Recuperação Multimodal de Imagens Com Realimentação de Relevância Baseada em Programação Genética

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Rodrigo Tripodi Calumby e aprovada pela Banca Examinadora.

Campinas, 12 de fevereiro de 2010.

Ricardo da Silva Torres (Orientador)

Este exemplar corresponde à redação final da Tese/Dissertação devidamente corrigida e defendida por: RODRIGO TRIPODI CALUMBY
e aprovada pela Banca Examinadora.
Campinas, 29 de JUNHO de 10
COORDENADOR DE PÓS-GRADUAÇÃO
IC-PG

Prof. Dr. Julio César Lopez Hernández
Coord. Subst. de Pós-Graduação
Instituto de Computação/Unicamp
Matr. 28.620-1

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciencia da Computação.

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO IMECC DA UNICAMP
Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Calumby, Rodrigo Tripodi
C138r Recuperação multimodal de imagens com realimentação de relevância baseada em programação genética/Rodrigo Tripodi Calumby-- Campinas, [S.P. : s.n.], 2010.

Orientador : Ricardo da Silva Torres.
Dissertação (mestrado) - Universidade Estadual de Campinas, Instituto de Computação.

1.Sistemas de recuperação da informação. 2.Processamento de imagens. 3.Programação genética. 4.Descritores. 5.Recuperação da informação. I. Torres, Ricardo da Silva. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Título em inglês: Multimodal image retrieval with relevance feedback based on genetic programming

Palavras-chave em inglês (Keywords): 1.Information storage and retrieval systems. 2.Image processing. 3.Evolutionary programming (Computer science). 4.Descriptors. 5.Information retrieval.

Área de concentração: Sistemas de Recuperação da Informação

Titulação: Mestre em Ciência da Computação

Banca examinadora: Prof. Dr. Ricardo da Silva Torres (IC-UNICAMP)
Profa. Dra. Gisele Lobo Pappa (DCC/UFMG)
Prof. Dr. Neucimar Jerônimo Leite (IC-UNICAMP)

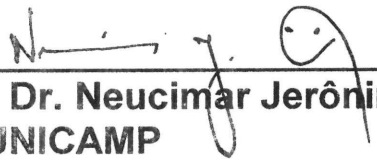
Data da defesa: 26/03/2010

Programa de Pós-Graduação: Mestrado em Ciência da Computação

TERMO DE APROVAÇÃO

Dissertação Defendida e Aprovada em 26 de março de 2010, pela Banca examinadora composta pelos Professores Doutores:

Profª. Drª. Gisele Lobo Pappa
DCC / UFMG



Prof. Dr. Neucimar Jerônimo Leite
IC / UNICAMP

Prof. Dr. Ricardo da Silva Torres
IC / UNICAMP

Instituto de Computação
Universidade Estadual de Campinas

Recuperação Multimodal de Imagens Com Realimentação de Relevância Baseada em Programação Genética

Rodrigo Tripodi Calumby[1]

Fevereiro de 2010

Banca Examinadora:

- Ricardo da Silva Torres (Orientador)
- Gisele Lobo Pappa
  Departmento de Ciência da Computação - UFMG
- Neucimar Jerônimo Leite
  Instituto de Computação - UNICAMP
- Hélio Pedrini (Suplente)
  Instituto de Computação - UNICAMP
- Adriano Veloso (Suplente)
  Departmento de Ciência da Computação - UFMG

[1]Suporte financeiro de: Bolsa do CNPq (edital MCT/CNPq nº 27/2007), Projeto 557430/2008-9.

Resumo

Este trabalho apresenta uma abordagem para recuperação multimodal de imagens com realimentação de relevância baseada em programação genética. Supõe-se que cada imagem da coleção possui informação textual associada (metadado, descrição textual, etc.), além de ter suas propriedades visuais (por exemplo, cor e textura) codificadas em vetores de características. A partir da informação obtida ao longo das iterações de realimentação de relevância, programação genética é utilizada para a criação de funções de combinação de medidas de similaridades eficazes. Com essas novas funções, valores de similaridades diversos são combinados em uma única medida, que mais adequadamente reflete as necessidades do usuário.

