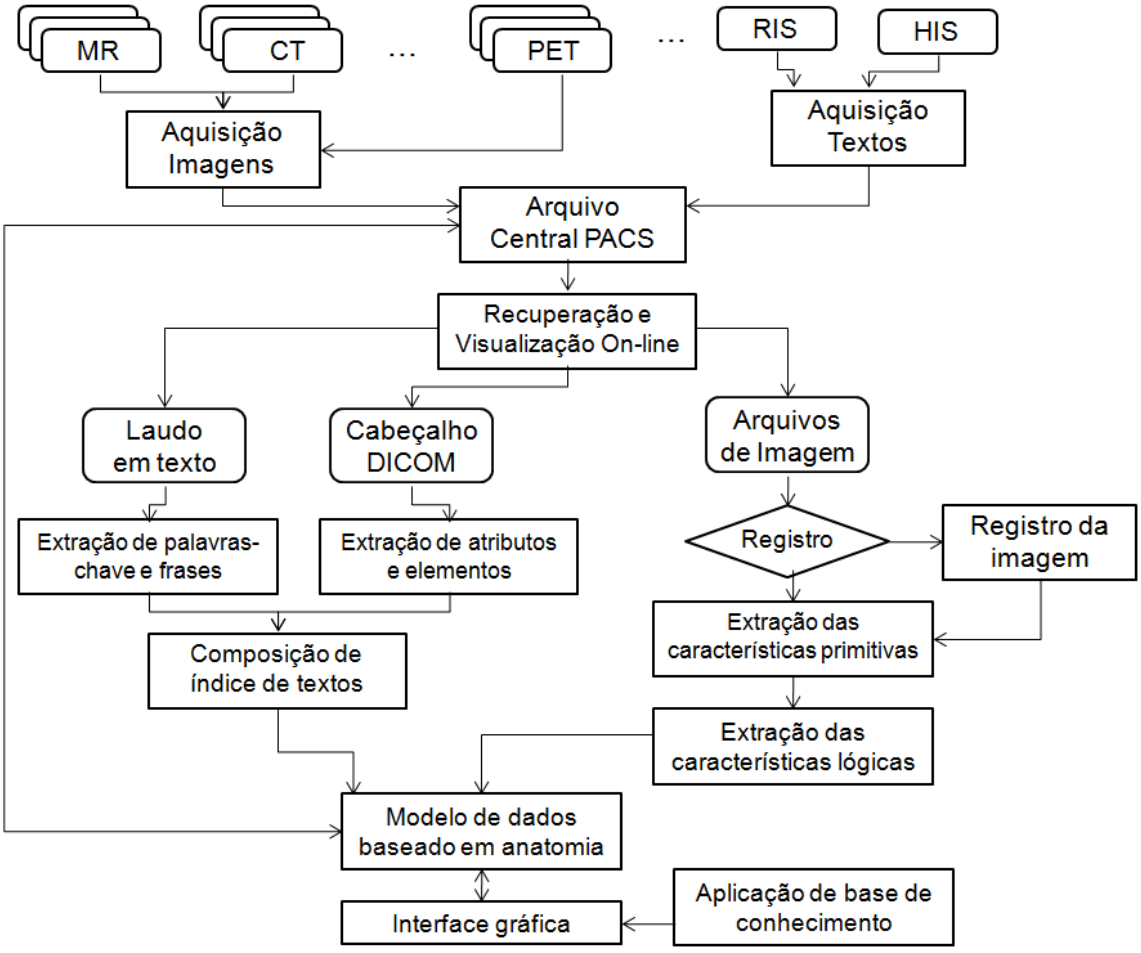

Figura 3.4: Fluxo operacional de extração de características do NIDS (adaptado de WONG, et. al. 2004).

Neste artigo, porém, não foram especificados os descritores de conteúdo utilizados, tão pouco, o modelo lógico de seu IDW e processamento de consultas

IOLAP. Pouco se comentou sobre a existência de índices sobre os vetores de características, visões ou mesmo as organizações desses dados no repositório, também não foi descrito sobre a real comparação de desempenho com medidas deste sistema em relação aos sistemas concorrentes.

3.5 Anne-Muriel Arigon

O problema abordado em (ARIGON; TCHOUNIKINE; MIQUEL, 2006) refere-se à dificuldade de escolher a priori os descritores mais apropriados para caracterizar os dados analisados. Dados multimídia geralmente são caracterizados por descritores que podem ser calculados de várias maneiras e por várias perspectivas de observação. Deste modo, a escolha de descritores é uma difícil tarefa e que está sujeita a más escolhas. Como medida para solucionar este problema, é proposto um modelo multidimensional de multiversões funcionais capaz de gerenciar dados multimídia caracterizados por descritores obtidos por vários modelos computacionais. Assim, o modelo permite que o usuário possa escolher durante a sua análise a melhor função de representação e processamento dos dados (e.g., descritor), em outras palavras, definir a melhor camada perceptual.

O estudo de caso abordado teve como finalidade utilizar esse cubo de dados multiversões funcionais como ferramenta de análise do efeito de um medicamento e um placebo a partir de sinais de eletrocardiograma (ECG), ou seja, foi utilizada como uma ferramenta de auxílio à tomada de decisão. Para tanto, foram extraídas as características intrínsecas dos sinais de ECG, tais como a duração do QT e o nível de ruído presente no exame. A duração de QT é o intervalo de tempo em que os ventrículos são repolarizados, enquanto o nível de barulho é um critério de seleção de grupos de estudo baseando-se na qualidade do sinal. A extração da característica duração de QT foi realizada por dois algoritmos, método de três limiares (T1, T2, T3) e o método de duas ondas T (S1, S2), este fato resgata e justifica o conceito de multiversões funcionais.

O MDW proposto foi modelado em um esquema floco de neve com sete dimensões:

- Três dimensões são referentes ao paciente: patologia, idade e sexo;

- Duas dimensões referentes ao conteúdo dos ECGs: duração de QT e nível de barulho; e
- Duas dimensões relativas a aquisição dos ECGs: tempo e tecnologia.

É importante ressaltar que para cada dimensão os dados foram representados em diversos descritores, assim a organização dos dados é feita pela definição dos membros em *bottom-up* (ou seja, inicia-se com alta granularidade) e estes são ligados hierarquicamente.

A tabela de fatos é composta pelas chaves-estrangeiras das dimensões associadas e por um campo, #ECG, que representa três funções de agregação: ECG-count, ECG-list, Average-ECG. Na Figura 3.5 este esquema floco de neve é ilustrado.

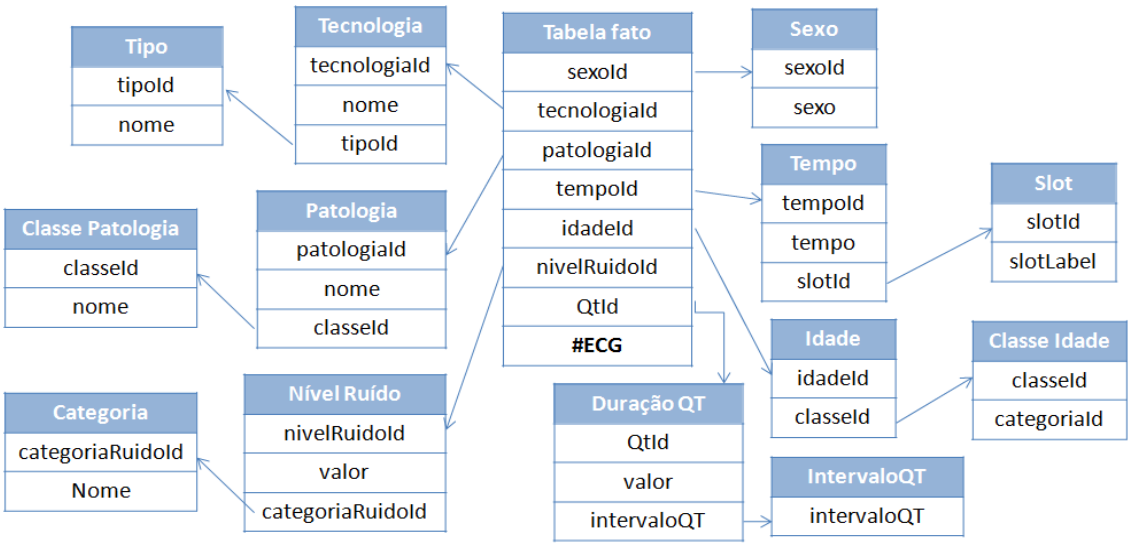

Figura 3.5: Esquema floco de neve do data warehouse do projeto EMIAT (adaptado de (ARIGON; MIQUEL; TCHOUNIKINE, 2007)).

Outro estudo de caso apresentado sucintamente pelos autores refere-se à aplicação desta estrutura de dados para dados de imagens, no caso imagens de cromossomos. Este estudo teve como objetivo comparar tecidos saudáveis com tecidos cancerígenos ou com possíveis anomalias entre si. A análise realizada com o auxilio do MDW consiste na investigação sobre a expressão gênica dos cromossomos.

A expressão de um gene em uma célula pode ser relacionada ao tempo (i.e., idade celular), ao espaço (i.e., localização desta célula no organismo), e/ou ao estado da célula (i.e., normal, doente, respondendo a algum estímulo). Um gene

pode ser caracterizado por vários níveis de expressão calculados de diferentes maneiras e por seu conteúdo de GC (i.e., conteúdo de guanina-citosina). Os métodos de identificação do conteúdo de GC baseiam-se na posição dos pares de GC (GC1, GC2, GC3) ou em regiões do gene onde os pares estão presentes (regiões exon e intron).

O modelo de dados para este estudo de caso possui uma tabela de fatos correspondente ao cromossomo com medidas (número de cromossomos e lista de cromossomos) e chaves estrangeiras das dimensões associadas. As dimensões são compostas por vários descritores das imagens de cromossomo, como táxon, genes, funções da proteína, conteúdo de CG e nível de expressão. Na Figura 3.6, este modelo multidimensional é ilustrado.

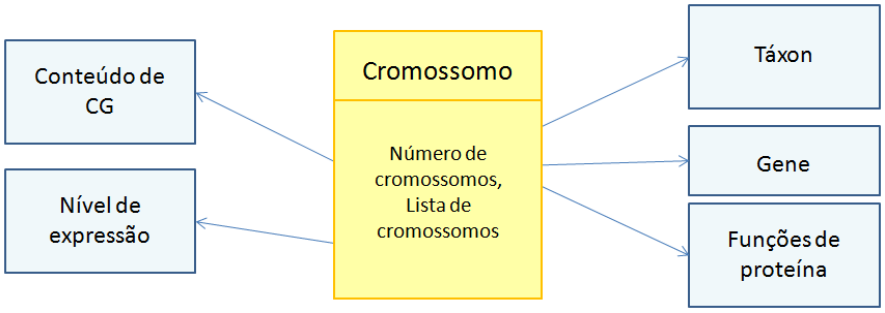

Figura 3.6: Esquema do Bioinformatics Data Warehose.

Mesmo citando alguns métodos de representação do conteúdo do cromossomo, pouco foi detalhado sobre o funcionamento dos descritores de conteúdo da imagem, assim como a organização destes dados no MDW.

A proposta de Arigon se assemelha a este projeto de pesquisa por se tratar de uma ferramenta de auxílio à tomada de decisão utilizando dados multimídia sobre múltiplos descritores. No entanto, como afirmado pela própria autora, esses modelos de dados em que os dados multimídias são considerados medidas da tabela de fatos podem ser melhorados por possuírem algumas redundâncias e por não realizarem o armazenamento das multiversões funcionais na tabela de fatos de maneira otimizada.

Um ponto pouco detalhado no artigo consiste na maneira com que as funções de multiversões são executadas, ou seja, não é possível afirmar em que momento os dados brutos são processados pela função de multiversão, se esta é executada durante a etapa de ETL ou quando as consultas são solicitadas.

Outro fato importante, mas pouco comentado pela autora, refere-se ao desempenho no processamento do consultas IOLAP. Por se tratar de processamento utilizando dados multimídia, o tempo de resposta destas consultas deve ser tratado como um elemento crucial. Nenhum resultado sobre o desempenho do modelo de Anne foi apresentado, apenas uma referência a uma dissertação foi citada como proposta de aperfeiçoamento.

3.6 Menchú Chen

Menshu Chen cita em (CHEN et al., 2008) que trabalhar com vídeos é um grande desafio, mas necessário. É considerado um desafio, pois vídeos são dados não estruturados, que possuem grande quantidade de informações relacionadas, além da descrição de seu conteúdo ser um processo subjetivo e, consequentemente, métodos de recuperação baseados em palavras-chave não são aconselhados.

A necessidade de trabalhar com este tema é justificada pelo autor devido ao crescimento de tecnologias de computadores e a popularização da internet. A quantidade de dados multimídia cresceu surpreendentemente, contudo, existem problemas relativos ao seu armazenamento, gerenciamento e recuperação que ainda não foram resolvidos.

Nesse contexto, os autores propõem uma estrutura em XML para organizar dados multimídia em um cubo de dados, pois DW são muito eficientes no gerenciamento de dados e adequados para volumosos bancos de dados. O esquema estrela proposto por Chen, ilustrado na Figura 3.7, possui quatro dimensões: frames chaves/líderes, informações de legenda, informações de tempo e informações de texto. Os dados que alimentam o DW são organizados em C categorias pelo método *k-means*. O dado líder/chave da categoria (i.e., o mais próximo à média do grupo) é armazenado na tabela de dimensão e utilizado para representá-la.

As dimensões deste esquema são classificadas pelo autor como não-espacial (i.e., dimensão tempo), espacial-não-espacial (i.e., em alguns níveis a dimensão é espacial e em outros não) e espacial-espacial (i.e., em todos os níveis a dimensão é

espacial). Medidas de dimensões não-espaciais são compreendidas como aditivas, sendo possível realizar funções de agregação como soma e média. Já as medidas de dimensões espacial-espacial e espacial-não-espacial são, respectivamente, não-aditivas e semi-aditivas. As funções de agregação sobre essas medidas não-aditiva e semi-aditiva são simuladas a partir do método de agrupamento *k-means*, que é um algoritmo de agrupamento de dados que os organiza baseado nas médias próximas.

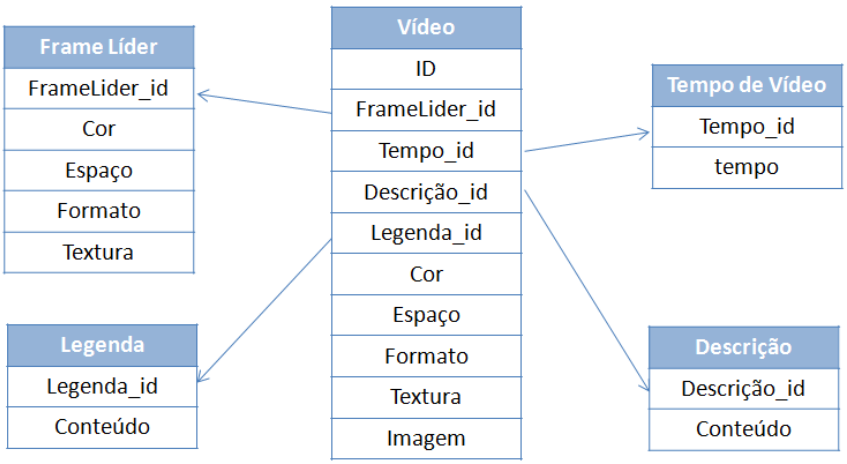

Figura 3.7: Esquema estrela do data warehouse proposto por CHEN (adaptado de CHEN, et. al. 2008).

Consultas realizadas nesta estrutura são baseadas no algoritmo de consulta por abrangência. A análise de similaridade entre os vídeos inicia com o cálculo da distância Euclidiana entre a frame de consulta e o *frame* líder/chave da categoria Cj. Dentre as categorias com a distância menor que o fator de abrangência, é selecionada a categoria com menor distância e, em seguida, aplicado o algoritmo de consulta pelos k-vizinhos mais próximos entre a imagem de consulta e as imagens pertencentes a esta categoria. A similaridade é medida em relação às características intrínsecas dos frames, tais com cor, textura, forma e espaço.

O estudo de caso realizado sobre esta estrutura teve o objetivo de comparar os métodos de recuperação de vídeos baseados em conteúdo entre uma base de dados e um cubo de dados multidimensional. Com este estudo foi possível concluir que há uma relação inversamente proporcional entre o número de categorias estabelecidas e o tempo de execução de processamentos da consultas.

Mesmo relatando que a similaridade entre os *frames* é calculada a partir das características de conteúdo (i.e., cor, textura, forma e espaço), nesse artigo não

foram especificados os descritores de conteúdo utilizados para extrair essas características dos frames, assim como não foi detalhada a maneira com que esses vetores de características foram organizados na estrutura multidimensional. Devido ao fato desse cubo de dados multimídia ser um sistema de recuperação de vídeos baseado em conteúdo e operação OLAP assim como outros tipos de consultas para o auxílio à tomada de decisão não serem abordadas pelo autor; torna-se evidente que o propósito do trabalho de Chen e este trabalho de mestrado são distintos.

3.7 Teh Wah

Em (WAH; SIM, 2009) são descritas as dificuldades enfrentadas no diagnóstico e no tratamento de linfomas, em que na maioria das vezes são utilizados sistemas de auxílio à decisão médica para a execução eficaz de tais tarefas. Como uma alternativa para aperfeiçoar esses processos, Wah propõe a utilização de sistemas de DW de coleta e integração de dados relacionados ao paciente e seu tratamento. Esse sistema realiza análises estatísticas sobre os dados para observar tendências, validando hipóteses e realizando descoberta de conhecimento, além disso, esse sistema tem por objetivo de melhorar o processo de tomada de decisão sobre o diagnóstico e tratamento de linfomas.

Neste trabalho não foram especificados os tipos de dados multimídia que podem ser utilizados, assim como não foram especificadas técnicas de extração, armazenamento e manipulação de características dos dados multimídia. Os autores apresentaram essa proposta apenas em nível conceitual e não foi definida nenhuma característica que permita a representação do cubo de dados nos níveis lógico (i.e., esquema estrela) e físico (i.e., estruturas de árvores que permitem a redução do espaço para representar o cubo de dados).

3.8 Considerações Finais

Considerando o enfoque deste trabalho de mestrado e as suas especificações, a Tabela 3.1 foi elaborada para comparar o contexto de aplicação do iCube com o contexto de aplicação dos trabalhos correlatos descritos nas seções anteriores. Dentre os trabalhos correlatos descritos, apenas três pesquisas são realizadas no domínio médico, destacando o trabalho de Arigon e Wong por utilizarem imagens como dados multimídia para alimentar um IDW. No entanto, como descrito na Seção 3.5, a proposta de Arigon não utilizou descritores de conteúdo convencionais para a extração de características e existem pontos em seu modelo de cubo de dados que devem ser aperfeiçoados.

Mesmo que o trabalho de Wong aborde um domínio muito próximo ao domínio do iCube, em seu artigo, porém, não foram especificados os descritores de conteúdo utilizados, tão pouco, o modelo lógico de seu IDW e processamento de consultas IOLAP para as quais o cubo de dados provê suporte. Pouco se comentou sobre a sua lógica de armazenamento e não foi realizada uma comparação de desempenho deste sistema em relação aos sistemas concorrentes.

Tabela 3.1: Relação entre os trabalhos e os dados utilizados no IDW propostos.

	Imagem	**Esquema estrela**	**Características extraídas da imagem**		
			Múltiplos descritores	**De maneira Flexível***	**Outras**
iCube e iStar	X	X	X	X	Referencia as imagens segundo valores de distância a representantes globais
Arigon	X		X		Conteúdo de GC
You	X		X	-	Pontos de interesse, pontos de borda, etc
Chen		X	X	-	Espaço
Wong	X	-	X	-	Volume, atividade metabólica, etc
Wah	-	X	-	-	-
Zaïane	X	X			

***Flexível a qualquer descritor de conteúdo intrínseco da imagem. IDW não restrito a um conjunto de descritores**

Com o intuito de analisar o objetivo dos artigos propostos a Tabela 3.2 também foi construída. Sabendo que este projeto de pesquisa propõe um IDWing para auxílio à tomada de decisão estratégica, apenas a metade dos trabalhos apresentou essa finalidade. Destacando-se, novamente, Arigon e Wong. Assim, uma vez desenvolvido o iCube, o desempenho e eficácia de alguns destes trabalhos serão comparados.

Como visto, segundo o nosso conhecimento, não existe na literatura um trabalho que apresente e detalhe o processamento de consultas IOLAP (i.e., consultas compostas por predicados convencionais e visuais) utilizando mecanismos de melhoria de desempenho no processamento de consultas.

Tabela 3.2: Assuntos abordados nos trabalhos e o objetivo do IDW proposto.

	Trabalhos abordam			**Uso do IDW para**	
	ETL para imagens	**Modelo lógico do IDW**	**Uso de índices**	**Consultas IOLAP baseadas em similaridade**	**Mineração e classificação**
iCube e iStar	X	X	X	X	
Arigon	-	X		X	
You		X			X
Chen		X		X	
Wong	X			X	X
Wah				X	
Zaïane	X	X			X

Capítulo 4

A Estrutura de Indexação Onion-Tree Aplicada em ambiente de IDWing

Neste capítulo, são descritas as adaptações necessárias para o uso da Onion-Tree em ambientes de IDWing. Estas adaptações são necessárias para que este MAM possa processar consultas IOLAP.

4.1 Considerações Iniciais

A Onion-Tree (CARÉLO et al., 2010) foi adaptada para prover suporte à análise multidimensional, ou seja, para prover suporte a um ambiente de *image data warehousing*. Como descrito nas seções anteriores, a Onion-Tree é um método de acesso métrico robusto e dinâmico, que apresenta atualmente os melhores resultados no processamento de consulta por similaridade. As adaptações realizadas e propostas nesta dissertação preservam a indexação dos vetores de características ao mesmo tempo em que estabelecem um relacionamento de cada imagem aos seus dados convencionais, como exemplo, dados de descrições do paciente, do hospital e da data. Devido a facilidade e aplicabilidade destas adaptações consideramos o uso da Onion-Tree em DWing como uma das tecnologias de *image data warehousing* que existem atualmente e que pode ser comparada a nossa proposta do iStar.

Identificamos três adaptações na Onion-tree que proporcionam suporte à análise multidimensional e que mantiveram as propriedades de indexação deste MAM. Estas adaptações foram denominadas como DWOnion, MultiOnion e SingleOnion. As seções a seguir descrevem quais foram essas adaptações e como as consultas IOLAP podem ser executadas em um ambiente de IDWing. Nestas seções, para fins didáticos foi utilizado o seguinte cenário exemplo:

- O ambiente IDWing deve responder a seguinte consulta: "*Quantas imagens são similares a imagem de consulta* s_q *segundo um raio de abrangência r e concomitantemente são imagens geradas no hospital da macrorregião da Grande São Paulo, nos anos de 1992 e 1993, referentes a pacientes com suspeita de tumor, do estado de São Paulo e com idade entre 30 a 40 anos?*".

4.2 DWOnion

A configuração DWOnion consiste na utilização conjunta de um índice métrico, no caso a Onion-Tree (Figura 4.2), e de um esquema estrela convencional (Figura 4.1), isto é, um esquema estrela que armazena apenas dados convencionais nas duas dimensões e na tabela de fatos e, portanto não armazena dados sobre o conteúdo intrínseco das imagens, os quais são atributos específicos de um IDW. Nesta configuração, enquanto o esquema estrela proporciona o suporte às consultas OLAP (Figura 4.1), o índice permite o processamento de consultas baseadas na similaridade entre as imagens, pois armazena os vetores de características das imagens (Figura 4.2).

Para processar a consulta do cenário exemplo, investigamos três métodos para o processamento de uma consulta IOLAP utilizando a configuração DWOnion chamados de Método de seleção baseado na interseção dos resultados, Método de seleção iniciado pela Onion-tree e Método de seleção iniciado pelo DW.

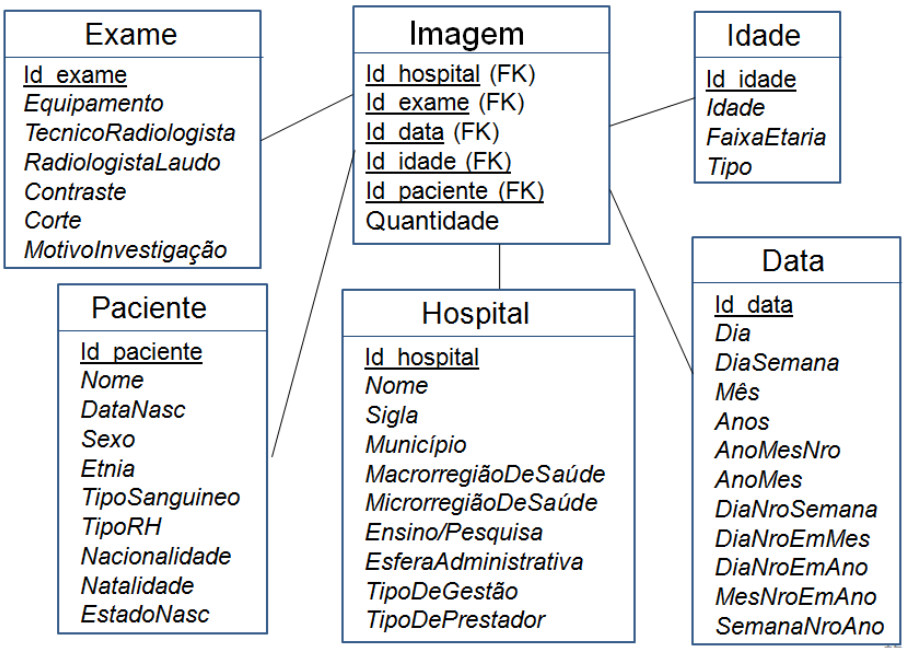

Figura 4.1: Esquema estrela com dados convencionais. Não há o armazenamento de dados sobre o conteúdo intrínseco das imagens.

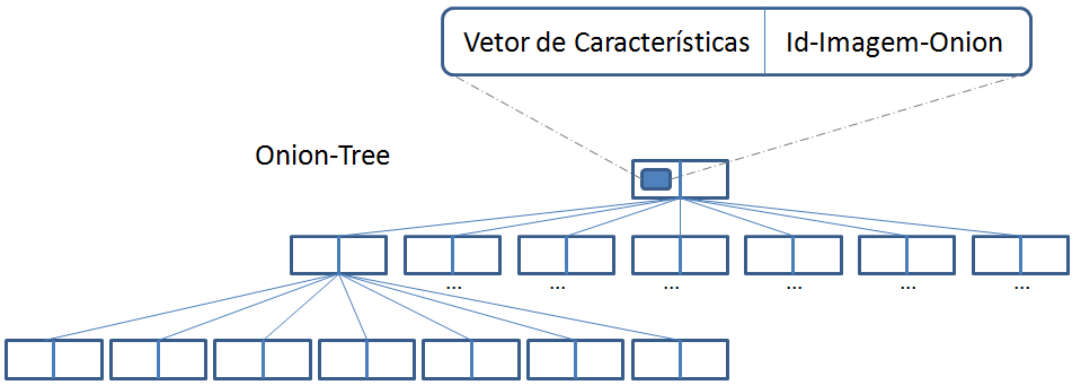

Figura 4.2: Estrutura de indexação da Onion-tree convencional.

