

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



**UMA PROPOSTA DE CONSTRUÇÃO DE ÍNDICE
INVERTIDO PARA RECUPERAÇÃO DE IMAGENS
BASEADA EM CONTEÚDO**

TAULLER AUGUSTO DE ARAÚJO MATOS

Uberlândia - Minas Gerais

2009

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO



TAULLER AUGUSTO DE ARAÚJO MATOS
Orientador: PROF. DR. ILMÉRIO REIS DA SILVA
Co-orientadora: PROFA. DRA. CELIA A. ZORZO BARCELOS

**UMA PROPOSTA DE CONSTRUÇÃO DE ÍNDICE
INVERTIDO PARA RECUPERAÇÃO DE IMAGENS
BASEADA EM CONTEÚDO**

Dissertação de Mestrado apresentada à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como parte dos requisitos exigidos para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Banco de Dados.

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada “**Uma Proposta de Construção de Índice Invertido para Recuperação de Imagens Baseada em Conteúdo**” por **Tauller Augusto de Araújo Matos** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

Uberlândia, 13 de Fevereiro de 2009

Orientador:

Prof. Dr. Ilmério Reis da Silva
Universidade Federal de Uberlândia/Minas Gerais

Co-orientadora:

Profa. Dra. Celia A. Zorzo Barcelos
Universidade Federal de Uberlândia/Minas Gerais

Banca Examinadora:

Prof. Dr. Ricardo da Silva Torres
Universidade Estadual de Campinas UNICAMP/São Paulo

Profa. Dra. Denise Guliato
Universidade Federal de Uberlândia UFU/Minas Gerais

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Data: Fevereiro de 2009

Autor: **Tauller Augusto de Araújo Matos**
Título: **Uma Proposta de Construção de Índice Invertido para Recuperação de Imagens Baseada em Conteúdo**
Faculdade: **Faculdade de Computação**
Grau: **Mestrado**

Fica garantido à Universidade Federal de Uberlândia o direito de circulação e impressão de cópias deste documento para propósitos exclusivamente acadêmicos, desde que o autor seja devidamente informado.

Autor

O AUTOR RESERVA PARA SI QUALQUER OUTRO DIREITO DE PUBLICAÇÃO DESTE DOCUMENTO, NÃO PODENDO O MESMO SER IMPRESSO OU REPRODUZIDO, SEJA NA TOTALIDADE OU EM PARTES, SEM A PERMISSÃO ESCRITA DO AUTOR.

Agradecimentos

Agradeço a Deus pela força e inspiração nos momentos mais difíceis.

Mamãe, como seria a minha vida sem a senhora? Não há palavras na língua, literatura suficientes para descrever todo o amor e apoio que você me deu.

Ao professor Ilmério, que com sua calma e tranquilidade, conseguiu me passar muito conhecimento nas horas mais próprias.

José Carlos, além da ajuda financeira que foi primordial a compaixão e a adoção inexplicáveis.

Carol, não tem como deixar de agradecer essa pessoa maravilhosa que colocou rumo na minha vida. Muito apoio em todos os momentos que passamos juntos.

Às minhas irmãs que ajudaram a segurar a barra nesses anos longe de casa.

À professora Célia, pelos conselhos e dicas preciosas durante o andamento deste trabalho.

À professora Sandra, que me fez crescer muito como profissional.

À minha tia Cida, que sempre me mostrou alegria e orgulho por este projeto de vida.

Ao Stéfano, por sofrer e sorrir junto comigo nos momentos tristes e alegres destes quase 2 anos. Além de todo apoio com seus conhecimentos.

À Patrícia, por estar sempre que possível ali comigo. Sofrendo nas implementações.

À professora Alessandra, por sempre ter acreditado em mim.

À CAPES pelo apoio financeiro.

A todos os amigos do Mestrado em Ciência da Computação da UFU. Ajudas em todos os sentidos, até quando vocês me matavam de raiva quando eu queria saber uma coisa e vocês respondiam: - Olha no Google!. Mesmo assim, foi um prazer ter convivido com vocês neste período. Espero não esquecer de ninguém, este é o risco, mas vamos tentar: Adriano Fiad Farias, Cricia Zilda Felicio, Eduardo Ferreira Ribeiro, Éverton Hipólito de Freitas, Felipe César de Castro Antunes, Jean Carlo de Souza Santos, Juliana de Fátima Franciscani, Italo Tiago da Cunha, Klérison Vinícius Ribeiro da Paixão, Liliane do Nascimento do Vale, Lúcio Agostinho Rocha, Marcos Roberto Ribeiro, Mirella Silva Junqueira, Núbia Rosa da Silva, Ricardo Soares Boaventura, Robson da Silva Lopes, Rodrigo Reis Pereira, Valquíria Aparecida Rosa Duarte, Vinícius Borges Pires, Victor Sobreira, Waldecir Pereira Junior.

Enfim, a todos os meus amigos, familiares e professores que direta e indiretamente contribuíram para que mais este objetivo de minha vida possa ter sido vencido.

Muito Obrigado a todos!

Um beijo do Panda...

*"O soldado da paz nunca pode ser derrotado, ainda que a guerra pareça perdida, pois
quanto mais se sacrifica a vida mais a vida e o tempo são os seus aliados"*
(Herbert Vianna)

Resumo

Este trabalho apresenta uma proposta de construção do Índice Invertido para recuperação de imagens baseado em conteúdo (CBIR). O objetivo é acelerar o processamento de consultas, sem perda de qualidade na resposta. Para avaliar a eficácia da proposta foram desenvolvidos dois sistemas de recuperação de imagens. O primeiro sistema, usado como base de comparação, utiliza técnicas de extração de características e cálculo de similaridade, normalmente utilizadas em CBIR, a saber, momentos de cor e distância Euclidiana. Esta abordagem apresenta como problema, o baixo desempenho no tempo de processamento de grandes coleções de imagens como por exemplo, a Web. Para obter um melhor desempenho no processamento de consultas, é proposto um sistema que indexa as imagens utilizando-se de uma estrutura muito utilizada na recuperação textual, conhecida como índice invertido. Outra mudança deste sistema, se refere ao cálculo da similaridade. Para permitir o uso do índice invertido como acelerador do processo de cálculo do *ranking* a similaridade entre a imagem de consulta e a coleção de imagens é medida por meio do cálculo do cosseno entre vetores representantes da imagem de consulta e das imagens do banco de dados. Foram feitas duas análises: a avaliação da qualidade de recuperação e análise do número de operações aritméticas efetuadas pelos dois sistemas no cálculo da similaridade. É mostrado o ganho significativo no número de operações aritméticas sem perda significativa na qualidade de recuperação.

Palavras chave: Recuperação de informação, Recuperação de imagens, índice invertido, modelo vetorial, palavras comuns, peso do termo, coseno, distância euclidiana.

Abstract

This work shows an approach to construct an inverted index for Content-Based Image Retrieval. The objective is to speed up the process of querying without loss of quality in the result. To evaluate the effectiveness of the proposal, two systems for retrieval of images were developed. The first system was used as a basis for comparison. It uses techniques for extraction of features and for computing similarity of images usually used in CBIR, namely color moments and Euclidian distance. This approach presents problems such as the poor performance in processing time of large collections of images, such as the Web. To improve performance in the querying process, we proposed a system that indexes the images through a structure often used in textual retrieval, known as inverted index. We also propose the use of a cosine-base distance function to compute the similarity among a given query image and database images. We conducted two tests: the quality of retrieval, and the analysis of the number of arithmetic operations performed by the two systems the calculation of similarity. It is shown significant gain in the number of arithmetic operations reduction without significant loss in quality of retrieved.

Keywords: Information retrieval, image retrieval, inverted index, vector space, stop words, term weighting, cosine, euclidean distance.

Sumário

Lista de Figuras	x
Lista de Tabelas	xii
Lista de Abreviaturas e Siglas	xiii
1 Introdução	1
2 Recuperação de Informação	5
2.1 Introdução aos Modelos Clássicos	7
2.2 Modelo Booleano	8
2.3 Índice Invertido	9
2.4 Stop words	13
2.5 Peso do Termo	13
2.6 Modelo Vetorial	15
2.7 Modelo Probabilístico	18
2.8 Medidas de Avaliação	20
2.9 Conclusão	22
3 Recuperação de Imagens	24
3.1 Definindo uma Imagem	24
3.2 Sistemas de Recuperação de Imagens	25
3.3 Recuperação de Imagens Baseada em Anotações Textuais	26
3.4 Recuperação de Imagens Baseada no Conteúdo	27
3.4.1 Extração de Característica	28
3.4.2 Propriedades da Imagem	29
3.5 Vetor de Característica	36
3.6 Medidas de Similaridade	36
3.7 Avaliação dos Resultados	38
3.8 Considerações Finais	39

4	Um Estudo da Característica Cor para Construção do Índice Invertido	40
4.1	Trabalhos Correlatos	41
4.2	CBIR-I	44
4.2.1	Análise dos vetores de características do CBIR-I	45
4.2.2	Distribuição Normal	48
4.3	CBIR-II	51
4.3.1	Indexação das imagens - Índice Invertido	51
4.3.2	Cálculo da Similaridade	54
4.4	Considerações Finais Sobre a Proposta	55
5	Experimentos e Resultados	57
5.1	Avaliação Experimental no Corel-1000 - Coleção de Treinamento	57
5.2	Avaliação Experimental no BD-10000 - Coleção de teste	60
5.3	Avaliação Experimental do número de operações Aritméticas	62
5.4	Considerações Finais	62
6	Conclusão e Perspectivas Futuras	70
	Referências Bibliográficas	72
A	Tabela Completa da Distribuição Normal Padronizada	77

Lista de Figuras

2.1	Uma taxonomia de modelos de RI (Adapatado de [Baeza-Yates and Ribeiro-Neto, 1999])	7
2.2	As duas partes que formam o Índice Invertido, a saber: dicionário e lista invertida	10
2.3	Indexação dos termos, parte um	11
2.4	Indexação dos termos, parte dois	12
2.5	Frequência na coleção (cf) e frequência de documentos (df)	14
2.6	Representação vetorial de um documento com dois termos de indexação .	16
2.7	Representação vetorial de um documento com três termos de indexação .	16
2.8	Espaço vetorial contendo dois documentos	16
2.9	Representação de uma expressão de busca em um espaço vetorial	17
2.10	Subconjunto de documentos após a execução de uma busca relativa a uma consulta q [Ferneda, 2003].	19
3.1	Fluxograma de um sistema CBIR típico.	27
3.2	Espaço RGB representado em Cubo.	30
3.3	Espaço de cor HSV [Gonzales et al., 2004].	32
3.4	Imagens com textura (www.fotosearch.com.br)	34
3.5	Taxinomia da característica forma e suas técnicas [Zhang and Lu, 2004]. .	35
4.1	Consulta por "Puma" no site Google retorna um resultado misturado de imagens.	43
4.2	Amostra do Banco de Dados Corel 1000.	46
4.3	Distribuição de ocorrência das 9 característica após a extração da característica cor, momentos de cor, da base de dados corel-1000. O eixo x corresponde ao valor da característica e o eixo y à frequência de cada valor.	46
4.4	Amostra do Banco de Dados BD-10000.	47
4.5	Distribuição de ocorrência das 9 característica após a extração da característica cor, momentos de cor, da base de dados BD-10000. O eixo x corresponde ao valor da característica e o eixo y à frequência de cada valor.	47

4.6	Distribuição adotada para classificação das faixas. \bar{X} é a média e s o desvio padrão da amostra.	50
4.7	Vetor de característica gerado pelo CBIR-I correspondente a imagem Africa1.jpg.	51
4.8	Vetor de característica gerado pelo CBIR-II correspondente a imagem Africa1.jpg.	53
4.9	Representação esparsa do vetor de característica do CBIR-II correspondente a imagem Africa1.jpg.	53
4.10	(a) Banco de dados de Característica (b) Estrutura de busca (vocabulário) e a lista invertida.	54
5.1	Resultados de busca obtido para a categoria Dinossauro, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	64
5.2	Resultados de busca obtido para a categoria ônibus, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	65
5.3	Resultados de busca obtido para a categoria praia, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	66
5.4	Resultados de busca obtido para a categoria montanha, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	67
5.5	Resultados de busca obtido para a categoria avião, no banco de dados BD-10000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	68
5.6	Resultados de busca obtido para a categoria avião, no banco de dados BD-10000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.	69
A.1	Distribuição Normal Padronizada	78

Lista de Tabelas

2.1	Tabela de contingência	21
4.1	Descrição do vetor de característica resultante da extração de característica da imagem usando os três momentos do espaço de cor HSV	45
4.2	Obtendo uma área sob a curva normal	49
4.3	Construção do vetor de Características do CBIR-II	52
5.1	Precisão média no sistema CBIR-I. Coleção Corel-1000.	59
5.2	Precisão média no sistema CBIR-IIa. Coleção Corel-1000.	59
5.3	Precisão média no sistema CBIR-IIb. Coleção Corel-1000.	59
5.4	Precisão média no sistema CBIR-I. Coleção BD-10000.	61
5.5	Precisão média no sistema CBIR-IIa. Coleção BD-10000.	61
5.6	Precisão média no sistema CBIR-IIb. Coleção BD-10000.	61

Lista de Abreviaturas e Siglas

BD	Banco de dados
CBIR	Recuperação de Imagens Baseada no Conteúdo
cf	Frequência na coleção
df	Frequência de documentos
ID	Identificador do documento
IDF	Frequência Inversa dos Documentos
fn	Falso negativo
fp	Falso positivo
HSV	Tonalidade, Saturação e Brilho
LSI	Indexação semântica latente
NR	Conjuntos dos documentos não relevantes e não recuperados
PLN	Processamento de Linguagem Natural
Rec	Conjunto dos documentos recuperados
Rel	Conjunto dos documentos relevantes
RGB	Vermelho, Verde e Azul
RI	Recuperação de Informação
RR	Conjunto dos documentos relevantes recuperados
SGBD	Sistemas de Gerenciamento de Banco de Dados
SRI	Sistemas de Recuperação de Informação
TF	Frequência do termo
tn	Verdadeiro negativo
tp	Verdadeiro positivo

Capítulo 1

Introdução

Com o avanço tecnológico dos computadores, aumento do número e qualidade das conexões de Internet e com a capacidade de armazenamento, cada vez maior, a informação na Web tem se expandido com uma diversidade enorme tanto em forma como em conteúdo. O crescimento desenfreado do número de publicações, a liberdade de organização das informações e o número de usuários fizeram da Internet uma grande fonte de informações heterogêneas. Um usuário específico está interessado apenas em uma pequena parcela destas informações. Sendo assim, ele precisa de formas efetivas de acesso, que garantam que as informações para ele disponibilizadas estejam dentro de sua área de interesse. Este problema é uma motivação para a área de Recuperação da Informação (RI), passando esta a ter uma grande importância no campo da ciência da computação. Ela estuda o armazenamento e a recuperação automática de documentos. Uma definição interessante de RI encontra-se em [Manning et al., 2007]:

"Recuperação de Informação é uma tarefa de encontrar informações relevantes (usualmente documentos) de uma natureza não estruturada (geralmente texto) que satisfazem uma necessidade de informação dentro de grandes coleções (que podem ser em servidores de computadores locais ou na Internet)."

Outra definição de RI é encontrada em [Baeza-Yates and Ribeiro-Neto, 1999]: RI consiste na recuperação, representação, armazenamento, organização e acesso de documentos. Assim RI, consiste em identificar em uma coleção de dados, quais documentos atendem à necessidade de informação do usuário. O problema central de RI é encontrar estas informações úteis e relevantes para o usuário. Para resolver este problema, a principal ferramenta utilizada é o emprego de sistemas de recuperação de informação (SRI). Inicialmente estes eram utilizados em sistemas bibliotecários e em sistemas com pequena coleção de documentos. Com a expansão da Web no início dos anos 90, SRI, ganhou um espaço enorme e contribuiu para a criação das máquinas de busca na Web.

Os usuários destes sistemas ou de máquinas de busca na Web normalmente expressam sua necessidade de informação via uma consulta. Esta pode ser especificada por um conjunto de palavras-chaves, que também são denominados de termos, utilizados para

recuperar documentos em uma coleção.

A Web apresenta características exclusivas como: a maioria dos dados serem voláteis, documentos são constantemente modificados, removidos ou adicionados. Outra característica importante se dá pelo fato dos documentos não conterem apenas textos, mas também vídeos, imagens, etc. Recuperação de imagens por conteúdo é o foco deste trabalho.

Da mesma forma, que os documentos, nos últimos anos tem-se visto um rápido aumento no tamanho das coleções de imagens digitais. Com o surgimento da *Internet* tornou-se possível o acesso e disponibilização de imagens em todas as partes do mundo, seja qual for sua finalidade. Imagens são produzidas por satélites, inspeção militares e operações de vigilância, sistemas biométricos (impressões digitais, faces, etc), experimentos científicos, sistemas de entretenimento e informação, entre outros. Há também várias formas de busca de imagens.

Dentre os modelos de recuperação de imagens o primeiro apresentado neste trabalho é o de recuperação de imagens baseada em anotações textuais. Este utiliza anotações manuais para descrever o conteúdo das imagens. Neste caso, consultas são expressas por palavras chaves e efetuadas por técnicas de gerenciamento de banco de dados. Esta abordagem é limitada, pois é inviável prover anotações de imagens para grandes coleções, além disso, anotações manuais são subjetivas - uma mesma imagem pode ser interpretada de diferentes maneiras por diferentes pessoas.

Outra abordagem muito utilizada em recuperação de imagens é o serviço provido pelo *site* de busca Google. Neste sistema as imagens são extraídas de documentos e são indexadas com base nas informações textuais ou rótulos que as acompanham. Para que, no processo de recuperação das imagens, possam ser utilizadas as técnicas tradicionais da recuperação de informação textual. As consultas aqui, também são expressas por palavra-chave. Esta abordagem apresenta um bom desempenho em tempo de processamento, mas uma baixa eficiência na qualidade de recuperação, proporcionada pelo fato de que na maioria das vezes o texto próximo a uma imagem não a descreve fielmente.

Por fim, uma técnica bastante utilizada para recuperação de imagens é conhecida como Recuperação de Imagens Baseada no Conteúdo (CBIR). Seu principal objetivo é encontrar imagens relevantes conforme a necessidade do usuário, por meio de características visuais automaticamente extraídas das imagens. Dentro do escopo das características visuais, estas podem ser classificadas como características gerais e características de domínio. A primeira inclui características principalmente do tipo cor, forma e textura. Para tornar a recuperação de imagens mais precisa estas podem ser combinadas em um vetor de características. As características de domínio são dependentes da aplicação, incluindo características específicas, por exemplo, faces humanas e impressões digitais em sistemas de reconhecimento de criminosos. Estas características são cobertas em artigos com temas relacionados a reconhecimento de padrões e podem incluir temas referentes ao domínio do conhecimento [Castañón, 2003]. Este trabalho abordará as características

visuais genéricas baseada em cor.

CBIR apresenta resultados melhores em relação às técnicas de recuperação de imagens baseadas em anotações textuais, mas apresentam problemas em relação ao tempo de processamento para grandes coleções. Este problema é consequente do cálculo da medida de similaridade. Uma vez que as características das imagens tenham sido extraídas e armazenadas em vetores de característica, fazem-se necessárias medidas que comparem vetores de característica das imagens do banco de dados com o vetor de característica da imagem exemplo. Essas medidas são normalmente baseadas na distância entre vetores das imagens. Muitos sistemas de recuperação de imagens fazem uso desta forma de medida de similaridade o que torna o processo de recuperação lento para grandes bases de dados, como por exemplo, a Web. Isto ocorre porque o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os vetores das imagens do banco de dados.

Visando acelerar este processo de recuperação propõe-se uma nova abordagem, tanto para indexação das imagens quanto para o cálculo da similaridade em CBIR. Pretende-se utilizar técnicas da recuperação textual, a saber, índice invertido, modelo vetorial, peso do termo juntamente com a medida de similaridade do cosseno com o intuito de acelerar o processo da consulta sem perda de qualidade na recuperação.

O objetivo geral deste trabalho é propor e avaliar experimentalmente um método para classificar grupos de valores de características de baixo nível de imagens digitais, baseado na característica cor, que possibilite o mapeamento para um identificador de indexação da coleção permitindo ganho de tempo no processamento da consulta sem perda de qualidade na recuperação.

O conteúdo deste trabalho é dividido em 6 capítulos.

No capítulo 2 são descritos os principais conceitos de RI textual, incluindo os modelos clássicos: booleano, probabilístico e vetorial. Este último é utilizado neste trabalho. São descritas as técnicas empregadas no desenvolvimento desta dissertação, tais como índice invertido, peso do termo e *stop words*. Além disso são apresentadas algumas medidas de avaliação para quantificar a qualidade dos Sistemas de Recuperação de Imagem.

No capítulo 3 são descritos conceitos para a recuperação de imagens por conteúdo. É mostrada a diferença dos sistemas de recuperação de imagens baseada em anotações textuais e recuperação de imagens baseadas em conteúdo. Esta segunda técnica é abordada com mais detalhes. Cor, textura e forma são descritos como as principais propriedades visuais usadas no processo de caracterização de imagens. Vetores de características e medidas de similaridade são abordados por serem importantes para o entendimento deste trabalho.

Os capítulos 2 e 3 apresentam uma revisão de literatura em áreas correlatas as presentes neste trabalho. A segunda parte desta dissertação é dedicada à descrição da técnica proposta, análise e comparações dos resultados obtidos, bem como propostas de trabalhos

futuros:

No capítulo 4 são apresentados os dois sistemas implementados nesta dissertação. Aqui chamados de CBIR-I e CBIR-II. O primeiro sistema foi desenvolvido com as técnicas tradicionais de CBIR. O segundo método, proposto neste trabalho, considera modificações no processo de indexação e no cálculo de similaridade.

No capítulo 5 são apresentadas as avaliações experimentais: comparação do CBIR-II com o CBIR-I a fim de verificar se não houve perda na qualidade da recuperação. Esta comparação será medida utilizando-se a medida de Precisão. Também é comparado o número de operações aritméticas realizada em cada um destes sistemas. O bom desempenho do CBIR-II é mostrado via uma série de experimentos.

Por fim, no capítulo 6, são apresentadas as conclusões gerais deste trabalho, uma discussão referente a vantagem de se utilizar o índice invertido no processo de indexação e recuperação das imagens por conteúdo e algumas direções para pesquisas futuras.

Capítulo 2

Recuperação de Informação

O primeiro passo para a compreensão de métodos para Recuperação de Informação é entender a diferença entre um SRI com Sistemas de Gerenciamento de Banco de Dados (SGBD).

Um SGBD é uma coleção de programas que possibilita que os usuários criem e mantenham bancos de dados. Ou seja, é um sistema cujo objetivo principal é gerenciar o acesso e a correta manutenção dos dados armazenados em bancos de dados [Elmasri and Navathe, 2000]. Manipular um SGBD inclui funções de como fazer consultas ao banco de dados para recuperar dados específicos.

Uma das principais diferenças entre Sistemas de Recuperação de Informação e Sistemas de Gerenciamento de Banco de Dados, esta no foco da consulta do usuário. O primeiro recupera informações sobre um determinado assunto, já o segundo recupera dados que satisfazem a uma expressão de busca. SGBD tem por objetivo recuperar todos objetos ou itens que satisfazem precisamente às condições formuladas por meio de uma expressão de busca, na qual os dados estão estruturados por meio de uma semântica bem definida. Já em RI, esta precisão não é tão rigorosa, pelo fato dos SRI tratarem com objetos linguísticos (textos), e herdarem toda a problematicidade do tratamento da linguagem natural [Ferneda, 2003]. O objetivo de um SRI, é recuperar informação útil e relevante para o usuário. A ênfase é dada para a recuperação da informação contida em documentos não estruturados, e não para a recuperação de dados, armazenados em arquivos estruturados.

O usuário de um SRI, expressa a sua necessidade de informação em uma consulta. Tradicionalmente, as consultas são expressas por meio de palavras-chaves, que são utilizadas por meio de heurísticas para recuperar documentos em uma coleção. Após a consulta, SRIs devem recuperar documentos de uma maneira rápida e estes documentos devem satisfazer a necessidade de informação do usuário. A recuperação destes documentos deve seguir uma ordem de relevância. Relevância indica a importância de um documento para uma consulta sendo um componente chave para determinar a ordem de representação dos documentos, tradicionalmente chamada de *ranking*.