As principais contribuições deste trabalho consistem na proposta e implementação de dois arcabouços. O primeiro, RFCore, é um arcabouço genérico para atividades de realimentação de relevância para manipulação de objetos digitais. O segundo, MMRF-GP, é um arcabouço para recuperação de objetos digitais com realimentação de relevância baseada em programação genética, construído sobre o RFCore.

O método proposto de recuperação multimodal de imagens foi validado sobre duas coleções de imagens, uma desenvolvida pela Universidade de Washington e outra da *ImageCLEF Photographic Retrieval Task*. A abordagem proposta mostrou melhores resultados para recuperação multimodal frente a utilização das modalidades isoladas. Além disso, foram obtidos resultados para recuperação visual e multimodal melhores do que as melhores submissões para a *ImageCLEF Photographic Retrieval Task 2008*.

Abstract

This work presents an approach for multimodal content-based image retrieval with relevance feedback based on genetic programming. We assume that there is textual information (e.g., metadata, textual descriptions) associated with collection images. Furthermore, image content properties (e.g., color and texture) are characterized by image descriptores. Given the information obtained over the relevance feedback iterations, genetic programming is used to create effective combination functions that combine similarities associated with different features. Hence using these new functions the different similarities are combined into a unique measure that more properly meets the user needs.

The main contribution of this work is the proposal and implementation of two frameworks. The first one, RFCore, is a generic framework for relevance feedback tasks over digital objects. The second one, MMRF-GP, is a framework for digital object retrieval with relevance feedback based on genetic programming and it was built on top of RFCore.

We have validated the proposed multimodal image retrieval approach over 2 datasets, one from the University of Washington and another from the ImageCLEF Photographic Retrieval Task. Our approach has yielded the best results for multimodal image retrieval when compared with one-modality approaches. Furthermore, it has achieved better results for visual and multimodal image retrieval than the best submissions for ImageCLEF Photographic Retrieval Task 2008.

Agradecimentos

Eu gostaria de agradecer a...
Deus por cada instante de presença em minha vida e pela sabedoria renovada a cada dia. À minha avó Evanildes Tripodi, por todos estes anos de dedicação apaixonada. Ao meu avô Aldo Tripodi, pelos ensinamentos e exemplos de comportamento e conduta. Aos meus pais, cuja distância não ofusca o amor incondicional. À minha irmã Maria Luiza, pela confiança e apoio. Aos meus tios, tias, primos e primas, por cada palavra e gesto de orientação e incentivo. À Isaura Rennaly, por todo carinho, amor e dedicação que me sustentaram até aqui. Aos amigos e colegas do Instituto de Computação, pelos momentos de descontração e orientação mútua. Aos colegas, amigos, parceiros, colaboradores do LIS, pela convivência diária, pelas constantes e valiosas palavras e atitudes solidárias, por abraçarem com carinho, naturalidade e descontração cada um dos apelidos criados nestes 2 anos e pelo 1º lugar no Campeonato Interlabs IC/2009. Em especial, os amigos Fábio Faria (Gordinho ou Fabíola), Jerfersson Alex (Gerso), Otávio Penatti (Otaviano ou Otaviô), Ricardo Panaggio (Pâna ou Penacho), por todas as atividades colaborativas de pesquisa e pessoais. À profa. Cláudia Bauzer Medeiros, por todas as lições, ensinamentos, exemplos, apoio e por sua fiel dedicação à computação e à ciência. À CAPES, FAPESP e CNPq (Edital MCT/CNPq no 27/2007 - Mestrado, Projeto número 557430/2008-9). Por fim e não menos importante, ao meu orientador prof. Ricardo Torres, peça fundamental na realização deste trabalho, por toda atenção dispensada, paciência, ponderação e bom senso comportamental e pelas importantes e incontáveis sugestões.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Atualmente, com os avanços tecnológicos, um grande conjunto de imagens digitais é gerado, manipulado e armazenado em grandes bancos de imagens. Esses acervos são empregados em várias aplicações, tais como sensoriamento remoto, medicina e bibliotecas digitais [21, 64, 72]. Dado o tamanho desses bancos, prover meios de recuperar imagens de tais acervos de forma eficiente e eficaz é essencial.