4.2.1 Método Seleção baseado na Interseção dos Resultados

O método de seleção baseado na interseção dos resultados é iniciado com o processamento do predicado visual da consulta IOLAP na Onion-Tree para encontrar imagens similares a uma determinada imagem de consulta s_q, gerando a lista ListaOnion de identificadores destas imagens (i.e. lista de id's), como apresentado no Algoritmo 1. Em seguida, o esquema estrela da Figura 4.1 é utilizado para selecionar as imagens conforme os predicados convencionais, idade, data, hospital e paciente). Esta seleção também resulta em uma lista de id's de

imagens, denominada ListaDW (Consulta SQL 1). A resposta final para a consulta do cenário exemplo é obtida pela interseção dos resultados das listas ListaOnion e ListaDW, e após calcula-se o número de id's resultantes dessa interseção.

Algoritmo 1: Range (N, s_q, r_q)	
	Input: N (nó), s_q (centro de consulta), r_q (raio de consulta)
	Output: (ids dos elementos que satisfazem a consulta por abrangência)
1	**if** N **= null then return;**
2	**else**
3	$d_1 \leftarrow d(s_q, N.s_1)$;
4	$d_2 \leftarrow d(s_q, N.s_2)$;
5	**if** $d_1 \leq r_q$ **then** Add $N.s_1$.id **to** the ListOnion;
6	**if** $d_2 \leq r_q$ **then** Add $N.s_2$.id **to** the ListOnion;
7	**for** Region ← 1 **to** N.Region **do**
8	**if** Query radius interesects region N.Son[Region] **then**
9	Range (N.Son[Region], s_q, r_q);
10	**end**
11	**end**
12	**end**

Consulta SQL 1: SelectDW (age1, age2, year1, year2, macro, st)	
1	SELECT id_image AS ListDW
2	FROM Idade, Data, Hospital, Paciente, Exame
3	WHERE Idade BETWEEN age1 and age2
4	AND Ano BETWEEN year1 and year2
5	AND MacrorregiaoDeSaude = macro
6	AND EstadoNasc = st
7	AND Exame.Id_hospital = Hospital.Id_hospital
8	AND Exame.Id_paciente = Paciente.Id_paciente
9	AND Exame.Id_data = Data.Id_data
10	AND Exame.Id_idade = Idade.Id_idade

4.2.2 Método de Seleção Iniciado pela Onion-tree (FirstOnion)

O método de seleção iniciado pela Onion-Tree inicia com o processamento do predicado visual da consulta IOLAP na Onion-Tree, para encontrar imagens similares a uma determinada imagem s_q, resultando na lista de id's de imagens ListaOnion (Algoritmo 1). Em seguida, é consultada quantas imagens do DW estão de acordo com os predicados convencionais e que tenha seu id presente em ListaOnion, tal como no Consulta SQL 2. A consulta é finalizada com o resultado dessa busca ao DW.

Consulta SQL 2: SelectDWM2 (age1, age2, year1, year2, macro, st, ListId)	
1	SELECT id_imagem AS ListDW
2	FROM Idade, Data, Hospital, Paciente, Exame
3	WHERE Idade BETWEEN age1 and age2
4	AND Ano BETWEEN year1 and year2
5	AND MacrorregiaoDeSaude = macro
6	AND EstadoNasc = st
7	AND Exame.Id_hospital = Hospital.Id_hospital
8	AND Exame.Id_paciente = Paciente.Id_paciente
9	AND Exame.Id_data = Data.Id_data
10	AND Exame.Id_idade = Idade.Id_idade
11	AND id_imagem IN ListId

4.2.3 Método de Seleção Iniciado pelo DW (FirstDW)

O Método FirstDW difere do Método FirstOnion por realizar as consultas de forma invertida, pois inicia com a consulta ao DW para identificar quais imagens do DW estão de acordo com os predicados convencionais, gerando a lista ListaDW, como apresentado no Consulta SQL 1. Em seguida, a Onion-Tree é consultada para identificar quais imagens são similares à imagem de consulta s_q e que possuem seu id presente em ListaDW. Como pode ser observado no Algoritmo 2, o método de consulta por abrangência à Onion-Tree foi adaptado para permitir a verificação sobre o identificador da imagem em ListaDW. A consulta é finalizada com o resultado dessa busca na Onion-Tree.

Algoritmo 2: RangeM3 (N, sq, rq, ListId)	
	Input: *N* (nó), s_q (centro de consulta), r_q (raio de consulta), *ListId* (lista de elementos de acordo com os predicados convencionais)
	Output: (ids dos elementos que satisfazem a consulta por abrangência e pela listagem de ids)
1	**if** *N* = **null then return**;
2	**else**
3	$d_1 \leftarrow d(s_q, N.s_1)$;
4	$d_2 \leftarrow d(s_q, N.s_2)$;
5	IdS1 ← checkId($N.s_1$, *ListId*);
6	IdS2 ← checkId($N.s_2$, *ListId*);
7	**if** $d_1 \leq r_q$ ***and*** *IdS1=true* ***then*** *Add* $N.s_1$*.id to the ListResult;*
8	**if** $d_2 \leq r_q$ ***and*** *IdS2=true* ***then*** *Add* $N.s_2$*.id to the ListResult;*
9	**for** *Region* ← *1* ***to*** *N.Region* ***do***
10	**if** *Query radius interesects region N.Son[Region]* **then**
11	RangeM3 (N.Son[Region], s_q, r_q, ListId);
12	**end**
13	**end**
14	**end**

4.3 MultiOnion

Na configuração MultiOnion, exemplificada na Figura 4.3, não há o uso do DW instanciado segundo o esquema estrela. No entanto, são construídas *n* Onion-Trees, onde o valor *n* refere-se ao número de dimensões convencionais existentes no esquema estrela. Portanto, cada árvore desta configuração é responsável por indexar as imagens e armazenar os dados de uma dimensão convencional. Dessa maneira, para o cenário exemplo foram construídas as árvores DataOnion, PacOnion, IdadeOnion, HospOnion e ExamOnion ilustradas na Figura 4.3 e que armazenam, respectivamente, os dados das dimensões Data, Paciente, Idade, Hospital e Exame.

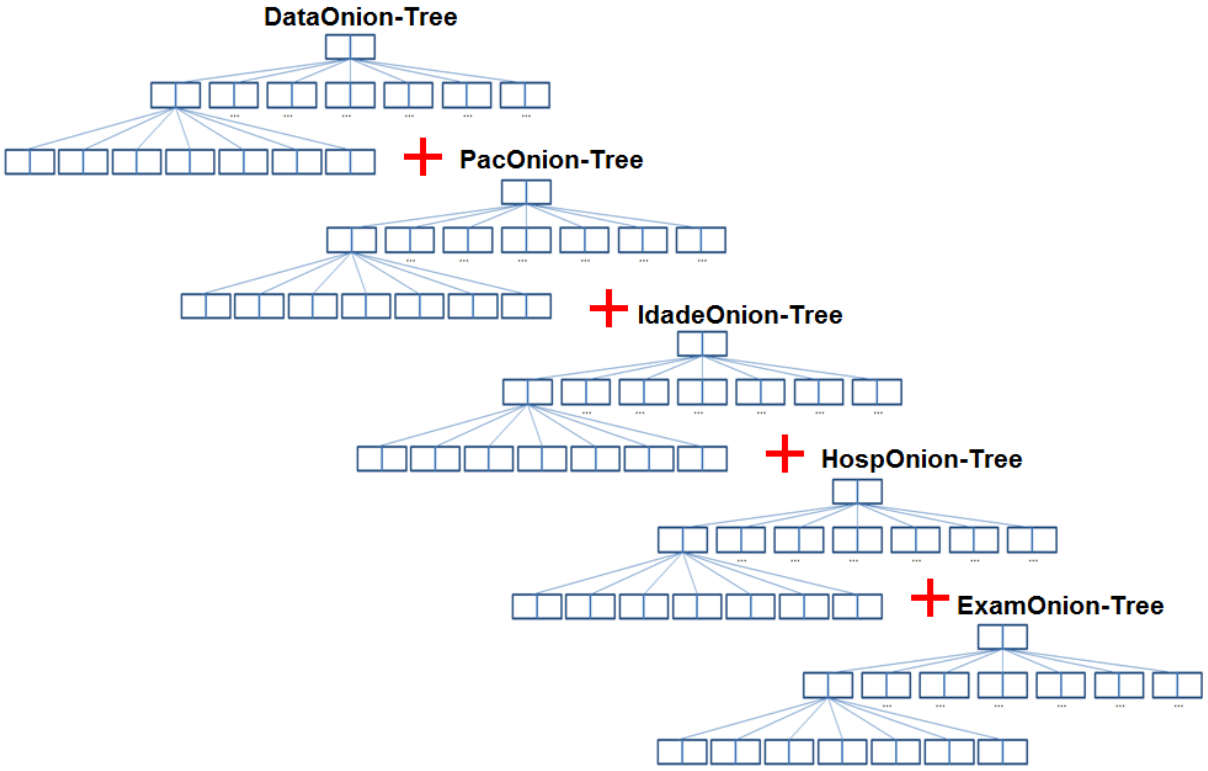

Figura 4.3: Estrutura de indexação da MultiOnion.

Mais detalhadamente cada objeto da IdadeOnion-Tree, ilustrada na Figura 4.4, além de conter o vetor de características e o identificador da imagem, também armazena dados convencionais sobre idade, faixa etária e tipo de idade. Na seção Apêndice A, encontra-se a ilustração das árvores DataOnion-Tree, ExamOnion-Tree, HospOnion-Tree e PacOnion-Tree.

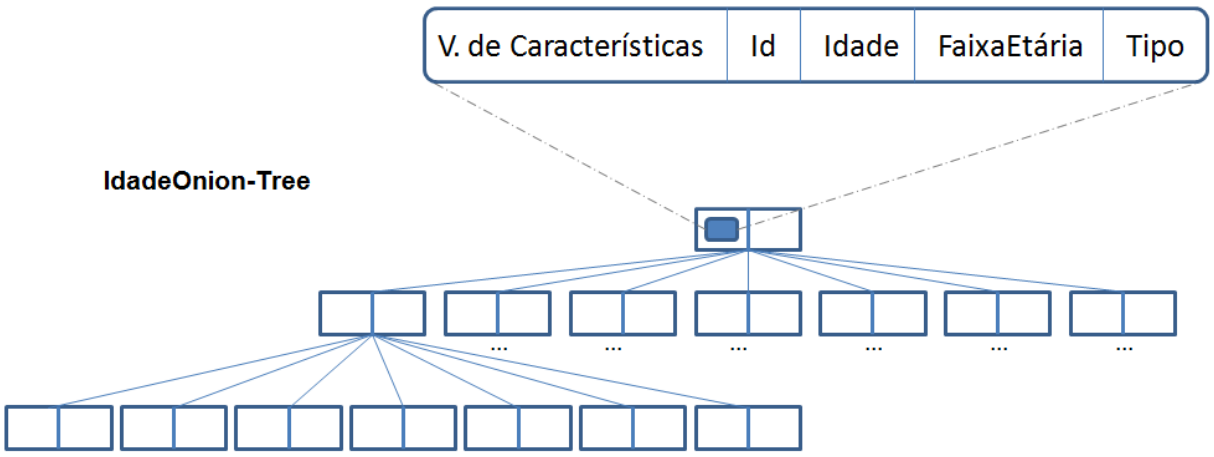

Figura 4.4: Onion-Tree adaptada para o armazenamento de dados sobre Idade do paciente.

O processamento do cenário exemplo segundo a MultiOnion ocorre acessando cada uma das árvores, sendo finalizado com a interseção dos resultados de cada uma das árvores. O acesso a uma das árvores resulta em uma lista de id's de imagens que são referentes às imagens similares intrinsecamente a imagem de consulta s_q e que concomitantemente estão de acordo com o predicado convencional referente à árvore acessada. Por exemplo, ao acessar a IdadeOnion as imagens similares à imagem de consulta s_q e que forem relacionadas a pacientes com idade entre 30 e 40 anos são selecionadas e seus id's são acrescentados a ListaIdade. Adaptações ao método de consulta por abrangência da Onion-Tree foram realizadas para prover suporte à verificação do predicado convencional. O novo método de busca é apresentado no Algoritmo 5. Ao acessar as quatro árvores, o processamento segue com a intersecção das listas ListaIdade, ListaData, ListaPac, ListaExam e ListaHosp, e a consulta é finalizada com o cálculo do número de id's resultantes dessa interseção. As alterações no pseudocódigo estão presentes nas linhas 5 e 6 do Algoritmo 5 para a dimensão de idade, para a qual é verificado o predicado convencional.

Algoritmo 5: RangeAge (N, s_q, r_q, age1, age2)

	Input: *N* (nó), s_q (centro de consulta), r_q (raio de consulta), *age1 e age2* (predicados idade)
	Output: (ids dos elementos que satisfazem a consulta por abrangência e por idade)
1	**if** *N* = **null then return**;
2	**else**
3	$d_1 \leftarrow d(s_q, N.s_1)$;
4	$d_2 \leftarrow d(s_q, N.s_2)$;
5	**if** $d_1 \leq r_q$ ***and*** *N.s1.age between age1 and age2* ***then*** *Add N.s1.id to the ListAge;*
6	**if** $d_2 \leq r_q$ ***and*** *N.s2.age between age1 and age2* ***then*** *Add N.s2.id to the ListAge;*
7	**for** *Region* $\leftarrow$ *1* ***to*** *N.Region* ***do***
8	**if** *Query radius interesects region N.Son[Region]* **then**
9	RangeAge (N.Son[Region], s_q, r_q, age1, age2);
10	e**nd**
11	e**nd**
12	**end**

4.4 SingleOnion

A configuração SingleOnion também não possui um esquema estrela para prover suporte direto ao armazenamento e consulta dos dados convencionais, no entanto difere da configuração MultiOnion pois armazena todos os dados convencionais relacionados a imagem em uma única árvore, tal como exemplificado na Figura 4.5.

Neste contexto, para o cenário exemplo, a consulta é processada percorrendo uma única árvore e verificando quantas imagens são similares à imagem de consulta s_q e que concomitantemente possuem seu atributo “Ano” com valor entre 1992 e 1993, o atributo idade com valor entre 30 e 40 anos, o atributo UF igual a “São Paulo” e macrorregião igual a “Grande São Paulo”, conforme descrito no Algoritmo 6. A consulta é finalizada com o resultado dessa consulta a SingleOnion-Tree. As alterações no pseudocódigo estão presentes nas linhas 5 a 13 e na linha 16 do Algoritmo 6, nas quais os predicados convencionais são avaliados.

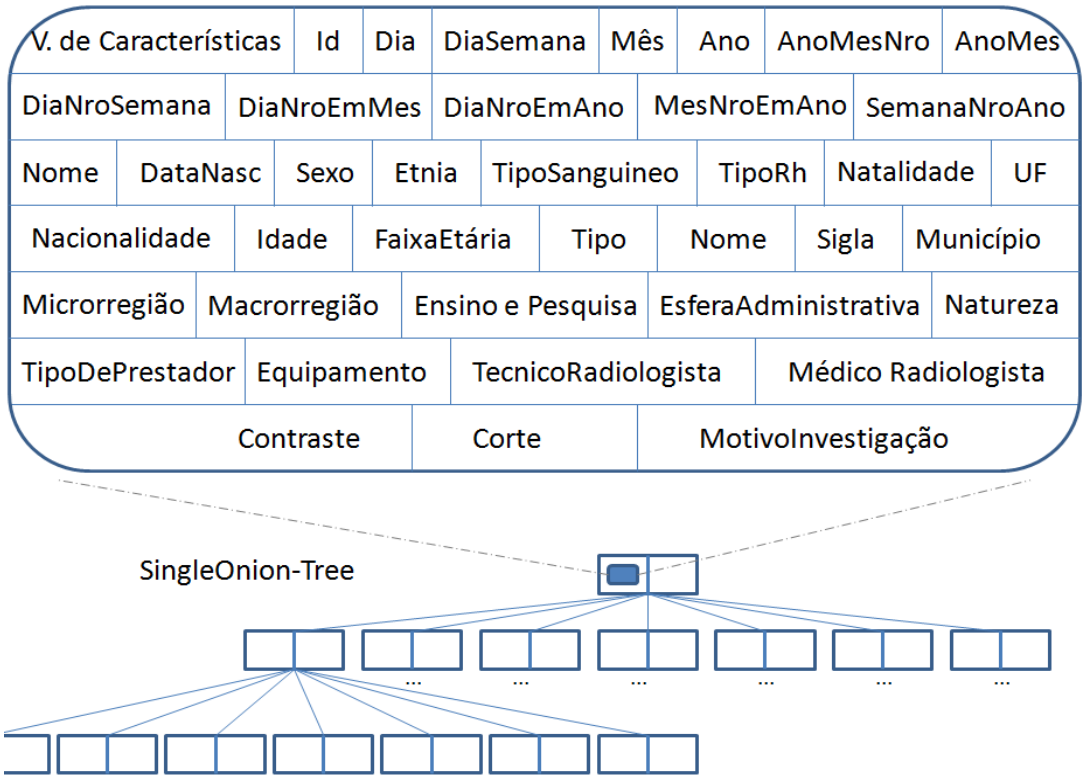

Figura 4.5: Onion-Tree adaptada para o armazenamento de todos os dados convencionais sobre Data, Paciente, Idade, Hospital e Exame.

Algoritmo 6: RangeSingleOnion (N, s_q, r_q, age1, age2, date1, date2, macro, st)

Input: *N* (nó), s_q (centro de consulta), r_q (raio de abrangência), *age1* e *age2* (predicado idade), *date1* e *date2* (predicado ano), *macro* (predicado macrorregião), *st* (predicado estado/UF)
Output: (ids dos elementos que satisfazem a consulta por abrangência e pelos predicados)

```
1  if N = null then return;
2  else
3     d1 ← d(sq, N.s1);
4     d2 ← d(sq, N.s2);
5     AgeS1 ← checkAge(N.s1, age1, age2);
6     AgeS2 ← checkAge(N.s2, age1, age2);
7     DateS1 ← checkDate(N.s1, date1, date2);
8     DateS2 ← checkDate(N.s2, date1, date2);
9     HospS1 ← checkHosp(N.s1, macro);
10    HospS2 ← checkHosp(N.s2, macro);
11    PatS1 ← checkPat(N.s1, st);
12    PatS2 ← checkPat(N.s2, st);
13    if d1 ≤ rq and AgeS1=true and DateS1=true and HospS1=true and PatS1=true
14       then Add N.s1.id to the ListId;
15    end
16    if d2 ≤ rq and AgeS2=true and DateS2=true and HospS2=true and PatS2=true
17       then Add N.s2.id to the ListId;
18    end
19    for Region ← 1 to N.Region do
20       if Query radius interesects region N.Son[Region] then
21          RangeAge (N.Son[Region], sq, rq);
22       end
23    end
24 end
```

4.5 Considerações Finais

O processamento de consultas IOLAP não é possível de ser feito usando apenas o MAM Onion-tree tradicional. Este capítulo primeiramente apresentou um ambiente de IDWing que provê suporte ao processamento de consultas IOLAP utilizando a estrutura de dados Onion-tree com o auxílio de um DW convencional. Para tanto, três métodos de processamento foram visualizados, os quais diferem entre si na ordem de processamento dos predicados, isto é, o primeiro método acessa as estruturas Onion-Tree e DM, que processam a consulta independente conforme seu domínio de predicado e em seguida os resultados são combinados a fim de obter o resultado final da consulta IOLAP. O segundo método primeiro processa a consulta segundo o predicado visual, acessando a Onion-tree e depois verifica se as imagens selecionadas como similares estão de acordo com os predicados convencionais acessando o DW. O terceiro método verifica primeiro os predicados convencionais e depois a checagem sobre a similaridade das imagens com relação a imagem de consulta é feita acessando a Onion-Tree. Além disso, foram propostas a MultiOnion e a SingleOnion para permitir o processamento de consultas IOLAP usando apenas a Onion-tree. Alterações na estrutura de dados da Onion-tree foram realizadas e consequentemente novos algoritmos foram propostos. Os ambientes de IDWing descritos neste capítulo compõem a base para a comparação de desempenho com a proposta do iCube, descrita no próximo capítulo.

Capítulo 5

ICUBE

Neste capítulo, é descrita a proposta do ambiente iCube. Este ambiente de IDWing aborda a extensão da fase de ETL para o suporte a imagens, detalha o esquema de um IDW que relaciona o conteúdo intrínseco das imagens a dados convencionais, e apresenta um processo genérico de processamento de consultas IOLAP. Neste capítulo, também são discutidos os resultados obtidos pelo iCube, os quais são comparados com a tecnologia atual de IDWing.

5.1 Considerações Iniciais

Neste trabalho propomos o iCube, um ambiente de *image data warehousing* que possui as seguintes propriedades:

- Permite o processamento de consultas IOLAP (*Image On-Line Analytical Processing*) com predicados visuais baseados em similaridade de imagens em ambientes de DWing;
- Estende a fase de ETL para permitir o processamento de imagens;
- Faz uso de um novo esquema estrela especialmente adaptado para a representação de imagens;
- Consiste em um ambiente flexível, pois não é restrito a um conjunto fixo de imagens de consulta, tampouco restrito a um valor pré-determinado para o raio de abrangência em consultas por similaridade;
- Introduz o uso da técnica Omni em ambiente de IDWing para acelerar o processamento de consultas IOLAP;

- Estende ao IDW as vantagens de multidimensionalidade já proporcionadas para ambientes clássicos de DWing, tanto para dados convencionais quanto para dados sobre o conteúdo intrínseco da imagem.

Em função das propriedades anteriormente descritas, o iCube permite que uma nova gama de consultas OLAP possa ser respondida por um IDWing. Alguns exemplos de consultas processadas por esse ambiente são:

- *"Quantas imagens são similares a uma determinada imagem de câncer de mama e que ocorreram no ano de 2010 na cidade de Ribeirão Preto?";*
- *"Fornecida uma imagem com certa anomalia, quais são as cidades e as datas que geraram imagens similares a essa?".*

As seções a seguir descrevem o cenário em que o iCube deve realizar o processamento para a consulta exemplo QE: "*Qual a quantidade de imagens que são similares a uma determinada imagem de câncer de mama e que concomitantemente ocorreram no ano de 2010 na cidade de Ribeirão Preto em pacientes de 30 a 40 anos?*". Esta consulta possui três predicados que atuam sobre dimensões convencionais (i.e. ano, cidade e idade), um predicado não-convencional e retorna uma medida numérica de quantidade.

5.2 Proposta de ETL para o iCube

A camada de ETL de um ambiente de IDWing deve realizar a extração, a transformação e o armazenamento de todos os dados que povoarão o IDW, incluindo também os dados sobre as imagens. Em especial, o módulo de transformação foi estendido para que consultas por similaridade possam ser realizadas no iCube. Com o uso desta extensão, as imagens são representadas no IDW por meio de seus respectivos vetores de características e são referenciadas por representantes globais.

Propomos um processo de três etapas para o módulo de transformação de imagens que está ilustrado na Figura 5.1. As três etapas consistem em:

- Etapa 1: Extração de características intrínsecas da imagem s_i;
- Etapa 2: Identificação dos representantes globais;
- Etapa 3: Cálculo da distância de cada imagem do IDW com relação a cada um dos representantes globais.

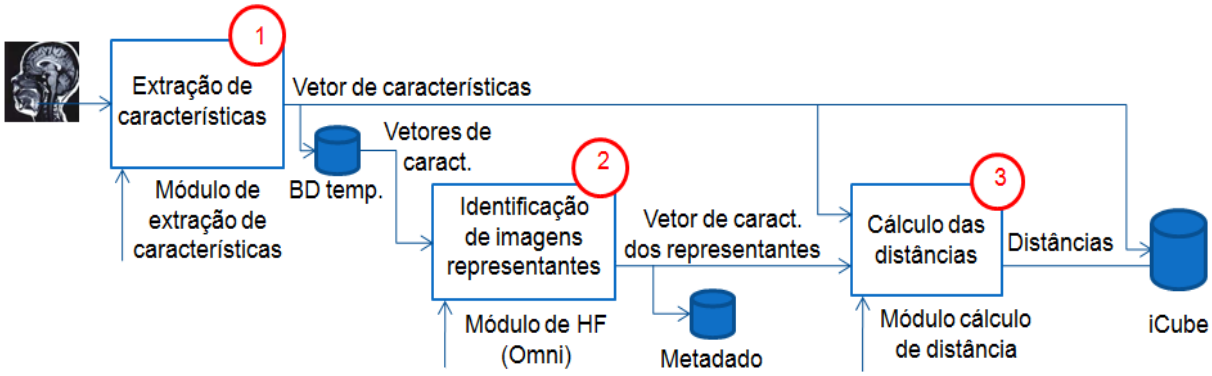

Figura 5.1: Etapas do processo de transformação de imagens na camada de ETL de um ambiente IDWing.

Na primeira etapa (etapa 1 da Figura 5.1), todas as imagens que serão armazenadas no iCube são processadas por um descritor de características. Vale ressaltar que esse módulo consiste na adaptação de um descritor já apresentado na literatura, que gera o vetor de características conforme a intenção de consulta dos usuários. A escolha do descritor, assim como possíveis pré-processamentos sobre as imagens e sobre os vetores, são elementos estritamente relacionados ao assunto escolhido pelos usuários. Dessa maneira, tratamos esses elementos como um módulo caixa-preta que possui como entrada um conjunto de imagens e gera como saída os respectivos vetores de características destas imagens baseado no descritor escolhido pelo usuário.

As etapas 2 e 3 do módulo de transformação geram dados que referenciam as imagens a partir da técnica Omni. Como descrito na Seção 2.4.2, a técnica Omni consiste em um método de aceleração do processamento de consultas por similaridade, que identifica imagens candidatas a similares a partir da definição de uma região de aproximação a imagem de consulta s_q.

Uma vez extraídas as características intrínsecas de todas as imagens a serem armazenadas no iCube, a etapa 2 identifica os representantes globais deste conjunto de imagens. O número de representantes identificados é determinado pela dimensionalidade intrínseca D2 do conjunto e o processo de identificação dos representantes é realizado com base no algoritmo "*Hull of Foci*" (HF) (Seção 2.4.2).

A etapa 3 do módulo de transformação ETL consiste em calcular a distância de cada imagem a ser representada no iCube para cada um dos representantes. Logo, para cada imagem do IDW é calculada e armazenada a sua distância para cada um dos *h* representantes globais.

5.3 Esquema Estrela do iCube

O *image data warehouse* do iCube consiste de um esquema estrela composto por várias tabelas de dimensão convencionais e por uma tabela de dimensão visual voltada ao tratamento de imagens. Pode-se perceber, portanto, que o image data warehouse do iCube é projetado segundo um esquema estrela diferenciado, pois, além de possuir dimensões com dados convencionais, este esquema é caracterizado por possuir uma dimensão dedicada ao armazenamento de dados sobre imagens, denominada dimensão "*Imagem*". Na Figura 5.2, é ilustrado o esquema estrela do iCube que foi modelado para atender o processamento da consulta exemplo QE. Este é apenas um exemplo de instância do esquema estrela proposto pelo iCube.