Especificamente na Web, um novo ambiente digital é observado, com uma gama de materiais digitais como: textos, imagens, sons, vídeos, páginas Web, que requerem diferentes tipos de tratamento e representação para uma recuperação de informação eficaz. Observa-se que os documentos não são formados apenas por textos. Outra particularidade da Web se dá pelo fato dos dados serem voláteis, estes são modificados constantemente, removidos ou adicionados. Em relação aos usuários da Web, estes não são especializados e possuem interesse diversificado, o que torna o processo de recuperação mais complexo, pelo fato das consultas expressas serem vagas. Isto é, o sistema deve ser capaz de identificar o tipo de informação contidas em cada documento, para então, recuperar somente os documentos que satisfazem uma determinada necessidade. Recuperar esta informação para um ser humano é possível, pois ele consegue estabelecer a relevância de um determinado documento para uma necessidade específica, mas consiste no maior problema em SRI. Para que os sistemas de recuperação de informação possa estabelecer isto, torna-se necessária a construção de modelos de decisões relevantes que possam ser determinados de maneira dinâmica.

Os modelos de recuperação de informação são baseados na comparação de uma determinada palavra-chave com os documentos armazenados. Desta forma, diversos modelos têm sido propostos na literatura. Uma taxonomia dos modelos de recuperação é apresentada na Figura 2.1. Neste trabalho, os modelos clássicos: booleano, vetorial e probabilísticos são abordados. Demais modelos também podem ser encontrados, mas não são temas desta dissertação. Mais detalhes referentes a estes modelos podem ser encontrados em [Baeza-Yates and Ribeiro-Neto, 1999].

Este capítulo explora os seguintes temas: primeiramente é feita uma introdução aos modelos clássicos da recuperação de informação. A seguir, aborda-se o primeiro dos modelos clássicos que é o modelo booleano (Seção 2.2). Neste, documentos e consultas são representados como conjunto de termos indexados, em que cada índice é uma variável booleana. Para obter ganho na velocidade do processamento da consulta, é abordado um dos maiores conceitos em recuperação de informação na Seção 2.3, que é o índice invertido. Palavras com pouco ou nenhum poder de discriminação são tratadas na Seção 2.4, estes são chamados de *Stop Words*. Para que os índices não suportem apenas consultas do tipo booleano, a associação de peso aos termos é abordado na Seção 2.5. Com intuito de aumentar o desempenho do processo de recuperação atribuindo pesos não binários aos termos de uma coleção é abordado o segundo modelo clássico, conhecido como Modelo Vetorial (Seção 2.6). Neste, documentos e consultas são representados como vetores. Outro modelo clássico é o Probabilístico, cuja representação de documentos e consultas é baseada na teoria probabilística. Por fim, são apresentadas formas de medidas de avaliação da qualidade do processo de recuperação de informação para os três modelos apresentados (ver Seção 2.8).

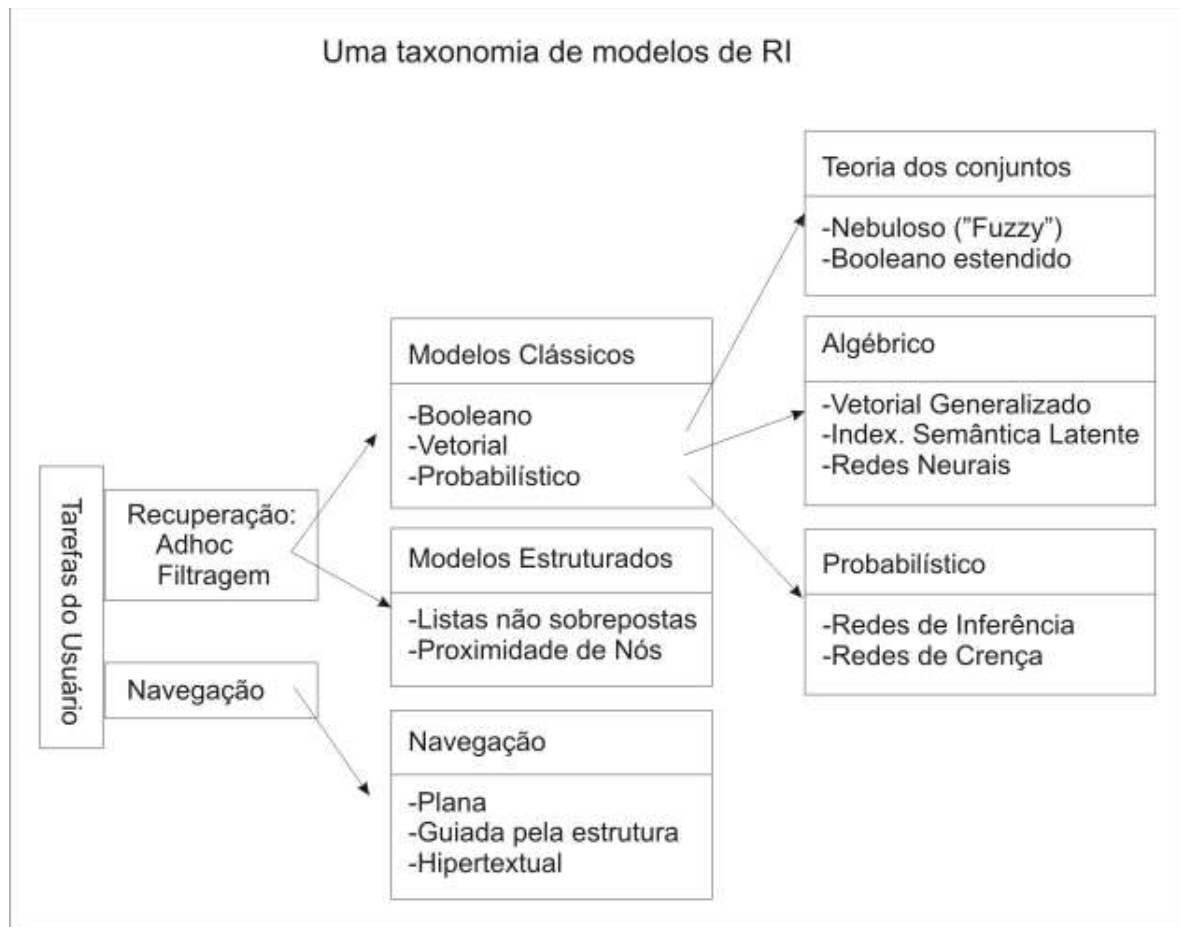


Figura 2.1: Uma taxonomia de modelos de RI (Adaptado de [Baeza-Yates and Ribeiro-Neto, 1999])

2.1 Introdução aos Modelos Clássicos

A maioria dos modelos de recuperação de informação são de natureza quantitativa, baseados na lógica, estatística e teoria dos conjuntos. Isto ocorre pelo fato de que a determinação de um modelo matemático geralmente pressupõe uma cuidadosa análise formal do problema e especificações de hipóteses, além de uma formulação explícita da forma como o modelo depende das hipóteses.

Nos modelos clássicos, cada documento é representado por um conjunto de palavras-chaves representativas - também chamadas de termos de indexação - considerados mutuamente independentes, que busca representar o assunto do documento e sumarizar seu conteúdo de forma significativa. Associa-se a cada termo de indexação t_i em um documento d_j a um peso $w_{ij} \geq 0$, que quantifica a correlação entre os termos e o documento, ou seja, reflete a importância do termo t_i no documento d_j . Analogamente a cada par termo-consulta (t_i, q) , associa-se o peso $w_{iq} \geq 0$. Como calcular estes pesos é abordado na Seção 2.5.

Os modelos clássicos de recuperação são três:

- booleano: documentos e consultas são representados como conjunto de índices, baseado na teoria de conjuntos;
- vetorial: documentos e consultas são representados como vetores em um espaço t -dimensional, onde t é o número de termos. Assim, este modelo é algébrico.
- probabilístico: o modelo de representação dos documentos e consultas é baseado na teoria probabilística. Possui esta denominação justamente por trabalhar com conceitos provenientes da área de probabilidade e estatística.

2.2 Modelo Booleano

É um modelo que se fundamenta na teoria dos conjuntos. A recuperação é baseada nos conceitos de lógica ou álgebra booleana. No modelo booleano os termos presentes na consulta e no documento são ligados por meio de conectores lógicos: *AND*, *OR* e *NOT* (E, OU, NÃO).

No modelo booleano, os documentos são representados por um conjunto de termos indexados, onde cada termo é uma variável booleana o que resulta em uma associação de peso a cada termo. Desta forma, se o termo estiver presente no documento terá um peso igual a 1 (um), caso contrário, 0 (zero). Resumidamente, pode-se dizer que, cada documento da coleção é representado por um vetor t -dimensional, onde t é o número de termos presentes no documento, e 1/0 refere-se a presença ou ausência do termo no documento respectivamente.

Uma consulta que utiliza apenas um termo, por exemplo, "animal" obtém como resultado o conjunto de documentos que contiverem a palavra "animal". Uma consulta conjuntiva que utiliza o conector *AND* requer que ambos os termos ligados por este conector estejam presentes no documento. Por exemplo, a expressão "animal *AND* puma" recupera somente documentos que possuem ambos os termos (animal e puma).

Outra forma de expressar uma consulta é baseado na utilização de uma expressão disjuntiva do tipo *OR*. Desta maneira, a consulta "animal *OR* puma" recupera todos os documentos que possuem o termo "animal" ou o termo "puma".

Por fim, pode-se utilizar o conector *NOT*, que solicita que um termo específico não esteja presente no documento. "Animal *NOT* puma" representa uma consulta em que os documentos recuperados possuem o termo "animal", mas não possuem a palavra "puma".

É importante ressaltar que, no modelo booleano, não existe a noção de "casamento" parcial com as condições de consulta. Suponha a consulta "animal *AND* puma *AND* América *AND* norte", o documento que contiver 3 ou menos destes termos, não será recuperado. Outras observações importantes sobre este modelo se devem ao fato dele não permitir a existência de resultados parciais e nem gerar informações suficientes para ordenação das consultas. Com isto, surge uma de suas principais desvantagens. Por trabalhar de forma

binária, não é criada nenhuma espécie de ordenação dos resultados por ordem de relevância, que atendam às condições de consulta do usuário. E esta forma de classificação é uma característica essencial em muitos SRI modernos.

É importante ressaltar também a semântica precisa destes modelos. De acordo com a semântica utilizada a quantidade de documentos recuperados pode sofrer uma variação muito grande. Por exemplo, caso o usuário faça uso da expressão disjuntiva *OR*, muitos documentos podem ser recuperados. O que torna a análise da qualidade de recuperação por parte do usuário crítica. Mas caso contrário, se fizer uso do conector *AND*, é provável que poucos documentos sejam recuperados não satisfazendo as necessidades do usuário. Portanto, é necessário que o usuário tenha um pré conhecimento da lógica inerente do sistema juntamente com o conhecimento destes operadores lógicos.

Como vantagem pode-se citar o formalismo e a simplicidade conceitual existente nos fundamentos deste modelo.

Com o intuito de ordenar as respostas da consulta por relevância utilizando o paradigma do modelo booleano, variações deste modelo, foram propostas na literatura, como por exemplo, os modelos fuzzy e booleano estendido [Baeza-Yates and Ribeiro-Neto, 1999].

Uma estratégia de recuperação do modelo booleano é comparar a lista de palavras de uma consulta com a lista de palavras de registros. Registros são coleções de documentos e cada registro contém uma lista de palavras. Entretanto, em grandes coleções a comparação direta da consulta com todos os registros é computacionalmente cara. Considere o seguinte cenário de um sistema de recuperação de informação: suponha uma coleção com 1 milhão de documentos e 500.000 termos, sendo em média 1.000 termos por documento. A Construção de uma matriz termo-documento de 500.000 X 1.000.000 gera uma matriz de meio milhão de zeros (0, ausência do termo) e uns (1, presença do termo). Mas, observe que esta matriz terá poucas entradas de valor 1 e muitos valores igual a 0, pois cada documento tem em média 1.000 palavras, e a matriz não terá mais de um bilhão de 1's, desta forma, um mínimo de 99,8% das células são zeros. Uma melhor representação seria gravar somente os termos que ocorrem, ou seja, os valores iguais a 1.

Este cenário inspira a criação para um dos principais conceitos em recuperação de informação, que é o índice invertido. Com o objetivo de acelerar o cálculo da similaridade, ele limita a comparação a um subconjunto de registros [Manning et al., 2007].

2.3 Índice Invertido

Um banco de dados textual é uma coleção de documentos, que pode ser representada por um conjunto de registros. Cada registro contém uma lista de palavras. Uma estratégia de recuperação textual é comparar a lista de palavras de uma consulta com a lista de palavras dos registros. Entretanto, em grandes coleções, a comparação direta da consulta com todos os registros é computacionalmente cara. Com o objetivo de acelerar este cálculo,

limitando a comparação a um subconjunto de registros, utiliza-se o índice invertido¹. A ideia é limitar a comparação somente com o subconjunto de registros que contenham pelo menos um termo da consulta.

O índice invertido possui duas partes principais: uma estrutura de busca, chamada de vocabulário, contendo todos os termos distintos existentes nos textos indexados e, para cada termo, uma lista invertida que armazena os identificadores dos registros contendo o termo, a saber, documentos onde a palavra ocorre [Baeza-Yates and Ribeiro-Neto, 1999]. O dicionário é ordenado por ordem alfabética e a lista invertida organizada pelo identificador do documento (ID) como ilustrado na Figura 2.2.

Doc1 O homem medíocre espera tudo dos outros. O homem é covarde.		Doc2 O homem de bem exige tudo de si próprio. O homem é corajoso.	
Coluna 1		Coluna 2	
Termo	DocID	Termo	DocID
o	1	bem	1
homem	1	corajoso	2
medíocre	1	covarde	1
espera	1	de	2
tudo	1	de	2
dos	1	dos	1
outros	1	é	1
o	1	é	2
homem	1	espera	1
é	1	exige	2
covarde	1	homem	1
o	2	homem	1
homem	2	homem	2
de	2	homem	2
bem	2	medíocre	1
exige	2	o	1
tudo	2	o	1
de	2	o	2
si	2	o	2
próprio	2	outros	1
o	2	próprio	2
homem	2	si	2
é	2	tudo	1
corajoso	2	tudo	2

Coluna 3	
Termo	DocID Freq. Termo
bem	1 1
corajoso	2 1
covarde	1 1
de	2 2
dos	1 1
é	1 1
é	2 1
espera	1 1
exige	2 1
homem	1 2
homem	2 2
medíocre	1 1
o	1 2
o	2 2
outros	1 1
próprio	2 1
si	2 1
tudo	1 1
tudo	2 1

Figura 2.2: As duas partes que formam o Índice Invertido, a saber: dicionário e lista invertida

As principais tarefas para a construção de um índice invertido serão descritas a seguir. Esta explicação é baseada em [Manning et al., 2007].


Para obter ganho de tempo no processo de recuperação e indexação, é necessário que

¹Alguns pesquisadores em recuperação da informação preferem o termo arquivo invertido. Mas neste trabalho será abordada a expressão índice invertido devido às referências principais adotadas nesta dissertação.

o índice invertido seja construído antes do processo que envolve a consulta do usuário, ou seja, o índice invertido é um processo de pré-processamento.

Devidamente escolhida a coleção de documentos a ser indexada, é necessário assumir que cada documento tenha um identificador do documento (ID) único na coleção. A cada documento encontrado é atribuído a ele um número serial. Os dados inseridos no processo de indexação são uma lista de termos contidos no documento, que podem também ser vistos como uma lista com dois dados possíveis, que são, o termo e o ID do documento, conforme ilustrado na Figura 2.3.

Como modo ilustrativo, é considerado uma coleção com 2 documentos (Doc1, Doc2). Na primeira coluna da Figura 2.3 os termos são indexados na ordem em que eles ocorrem no documento juntamente com o ID do documento correspondente. Termos que ocorrem mais de uma vez em um mesmo documento, nesta fase, ainda são indexados quantas vezes eles estiverem presentes no documento (coluna 1 da Figura 2.3). A seguir, os termos são ordenados por ordem alfabética (coluna 2 da Figura 2.3). Para finalizar esta etapa de construção, as várias ocorrências do mesmo termo, a partir do mesmo documento são, em seguida, agrupados, e uma coluna extra é adicionada para gravar a frequência do termo no documento (coluna 3 da Figura 2.3).



Termo	DocID	Freq. Termo
bem	1	1
corajoso	2	1
covarde	1	1
de	2	2
dos	1	1
é	1	1
é	2	1
espera	1	1
exige	2	1
homem	1	2
homem	2	2
mediocre	1	1
o	1	2
o	2	2
outros	1	1
próprio	2	1
si	2	1
tudo	1	1
tudo	2	1

Termo	Freq. Termo	Lista Invertida
bem	1 → 1	
corajoso	1 → 2	
covarde	1 → 1	
de	2 → 2	
dos	1 → 1	
é	2 → 1, 2	
espera	1 → 1	
exige	1 → 2	
homem	4 → 1, 2	
mediocre	1 → 1	
o	4 → 1, 2	
outros	1 → 1	
próprio	1 → 2	
si	1 → 2	
tudo	2 → 1, 2	

Figura 2.3: Indexação dos termos, parte um

A frequência do termo no documento é desnecessária para SRI que utiliza o modelo booleano apresentado na seção 2.2, mas como será abordado mais adiante, é útil em muitos outros modelos de recuperação, no qual destaca-se, o modelo vetorial, que foi utilizado na

implementação deste trabalho.

Com isto, instâncias do mesmo termo são agrupadas e o resultado é dividido em um dicionário e uma lista invertida, como ilustrado na Figura 2.4. Considerando-se que um termo geralmente está presente em diversos documentos, este tipo de estrutura reduz o número de comparações necessárias para o processamento de consultas, pois somente os documentos que contiverem os termos da consulta serão examinados. A estrutura de índice invertido é a melhor e mais eficiente estrutura para suportar uma busca de texto *ad-hoc* [Frakes and Baeza-Yates, 1992].

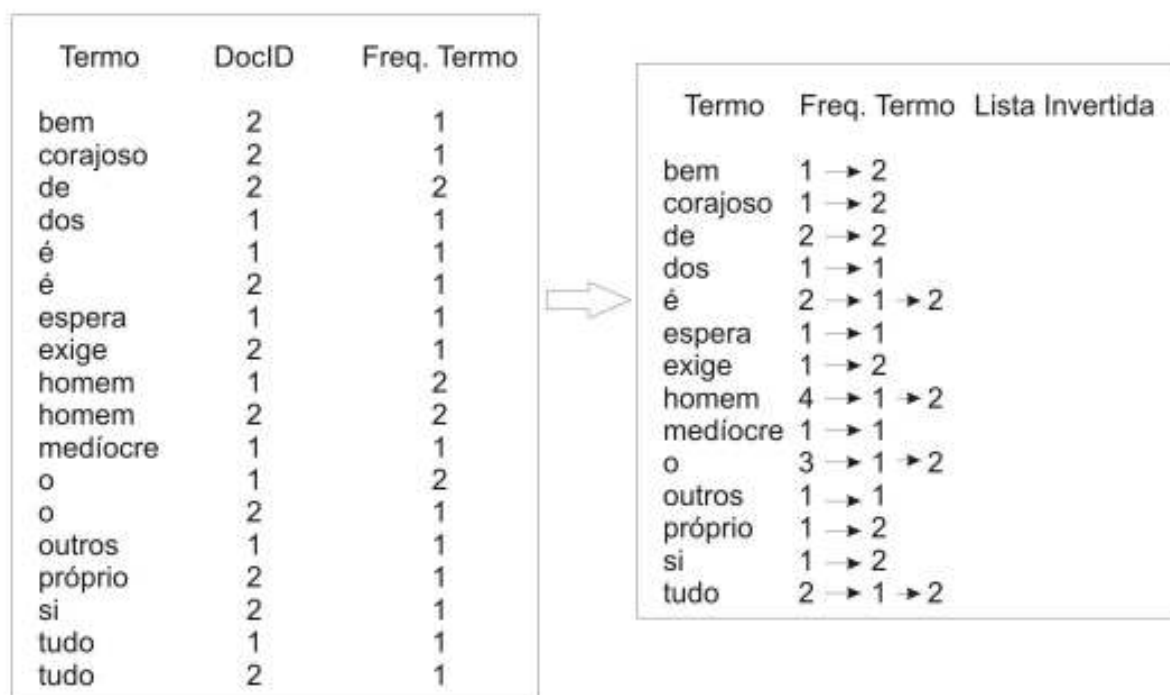


Figura 2.4: Indexação dos termos, parte dois

Consultas nos arquivos invertidos são feitas tomando-se a lista invertida correspondente ao termo procurado. Uma busca por documentos com os termos "próprio" e "corajoso" pode ser processada por meio das duas listas invertidas, sem a necessidade de percorrer todos os registros.

Normalmente o dicionário é mantido na memória do computador, enquanto que as listas são mantidas em disco, assim o tamanho de cada um deve ser considerado relevante e o estudado da forma de se otimizar o armazenamento destes componentes. No capítulo 5 de [Manning et al., 2007], este estudo é bem detalhado e o leitor terá grandes informações.

Uma maneira de se reduzir o número de termos armazenados no índice invertido é considerar o uso das *stop words*.

2.4 Stop words

Stop Words são palavras extremamente comuns e semanticamente não seletivas. Por serem comuns, geram grandes listas invertidas difíceis de processar. Mas como não são seletivas podem ser excluídas do dicionário. Exemplos de *stop words* são: artigos, preposições, conjunções e outras. Uma relação de *stop words* para o português do Brasil é proposta por [Balinski, 2002]. Já em [Frakes and Baeza-Yates, 1992] encontra-se uma lista de palavras comuns da língua inglesa, juntamente com o algoritmo para detecção destes termos comuns.

A estratégia geral para determinação de uma *stop list* é classificar os termos de acordo com a sua frequência na coleção de documentos, e em seguida, os termos mais frequentes são considerados como termos comuns e são descartados durante o processo de indexação. Esta classificação pode ser manual, neste caso, um especialista avalia quais as palavras que não devem ser indexadas, (o que pode variar de sistema para sistema) ou utilizar listas sugeridas como nos trabalhos citados acima.

Uma propriedade importante dentro deste paradigma refere-se à regra dos 30 [Manning et al., 2007] e [Frakes and Baeza-Yates, 1992]. Esta regra diz que, os 30 termos mais comuns de uma coleção correspondem a 30% dos termos presentes nos documentos.

Na verdade indexar ou não *stop words* provavelmente (teoricamente) não afeta muito nos resultados, porque estas inerentemente tem um peso (importância) muito baixo, já que são extremamente comuns (peso de termos é apresentando na Seção 2.5). Termos comuns não são representativos ao documento, por isto, não fornecem nenhuma contribuição na identificação do conteúdo do texto. Mas, no entanto, para pesquisas com frases esta afirmação não pode ser considerada verdadeira. Algumas consultas podem ser afetadas com o uso desta técnica. No capítulo 7 de [Manning et al., 2007] é feito um estudo mais detalhado deste assunto. Este assunto mostra que o custo da utilização de *stop words* não é tão grande, nem em relação a dimensão do índice, nem em relação ao tempo de processamento.

2.5 Peso do Termo

O índice apresentado anteriormente considera apenas a presença ou ausência do termo no documento. Sendo assim, um documento pode ou não estar relacionado a uma consulta. Como dito anteriormente, caso a consulta seja pouca específica, o modelo booleano, pode obter um número elevado de documentos como resposta o que torna análise humana complexa e muitas vezes inviável. Uma solução para esse problema é que os SRI classifiquem ou ordenem os documentos correspondentes a uma consulta.

Dada uma consulta q , é necessário que os SRI a partir de uma medida de similaridade, para calcular quais documentos da coleção são mais similares à consulta q com o intuito

de criar uma ordenação dos documentos da coleção, de modo que os documentos mais similares à consulta q apareçam nas primeiras posições dessa ordenação (*ranking*).

Uma abordagem para construção desse *ranking* é considerar a importância de um termo no documento com base na frequência com que o termo ocorre no documento e na coleção. Esta abordagem é conhecida como peso do termo. A questão é, como determinar este peso entre um termo e cada documento?

O peso de cada termo presente no documento depende do número de ocorrências do termo no documento. A abordagem mais simples para este caso, consiste em atribuir um peso igual ao número de ocorrências do termo t no documento d . Este esquema de peso é chamado de frequência do termo (tf) e é denotado por $tf_{t,d}$, onde t representa o termo e d o documento.