A abordagem mais comum para recuperação de imagens é baseada na criação de descrições textuais (metadados ou palavras-chaves definidas por usuários) e no uso de técnicas tradicionais de bancos de dados para recuperá-las [75, 85], operação que exige a anotação prévia das imagens. Entretanto, o processo de anotação costuma ser ineficiente pois é comum que os usuários não façam anotações de forma sistemática. Usuários diferentes acabam usando palavras distintas para uma mesma característica. Esta falta de sistematização prejudica o desempenho da busca por palavras-chaves, uma vez que ela se baseia na igualdade entre as palavras anotadas na imagem e as fornecidas como parâmetros da busca. Apesar isso, a abordagem baseada em palavras-chaves possui a vantagem de abranger qualquer descrição que o usuário deseje [46].

Um outro paradigma para recuperação utiliza a descrição do conteúdo de imagens para indexá-las e manipulá-las [22, 45]. Nesses sistemas de *recuperação de imagens por conteúdo*, o processo de busca consiste em, dada uma imagem, calcular a sua similaridade em relação a outras armazenadas numa dada coleção de imagens. Para o cálculo da similaridade, são extraídos vetores de características. Esses vetores caracterizam propriedades da imagem, como cor, textura e forma, a partir da análise de seu *conteúdo*, ou seja, seus *pixels*. Assim, a similaridade entre duas imagens pode ser medida pela distância entre seus respectivos vetores de características. Várias técnicas são utilizadas para implementar esse processo [28, 92, 98].

A tarefa de comparar duas imagens é realizada por descritores [36,76,101,103,127]. Um descritor pode ser caracterizado por [22]: (i) um *algoritmo de extração de características*,

baseado em técnicas de processamento de imagens, que codifica as propriedades da imagem em um *vetor de característica*; e (ii) uma *medida de similaridade* (função de distância) que computa a similaridade entre duas imagens como uma função de distância entre seus vetores de características correspondentes. No domínio de recuperação de imagens por conteúdo, usualmente um descritor é considerado melhor que outro se sua utilização resulta em um número maior de imagens relevantes retornadas para uma dada consulta.

Descrições visuais e textuais codificam diferentes propriedades de uma imagem. Porém, na maioria das vezes, deseja-se recuperar uma imagem em função de múltiplas propriedades. Assim, descritores são combinados com o intuito de suprir tais necessidades. Vários métodos são utilizados para essa combinação [33, 111]. Grande parte desses métodos são baseados na atribuição de pesos aos descritores. Esses pesos determinam a relevância de cada descritor na composição.

Recentemente foi proposta em [24,25] uma estratégia para a composição de descritores visuais. Essa estratégia é baseada em uma técnica de otimização da *Inteligência Artificial* chamada *Programação Genética* [10], que é utilizada em várias aplicações de mineração de dados, processamento de sinais, evolução interativa, dentre outras [11, 37, 69, 125]. Essa técnica busca soluções ótimas se baseando na seleção natural das espécies. Ou seja, os indivíduos mais aptos (melhores soluções) tendem a se reproduzir e evoluir nas gerações futuras. No trabalho apresentado em [24, 25], essa técnica é empregada na combinação dos valores de similaridade obtidos a partir de descritores, gerando funções de similaridade mais eficazes. Assim, a Programação Genética é utilizada para combinar descritores pré-definidos, com o intuito de obter uma melhor composição desses descritores.