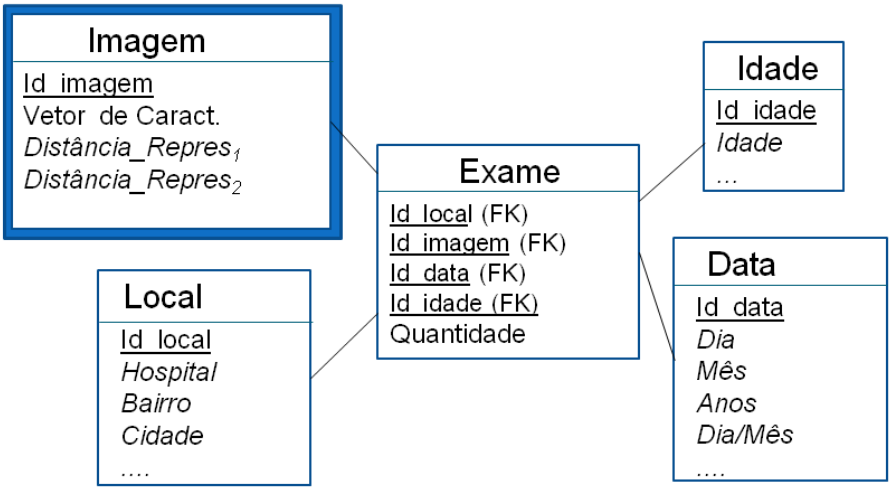

Figura 5.2: Exemplo de esquema estrela proposto para o iCube.

Neste contexto, o esquema estrela do iCube é caracterizado por possuir uma tabela de dimensão denominada *Imagem*, que é o principal elemento de diferenciação deste esquema estrela e que possibilita a execução de consultas por

similaridade no IDWing. No iCube toda imagem $s_i \in S$ é representada na dimensão *Imagem* por uma tupla composta por *h*+2 atributos, onde:

- *h* é igual ao número de representante globais do conjunto (*h* > 0) e armazena o valor de distância da imagem s_i ao representante f_i. Como exemplificado na Figura 5.2, o atributo Distância_Repres$_1$ armazena a distância da imagem s_i ao representante f_1, o atributo Distância_Repres$_2$ armazena a distância da imagem s_i ao representante f_2.
- O valor 2 (de h+2) consiste em uma chave identificadora (id) da imagem s_i e de um atributo para armazenar o vetor de características de s_i, os quais são respectivamente exemplificados na Figura 5.2 por Id_imagem e Vetor de Caract.

As dimensões convencionais são dimensões tradicionais, definidas de acordo com o assunto do IDWing e que armazenam dados que possuem relação de ordem total. Estas dimensões também podem conter hierarquias como em um DW convencional, como exemplo, hospital → bairro → cidade → região.

No iCube a tabela de fatos foi projetada para preservar o relacionamento de cada imagem com os dados convencionais. Para tanto, um campo artificial foi criado para esse tipo de tabela de fatos visando melhorar o desempenho no cálculo de agregações, o qual é sempre povoado com o valor 1. Testes experimentais realizados mostraram que o uso do operador *SUM(*)* é mais eficiente que *COUNT(*)*. Ademais, para níveis superiores do DW, o campo artificial permite o armazenamento da quantidade agregada sem a necessidade de recomputação a partir do nível inferior do DW.

Como é possível observar, o esquema estrela proposto neste trabalho é adaptável a diferentes assuntos, podendo ser aplicado a qualquer contexto. Por exemplo, o iCube pode ser projetado no contexto médico para pesquisar e avaliar tendências de quadros clínicos segundo janelas de tempo, ou perfil dos pacientes (e.g, pacientes do sexo masculino com determinada escolaridade), ou mesmo avaliar a eficácia ou eficiência de um tratamento. No contexto ambiental, este esquema estrela pode ser utilizado para avaliar, a partir de imagens de satélites, o desmatamento em um região ao longo de um período de tempo; ou no contexto agronômico, em que o usuário pode avaliar a ocorrência de uma doença na lavoura segundo a distribuição geográfica, as condições climáticas ou outros fatores.

Ademais, a maneira segundo a qual os dados são organizados permite que qualquer imagem possa ser utilizada como imagem de consulta desde que seja representada pelo mesmo descritor utilizado do IDW. Por fim, o iCube utiliza um filtro simples e eficiente por meio da técnica Omni.

Outra vantagem do esquema estrela proposto consiste no fato de que qualquer tipo de consulta por similaridade pode ser submetido, pois os vetores de características são armazenados univocamente e a técnica Omni pode ser aplicada a diferentes algoritmos de consulta (TRAINA-JR. et al., 2007). Para tanto, é necessário apenas que o servidor IOLAP esteja adaptado para o algoritmo de consulta escolhido. Neste contexto, a seção a seguir descreve o processamento de consultas IOLAP no iCube baseadas em consultas por abrangência.

5.4 Processamento de consultas no iCube com o uso de filtros

Neste trabalho também propomos um processo genérico para o processamento de consultas IOLAP (Figura 5.3). Este processo é realizado em, no máximo, seis etapas, as quais são descritas a seguir:

- Etapa 1: etapa de filtragem baseada em predicados convencionais;
- Etapa 2: etapa de extração do conteúdo intrínseco da imagem de consulta s_q;
- Etapa 3: etapa de cálculo de distância de s_q aos representantes globais do conjunto;
- Etapa 4: etapa de filtragem baseada na mbOr;
- Etapa 5: etapa de refinamento (i.e., consulta por similaridade); e
- Etapa 6: etapa de formulação da resposta à consulta IOLAP.

Nesta proposta, o desempenho no processamento de consultas IOLAP é melhorado por dois módulos de filtragem. O primeiro reduz o número de cálculos de similaridade de acordo com predicados sobre os atributos das dimensões convencionais (i.e., predicados convencionais) e o segundo melhora o desempenho com o uso da mbOr. A descrição a seguir foi elaborada utilizando como base o exemplo corrente, adaptado da consulta exemplo QE: "*Quantas imagens que são similares a uma imagem de consulta s_i (segundo o raio de abrangência r = 10) e que*

concomitantemente ocorreram no ano de 2010 na cidade de Ribeirão Preto em pacientes de 30 a 40 anos?".

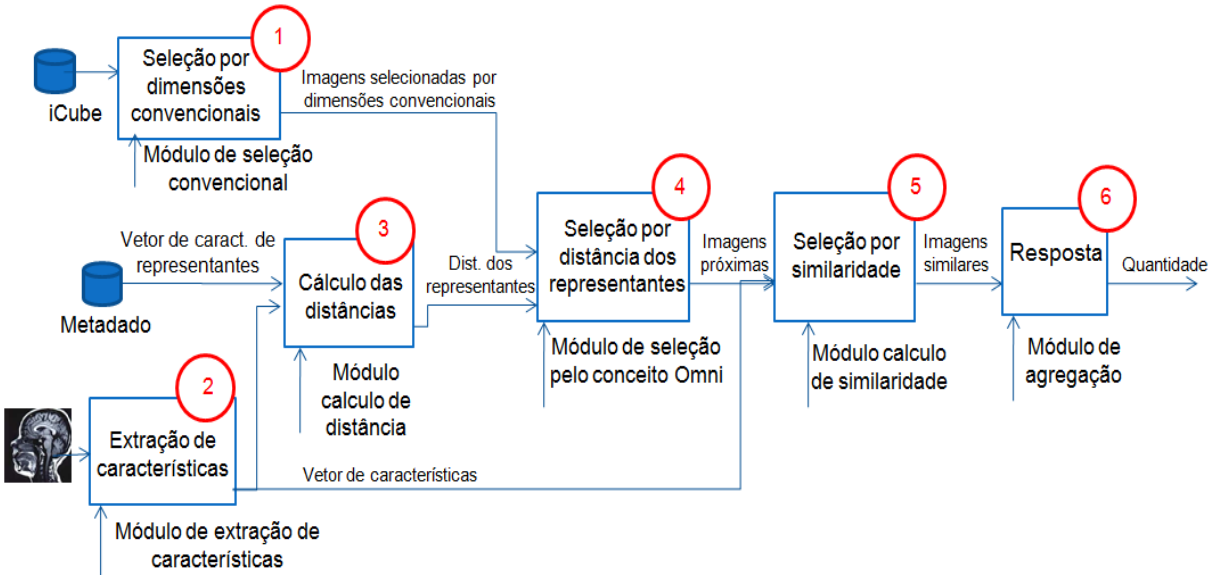

Figura 5.3: Etapas do processamento de consultas IOLAP no iCube.

No iCube o processamento de consulta IOLAP é iniciado com a filtragem do conjunto de imagens do IDW de acordo com os predicados convencionais. Seguindo o exemplo corrente, as imagens que ocorreram no ano de 2010, em pacientes de 30 a 40 anos e de Ribeirão Preto (Figura 5.3 etapa 1) são selecionadas. Essa redução é exemplificada na Tabela 5.1.

Caso nenhuma imagem atenda aos predicados convencionais da consulta, as etapas 2, 3, 4 e 5 não são executadas, e a etapa 6 apresenta o valor zero como resultado para a quantidade retornada pela consulta QE, pois não há imagens que satisfaçam estas condições. Em contrapartida, caso o conjunto resultante da etapa 1 não seja vazio, o processamento da consulta prossegue com a extração do conteúdo intrínseco da imagem de consulta s_q (Figura 5.3 etapa 2). Lembrando que a imagem s_q é processada por um extrator de características e, em seguida, é calculada a sua distância com relação aos representantes globais (Figura 5.3 etapa 3), tal como foi feito para as imagens armazenadas no IDW com o uso da camada de ETL.

Na etapa 4, o conjunto resultante da filtragem por predicados convencionais é submetido a outra filtragem, que ocorre sob os conceitos da técnica Omni. Como descrito na Seção 2.4.2, a técnica Omni permite que as imagens similares à imagem

de consulta s_q sejam filtradas com o estabelecimento de uma região de aproximação, denominada mbOr.

Tabela 5.1: Filtragem baseada nos predicados convencionais.

Representação inicial dos dados referente ao esquema estrela da Figura 5.2 Este cuboid ilustra o conjunto de dados utilizados como entrada para a etapa 1 (filtragem baseada nos predicados convencionais, Figura 5.3).	Seleção decorrente da etapa 1 Figura 5.3. Seguindo o exemplo corrente, as imagens que ocorreram no ano de 2010, em pacientes de 30 a 40 anos e de Ribeirão Preto são filtradas pela etapa 1 e as demais são desconsideradas (tuplas tachadas).

IdImagem	Idade	Cidade	Ano
1	54	Ribeirão Preto	2010
2	33	São Paulo	2010
3	39	Ribeirão Preto	2010
4	32	Ribeirão Preto	2010
5	45	Ribeirão Preto	2010
6	33	Piracicaba	2010
7	38	Ribeirão Preto	2010
8	44	Ribeirão Preto	2010
9	33	Ribeirão Preto	2010
10	31	Ribeirão Preto	2010
11	32	Ribeirão Preto	2009
12	35	Ribeirão Preto	2010
13	40	São Paulo	2010
14	39	Ribeirão Preto	2008
15	37	Ribeirão Preto	2010
16	32	Ribeirão Preto	2010
17	31	Ribeirão Preto	2010
18	34	Ribeirão Preto	2010
19	35	Ribeirão Preto	2010
20	32	Ribeirão Preto	2010

IdImagem	Idade	Cidade	Ano
~~1~~	~~54~~	~~Ribeirão Preto~~	~~2010~~
~~2~~	~~33~~	~~São Paulo~~	~~2010~~
3	39	Ribeirão Preto	2010
4	32	Ribeirão Preto	2010
~~5~~	~~45~~	~~Ribeirão Preto~~	~~2010~~
~~6~~	~~33~~	~~Piracicaba~~	~~2010~~
7	38	Ribeirão Preto	2010
~~8~~	~~44~~	~~Ribeirão Preto~~	~~2010~~
9	33	Ribeirão Preto	2010
10	31	Ribeirão Preto	2010
~~11~~	~~32~~	~~Ribeirão Preto~~	~~2009~~
12	35	Ribeirão Preto	2010
~~13~~	~~40~~	~~São Paulo~~	~~2010~~
~~14~~	~~39~~	~~Ribeirão Preto~~	~~2008~~
15	37	Ribeirão Preto	2010
16	32	Ribeirão Preto	2010
17	31	Ribeirão Preto	2010
18	34	Ribeirão Preto	2010
19	35	Ribeirão Preto	2010
20	32	Ribeirão Preto	2010

Na Tabela 5.2, é exemplificada a filtragem do conjunto de imagens com relação à mbOr. Como a mbOr é definida pela intersecção dos intervalos de distância aos representantes (Equação 1; Seção 2.4.2), logo a operação lógica que seleciona as imagens é AND.

É interessante o uso desta técnica de aproximação em ambientes de *image data warehousing*, pois gera a imersão do espaço métrico provido pelos vetores de

características em um espaço dimensional definido pelos representantes globais. Dessa maneira, as imagens que até então eram representadas em um espaço métrico, com a técnica Omni, são filtradas por dados de relação de ordem total (i.e., valores de distância aos representantes globais).

Tabela 5.2: Filtragem baseada na mbOr.

Cuboid resultante da filtragem realizada na etapa 1 (filtragem pelos predicados convencionais) do processamento de consultas IOLAP.

Os valores de distância da imagem de consulta sq aos representantes f_1, f_2 e f_3 são, respectivamente

$$df_1(s_q) = 43{,}28$$

$$df_2(s_q) = 30{,}48$$

$$df_3(s_q) = 33{,}54$$

IdImagem	Repr1	Repr2	Repr3	...
3	60,73	62,80	0	
4	15,65	42,01	51,16	
7	39,60	34,41	31,62	
9	52,95	63,66	10	
10	34,21	39,83	31,15	
12	66,22	33,24	38,01	
15	28,46	50,70	32,28	
16	45,22	26,17	37,22	
17	46,86	29,73	33,29	
18	53,23	17	46,1	
19	32,14	24,52	56,75	
20	46,14	21,93	41,34	

Filtragem de imagens baseada na mbOr. Neste exemplo, a mbOr(s_q, r_q) foi definida pelos intervalos:

$$I_1 = [33{,}28;\ 53{,}28]$$

$$I_2 = [20{,}48;\ 40{,}48]$$

$$I_3 = [23{,}54;\ 43{,}54]$$

onde $r_q = 10$

Logo, as imagens condizentes concomitantemente aos intervalos I_1, I_2 e I_3 são filtradas e as demais (tuplas tachadas) são desconsideradas.

IdImagem	Repr1	Repr2	Repr3	...
~~3~~	~~60,73~~	~~62,80~~	~~0~~	
~~4~~	~~15,65~~	~~42,01~~	~~51,16~~	
7	39,60	34,41	31,62	
~~9~~	~~52,95~~	~~63,66~~	~~10~~	
10	34,21	39,83	31,15	
~~12~~	~~66,22~~	~~33,24~~	~~38,01~~	
~~15~~	~~28,46~~	~~50,70~~	~~32,28~~	
16	45,22	26,17	37,22	
17	46,86	29,73	33,29	
~~18~~	~~53,23~~	~~17~~	~~46,1~~	
~~19~~	~~32,14~~	~~24,52~~	~~56,75~~	
20	46,14	21,93	41,34	

Como resultado da etapa 4 (etapa de filtragem por mbOr), o conjunto de imagens é reduzido mais uma vez. Caso o resultado desta seleção seja um conjunto vazio, a etapa 5 (i.e. etapa de refinamento) não é executada e a etapa 6 (i.e. etapa de resposta) apresenta o valor zero como resultado desta consulta, pois não há imagens na mbOr. Caso contrário, a etapa 5 é realizada, pois o subconjunto de

imagens selecionadas pela etapa 4 pode conter falsos positivo, ou seja, além de haver imagens realmente similares à imagem de consulta segundo o raio de abrangência r_q, o subconjunto de mbOr também pode possuir imagens que não são similares à imagem de consulta segundo o raio r_q.

Na etapa 5 (etapa de refinamento) do processamento da consulta IOLAP ocorre o cálculo da similaridade entre a imagem de consulta s_q e as imagens resultantes da etapa 4. O cálculo da similaridade entre as imagens é realizado com base nos valores de distância obtidos a partir dos vetores de características dessas imagens e com base no algoritmo de consulta por abrangência com raio igual a r_q.

Por fim, a etapa 6 (etapa de resposta) tem como objetivo contar o número de imagens resultantes da etapa 5, e apresentar o valor resultante da contagem.

5.5 Experimentos com o iCube

As vantagens do iCube sobre a atual tecnologia de *data warehousing* foram analisadas em testes de desempenho usando imagens reais concedidas pelo KDD Cup 2008 (http://www.kddcuo2008.com). Este conjunto de imagens é composto por 102.240 imagens de câncer de mama, representadas em um vetor 177 características.

O IDW do iCube foi alimentado tanto com dados reais como com dados sintéticos. A dimensão *Imagem* (Figura 5.2) foi alimentada com dados sintéticos sobre o exame (e.g. tipo de corte, uso de contraste), e com dados reais a partir dos vetores de características e dos valores de distância aos representantes (dados obtidos pelo processo de ETL sobre as imagens do KDD cup).

Dados reais sobre o Sistema de Saúde Brasileiro, obtidos pelo site www2.datasus.gov.br/datasus/índex.php, foram utilizados para alimentar a dimensão Hospital e dados sintéticos foram utilizados para alimentar as demais tabelas da Figura 5.2. A geração de dados sintéticos foi realizada com base em uma distribuição uniforme dos dados. Por exemplo, a tabela de dimensão "*Idade"* foi alimentada com 121 valores de idade, de zero anos a 120 anos. Estes valores de idade eram relacionados às imagens pela composição da tabela de fatos conforme a cardinalidade da tabela de dimensão "*Idade"*, ou seja, a primeira imagem era

relacionada à primeira tupla da tabela "*Idade"* e assim sucessivamente. Como o número de imagens é maior que o número de tuplas em "*Idade"*, após o relacionamento com a última tupla da tabela "*Idade"*, o relacionamento seguinte foi composto pela primeira tupla de "*Idade"*, reiniciando a sequência.

Os experimentos discutidos nessa seção foram realizados com base no processamento da consulta QE2 que é uma adaptação da consulta exemplo QE: "*Dada uma imagem de consulta s_q, quantas imagens são similares a s_q segundo um critério de abrangência r_q e que concomitantemente foram geradas por exames nos anos 1993 e 1994, por hospitais da macrorregião da Grande São Paulo, em pacientes do estado de São Paulo e com idade entre 30 a 40 anos?*".

Para a execução dos experimentos, foram selecionadas aleatoriamente 500 imagens do conjunto KDD Cup 2008, as quais foram submetidas a função de distância Euclidiana (L_2) com o raio de abrangência r_q variando entre 31% a 40% sobre a metade do diâmetro do conjunto.

Os experimentos foram realizados comparando duas configurações. A configuração CONF01 usou o iCube para processar consultas IOLAP e a configuração CONF02 utilizou a atual tecnologia de data warehousing. A CONF02 consiste na implementação da DWOnion (Seção 4.2), que é ambiente de data warehousing composto por um esquema estrela e pelo MAM Onion-Tree. O processamento das consultas IOLAP utiliza a Onion-Tree para identificar as imagens similares a s_q ($Lista_1$) e concomitantemente um esquema estrela convencional para analisar os predicados convencionais ($Lista_2$), como UF = São Paulo. O resultado da consulta IOLAP é obtido como a intersecção das respostas presentes em $Lista_1$ com a $Lista_2$. Isso corresponde ao método de seleção baseado na interseção dos resultados descrito em Seção 4.2.1.

No que diz respeito às configurações de *hardware* e *software*, os experimentos foram executados em um computador com Processador Intel Core 2 Duo, 4 GB de memória RAM, com um disco rígido SATA 250 GB 7200 RPM, sistema operacional Windows Vista, com uso do PostgreSQl 8.4.4. como Sistema Gerenciador de Banco de Dados e a IDE Borland C++ Builder 6 para compilar e executar os ambientes de IDWing.

Nestes experimentos, foram analisados o custo de construção da estrutura e o custo de processamento de consultas IOLAP. A medida de desempenho para ambos os testes foi o tempo total gasto na execução dessas operações.

5.5.1 Resultados para a construção da estrutura

O experimento para avaliar o tempo de construção da estrutura foi realizado com a intenção de avaliar o tempo despendido para a adaptação da técnica Omni ao ambiente de *data warehousing* versus o tempo de construção da estrutura de indexação Onion-Tree.

O tempo decorrido sobre a configuração DWOnion (CONF02 Figura 5.4) foi obtido a partir da soma do tempo gasto para a construção e alimentação do DW, mais o tempo de construção da Onion-Tree. Já o tempo de construção decorrido sobre a configuração iCube (CONF01) consiste na soma do tempo gasto na identificação dos representantes globais, mais o tempo gasto no cálculo da distância de cada imagem do IDW aos representantes, mais o tempo de construção e alimentação do IDW.

Os resultados de desempenho das configurações CONF01 e CONF02 sobre a construção das estruturas estão ilustrados na Figura 5.4. Estes resultados não incluem o tempo decorrido para extrair os vetores de características das imagens, uma vez que o KDD Cup 2008 já forneceu estes vetores.

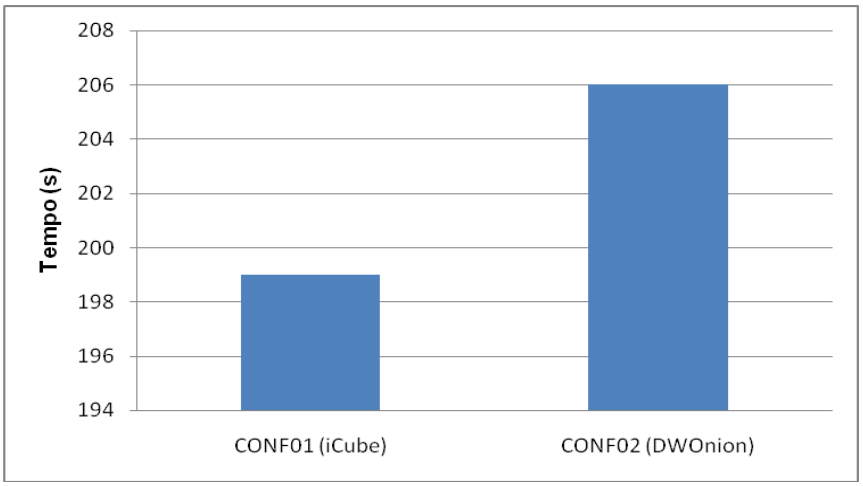

Figura 5.4: Gráfico com resultado sobre o desempenho de construção das estruturas.

Como ilustrado na Figura 5.4, a configuração DWOnion (CONF02) utilizou em média 206 segundos para ser construída, enquanto o iCube, representado por CONF01, apresentou um ganho de 3.53% ao ser construído em 199 segundos. Portanto, podemos concluir que o custo referente à identificação dos representantes

e o tempo despendido para calcular os valores de distância aos representantes não prejudicam o desempenho do iCube para a construção da estrutura.

5.5.2 Resultados para o processamento de consultas IOLAP

Experimentos sobre o processamento de consultas IOLAP foram realizados com o intuito de avaliar qual configuração apresenta melhores resultados no processamento de consultas IOLAP. Lembrando que consultas IOLAP são consultas que contem predicados convencionais e predicados visuais sobre a similaridade entre as imagens.

Na Figura 5.5, estão ilustrados os resultados sobre o tempo decorrido no processamento de consultas IOLAP. A abscissa refere-se à média do tempo decorrido em segundos no processamento de 500 consultas variando a imagem de consulta s_q, enquanto a ordenada refere-se à variação do raio de abrangência r_q.

Como observado no gráfico, o iCube apresentou um expressivo ganho no desempenho do processamento de consultas IOLAP. O custo de processamento de consultas IOLAP com o iCube apresentou-se praticamente invariante à seletividade estabelecida pelo raio de abrangência, enquanto o desempenho da configuração DWOnion é degradado rapidamente conforme o raio de abrangência é aumentado. Para todos os valores de raio de abrangência, o iCube reduziu o tempo de processamento das consultas IOLAP em mais de 40%, e com o aumento do raio os ganhos do iCube ficam ainda maiores. Como observado na Figura 5.5, com raio 31% o iCube teve um ganho de 43% e com o aumento do raio para 40% teve um ganho de 76,70%. Devido aos excelentes resultados pode-se concluir que o ambiente iCube é aplicação viável e aconselhável de *image data warehousing*.

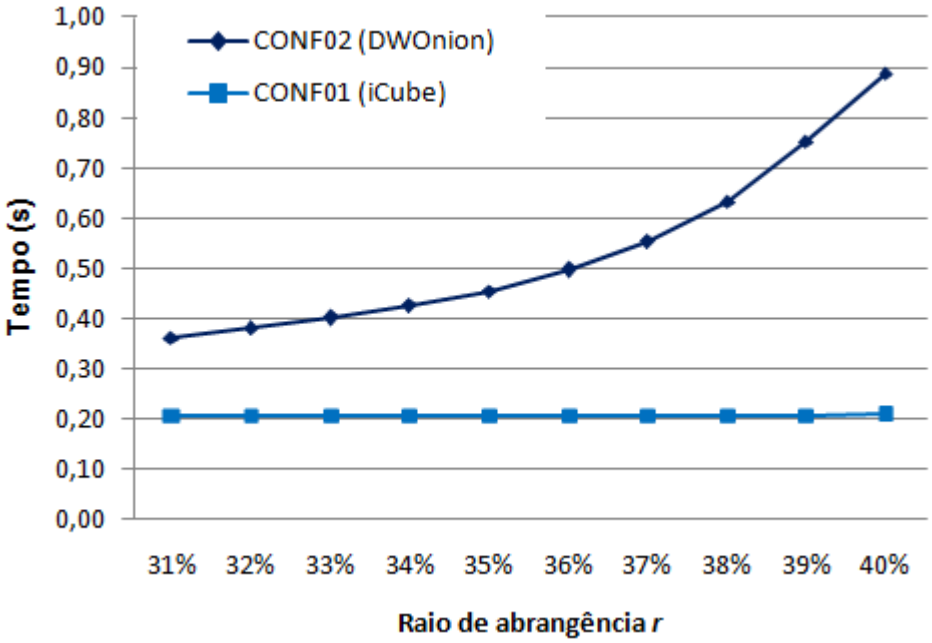

Figura 5.5. Tempo decorrido para o processamento de consultas IOLAP com as configurações iCube e DWOnion.

5.6 Considerações Finais

A proposta do iCube incorpora uma nova funcionalidade a ambientes de *data warehousing*, ao permitir que sejam realizados levantamentos de imagens baseados em similaridade, aproveitando-se de todas as facilidades oferecidas por esses ambientes. Como resultado, atividades de análise e gestão de recursos podem ser realizadas de maneira ágil e organizadas, além de utilizar imagens para tomada de decisão estratégica.

O presente trabalho propôs um esquema estrela diferenciado, que possui uma dimensão dedicada ao armazenamento de dados sobre o conteúdo intrínseco das imagens, permitindo assim que consultas complexas baseados em similaridade entre imagens sejam executadas. O trabalho também apresentou uma extensão do processo de ETL para a transformação de imagens com base no esquema proposto, assim como foi apresentado uma estratégia para o processamento de consultas IOLAP. O processamento ocorre em no máximo seis etapas, as quais foram elaboradas visando a otimização do processamento das consultas, uma vez que apresentam mecanismos de eliminação de comparações desnecessárias.

A proposta iCube foi validada a partir de testes experimentais que compararam seu desempenho à atual tecnologia de *IDWing*. Os resultados indicaram que o iCube requer um pequeno tempo de resposta no processamento de consultas IOLAP.