Esta abordagem apresenta um problema. Todos os termos presentes na coleção com a mesma frequência são considerados igualmente importantes. Como discutido anteriormente, existem termos com baixo poder discriminatório ou até mesmo com nenhum poder de discriminação para determinar a relevância do documento em relação à consulta. Por exemplo, em uma coleção de documentos sobre recuperação de informação, o termo "recuperação" aparecerá em quase todos os documentos. O termo "recuperação", neste caso, terá um alto poder de discriminação? A resposta é não. Neste caso, é necessário introduzir mecanismos que atenuem o efeito de um termo ocorrer muitas ou poucas vezes na coleção. Assim, a ideia inicial seria reduzir o peso dos termos que ocorressem com maior frequência na coleção. É comum utilizar para esse fim a frequência de documentos (df) df_t , definido como o número de documentos da coleção em que o termo t ocorre. A razão desta preferência pode ser vista no exemplo ilustrado na Figura 2.5. Define-se frequência na coleção (cf) como o número de vezes que o termo ocorre em toda coleção. Assim, a frequência na coleção para os termos *Recuperação* e *Seguro* são bastantes similares, mas os seus valores df variam significativamente. Isto indica que, o termo *Seguro* aparece em poucos documentos. Mas nos poucos documentos em que ele ocorre a sua frequência é muito alta. Assim, sua frequência na coleção é alta enquanto sua frequência nos documentos é baixa. Desta forma, na coleção o termo *Seguro* deve receber um peso maior do que o termo *Recuperação*.

Termo	cf	df
Recuperação	15.613	5.976
Seguro	15.600	13

Figura 2.5: Frequência na coleção (cf) e frequência de documentos (df)

Desta forma, define-se a Frequência Inversa do Documento (IDF) de um termo t da

seguinte forma:

$$idf_t = \log \frac{N}{df_t} \quad (2.1)$$

onde N é o número total de documentos da coleção e df_t é o número de documentos da coleção em que o termo ocorre. Assim, o IDF de um termo raro é elevado, enquanto que o idf de termos comuns decresce em escala logarítmica.

Resumidamente $tf_{t,d}$ reflete características intra-documento e idf_t dá uma medida de distinções inter-documentos. Pesos baseados no produto $tf_{t,d} \times idf_t$, são chamados de abordagem $tf - idf_{t,d}$. Estes servem para produzir um peso composto para cada termo em cada documento.

Em outras palavras, $tf - idf_{t,d}$ atribui ao termo t um peso no documento d que é:

- maior quando o termo t ocorre muitas vezes dentro de um pequeno número de documentos (assim terá um alto poder para discriminar aqueles documentos);
- diminui quando o termo t ocorre poucas vezes em um documento ou ocorre em muitos documentos (oferecendo assim um sinal menos acentuado de relevância);
- aproxima-se de zero quando o termo t ocorre em quase todos os documentos;
- é igual a zero quando o termo ocorrer em todos os documentos, por exemplo, alguns artigos e preposições.

Neste momento, pode-se visualizar cada documento como um vetor onde cada índice correspondente a um termo, juntamente com um peso dado pela equação do $tf \times idf_t$. Este vetor será crucial para a geração do *ranking* que iremos apresentar na seção seguinte.

2.6 Modelo Vetorial

Proposto inicialmente por [Salton and Buckley, 1988], este modelo representa documentos e consultas como vetores de termos. Os termos que compõem o sistema de recuperação são modelados como elementos pertencentes a um espaço vetorial. Cada termo possui um valor (peso) associado a um documento que indica o grau de importância deste termo neste documento. Assim, cada documento é constituído por pares de elementos na forma $[termo_i, peso_i]$. Cada elemento do vetor de termos é considerado uma coordenada dimensional. Desta forma, os documentos são representados como vetores de termos em um espaço t -dimensional, onde t é o número de termos ou o tamanho do vocabulário.

Observe na Figura 2.6 uma representação gráfica de um documento Doc1 com termos de indexação $t1$ e $t3$, com pesos 0.7, 0.6, respectivamente.

A Figura 2.7 representa graficamente um documento tridimensional doc2 com três termos, $t1$, $t2$, $t3$, com pesos 0.6, 0.8, 0.5 respectivamente.

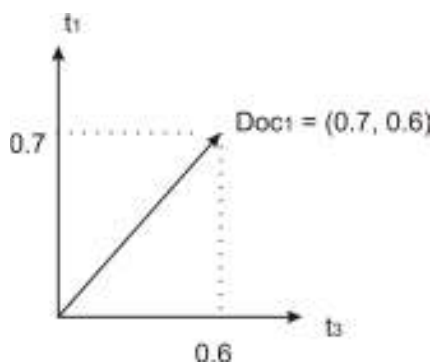


Figura 2.6: Representação vetorial de um documento com dois termos de indexação

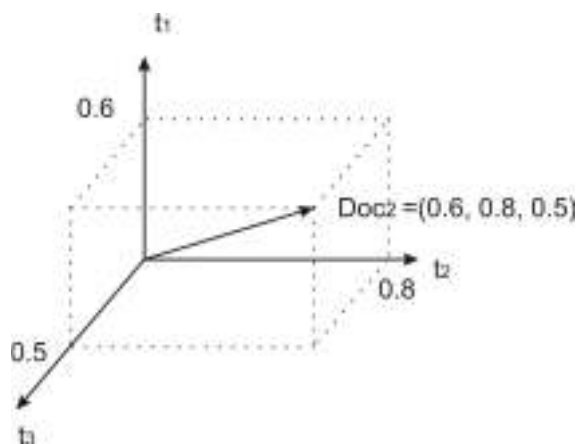


Figura 2.7: Representação vetorial de um documento com três termos de indexação

Os dois documentos são representados em um mesmo espaço vetorial. É interessante lembrar que, os termos que não estão presentes em um determinado documento recebem peso igual a zero, veja Doc1 na Figura 2.8.

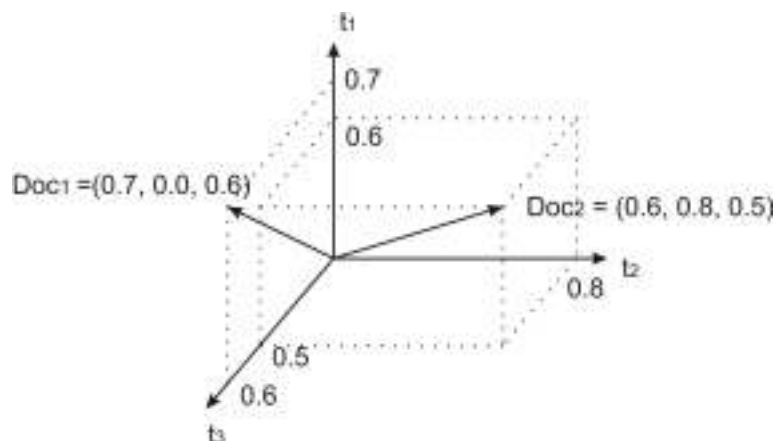


Figura 2.8: Espaço vetorial contendo dois documentos

Da mesma forma que os documentos, as consultas também são representadas por vetores de termos $q = (W1_q, W2_q, \dots, Wn_q)$. A Figura 2.9 esboça a expressão de busca eBusca1 com três termos, $t1$, $t2$ e $t3$, com pesos 0.8, 0.6 e 0.2, respectivamente e eBusca1

é representado juntamente com dos documentos Doc1 e Doc2 em um espaço vetorial formados pelos termos $t1$, $t2$ e $t3$.

É importante lembrar que esta representação tridimensional de um espaço vetorial foi feita apenas como modo de visualização. Em um sistema real, existem inúmeros termos de indexação e documentos.

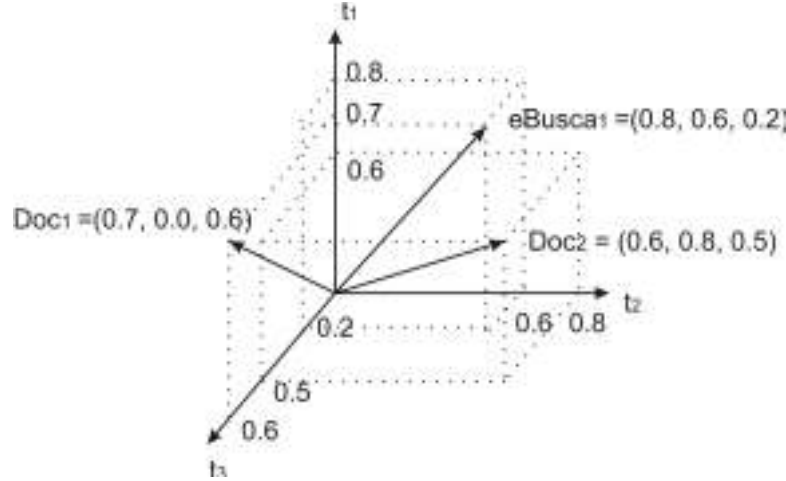


Figura 2.9: Representação de uma expressão de busca em um espaço vetorial

A utilização de uma mesma representação para documentos e consultas permite o cálculo da similaridade entre uma consulta q e um documento d_j . A similaridade é dada pela correlação entre os vetores que os representam, quantificada pelo cosseno do ângulo formado por d_j e q . Esta métrica é conhecida como medida de similaridade do cosseno. Em um espaço vetorial de dimensão n , a similaridade (sim) entre dois vetores q e d é calculada através do cosseno do ângulo formado por estes vetores, através da seguinte fórmula [Manning et al., 2007]:

$$\text{sim}(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^n W_{qi} W_{di}}{\sqrt{\sum_{i=1}^n W_{qi}^2} \sqrt{\sum_{i=1}^n W_{di}^2}} \quad (2.2)$$

onde \mathbf{q} é o vetor de termos da consulta; \mathbf{d} é o vetor de termos do documento; W_{qi} é o peso do termo i da consulta q e W_{di} é o peso do termo i no documento d .

Desta forma é possível obter o grau de similaridade entre Doc1 e Doc2, exibidos na Figura 2.8, conforme exemplo abaixo:

$$\text{sim}(\vec{Doc1}, \vec{Doc2}) = \frac{(0.7 \times 0.6) + (0.0 \times 0.8) + (0.6 \times 0.5)}{\sqrt{(0.7^2 + 0.0^2 + 0.6^2)} \sqrt{0.6^2 + 0.8^2 + 0.5^2}} = 0.70 \quad (2.3)$$

Portanto, o grau de similaridade entre o Doc1 e Doc2 é igual a 0.70 ou 70%.

Analogamente pode-se calcular a similaridade entre a expressão eBusca1 e os docu-

mentos Doc1 e Doc2 mostrados na Figura 2.9.

$$\text{sim}(e\vec{Busca1}, \vec{Doc1}) = \frac{(0.7 \times 0.8) + (0.0 \times 0.6) + (0.8 \times 0.2)}{\sqrt{(0.7^2 + 0.0^2 + 0.8^2) \sqrt{0.8^2 + 0.6^2 + 0.2^2}}} = 0.67 \quad (2.4)$$

$$\text{sim}(e\vec{Busca1}, \vec{Doc2}) = \frac{(0.6 \times 0.8) + (0.8 \times 0.6) + (0.5 \times 0.2)}{\sqrt{(0.6^2 + 0.8^2 + 0.5^2) \sqrt{0.6^2 + 0.8^2 + 0.5^2}}} = 0.94 \quad (2.5)$$

Portanto a expressão eBusca1 possui um grau de similaridade de 67% com o Doc1 e de 94% com o Doc2.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos da coleção são utilizados na ordenação dos documentos resultantes. Portanto, no modelo vetorial os resultados são ordenados de acordo com o grau de similaridade. Com o intuito de construir um *ranking*, ou seja, uma lista dos documentos ordenados por seus respectivos graus de relevância em relação à consulta. Com isto, no modelo vetorial, um documento pode ser recuperado mesmo se ele satisfizer a consulta parcialmente. Os documentos com maior grau de similaridade ficam na parte superior do *ranking*. No exemplo acima, a ordem dos documentos no *ranking* seria a seguinte: Doc2, Doc1.

Uma desvantagem teórica deste modelo se dá pelo fato dos termos serem considerados mutuamente independentes, ou seja, não são considerados os relacionamentos existentes entre eles. Porém [Baeza-Yates and Ribeiro-Neto, 1999], diz que na prática, não há evidências conclusivas que apontem que a dependência dos termos traga melhoria significativa nos resultados de um sistema de recuperação de informação.

As principais vantagens do modelo vetorial são: simplicidade, facilidade para calcular a similaridades com eficiência, e pelo fato de se comportar bem em coleções genéricas [Baeza-Yates and Ribeiro-Neto, 1999]. Por este modelo ter um bom comportamento em coleções genéricas surgiu a motivação de aplicá-lo no contexto de imagens. Mais detalhes no Capítulo 4.

2.7 Modelo Probabilístico

O modelo probabilístico [Rijsbergen, 1976], tenta representar o processo de recuperação de informação sob um ponto de vista probabilístico. Assim, estima a probabilidade do usuário encontrar um documento d_j relevante para uma consulta q .

Dada uma consulta, existe um conjunto de documentos que contém exatamente os documentos relevantes, e nenhum outro mais. Desta maneira é suposto que exista um conjunto ideal de documentos que satisfaça a cada uma das consultas do usuário, e que este conjunto possa vir a ser recuperado.

Ele é baseado no Princípio Probabilístico de Ordenação (*Probability Ranking Principle*), que estabelece que este modelo pode ser usado de forma ótima. Este princípio considera que a relevância de um documento para uma determinada consulta é independente de outros documentos. Sua definição é a seguinte:

"Se a resposta de um sistema de recuperação de referência a cada consulta, é uma ordem de documentos classificada de forma decrescente pela probabilidade de relevância para o usuário que submeteu a requisição, onde as probabilidades são estimadas com a melhor precisão com base nos dados disponíveis, então a efetividade geral do sistema para o seu usuário será a melhor que pode ser obtida com base naqueles dados".

Então dada uma expressão de busca, podemos dividir a coleção (com N documentos) em quatro subconjuntos distintos: i) conjunto dos documentos relevantes (Rel); ii) o conjunto dos documentos recuperados (Rec); iii) conjunto dos documentos relevantes recuperados (RR); iv) conjuntos dos documentos não relevantes e não recuperados (NR) conforme a Figura 2.10.

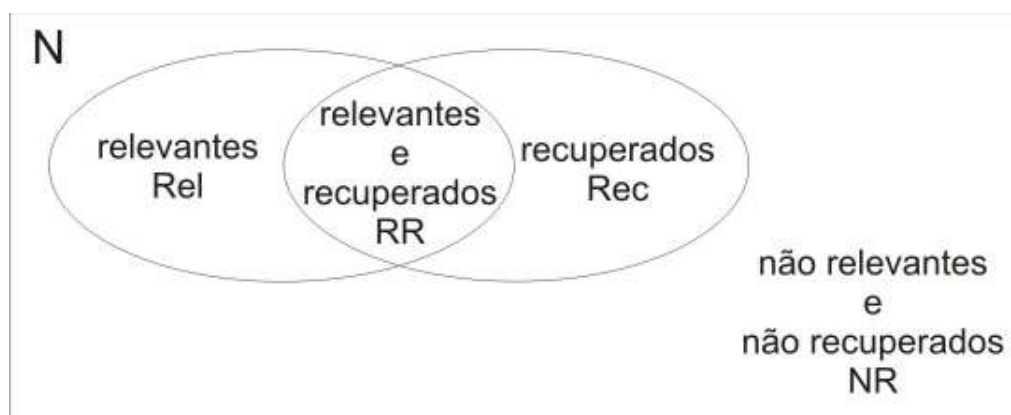


Figura 2.10: Subconjunto de documentos após a execução de uma busca relativa a uma consulta q [Ferneda, 2003].

A partir desta análise, observa-se que o resultado ideal de uma consulta seria o conjunto que contenha todos os documentos relevantes (Rel_q). O problema é que não se sabe exatamente quais são estes documentos relevantes. Sabe-se que existem termos de índice cujas semânticas podem ser utilizadas para caracterizar estas propriedades. Como estas propriedades não são conhecidas ao se fazer a consulta, são necessárias estimá-las. Com os resultados obtidos após a primeira execução é possível gradativamente melhorar o resultado a partir das interações do usuário com o sistema de busca.

Os pesos neste modelo são todos binários, tanto para os documentos quanto para a consulta indicando presença ou ausência do termo no documento e/ou consulta. Seja o conjunto Rel_q o conjunto de documentos que foram estimados como relevantes para a consulta q , e seja NR_q o complemento de Rel_q , ou seja, o conjunto dos documentos não relevantes, tem-se que: $P(t_i|Rel_q)$ é a probabilidade do termo t_i estar presente no documento escolhido aleatoriamente do conjunto Rel_q ; e $P(t_i|NR_q)$ a probabilidade do

termo t_i estar presente em um documento escolhido aleatoriamente do conjunto NR_q . Seja $P(Rel_q|d_j)$ a probabilidade de o documento Rel_q ser relevante para a consulta q e $P(NR_q|d_j)$ a probabilidade do documento d_j não ser relevante para a consulta q . No modelo probabilístico, a similaridade entre um documento d_j e uma consulta q é definida por:

$$Sim_{d_j,q} = \frac{P(Rel_q|d_j)}{P(NR_q|d_j)} \quad (2.6)$$

Por motivos de simplificação do cálculo, no modelo probabilístico é assumida a independência dos termos. Além disso, para calcular $P(Rel_q|d_j)$ e $P(NR_q|d_j)$, o modelo aplica a elas uma série de transformações algébricas preservando a ordem com o objetivo de obter uma estimativa numérica para o *ranking* do documento d_j . Após várias operações utilizando regra de Bayes e logaritmos (para mais detalhes ver em [Rijsbergen, 1976] e [Manning et al., 2007] tem-se a expressão clássica para determinar o *ranking* no modelo probabilístico:

$$Sim_{d_j,q} = \sum_{i=1}^t w_{i,j} \cdot w_{i,q} \cdot \sigma_{i/r} \quad (2.7)$$

onde

$$\sigma_{i/r} = \ln \frac{P(t_i|Rel_q)}{1 - P(t_i|Rel_q)} + \ln \frac{1 - P(t_i|NR_q)}{P(t_i|NR_q)} \quad (2.8)$$

O modelo probabilístico apresenta como vantagem, além do bom desempenho prático, o princípio probabilístico de ordenação, que uma vez garantido, resulta em um comportamento ótimo do método. Entretanto, a desvantagem é que este comportamento depende da precisão das estimativas de probabilidade, que é uma precisão difícil de ser estimada. Além disso, o método não explora a frequência do termo no documento [Baeza-Yates and Ribeiro-Neto, 1999].

2.8 Medidas de Avaliação

Conhecer a eficácia de sistemas de Recuperação de Informação é importante não só para os pesquisadores de RI, mas também para quem usa estes sistemas. Pesquisadores e usuários precisam ter maneiras efetivas de medir a qualidade dos sistemas e como estes podem ser melhorados. Desta forma, SRI são avaliados segundo o ponto de vista do sistema ou sob o ponto de vista do usuário.

Em RI, quando os sistemas são avaliados sob o ponto de vista do sistema, as duas medidas mais utilizadas são precisão e revocação [Rijsbergen, 1976].

Precisão (*precision*): a quantidade de documentos relevantes recuperados pelo sistema, divididos pelo número total de documentos recuperados. Por exemplo, se para uma

busca realizada por um SRI forem recuperados 12 documentos e destes apenas 6 forem realmente relevantes, a precisão do sistema é 0,5 ou 50%.

$$Pr_q = \frac{RR_q}{Rec_q} \quad (2.9)$$

onde: RR_q é o número de documentos relevantes recuperadas e Rec_q o total de documentos recuperados.

Revocação (*recall*): existe a possibilidade de que ocorram diversos documentos na base de dados que o usuário do SRI considere relevante, mas somente alguns deles podem ser recuperados pelo sistema. A taxa de revocação de uma consulta é dada pelo número de documentos relevantes recuperados pelo sistema dividido pelo número total de documentos relevantes existente na base de dados. Seguindo o exemplo, caso fossem recuperados 12 documentos e destes 6 apenas fossem relevantes e o total de documentos relevantes da coleção fossem 10 então a revocação seria de 0,6 ou 60%.

$$Re_q = \frac{RR_q}{Rel_q} \quad (2.10)$$

onde: RR_q é o número de documentos relevantes recuperadas e Rel_q o total de documentos relevantes da coleção.

Estas equações podem ser feitas por meio de um exame da seguinte Tabela de contingência:

Tabela 2.1: Tabela de contingência

-----	Relevantes	Irrelevantes
Relevantes	verdadeiro positivo (tp)	falso positivo (fp)
Irrelevantes	falso negativo (fn)	verdadeiro negativo (tn)

onde:

$$Pr_q = \frac{tp}{tp + fp} \quad (2.11)$$

$$Re_q = \frac{tp}{tp + fn} \quad (2.12)$$

Outras medidas podem ser encontradas na literatura, no qual apresenta-se algumas delas aqui e outras podem ser encontradas em [Manning et al., 2007] e [Baeza-Yates and Ribeiro-Neto, 1999].

Precisão R

Esta medida de precisão busca encontrar a precisão na posição X do *ranking*. Esta abordagem visa focar a avaliação nos documentos efetivamente observados pelo usuário,

pois embora os SRI recupere grande quantidade de documentos, em geral, o usuários só observam as primeiras posições do *ranking*. A Precisão-R (Pr-R) de uma consulta q é dada por:

$$Pr - R_q = \frac{Relevantes(RR_q)}{R} \quad (2.13)$$

onde $Relevantes(RR_q)$ é a quantidade de imagens relevantes encontradas até a posição R do *ranking*.

Medida-E

Esta combina a medida de revocação e precisão. A medidaE é definida como:

$$MedidaE = 1 - \frac{(1 + b^2)P \times R}{b^2P + R} \quad (2.14)$$

onde P é a precisão, R a revocação, e b é uma medida de importância relativa para o usuário, no caso dele optar por expressar uma maior ênfase na precisão ou na revocação. Quanto maior o valor de b maior é o interesse do usuário na precisão. A medida a ser utilizada em uma avaliação depende da aplicação e do contexto [Frakes and Baeza-Yates, 1992].

2.9 Conclusão

Este capítulo abordou conceitos e técnicas de recuperação de informação. Foram descritos os três modelos clássicos.

Foi visto que o modelo booleano apresenta como vantagem a expressividade completa se o usuário souber exatamente o que quer, além de ter uma complexidade baixa de implementação. Como desvantagem, destaca-se a semântica precisa deste modelo ocasionando a recuperação de nenhum documento (poucos) ou muitos documentos recuperados (*overload*). Outro problema deste modelo é que ele não ordena os documentos de acordo com o grau de relevância com a consulta.

Com o intuito de acelerar o processo de recuperação, a arquitetura conhecida como índice invertido foi apresentada. Este conceito visa limitar a comparação da consulta somente com o subconjunto de registros que contenham pelo menos um termo da consulta. Outro tópico estudado neste capítulo foi o de *Stop words*. Estas são utilizadas para reduzir o número de termos armazenados no índice invertido. Nesta Seção, foi visto um conceito importante que será aplicado em nossos experimentos, que se refere à regra dos 30. Esta diz que os 30 termos mais comuns da coleção representam 30% da coleção.

Para que os documentos possam ser ordenados de acordo com uma consulta, foi utilizado o conceito peso do termo. É atribuído para cada termo presente no documento um peso. Heurísticas para este cálculo foram apresentadas.

Uma forma de aumentar o desempenho do processo de recuperação é trabalhar com estes pesos no processo de indexação dos documentos. O segundo modelo clássico apresentado aqui faz uso destes pesos. Este modelo é chamado de Modelo Vetorial. Este modelo apresenta como vantagem: i) o uso de pesos aos termos, o que melhora o desempenho do processo de recuperação; ii) utiliza a estratégia de encontro parcial (função de similaridade), que é melhor que a exatidão do modelo booleano; iii) os documentos são ordenados de acordo com seu grau de similaridade com a consulta; iv) apresenta um bom comportamento em coleções genéricas. Como desvantagem foi visto que, no modelo vetorial ocorre a ausência de ortogonalidade entre os termos. Este modelo pode encontrar relações entre termos que aparentemente não têm nada em comum.

Dentro da Seção do modelo Probabilístico foi visto que, uma vez garantido o princípio probabilístico e de ordenação, este método resulta em um ótimo comportamento. Entretanto, seu comportamento depende da precisão das estimativas de probabilidade. Outro problema está no fato de não explorar a frequência do termo no documento.

Medidas para avaliar a qualidade de recuperação destes modelos foram apresentadas. As mais utilizadas são precisão e revocação.