Porém, nesse tipo de abordagem não há nenhuma interação entre o usuário e o sistema. Esse fato dificulta a aquisição de imagens relevantes quando se consideram vários usuários. Isso acontece porque diferentes pessoas podem ter percepções visuais distintas diante de uma mesma imagem. Com isso, surge a necessidade de aprimorar essa estratégia para que ela se adapte a diferentes usuários. Para suprir essa necessidade pode-se utilizar uma técnica de recuperação de imagem interativa chamada realimentação de relevância [93] em que um usuário interage com o sistema de busca e assim refina o resultado de uma determinada consulta de acordo com sua necessidade.

Realimentação de relevância é uma técnica inicialmente utilizada na recuperação de informações por texto [52, 120], mas que atualmente é alvo de pesquisa na área de recuperação de imagem [62, 65, 74, 77, 93, 130]. O processo de recuperação por meio dessa técnica consiste basicamente na: (i) realização de uma consulta em que é retornado um pequeno número de imagens; (ii) indicação das imagens relevantes e/ou irrelevantes pelo usuário; (iii) re-combinação dos descritores de forma automática, considerando a indicação do usuário. Esse processo é repetido até que um resultado satisfatório para o usuário seja obtido. Em geral, uma grande eficácia no resultado de uma consulta é obtida após poucas

iterações [72, 93].

Recentemente, foi proposta no IC-Unicamp uma nova técnica de realimentação de relevância baseada em programação genética que vem sendo utilizada no processo de recuperação de imagem por conteúdo [34, 44]. Essa técnica busca soluções ótimas baseando-se na seleção natural de indivíduos que representam soluções do problema-alvo dentro de uma população. Assim, a Programação Genética é utilizada para combinar descritores pré-definidos, com o intuito de obter uma melhor eficácia no processo de recuperação de imagens.

Este projeto de pesquisa propõe a extensão desta técnica de realimentação de relevância e tem como objetivo especificar e implementar parcialmente um sistema de recuperação de imagens que empregue a técnica de realimentação de relevância baseada em programação genética para combinar descrições textuais e visuais. O uso de programação genética visa melhorar o processo de combinação de descritores, agregando as informações fornecidas pelo usuário, e assim conseguir resultados mais relevantes em uma busca em um dado banco de imagens após poucas iterações.

As principais contribuições deste projeto de pesquisa são:

- Estudo comparativo de diferentes técnicas de realimentação de relevância;
- Estudo comparativo de técnicas de recuperação de imagens multimodais e por conteúdo;
- Especificação e implementação de um arcabouço genérico de realimentação de relevância para manipulação de objetos digitais;
- Especificação e implementação de um modelo de realimentação de relevância multimodal baseada em programação genética;
- Implementação parcial de um sistema de recuperação multimodal de imagens com a utilização de realimentação de relevância baseada em programação genética para combinação de evidências textuais e visuais.

O restante deste documento é organizado como segue: o próximo capítulo apresenta os conceitos básicos e trabalhos correlatos; O Capítulo 3 descreve o arcabouço proposto para realimentação de relevância. O Capítulo 4 descreve o modelo de realimentação de relevância multimodal baseada em programação genética. O Capítulo 5 apresenta o sistema de recuperação de imagens baseada em evidências textuais e visuais bem como sua validação experimental e análise dos resultados. Finalmente, o Capítulo 6 apresenta as conclusões e discute possíveis extensões para este trabalho.

Capítulo 2

Conceitos e Trabalhos Correlatos

Este capítulo apresenta os conceitos fundamentais utilizados ao longo desta dissertação, além de trabalhos correlatos das áreas de abragência do trabalho. A seção 2.1 apresenta conceitos relacionados à recuperação textual de imagens. A seção 2.2 descreve sistemas de recuperação de imagens por conteúdo, bem como formalização de descritores de imagens. Esta formalização é usada ao longo da dissertação. A seção 2.3 apresenta as caracterísitcas da recuperação de imagens baseada em múltiplas modalidades. A seção 2.4 descreve a meta-heurística de programação genética e sua aplicação para a combinação de medidas de similaridade. A seção 2.5 apresenta conceitos da técnica de realimentação de relevância e sua aplicação na recuperação de imagens. Por fim, a seção 2.6 descreve diferentes trabalhos que exploram a técnica de realimentação de relevância para recuperação de imagens por conteúdo.