O objetivo do iCube é tornar viável o desenvolvimento e a implantação de aplicações de *image data warehousing* permitindo que consultas IOLAP sejam realizadas em um ambiente multidimensional e de alto desempenho. Neste contexto, a proposta do iCube aqui descrita no Capítulo 5 foi estendida de modo a ampliar a propriedade de multidimensionalidade sobre os predicados visuais, ou seja, para permitir que o ambiente disponibilize para o usuário diversas perspectivas das imagens sobre diferentes descritores. O capítulo a seguir descreve a extensão do ambiente iCube para múltiplos descritores.

Capítulo 6

iSTAR

Neste capítulo, é descrita a proposta iStar, que é um esquema estrela composto por camadas perceptuais. Neste capítulo também são apresentados experimentos que avaliam esta proposta.

6.1 Considerações Iniciais

Como descrito na Seção 2.2, um dos grandes desafios sobre sistemas de recuperação de imagens baseado em conteúdo é a redução do *gap* semântico, ou seja, aproximar o resultado gerado pelo sistema computacional ao resultado esperado pelo usuário. O *gap* semântico ocorre pois diferentes usuários, ou o mesmo usuário em diferentes intenções de consulta, podem optar por descritores visuais diferentes para determinar a similaridade entre as imagens. Uma maneira simples para reduzir essa diferença consiste na representação do conteúdo intrínseco das imagens por múltiplos descritores, permitindo que o usuário escolha quais descritores lhe é mais adequado para uma determinada intenção de consulta (PONCIANO-SILVA et al., 2009).

Neste contexto, este projeto realizou novas investigações no contexto de *image data warehousing* a fim de permitir que múltiplos descritores sejam disponibilizados no ambiente iCube. Para tanto, foram investigados modelos lógicos com o intuito de identificar o esquema estrela que tenha o melhor desempenho no processamento de consultas IOLAP sobre diversas composições de predicado, tanto sobre os predicados convencionais quanto sobre os múltiplos predicados visuais.

Como discutido por Kimball em (KIMBALL; ROSS, 2002), o desempenho de uma aplicação de DWing pode ser influenciado segundo as estratégias de definição de esquema do DW:

- Redundância dos dados no DW *versus* o custo de junções-estrela;
- Granularidade definida pelas tabelas de dimensão; e
- Medidas e atributos da tabela de fatos.

Com base nesses argumentos, este trabalho propôs oito configurações de esquemas estrela compostos por camadas perceptuais. Estas configurações foram avaliadas de modo a investigar o impacto dessas estratégias no desempenho do processamento de consultas IOLAP em um ambiente de IDWing e identificar qual é o esquema mais adequado para o desenvolvimento de um iCube. Vale ressaltar que cada uma das oito configurações de esquema estrela propostas neste trabalho possuem as seguintes contribuições:

- Introduz um esquema estrela adaptado a múltiplos descritores. Isto torna o ambiente ainda mais flexível que a proposta anterior descrita no capítulo 5, por ampliar as possibilidades de uso de vários predicados visuais e reduzir o *gap* semântico;
- Mantém o suporte a consultas IOLAP em que os critérios de similaridade são definidos de forma ad-hoc; ou seja, a imagem de consulta e o critério de abrangência são desconhecidos durante a construção e alimentação do IDW;
- Mantém o uso da técnica Omni para melhorar o desempenho no processamento de consultas por similaridade; e
- Estende as vantagens de multidimensionalidade de um IDW aos predicados visuais. Consequentemente, permite o uso de múltiplos descritores em uma mesma consulta IOLAP, como por exemplo:
    - *"Quantas imagens são similares a uma imagem consulta s_q segundo os descritores de forma e textura sobre um raio de abrangência $r_{q;}$ e que concomitantemente ocorreram no ano de 2010 na cidade de Ribeirão Preto?";*

6.2 Camadas Perceptuais

O iStar se diferencia do esquema estrela proposto no capítulo anterior, pois permite que as imagens do IDWing sejam representadas por múltiplos descritores. Esta representação é feita de maneira organizada com a definição de camadas perceptuais. Compreende-se como uma camada perceptual o conjunto de dados relacionados à representação do conteúdo intrínseco de uma imagem segundo um determinado descritor. Por exemplo, ao processar um conjunto de imagens com os descritor de momentos de Zernike, de Haralick e com o histograma de níveis de cinza, é possível analisar esse conjunto de imagens segundo três camadas perceptuais, uma para cada descritor utilizado. Estas camadas podem ser usadas separada ou conjuntamente. Logo, em um iStar com esta composição é possível analisar esse conjunto de imagens sob a perspectiva da camada perceptual de *textura* gerada pelos descritores de Haralick, sob o ponto de vista da camada perceptual de *forma* gerada pelos momentos de Zernike, sob a camada perceptual de *cor* gerada pelo histograma de níveis de cinza, ou sob a combinação dessas camadas perceptuais (e.g., sob a forma e textura das imagens).

Com o advento das camadas perceptuais, o iStar estende os conceitos de multidimensionalidade à perspectiva do conteúdo intrínseco das imagens, disponibilizando o acesso a diferentes níveis perceptuais, além de ser uma alternativa organizada para a redução do *gap* semântico em ambiente de IDWing.

6.3 Características dos Dados do Estudo de Caso

Para facilitar o entendimento das configurações, as figuras dos esquemas estrela apresentadas nas seções a seguir são modelos para um IDW contendo os seguintes assuntos:

Imagens descritas segundo os dados convencionais:

- Data, com os atributos: dia, dia da semana, mês, ano, ano e mês em número, ano e mês literal, dia numerado pela semana, dia numerado

pelo mês, dia numerado pelo ano, mês numerado pelo ano, semana numerada pela ano.

- Paciente, com os atributos: nome, data de nascimento, sexo, etnia, tipo sanguíneo, tipo Rh, nacionalidade, natalidade e o UF natal.
- Hospital, com os atributos: nome do hospital, sigla, município, macrorregião, microrregião, se o hospital é de ensino e pesquisa, qual a esfera administrativa, tipo de gestão, tipo de prestador.
- Exame, com os atributos: equipamento utilizado, técnico radiologista, médico radiologista, uso de contraste, corte, motivo da investigação/exame.
- Idade, com os atributos: idade, faixa etária e tipo.

Imagens foram descritas segundo as camadas perceptuais:

- Haralick, descrita por vetor de características e por *dois* representantes globais. A quantidade de representantes globais foi determinada pela dimensão fractal $\lceil D2 \rceil$ + 1 e os representantes escolhidos pelo algoritmo HF, conforme descritos na Seção 2.4.2. O mesmo se aplica para todas as camadas perceptuais.
- Wavelet, descrita por vetor de características e por *dois* representantes globais.
- Zernike, descrita por vetor de características e por *três* representantes globais.
- Histograma, descrita por vetor de características e por *três* representantes globais.

6.4 Esquemas Estrela Propostos EBR e EBM

Por ser uma extensão do esquema estrela proposto pelo iCube, as oito configurações baseadas em camadas perceptuais também são compostas por uma tabela de fatos, por *dimensões convencionais* e por *dimensões visuais*. A tabela de fatos destas configurações também possui o campo artificial “quantidade” para facilitar a formulação de consultas SQL envolvendo soma e agrupamento.

A investigação do melhor esquema estrela foi conduzida com a definição de dois grupos de esquemas: *esquemas baseados em representantes* e *esquemas baseados em mbOr*. Estes grupos diferem no nível de detalhamento que as *dimensões visuais* descrevem as imagens do IDW, sendo então investigado qual o melhor nível de granularidade para o iStar.

O grupo de *esquemas baseados em representantes* é caracterizado por possuir uma *dimensão visual* para cada representante e assim ter um nível mais granular sobre as *dimensões visuais*. Por outro lado, o grupo de *esquemas baseados em mbOr* tem um nível menos detalhado, determinado pela mbOr de uma camada perceptual. No grupo de *esquemas baseados em mbOr,* os dados relacionados aos representantes, isto é, os dados que compõem uma mbOr, são mantidos em uma única *dimensão visual*. Consequentemente, este grupo a priori produzirá um melhor desempenho no processamento de consultas IOLAP, pois os dados estão organizados de forma a reduzir o número de junções necessárias para compor uma mbOr. Lembrando que a mbOr é um mecanismo de aceleração de IOLAP que é proposta por meio da aplicação iCube. A seguir apresentamos a primeira configuração dos *esquemas baseados em representantes* (EBR) e a primeira configuração dos *esquemas baseados em mbOr* (EBM).

A primeira configuração proposta para o grupo de *esquemas baseados em representantes* (EBR) foi a configuração EBR1 (esquema baseado em representantes 1), que possui uma tabela de fatos, chamada *Imagem*, composta por $k+n+1$ atributos, onde:

- $k > 0$ e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- $n > 0$ e refere-se ao número de chaves estrangeiras de *dimensões visuais; e*
- O número *"1"* refere-se à medida "*quantidade".*

Sobre as demais tabelas do esquema, o EBR1 possui k dimensões convencionais, determinadas conforme o assunto do IDW, e n dimensões visuais, em que o valor n corresponde ao total de representantes, obtido pela Equação 6.1:

$$n = \sum_{i=1}^{v} Tr_i \qquad (6.1)$$

Em que $v \leq n$ e refere-se à quantidade de camadas perceptuais, enquanto Tr_i consiste no total de representantes identificados para a camada perceptual *i.* Na

Figura 6.1, é ilustrado o esquema estrela de um EBR1 para o cenário exemplo especificado anteriormente, o qual contém quatro camadas perceptuais, Wavelet, Haralick, Zernike e Histograma, que estão destacadas pelo retângulo acinzentado. Nesse cenário exemplo, o valor *n* é obtido pela soma do número de representantes dessas quatro camadas perceptuais, ou seja, n = 10 devido à soma de 2 representantes de Wavelet, 2 representantes de Haralick, 3 representantes de Zernike e 3 representantes de Histograma.

É importante ressaltar que a característica que diferencia o EBR1 dos demais esquemas no grupo de *esquemas baseados em representantes* (EBR) é a redundância de vetores de características, pois em cada uma das *n dimensões visuais* o vetor de características referente à sua respectiva camada perceptual é empregado como atributo em cada tabela de *dimensão visual* (e.g. WVetorCaract é o mesmo nas tabelas Wavelet Rep1 e Wavelet Rep2). Na Figura 6.1, esses atributos são identificados por WVetorCaract, HVetorCaract, ZVetorCaract e HiVetorCaract, correspondente respectivamente ao vetor de características da camada perceptual de Wavelet, Haralick, Zernike e Histograma.

Por meio do EBR1, este trabalho investigou o impacto da redundância dos vetores de características, que é um tipo de dado não convencional em ambiente de DWing, sobre o desempenho no processamento de consultas IOLAP. Como descrito por Kimball em (KIMBALL; ROSS, 2002), usualmente a redundância de dados em um DW gera menores tempos de resposta em consultas OLAP. Esses resultados se justificam pois a redundância dos dados é uma abordagem para reduzir o número de junções e consequentemente reduzir o custoso impacto desse tipo de operação. No entanto, por vetores de características serem dados não-convencionais, este consenso apresentado por Kimball pode não se aplicar.

Siqueira et. al abordam a complexidade do processamento de consultas SOLAP em (SIQUEIRA et al., 2009). Por meio desse trabalhos os autores puderam observar que o armazenamento redundante de dados não-convencionais, como os dados espaciais, pode ser mais custoso do que o custo das junções. De maneira semelhante, este trabalho visa observar qual tendência que dados de imagens seguem. Em especial, este trabalho visa investigar se há algum cenário ou composição de predicado em que o custo de armazenamento redundante de dados sobre imagens em tabelas de dimensões causa maior degradação do tempo de processamento de consultas IOLAP do que o custo de junções.

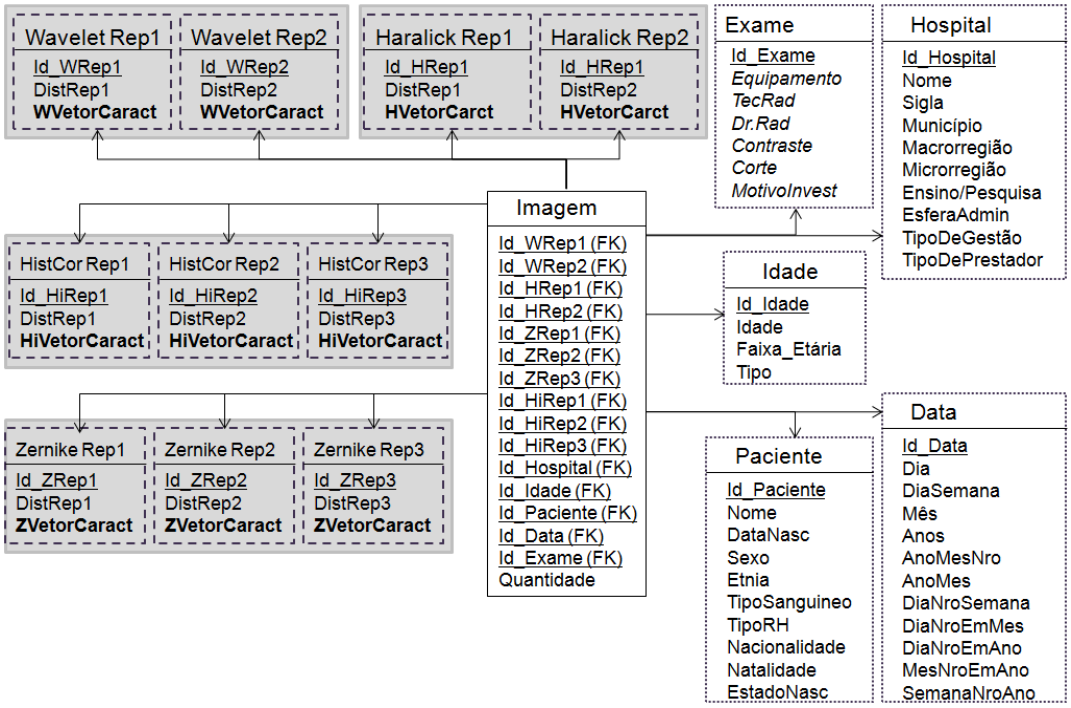

Figura 6.1: Esquema estrela baseado na configuração EBR1 para o cenário exemplo.

A primeira configuração proposta para o grupo de *esquemas baseados em mbOr* (EBM) foi a configuração EBM1 (esquema baseado em mbOr 1). Como dito anteriormente, este grupo de configuração foi modelado com o intuito de minimizar o número de junções entre as dimensões visuais necessárias para compor a mbOr.

Este esquema possui uma tabela de fatos, *Imagem*, composta por *k+v+1* atributos, onde:

- *k* > 0 e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- *v* > 0 e refere-se ao número de camadas perceptuais existentes nesse modelo, ou seja, ao número de descritores utilizados para extrair o conteúdo intrínseco das imagens;
- O número "*1*" refere a medida "*quantidade*".

Sobre as demais tabelas do modelo, o EBM1 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de *v dimensões visuais*. As *dimensões visuais* deste modelo são caracterizadas como tabelas horizontalmente longas por atribuírem em sua composição todos os dados referentes a uma camada perceptual. Dessa maneira, dados relacionados aos

valores de distância aos representantes, assim como o vetor de características gerado para essa camada perceptual, estão armazenados em uma mesma tabela.

Vale ressaltar que se espera que esta configuração proporcione um menor custo de junção, além de gerar um ganho físico, pois os vetores de características não são armazenados redundantemente nas dimensões visuais como ocorre na EBR1. Na Figura 6.2, é ilustrado o esquema estrela de um EBM1 para o cenário exemplo especificado anteriormente, o qual contém quatro camadas perceptuais, Wavelet, Haralick, Zernike e Histograma, que estão destacadas pelo retângulo acinzentado. Nesse cenário exemplo, o valor *v* é obtido pela soma do número de camadas perceptuais, ou seja, n = 4 devido à soma 1 (descritor de Wavelet), 1 (descritor de Haralick), 1 (descritor de Zernike) e 1 (descritor de Histograma).

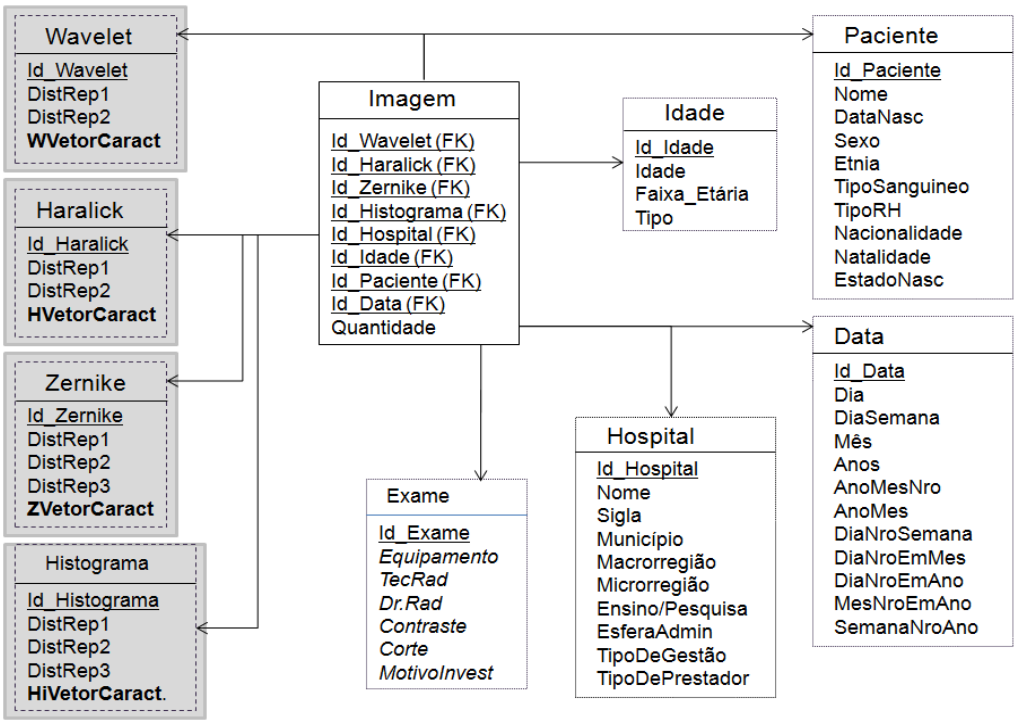

Figura 6.2: Esquema estrela baseado na configuração EBM1 para o cenário exemplo.

As configurações descritas a seguir são adaptações destes dois grupos de esquemas realizadas para investigar o impacto do desempenho no tempo de processamento de consultas IOLAP sob a redundância dos dados, sob o custo de junções-estrela e sob o custo da “degeneração” da tabela de fatos versus a eliminação de junções-estrela.

6.4.1 Proposta de Esquemas com Dimensões Dedicadas para o Armazenamento de Vetores de Características

O *Esquema baseado em representantes 2* (EBR2) consiste em uma adaptação do EBR1 com o intuito de eliminar a redundância sobre os atributos de vetor de características. Este trabalho abordou a eliminação deste tipo de redundância, pois vetores de características não são dados convencionais ao ambiente de DWing e podem impactar o desempenho do ambiente quando estes dados possuírem um conjunto muito grande de características (i.e., quando o vetor de características for muito longo). Neste contexto, o EBR2 foi proposto para comparar o custo de junção desta configuração versus o impacto gerado pela redundância dos vetores de características existente na configuração EBR1.

O esquema EBR2 proposto neste trabalho é caracterizado por possuir uma tabela auxiliar (i.e., uma tabela de dimensão visual adicional) para cada camada perceptual, que armazena o seu respectivo vetor de características. Na Figura 6.3 é ilustrado o EBR2 para o cenário exemplo. Nesta figura, as tabelas auxiliares são destacadas com o contorno *"ponto-ponto-traço"* e devidamente agrupadas em sua camada perceptual pelos retângulos acinzentados.

O EBR2 possui uma tabela de fatos, *Imagem*, composta por $k+n+v+1$ atributos, onde:

- $k > 0$ e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- $n > 0$ e refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes;
- $v > 0$ e refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam os vetores de características (i.e., tabelas auxiliares); e
- O número *"1"* refere-se à medida "*quantidade".*

Sobre as demais tabelas do esquema, o EBR2 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de $n+v$ *dimensões visuais*, onde n também corresponde ao total de representantes (obtido pela Equação 6.1), enquanto v refere-se ao número de *dimensões visuais* dedicadas ao armazenamento do vetor de características de sua respectiva camada perceptual.

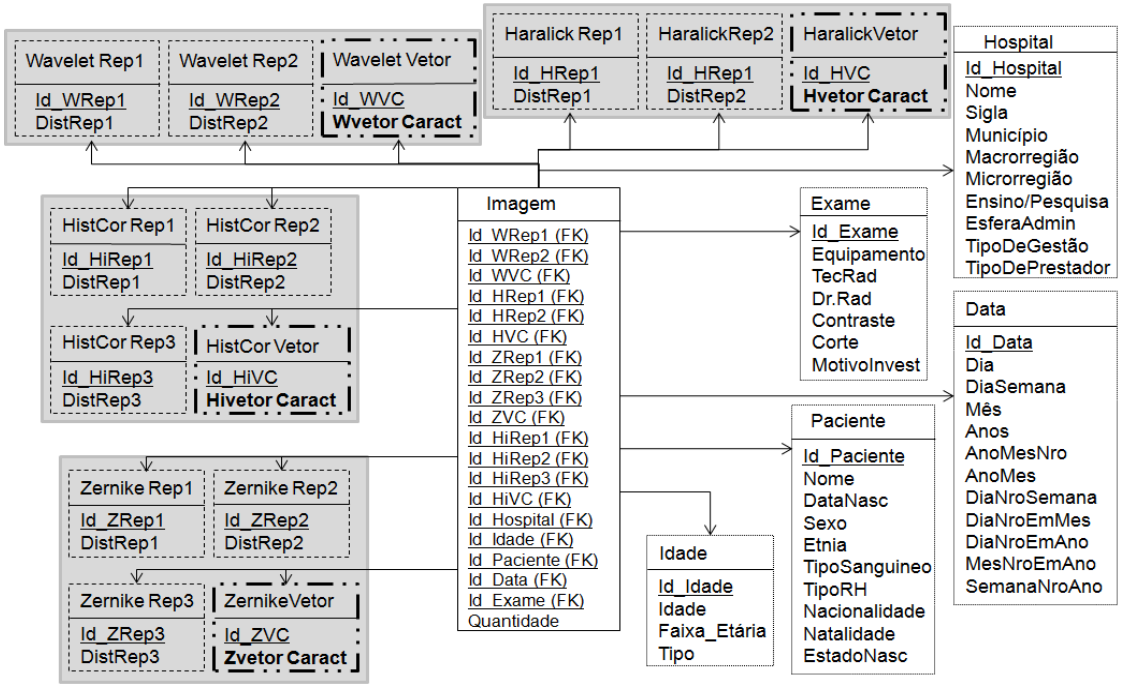

Figura 6.3: Esquema estrela baseado na configuração EBR2 para o cenário exemplo.

O *Esquema Baseado em mbOr 2* (EBM2) confere uma adaptação e investigação equivalente ao EBR2, pois também possui tabelas auxiliares para o armazenamento dos vetores de características. Logo, o principal conceito que diferencia EBM2 do EBM1 é a forma de armazenar os vetores de características. Esta diferença pode ser visualizada na Figura 6.4, em que o esquema estrela possui as tabelas auxiliares destacadas com o contorno *"ponto-ponto-traço"* e devidamente agrupadas à sua camada perceptual pelos retângulos acinzentados.

O EBM2 possui uma tabela de fatos, *Imagem*, composta por $k+v_1+v_2+1$ atributos, onde:

- $k > 0$ e refere-se ao número de chaves estrangeiras de *dimensões de dados convencionais*;
- $v_1 > 0$ refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes;
    - *O valor v_1 representa, portanto,* o número de camadas perceptuais existentes nesse esquema.
- $v_2 > 0$ refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam vetores de características;
    - *O valor v_2 representa, portanto,* o número de camadas perceptuais existentes nesse esquema; e

- O número "*1*" refere-se à medida *quantidade*.

Sobre as demais tabelas do esquema, o EBM2 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de *2*v dimensões visuais, onde v* é igual ao número de camadas perceptuais existentes nesse esquema. Como pode-se observar $v = v_1 = v_2$.

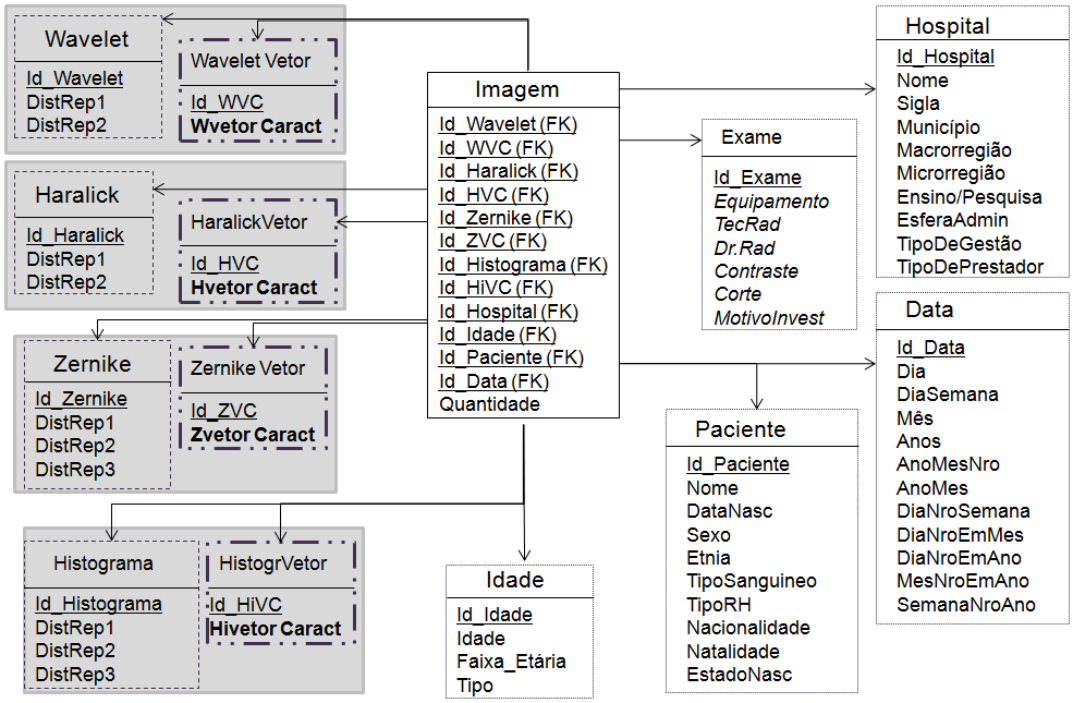

Figura 6.4: Esquema estrela baseado na configuração EBM2 para o cenário exemplo.

6.4.2 Proposta de Esquemas que Visam o Armazenamento Conjunto dos Vetores de Características

O *Esquema baseado em representantes 3* (EBR3) consiste em um esquema estrela adaptado para armazenar conjuntamente todos os vetores de características de todas as camadas visuais. Dessa maneira, este esquema diferencia-se dos demais esquemas EBR por possuir apenas uma tabela auxiliar dedicada ao armazenamento de todos os vetores de características. Esta tabela auxiliar é ilustrada na Figura 6.5, destacada pelo contorno *"ponto-ponto-traço"*.