Além de documentos, outras áreas da recuperação de informação são de grande utilidade para a comunidade das ciências da computação em geral. Com a explosão do número de imagens na Web, modelos para recuperação precisa de imagens passaram a ter uma importância muito maior. Recuperação de Imagens é o tema a ser estudado no próximo capítulo.

Capítulo 3

Recuperação de Imagens

Além da busca de documentos textuais descrita no Capítulo 2, o usuário pode se interessar na busca por outros conteúdos, tais como sons, imagens e vídeos. Com o foco nas imagens presentes na Web, este capítulo apresenta a recuperação de imagens.

Este capítulo é dividido da seguinte maneira: inicialmente é feita uma contextualização dos Sistemas de Recuperação de Imagens. A primeira modalidade abordada é a Recuperação de Imagens Baseadas em Anotações Textuais (Seção 3.3). Um foco especial é dado aos Sistemas de Recuperação de Imagens Baseada no Conteúdo (Content-Based Image Retrieval - CBIR) por ser utilizado neste trabalho (ver Seção 3.4). Métodos automáticos de indexação e recuperação baseados em algum tipo de característica cor, textura e forma são interessantes neste contexto uma vez que podem reduzir a intervenção humana possibilitando, desta forma, maior eficácia e, em muitos casos, uma significativa diminuição da margem de erro. Vetor de característica será tema tratado na Seção 3.5. Medidas de similaridade são estudadas na Seção 3.6. Um foco maior é dado para a medida de distância *Minkowski*, por ser utilizada neste trabalho. Por fim, medidas de avaliação dos resultados de recuperação de imagens são apresentadas.

3.1 Definindo uma Imagem

Uma imagem digital é uma função bidimensional $f(x, y)$ discretizada tanto em coordenadas espaciais quanto em brilho. Esta pode ser considerada como sendo uma matriz cujos índices de linhas e de colunas identificam um ponto na imagem e o correspondente valor de elemento da matriz identifica o nível de cinza naquele ponto. Quando os valores de x e y juntamente com os valores da amplitude f são todos finitos, a imagem é chamada de imagem digital. Os elementos dessa matriz digital são chamados de elementos da imagem, elementos da figura, *pixels* ou *pels*, estes dois últimos, abreviações de "*picture elements*" (elementos de figura). Neste trabalho é utilizado o termo *pixel* [Gonzales and Woods, 2002].

Embora o tamanho de uma imagem digital varie de acordo com sua aplicação, é

sugerido de se trabalhar com imagens de matrizes quadradas com tamanhos e número de níveis de cinza que sejam potências inteiras de 2 [Gonzales and Woods, 2002]. Por exemplo, na coleção Corel-1000 todas as imagens estão na resolução 384x256 ou 256x384 *pixels*.

De uma forma mais simplista, um *pixel* é o menor ponto que forma uma imagem digital, sendo que o conjunto de pixels forma a imagem inteira.

3.2 Sistemas de Recuperação de Imagens

O modo como o usuário expressa sua necessidade de informação por meio de uma consulta e a maneira de calcular a similaridade entre as imagens da coleção em relação à consulta depende, dentre outras coisas, da abordagem utilizada pelo sistema de recuperação de imagens [Datta et al., 2008]. A seguir, descrevemos as principais modalidades do usuário expressar a sua consulta em um sistema de recuperação de imagens.

- palavras-chaves: a consulta é expressa sob a forma de palavras-chaves. Atualmente este tipo de consulta é a forma mais popular dos motores de busca da Web, como por exemplo, o Google e o Yahoo!;
- imagem: a consulta é expressa por meio de uma imagem exemplo. Este tipo de consulta é a forma mais usual nos sistemas de recuperação de imagens baseada no conteúdo;
- gráficos: a consulta é expressa usando uma imagem desenhada à mão ou gerada pelo computador;
- composto: envolve o uso de uma ou mais das modalidades citadas anteriormente para realizar a consulta em um sistema.

As modalidades de consultas acima exigem diferentes métodos de processamento e/ou suporte para interação com o usuário. Os métodos de processamento em sistemas de recuperação de imagens são:

- baseada em textos: envolve uma combinação de palavras-chaves utilizando conjuntos teóricos de operações.
- baseada no conteúdo: este tipo de processamento é o cerne dos sistemas CBIR. O processamento da consulta (imagens ou gráficos) envolve a extração das características visuais das imagens.
- composto: envolve a recuperação baseada em texto e no conteúdo. Um exemplo de um sistema que suporta esse tipo de processamento é encontrado em [Joshi et al., 2006].

- Interativo: faz uso de interações do usuário no processo de recuperação. Um exemplo é o sistema de recuperação de imagens baseado na realimentação de relevantes. Um exemplo deste sistema pode ser encontrada em [Silva et al., 2006].
- interativo-composto: usuário interage usando mais de uma modalidade, por exemplo, texto e conteúdo.

3.3 Recuperação de Imagens Baseada em Anotações Textuais

Esta modalidade de sistemas de recuperação de imagens utiliza-se de anotações textuais para descrever o conteúdo destas imagens. Estas anotações podem ser extraídas de diversas formas, tais como:

- manual: descrição textual (metadados ou palavras-chaves definidas pelo usuário);
- semi-automático: associam entidades nomeadas do texto a termos da ontologia, utilizando-se de julgamento humano;
- automático: aplicam técnicas de Processamento de Linguagem Natural (PLN)¹, aprendizado de máquinas, extração de informação, entre outras, para associar automaticamente as entidades nomeadas ao texto.

As consultas nestes sistemas são expressas por meio de palavras-chaves, às vezes combinadas por operadores booleanos, e efetuadas por meio de técnicas de gerenciamento de banco de dados. Esta abordagem é limitada, pois é inviável, prover anotações de imagens para grandes coleções. Além disso, anotações manuais são subjetivas - uma mesma imagem pode ser interpretada de diferentes maneiras por diferentes pessoas. Revisões dos trabalhos de recuperação de imagens baseados em anotações textuais podem ser encontrados em [Chang and Hsu, 1992] e [Tamura and Yokoya, 1984]. Apesar do seu improvável sucesso para muitas aplicações, esta abordagem permanece aplicável para imagens de significado semântico especial como coleções de fotos de museus e de pinturas famosas.

Segundo [Veltkamp and Tanase, 2000] existe também uma outra categoria de sistemas de recuperação de imagens. Esta categoria inclui o serviço provido pela máquina de busca Google. Neste tipo de sistema, as imagens são extraídas de documentos (geralmente páginas html) e são indexadas com base nas informações textuais ou na *label* que as acompanham. Posteriormente, utiliza-se de técnicas tradicionais de recuperação de informação textual, para comparar a consulta provida pelo usuário por meio de palavras-chaves, com os índices das imagens. Os problemas referentes a esta técnica têm origem

¹Sistemas que trabalha com este paradigma pode ser encontrado em: <http://www.powerset.com/>

na fraca eficiência, proporcionada principalmente pelo fato de que na maioria das vezes o texto próximo a uma imagem não a descreve fielmente.

Pelo fato da técnica de anotações textuais sobre as imagens não conseguir extrair todas as suas características, a Recuperação de Imagens Baseada no Conteúdo se apresenta como solução e tem ganho importância como tema de pesquisa junto à comunidade acadêmica. CBIR é baseado no conceito de busca de imagens por similaridade, ou seja, quando um usuário submeter uma consulta (imagem exemplo), o sistema deve recuperar o conjunto de imagens mais similares à imagem de referência, organizadas em ordem decrescente de similaridade (*ranking*).

3.4 Recuperação de Imagens Baseada no Conteúdo

Como visto anteriormente a dificuldade e o custo de se produzir anotações textuais ricas e confiáveis para grandes bancos de dados de imagens, como também a subjetividade associada a estas anotações, explicam o fato do CBIR, ser de grande interesse nos dias atuais [Crucianu et al., 2004].

Os sistemas para a Recuperação de Imagens Baseada no Conteúdo, em princípio, ajudam a organizar arquivos de imagens digitais considerando o conteúdo visual extraídos automaticamente da própria imagem. Estas características podem ser de baixo nível ou alto nível. Baixo nível são as representação das características visuais das imagens como cor, textura e forma. Alto nível refere-se à semântica, como objetos, interpretações de cenas, ações [Datta et al., 2008].

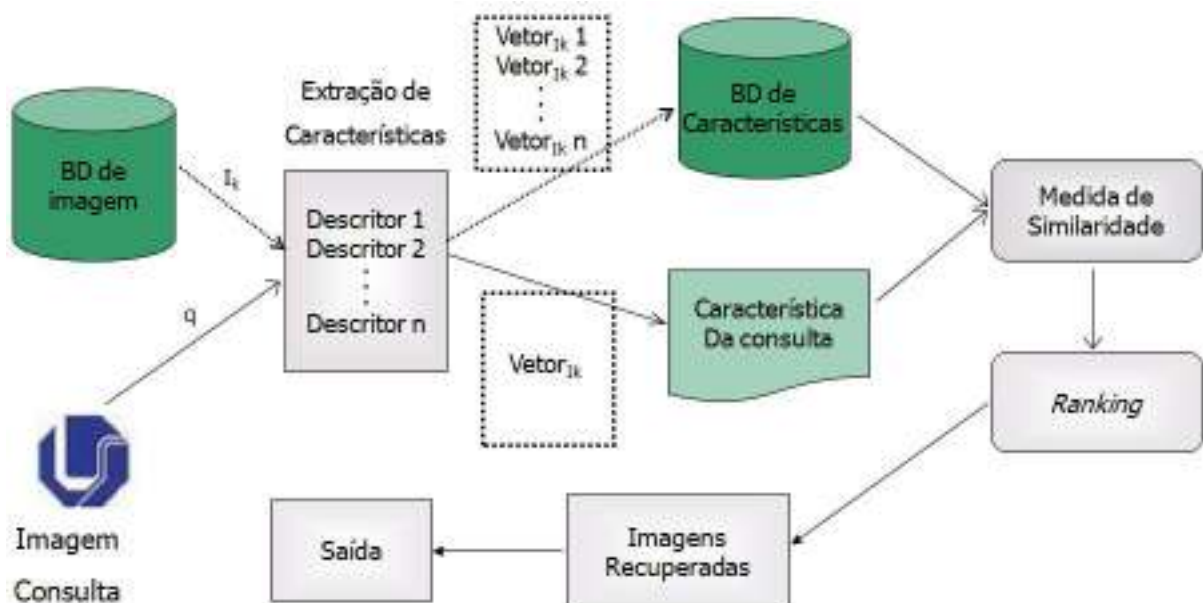


Figura 3.1: Fluxograma de um sistema CBIR típico.

A Figura 3.1 corresponde a um fluxograma típico de um sistema CBIR. Inicialmente,

a partir de um banco de dados de imagens digitais, extraem-se características sintáticas e/ou semânticas, obtendo um banco de dados de vetores de características, onde cada vetor representa uma imagem. O tamanho deste vetor dependerá de quantas características forem utilizadas na representação da imagem. Até este momento, todas as etapas são realizadas antes da interação com o usuário. A partir do momento que o usuário expressa sua consulta através de uma imagem-exemplo, o mesmo processo de extração das características feito para o banco de dados é aplicado à consulta do usuário, obtendo-se o vetor de características da imagem consulta. Então, compara-se o vetor de características da imagem consulta com os vetores de características das imagens da coleção através de alguma medida de similaridade. Com base nos valores de similaridade um *ranking* é formado e as imagens são recuperadas e exibidas ao usuário. As etapas citadas acima serão temas das próximas seções.

Muitos trabalhos e sistemas CBIR fornecem ao usuário a possibilidade de refinar suas buscas através de realimentação de relevância. A recuperação de imagens utilizando este paradigma consiste na interação do usuário com o sistema de busca com intuito de melhorar a qualidade da recuperação a partir de indicação de exemplos positivos e/ou negativos das imagens recuperadas. Este tipo de metodologia pode ser encontrado em [Silva et al., 2006] e [Zhou and Huang, 2003]. Realimentação de relevantes não faz parte do escopo deste trabalho.

É importante reconhecer as deficiências de CBIR como uma tecnologia do mundo real. Um problema com todas as abordagens atuais é a dependência em relação a semelhança visual para julgar a similaridade semântica, que pode ser problemática, devido ao "*gap semântico*"² [Smeulders et al., 2000] entre as características de baixo nível e de alto nível.

[Datta et al., 2008] acredita que as pesquisas em CBIR são promissoras o suficiente para serem úteis nas aplicações do mundo real. Os grandes motores de busca da Web, por exemplo, o Google e Yahoo!, são nomes fortes no processo de recuperação de imagens, apesar dos problemas referentes à descrição textual das imagens.

Outros *sites* de compartilhamento de imagens também têm se tornado extremamente populares como Flickr³ que possuem milhões de fotografia com conteúdo diversificado. Sistemas de gerenciamento de videos também têm feito muito sucesso na Web como o Youtube⁴. O levantamento atual se restringe a discussão relacionada apenas à imagem.

3.4.1 Extração de Característica

Extração de característica é uma das etapas mais importantes em um projeto de construção de sistemas de recuperação de imagens baseada no conteúdo. As características

²perda da informação real da imagem, que não é preservada/capturada pelas características (algoritmos de Processamento de Imagens) e a expectativa de informação total desejada pelo usuário.

³<http://www.flickr.com>

⁴<http://www.youtube.com>

extraídas das imagens devem sintetizar suas propriedades inerentes. O objetivo nessa etapa é expressar numericamente as propriedades das imagens, ou seja, obter um tipo de assinatura, uma transformação no sinal gráfico do domínio espacial para outro domínio mais apropriado (frequência dos *pixels*, por exemplo) da imagem-exemplo e de toda a coleção de imagens.

Atualmente, os métodos de representação de características das imagens mais utilizados usam: cor, textura e forma como atributos de indexação, os quais são extraídos da imagem de maneira independente. A maior parte das pesquisas atuais em CBIR são voltadas para a exploração de características de baixo nível. [Choras et al., 2007].

Boa parte dos sistemas atuais utilizam a extração da característica de baixo nível em duas fases [Bender, 2003]:

- para obter um vetor de característica que identifique algumas propriedades da imagem e armazenar este vetor em um banco de dados de características;
- a partir da imagem exemplo, calcular o seu vetor de característica e comparar com os vetores armazenados no banco de dados recuperando os que possuem os melhores índices de similaridades.

O vetor de característica é uma representação numérica da imagem, caracterizando medidas representativas dos seus aspectos visuais significativos. Ele deve atender a três considerações: reduzir a dimensionalidade dos dados, ressaltar aspectos da imagem para facilitar a percepção humana e ser invariante às transformações da imagem [Paris, 2008].

Existem duas formas de se extrair características das imagens: global e local. Uma característica global, como o nome já diz, refere-se à extração das características visuais da imagem inteira. Já uma característica local usa características de regiões ou objetos para descrever o conteúdo da imagem. Este trabalho foca-se na extração de características globais.

A seguir são mostrados os principais descritores utilizados para representar as características visuais de baixo nível das imagens.

3.4.2 Propriedades da Imagem

Muitos sistemas de recuperação de imagens utilizam, cor, forma e textura para representar uma imagem e a recuperação está baseada na similaridade das características derivadas delas. Apesar da característica cor ser um atributo confiável na recuperação de imagens, existem situações em que esta característica não discrimina imagens relevantes das não relevantes de uma forma satisfatória sendo necessário o uso de outros atributos tais como, forma e textura. A seguir, estes atributos são descritos.

COR

Ao se utilizar a característica cor no processo de recuperação de imagens deseja-se recuperar todas as imagens que possuem uma composição de cor similar, mesmo que elas sejam de contexto diferente. Conforme visto em [Antani et al., 2002], cor é uma das características mais utilizadas nos sistemas de recuperação de imagens. Além de ser fácil de se representar, é de grande importância na diferenciação de imagens por meio de máquinas.

A primeira decisão para se definir uma caracterização da cor de uma imagem é escolher o modelo de cor. Imagens coloridas podem ser armazenadas em três componentes primários formando um espaço de cor. Este espaço é uma especificação de um sistema de coordenadas tridimensionais e um subespaço dentro deste sistema onde cada cor é representada por um único ponto. Segundo [Schettinia et al., 2001], os modelos de cores mais comuns na recuperação de imagens são o RGB, CIE L^*a^*b , CIE L^*u^*v , HSV e HSI.

Em [Bender, 2003] os modelos de cor são classificados em quatro categorias de acordo com o seu propósito. Neste trabalho é focado no Sistema de Interface, que são os sistemas projetados para interagir com o usuário, e portanto, são apropriados para o CBIR. Mais detalhes das outras categorias podem ser encontrados em [Gomes and Velho, 1994]. A seguir são abordados dois modelos de cor, que são: RGB (*red*, *green* e *blue* respectivamente) e HSV (*tonalidade*, *saturação*, *brilho*) adotados neste trabalho.

RGB: o modelo de cor RGB foi utilizado neste trabalho por ser amplamente suportado em CBIR e ser um sistema nativo de diversas coleções gráficas. É constituído por um espaço tridimensional cujos componentes representam a intensidade de cada cor primária: vermelho, verde e azul, que compõem uma dada cor. Essas três imagens combinam-se de forma aditiva quando visualizadas em um monitor RGB, produzem uma imagem em cores. O valor de cada *pixel* é representado pela composição de valores $(R, G \text{ e } B) \in \{0,1\}$. Sendo assim, o espaço de cores RGB toma a forma de um cubo como pode ser visto na Figura 3.2.

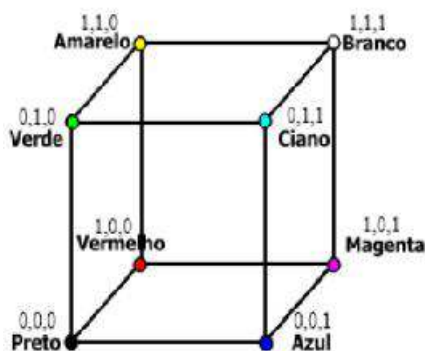


Figura 3.2: Espaço RGB representado em Cubo.

Cada uma das cores primárias corresponde a um dos vértices do cubo localizados sobre os eixos do espaço, em que apenas uma das coordenadas não é nula [Gonzales and Woods, 2002].

HSV: o modelo de cor HSV decompõe a cor em sua tonalidade predominante, sua saturação e seu brilho.

A tonalidade predominante de uma imagem verifica o tipo de cor, abrangendo todas as cores do espectro, desde o vermelho até o violeta, mais o magenta. Atinge valores de 0 a 360, mas para algumas aplicações, esse valor é normalizado de 0 a 100%.

A Saturação que também pode ser chamada de "pureza" calcula os valores de tom de cinza da imagem. Quanto menor o valor da saturação, mais com tom de cinza aparecerá a imagem. Quanto maior o valor, mais "pura" é a cor. Por fim, o brilho ou valor define o brilho da cor.

Este modelo de cor é considerado o mais próximo do sistema RGB. Ele é utilizado pelo fato de decompor a cor em sua tonalidade predominante e pureza (o que permite uma função de comparação de características baseadas em cor uma discriminação adequada de tonalidades semelhantes) e na componente brilho, que pode variar em diferentes cenas ou condições.

Uma representação tridimensional do espaço HSV é um cone, onde o eixo vertical central representa a variação do brilho. A tonalidade é definida como sendo o ângulo, variando de $[0, 2\pi]$, relativo ao eixo do Vermelho, sendo o ângulo vermelho igual a 0, o verde a $2\pi/3$, o azul $4\pi/3$ e o vermelho novamente em 2π . A saturação indica a pureza da cor, e é medida pela distância radial em relação ao eixo central, tendo como valor 0 no centro e valor 1 na face externa do cone. Um exemplo gráfico do modelo de cor HSV pode ser visualizado na Figura 3.3.

A vantagem de se trabalhar com essa representação de cor encontra-se na possibilidade de separar a intensidade da informação, tonalidade e saturação, bem como, na relação que existe entre essas componentes, muito próxima da forma na qual o homem percebe a cor.

Conversão de RGB para HSV: é possível transformar qualquer imagem do espaço de cor RGB para uma imagem no espaço de cor HSV aplicando as seguintes regras: as cores nos modelos HSV são obtidas a partir das informações RGB com respeito aos valores normalizados do vermelho (r), verde (g) e azul (b), dados por:

$$r = \frac{\text{vermelho}}{255} \quad (3.1)$$

$$g = \frac{\text{verde}}{255} \quad (3.2)$$

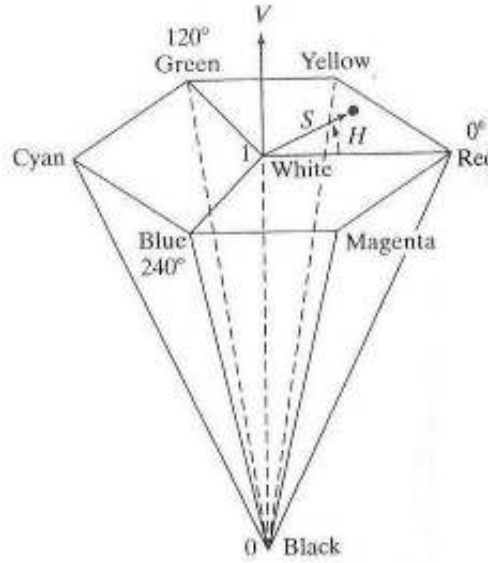


Figura 3.3: Espaço de cor HSV [Gonzales et al., 2004].

$$b = \frac{azul}{255} \quad (3.3)$$

Após obter os canais r , g e b normalizados, as equações abaixo são utilizadas para o cálculo do H (tonalidade), S (saturação) e V (*brilho*).

$$H = \arctan \frac{\sqrt{(r-g)^2 + (r-b)(g-b)}}{\sqrt{(r-g)^2 + (r-g)(g-b)}} \quad (3.4)$$

$$S = 1 - \frac{3}{r+g+b} \times \min(r, g, b) \quad (3.5)$$

$$V = \frac{(r+g+b)}{3} \quad (3.6)$$

A vantagem dessa representação de cor está na possibilidade de separar os canais. Esse fato torna o modelo HSV uma ferramenta ideal para o desenvolvimento de algoritmos de processamento de imagens. Portanto, têm-se 3 canais de cor (H, S, V), que serão utilizados pelo descritor de cor, chamado Momento de Cor, que é abordado a seguir.

Momento de Cor: momentos de cor [Stricker and Orengo, 1995] caracterizam as imagens em termos da distribuição das cores nos *pixels* da imagem. São normalmente dados por três medidas estatísticas: média, desvio padrão e obliquidade. Momentos de cor podem ser derivados do espaço de cor HSV correspondendo aos canais tonalidade, saturação e brilho da seguinte forma: sendo N o número de *pixels* de uma imagem e p_{ij} o valor do j -ésimo *pixel* no i -ésimo canal de cor, os três primeiros momentos: média (E_i), desvio

padrão (σ_i) e obliquidade (s_i) são computados por:

$$E_i = \frac{1}{N} \sum_{j=1}^n p_{ij} \quad (3.7)$$

$$\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^n (p_{ij} - E_i)^2 \right)} \quad (3.8)$$

$$s_i = \sqrt[3]{\left(\frac{1}{N} \sum_{j=1}^n (p_{ij} - E_i)^3 \right)} \quad (3.9)$$

A utilização deste descritor de cor gera um vetor de característica de nove dimensões, representando respectivamente a média, desvio padrão e obliquidade do canal H (tonalidade); média, desvio padrão e obliquidade do canal S (saturação) e por fim a média, desvio padrão e obliquidade do canal V (brilho).

Demais Descritores de Cor

Outros descritores de cor comumente usados para descrever as informações de cores das imagens são encontrados na literatura entre os quais destaca-se o histograma de cor [Swain and Ballard, 1991] que descreve a frequência global das cores de uma imagem. Sua vantagem esta na compactabilidade e performance. Já a desvantagem esta na sua alta dimensionalidade (necessária para sua eficácia) e o fato destes não considerarem a localização espacial das cores. Assim, imagens muito diferentes podem ter representações semelhantes.

No intuito de incorporar informações espaciais em um descritor de cor [Pass and Zabith, 1996] propôs o método denominado vetor de coerência de cores. Outro descritor conhecido, correlograma de cores, pode ser encontrado em [Huang et al., 1997].

Textura

Outro modelo de extração de característica de baixo nível utilizado em CBIR é conhecido como textura. Apesar de não haver uma definição estrita do conceito de textura de imagem, para a visão humana a textura pode ser bem perceptível e representa uma fonte de informação visual muito rica acerca da natureza e estrutura tridimensional das imagens. São aspectos presentes nas imagens, caracterizados pela relação entre *pixels*, ao contrário da característica cor que é uma propriedade individual de cada *pixel*.