2.1 Recuperação Textual de Imagens

Sistemas tradicionais de recuperação de imagens baseiam-se no cálculo de relevância baseado em informações textuais associadas às imagens da coleção [14, 17, 59, 75, 85]. Essas informações podem ser obtidas a partir de diferentes fontes como, por exemplo, texto em páginas web, legendas, metadados, palavras-chaves, descrições textuais e até mesmo reconhecimento ótico de caracteres.

Para comparar as descrições textuais diferentes medidas de similaridades podem ser utilizadas. Para obter essas similaridades, uma das técnicas mais utilizadas consiste em representar os textos (documentos da coleção e consultas) no Modelo Vetorial (Vector Space Model [9]). Segundo este modelo, documentos são analisados como vetores. Suponha a existência de uma coleção com t termos distintos de índice t_j. Um documento d_i pode ser representado da seguinte maneira: $d_i = (w_{i1}, w_{i2}, \ldots, w_{it})$, onde w_{ij} representa o peso do termo t_j no documento d_i.

O peso para um dado termo t_j do documento d_i pode ser calculado como o respectivo valor $tf \times idf$, onde tf indica a frequência do termo no documento e idf a frequência inversa do termo na coleção. O valor de idf é dado por $log(N/nt)$, sendo N o número total de documentos na coleção e nt o número de documentos em que o termo t_j ocorre. Para um processo de recuperação simples, dado um termo de consulta, a coleção de documentos pode ser ordenada apenas de acordo com o peso do termo de consulta nos documentos. Para consultas com múltiplos termos, uma função de agregação pode ser utilizada para combinar os pesos dos termos e assim gerar uma medida de relevância final.

Considerando o espaço vetorial e o modelo de ponderação tf-idf, algumas medidas de similaridades são descritas a seguir.

$$Bag - of - words(d_1, d_2) = \frac{|\{d_1\} \bigcap \{d_2\}|}{|d_1|}, \tag{2.1}$$

onde $\{d_i\}$ representa o conjunto de termos que ocorrem no documento d_i. Essa é uma medida simples da porcentagem de palavras em comum dos dois documentos analisados.

$$Cosseno(d_1, d_2) = \frac{\sum_{i=1}^{t} w_{1i} \times w_{2i}}{\sqrt{\sum_{i=1}^{t} w_{1i}^2 \times \sum_{i=1}^{t} w_{2i}^2}} \tag{2.2}$$

onde w_{ij} é o documento como definido anteriormente. Essa fórmula basicamente calcula o cosseno entre os dois vetores que representam os documentos, de forma que quanto mais próximo de 1 seja esse cosseno, mais similares são os documentos.

$$Okapi(d_1, d_2) = \sum_{t \in d1 \bigcap d2} \frac{3 + tf_{d2}}{0,5 + 1,5 \times \frac{tam_{d2}}{tam_{med}} + tf_{d2}} \times log \frac{N - df + 0,5}{df + 0,5} \times tf_{d1} \tag{2.3}$$

onde tf é a freqüência do termo no documento, df é a freqüência do termo do documento na coleção inteira, N é o número de documentos na coleção inteira, tam_{di} é o tamanho do documento i, e tam_{med} é o tamanho médio de todos os documentos da coleção.

2.2 Recuperação de Imagens por Conteúdo

A recuperação de imagens por conteúdo constitui uma área de pesquisa de ponta. Essa abordagem é centrada na noção de similaridade de imagens — dado um banco de dados com um grande número de imagens, o usuário deseja recuperar as imagens mais similares a um padrão de consulta (normalmente uma imagem). A maioria destes sistemas está alicerçada na extração de características visuais (cor, textura e forma) e comparação das imagens por meio de descritores [5, 45, 47, 79, 85].