O *esquema baseado em representante 3* possui uma tabela de fatos, *Imagem*, composta por *k+n+2* atributos, onde:

- *k* > 0 e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- *n* > 0 e refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes; e
- O número *"2"* refere-se à *1* medida *quantidade* e *1* à chave estrangeira da *tabela auxiliar* que armazena todos os vetores de características.

Sobre as demais tabelas do esquema, o EBR3 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de *n+1 dimensões visuais*. No EBR3 a quantidade *n* também corresponde ao total de representantes, podendo ser calculada pela Equação 6.1, e o valor 1 refere-se à tabela auxiliar que armazena conjuntamente todos os vetores de características de todas as camadas visuais.

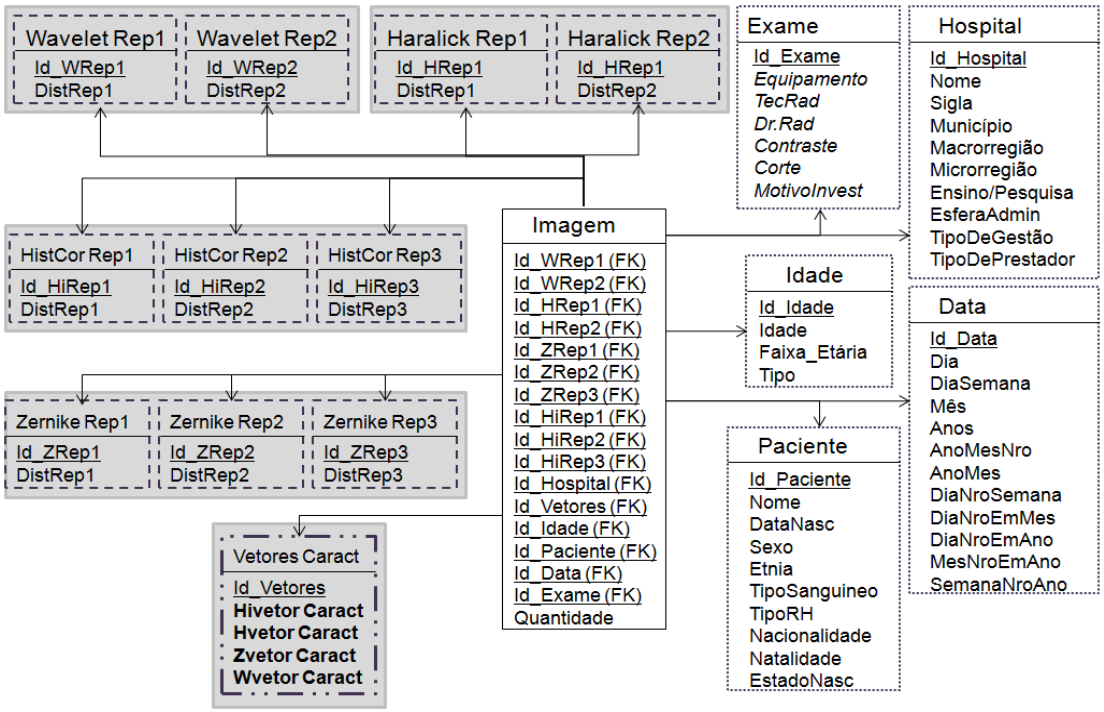

Figura 6.5: Esquema estrela baseado na configuração EBR3 para o cenário exemplo.

O *Esquema baseado em mbOr 3* (EBM3) é uma equivalência do EBR3 para o grupo de *esquemas baseados em descritores*, pois possui apenas uma tabela auxiliar dedicada ao armazenamento de todos os vetores de características. A tabela auxiliar proposta nesta configuração é ilustrada na Figura 6.6, destacada pelo contorno *"ponto-ponto-traço"*.

O *esquema baseado em mbOr 3* possui uma tabela de fatos, *Imagem*, composta por *k+v+2* atributos, onde:

- *k* > 0 e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- *v* > 0 e refere-se o número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes;
    - O valor v representa, portanto, o número de camadas perceptuais existentes nesse esquema;
- O número *"2"* refere-se à *1* medida quantidade e *1 à* chave estrangeira da *dimensão visual* que armazena todos os vetores de características.

Sobre as demais tabelas do modelo, o EBM3 possui *k dimensões convencionais*, determinadas conforme o assunto do DW, e um total de *v+1 dimensões visuais*, onde a quantidade *v* corresponde ao número total de camadas perceptuais e o valor 1 refere-se à tabela auxiliar que armazena todos os vetores de características.

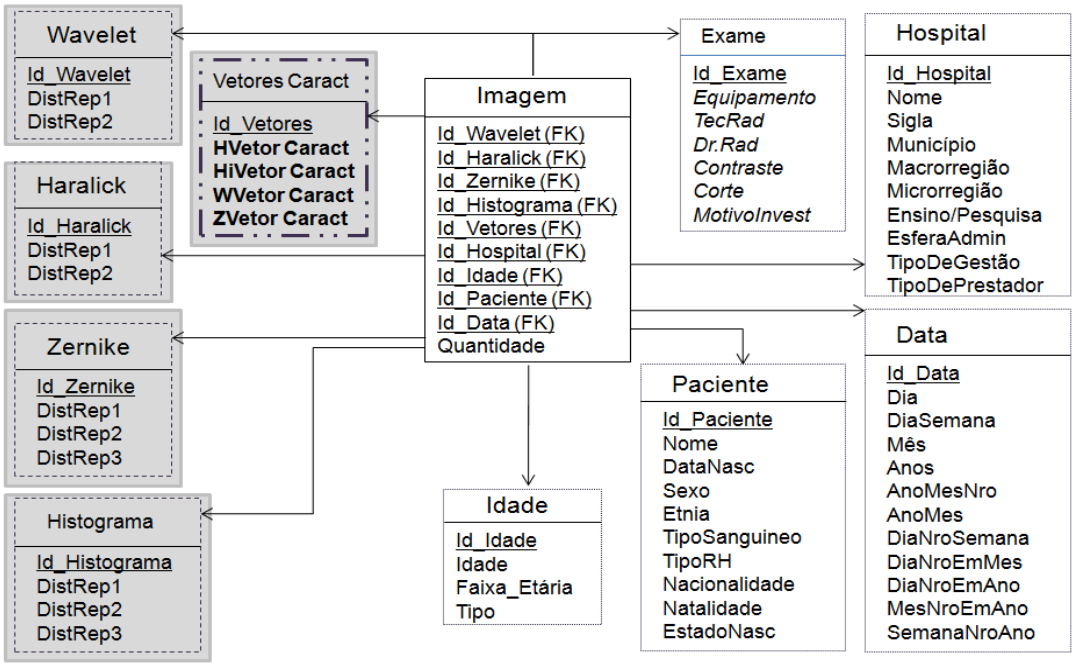

Figura 6.6: Esquema estrela baseado na configuração EBM3 para o cenário exemplo.

6.4.3 Proposta de Esquemas que Armazenam os Vetores de Características como Medidas na Tabela de Fatos

O último esquema proposto para o grupo de *esquemas baseados em representantes* por este trabalho foi o EBR4 (*Esquema Baseado em Representantes 4*), que difere dos demais esquemas EBR por armazenar todos os vetores de características na tabela de fatos e, portanto, eliminar o armazenamento destes vetores em tabelas de dimensão e consequentemente eliminar a necessidade de realizar junções para o acesso aos vetores de características. Dessa maneira, além da tabela de fatos ser verticalmente extensa, essa também pode se tornar extensa horizontalmente, conforme a dimensionalidade dos vetores de características.

Na Figura 6.7 é ilustrado o EBR4, sendo que os vetores de características nesse esquema, são destacados pelo retângulo acinzentado na tabela de fatos. Esta configuração possui uma tabela de fatos, *Imagem*, composta por $k+n+1+v$ atributos, onde:

- $k > 0$ e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- $n > 0$ e refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes;
- O número "*1*" refere-se à medida *quantidade* e
- $v \leq n$ e refere-se ao número de camadas perceptuais.

Sobre as demais tabelas do modelo, o EBR4 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de *n dimensões visuais,* que corresponde ao total de representantes, sendo que o valor de *n* pode ser calculado pela Equação 6.1.

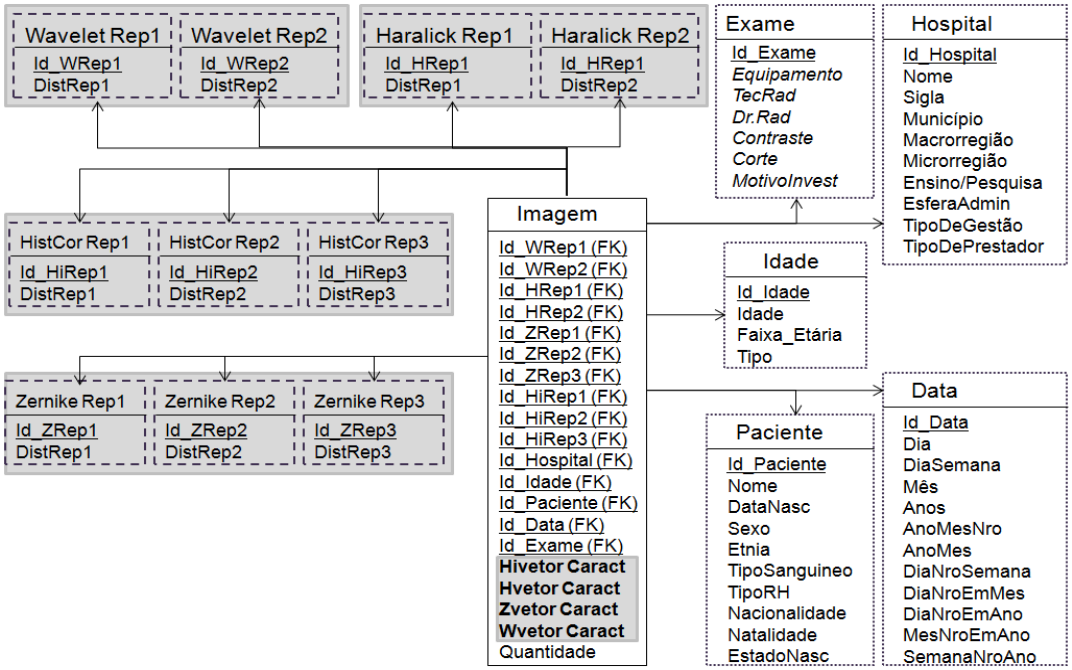

Figura 6.7: Esquema estrela baseado na configuração EBR4 para o cenário exemplo.

O *Esquema baseado em mbOr 4* (EBM4) consiste em uma adaptação do EBM3 com o intuito de reduzir o número de junções-estrela necessárias para a obtenção dos vetores de características. Este esquema diferencia-se dos demais esquemas por estender a tabela de fatos com o armazenamento de todos os vetores de características nessa tabela. Assim como no EBR4, a tabela de fatos, além de ser verticalmente extensa, também pode se tornar extensa horizontalmente. Na Figura 6.8, é ilustrado o EBM4 sendo que os vetores de características nesse esquema são destacados pelo retângulo acinzentado na tabela de fatos.

O *esquema baseado em mbOr 4* possui uma tabela de fatos, *Imagem*, composta por $k+v_1+1+v_2$ atributos, onde:

- $k > 0$ e refere-se ao número de chaves estrangeiras de *dimensões convencionais*;
- $v_1 > 0$ e refere-se ao número de chaves estrangeiras de *dimensões visuais* que armazenam dados sobre os representantes;
    - *O valor v_1 representa, portanto,* o número de camadas perceptuais existentes nesse esquema.
- O número "*1*" refere-se à medida *quantidade;* e
- $v_2 > 0$ e refere-se aos atributos de vetores de características armazenados na tabela de fatos;

- *O valor v_2 representa, portanto,* o número de camadas perceptuais existentes nesse esquema.

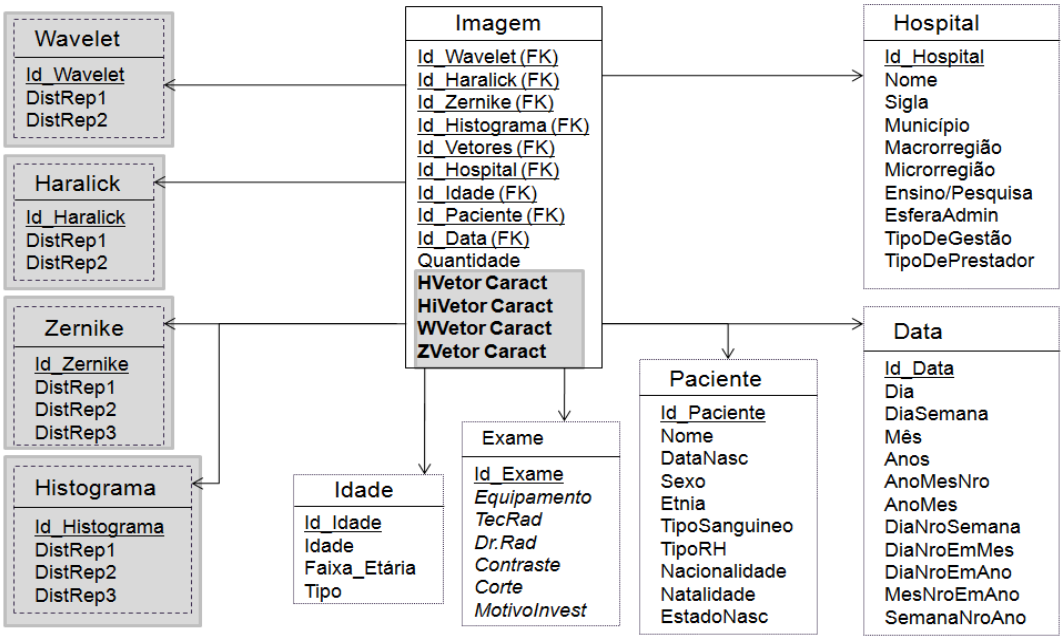

Figura 6.8: Esquema estrela baseado na configuração EBM4 para o cenário exemplo.

Sobre as demais tabelas do modelo, o EBM4 possui *k dimensões convencionais*, determinadas conforme o assunto do IDW, e um total de *v dimensões visuais* que representam cada uma das camadas perceptuais composta por dados sobre os valores de distantes a seus representantes.

6.5 Testes de Desempenho

Os testes de desempenho descritos nesta seção possuem como objetivo investigar quais das duas abordagens (EBR *versus* EBM) proporciona um melhor desempenho no tempo de processamento de consultas IOLAP e também dentre os esquemas propostos em cada abordagem, qual o esquema mais eficiente com relação ao tempo de processamento. As melhores abordagens também são comparadas aos atuais recursos de IDWing.

Os experimentos foram conduzidos em um computador executando o sistema operacional Linux Ubuntu 9.04 (32 bits) com processador Intel Core i7 2.67 GHz, 12 GB de memória primária e 2 TB de memória secundária. A linguagem de

programação C++ foi escolhida para a implementação dos ambientes e o PostgreSQL 8.3 foi selecionado como SGBD por ser um software livre reconhecido e eficiente.

Os esquemas propostos foram testados contra os recursos disponíveis atualmente nos ambientes de DWing para imagens. Dessa forma, um total de treze configurações de IDWing foram avaliadas, em que cinco destas configurações são referentes aos atuais recursos de IDWing disponíveis (i.e. trabalhos correlatos): DWOnion com método de consulta baseado na interseção (DWOM1); DWOnion com método de consulta iniciado pela Onion-Tree (DWOM2); DWOnion com método de consulta iniciado pelo DW (DWOM3); SingleOnion; MultiOnion. As demais oito configurações referem-se às propostas de esquema estrela apresentadas neste trabalho: iStar-EBR1; iStar-EBR2; iStar-EBR3; iStar-EBR4; iStar-EBM1; iStar-EBM2; iStar-EBM3; e iStar-EBM4.

Seguindo o cenário exemplo descrito no início deste capítulo, em que os ambientes de IDWing avaliados devem prover suporte a múltiplas camadas perceptuais, cada configuração estendida da Onion-Tree (DWOM1, DWOM2, DWOM3, SingleOnion e MultiOnion) foi implementada pela combinação de múltiplas estruturas, onde cada estrutura refere-se à uma camada perceptual. Por exemplo, a implementação da SingleOnion para as camadas perceptuais de Zernike, Histograma, Haralick-Variância, Haralick-Entropia e Haralick-Uniformidade foi realizada com a construção de cinco estruturas SingleOnion, em que cada uma das estruturas era responsável pela indexação de uma camada perceptual.

Para povoar o IDW foram utilizadas 131.656 imagens de ressonância magnética disponibilizadas pelo Grupo de Banco de Dados e Imagens (GBdi) da USP - São Carlos. Estas imagens foram processadas pelos descritores de Uniformidade de Haralick (média dos ângulos), Variância de Haralick (média dos ângulos), Entropia de Haralick (média dos ângulos), pelo extrator de Histograma de níveis de cinza e momentos de Zernike na ordem 10. A partir desse processamento foram identificados usando a técnica HF (Seção 2.4.2) 3 representantes para a camada perceptual Haralick-Uniformidade, 4 representantes para Haralick-Variância, 4 representantes para Haralick-Entropia, 5 representantes para Zernike e 2 representantes para o Histograma.

A dimensão *Hospital* foi alimentada com dados reais sobre o sistema de saúde Brasileiro, obtidos pelo site www2.datasus.gov.br/datasus/índex.php. Para a

dimensão *Data* foram utilizados dados do benchmark TPC-H. As demais tabelas convencionais foram alimentadas como dados sintéticos com distribuição uniforme conforme a cardinalidade das tabelas de dimensão.

Para avaliar o desempenho do processamento de consultas IOLAP, as configurações foram submetidas a até 4 repetições na submissão de consultas. Nestes experimentos, as imagens de consulta foram definidas aleatoriamente e o tempo médio decorrido desse processamento foi calculado. Esta avaliação foi realizada em três baterias de testes, especificados conforme a seletividade de predicados convencionais, a seletividade gerada pelo raio de abrangência, e a dimensionalidade do vetor de características.

6.5.1 Experimentos baseados na Seletividade dos Predicados Convencionais

Nestes experimentos foram projetados para explorar a multidimensionalidade do IDW, supondo uma situação extrema em que o usuário deseja utilizar todas as camadas perceptuais como critério de similaridade entre as imagens. Neste contexto, estes experimentos serão realizados para avaliar o desempenho das configurações propostas e correlatas segundo a característica de multidimensionalidade (i.e., a análise da imagem segundo diversos predicados). Para tanto, seis tipos de consultas foram determinados, as quais gradativamente submetiam a configuração testada a maiores esforços de processamento.

Estes experimentos foram conduzidos eliminando progressivamente os predicados convencionais mais seletivos até a ausência total de predicado convencional (i.e., condições sobre os dados pertencentes às dimensões convencionais). A Tabela 6.1 detalha a seletividade de cada predicado utilizado nas consultas IOLAP, o qual consiste no número de registros na tabela de fatos que atendem ao predicado.

Para esta bateria de testes, o raio de abrangência foi fixado no valor de 30% do raio de magnitude do conjunto, 10 imagens aleatórias foram submetidas como imagem de consulta s_q e o predicado visual foi composto pela condição de similaridade à imagem de consulta s_q segundo as cinco camadas visuais, isto é, o predicado visual selecionava as imagens que são similares à s_q com relação à representação de Haralick Uniformidade, de Haralick Variância, de Haralick Entropia,

de Histograma de níveis de cinza e Zernike. Vale ressaltar que a extração do conteúdo intrínseco dessas imagens de consulta foi realizada previamente de maneira a não ser contabilizado no tempo decorrido do experimento.

Tabela 6.1: Seletividade do conjunto determinada pelos predicados convencionais Macrorregião, Razão da Investigação, Idade, UF e Ano.

Dimensão	Predicado convencional	Registros selecionados
Hospital	Macrorregião = “Grande São Paulo”	7.995
Exame	Razão da Investigação = “suspeita de tumor”	26.331
Idade	Idade entre 0 a 30	33.736
Paciente	UF = “SP”	58.516
Data	Ano entre 1992 e 1993	75.315

Considerando as especificações desta bateria de testes e considerando a seletividade de cada predicado convencional, o predicado “*Macrorregião”* foi o primeiro predicado eliminado (Consulta *4Conv*) seguido por *“Razão da investigação”* (Consulta *3Conv*), *“Idade”* (Consulta *2Conv*), “*UF”* (Consulta *1Conv*) e por fim o predicado *“Ano”* (Consulta *0Conv*).

As consultas submetidas são adaptações da consulta QE a múltiplos predicados visuais, as quais são apresentadas a seguir:

- Consulta *5Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas no hospital da macrorregião da Grande São Paulo, nos anos de 1992 e 1993, referentes a pacientes com suspeita de tumor, do estado de São Paulo e com idade entre 0 a 30 anos? ”*.
- Consulta *4Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas nos anos de 1992 e 1993, referentes a pacientes com suspeita de tumor, do estado de São Paulo e com idade entre 0 a 30 anos? ”*.
- Consulta *3Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas nos anos de*

1992 e 1993, referentes a pacientes do estado de São Paulo e com idade entre 0 a 30 anos? ” .

- Consulta *2Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas nos anos de 1992 e 1993 e referentes a pacientes do estado de São Paulo?* ” .
- Consulta *1Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas nos anos de 1992 e 1993?* ” .
- Consulta *0Conv*: “*Quantas imagens são similares à imagem de consulta* s_q *segundo um raio de abrangência de 30% nas cinco camadas perceptuais?* ”.

6.5.2 Resultados dos Experimentos baseados na Seletividade dos Predicados Convencionais

Nesta seção, são apresentados e discutidos os experimentos baseados na seletividade dos predicados convencionais. Nas Figuras 6.9 e 6.10 são apresentados os tempos médios para o processamento das consultas 5Conv, 4Conv, 3Conv, 2Conv, 1Conv e 0Conv sobre as configurações propostas para o iStar (Seção 6.4).

Por meio dos resultados apresentados nesta tabela foi possível observar interessantes consequências da estratégia de esquema, segundo esta composição de predicado. Como destacada nas Figuras 6.9 e 6.10, a configuração EBM3, que é composta por apenas uma tabela auxiliar para o armazenamento de todos os vetores de características e por uma dimensão visual com todos os valores de distância aos representantes para cada camada visual, apresentou no geral menor tempo de resposta do que as demais configurações. Esta é uma interessante análise, pois foge do esperado, que determina que o custo de junção degrada o desempenho no processamento de consultas OLAP.

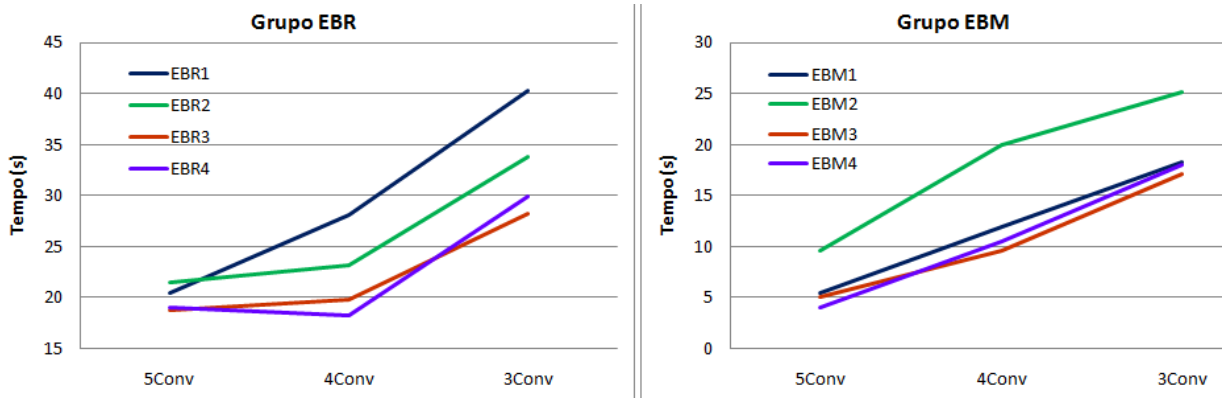

Figura 6.9: Tempo médio decorrido em segundos para os experimentos 5Conv, 4Conv e 3Conv sobre as propostas de extensão do iCube.

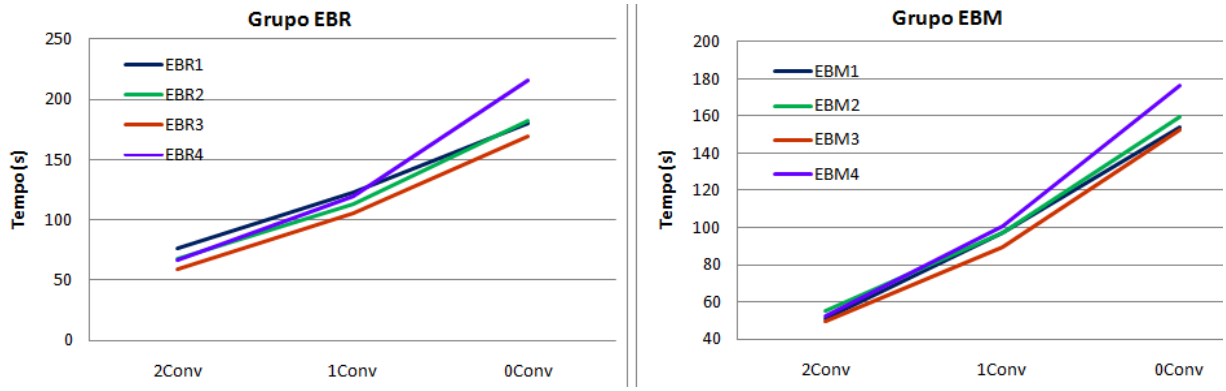

Figura 6.10: Tempo médio decorrido em segundos para os experimentos 2Conv, 1Conv e 0Conv sobre as propostas de extensão do iCube

Como observado nas comparações entre as configurações EBR1 *versus* EBR3 e EBM1 *versus* EBM3, para estas composições de predicados, o custo do acréscimo de *uma* junção para obter os vetores de características é menor que o custo de manter esse tipo de dado não convencional junto dos dados de composição da *mbOr*. Por exemplo, a configuração EBM1 armazena conjuntamente o vetor de características com todos os valores de distância aos representantes em sua respectiva camada perceptual, mas apesar de evitar o custo de junção para recuperar o vetor de características, a configuração EBM1 perde da configuração EBM3.

Outra observação atípica refere-se ao desempenho das configurações que degeneram a dimensão da tabela de fatos (configurações EBR4 e EBM4). Para estas composições de predicados, as configurações EBR4 e EBM4 apresentaram dentro de seu respectivo grupo um baixo tempo de resposta em consultas muito restritivas, como para as consultas 5Conv, 4Conv e 3Conv demonstradas na Figura

6.9. No entanto, esta é uma abordagem que deve ser utilizada com muita cautela, pois, como observado na Figura 6.10, o desempenho destas configurações (EBR4 e EBM4) degenera com maior facilidade do que as demais configurações em consultas pouco restritivas, como é o caso das consultas 2Conv, 1Conv e 0Conv. Nestas consultas o uso dos vetores de características como atributos da tabela de fatos demonstrou ser uma estratégia pouco aconselhável para IDW.