De acordo com [Haralick, 1973], as texturas definem uniformidade, densidade, aspereza, regularidade, intensidade, entre outras características em uma imagem e são encontradas em muitos tipos de superfícies tais como, superfícies de madeira, tecidos, vegetação e nuvens. A Figura 3.4 mostra alguns exemplos de imagens com textura. Textura se caracteriza pela repetição de um modelo sobre uma região, sendo este modelo repetido

em sua forma exata ou com pequenas variações. A partir de sua análise é possível distinguir regiões que apresentem as mesmas características de refletância, e portando, mesmas cores em determinada combinação de bandas. Isso torna a textura um excelente descritor regional, contribuindo para uma melhor precisão dos processos de reconhecimento, descrição e classificação de imagens. Apesar de seus benefícios, seu processo de reconhecimento exige um alto nível de sofisticação e complexidade computacional [Ebert, 1994].

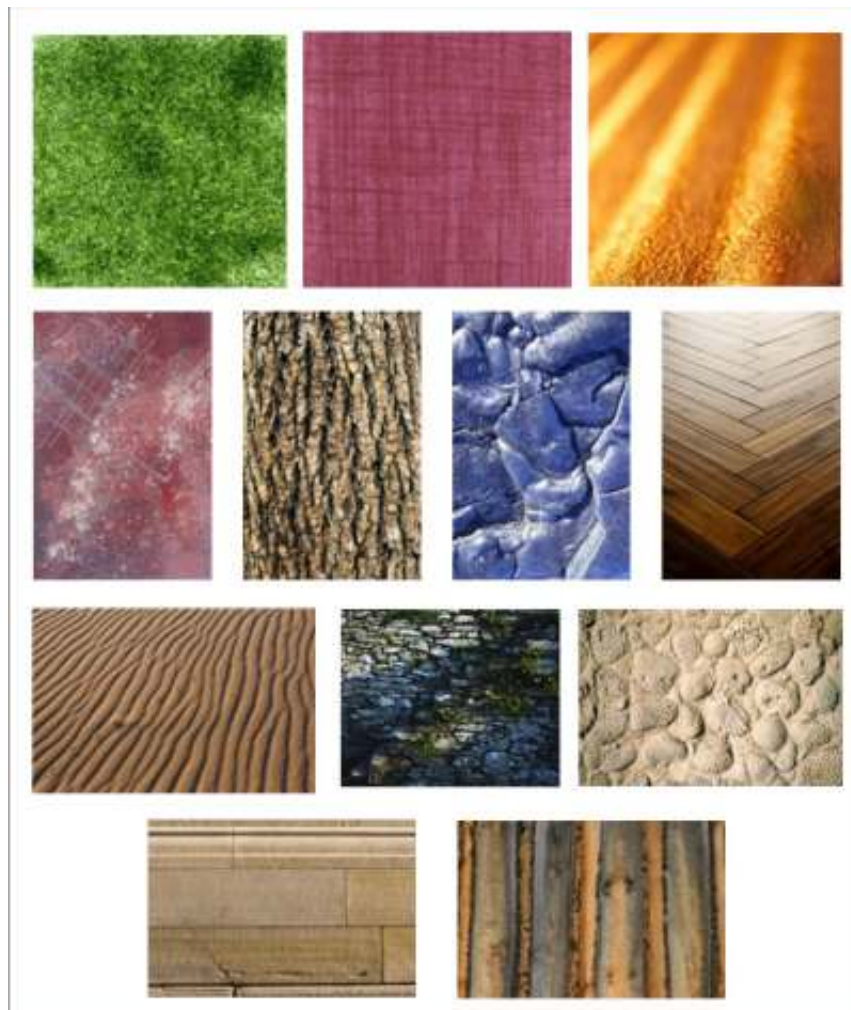


Figura 3.4: Imagens com textura (www.fotosearch.com.br)

Existem três abordagens principais usadas em classificação de imagens para a descrição de texturas, que são: a estatística, a estrutural e a espectral [Long and Leow, 2000]

- Abordagens estatísticas: categorizam texturas de acordo com medidas estatísticas de características visuais, tais como, grossura, granularidade, regularidade, entre outros. Estas medidas são fortemente baseadas nos aspectos da percepção humana da textura;
- Abordagens estruturais: caracterizam texturas de acordo com o relacionamento espacial entre *pixels* de imagens. Utilizam a idéia de que texturas são compostas de

primitivas dispostas de forma aproximadamente regular e repetitiva, de acordo com regras bem definidas. Como exemplo, pode-se citar a descrição da textura baseada em linhas paralelas regularmente espaçadas;

- Abordagens espectrais: caracterizam textura como propriedades de transformadas de Fourier ou nos resultados de filtragem das texturas por filtros apropriados.

Para mais detalhes destas abordagens consulte [Turceyan and Jain, 1998].

Forma

A última propriedade usada para extração de característica apresentado neste trabalho é denominado de característica forma. A caracterização da forma de objetos exige o desenvolvimento de técnicas que concedam uma descrição total da borda do objeto ou que descrevam as características morfológicas de uma região. A cognição humana tem na percepção da forma a melhor alternativa para reconhecer e classificar um objeto. Entretanto, em sistemas CBIR, esta é uma abordagem que apresenta a maior dificuldade, pelo fato de ter que segmentar e conhecer o tamanho dos objetos contidos na imagem.

Várias técnicas de extração de bordas são encontradas na literatura. No trabalho de [Zhang and Lu, 2004], estas técnicas são classificadas em dois grupos: métodos baseados no contorno e métodos baseados em regiões. Já estes podem ser divididos em global e estrutural e cada um apresenta várias técnicas de extração da forma específica, conforme pode-se ver na Figura 3.5.

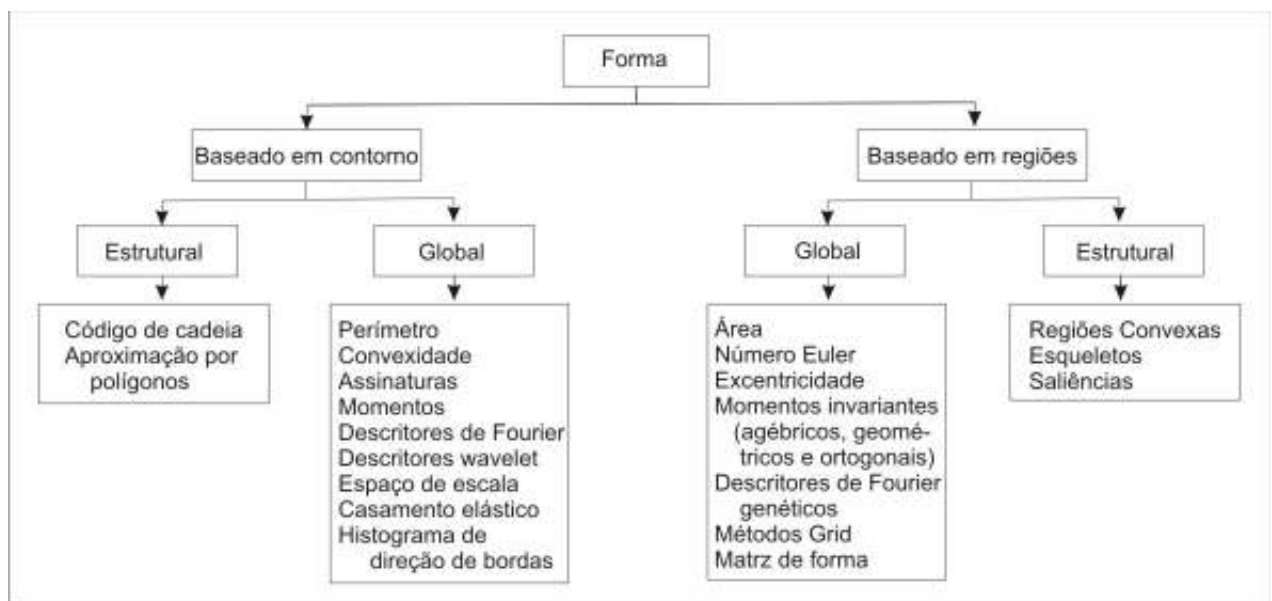


Figura 3.5: Taxinomia da característica forma e suas técnicas [Zhang and Lu, 2004].

Dentro das técnicas baseadas em contornos são exploradas somente as informações de bordas (fronteiras), que podem ser global (abordagem contínua) ou estrutural (abordagem

discreta). O método global apresenta técnicas como Histograma de direção de bordas que é um dos descritores de forma mais populares [Brandt, 1999], pois principalmente não necessita de segmentar o objeto. Mais detalhes sobre as diversas técnicas de extração de forma podem ser encontradas em [Zhang and Lu, 2004] e [Zhang, 2002].

3.5 Vetor de Característica

Vetor de característica é a representação numérica da imagem, caracterizando medidas dos aspectos representativos do objeto. Portanto, tem-se uma nova representação do objeto ou imagem e essa deve atender a três considerações: i) reduzir a dimensionalidade dos dados; ii) ressaltar aspectos da imagem para facilitar a percepção humana; iii) ser invariante às transformações da imagem.

Sua importância se deve pela dificuldade de se manipular grandes quantidades de informações contidas em uma imagem. Assim, a geração do vetor de características é um processo que calcula novas variáveis, a partir da imagem original. Esse processo visa extrair informações contidas na imagem que permitirão codificá-la adequadamente [Paris, 2008].

3.6 Medidas de Similaridade

Em um sistema de recuperação de imagens, após a extração das características da imagem e seu respectivo armazenamento em um vetor de característica, fazem-se necessários medidas que comparem a similaridade entre os vetores de característica da base de dados bem como o vetor de característica da imagem de consulta.

A quantificação do valor de similaridade é obtida por meio de medidas de distância entre dois objetos e retorna um valor real e positivo que indica o grau de semelhança entre eles. Quanto maior este valor, menor a similaridade entre os objetos comparados, quanto mais próximo de zero maior a similaridade dos objetos comparados. Se a distância for igual a zero indica que as imagens são iguais ou similaridade total.

As funções de distâncias utilizadas em sistemas CBIR podem estar definidas dentro de um espaço métrico e podem ser definidas da seguinte maneira: $\{P, d()\}$, onde P é conjunto de elementos do domínio e d é a distância entre os elementos. Considere o conjunto dos elementos s_1, s_2, s_3 , pertencentes ao domínio P , uma função de distância $d(s_1, s_2)$ para o espaço métrico $M \{P, d()\}$ deve satisfazer as seguintes propriedades:

- i) $d(s_1, s_2) \geq 0$;
- ii) $d(s_1, s_2) = 0$, se e somente se, $s_1 = s_2$;
- iii) $d(s_1, s_2) = d(s_2, s_1)$;

- iv) $d(s_1, s_2) \leq d(s_1, s_3) + d(s_3, s_2)$.

Em relação as quatro propriedades citadas acima pode-se concluir os seguintes aspectos: pela propriedade i) a distância entre s_1 e s_2 é igual a zero se (s_1 e s_2 são idênticas) ou maior do que zero se (s_1 e s_2) possuem um grau de diferença determinado por $d(s_1, s_2)$. Caso $d(s_1, s_2)$ for menor do que zero, s_2 seria mais semelhante a s_1 do que a própria imagem s_1 . Caso a propriedade ii) não seja satisfeita existem imagens na coleção que são diferentes, mas não é possível discriminá-las com a função de distância utilizada. Caso a propriedade iii), que se refere à simetria entre as imagens, não seja satisfeita ocorreria da imagem s_1 ser similar a s_2 , mas s_2 não ser igual a s_1 . Já a propriedade da desigualdade triangular iv) diz que entre 3 imagens s_1, s_2, s_3 , se s_1 é similar a s_2 e s_1 é igual a s_3 então s_2 e s_3 devem ser similares também.

A seguir são apresentadas algumas das medidas de distâncias mais utilizadas em CBIR. A escolha de uma função de distância adequada é de extrema importância. No entanto, não existe uma formalização definida e a função é identificada por meio de heurísticas dependente das características dos dados [Paris, 2008].

Distância Minkowski - É uma das medidas de distâncias mais utilizadas, por ser aplicada a domínios vetoriais e por cada coordenada dos vetores de características serem independentes entre si e de igual importância. Ela é definida de acordo com a equação 3.10:

$$D(s_1, s_2) = \sqrt[p]{\sum_{i=1}^n |s_{1i} - s_{2i}|^p} \quad (3.10)$$

onde n é a dimensão do espaço vetorial e quando $p = 1, 2$ temos as seguintes medidas de distância:

Distância Euclidiana: é um caso particular da Distância Minkowski, quando $p = 2$.

$$D(s_1, s_2) = \sqrt{\sum_{i=1}^n |s_{1i} - s_{2i}|^2} \quad (3.11)$$

Distância em Quadras: (*city-block*) - outro caso particular da Distância Minkowski, quando $p = 1$.

$$D(s_1, s_2) = \sum_{i=1}^n |s_{1i} - s_{2i}| \quad (3.12)$$

Outra abordagem é conhecida como a medida de similaridade do cosseno. A distância é calculada usando a correlação entre os vetores que representam as imagens, quantificada pelo cosseno do ângulo formado por s_1 e s_2 . Desta forma, em um espaço vetorial de dimensão n , a similaridade (sim) entre dois vetores s_1 e s_2 é calculada pelo do cosseno do ângulo formado por estes vetores, por meio da seguinte fórmula:

$$\text{sim}(s_1, s_2) = \frac{\sum_{i=1}^n W_{s1i} W_{s2i}}{\sqrt{\sum_{i=1}^n W_{s1i}^2} \sqrt{\sum_{i=1}^n W_{s2i}^2}} \quad (3.13)$$

onde s_1 é o vetor da imagem 1; s_2 é o vetor da imagem 2. W_{s1i} é o peso da característica i da imagem s_1 e W_{s2i} é o peso da característica i da imagem s_2 .

Distância Quadrática, *Mahalanobis*, *Canberra*, *Chebyshev*, *Bottleneck*, *Hausdorff*, *Earth Mover's* e *Banach-Mazar* fazem parte de uma imensa lista de medidas de distância utilizadas na comparação de vetores de características.

Destas medidas citadas anteriormente a mais utilizada é a distância Minkowski. Dos trabalhos pesquisados, foi encontrado uma quantidade significativa que faz uso desta medida de distância para calcular a similaridade entre a imagem de consulta e as imagens do banco de dados. Dentre os quais, destacam-se o sistema MARS [Rui et al., 1997] que utiliza a distância Euclidiana para calcular a similaridade de características de textura entre imagens. O sistema Netra [Ma and Manjunath, 1999], também a utiliza para calcular a similaridade baseada em cor e forma e distância em Quadras para computar a similaridade de características de textura. Em CIRES [Iqbal and Aggarwal, 2002] e Blob-world [Carson et al., 1999] a comparação de características de forma e textura também é obtida por meio da distância Euclidiana e por fim, em [Silva et al., 2006], a distância Euclidiana foi utilizada para computar as similaridades da característica cor e a distância em Quadras para computar a similaridade de características de textura e forma.

3.7 Avaliação dos Resultados

Com a existência de vários sistemas CBIR é necessário que haja uma avaliação da qualidade de recuperação destes sistemas. Portanto, surge a necessidade de métricas para avaliar os resultados do processo de recuperação. Do ponto de vista dos pesquisadores, uma avaliação eficiente tornaria possível a comparação dos resultados entre duas abordagens de sistemas CBIR.

CBIR é considerado essencialmente um problema de recuperação de informação. Portanto, as métricas de avaliação foram aprovadas automaticamente a partir dos sistemas de recuperação da informação. Como no campo da recuperação da informação como em CBIR uma das principais medidas de eficácia de recuperação utilizadas são a precisão e a revocação [Datta et al., 2008].

Outras medidas de avaliação como visto na Seção 2.8 podem ser utilizadas em CBIR da mesma forma que na recuperação textual.

3.8 Considerações Finais

Neste capítulo técnicas de recuperação de imagens foram apresentadas. A recuperação de imagens pode ser dividida em duas grandes abordagens: baseada em anotações textuais e baseada no conteúdo (extração automática de características visuais). A primeira apresenta dois problemas cruciais: a subjetividade das anotações e a dificuldade desta técnica para grandes coleções como a Web, por exemplo. Dentro da Recuperação de Imagens Baseada no Conteúdo, foram apresentadas as principais técnicas para extração de características considerando-se atributos classificados como primitivos da imagem (cor, textura e forma). Estas técnicas tentam contornar os problemas das anotações manuais. Finalizando o Capítulo foi apresentado as principais medidas avaliação de recuperação usadas para avaliar os sistemas de recuperação de imagens.

O próximo capítulo apresenta a nossa contribuição para a área de CBIR. Como os sistemas de recuperação de imagens baseadas no conteúdo apresentam uma qualidade satisfatória em seu processo de recuperação, mas baixo desempenho na velocidade de recuperação iremos propor uma adaptação nos sistemas CBIR. Com intuito de acelerar o processo de recuperação sem perda de qualidade na recuperação.

Capítulo 4

Um Estudo da Característica Cor para Construção do Índice Invertido

Neste capítulo apresenta-se uma proposta de utilização do índice Invertido para recuperação de imagens baseado em conteúdo. O objetivo é acelerar o processamento das consultas, sem perda de qualidade na resposta. Sistemas clássicos de CBIR tendem a serem lentos para grandes coleções de imagens, como por exemplo, a Web, devido ao fato de utilizarem em seu cálculo de similaridade, na maioria dos casos, a distância de *Minkowski*, especificamente a distância euclidiana. Nestes sistemas, quando não se utiliza uma estrutura de indexação multidimensional, o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os vetores das imagens do banco de dados, tornando a recuperação computacionalmente cara.

Este trabalho lida com este problema grave dos sistemas de Recuperação de Imagens Baseado no Conteúdo, utilizando como ferramenta principal o índice invertido para melhorar o desempenho no processo de recuperação.

Com objetivo de avaliar a proposta, foi necessário a construção de dois sistemas aqui chamados de CBIR-I e CBIR-II, que serão detalhados neste capítulo. Por meio de experimentos descritos no próximo capítulo, foram feitas análises comparativas do número de operações aritméticas efetuadas pelos dois sistemas no cálculo da similaridade mostrando o ganho significativo no tempo de processamento e sem impactos negativos na qualidade da recuperação.

Cabe destacar, que a maior contribuição deste trabalho, é mapear um identificador de indexação da coleção para implementar um sistema de recuperação de imagens que utilize este identificador.

Este capítulo apresenta inicialmente um levantamento dos trabalhos correlatos, além de detalhar os dois sistemas construídos neste trabalho.

4.1 Trabalhos Correlatos

Muitos projetos foram iniciados nos últimos anos para investigar e desenvolver métodos eficientes para CBIR. Como resultado, vários sistemas CBIR acadêmicos e comerciais foram desenvolvidos. Nesta seção, alguns sistemas são introduzidos brevemente. O desempenho global das implementações atuais permanecem ainda muito modesto quanto à tarefa de desenvolver um sistema de recuperação de imagem, comparável aos sistemas de recuperação textuais [Veltkamp and Tanase, 2000].

Sistemas Gerais de Recuperação de Imagens Baseada no Conteúdo

Estes sistemas são implementados em diversos ambientes. Estes sistemas operam com banco de dados de imagens predeterminados, oposto aos sistemas de busca de imagens da Web, que serão apresentados brevemente.

O mais conhecido sistema de recuperação de imagens baseado no conteúdo é provavelmente o sistema da IBM chamado de *Query By Image Content* (QBIC) [Niblack et al., 1993]. Ele foi o primeiro sistema comercial de CBIR. Este sistema foi muito utilizado para avaliação e comparação de sistemas de recuperação de imagens baseado no conteúdo. Avaliações de características incluem cor, textura e forma. Utiliza a distância Euclidiana para o cálculo da similaridade.

O sistema KIWI [Louupias and Bres, 2001] utiliza momentos de cor e a distância Euclidiana para computar a similaridade. A consulta neste sistema pode ser feita por imagem exemplo. Já em [Rui et al., 1997], o sistema chamado de MARS foi implementado com a característica de baixo nível cor, com o espaço de cor HSV, além da textura e forma. Faz uso da distância Euclidiana para computar a similaridade de características de textura entre imagens. Outros sistemas usando a característica cor são apresentados em [Srihari et al., 2000], [Laaksonen et al., 2000] e [Mills et al., 2000].

Todos estes sistemas apresentam bons resultados de avaliação. Mas apresentam problemas em relação ao tempo de processamento para grandes coleções como a Web. Pois, em todos, o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os vetores das imagens do banco de dados, tornando a recuperação computacionalmente cara. Por este motivo sistemas CBIR não são usualmente utilizados em motores de busca da Web.

Outra medida bastante utilizada é o cálculo do cosseno em suas implementações. Em Circus [Pecenovie, 1998], as características utilizadas neste sistema são cor e textura. A similaridade entre as imagens é dada pelo cosseno do ângulo formado entre os dois vetores. Já o sistema DrawSearch [Huijsmans and Smeulders, 1999], trabalha com cor, forma e textura. Em Surfimage [Nastar et al., 1998] a combinação de várias características de baixo nível, quanto de alto nível é usada. Ambos os sistemas utilizam o cálculo do cosseno para medir a similaridade entre a imagem de consulta e a coleção de imagens.

Estes trabalhos fazem uso da medida de similaridade do cosseno, mas não utilizam o índice invertido.

O uso do índice invertido em recuperação de imagem pode ser encontrado em [Squire et al., 1999]. Foi constatado que, com o uso do índice invertido, é possível utilizar vetores de características de $O(10^4)$ dimensões, ao restringir a pesquisa apenas para as características presentes na consulta. Nesse trabalho, foi utilizada as características cor e textura para construção do vetor de característica. Além da abordagem do peso do termo juntamente com a realimentação de relevantes. Os autores citam o uso do índice invertido sem detalhar o processo de geração dos "termos", isto é, o mapeamento das características de baixo nível para as entradas no índice invertido. Lá são gerados $O(10^3)$ entradas. Aqui, conforme apresentado na seção 4.2.1, é descrito em detalhes este mapeamento e, além disso, usa-se um número bem menor de entradas ("termos"). Outro trabalho que utilizou a estrutura do índice invertido pode ser visto em [Müller et al., 1999].

Com intuito de acelerar o processo de recuperação de imagens foram propostas diversas estruturas de indexação de imagens. Embora se tenha um elevado número de técnicas de indexação, não há ainda um consenso universal de um extrator de característica, indexação e técnica de recuperação disponível. Em [Sudhamani and Venugopal, 2008] apresenta-se uma revisão atualizada de diversas estruturas de indexação de recuperação de imagens mencionadas na literatura. Um exemplo de trabalho que faz uso de técnicas de indexação multidimensional pode ser visto em [El-Kwae and Kabuka, 2000] no qual denomina-se *Two Signature Multi-Level Signature File* (2SMLSF). A 2SMLSF codifica a informação da imagem em uma assinatura binária e cria uma estrutura de árvore que pode ser utilizada na pesquisa de forma eficiente para satisfazer uma necessidade do usuário. Em [El-Kwae and Kabuka, 2000] é citado um exemplo onde em uma grande coleção de imagens, o custo de armazenamento pode ser reduzido em 78% e o desempenho de processamento pode chegar a ter um aumento de até 98%.

Como mencionado anteriormente o foco maior deste trabalho é mapear um identificador de indexação da coleção para programar um sistema de recuperação de imagens que utilize este identificador. Portanto, o uso de estruturas de indexação multidimensionais, não faz parte do escopo desta dissertação.

Busca de imagens na Web

Outra área importante está relacionada com a recuperação de informações visuais e de multimídia a partir da Web. Os atuais motores de busca de textos, como o AltaVista ¹ e Google ², não são adequados para CBIR. Motores de busca de imagens enfrentam os mesmos problemas que o texto com base nos mecanismos de busca, como o enorme tamanho da coleção, a diversidade e a natureza dinâmica da Web [Koskela, 1999].

¹<http://www.altavista.com/image/default>

²<http://images.google.com.br>

Dentre os sistemas de busca de imagens da Web, destaca-se o site Google, baseado no texto adjacente à imagem, presente na legenda da imagem e dezenas de outros fatores que determinam o conteúdo da imagem. A pesquisa de Imagens do Google tem mais de 390 milhões de imagens indexadas e disponíveis para visualização.