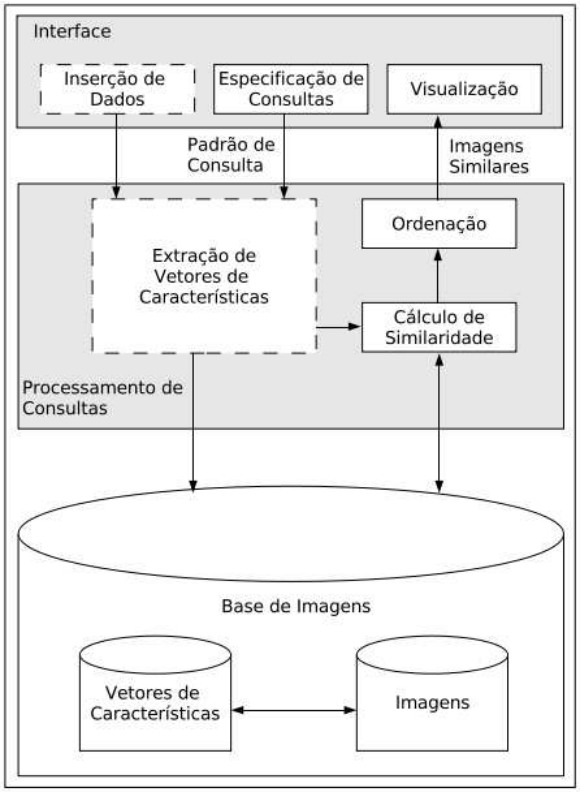


Figura 2.1: Arquitetura típica de um sistema de recuperação de imagens por conteúdo [26].

A Figura 2.1 mostra a arquitetura típica de um sistema de recuperação de imagens por conteúdo. A interface permite ao usuário especificar uma consulta a partir de um padrão (por exemplo, a partir da definição de uma imagem de consulta – *query by visual example* [54]) e visualizar as imagens recuperadas. O módulo de processamento de consultas extrai o vetor de características do padrão de consulta e aplica uma métrica de distância para avaliar a similaridade entre a imagem de consulta e as imagens da base. Depois, esse módulo ordena as imagens da base de acordo com a similaridade e retorna as mais similares para o módulo de interface.

Para informações detalhadas, em [113] são encontradas descrições de um conjunto significativo de sistemas de recuperação de imagens por conteúdo.

A seguir será mostrada a formalização de um método para a recuperação de imagens por conteúdo, conforme exposto em [24, 25]. Esta formalização será usada ao longo da dissertação.

Definição 1. *Uma* **imagem** $\hat{I}$ *é um par* $(D_I, \vec{I})$*, onde:*

- D_I *é um conjunto finito de* pixels *(pontos em* $\mathbb{Z}^2$*, tal que,* $D_I \subset \mathbb{Z}^2$*), e*
- $\vec{I} : D_I \rightarrow \mathsf{D}'$ *é uma função que atribui a cada* pixel p *em* D_I *um vetor* $\vec{I}(p)$ *de valores em algum espaço arbitrário* D' *(por exemplo,*$\mathsf{D}' = \mathbb{R}^3$ *quando uma cor é atribuída a um* pixel *no sistema RGB).*

Definição 2. *Um* ***descritor simples*** *(ou simplesmente,* ***descritor****) D é definido como um par* (ϵ_D, δ_D)*, onde:*

- $\epsilon_D : \hat{I} \rightarrow \mathbb{R}^n$ *é uma função que extrai um* vetor de características $\vec{v}_{\hat{I}}$ *de uma* imagem $\hat{I}$*.*
- $\delta_D : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ *é uma* função de similaridade *(por exemplo, baseada em uma medida de distância) que computa a similaridade entre duas imagens a partir da distância entre seus* vetores de características *correspondentes.*

Definição 3. *Um* ***vetor de características*** $\vec{v}_{\hat{I}}$ *de uma imagem* $\hat{I}$ *é um ponto no espaço* $\mathbb{R}^n$*:* $\vec{v}_{\hat{I}} = (v_1, v_2, ..., v_n)$*, onde n é a dimensão do vetor.*

A Figura 2.2 ilustra o uso de um descritor simples D para computar a similaridade entre duas imagens $\hat{I}_A$ e $\hat{I}_B$. Primeiro, o algoritmo de extração ϵ_D é usado para computar os vetores de características $\vec{v}_{\hat{I}_A}$ e $\vec{v}_{\hat{I}_B}$ associados com as imagens. Depois, a função de similaridade δ_D é utilizada para o valor da similaridade d entre as imagens.