A partir dos resultados apresentados nestes experimentos também foi possível observar reações esperadas devido à estratégia de esquema estrela. Como esperado, a granularidade do grupo EBR impacta negativamente o desempenho do processamento das consultas IOLAP quando comparado ao grupo EBM. Este impacto ocorre devido ao custo de junções para compor os predicados de *mbOr*, que pode ser melhor observado na Figura 6.11, em que as configurações equivalentes são comparadas para o processamentos das consultas desta bateria de teste. O impacto resultante pelo custo de junção é acentuado em cenários em que o predicado convencional é muito seletivo gerando em média uma redução de 70,18% do tempo decorrido de processamento. No total a mudança de nível de descrição resultou em uma redução média de 33,91% no custo do processamento das consultas.

O custo de junção existente no grupo EBR degrada de tal modo o desempenho do tempo de processamento de consultas IOLAP que mesmo quando comparado o melhor resultado de EBR ao pior caso de EBM, o EBM na maioria das vezes apresentou melhores resultados do que o EBR. Em média a redução do tempo do EBM sobre o EBR foi de 9,57% e de no máximo de 48,45%.

Estes resultados acentuaram a hipótese de que as adaptações realizadas no grupo EBM geram as melhores configurações de esquema estrela para o iCube, tornando estas configurações fortes candidatas a iStar.

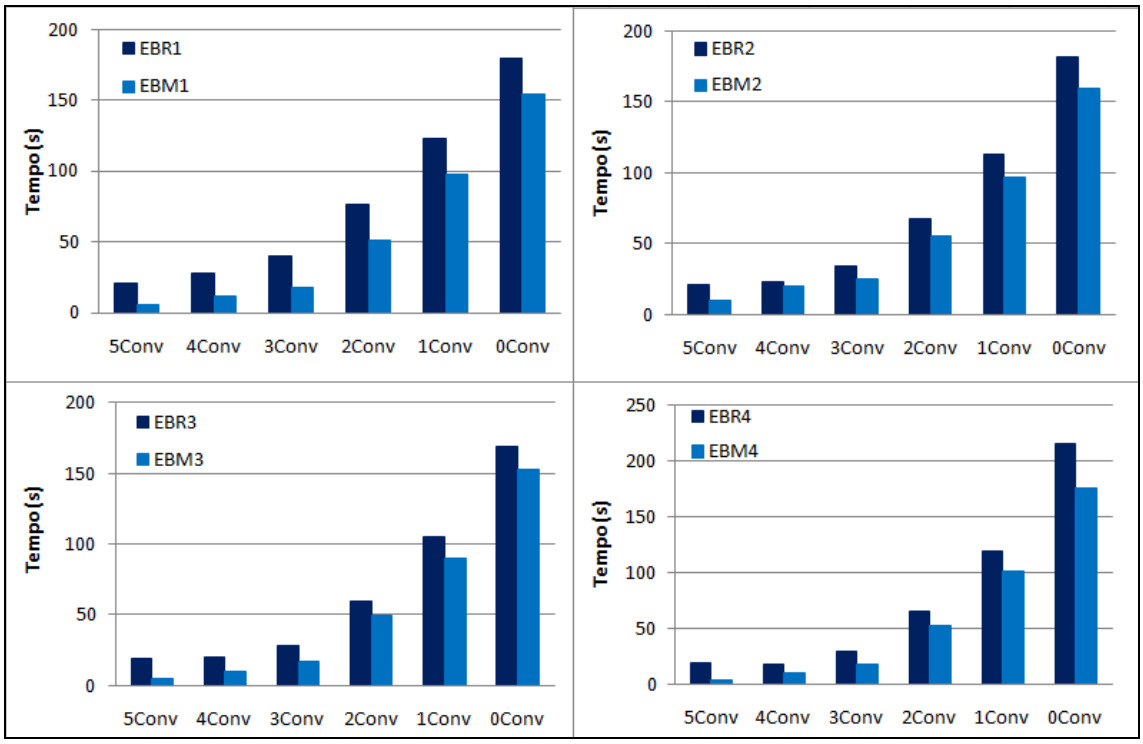

Figura 6.11: Comparação par a par entre as configurações equivalentes do grupo EBR e EBM para avaliar a redução de tempo gerada pela eliminação do custo de junção.

Tabela 6.2: Tempo decorrido do melhor caso para o grupo EBR e do pior caso para o grupo EBM, e sobre o ganho de desempenho gerado pelo grupo EBM.

	Melhor EBR	Pior EBM	Redução de Tempo
5Conv	18,71 (EBR3)	9,64 (EBM2)	**48,45%**
4Conv	18,27 (EBR4)	20,02 (EBM2)	**-9,57%**
3Conv	28,26 (EBR3)	25,13 (EBM2)	**11,07%**
2Conv	59,40 (EBR3)	54,97 (EBM2)	**7,46%**
1Conv	105,42 (EBR3)	101,04 (EBM4)	**4,16%**
0Conv	169,08 (EBR3)	176,08 (EBM4)	**-4,14%**

Esta bateria de testes também foi realizada sobre os atuais recursos de IDWing. Na Figura 6.12, as medidas de tempo médio decorrido das configurações compostas pela Onion-tree são apresentadas e como observado a partir desta figura a configuração SingleOnion apresentou no geral os melhores resultados quando comparada as demais configurações adaptadas da Onion-Tree.

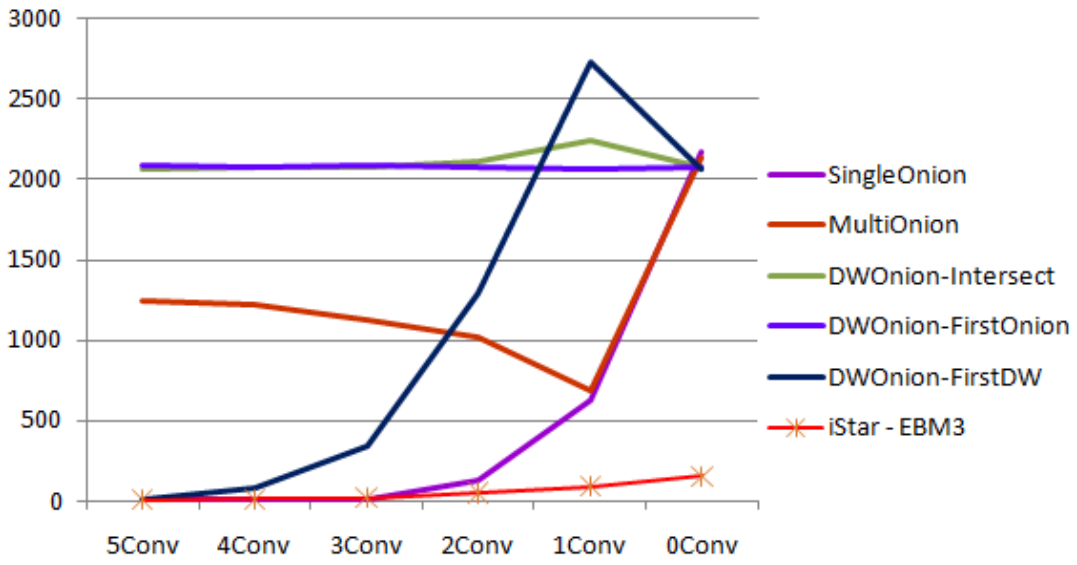

Figura 6.12: Tempo decorrido em segundos para os experimentos baseados na seletividade dos predicados convencionais sobre os atuais recursos de IDWing e sobre a configuração EBM3 do iStar.

Com o intuito de comparar a configuração dos atuais recursos de IDWing com menor tempo de resposta (SingleOnion) às configurações propostas por este trabalho, na Tabela 6.3 é apresentada a porcentagem de redução do tempo gerada pela configuração EBM3 quando comparada às demais configurações do grupo EBM e à SingleOnion. Como apresentado nesta tabela, o EBM3 apresentou uma expressiva redução do tempo quando comparado à SingleOnion, que em média resultou em uma redução de 40,59% do tempo decorrido de processamento das consultas IOLAP.

Tabela 6.3: Redução do tempo gerada pela configuração EBM3 em relação a melhor configuração dos atuais recursos de IDWing e em relação as demais configurações do grupo EBM.

	SingleO	**EBM1**	**EBM2**	**EBM4**
5Conv	38,90%	6,74%	47,56%	-27,48%
4Conv	-16,80%	19,58%	52,05%	8,57%
3Conv	-18,86%	5,83%	31,65%	4,69%
2Conv	61,55%	2,61%	9,78%	5,40%
1Conv	85,79%	7,94%	7,76%	11,30%
0Conv	92,98%	1,02%	4,52%	13,34%
Média	**40,59%**	**7,29%**	**25,55%**	**2,64%**

Este trabalho também realizou experimentos baseados na seletividade gerada pelo raio de abrangência e baseados no impacto da dimensionalidade dos vetores

de características. Estes experimentos também foram realizados para avaliar as configurações propostas e compará-las aos atuais recursos de IDWing, no entanto daremos preferência às configurações que apresentaram melhores resultados sobre a bateria de testes baseados na seletividade de predicados convencionais. Esta preferência é devido à característica de multidimensionalidade do ambiente de DWing, de modo a permitir que as imagens sejam analisadas sobre diferentes predicados convencionais e por qualquer predicado convencional. Portanto, para a bateria de testes baseada na dimensionalidade dos vetores de características, foram testadas apenas as configurações do grupo EBM e a configuração SingleOnion como trabalho correlato.

6.5.3 Experimentos baseados na Seletividade Determinada pelo Raio de Abrangência

Esta bateria de testes foi planejada para explorar a característica de flexibilidade de um ambiente de IDWing (i.e., o ambiente oferece suporte a qualquer predicado visual, sendo flexível tanto para a imagem de consulta quanto para o raio de abrangência). Dessa maneira, este trabalho pretende avaliar o desempenho das configurações propostas e correlatas segundo diferentes restrições de similaridade. Para tanto, um conjunto de dez tipos de consultas foi estabelecido, o qual gradativamente submetia a configuração testada a maiores esforços de processamento.

Estes experimentos foram conduzidos aumentando progressivamente o valor de raio de abrangência de 10% a 55% (da metade do diâmetro determinada pelo conjunto $\mathcal{S}$). O predicado convencional foi desconsiderado (i.e., estas consultas não possuem predicados convencionais apenas possuem predicados visuais) e o predicado visual foi determinado supondo uma situação extrema em que o usuário deseja utilizar todas as camadas perceptuais como critério de similaridade entre as imagens. Seguindo o cenário exemplo, o predicado visual refere-se às imagens que eram similares a s_q com relação à representação de Haralick Uniformidade, de Haralick Variância, de Haralick Entropia, de Histograma e Zernike.

Ademais, os experimentos foram realizados submetendo 10 imagens aleatoriamente como imagem de consulta e o tempo decorrido para o processamentos dessas consultas foi medido. Vale ressaltar que a extração do

conteúdo intrínseco dessas imagens de consulta também foi realizada previamente de maneira a não ser contabilizado no tempo decorrido do experimento.

Neste contexto, por exemplo, as consultas determinadas pelo raio de abrangência de 10% e pelo raio de 55% foram:

- Consulta r_q:10%: "*Quantas imagens são similares à imagem de consulta s_q segundo um raio de abrangência de 10% nas cinco camadas perceptuais?"*.
- Consulta r_q:55%: "*Quantas imagens são similares à imagem de consulta s_q segundo um raio de abrangência de 55% nas cinco camadas perceptuais?"*.

6.5.4 Resultados dos Experimentos baseados na Seletividade Determinada pelo Raio de Abrangência

De acordo com as premissas determinadas na seção anterior, os experimentos foram realizados e na Tabela 6.4 são apresentados os tempos médios decorridos para o processamento das consultas com raio de abrangência de 10% a 55%.

Tabela 6.4: Tempo decorrido em segundos para os experimentos baseados na seletividade dos predicados convencionais sobre as propostas de extensão do iCube

	EBR1	**EBR2**	**EBR3**	**EBR4**	**EBM1**	**EBM2**	**EBM3**	**EBM4**
r_q: 10%	63,94	42,17	40,77	34,69	35,03	30,99	28,78	*22,85*
r_q: 15%	91,23	75,58	72,00	70,12	69,70	63,52	59,98	*56,21*
r_q: 20%	127,79	116,36	110,82	118,71	97,68	102,01	*95,04*	95,29
r_q: 25%	159,68	149,69	144,87	172,03	135,66	133,17	*127,34*	135,67
r_q: 30%	180,66	181,87	167,79	217,09	153,53	158,86	*152,68*	175,77
r_q: 35%	210,99	205,07	197,13	260,92	*171,67*	179,33	178,98	200,14
r_q: 40%	234,40	229,15	218,91	294,95	201,64	202,97	*195,70*	223,50
r_q: 45%	280,96	275,45	269,74	363,12	248,42	257,19	*244,87*	269,50
r_q: 50%	298,40	295,34	294,45	395,26	273,07	281,82	*269,61*	288,19
r_q: 55%	319,86	317,17	315,04	423,84	*292,62*	297,70	294,28	305,95

A partir das medidas apresentadas na Tabela 6.4, é possível observar que o grupo EBM novamente apresentou melhores resultados que o grupo EBR, o que confirma a hipótese que motivou este trabalho a estender as propostas de EBR aos EBM, ou seja, descrever as imagens em um nível mais granular degrada o

desempenho do tempo de processamento de consultas IOLAP devido ao custo de junção necessário para obter os dados que compõem a *mbOr*. Consequentemente, é possível afirmar que o grupo EBM possui os melhores candidatos a iStar. Este impacto pode ser melhor observado na Tabela 6.5, em que é apresentado a porcentagem de redução do tempo gerado pelas configurações EBM sobre suas equivalentes no grupo EBR.

O impacto resultante pelo custo de junção é acentuado em cenários que em que o predicado visual é muito seletivo, gerando em média uma redução de 33,82% do tempo decorrido de processamento. No total, a mudança de nível de descrição resultou em uma redução média de 16,78%.

Tabela 6.5: Comparação par a par entre as configurações equivalentes do grupo EBR e EBM para avaliar a redução de tempo gerada pela eliminação do custo de junção.

	EBR1 x EBM1	EBR2 x EBM2	EBR3 x EBM3	EBR4 x EBM4	Média
r_q: 10%	45,22%	26,51%	29,41%	34,15%	**33,82%**
r_q: 15%	23,60%	15,96%	16,69%	19,84%	**19,02%**
r_q: 20%	23,56%	12,33%	14,24%	19,73%	**17,47%**
r_q: 25%	15,04%	11,03%	12,10%	21,14%	**14,83%**
r_q: 30%	15,02%	12,65%	9,01%	19,04%	**13,93%**
r_q: 35%	18,64%	12,55%	9,20%	23,30%	**15,92%**
r_q: 40%	13,98%	11,43%	10,60%	24,22%	**15,06%**
r_q: 45%	11,58%	6,63%	9,22%	25,78%	**13,30%**
r_q: 50%	8,49%	4,58%	8,44%	27,09%	**12,15%**
r_q: 55%	8,51%	6,14%	6,59%	27,81%	**12,26%**
Média	**18,36%**	**11,98%**	**12,55%**	**24,21%**	**16,78%**

Tabela 6.6: Tempo decorrido do melhor caso para o grupo EBR e do pior caso para o grupo EBM, e sobre o ganho de desempenho gerado pelo grupo EBM.

	Melhor EBR	Pior EBM	Redução do tempo
r_q: 10%	34,69 (EBR4)	35,03 (EBR1)	**-0,95%**
r_q: 15%	70,12 (EBR4)	69,70 (EBM1	**0,60%**
r_q: 20%	110,82 (EBR3)	102,01 (EBM2)	**7,95%**
r_q: 25%	144,87 (EBR3)	135,67 (EBM4)	**6,35%**
r_q: 30%	167,79 (EBR3)	175,77 (EBM4)	**-4,75%**
r_q: 35%	197,13 (EBR3)	200,14 (EBM4)	**-1,53%**
r_q: 40%	218,91 (EBR3)	223,50 (EBM4)	**-2,09%**
r_q: 45%	269,74 (EBR3)	269,50 (EBM4)	**0,09%**
r_q: 50%	294,45 (EBR3)	288,19 (EBM4)	**2,13%**
r_q: 55%	315,04 (EBR3)	305,95 (EBM4)	**2,89%**

Para esta bateria de testes, a comparação do melhor caso do EBR ao pior caso do EBM foi um pouco menos expressivo, todavia a redução média manteve-se positiva de 1,07% e de no máximo 7,95%. Esta comparação é apresentada na Tabela 6.6.

Em relação às medidas de tempo de processamento das consultas IOLAP apresentadas na Tabela 6.4, é possível observar que o EBM3 se manteve como um dos melhores candidatos para o iStar. Na Tabela 6.7 é apresentada a porcentagem da redução de tempo gerada pelo EBM3 quando comparado às demais configurações propostas. Como observado, a configuração EBM3 apresenta uma positiva redução do tempo de processamento de consulta: em média o EBM3 gera uma redução média sobre todas as demais configurações de 10,93%.

Nos resultados apresentados na Tabela 6.4 também é possível observar que para consultas muito restritivas, como com raio igual a 10% e 15%, as configurações que degeneram a tabela de fatos (EBR4 e EBM4) apresentam baixo tempo de resposta quando comparadas às demais configurações do seu grupo, e mais uma vez estas configurações demonstraram ser muito reativas à seletividade da consulta, pois seu desempenho é muito degradado conforme a seletividade da consulta é reduzida.

Tabela 6.7: Redução do tempo gerada pela configuração EBM3 em relação as demais configurações propostas para o iCube.

	EBR1	EBR2	EBR3	EBR4	EBM1	EBM2	EBM4	Média
r_q: 10%	54,987%	31,748%	29,405%	17,040%	17,824%	7,125%	-25,974%	**16,519%**
r_q: 15%	34,251%	20,636%	16,689%	14,461%	13,944%	5,566%	-6,707%	**12,355%**
r_q: 20%	25,631%	18,325%	14,242%	19,945%	2,711%	6,834%	0,266%	**10,994%**
r_q: 25%	20,256%	14,933%	12,102%	25,983%	6,137%	4,381%	6,145%	**11,242%**
r_q: 30%	15,487%	16,052%	9,008%	29,671%	0,556%	3,892%	13,136%	**10,975%**
r_q: 35%	15,171%	12,721%	9,205%	31,405%	-4,258%	0,194%	10,570%	**9,376%**
r_q: 40%	16,513%	14,599%	10,604%	33,650%	2,946%	3,583%	12,438%	**11,792%**
r_q: 45%	12,844%	11,100%	9,220%	32,565%	1,429%	4,792%	9,140%	**10,136%**
r_q: 50%	9,649%	8,711%	8,438%	31,790%	1,268%	4,334%	6,448%	**8,830%**
r_q: 55%	7,996%	7,215%	6,589%	30,568%	-0,567%	1,149%	3,814%	**7,096%**
Média	**21,279%**	**15,604%**	**12,550%**	**26,708%**	**4,199%**	**4,185%**	**2,928%**	**10,932%**

Esta bateria de testes também foi realizada sobre os atuais recursos de IDWing. Na Figura 6.13, são apresentadas as medidas de tempo decorrido do

processamento das consultas para as configurações compostas pela Onion-tree. Para esta bateria de testes, os atuais recursos de IDWing apresentaram desempenhos próximos, o que dificulta destacar uma destas configurações com a mais indicada.

A redução máxima entre as configurações Onion-Tree foi de 7,4% (i.e., a redução da melhor configuração sobre a pior configuração), enquanto a redução mínima para as configurações Onion-Tree entre a melhor configuração e a pior foi de 2,4%. Neste contexto, devido à preferência pela SingleOnion segundo seus resultados nos experimentos baseados na seletividade dos predicados convencionais, esta configuração será utilizada como base de comparação as proposta de iCube nos experimentos sobre a dimensionalidade dos vetores de características.

Ainda detalhando a Figura 6.13, é possível observar que o iStar representado pela configuração EBM3 apresentou um expressivo ganho sobre o tempo de processamento de consulta.

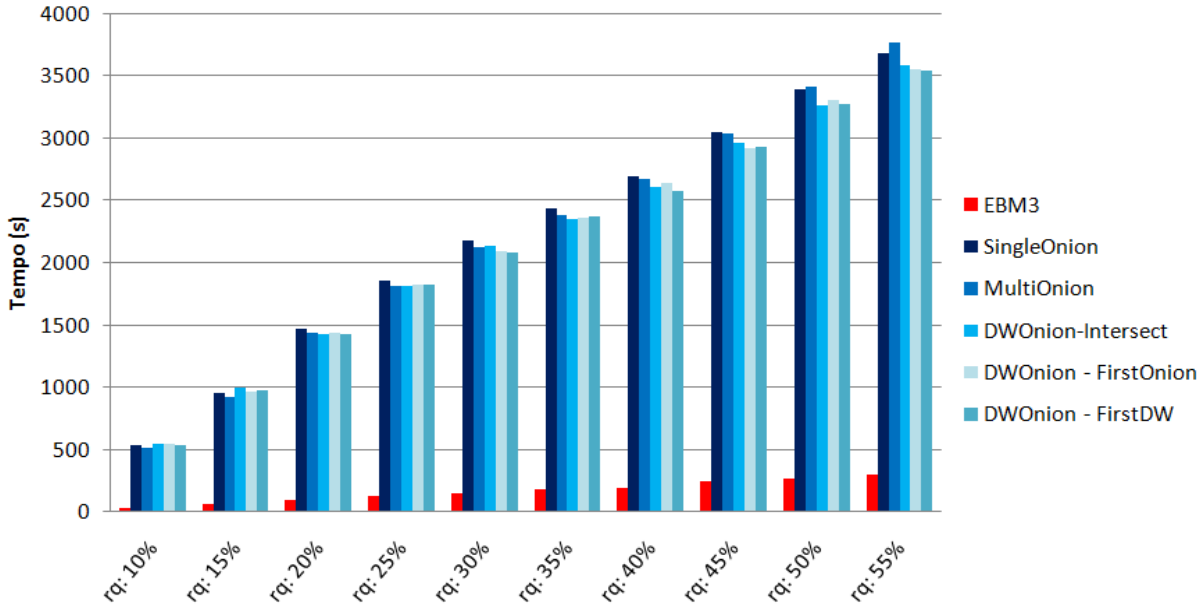

Figura 6.13: Tempo decorrido em segundos para os experimentos baseados na seletividade dos predicados convencionais sobre os atuais recursos de IDWing e a configuração EBM3 do iStar.

Na Tabela 6.8 é apresentada a porcentagem de redução do tempo gerada pela configuração EBM3 com relação aos atuais de recursos de IDWing. O EBM3 atuou de maneira extremamente eficiente, reduzindo o tempo de processamento a

92,82% em média. Fato este que reafirma a eficiência da proposta do iStar sobre os atuais recursos de IDWing.

Tabela 6.8: Redução do tempo gerada pela configuração EBM3 em relação as atuais recursos de IDWing.

	SingleOnion	MultiOnion	DWOM1	DWOM2	DWOM3
r_q: 10%	94,59%	94,44%	94,72%	94,70%	94,64%
r_q: 15%	93,69%	93,49%	93,97%	93,79%	93,84%
r_q: 20%	93,52%	93,38%	93,32%	93,38%	93,35%
r_q: 25%	93,16%	92,99%	92,99%	93,02%	93,03%
r_q: 30%	92,98%	92,82%	92,86%	92,72%	92,66%
r_q: 35%	92,66%	92,49%	92,38%	92,41%	92,44%
r_q: 40%	92,74%	92,67%	92,49%	92,58%	92,41%
r_q: 45%	91,97%	91,95%	91,74%	91,62%	91,63%
r_q: 50%	92,04%	92,09%	91,73%	91,85%	91,76%
r_q: 55%	92,01%	92,20%	91,80%	91,73%	91,68%

6.5.5 Experimentos baseados na Dimensionalidade dos Vetores de Características

Esta bateria de testes tem o intuito de investigar uma possível relação entre a estratégia de geração do esquema estrela e a composição dos predicados visuais. Ademais esta bateria foi realizada para avaliar se a complexidade dos dados de imagens, ou seja, a dimensionalidade dos vetores, gera algum impacto sobre o desempenho no processamento de consultas IOLAP. Para esta análise, experimentos foram conduzidos com a eliminação progressiva das camadas perceptuais segundo a dimensionalidade dos vetores de características. Desse modo, para cada série de experimentos as camadas perceptuais com os vetores de características mais longos eram desconsideradas até que restasse apenas uma camada perceptual.

Nestes experimentos, o raio de abrangência foi fixado em três valores: 10%, 30% e 55%; dez imagens aleatórias foram submetidas como imagem de consulta, e o predicado convencional não foi utilizado para compor as consultas IOLAP. Vale ressaltar que a extração do conteúdo intrínseco dessas imagens de consulta foi realizada previamente de maneira a não ser contabilizado no tempo decorrido do experimento.

Considerando as especificações desta bateria de testes e considerando a dimensionalidade dos vetores de características, a camada perceptual de Histograma de níveis de cinza foi a primeira camada desconsiderada (Consulta "*Variância, Uniformidade, Entropia e Zernike"*) seguida pela camada perceptual de Zernike (Consulta "*Variância, Uniformidade e Entropia"*). Como as camadas Variância, Uniformidade e Entropia possuem vetores de características de mesma dimensionalidade, os experimentos foram realizados sobre consultas: "*Variância e Uniformidade"*, "*Variância e Entropia"*, e "*Uniformidade e Entropia"*. Por fim, esta bateria de testes submeteu as configurações avaliadas a consultas definidas por apenas uma camada perceptual (Consultas "*Variância*", "*Uniformidade*", "*Entropia*", "*Zernike*" e "*Histograma*").

Devido aos resultados apresentados anteriormente (Seções 6.5.2 e 6.5.4), esta bateria de testes foi realizada apenas para as configurações do grupo EBM e para a SingleOnion. A SingleOnion foi selecionada para esta bateria de testes por ser uma configuração com resultados muito semelhante às demais configurações de Onion-Tree para os experimentos baseados em seletividade determinada pelos raios de abrangência, e principalmente devido ao seu excelente desempenho na bateria de testes baseados nos predicados convencionais, em que as configurações foram avaliadas sobre uma característica fundamental do um ambiente de IDWing, a multidimensionalidade. Enquanto as configurações do EBM foram selecionadas para esta bateria de testes por apresentarem melhores resultados que as configurações do grupo EBR.

6.5.6 Resultados dos Experimentos baseados na Seletividade Determinada pelo Raio de Abrangência

Considerando a dimensionalidade dos vetores de características, na Tabela 6.9 são apresentados os tempos médios decorridos para o processamento das consultas IOLAP determinadas pela composição de predicado visual: imagens similares a s_q segundo as camadas perceptuais de Variância, Uniformidade, Entropia e Zernike. Como determinado pela premissa desta bateria de testes, a camada perceptual Histograma foi desconsiderada neste experimento por possuir o vetor de características mais longo, com 254 atributos.

Tabela 6.9: Tempo decorrido em segundos para a execução da consulta IOLAP segundo as camadas perceptuais Variância, Uniformidade, Entropia e Zernike.

	Variância, Uniformidade, Entropia e Zernike				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	11,30	13,19	18,37	11,54	516,93
r_q: 30%	66,84	69,45	72,15	73,78	1748,88
r_q: 55%	119,76	124,16	128,35	125,22	3023,80

Diferentemente dos resultados apresentados nas baterias anteriores (Figuras 6.9 e 6.10, e Tabela 6.4), a configuração EBM3 apresentou os maiores tempos de resposta sobre o processamento de consultas IOLAP. Desconsiderar a camada perceptual com maior custo físico alterou significativamente o desempenho das configurações EBM. Os resultados apresentados na Tabela 6.9 seguiram uma reação esperada segundo os conceitos de DWing convencionais, em que o EBM1, esquema estrela projetado para realizar poucas junções, apresenta os melhores resultados no processamento das consultas.