Outros sistemas podem ser encontrados na literatura como o WebSeek [Smith, 1997]. Ele utiliza palavras-chaves tanto textuais, por exemplo, a partir de endereços URL e *tags* HTML, e informações da característica cor para categorizar as imagens. Possui 665.000 imagens em sua coleção.

ImageRover [Sclaroff et al., 1997] combina características textuais e visuais em um único índice para busca baseada no conteúdo de uma coleção de imagens da Web. Característicos textuais são capturados em um vetor usando indexação semântica latente (LSI), baseada no texto em que contenham o documento HTML.

Observa-se ao usar esses sistemas na Web uma melhora significativa na velocidade de recuperação, mas conforme visto em alguns exemplos ver Figura 4.1) apresentam o problema em relação a sua fraca eficiência e qualidade de recuperação, proporcionada pelo fato de que na maioria das vezes o texto próximo a uma figura não a descreve fielmente.

Considere, por exemplo, uma consulta no sistema de busca de imagens do Google, que o usuário esteja interessado em pesquisar imagens do animal puma. Desta forma, a palavra-chave "puma" deverá ser digitada. As 10 primeiras imagens recuperadas são mostradas na Figura 4.1. Observe que as imagens são de diversas categorias como marca de material esportivo, carro, etc.



Figura 4.1: Consulta por "Puma" no site Google retorna um resultado misturado de imagens.

Um levantamento de sistemas de recuperação de imagens para a Web pode ser visto em [L.Kherfi et al., 2004]. O artigo analisa as principais características dos sistemas existentes e mais citados na literatura abordando a captura, indexação e recuperação das imagens nesses sistemas.

Este trabalho propõe uma nova abordagem, tanto para indexação das imagens quanto para o cálculo da similaridade em CBIR. A ideia é aproveitar a vantagem dos sistemas CBIR, que conseguem um bom desempenho na qualidade de recuperação, junto com a vantagem dos sistemas de recuperação de imagens da Web, que apresentam um alto desempenho na velocidade de recuperação.

Para isto, dois sistemas foram construídos, aqui chamados de CBIR-I e CBIR-II. O primeiro reproduz uma forma clássica ao lidar com o problema e será utilizado para comparações, o segundo é o principal objeto deste trabalho.

4.2 CBIR-I

Em [Veltkamp and Tanase, 2000], foram analisados 56 sistemas de recuperação de imagens baseado no conteúdo. Destes, 46 utilizam a característica cor, 38 utilizam textura e 29 a característica forma. Esta preferência pela característica cor, se deve pelo fato de sua simples implementação, comparada com as outras características, além de obter bons resultados.

Baseado neste argumento, o primeiro sistema implementado para análise e testes da proposta foi um sistema de recuperação de imagens baseado no conteúdo, que faz uso da característica de baixo nível cor. Foi utilizado o extrator de cor, Momentos de Cor. Neste extrator, inicialmente, as imagens são convertidas do espaço de cor RGB para o espaço de cor HSV, conforme mostrado na Seção 3.4.2. Momentos de cor caracterizam as imagens em termos da distribuição das cores nos *pixels* da imagem. São normalmente gerados por três medidas estatísticas: média, desvio padrão e obliquidade. Os momentos de cor podem ser derivados do espaço de cor HSV correspondendo aos canais tonalidade, saturação e brilho conforme as fórmulas 3.7, 3.8, 3.9.

Calcula-se então os três momentos acima para cada canal de cor ($E_H, \sigma_H, s_H; E_S, \sigma_S, s_S; E_V, \sigma_V, s_V$). Assim, o vetor de característica da imagem terá nove posições distribuídas conforme a Tabela 4.1.

No CBIR-I o banco de dados de característica é construído com os valores encontrados nas equações 3.7, 3.8 e 3.9, para cada canal de cor HSV e calcula a similaridade baseada na distância Euclidiana, dada pela equação 4.1.

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{k=1}^n |x_k - y_k|^2} \quad (4.1)$$

onde x_k é o valor correspondente a um elemento do vetor \vec{x} que representa uma imagem do banco de dados de característica e y_k é a k -ésima posição do vetor de característica \vec{y} que representa a imagem de consulta.

Como frisado anteriormente, para grandes coleções, esta forma de calcular a simi-

Tabela 4.1: Descrição do vetor de característica resultante da extração de característica da imagem usando os três momentos do espaço de cor HSV

Vetor	Descrição
1	média da tonalidade (MH)
2	desvio padrão da tonalidade (DH)
3	obliquidade da tonalidade (IH)
4	média da saturação (MS)
5	desvio padrão da saturação (DS)
6	obliquidade da saturação (IS)
7	média do brilho (MV)
8	desvio padrão do brilho (DV)
9	obliquidade do brilho (IV)

laridade torna a recuperação das imagens computacionalmente cara, pois o processo de recuperação inclui o cálculo da distância entre o vetor da imagem de consulta e todos os vetores das imagens do banco de dados. Por este motivo, este processo se torna inviável para grandes coleções de imagens.

Com o objetivo de construir um sistema viável para recuperação de imagens em grandes bases de dados como a Web, propõe-se uma nova abordagem tanto para indexação das imagens quanto para o procedimento do cálculo da similaridade em CBIR.

Para tornar a nova abordagem possível foi necessário estudar e analisar a distribuição dos valores dos vetores de características gerados pelo CBIR-I de acordo com a Tabela 4.1.

4.2.1 Análise dos vetores de características do CBIR-I

Foram analisados a distribuição dos vetores de características de 2 bancos de dados, descritos a seguir.

O primeiro banco de dados de imagens utilizado foi a base de dados Corel-1000, que é um subconjunto do banco de dados Corel. As imagens coloridas possuem uma resolução padrão de 384X256 ou de 256X384 *pixels*. O banco de dados é composto por 10 categorias (África, praia, edifícios, ônibus, dinossauros, elefantes, flores, comidas, cavalos e montanhas), com 100 imagens em cada. A Figura 4.2 contém uma amostra deste banco de dados.

Gerado o banco de dados de características é feito um mapeamento destes vetores e calcula-se a distribuição dos valores em cada posição do vetor separadamente. Com isto, têm-se nove gráficos de ocorrência dos valores do momento de cor, conforme mostrado na Figura 4.3.

A segunda base de dados utilizada foi a coleção BD-10000. É um banco de dados



Figura 4.2: Amostra do Banco de Dados Corel 1000.

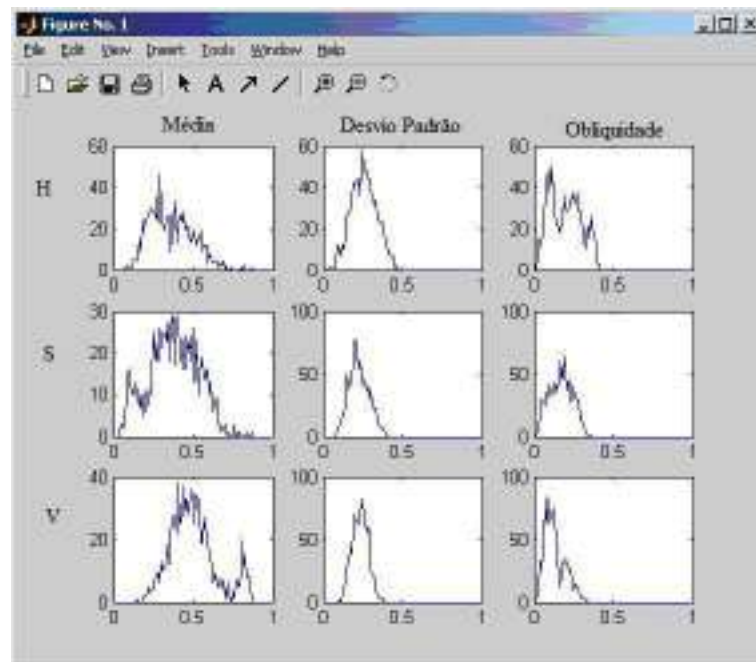


Figura 4.3: Distribuição de ocorrência das 9 característica após a extração da característica cor, momentos de cor, da base de dados corel-1000. O eixo x corresponde ao valor da característica e o eixo y à frequência de cada valor.

de 10000 imagens reais cobrindo uma ampla variedade de categorias semânticas. Sendo que 1000 dessas imagens são transportadas da coleção Corel-1000. As outras 9000 imagens foram coletadas de bases de dados públicas disponíveis na Web, principalmente de [CalPhotos,] [Silva, 2007]. Além das classes existentes do Corel-1000 esta coleção apresenta 3 novas classes (avião, carro e moto). A Figura 4.4 contém uma amostra deste banco de dados

Do mesmo modo que a coleção anterior, o mapeamento dos vetores de características desta coleção foi realizado. A distribuição dos valores em cada posição do vetor pode ser visualizada na Figura 4.5.



Figura 4.4: Amostra do Banco de Dados BD-10000.

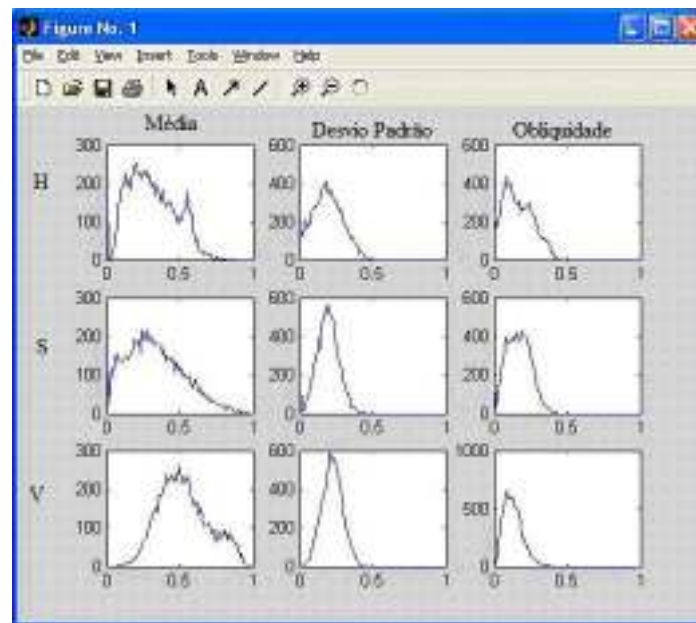


Figura 4.5: Distribuição de ocorrência das 9 características após a extração da característica cor, momentos de cor, da base de dados BD-10000. O eixo x corresponde ao valor da característica e o eixo y à frequência de cada valor.

Ao analisar matematicamente os gráficos das Figuras 4.3 e 4.5, observa-se que estas distribuições são semelhantes à distribuição normal da estatística e probabilidade.³

³No intuito de verificar se estas distribuições podem ser consideradas como uma distribuição normal é feito um estudo desta categoria de distribuição na Seção seguinte deste trabalho.

4.2.2 Distribuição Normal

A distribuição normal é a distribuição de probabilidade mais freqüente em estatística e probabilidade. Esta distribuição é unimodal e simétrica em relação a sua média. Suas características fundamentais são a média e o desvio padrão.

Segundo [Levine et al., 2000] a distribuição normal é importante na estatística por: i) inúmeros fenômenos contínuos tendem a segui-la ou se aproximar; ii) utilizá-la para aproximar várias distribuições de probabilidade discreta; iii) oferece a base para a inferência estatística clássica devido à sua afinidade com o teorema do limite central.

Propriedades da Distribuição Normal

A distribuição normal possui várias propriedades teóricas importantes, mas [Levine et al., 2000] destaca as seguintes: i) em termos de aparência ela é simétrica e tem formato de um sino; ii) suas medidas de tendência central (média aritmética, mediana, moda, média de intervalo e média das juntas) são todas idênticas; iii) sua "dispersão média" é igual a 1.33 desvio padrão. Isto significa que o intervalo interquartil está contido dentro de um intervalo de dois terços de um desvio padrão abaixo da média aritmética e dois terços de um desvio padrão acima da média; iv) Sua variável aleatória associada possui um intervalo infinito ($-\infty < x < +\infty$).

Na prática, algumas das distribuições observadas podem somente aproximar destas propriedades teóricas. Isso ocorre pela distribuição da população subjacente ser apenas aproximadamente normal ou a amostra real desviar das características teóricas esperadas. Para um fenômeno ser aproximado de um modelo da distribuição normal: i) seu polígono pode ser semelhante ao formato de um sino e ter aparência simétrica; ii) suas medidas de tendência central podem divergir pouco uma da outra; iii) o valor do seu intervalo interquartil pode diferir ligeiramente de 1,33 desvio padrão; iv) seu intervalo prático não será infinito, mas geralmente estará entre 3 desvios padrões acima e abaixo da média aritmética (isto é, intervalo de aproximadamente 6 desvios padrões).

Após este estudo, chegou-se a conclusão que as distribuições mostradas nas Figuras 4.3 e 4.5, são consideradas normais ou próximas das normais, portanto, as propriedades teóricas da distribuição normal podem ser aplicadas nestas distribuições.

Com intuito de classificar/agrupar as imagens semelhantes é feito o uso da propriedade do desvio padrão dentro das propriedades da distribuição normal.

Propriedade do Desvio Padrão

Dentro das propriedades da distribuição normal é encontrado o uso do desvio padrão para determinar a porcentagem de ocorrências dentro de um intervalo. Este processo é descrito a seguir.

Através da fórmula da transformação, qualquer variável aleatória normal Y é conver-

tida em uma variável normal padronizada Z .

$$Z = \frac{Y - \bar{X}}{s} \quad (4.2)$$

onde: \bar{X} é a média aritmética dos valores e s é o desvio padrão.

O valor da variável normal padronizada (Z) é encontrado com um auxílio da tabela da distribuição normal padronizada. Esta representa as probabilidades ou áreas sob a curva normal calculadas a partir da média aritmética (s) para os determinados valores de interesse Y . Uma parte desta tabela é apresentada na Tabela 4.2. A tabela de distribuição normal padronizada completa é mostrada no Apêndice 1.

Tabela 4.2: Obtendo uma área sob a curva normal

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224

Aplicando-se a Equação 4.2, isso corresponde às probabilidades ou áreas sob a curva normal padronizada a partir da média $\bar{X} = 0$ para os valores de interesse transformados Z . Somente entradas positivas para Z são listadas na tabela, uma vez que, para uma distribuição simétrica que tenha uma média aritmética igual a 0, a área desde a média aritmética até $+Z$ (isto é, Z desvios padrões acima da média) deve ser idêntica à área desde a média aritmética até $-Z$ (isto é, Z desvios padrões abaixo da média).

O primeiro parâmetro então que precisa ser determinado é o valor de Z . Pois a média e o desvio padrão já são calculados através dos valores das amostras. Para determinar o valor de Z baseou-se na regra dos 30 (ver Seção 2.4). Esta define que, os 30 termos mais comuns de uma coleção correspondem a 30% desta coleção. Desta forma, pretende-se encontrar uma faixa que corresponda a $Z = +30\%$ da amostra.

Para aplicar a Tabela 4.2, nota-se que todos os valores Z devem primeiramente ser atualizados para duas casas decimais. Assim, o valor de interesse de Z é atualizado como $+0,3$. A fim de entender a probabilidade ou a área sob a curva a partir da média até $Z = +0,3$, percorre-se a coluna Z da Tabela 4.2 até localizar o valor de interesse Z (em decimais). Conseqüentemente, para-se na Linha $Z = +0,09$. Em seguida, lê-se ao longo dessa linha até achar-se a interseção da coluna que contém a casa de centésimos do valor Z . Portanto, no corpo da tabela, a probabilidade tabulada para $Z = +30\%$ corresponde à interseção entre a linha $Z = 0,3$ com a coluna $Z = 0,09$, conforme mostra a Tabela 4.2. Essa probabilidade é igual a 0,1517. Então, existe uma chance de 15,17% de uma imagem

Definiu-se então que estas faixas passam a ser os identificadores de indexação da coleção chamados de termo de indexação no CBIR-II. Mais detalhes na seção seguinte.

4.3 CBIR-II

[Veltkamp and Tanase, 2000] faz um comentário muito interessante em relação aos 56 sistemas analisados. Nestes sistemas, não há preocupação em relação à estrutura de indexação das imagens. Pois eles não são projetados para trabalhar em grandes coleções como a Web.

A partir deste momento esta preocupação vira foco deste trabalho. Com isto, desenvolveu-se um sistema chamado de CBIR-II.

Este foi implementado com o mesmo paradigma do CBIR-I. Primeiramente, as imagens são convertidas do espaço de cor RGB para o espaço de cor HSV. Então, o extrator de característica de cor, momentos de cor, caracteriza as imagens em termos da distribuição das cores nos *pixels* da imagem. A diferença deste sistema em relação ao CBIR-I está no processo de indexação das imagens e no cálculo da similaridade.

Para o contexto deste trabalho as seguintes analogias entre recuperação de imagens e recuperação textual foram feitas: i) imagens correspondem aos documentos; ii) **as faixas obtidas pela análise dos vetores de características da Figura 4.6 são os termos**; iii) imagens exemplos ou imagens de consulta são as consultas ou *query* na recuperação textual.

Portanto, esta seção tem como objetivo apresentar a proposta deste trabalho que é a identificação de um termo para indexação das imagens, assim técnicas da recuperação textual podem ser utilizadas em sistemas de Recuperação de Imagens Baseada no Conteúdo. Inicialmente descrevemos o índice invertido da recuperação textual adaptado para o contexto da imagem.

4.3.1 Indexação das imagens - Índice Invertido

No CBIR-I, após a extração das características, é gerado um banco de dados de característica conforme visto na Seção 3.4, onde consta o nome da imagem e nove valores entre 0 e 1 correspondentes a média, desvio padrão e obliquidade dos canais HSV, conforme a Figura 4.7.

MH	DH	IH	MS	DS	IS	MV	DV	IV
0.317627	0.267757	0.232397	0.391507	0.195728	0.119390	0.426673	0.242813	0.181251

Figura 4.7: Vetor de característica gerado pelo CBIR-I correspondente a imagem Africa1.jpg.

Já no CBIR-II, inicialmente cria-se o mesmo banco de dados de características do

CBIR-I. Posteriormente os valores da média, desvio padrão e obliquidade são mapeados para faixas de acordo com a sua ocorrência determinada pela propriedade do desvio padrão (ver Figura 4.6). A construção de um destes vetores pode ser visualizada pela Tabela 4.3. Suponha que os valores da Tabela 4.3 correspondem a primeira característica do vetor de características, que é a Média da Tonalidade (MH). De acordo com o valor do MH, será mapeado para um dos seguintes termos: MHA, MHB, MHC, MHD, MHE, MHF, MHG, MHH, MHI, MHJ. Onde M significa a média dos *pixels* do canal de cor tonalidade, H significa o canal de cor tonalidade; e A até J, é o valor correspondente à faixa onde a característica estiver presente. Este exemplo é ilustrado na Tabela 4.3.

Tabela 4.3: Construção do vetor de Características do CBIR-II

De	a	Faixa	termo
0	0.086547	A	MHA
0.086548	0.197300	B	MHB
0.197301	0.257346	C	MHC
0.257347	0.309386	D	MHD
0.309387	0.361427	E	MHE
0.361428	0.413467	F	MHF
0.413468	0.465507	G	MHG
0.465508	0.525554	H	MHH
0.525555	0.636306	I	MHI
0.636307	1	J	MHJ

O valor da primeira posição do vetor de característica do CBIR-I é 0.317627 conforme Figura 4.7. Analisando a Tabela 4.3, é possível notar que este valor corresponde a Faixa E. Sendo assim, esta característica será mapeada para o termo MHE ao invés de 0.317627.

As outras oito posições do vetor de característica são construídas de forma análoga a esta. Portanto, o CBIR-II terá um vetor de característica com 90 posições. Onde o valor igual a 1 corresponde a presença do termo correspondente ao momento/canal/faixa e valor igual a 0 a ausência do termo. Como toda imagem terá um vetor original com 9 características o vetor mapeado será formado por 9 valores iguais a 1 e 81 valores iguais a 0. Assim, o vetor de característica para a imagem *Africa1.jpg* no sistema CBIR-II é representado na Figura 4.8.

Este vetor de característica foi mapeado para uma representação esparsa. Sendo assim, somente os termos que estiverem presentes na imagem são inseridos no vetor de característica do CBIR-II. Então, uma representação esparsa do vetor de características da imagem *Africa1.jpg* pode ser visualizado na Figura 4.9.

Portanto, os 90 termos criados pelo mapeamento de faixas são utilizados para indexação das imagens através da estrutura conhecida como índice invertido.

Como visto, na Seção 2.3, o índice invertido possui duas partes principais: uma estrutura de busca, chamada de vocabulário, contendo todos os termos distintos existentes nas imagens indexadas e, para cada termo, uma lista invertida que armazena os identificadores dos registros contendo o termo, a saber, imagens onde o termo ocorre, como ilustrado na

0	0	0	0	1	0	0	0	0	0
MHA	MHB	MHC	MHD	MHE	MHF	MHG	MHH	MHI	MHJ
0	0	0	0	0	1	0	0	0	0
DHA	DHB	DHC	DHD	DHE	DHF	DHG	DHH	DHI	DHJ
0	0	0	0	0	0	0	1	0	0
IHA	IHB	IHC	IHD	IHE	IHF	IHG	IHH	IHI	IHJ
0	0	0	0	0	1	0	0	0	0
MSA	MSB	MSC	MSD	MSE	MSF	MSG	MSH	MSI	MSJ
0	0	0	0	1	0	0	0	0	0
DSA	DSB	DSC	DSD	DSE	DSF	DSG	DSH	DSI	DSJ
0	0	0	0	0	1	0	0	0	0
ISA	ISB	ISC	ISD	ISE	ISF	ISG	ISH	ISI	ISJ
0	0	0	0	1	0	0	0	0	0
MVA	MVB	MVC	MVD	MVE	MVF	MVG	MVH	MVI	MVJ
0	0	0	1	0	0	0	0	0	0
DVA	DVB	DVC	DVD	DVE	DVF	DVG	DVH	DVI	DVJ
0	0	0	0	0	1	0	0	0	0
IVA	IVB	IVC	IVD	IVE	IVF	IVG	IVH	IVI	IVJ

Figura 4.8: Vetor de característica gerado pelo CBIR-II correspondente a imagem Africa1.jpg.

MHE	DHF	IHH	MSF	DSE	ISF	MVE	DVD	IVF
-----	-----	-----	-----	-----	-----	-----	-----	-----

Figura 4.9: Representação esparsa do vetor de característica do CBIR-II correspondente a imagem Africa1.jpg.

Figura 4.10.

Consultas nos arquivos invertidos são feitas tomando-se a lista invertida correspondente aos termos procurados. Por exemplo, caso a imagem de consulta tivesse os seguintes termos: MHA, DSC e IVJ. Apenas as listas invertidas destes três termos seriam percorridas.

Em relação ao peso dos termos, observe na Figura 4.10 que no vetor mapeado, são binários. Com o intuito de melhorar a eficácia do processo de recuperação, pesos não binários podem ser atribuídos a estes termos. A ideia é associar a cada termo do vetor de característica um peso que indique o grau de importância do termo para definir o conteúdo da imagem. Portanto, cada imagem possui um vetor associado que é constituído por elementos formados pelo termo e o seu peso. Na abordagem anterior (Figura 4.10) esses pesos eram binários, indicando apenas presença/ausência.