O baixo desempenho do EBM1 nas baterias de testes anteriores (Seção 6.5.2 e 6.5.4) é causado devido à dimensionalidade dos vetores de características de histograma. A hipótese para explicar este fato é que os vetores de características de histograma, ao estarem presentes na mesma tabela utilizada como mecanismo de filtragem, podem causar uma paginação de poucas tuplas e consequentemente exigir um maior número de acessos a disco para verificar as condições de composição da *mbOr*. Em outras palavras, como as tuplas das dimensões visuais se tornam mais longas por conta do histograma de níveis de cinza, o número de acessos a disco pode ser maior e consequentemente o tempo despendido para processar a consulta também pode ser maior.

Outra interessante observação sobre os resultados da Tabela 6.9 é feita quando comparado o tempo de EBM2 ao tempo de EBM3. Como observado, o EBM2, que realiza várias junções para recuperar os vetores de características, apresentou melhores resultados do que o EBM3. Mais uma vez, esta reação é justificada pela hipótese de relação entre a dimensionalidade do histograma e o número de acessos a disco. Mesmo estando todos os vetores em uma mesma tabela, o impacto gerado pelos vetores referentes ao histograma (i.e., possível

aumento no número de acessos a disco) gera uma maior degradação no desempenho de processamento do que o custo de junções no esquema EBM2.

Neste contexto, este trabalho destaca a necessidade de haver um esquema especial quando vetores de características longos são utilizados. Segundo os resultados apresentados nas seções anteriores (Seção 6.5.2 e 6.5.4) e Tabela 6.9, a configuração EBM3 demonstra ser a estratégia de esquema estrela mais adequada para situações em que o predicado visual de uma consulta IOLAP é determinado por múltiplos descritores, sendo que pelo menos um destes descritores representa a imagem em um longo vetor de características, como o vetor de histograma.

Ainda segundo os resultados apresentados na Tabela 6.9 foi também possível observar que para condições muita restritivas, como para o raio de abrangência de 10%, o esquema EBM4 apresentou resultado muito satisfatório, e assim como nos resultados das baterias anteriores esse desempenho é degradado em condições pouco seletivas.

Assim como nos experimentos sem predicados convencionais (i.e., bateria de testes anterior), o trabalho correlato, representado pela SingleOnion, apresentou um desempenho muito abaixo do que as configurações propostas por este trabalho (Tabela 6.9). Quando comparado o EBM1 à SingleOnion, o esquema proposto por este trabalho gerou uma redução de tempo média de 96,68%.

Seguindo as premissas determinadas para esta bateria de testes, outro conjunto de testes foi conduzido eliminando o segundo maior vetor de característica, que no cenário exemplo refere-se à camada perceptual de Zernike (32 atributos). Na Tabela 6.10 são apresentados os tempos decorridos para o processamento das consultas IOLAP determinadas pelo predicado visual: imagens similares a s_q segundo as camadas perceptuais de Variância, Uniformidade e Entropia.

Tabela 6.10: Tempo decorrido em segundos para a execução da consulta IOLAP segundo as camadas perceptuais Variância, Uniformidade e Entropia.

	Variância, Uniformidade e Entropia				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	9,47	11,08	15,98	11,50	525,06
r_q: 30%	32,20	36,82	40,15	41,16	1652,72
r_q: 55%	47,03	50,50	54,07	52,95	2451,50

Na Tabela 6.10, observou-se as mesmas relações entre o desempenho da configuração e a dimensionalidade do vetor de características, confirmando a

hipótese de que vetores de características longos geram efeitos atípicos, o que torna necessário um esquema especial. Também foi possível afirmar que, para predicados visuais compostos por pequenos vetores de características, o desempenho de um IDWing segue os conceitos tradicionais de DW, ou seja, o esquema EBM1 é o mais aconselhado.

Também sobre os resultados apresentados na Tabela 6.10, o EBM2 apresentou melhores resultados do que o EBM3, o que reafirma a hipótese de que o armazenamento de vetores muito longos degeneram o desempenho de processamento de consultas IOLAP. No entanto, segundo os resultados apresentados nesta tabela, a configuração EBM4 não manteve o seu desempenho. Para esta configuração de predicados visuais, o EBM4 deixou de apresentar um bom desempenho para uma condição muito restritiva, r_q = 10%. Esta observação permite concluir que a degeneração da tabela de fatos só é aconselhada em cenários em que a consulta é muito restritiva e em situações em que seu predicado visual seja composto por camadas perceptuais custosas (i.e., múltiplos descritores e vetores de características longos).

Nestes experimentos, as configurações propostas também apresentaram melhores desempenhos que o trabalho correlato SingleOnion (Tabela 6.10), em que o esquema EBM1 gerou uma redução média de tempo de 98,10% sobre os tempos gerados pela SingleOnion.

Mais uma vez, seguindo as premissas desta bateria de testes, experimentos foram realizados desconsiderando outro vetor de características. Como os demais vetores possuem a mesma dimensão, cinco atributos, realizamos experimentos sobre as consultas IOLAP determinadas pelos predicados visuais: imagens similares a s_q segundo Variância e Uniformidade; segundo Variância e Entropia; e segundo Entropia e Uniformidade. Nas Figuras 6.14, 6.15 e 6.16 são apresentados respectivamente os tempos médios decorridos para o processamento destas consultas IOLAP.

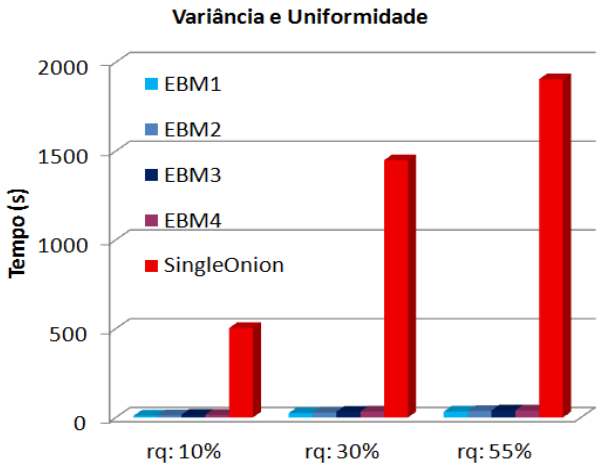

Figura 6.14: Tempo decorrido em segundos para a execução da consulta IOLAP segundo as camadas perceptuais Variância e Uniformidade.

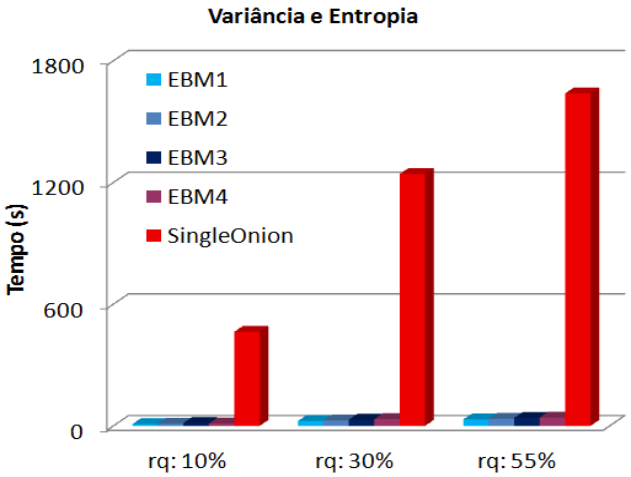

Figura 6.15: Tempo decorrido em segundos para a execução da consulta IOLAP segundo as camadas perceptuais Variância e Entropia.

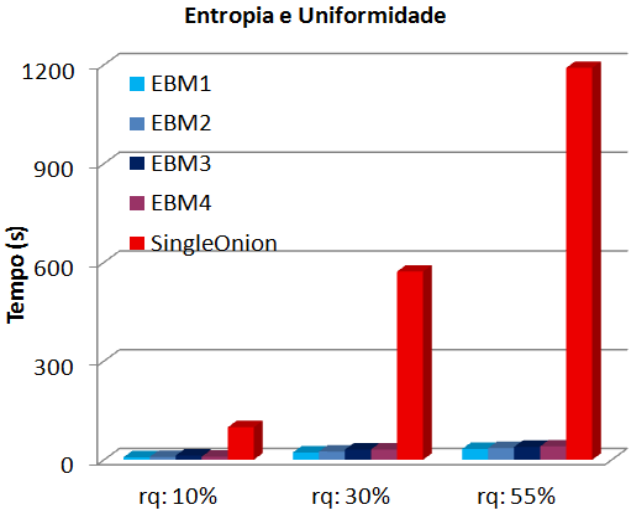

Figura 6.16: Tempo decorrido em segundos para a execução da consulta IOLAP segundo as camadas perceptuais Uniformidade e Entropia.

Nestes resultados, também observou-se as mesmas relações entre o desempenho da configuração e a dimensionalidade do vetor de características, reafirmando a hipótese de que vetores de características longos geram efeitos atípicos, o que torna necessário um esquema especial. Consequentemente, estratégias de esquemas estrela convencionais de DW são aplicadas para IDW com vetores de características com tamanho curto e médio.

Nestes resultados também observa-se a tendência do EBM2 apresentar melhores resultados do que o EBM3, e a configuração EBM4 não gerar resultados significativos para uma condição muito restritiva, r_q = 10%.

Mais uma vez, as configurações propostas apresentaram melhores desempenhos que o trabalho correlato SingleOnion, destacando o esquema EBM1, que gerou sobre a SingleOnion uma redução de tempo média de 98,21%, 98,10% e 95,20% respectivamente para os predicados Variância e Uniformidade; Variância e Entropia, e Entropia e Uniformidade.

Por fim, este trabalho realizou experimentos sobre cada uma das cinco camadas perceptuais, seguindo as premissas determinadas na seção anterior, e os resultados são apresentados nas tabelas 6.11, 6.12 e 6.13.

Tabela 6.11: Tempo decorrido em segundos para a execução da consulta IOLAP segundo a camada perceptual Uniformidade de Haralick.

	Uniformidade				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	6,248	7,472	12,514	9,272	58,458
r_q: 30%	13,572	14,895	23,325	20,835	342,416
r_q: 55%	17,893	19,127	27,149	27,368	712,688

Tabela 6.12: Tempo decorrido em segundos para a execução da consulta IOLAP segundo a camada perceptual Entropia de Haralick.

	Entropia				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	5,930	6,851	9,952	7,967	39,354
r_q: 30%	13,087	14,402	20,923	20,011	269,185
r_q: 55%	17,851	19,119	27,077	26,279	599,943

Tabela 6.13: Tempo decorrido em segundos para a execução da consulta IOLAP segundo a camada perceptual Variância de Haralick.

	Variância				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	14,313	15,933	24,427	23,038	414,315
r_q: 30%	18,114	19,342	27,779	26,930	958,954
r_q: 55%	18,636	19,876	27,537	27,052	1061,185

Como observado nos resultados apresentados nas tabelas 6.11, 6.12 e 6.13, vetores de características curtos permitem que os IDW sejam projetados segundo as estratégias convencionais de DW. Estes conceitos são aplicados pois os vetores de características se comportam como textos curtos armazenados nas tabelas e dessa maneira não degradam o desempenho de processamento das consultas.

Como observado nos experimentos anteriores (tabelas tabelas 6.9 e 6.10, e Figuras 6.14, 6.15 e 6.16), o EBM1 demonstrou ser o melhor esquema estrela de IDW para vetores curtos e médios, o EBM2 gerou melhores resultados que o EBM3, uma vez que as tuplas da tabela auxiliar do EBM3 são muito longas e consequentemente, a nossa hipótese é que resultam em mais acessos a disco, e para cenários em que o predicado visual é determinado por pequenos vetores, a degeneração da tabela de fatos (esquema EBM4) não resulta em melhor desempenho, mesmo para predicados muito restritivos, como para raio de abrangência igual a 10%.

Também para estes três experimentos a configuração EBM1 sempre gerou desempenhos mais expressivos do que a SingleOnion (i.e., atuais recursos de IDWing), apresentando uma redução média de 94,28% para as consultas sobre Uniformidade, 92,37% para as consultas sobre Entropia e 97,63% para as consultas sobre Variância.

Tabela 6.14: Tempo decorrido em segundos para a execução da consulta IOLAP segundo a camada perceptual Zernike.

	Zernike				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	22,125	22,487	24,927	24,719	3,900
r_q: 30%	70,885	71,411	75,364	75,337	184,865
r_q: 55%	95,519	95,874	101,172	99,977	810,333

Também para os experimentos realizados sobre a camada perceptual de Zernike, que possuem um vetor de tamanho médio com 32 atributos (Tabela 6.14), a

configuração EBM1 apresentou menor tempo de resposta do que o esquema EBM3. No entanto, para o raio de abrangência de 10%, as configurações iStar apresentaram piores resultados do que o trabalho correlato SingleOnion. A fim de justificar esta queda de desempenho, a seletividade da *mbOr* para o predicado visual determinado pela camada perceptual de Zernike foi analisado.

Na Tabela 6.15 é apresentado o número de imagens selecionadas pela *mbOr* e o real número de imagens similares, assim como a precisão do *mbOr* com relação ao número de imagens similares. Como é possível observar, a mbOr para o raio de abrangência de 10% apresentou um baixo desempenho. Esta baixa precisão é o principal elemento que justifica o baixo desempenho das configurações EBM quando comparadas à SingleOnion para consultas determinadas pelo raio r_q = 10%.

Esta hipótese sobre a precisão da mbOr e o desempenho das configurações EBM pode ser confirmada quando comparado o tempo de processamento das configurações EBM à SingleOnion para os raios 30% e 55%. Para estes testes a mbOr apresentou uma significativa precisão e consequentemente o desempenho do processamento das consultas foi muito melhor do que os experimentos em que a mbOr gerou uma baixa precisão. Na Tabela 6.16 é apresentado para os raios igual a 30% e 55% o número de imagens selecionadas pela mbOr, o número de imagens realmente similares a $s_{q,}$ assim como a precisão da mbOr segundo o conjunto de imagens similares.

Tabela 6.15: Número de imagens selecionadas pela mbOr, número de imagens similares a s_q segundo o raio de abrangência de 10% e a precisão da mbOr nesta consulta segundo a camada perceptual de Zernike.

		mbOr	Similar	Precisão
Raio:10	Imagem s_1	50756	1654	0,03
	Imagem s_2	8045	14	0,00
	Imagem s_3	5901	15	0,00
	Imagem s_4	2347	735	0,31
	Imagem s_5	27275	130	0,00
	Imagem s_6	6289	435	0,07
	Imagem s_7	50400	19761	0,39
	Imagem s_8	12651	81	0,01
	Imagem s_9	51054	24213	0,47
	Imagem s_{10}	41175	2048	0,05

Tabela 6.16: Número de imagens selecionadas pela mbOr, número de imagens similares a s_q segundo o raio de abrangência de 30% e 55%, e a precisão da mbOr nesta consulta segundo a camada perceptual de Zernike.

	Raio 30%			Raio 55%		
	mbOr	**Similar**	**Precisão**	**mbOr**	**Similar**	**Precisão**
Imagem s_1	90371	76863	0,85	118596	110022	0,93
Imagem s_2	99015	62476	0,63	126998	117841	0,93
Imagem s_3	74845	14616	0,20	130150	119972	0,92
Imagem s_4	80760	1475	0,02	112417	80175	0,71
Imagem s_5	97098	71691	0,74	123535	111135	0,90
Imagem s_6	96682	39596	0,41	124339	107132	0,86
Imagem s_7	84287	71215	0,84	113372	101128	0,89
Imagem s_8	94942	71092	0,75	121741	111961	0,92
Imagem s_9	85249	72660	0,85	114250	103212	0,90
Imagem s_{10}	81469	70936	0,87	109149	98883	0,91

A partir dos experimentos realizados sobre a camada perceptual Histograma de níveis de cinza (Tabela 6.17), foi possível observar efeitos interessantes entre o esquema estrela definido para o IDW e a dimensionalidade do vetor de características. Pelo fato do histograma de níveis de cinza se comportar como um texto longo (i.e., um atributo não convencional em ambientes de DWing), a estratégia de esquema estrela convencional de DW não se aplica em cenários que este atributo é utilizado como predicado de consultas IOLAP.

Segundo estes resultados, também foi observado que a configuração EBM1 apresenta piores resultados do que as configurações que precisam realizar junção para recuperar o vetor (EBM2 e EBM3), o que confirma a relação existente entre a dimensionalidade dos vetores de características e o desempenho de esquemas do tipo EBM1. Mais uma vez, para condições menos restritivas, como para o raio igual a 30% e 55%, a configuração EBM4 apresentou os piores resultados.

Os esquemas EBM2 e EBM3, que são esquemas que precisam realizar junção para recuperar o vetor, apresentaram os melhores tempos de processamento. Vale ressaltar ainda que o EBM3 apresentou resultados discretamente inferiores do que o EBM2, pois a tabela auxiliar desta configuração possui tuplas ligeiramente maiores do que as tuplas da tabela auxiliar do EBM2.

Assim como nos experimentos sobre a camada perceptual de Zernike, para o raio de abrangência de 10%, o desempenho das configurações EBM foi inferior que o desempenho obtido pela SingleOnion (Tabela 6.17). Também para estes experimentos o baixo desempenho das configurações EBM é justificado pela baixa

precisão da mbOr. Na Tabela 6.18, é apresentado o número de imagens selecionadas pela mbOr e o real número de imagens similares, assim como a precisão da mbOr para o predicado visual determinado pelo Histograma para r=10%, 30% e 55%.

Tabela 6.17: Tempo decorrido em segundos para a execução da consulta IOLAP segundo a camada perceptual Histograma.

	Histograma				
	EBM1	**EBM2**	**EBM3**	**EBM4**	**SingleO**
r_q: 10%	206,148	197,987	198,824	200,677	15,352
r_q: 30%	211,380	207,632	209,790	214,556	471,857
r_q: 55%	236,420	245,688	234,309	249,967	748,105

No entanto, para situações em que a precisão da mbOr é alta (i.e., para os raios 30% e 55%), as configurações EBM apresentaram resultados muito mais significativos do que a SingleOnion, onde a configuração EBM3 gerou uma redução média de 55,54% para consultas com raio igual a 30% e de 68,68% para as consultas com raio igual a 55%.

Tabela 6.18: Número de imagens selecionadas pela mbOr, número de imagens similares a s_q segundo o raio de abrangência de 30% e 55%, e a precisão da mbOr nesta consulta segundo a camada perceptual de Histograma.

	Raio = 10%			**Raio = 30%**			**Raio = 55%**		
	mbOr	**Similar**	**Precisão**	**mbOr**	**Similar**	**Precisão**	**mbOr**	**Similar**	**Precisão**
Imagem s_1	84964	1149	0,0135	88619	82744	0,93	89901	88654	0,99
Imagem s_2	86751	1656	0,0191	88644	83737	0,94	90003	88888	0,99
Imagem s_3	923	80	0,0867	7414	3858	0,52	118794	113048	0,95
Imagem s_4	84755	516	0,0061	88519	1153	0,01	89630	86963	0,97
Imagem s_5	86303	771	0,0089	88619	83635	0,94	89919	88796	0,99
Imagem s_6	86102	2332	0,0271	88619	83388	0,94	89917	88755	0,99
Imagem s_7	86729	37116	0,4280	88626	84030	0,95	89951	88897	0,99
Imagem s_8	85094	992	0,0117	88618	83587	0,94	89911	88712	0,99
Imagem s_9	86641	17479	0,2017	88621	83749	0,95	89939	88843	0,99
Imagem s_{10}	86707	32491	0,3747	88620	83930	0,95	89948	88883	0,99

6.5.7 Considerações finais

Neste capítulo, foram apresentadas alterações sobre a nossa proposta de esquema estrela para o IDW a fim de estender as funcionalidades do iCube ao suporte de camadas perceptuais, tornando o ambiente proposto de IDWing ainda mais flexível e mais adaptado às intenções de consulta do usuário. Também neste

capítulo foram descritas três baterias de testes que tiveram como objetivo avaliar o desempenho das configurações propostas, além de identificar relacionamentos entre a estratégia de esquema estrela e o desempenho da estratégia para o processamento de consultas IOLAP, e por fim comparar o desempenho das nossas propostas aos recursos de *image data warehousing* disponíveis atualmente.

Segundo os resultados apresentados neste capítulo torna-se evidente que os esquemas estrela do grupo EBM consistem em uma melhor estratégia de esquema estrela do que os esquemas estrela do grupo EBR, pois mesmo quando comparado o melhor resultado do grupo EBR com o pior resultado do EBM, o grupo EBM gerou uma redução média de até 48,45%. A partir desses resultados podemos concluir que modelar o iStar em um nível muito granular, como é o caso das configurações EBR, resulta em uma degradação no desempenho no processamento de consultas IOLAP devido ao maior custo de junção para compor as condições definidas pela mbOr (i.e. compor a interseção entre os intervalos de distância). Neste contexto, este trabalho adota e sugere a modelagem do iStar segundo os conceitos definidos pelo grupo EBM, em que os atributos referentes ao valor distância da imagens s_i aos representantes devem ser armazenados em uma mesma dimensão visual.

Em geral, as configurações que armazenam os vetores de características na tabela de fatos, como o esquema EBM4, são estratégias de esquema estrela para ambientes em que as consultas são muito restritivas e concomitantemente seu predicado visual é determinado por múltiplas camadas perceptuais, sendo que uma dessas camadas possui vetores de características longos ou médios. São consideradas consultas muito restritivas as consultas compostas por muitos predicados convencionais ou a um critério de similaridade determinado por um raio de abrangência pequeno. No entanto, devido ao fato desta configuração ter o desempenho degradado facilmente conforme as consultas se tornam menos restritivas, esta configuração não demonstra ser um bom candidato para o iStar.

Também por meio de análises sobre os resultados das baterias de testes apresentadas neste capítulo foi possível observar uma forte relação entre a dimensionalidade dos vetores de características e o desempenho das configurações com múltiplos descritores.

Para cenários em que as consultas IOLAP são determinadas por múltiplas camadas perceptuais e por vetores de características de dimensões variadas, a configuração EBM3 (i.e. que possui uma tabela auxiliar para armazenar todos os

vetores de características) apresentou baixo tempo de resposta (Figura 6.9 e 6.10, e Tabela 6.4). Neste cenário de processamento de consultas, as estratégias tradicionais de esquema estrela de DW não geram os resultados esperados devido ao impacto causado pela dimensionalidade do vetor de características longo.

A hipótese para explicar este fato é que o custo de uma junção para recuperar os vetores de características é menor que o número de acessos a disco que os vetores de características de histograma podem causar ao estarem presentes na mesma tabela utilizada no mecanismo de filtragem da mbOr. Logo, para cenários em que o ambiente de IDWing é predominantemente utilizado para o processamento de consultas com múltiplas camadas perceptuais de diversas dimensionalidade de vetores de características, a proposta EBM3 demonstrou ser a melhor candidata a iStar.

Para cenários em que o IDW é utilizado para processar consultas compostas por vetores de características curtos e médios, ou para cenários em que o usuário não demonstra seguir um padrão de composição de consultas IOLAP indicamos que o iStar seja modelado segundo o esquema EBM1 (i.e. todos os dados de uma camada perceptual são armazenados na mesma dimensão visual, ou seja, são armazenados conjuntamente as distâncias para os representantes e o vetor de característica).

Também com base nos resultados apresentados neste capítulo, a configuração EBM1 demonstrou ser a melhor estratégia de esquema estrela para cenários em que as consultas IOLAP são baseadas em vetores de características curtos ou médios, independentemente se o predicado visual é determinado por uma ou mais camadas perceptuais. Nestes cenários, a configuração EBM1 gerou os melhores resultados por dois motivos. Primeiro, devido ao fato de vetores de características curtos se comportarem como textos curtos, assim as estratégias tradicionais de esquema estrela (i.e., custo de junção resulta em degradação do desempenho) se aplicam de forma esperada no processamento de consultas IOLAP. Já o segundo motivo refere-se ao armazenamento dos vetores longos, pois na configuração EBM1 as dimensões que contêm os vetores de características longos não são utilizadas e consequentemente a configuração não sofre o impacto de um custo maior de paginação gerado por conta da dimensionalidade dos tuplas.

No geral o esquema EMB1 demonstrou ser o melhor candidato a iStar, pois quando a configuração EBM1 não apresentou os melhores resultados, esta

configuração gerou resultados próximos à melhor configuração de sua bateria de testes. Como observado nos experimentos baseados na seletividade dos predicados convencionais, a configuração EBM3 (a configuração que alcançou os menores tempos de resposta) gerou uma redução média sobre o EBM1 de 7,29%, enquanto para os experimentos baseados na seletividade determinada pelo raio de abrangência r_q, a configuração EBM3 gerou uma redução média sobre o EBM1 de 4,20%. Estes ganhos confirmam a indicação do EBM1 para cenários em que as consultas não seguem um padrão de composição.

Neste capítulo também foram realizadas comparações de desempenho sobre as propostas de iStar com relação aos atuais recursos de *image data warehousing*. Como observado nas baterias de testes realizadas, as nossas propostas de esquemas estrela apresentaram uma expressiva redução do tempo de processamento de consultas IOLAP mesmo quando comparado ao recurso de IDWing que produziu os melhores tempos (SingleOnion).

Para a bateria de testes baseada na seletividade dos predicados convencionais, o esquema EBM3 (melhor candidato para esta bateria de testes) gerou em média uma redução de 40,59% sobre o tempo obtido pela SingleOnion. A redução do esquema EBM3 foi ainda maior para os experimentos baseados na seletividade determinada pelo raio de abrangência, para os quais a redução de tempo média gerada sobre a SingleOnion foi de 92,94%.

Para os experimentos baseados na dimensionalidade dos vetores de características, em apenas dois casos as nossas propostas não apresentaram melhor desempenho do que a SingleOnion. Estes dois casos ocorreram para as consultas IOLAP com raio de abrangência igual a 10%, em que a similaridade entre as imagens foi determinada segundo a camada perceptual de Zernike e Histograma de níveis de cinza. O baixo desempenho de nossas propostas nestes dois casos é justificado devida à baixa precisão obtida pela mbOr. Por outro lado, para todos os demais experimentos baseados na dimensionalidade dos vetores de características, a configuração EBM1, que foi a configuração que apresentou os melhores resultados, gerou uma excelente redução média sobre a SingleOnion, variando de 55,54% até 98,21%.

Capítulo 7

CONCLUSÃO

Neste capítulo são descritas as conclusões deste trabalho de pesquisa em nível de mestrado, além de serem indicadas sugestões de trabalhos futuros.

7.1 Considerações Finais

Ambientes de *image data warehouse* são infraestruturas computacionais robustas, eficientes e desenvolvidas para prover suporte à tomada de decisão estratégica, possibilitando também a utilização de imagens digitais em consultas OLAP. O desenvolvimento e o aperfeiçoamento destas aplicações tem demonstrado ser a cada dia mais necessário devido à facilidade de geração de imagens digitais e devido ao fato de diversas áreas (como exemplo, as áreas médica, agropecuária e ambiental) tomarem decisões a partir da análise de um volume cada vez maior de dados complexos.

Um *image data warehouse* amplia as funcionalidades consolidadas e amplamente utilizadas de tecnologias de *data warehousing* a uma nova gama de consultas complexas, denominada neste trabalho por consultas IOLAP (*image on-line analytical processing)*, as quais podem ser compostas por predicados convencionais (i.e. predicados sobre os tradicionais dados de DW) assim como predicados visuais determinados por critérios de similaridade entre as imagens.