Pode-se usar diversas formas de se calcular o peso de um termo na imagem. Em analogia com a RI textual, considera-se aqui, cálculos fundamentados no número de ocorrências do termo na imagem e na coleção. Na imagem, é considerado o número de ocorrências do termo como binário, ou seja, se o termo estiver presente na imagem ele terá ocorrência igual a 1; caso contrário a ocorrência é 0. A frequência na coleção é dada pelo IDF

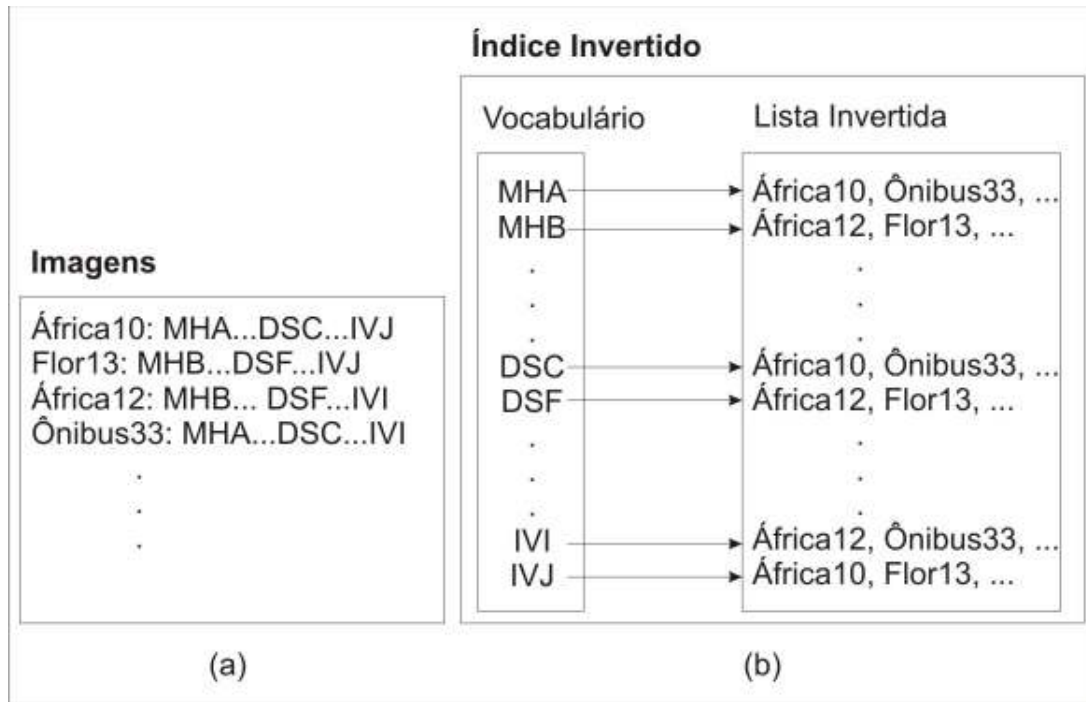


Figura 4.10: (a) Banco de dados de Característica (b) Estrutura de busca (vocabulário) e a lista invertida.

apresentado na Seção 2.5, que é calculado da seguinte forma:

$$idf_t = \log \frac{N}{df_t} \quad (4.3)$$

onde N é o número total de imagens da coleção e df_t é número de imagens da coleção em que o termo ocorre. Assim, o IDF de um termo raro é elevado, enquanto que o IDF de termos comuns decresce em escala logarítmica. Assim o peso será o produto de um indicador de presença/ausência, aqui chamado de tf pelo idf :

$$W_{it} = tf \times idf \quad (4.4)$$

onde $tf \in \{0,1\}$ e idf é dada pela equação 4.3.

Desta forma, os termos que estiverem na extremidade da distribuição de características da Figura 4.6, terão um idf elevado e os termos que se encontrarem nas faixas centrais terão um idf inferior.

4.3.2 Cálculo da Similaridade

A utilização de uma mesma representação para imagens e consultas permite o cálculo da similaridade entre uma consulta q e uma imagem i . Este cálculo é realizado através da correlação entre os vetores que os representam, quantificada pelo cosseno do ângulo formado por \vec{i} e \vec{q} . Esta métrica é conhecida como medida de similaridade do cosseno.

Desta forma, em um espaço vetorial de dimensão n , a similaridade (sim) entre dois vetores \vec{q} e \vec{i} é calculada através do cosseno do ângulo formado por estes vetores, através da seguinte fórmula:

$$\text{sim}(\vec{q}, \vec{i}) = \frac{\sum_{t=1}^n W_{qt} W_{it}}{\sqrt{\sum_{t=1}^n W_{qt}^2} \sqrt{\sum_{t=1}^n W_{it}^2}} \quad (4.5)$$

onde \vec{q} é o vetor de termos da consulta; \vec{i} é o vetor de termos da imagem; W_{qt} é o peso do termo t da consulta q e W_{it} é o peso do termo t na imagem i . Os valores de W_{qt} e W_{it} são dados pela equação 4.5. Note que os termos não presentes na consulta receberão peso igual a 0.

Os valores da similaridade entre uma expressão de busca e cada um dos documentos da coleção são utilizados na ordenação dos documentos resultantes. Portanto, no modelo vetorial os resultados são ordenados de acordo com a medida de similaridade do cosseno formando o *ranking*.

A vantagem deste processo e conseqüentemente o ganho de tempo no processo de recuperação (mais detalhes no Capítulo 5) está no fato de que apenas as imagens que contiverem no mínimo um dos termos de busca da consulta são recuperadas. Em relação ao cálculo do cosseno, os termos que não estiverem presentes receberão peso zero e, portanto, seus produtos não serão calculados.

4.4 Considerações Finais Sobre a Proposta

Neste capítulo foi visto que, muitos sistemas de recuperação de imagens baseado no conteúdo utilizam o espaço de cor HSV juntamente com o extrator de cor momentos de cor. Para calcular a similaridade entre a imagem de consulta com as imagens da coleção é utilizada a distância Euclidiana. Estes sistemas apresentam boas medidas de avaliação na qualidade de recuperação, mas apresentam um baixo desempenho na velocidade de recuperação para grandes coleções como a Web. Outros sistemas CBIR foram citados. Estes sistemas que utilizam o cálculo do cosseno em suas implementações, mas não fazem uso da estrutura do índice invertido.

Sistemas de recuperação de imagens na Web possuem desempenho melhor em relação a velocidade de recuperação, mas apresentam baixa qualidade na medida de avaliação destes sistemas, devido ao fato de que, na maioria das vezes, o texto próximo a uma imagem não descreve-lá fielmente.

Foram descritos os dois sistemas desenvolvidos neste trabalho. Aqui chamados de CBIR-I e CBIR-II. O primeiro foi implementado com as características tradicionais de sistemas CBIR, com intuito de ser utilizado como base de comparação tanto na qualidade quanto na velocidade do processo de recuperação com o CBIR-II. Para isto, foram

necessários estudos sobre o vetor de característica do CBIR-I, onde ficou constatado que a distribuição destes vetores para 2 coleções de imagens eram similares a distribuição normal da probabilidade e estatística. Assim, propriedades desta distribuição foram utilizadas para classificar grupos de valores, com o intuito de se criar termos de indexação para serem utilizados no CBIR-II. Este sistema foi implementado utilizando técnicas tradicionais da recuperação de informação textual, tais como, índice invertido, peso do termo e modelo vetorial, com o intuito de se obter um ganho de tempo no processo de recuperação sem grandes perdas na qualidade de recuperação.

O próximo capítulo esboça os resultados encontrados na comparação da qualidade de recuperação destes dois sistemas juntamente com o número de operações que cada sistema realiza no cálculo da similaridade quando uma consulta é solicitada pelo o usuário de um SRI.

Capítulo 5

Experimentos e Resultados

Este capítulo é dedicado às avaliações experimentais do sistema CBIR-II. Os experimentos aqui realizados têm dois objetivos principais: i) comparar os resultados do CBIR-II com aqueles obtidos pelo CBIR-I, e verificar a qualidade da recuperação; ii) comparar os dois sistemas em relação ao número de operações aritméticas efetuadas para o cálculo da similaridade, entre a imagem de consulta e a base de dados.

Para comparar este dois sistemas primeiramente foi utilizada uma base de treinamento. A coleção escolhida para esta etapa foi a Corel-1000. Posteriormente foi realizado outros experimentos em uma coleção de teste, BD-10000. Em ambos os experimentos a medida de avaliação da qualidade do *Ranking* foi a Precisão. A seguir descrevem-se os resultados da base de treinamento.

5.1 Avaliação Experimental no Corel-1000 - Coleção de Treinamento

Para a realização deste primeiro experimento foi utilizada uma base de dados de treinamento, que neste caso, foi a coleção Corel-1000. Esta escolha se deu pelo fato das imagens desta coleção serem classificadas. Assim, esta base contém 10 classes cada uma com 100 imagens, a saber: imagens da África, praia, construções, dinossauros, ônibus, cavalos, flores, comidas, elefantes e montanhas. Para a realização dos testes foram escolhidas de forma aleatória 80 imagens de consulta, sendo 8 imagens para cada classe.

São comparados três sistemas de recuperação de imagens. O primeiro é o CBIR-I, que é um sistema tradicional de CBIR. O segundo sistema, CBIR-II, foi dividido em dois experimentos. O primeiro consiste em trabalhar com peso binário, este será chamado de CBIR-IIa. A abordagem que utiliza o peso binário considera apenas o número de ocorrências do termo na imagem como binário. Se o termo estiver presente na imagem ele terá peso igual a 1; caso contrário o peso é 0. A similaridade neste sistema é calculada pela interseção entre os termos presentes na consulta e os termos presentes nas imagens

da coleção. As imagens que obtiverem os maiores valores da interseção estarão no topo do *ranking*. Caso haja empate, o critério de desempate é feito por ordem em que o algoritmo localizou as imagens. A imagem que for localizada primeiramente fica na frente. Por fim, o último experimento utiliza a abordagem do peso baseado em *idf* descrito na Seção 2.5. Este sistema foi chamado de CBIR-IIb.

Para avaliar a qualidade de recuperação, entre os dois sistemas implementados, a medida de desempenho precisão foi utilizada. Dada uma consulta, a precisão é definida como a fração entre o número de imagens relevantes recuperadas, sobre o número total de imagens recuperadas:

$$Pr(q) = \frac{RR_q}{Rec_q} \quad (5.1)$$

onde: RR_q é o número de imagens relevantes recuperadas e Rec_q o total de imagens recuperados.

Para efetuar a medida da precisão baseou-se em uma estatística encontrada em [Manning et al., 2007]. Foi constatado que os usuários de SRI, navegam na grande maioria das vezes até a terceira página de resultados. Portanto, é de fundamental importância, que os resultados da requisição do usuário mais relevantes estejam nas primeiras posições do *ranking*. Assim, calculou-se a Precisão em cinco posições (P0, P5, P10, P20 e P100). P0 significa calcular a precisão na posição do *ranking* onde encontra a primeira imagem relevante, com exceção da imagem de consulta. P5 é o cálculo da precisão na quinta posição do *ranking*, analogamente P10, P20 e P100 são a décima, vigésima e centésima respectivamente.

Uma imagem da coleção é considerada relevante à imagem de consulta se ambas pertencerem a mesma classe. Esta definição de relevância por classe, foi adotada, a fim de evitar um dos maiores problemas, na análise de resultados em CBIR, a subjetividade humana. Considera-se, portanto que as imagens no Corel-1000 foram corretamente classificadas.

A Tabela 5.1 mostra os valores encontrados das precisões P0, P5, P10, P20 e P100 do sistema CBIR-I, relativo à média de 8 consultas de cada classe, juntamente com o valor médio da precisão de 80 consultas da coleção. Já na Tabela 5.2 e Tabela 5.3 encontram-se estes mesmos resultados para o sistema CBIR-IIa e CBIR-IIb.

De acordo com os resultados, pode-se observar que o CBIR-IIa e o CBIR-IIb apresentam um ganho na qualidade de recuperação no ponto P0. A medida que se avança no *ranking* há uma pequena perda na qualidade de recuperação, mas esta perda permanece estável. Outra comparação importante, é em relação ao CBIR-IIa e o CBIR-IIb, ambos obtiveram resultados superiores ao CBIR-I no ponto P0, mas o CBIR-IIb, apresentou resultados superiores em todos os pontos em relação ao CBIR-IIa. Isto mostra que a abordagem utilizando peso baseado em *idf* favorece o melhor desempenho na qualidade de recuperação, principalmente nas regiões de alta precisão.

Tabela 5.1: Precisão média no sistema CBIR-I. Coleção Corel-1000.

Classe	P0	P5	P10	P20	P100
África	86,6%	84%	72%	67%	47%
Praia	46,5%	48%	46%	38%	28,8%
Construções	62%	64%	56%	46%	26,5%
Dinossauros	100%	100%	98%	97%	78,2%
Ônibus	100%	88%	82%	70%	51,2%
Cavalos	100%	100%	92%	82%	47,6%
Flores	90%	76%	52%	41%	24%
Comidas	86,6%	80%	74%	69%	36%
Elefantes	90%	88%	80%	65%	44,8%
Montanhas	36,8%	48%	46%	43%	34,2%
Média Geral	79,8%	77,6%	69,8%	61,8%	41,8%

Tabela 5.2: Precisão média no sistema CBIR-IIa. Coleção Corel-1000.

Classe	P0	P5	P10	P20	P100
África	71,6%	76%	68%	52%	35,4%
Praia	62,5%	52%	36%	26%	19,8%
Construções	56,6%	52%	46%	37%	19,5%
Dinossauros	100%	96%	94%	91%	61,4%
Ônibus	100%	96%	78%	62%	40,1%
Cavalos	100%	76%	74%	60%	34,4%
Flores	100%	76%	48%	42%	23,8%
Comidas	73,2%	64%	60%	50%	28,2%
Elefantes	100%	76%	66%	51%	28,8%
Montanhas	49,2%	28%	30%	22%	18,6%
Média Geral	81,3%	69,2%	60%	49,3%	31%

Tabela 5.3: Precisão média no sistema CBIR-IIb. Coleção Corel-1000.

Classe	P0	P5	P10	P20	P100
África	71,6%	72%	66%	58%	36,4%
Praia	51,8%	44%	34%	32%	19,6%
Construções	70,6%	48%	50%	39%	22,2%
Dinossauros	100%	100%	96%	95%	65%
Ônibus	100%	96%	84%	66%	39,6%
Cavalos	100%	84%	72%	58%	34,6%
Flores	100%	76%	62%	53%	28,8%
Comidas	86,6%	64%	60%	56%	31%
Elefantes	100%	76%	64%	57%	31,8%
Montanhas	76,6%	40%	38%	34%	20,8%
Média Geral	85,7%	70%	62,6%	54,8%	33%

É bom lembrar que o objetivo do sistema proposto não é ter resultados melhores no processo de recuperação e sim manter uma qualidade similar ao CBIR-I, pois a principal vantagem do CBIR-II está relacionada no ganho de tempo no processo de recuperação, que será mostrado em breve.

Uma abordagem que compara a distância Euclidiana com a distância do cosseno foi abordada em [Qian et al., 2002]. Provou-se que ambas as distâncias trazem resultados bem similares. A ordem das imagens recuperadas é que pode sofrer alterações devido à variância da amostra. Mas em [Qian et al., 2002] não foi utilizada a metodologia

de indexação das imagens proposta aqui, com o uso do índice invertido por meio do mapeamento em faixas. Neste sentido, avaliamos aqui a proposta de construção do índice invertido, considerando que a comparação da distância Euclidiana com o cosseno como medida de similaridade já foi estudada.

Alguns exemplos de consultas são mostrados para a coleção Corel-1000 realizada nos três sistemas (CBIR-I, CBIR-IIa, CBIR-IIb). Para todos os exemplos a primeira imagem do *ranking* corresponde à imagem de consulta.

A classe *dinossauro* foi a que apresentou melhor porcentagem na qualidade de recuperação. Pelo fato de todas as imagens presentes nesta classe terem a cor branca como predominante. Esta predominância contribui para o melhor resultado desta classe. Observe a Figura 5.1 como exemplo. Pelo fato do sistema CBIR-II não possuir um critério de desempate na interseção dos termos presentes na consulta e nas imagens, ele obteve um desempenho inferior aos demais na classe *dinossauro*. Um exemplo é a imagem da 19ª posição do *ranking*. As imagens presentes da 19ª a 25ª posição do *ranking* são da classe *dinossauro*. Elas apresentam o mesmo número de termos em comum com a imagem de consulta. A diferença na posição do *ranking* ocorre pelo fato do algoritmo ter localizado primeiramente a imagem que se encontra na 19ª posição. Ou seja, esse *ranking* é uma ordem parcial.

A classe *ônibus* obteve bons resultados na qualidade de recuperação. Um exemplo de consulta na categoria *ônibus* é mostrada na Figura 5.2.

As classes *praia* e *montanha* foram as que proporcionaram os piores resultados. Isto se deve ao fato das imagens destas classes apresentarem imagens bem similares em relação a característica cor. As Figuras 5.3 e 5.4 demonstram este fato.

O próximo passo deste experimento é aplicá-lo em uma coleção de teste. A próxima seção apresenta os resultados destes experimentos.

5.2 Avaliação Experimental no BD-10000 - Coleção de teste

Com intuito de verificar se a proposta apresentada é consistente, decidiu-se realizar novos experimentos, só que agora em uma coleção de testes. A coleção escolhida foi a base de dados BD-10000.

A coleção BD-10000 é uma coleção de imagens reais cobrindo uma ampla variedade de categorias semânticas. Dentre estas categorias três foram classificadas: avião, carro e moto. O mesmo método aplicado na Seção 5.1 foi utilizado aqui. Oito imagens de consulta de cada classe foram escolhidas aleatoriamente. A medida de avaliação Precisão foi aplicada da mesma forma também. Assim, os resultados para estas três classes são mostrados nas Tabelas 5.4, 5.5 e 5.6, que correspondem aos resultados do CBIR-I, CBIR-

IIa e CBIR-IIb, respectivamente.

Tabela 5.4: Precisão média no sistema CBIR-I. Coleção BD-10000.

Classe	P0	P5	P10	P20	P100
Avião	79,1%	77,5%	71,3%	68,8%	42,5%
Carro	79%	62,5%	58,8%	56,3%	46,3%
Moto	82,4%	72,5%	57,5%	46,9%	30,8%
Média Geral	80,1%	70,8%	62,5%	57,3%	39,9%

Tabela 5.5: Precisão média no sistema CBIR-IIa. Coleção BD-10000.

Classe	P0	P5	P10	P20	P100
Avião	93,8%	75%	73,8%	62,5%	41,6%
Carro	75%	62,5%	55%	49,4%	38,4%
Moto	55,9%	60%	41,3%	38,8%	29,5%
Média Geral	74,9%	65,8%	56,7%	50,2%	36,5%

Tabela 5.6: Precisão média no sistema CBIR-IIb. Coleção BD-10000.

Classe	P0	P5	P10	P20	P100
Avião	93,8%	77,5%	68,8%	60,6%	39,8%
Carro	85,4%	62,5%	61,3%	55,6%	40%
Moto	71,9%	65%	48,8%	36,9%	29%
Média Geral	83,7%	68,3%	59,6%	51%	36,3%

Como pode ser observado, as qualidades de recuperação continuam bem similares aos resultados encontrados na Seção 5.1, só que com algumas mudanças.

O CBIR-IIb continuou a apresentar o melhor resultado no ponto P0. Só que agora o CBIR-IIa não conseguiu manter o desempenho do anterior e perde em todos os pontos para o CBIR-I.

A diferença da porcentagem nos pontos P5, P10 e P100 entre o CBIR-I e o CBIR-II-b caiu, deixando-os bem mais próximos. A diferença no ponto P20 permaneceu praticamente a mesma.

Pode-se concluir então que a abordagem utilizando peso baseado em *idf* contribui para o melhoramento da qualidade de recuperação comparada com a abordagem que utiliza peso binário. O CBIR-IIb apresentou em todos testes uma qualidade de recuperação superior no ponto P0 e na medida que se avança no *ranking* a qualidade de recuperação do CBIR-IIb decresce comparada com a medida do CBIR-I.

A categoria *avião* apresenta suas imagens com predominância da cor azul (avião voando). Quando utilizadas uma destas imagens como consulta os resultados são satisfatórios. Mas a predominância da cor azul não é encontrada em todas as imagens. Existem algumas imagens em que o avião esta no chão. Nestas, a qualidade da recuperação foi inferior. Um exemplo da classe *avião* com e sem predominância da cor azul pode ser encontrado na Figura 5.5. e na Figura 5.6, respectivamente.

Esta pequena perda na qualidade de recuperação pode ser compensada pelo fato do CBIR-IIb apresentar um desempenho melhor no tempo de processamento que o CBIR-I. A próxima seção, apresenta esta comparação.

5.3 Avaliação Experimental do número de operações Aritméticas

Para estudar o ganho de tempo quanto processamento, em [Matos et al., 2008] foi feita uma comparação dos dois sistemas em relação ao número de operações aritméticas efetuadas para o cálculo da similaridade, entre a imagem de consulta e a base de dados.

No CBIR-I, de acordo com a equação 4.1 e o vetor de característica da Tabela 4.1, para cada consulta são feitas as seguintes operações: i) 9 subtrações; ii) 9 multiplicações; iii) 8 adições; iv) 1 raiz quadrada. Como mencionado anteriormente o CBIR-I faz a comparação do vetor de consulta com todas as imagens da coleção, com o objetivo de se obter o *ranking* de similaridade final. Tendo a base de dados 1000 imagens ao final de uma consulta totaliza-se 27.000 operações aritméticas.

Entretanto, no CBIR-II, por meio da lista invertida não é necessário varrer toda a coleção, mas somente aquelas que casarem com pelo menos um termo da consulta. Para cada característica (termo) presente faremos as seguintes operações: i) $z * 9$ somas, onde z é o tamanho médio das listas invertidas; ii) y divisões, onde y é o número total de imagens encontradas nas listas. Nos experimentos z e y tiveram valores médios de 1038 adições e 677 divisões respectivamente. Totalizando 1715 operações aritméticas por consulta. Valor bem abaixo dos 27.000 encontrados no CBIR-I.

Analogamente na coleção BD-10000 o número de operações aritméticas por consulta foi de 25.500, valor bem abaixo do encontrado no CBIR-I, que é igual a 270.000.

O denominador da equação 2.2 não é necessário calcular no momento da consulta, pois ele é calculado no momento da indexação das imagens. Isto indica um ganho de tempo no processamento da consulta utilizando o índice invertido e o cálculo do cosseno no processo de recuperação, similarmente em grandes coleções de imagens.

Uma questão não abordada anteriormente refere-se à escolha das 10 faixas na distribuição das amostras dos vetores de características. Foram feitos vários experimentos, usando 6, 8, 10 e 24 faixas. A divisão que trouxe melhor resultado foi a com 10 faixas. Então essa foi a utilizada na coleção de teste.

5.4 Considerações Finais

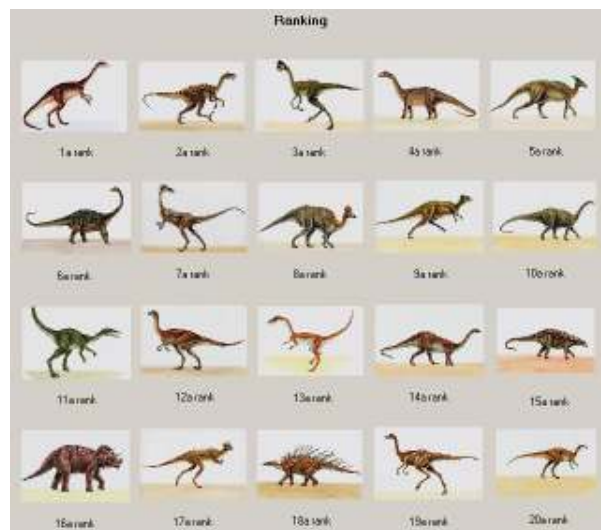
Neste capítulo verificou-se que o método de recuperação de imagens proposto neste trabalho produziu bons resultados, que quando comparados com os resultados de um

método tradicional de recuperação de imagens.

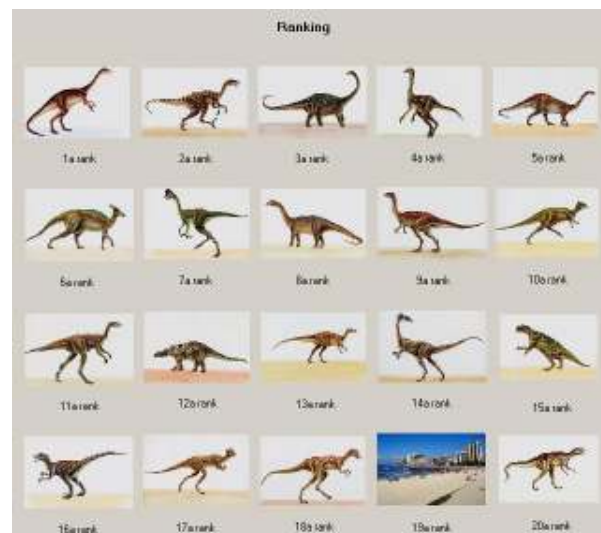
Foram mostrados duas avaliações experimentais. Na primeira foi utilizada a coleção Corel-1000 como treinamento. Nesta avaliação, pode-se notar que a qualidade de recuperação do sistema proposto no topo do *ranking* é superior ao sistema CBIR tradicional. Já na segunda avaliação, foi utilizada uma coleção de testes BD-10000. Os resultados nesta coleção foram similares ao da coleção de treinamento. Na primeira posição do *ranking* o sistema proposto apresenta melhor qualidade de recuperação e a medida que se avança no *ranking* o sistema proposto perde um pouco na qualidade.

Para compensar esta perda, foi mostrado que a abordagem proposta é bem mais rápida no cálculo da similaridade da imagem de consulta com a coleção de imagens pelo fato de realizar menos operações aritméticas.

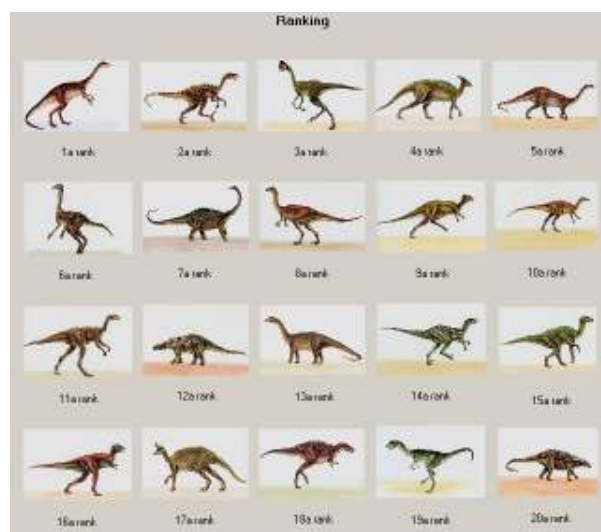
O próximo capítulo apresenta as considerações finais desta dissertação, além das perspectivas futuras deste trabalho.



(a)



(b)



(c)

Figura 5.1: Resultados de busca obtido para a categoria Dinossauro, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.

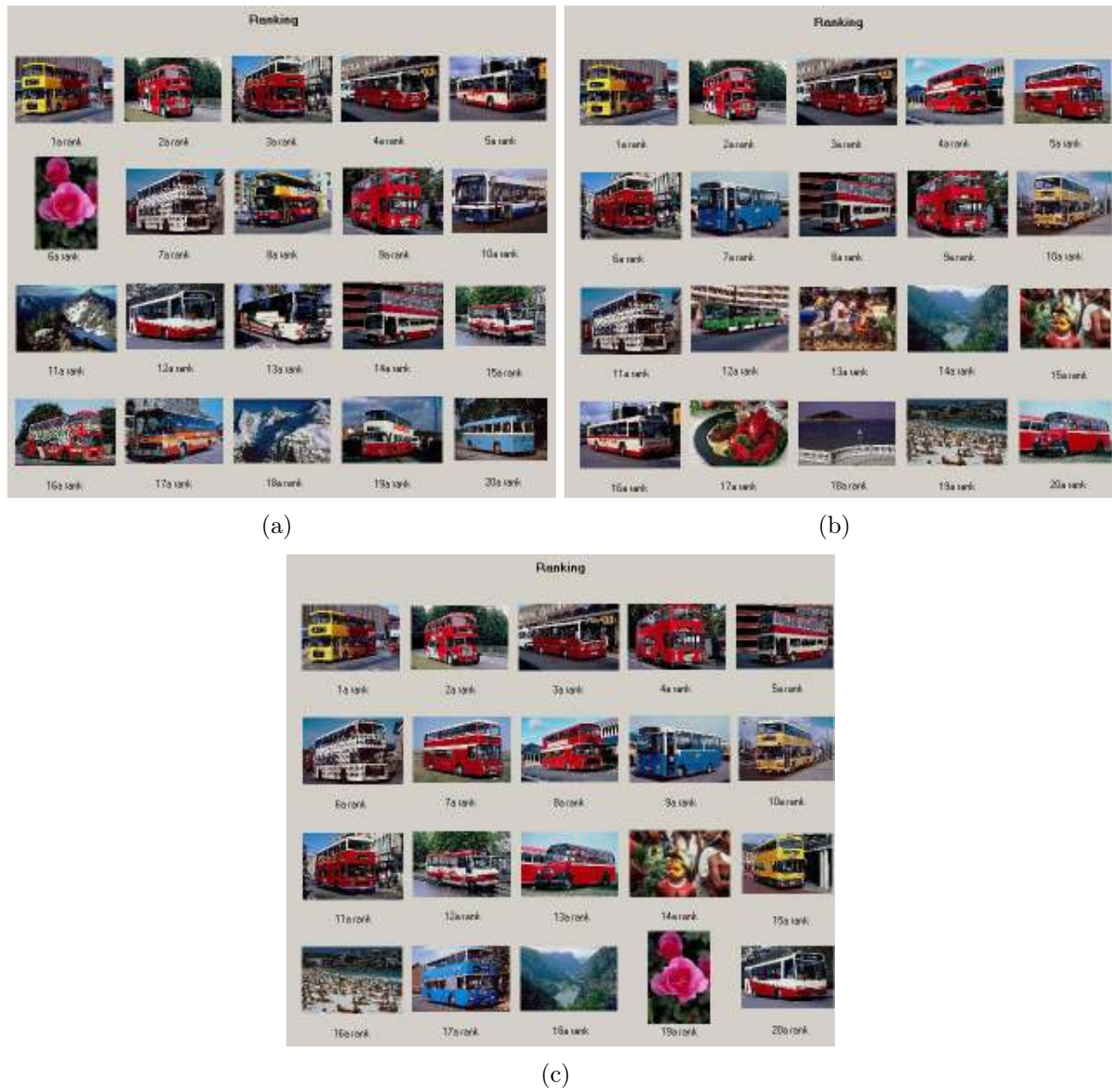
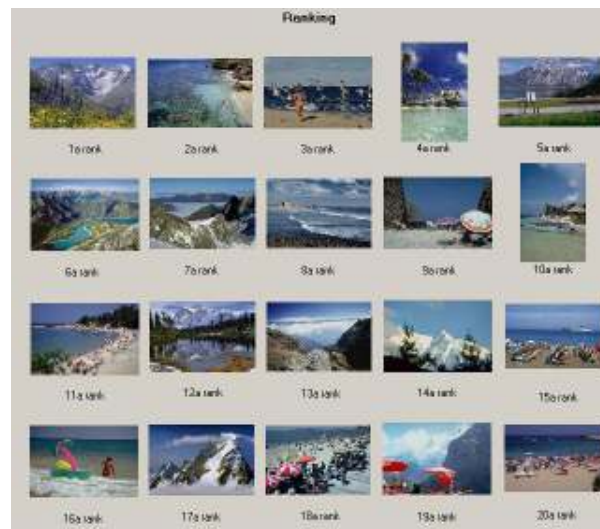


Figura 5.2: Resultados de busca obtido para a categoria ônibus, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.



Figura 5.3: Resultados de busca obtido para a categoria praia, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.



(a)



(b)



(c)

Figura 5.4: Resultados de busca obtido para a categoria montanha, no banco de dados Corel-1000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.



(a)



(b)



(c)

Figura 5.5: Resultados de busca obtido para a categoria avião, no banco de dados BD-10000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.



Figura 5.6: Resultados de busca obtido para a categoria avião, no banco de dados BD-10000. (a) No Sistema CBIR-I (b) No Sistema CBIR-IIa (c) No Sistema CBIR-IIb.

Capítulo 6

Conclusão e Perspectivas Futuras

Devido ao aumento crescente do volume de informação na Internet, é necessária a criação de mecanismos capazes de prover, de maneira rápida e eficaz, a informação requisitada pelo usuário. Foi neste aspecto que esta dissertação focou o seu estudo, apresentando uma abordagem que faz uso das técnicas tradicionais de Recuperação de Informação (que além de serem totalmente automáticas, aumentam a velocidade de processamento da consulta) com a eficácia na qualidade de recuperação dos sistemas de recuperação de imagens baseado no conteúdo.

Com o intuito de estabelecer um modelo apropriado para esta tarefa, foi necessário o estudo de arquiteturas de modelos de recuperação de informação textual, como por exemplo, índice invertido, peso do termo e modelo vetorial. Por meio de exemplos, as vantagens de cada técnica foi apresentada.

Na área de recuperação de imagens foram necessários estudos em relação aos sistemas de recuperação de imagens. Foram identificadas três categorias que são: recuperação de imagens baseada em anotações textuais, recuperação de imagens providas pelos *sites* Google e Yahoo. Neste tipo de recuperação as imagens são indexadas com base nas informações textuais ou rótulos que as acompanham. Finalmente, existe a recuperação de imagens por conteúdo. Foi mostrado que as duas primeiras técnicas de recuperação de imagens apresentam uma boa qualidade em relação à velocidade no processo de consulta, após a definição de uma consulta de um usuário com palavras-chaves. A desvantagem dessa técnica é que apresentam uma baixa qualidade na recuperação proporcionada pelo fato de que anotações manuais ou rótulos que acompanham as imagens não as descrevem fielmente. Já os sistemas de recuperação de imagens baseadas no conteúdo apresentam uma melhora na qualidade de recuperação, mas apresentam um problema em relação à velocidade no processo da consulta para grandes coleções.

Assim, o objetivo deste trabalho foi propor um método para separar valores de característica de baixo nível em grupos, priorizando a característica cor, que possibilitasse o mapeamento para um identificador de indexação da coleção.

Este mapeamento foi feito a partir do estudo da distribuição dos valores dos vetores

de características. Ficou constatado que a distribuição desses valores eram similares à distribuição normal da probabilidade e estatística. Pela propriedade existente nesta distribuição, conhecida como a propriedade da distribuição normal, foi possível determinar a porcentagem de ocorrências dentro de um intervalo. Sendo assim, estes intervalos de imagens similares passou a ser o identificador de indexação da coleção.

Diante disso, foi possível utilizar técnicas existentes em RI textual, a fim de acelerar o processo de consulta: índice invertido e similaridade baseada no cosseno.

Assim comparou-se o desempenho entre dois sistemas. O primeiro utiliza a medida da distância euclidiana para o cálculo da similaridade enquanto o outro usa a medida do cosseno, além de indexar as imagens por meio do índice invertido. Mostrou-se que o segundo sistema tem um ganho significativo no tempo de processamento por realizar menos operações aritméticas, pois não há necessidade de se varrer toda a base de dados. Na avaliação experimental do resultado de recuperação o sistema manteve a qualidade nas regiões de alta precisão, mas com uma pequena queda a medida que se avança no *ranking*.

Trabalhos Futuros

Como trabalhos futuros destacam-se: melhorar a qualidade do resultado por meio de novos algoritmos para definir as faixas dos vetores de características; e explorar outras características de baixo nível, como textura e forma. Embora não registrada nesta dissertação durante os estudos constatou-se que a distribuição dos vetores de característica de textura (utilizando a técnica vizinhança de textura) e da forma (com o uso do histograma de direção de bordas) também apresentam distribuições semelhantes a distribuição normal da probabilidade e estatística nas bases Corel-1000 e BD-10000. Diante disso, como trabalho futuro, essas características podem ser utilizadas, de forma análoga ao trabalho aqui apresentado. A extração dos vetores de características neste trabalho utilizou imagem inteira (global), caso seja utilizada a forma local para extrair o vetor de característica é possível trabalhar com a frequência do termo na imagem. Outro trabalho que pode melhorar o desempenho da recuperação é aplicar o conceito de *stop words* nos termos da imagem, com a exclusão dos termos comuns do índice invertido, verificando se não há perda na qualidade de recuperação.

Referências Bibliográficas

- [Antani et al., 2002] Antani, S., Kasturi, R., and Jain, R. (2002). A survey on the use of parttern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern Recognition*, (35):945–965.
- [Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (1999). *Modern Information Retrieval*. Addison-Wesley. Essex, UK.
- [Balinski, 2002] Balinski, R. (2002). Filtragem de informações no ambiente do direto. Master’s thesis, Universidade Federal do Rio Grande do Sul.
- [Bender, 2003] Bender, T. C. (2003). Classificação e recuperação de imagens por cor utilizando técnicas de inteligencia artificial. dissertação de mestrado. Master’s thesis, Universidade do Vale do Rio dos Sinos - UNISINOS.
- [Brandt, 1999] Brandt, S. (1999). Use of shape features in content-based image retrieval. *Master thesis, Helsinki University of Technology, Department of Engineering Physics e Mathematics, Epoo, Finland*.
- [CalPhotos,] CalPhotos. University of california, berkeley. available: <http://calphotos.berkeley.edu/>.
- [Carson et al., 1999] Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., and Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. *In Third International Conference on Visual Information Systems*.
- [Castañón, 2003] Castañón, C. A. B. (2003). Recuperação de imagens por conteúdo através de análise multiresolução por wavelets. Master’s thesis, USP-São Carlo. SP.
- [Chang and Hsu, 1992] Chang, S. K. and Hsu, A. (1992). Image information systems: where do we go from here. *IEEE Trans. on Knowledge and Data Engineering* 5, pages 431–442.
- [Choras et al., 2007] Choras, R. S., Andrysiak, T., and Choras, M. (2007). Integrated color, texture and shape information for content based image retrieval. *Pattern Anal Applic*.
- [Crucianu et al., 2004] Crucianu, M., Ferecatu, M., and Boujemaa, N. (2004). Relevance feedback for image retrieval: a short survey. *Le Chesnay Cedex. France, Italy*.
- [Datta et al., 2008] Datta, R., Dhiraj Joshi, J. L., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Transactions on Computing Surveys*, 40(2).

- [Ebert, 1994] Ebert, D. S. (1994). Texturing and modeling: A procedural approach. *Academic Press, Cambridge, MA*.
- [El-Kwae and Kabuka, 2000] El-Kwae, E. and Kabuka, M. R. (2000). Efficient content-based indexing of large image databases. *ACM Transactions on Information Systems*, 18(2):171–210.
- [Elmasri and Navathe, 2000] Elmasri, R. and Navathe, S. B. (2000). *Fundamentals of Database Systems*. Addison Wesley Longman, Inc., 3 edition.
- [Ferneda, 2003] Ferneda, E. (2003). Recuperação de informação: Análise sobre a contribuição da ciência da computação para a ciência da informação. Master's thesis, Escola de Comunicação e Artes da Universidade de São Paulo.
- [Frakes and Baeza-Yates, 1992] Frakes, W. B. and Baeza-Yates, R., editors (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall.
- [Gomes and Velho, 1994] Gomes, J. and Velho, L. (1994). *Computação Gráfica: Imagem*. IMPA/SBM. Rio de Janeiro-RJ. Brasil.
- [Gonzales and Woods, 2002] Gonzales, R. C. and Woods, R. E. (2002). *Digital Image Processing*, volume Second Edition. Pretenci Hall.
- [Gonzales et al., 2004] Gonzales, R. C., Woods, R. E., and Eddins, S. L. (2004). *Digital Image Processing Using MATLAB*. Pearson Education.
- [Haralick, 1973] Haralick, R. M. (1973). Texture features for image classification. *In IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6).
- [Huang et al., 1997] Huang, J., Kumar, S. R., Mitra, M., Zhu, W., and Zabih, R. (1997). Image indexing using color correlogram. *In IEEE International Conference on Computer Vision and Evolutionary Computation and Pattern Recognition, Puerto Rico*.
- [Huijsmans and Smeulders, 1999] Huijsmans, D. P. and Smeulders, A. W. M. (1999). Visual information and information systems. *Proceedings of the Third International Conference VISUAL, Amsterdam, The Netherlands*.
- [Iqbal and Aggarwal, 2002] Iqbal, Q. and Aggarwal, J. K. (2002). Cires: A system for content-based retrieval in digital image libraries. *In Seventh International Conference on Control, Automation, Robotics and Vision*, pages 205–210.
- [Joshi et al., 2006] Joshi, D., Datta, R., Zhuang, Z., Weiss, W., Friedenber, M., Li, J., and Wang, J. Z. (2006). Paragrab: A comprehensive architecture for web image management and multimodal querying. *In Proc. of Very large data bases, 32 (Seoul, Korea, 2006)*, 1163, page 1166.
- [Koskela, 1999] Koskela, M. (1999). *Content-based Image Retrieval with Self-Organizing Maps*. PhD thesis, Helsinki University of Technology.
- [Laaksonen et al., 2000] Laaksonen, J., Koskela, M., Laakso, S., and Oja, E. (2000). Pic-som: content-based image retrieval with self-organizing maps. *Pattern Recognition Letters* 21:1199-1207.

- [Levine et al., 2000] Levine, D. M., Berenson, M. L., and Stephan, D. (2000). *Estatística: Teoria e Aplicações: Usando Microsoft Excel. tr. It. de Teresa Cristina Padilha de Souza*. LTC, Rio de Janeiro.
- [L.Kherfi et al., 2004] L.Kherfi, M., Ziou, D., and Bernardi, A. (2004). Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Comput. Surv.*
- [Long and Leow, 2000] Long, H. Z. and Leow, W. K. (2000). Perceptual texture space for content-based image retrieval. *World Scientific*.
- [Louupias and Bres, 2001] Louupias, E. and Bres, S. (2001). Key point-based indexing for pre-attentive similarities: The kiwi system. *Pattern Analysis and Applications*, 4(2/3), pages 200–214.
- [Ma and Manjunath, 1999] Ma, W. Y. and Manjunath, B. S. (1999). Netra: A toolbox for navigating large image databases. *Multimedia Systems* 7(3), pages 184–198.
- [Manning et al., 2007] Manning, C., Raghavan, P., and Schütze, H. (2007). *An Introduction to information Retrieval*. Cambridge University Press, Cambridge, England.
- [Matos et al., 2008] Matos, T., Silva, I., Barcelos, C., and Proença, P. (2008). Avaliação de Índice invertido em busca de imagens por conteúdo. *XXXIV Conferencia Latinoamericana de Informática, 2008, Santa Fe-Argentina. Anais da XXXIV Conferencia Latinoamericana de Informática. Buenos Aires-Argentina : Sociedade Argentina de Informática*, 1:1–10.
- [Mills et al., 2000] Mills, T. J., Pye, D., Sinclair, D., and Wood, K. R. (2000). Shoebox: A digital photo management system. *Technical Report Technical Report*.
- [Müller et al., 1999] Müller, H., Squire, D. M., Müller, W., and Pun, T. (1999). Efficient access methods for content-based image retrieval with inverted files. In *Sethuraman Panchanathan, Shih-Fu Chang and C.-C. Jay Kuo eds., Multimedia Storage and Archiving Systems IV (VV02), Boston, Massachusetts, USA, Vol. 3846 of SPIE Proceedings SPIE Symposium on Voice, Video and Data Communications*), pages 20–22.
- [Nastar et al., 1998] Nastar, C., Mitschke, M., Meilhac, C., and Boujemaa, N. (1998). Surfimage: A flexible content-based image retrieval system. In *Proceedings of the ACM International Multimedia Conference*, 12-16:339–344.
- [Niblack et al., 1993] Niblack, C. W., Barber, R., Equitz, W., Flickner, M. D., Glasman, E. H., Petkovic, D., Yanker, P., Faloutsos, C., and Taubin, G. (1993). The qbic project: Querying images by content using color, texture, and shape. In *Pocceedings of the SPIE Conference on Storage and Retrieval for Image and Video Databases 2-3, February, San Jose, CA,,* pages 173–187.
- [Paris, 2008] Paris, A. C. (2008). Análise da eficiência de recuperação por conteúdo de imagens médicas, utilizando extratores de textura baseados em wavelet e wavelet packet. dissertação de mestrado. Master’s thesis, Universidade de São Paulo - São Carlos.
- [Pass and Zabith, 1996] Pass, G. and Zabith, R. (1996). Histogram refinement for content-based image retrieval. In *IEEE Workshop on Applications of Computer Vision*, pages 96–102.

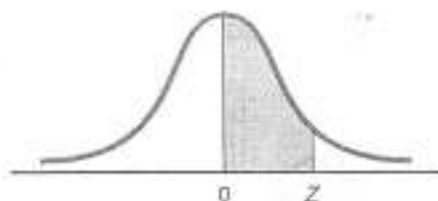
- [Pecenovie, 1998] Pecenovie, Z. (1998). Finding rainbows on the internet. *Technical report, Section of Communication Systems, Ecole Polytechnique Federal de Lausanne*.
- [Qian et al., 2002] Qian, G., Surat, S., and Pramanik, S. (2002). A comparative analysis of two distance measures in color image databases. *ICIP02*, pages 401–404.
- [Rijsbergen, 1976] Rijsbergen, V. C. J. (1976). *Information Retrieval*. Department of Computing Science University of Glasgow.
- [Rui et al., 1997] Rui, Y., Huang, T. S., and Mehrotra, S. (1997). Content-based image retrieval with relevance feedback in mars. *In Proceedings of International Conference on Image Processing*, 2:815–818.
- [Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523.
- [Schettinia et al., 2001] Schettinia, R., Ciocca, G., and Zu, S. (2001). A survey on methods for colour image indexing and retrieval in image databases. *In R. Luo and L. MacDonald (Eds.), Color Imaging Science: Exploiting Digital Media, Wiley, New York*.
- [Sclaroff et al., 1997] Sclaroff, S., Taycher, L., and Cascia, M. L. (1997). Imagerover: A content-based image browser for the world wide web. *In Proceedings IEEE Workshop on Content-based Access of Image and Video Libraries*.
- [Silva, 2007] Silva, S. F. (2007). Realimentação de relevância via algoritmos genéticos aplicada na recuperação de imagens. dissertação de mestrado. Master’s thesis, Universidade Federal de Uberlândia - UFU.
- [Silva et al., 2006] Silva, S. F., Barcelos, C. A. Z., and Batista, M. A. (2006). The effects of fitness functions on genetic algorithms applied to relevance feedback in image retrieval. *13th International Conference on Systems, Signals and Image Processing (IWSSIP). September, Budapest, Hungary*, 21-23:443–446.
- [Smeulders et al., 2000] Smeulders, A. W., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.
- [Smith, 1997] Smith, J. R. (1997). *Integrated Spatial and Feature Image Systems: Retrieval, Compression and Analysis*. PhD thesis, Graduate School of Arts and Sciences, Columbia University.
- [Squire et al., 1999] Squire, D. M., Müller, W., Müller, H., and Raki, J. (1999). Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. *Pattern Recognition Letters*, pages 143–149.
- [Srihari et al., 2000] Srihari, R., Zhang, Z., and Rao, A. (2000). Intelligent indexing and semantic retrieval of multimodal documents. *Information Retrieval*, 2(2):245-275.
- [Stricker and Orengo, 1995] Stricker, M. and Orengo, M. (1995). Similarity of color images. *In Proceedings of IS&T and SPIE Storage and Retrieval of Image and Video Databases III. San Jose, USA*.

- [Sudhamani and Venugopal, 2008] Sudhamani, M. V. and Venugopal, C. R. (2008). Multidimensional indexing structures for content-based image retrieval: A survey. *International Journal of Innovative Computing, Information and Control*, 4(4):867–881.
- [Swain and Ballard, 1991] Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision* 7 (1), pages 11–32.
- [Tamura and Yokoya, 1984] Tamura, H. and Yokoya, N. (1984). Image database systems: A survey. *Pattern Recognition* 17(1), pages 29–43.
- [Turceyan and Jain, 1998] Turceyan, M. and Jain, A. (1998). Texture analysis. *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, by C. H. Chen, L. F. Pau, P. S. P. Wang. World Scientific Publishing Co, pages 207–248.
- [Veltkamp and Tanase, 2000] Veltkamp, R. C. and Tanase, M. (2000). Content-based image retrieval systems: A survey.
- [Zhang, 2002] Zhang, D. (2002). *Image Retrieval Based on Shape*. PhD thesis, Faculty of Information Technology Monash University.
- [Zhang and Lu, 2004] Zhang, D. and Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, 37(1–19).
- [Zhou and Huang, 2003] Zhou, X. S. and Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems*, 8:536–544.

Apêndice A

Tabela Completa da Distribuição Normal Padronizada

Tabela E.2 A Distribuição Normal Padronizada



Os dados representam a área sob a distribuição normal padronizada, desde a média aritmética até Z

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49897	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952	0,49953	0,49955	0,49957	0,49958	0,49960	0,49961	0,49962	0,49964	0,49965
3,4	0,49966	0,49968	0,49969	0,49970	0,49971	0,49972	0,49973	0,49974	0,49975	0,49976
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983
3,6	0,49984	0,49985	0,49985	0,49986	0,49986	0,49987	0,49987	0,49987	0,49988	0,49989
3,7	0,49989	0,49990	0,49990	0,49990	0,49991	0,49991	0,49992	0,49992	0,49992	0,49992
3,8	0,49993	0,49993	0,49993	0,49994	0,49994	0,49994	0,49994	0,49995	0,49995	0,49995
3,9	0,49995	0,49995	0,49996	0,49996	0,49996	0,49996	0,49996	0,49996	0,49997	0,49997

(continua)

Figura A.1: Distribuição Normal Padronizada