Mesmo havendo um número significativo de trabalhos na literatura que apresentam modelos conceituais e lógicos para IDW com múltiplos descritores, em nenhum destes trabalhos foi abordada a proposta de um modelo flexível a qualquer tipo de descritor ou a qualquer intenção de consulta, tão pouco foi discutido o uso de

mecanismos para acelerar o processamento de consultas IOLAP ou mesmo discutidas estratégias de esquemas estrela que proporcionem melhores resultados para o processamento de consultas IOLAP. Neste contexto, os assuntos tratados neste trabalho consistem em uma interessante contribuição para o estado da arte em *image data warehouse*.

Nesta dissertação, apresentamos um ambiente de *image data warehouse* denominado iCube, que além de manter as funcionalidades tradicionais de DWing também incorpora novas funcionalidades a ambientes de IDWing ao permitir que análises multidimensionais sejam realizadas sobre a perspectiva de predicados visuais e também por predicados convencionais. Além disso, o iCube é um ambiente flexível às intenções de consulta do usuário, pois não é restrito a um conjunto fixo de imagens de consulta tampouco restrito a um valor pré-determinado para o raio de abrangência em consultas por similaridade. Como resultado, atividades de análise e gestão de recursos podem ser realizadas de maneira ágil e organizada, além de utilizar imagens para a tomada de decisão estratégica.

O iCube aborda três grandes desafios de desenvolvimento de ambientes de IDWing, que são referentes à fase de ETL, à representação lógica do DW e ao processamento de consultas IOLAP. O esquema estrela proposto neste trabalho para o DW é caracterizado por possuir dimensões visuais dedicadas à representação de imagens segundo seu conteúdo intrínseco e por referenciar as imagens segundo representantes globais, permitindo assim que consultas complexas baseadas na similaridade entre imagens sejam executadas de maneira ágil.

Este trabalho também apresentou uma adaptação sobre o processo de transformação da fase de ETL para permitir que as imagens sejam integradas aos dados convencionais no IDW, ao mesmo tempo em que gera os dados para alimentar a tabela de dimensão "*Imagem".* O processamento de consultas IOLAP proposto neste trabalho ocorre em no máximo seis etapas, as quais foram elaboradas visando a otimização do processamento das consultas, uma vez que apresentam mecanismos de eliminação de comparações desnecessárias.

Nesta dissertação estendemos nossa proposta de esquema estrela de IDW com o intuito de permitir que as imagens presentes no IDW sejam descritas segundo diversas camadas perceptuais. Para tanto, propomos um conceito de esquema estrela, denominado iStar, adaptável a diferentes assuntos de DW e a intenções de

consulta. O iStar é um esquema estrela composto por dois tipos de tabelas de dimensão: as *dimensões convencionais* que armazenam dados tradicionais de DW (e.g. data, textos curtos e atributos numéricos) e são determinadas conforme o assunto abordado pelo IDW; e *dimensões visuais* que são organizadas conceitualmente em camadas perceptuais e contêm dados sobre o conteúdo intrínseco das imagens e dados que referenciam essas a partir de representantes globais. Por ser uma extensão do IDW do ambiente iCube, o iStar mantém o suporte a consulta IOLAP em que os critérios de similaridade (i.e. predicados visuais) são definidos pelo usuário de maneira ad-hoc, também é mantido o uso da técnica Omni como mecanismo para acelerar o processamento das consultas. Ademais, o iStar estende as funcionalidades multidimensionais de um IDW às camadas perceptuais, permitindo que as consultas IOLAP sejam elaboradas por diversas composições de predicados visuais e predicados convencionais.

Com o intuito de identificar a melhor estratégia de esquema estrela para o iStar otimizado, propomos oito configurações, que foram organizadas em dois grupos: Esquemas Baseados em Representantes (EBR) e Esquemas Baseados em mbOr (EBM). Estas configurações foram projetadas segundo diferentes estratégia de esquema estrela do DW referentes: (i) a redundância dos dados no DW *versus* o custo de junção-estrela; (ii) a granularidade definida pelas tabelas de dimensão; e (iii) a composição de atributos e medidas da tabela de fatos. Baterias de testes foram realizadas e elaboradas com o objetivo de avaliar as configurações segundo diferentes composições de predicados convencionais, diversas seletividades gerada pelo raio de abrangência, e segundo diferentes composições de predicados visuais. Estas baterias de testes também foram utilizadas para comparar as propostas de iStar aos atuais recursos de IDWing.

Devido ao fato do processamento de consultas IOLAP não ser possível de ser feito usando apenas o MAM Onion-tree, este trabalho primeiramente usou um ambiente de IDWing que provê suporte ao processamento de consultas IOLAP sem alterar a estrutura de dados da Onion-tree com o auxílio de um DW convencional. Para tanto, três métodos de processamento de consultas IOLAP foram implementados de maneira a combinar os recursos dimensionais do DW e os recursos de consulta por similaridade da Onion-Tree.

Esses métodos diferem entre si na ordem de processamento dos predicados, onde no primeiro método a consulta é realizada independentemente em cada

estrutura, isto é o predicado convencional é processado pelo DW e o predicado visual é processado pela Onion-Tree, e em seguida os resultados são combinados a fim de obter o resultado final da consulta IOLAP. O segundo método primeiro processa a consulta segundo o predicado visual na Onion-tree e depois verifica se as imagens selecionadas como similares estão de acordo com os predicados convencionais acessando o DW. O terceiro método verifica primeiro os predicados convencionais e depois verifica sobre a similaridade das imagens com relação à imagem de consulta acessando a Onion-Tree.

Além disso, foram propostas nesta dissertação adaptações para a Onion-tree, chamadas de MultiOnion e a SingleOnion, para permitir o processamento de consultas IOLAP usando apenas a Onion-tree. Alterações na estrutura de dados da Onion-tree foram realizadas e consequentemente novos algoritmos foram propostos. Dessa maneira, estes ambientes de IDWing descritos compõem a base para a comparação de desempenho com a proposta do iCube e iStar.

A partir dos resultados apresentados nos capítulos 5 e 6, foi possível observar que o ambiente de IDWing apresentado por este trabalho (i.e. iCube e iStar) proporcionou melhores resultados de desempenho do que os atuais recursos de IDWing tanto em relação ao tempo despendido para a construção do ambiente quanto ao tempo gasto no processamento de consultas IOLAP.

Sobre os experimentos realizados com o iCube em sua representação lógica inicial (Capítulo 5) e a DWOnion, o ambiente de IDWing proposto apresentou um expressivo ganho de desempenho no processamento das consultas IOLAP quando comparado ao trabalho correlato, para o qual o ganho gerado variou entre 43% e 76.70%.

As configurações do iStar também apresentaram significativos resultados no processamento de consultas IOLAP. Para os experimentos baseados na seletividade dos predicados convencionais o iStar gerou em média uma redução de 40,59% sobre o tempo obtido pela SingleOnion (i.e. versão da Onion-tree mais eficiente) e para experimentos baseados na seletividade determinada pelo raio de abrangência, o ganho médio gerado pelo iStar foi ainda maior, de 92,94% sobre o tempo obtido pela SingleOnion. Para os experimentos baseados na dimensionalidade dos vetores de características, em apenas dois casos as nossas propostas não apresentaram melhores desempenho do que a SingleOnion. Estes dois casos ocorrem para as consultas IOLAP com raio de abrangência igual a 10% em que a similaridade entre

as imagens foi determinada segundo a camada perceptual de Zernike e Histograma de níveis de cinza. O baixo desempenho de nossas propostas nestes dois casos é justificado devida à baixa precisão obtida pela mbOr. Por outro lado, para todos os demais experimentos baseados na dimensionalidade dos vetores de características, a configuração EBM1, que foi a configuração que apresentou melhores resultados para o iStar, gerou uma excelente redução média sobre a SingleOnion, variando de 55,54% a 98,21%.

7.2 Trabalhos futuros

A partir deste trabalho de pesquisa em nível de mestrado foi possível identificar diversos trabalhos futuros, como:

- Estender as operações OLAP, como *drill-down* e *roll-up*, a níveis de agregação determinados pelas imagens. Vislumbra-se abordar a definição de níveis hierárquicos baseado no conteúdo intrínseco das imagens, assim como abordar o significado semântico de cada nível de maneira que o usuário compreenda como navegar entre esses níveis, permitindo assim que as consultas sejam executadas em diferentes níveis de granularidade;
- Permitir a atualização incremental no cubo de dados, de forma que a precisão da mbOr seja mantida, ou seja, estabelecer critérios para a reconstrução do IDW;
- Desenvolver uma interface gráfica que dê suporte aos recursos de servidores OLAP a ambientes de IDWing, assim como apresentar elementos mais semânticos para a definição do raio de abrangência; e
- Testar o iCube para outras tipos de consulta por similaridade, como exemplo consultas k-NN.

REFERÊNCIAS

ARIGON, A.-M.; MIQUEL, M.; TCHOUNIKINE, A. Multimedia data warehouses: a multiversion model and a medical application. **Multimedia Tools and Applications**, v. 35, n. 1, p. 91-108, 2007.

ARIGON, A.-M.; TCHOUNIKINE, A.; MIQUEL, M. Handling Multiple Points of View in a Multimedia Data Warehouse. **ACM Transactions Multimedia Computing, Communications and Applications**, v. 2, n. 3, p. 199-218, 2006.

BANEK, M.; TJOA, A. M.; STOLBA, N. Integrating Different Grain Levels in a Medical Data Warehouse Federation. **LECTURE NOTES IN COMPUTER SCIENCE**, p. 185-194 2006.

BECKERA, K.; RUIZ, D. D. An Aggregate-Aware Retargeting Algorithm for Multiple Fact Data Warehouses. In: (Eds.). **Data Warehousing and Knowledge Discovery**. Springer Berlin / Heidelberg, 2004. p. 118-128. (Lecture Notes in Computer Science; v. 3181/2004).

BÖHM, C.; BERCHTOLD, S.; KEIM, D. A. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. **ACM Computing Surveys**, v. 33, n. 3, p. 322-373, 2001.

BOZKAYA, T.; OZSOYOGLU, M. Indexing large metric spaces for similarity search queries. **ACM Transactions On Database Systems**, v. 24, n. 3, p. 361-404, 1999.

BRIN, S. Near Neighbor Search in Large Metric Spaces. **The VLDB Journal**, p. 574-584, 1995.

CARÉLO, C. C. M. et al. Slicing the Metric Space to Provide Quick Indexing of Complex Data in the Main Memory. **Information Systems (Oxford)**, p. 2010.

_____. The Onion-Tree: Quick Indexing of Complex Data in the Main Memory In: (Eds.). **Advances in Databases and Information Systems**. Springer Berlin / Heidelberg, 2009. p. 235-252. (Lecture Notes in Computer Science; v. 5739/2009).

CHAUDHURI, S.; DAYAL, U. An overview of data warehousing and OLAP technology. **SIGMOD Rec.**, v. 26, n. 1, p. 65-74, 1997.

CHÁVEZ, E. et al. Searching in metric spaces. **ACM Computing Surveys**, v. 33, n. 3, p. 273-321, 2001.

CHEN, M. et al. Multimedia database retrieval based on data cube. In: INTERNATIONAL CONFERENCE ON AUDIO, LANGUAGE AND IMAGE, 2008, Shanghai. **Proceedings.** 2008. p. 1265-1269.

CIACCIA, P.; PATELLA, M. Searching in metric spaces with user-defined and approximate distances. **ACM Transactions On Database Systems**, v. 27, n. 4, p. 398-437, 2002.

CIFERRI, C. D. A. et al. Data Warehousing na Saúde: Melhorando a Tomada de Decisão Médico-Analítica. In: XXXII CONFERENCIA LATINOAMERICANA DE INFORMÁTICA, 2006, Santiago, Chile. **Proceedings.** Memorias da XXXII Conferencia Latinoamericana de Informática (CLEI 2006), 2006. p. 354.1-354.12.

CIFERRI, C. D. D. A. **Distribuição dos Dados em Ambientes de Data Warehousing:** O Sistema WebD2W e Algoritmos Voltados à Fragmentação Horizontal dos Dados. 263 f. Tese de doutorado em Ciência da Computação – Centro de Informática, Universidade Federal de Pernambuco, 2002.

CIFERRI, C. D. D. A. et al. Horizontal fragmentation as a technique to improve the performance of drill-down and roll-up queries. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2007, Seoul, Korea. **Proceedings.** ACM, 2007. p. 494-499.

DATTA, R. et al. Image retrieval: Ideas, influences, and trends of the new age. **ACM Computing Surveys**, v. 40, n. 2, p. 1-60, 2008.

ELMASRI, R. E.; NAVATHE, S. **Sistemas de Banco de Dados** Addison-Wesley, 2005. p.

FELIPE, J. C. **Desenvolvimento de Métodos para Extração, Comparação e Análise de Características Intrínsecas de Imagens Médicas, Visando à Recuperação Perceptual por Conteúdo**. 176 f. Doutorado em Ciências de Computação e Matemática Computacional – ICMC-USP, Universidade de São Paulo, São Carlos, 2005.

FELIPE, J. C.; JR., C. T.; TRAINA, A. J. M. A New Family of Distance Functions for Perceptual Similarity Retrieval of Medical Images. **Journal Digital Imaging**, v. 22, p. 183-201, 2009.

FELIPE, J. C.; TRAINA, A. J. M.; TRAINA, C. J. Global Warp Metric Distance: Boosting Content-based Image Retrieval through Histograms. In: SEVENTH IEEE INTERNATIONAL SYMPOSIUM ON MULTIMEDIA, 2005, Irvine, California. **Proceedings.** 2005. p. 295-302.

FIDALGO, R. N. et al. GeoDWFrame: A Framework for Guiding the Design of Geographical Dimensional Schemas In: (Eds.). **Data Warehousing and Knowledge Discovery**. Springer Berlin / Heidelberg, 2004. p. 26-37. (Lecture Notes in Computer Science; v. 3181/2004).

FILHO, R. F. S. et al. Similarity search without tears: the OMNI-family of all-purpose access methods. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 17th 2001, **Proceedings.** 2001. p. 623-630.

FORLANI, D. T.; CIFERRI, C. D. D. A.; CIFERRI, R. R. Melhorando o Desempenho do Processamento de Consultas Drill-Across em Ambientes de Data Warehousing.

In: XXI SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 2006, Florianópolis, Santa Catarina, Brasil. **Proceedings.** UFSC, 2006. p. 161-175.

GARCIA-MOLINA, H.; ULLMAN, J. D.; WIDOW, J. **Database Systems Implementation** Upper Saddle River, New Jersey, USA: Prentice Hall: , 2000.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 3rd ed. Prentice Hall, 2008. 1-976 p.

GRAY, J. et al. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. In: ICDE, 12th, 1995, **Proceedings.** Institute of Electrical and Electronics Engineers, Inc., 1995. p. 152-159.

HAN, J. et al. Efficient computation of Iceberg cubes with complex measures. **ACM SIGMOD Record**, v. 30, n. 2, p. 1-12, 2001.

HARALICK, R. M.; SHANMUGAN, K.; DINSTEIN, I. Textural Features for Images Classification. In: IEEE Transactions on Systems, Man and Cybernetics, 1973, **Proceedings.** 1973. p. 610-621.

INMON, W. H. **Building the Data Warehouse**. 4th ed. Indianapolis: Wiley Publishing, 2005. 1-576 p.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit**. 2nd ed. New York: Wiley Computer Publishing, 2002. 1-436 p.

LAI, S. Y.; NICOLLET, P. A clinical data warehouse effort to support the measurable organizational patient care and health outcomes improvement initiative. In: AMIA SYMPOSIUM, 2000, **Proceedings.** 2000. p.

LEVENSHTEIN, V. I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. **Soviet Physics Doklady**, v. 10, p. 707-710, 1966.

MURPHY, S. Data Warehousing for Clinical Research. In: AND, L. L.; ÖZSU, M. T. (Eds.). **Encyclopedia of Database Systems**. Springer US, 2009. p. 679-684.

MURPHY, S. N. et al. Optimizing healthcare research data warehouse design through past COSTAR query analysis. In: AMIA SYMPOSIUM, 1999, **Proceedings.** 1999. p. 892-896.

NEUMUTH, T. et al. Data Warehousing Technology for Surgical Workflow Analysis. **Computer-Based Medical Systems, IEEE Symposium on**, v. 0, p. 230- 235, June 17-June 19, 2008.

PEDERSEN, T. B.; JENSEN, C. S. Research Issues in Clinical Data Warehousing. In: INTERNATIONA CONFERENCE ON SCIENTIFIC AND STATISTICAL DATABASE MANAGEMENT, 1998, **Proceedings.** IEEE Computer Society, 1998. p. 43-52.

PONCIANO-SILVA, M. et al. Including the Perceptual Parameter to Tune the Retrieval Ability of Pulmonary CBIR Systems. **Computer-Based Medical Systems. 22nd IEEE International Symposium on**, p. 1-8, 2009, 2009.

RIZZI, S. Conceptual Modeling Solutions for the Data Warehouse. In: KONCILIA, W. R. E. C. (Eds.). **Data Warehouse and OLAP:** Concepts, Architectures and Solutions. Hershey, Pennnsylvania, USA: IRM Press, 2006. p. 7-9. cap. 1.

SIQUEIRA, T. L. L. **SB-Index: Um Índice Espacial baseado em Bitmap para Data Warehouse Geográfico** SB-Index: Um Índice Espacial baseado em Bitmap para Data Warehouse Geográfico. 120 f. – Programa de Pós-Graduação em Ciência da Computação, Universidade Federal de São Carlos, São Carlos, 2009.

SIQUEIRA, T. L. L. et al. The Impact of Spatial Data Redundancy on SOLAP Query Performance. **Journal Of The Brazilian Computer Society (Impresso)**, v. 15, p. 19-34, 2009.

SONG, I.-Y. Data Warehousing Systems: Foundations and Architectures. In: AND, L. L.; ÖZSU, M. T. (Eds.). **Encyclopedia of Database Systems**. Springer US, 2009. p. 684-692.

SOUZA, M. F. D.; SAMPAIO, M. C. Efficient materialization and use of views in data warehouses. **ACM SIGMOD Record**, v. 28, n. 1, p. 78-83, 1999.

TRAINA-JR, C. et al. Fast Indexing and Visualization of Metric Data Sets using Slim-Trees. **IEEE Trans. on Knowl. and Data Eng.**, v. 14, n. 2, p. 244-260, 2002.

TRAINA-JR., C. et al. The Omni-family of all-purpose access methods: a simple and effective way to make similarity search more efficient. **The VLDB Journal**, v. 16, n. 4, p. 483-505, 2007.

_____. Fast Feature Selection using Fractal Dimension. In: XV Brazilian Database Symposium (SBBD), 2000, João Pessoa, Brasil. **Proceedings.** 2000. p. 158-171.

TRAINA, A. J. M. et al. The Metric Histogram: A New and Efficient Approach for Content-based Image Retrieval. In: PU., X. Z. P. (Eds.). **Visual and Multimedia Information Management**. Norwell, MA: Kluwer Academic Press, 2002. p. 297-311. v. 1).

_____. How to Cope with the Performance Gap in Content-based Image Retrieval Systems. **International Journal of Healthcare Information Systems and Informatics (IJHISI)**, v. 4, p. 47-67, 2009. Disponível em: <http://www.igi-pub.com/articles/details.asp?id=32968>.

VIEIRA, M. R. et al. DBM-Tree: A Dynamic Metric Access Method Sensitive to Local Density Data. In: XIX SIMPÓSIO BRASILERIO DE BANCO DE DADOS, 2004, Brasilia, Distrito Federal, Brasil. **Proceedings.** UnB, 2004. p. 163-177.

WAH, T. Y.; SIM, O. S. Development of a data warehouse for Lymphoma cancer diagnosis and treatment decision support. **WSEAS Trans. Info. Sci. and App.**, v. 6, n. 3, p. 530-543, 2009.

WANG, J. Z. et al. Diversity in multimedia information retrieval research. In: ACM INTERNATIONAL WORKSHOP ON MULTIMEDIA INFORMATION RETRIEVAL, 2006, Santa Barbara, California, USA. **Proceedings.** ACM, 2006. p. 5-12.

WONG, S. T. C. et al. A neuroinformatics database system for disease-oriented neuroimaging research. **Academic Radiology**, v. 11, n. 3, p. 345-358, 2004. Disponível em: <http://www.sciencedirect.com/science/article/B75BK-4C0HM0B-13/2/a5f6d55501498946d310ca7fbbc88d90>.

XÉXEO, G.; SANTOS, A. C. O. G. FBCDataWare: Um Ambiente de Dados Integrados para Cardiologia. In: INTERNATIONAL CONGRESS ON INFORMATIC ENGINEERING, 2000, Buenos Aires, Argentina. **Proceedings.** 2000. p. 1-5.

YIANILOS, P. N. Data structures and algorithms for nearest neighbor search in general metric spaces. In: FOURTH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS, 1993, Austin, Texas, United States. **Proceedings.** Society for Industrial and Applied Mathematics, 1993. p.

YOU, J. J. et al. On Hierarquical Content-based Image Retrieval by Dynamic and Guided Search. In: IEEE INTERNATIONAL CONFERENCE ON COGNITIVE INFORMATICS, 8th, 2004, Hong Kong, China. **Proceedings.** IEEE Computer Society, 2004. p. 188-195.

ZAÏANE, O. R. et al. MultiMediaMiner: a system prototype for multimedia data mining. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1998, Seattle, United State. **Proceedings.** 1998. p. 581-583.

Apêndice A

Ilustrações das Demais Árvores Presentes na MultiOnion

As figuras a seguir ilustram as árvores DataOnion-Tree, ExameOnion-Tree, HospOnion-Tree e PacOnion-Tree projetados para o ambiente MultiOnion do cenário exemplo descrito na Seção 4.3.

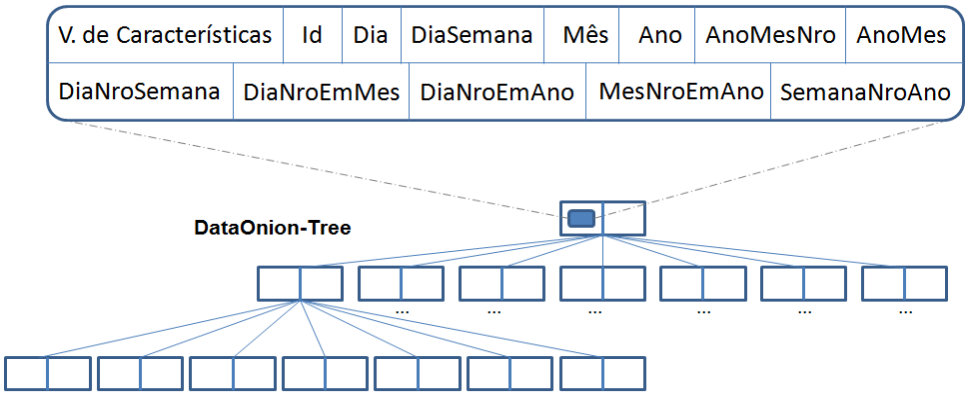

Figura 7.1: Onion-Tree adaptada para o armazenamento de dados sobre a dimensão Data.

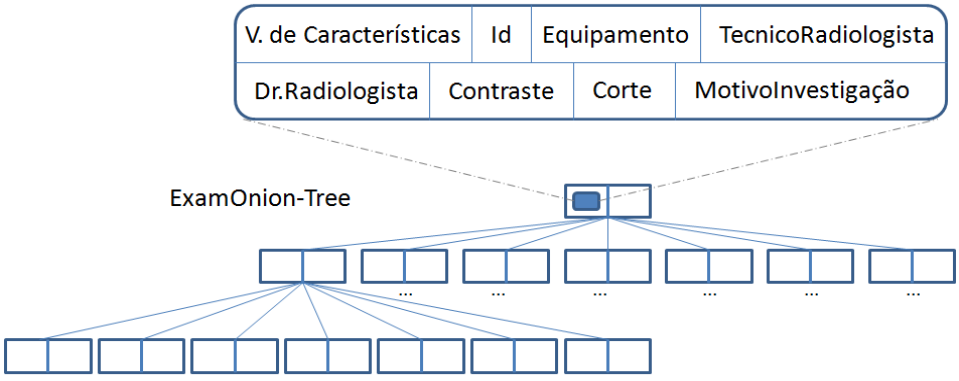

Figura 7.2: Onion-Tree adaptada para o armazenamento de dados sobre a dimensão Exame.

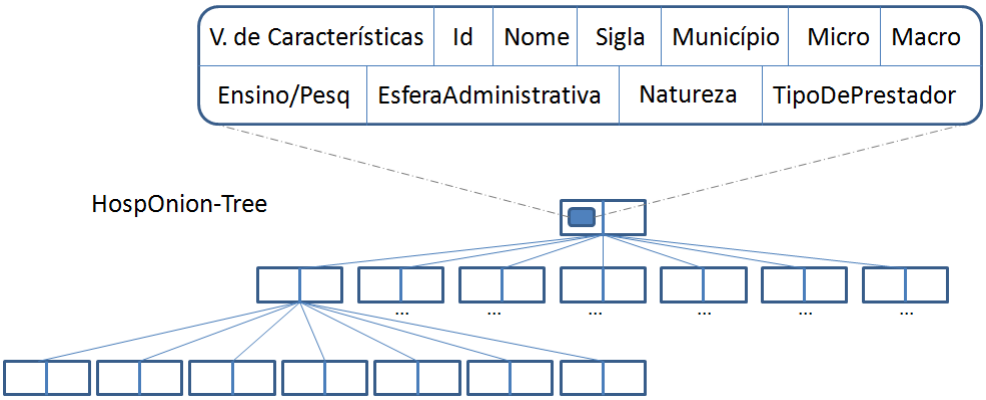

Figura 7.3: Onion-Tree adaptada para o armazenamento de dados sobre a dimensão Hospital.

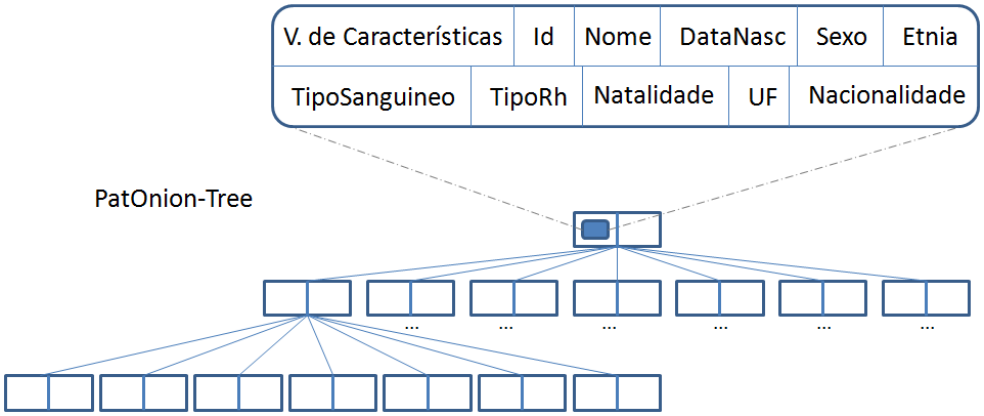

Figura 7.4: Onion-Tree adaptada para o armazenamento de dados sobre a dimensão Paciente.