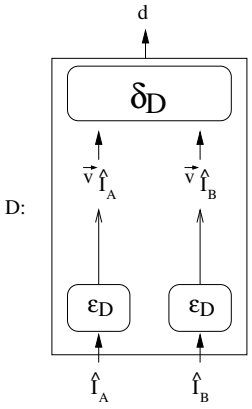


Figura 2.2: O uso de um descritor simples D para computar a similaridade entre duas imagens [22].

Definição 4. *Um* ***descritor composto*** $\hat{D}$ *é um par* $(\mathcal{D}, \delta_{\mathcal{D}})$ *(veja Figura 2.3), onde:*

- $\mathcal{D} = \{D_1, D_2, \ldots, D_k\}$ *é um conjunto de* k *descritores simples pré-definidos.*
- $\delta_{\mathcal{D}}$ *é uma função de similaridade que combina os valores de similaridade obtidos de cada descritor* $D_i \in \mathcal{D}$*,* $i = 1, 2, \ldots, k$*.*

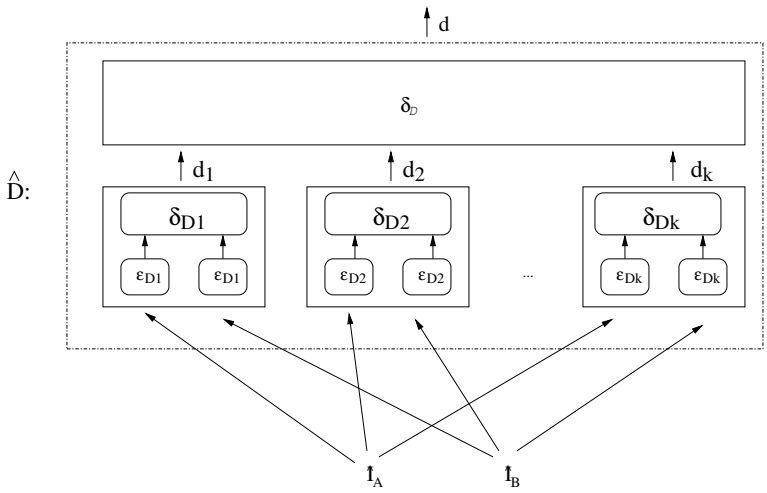


Figura 2.3: Descritor composto [22].

2.3 Recuperação Multimodal

A multimodalidade inerente aos dados multimídia atrai cada vez mais interesse de pesquisadores que buscam criar técnicas que se beneficiem do maior número de informação disponível sobre um dado item [12, 17, 106, 110, 115, 128]. Na recuperação de vídeos, por exemplo, podem ser exploradas informações presentes nos *frames* do vídeo, no som e em textos obtidos a partir do reconhecimento de fala, reconhecimento ótico de caracteres e/ou *closed captions* [121]. Na recuperação textual de imagens na web, trabalhos como [14, 17, 59, 95] estudam a combinação das evidências disponíveis sobre as imagens, sendo elas: URL, legendas, título da página, título da imagem, texto *alt* e o texto ao redor da imagem.

Na recuperação de imagens, uma questão importante é a modalidade da consulta a ser utilizada. Os primeiros sistemas utilizavam consultas baseadas em palavras-chave, criando a necessidade de anotação textual do conteúdo da imagem. Contudo, as descrições textuais estão sujeitas aos problemas de inconsistência da anotação, existência de sinônimos e polisemia [129]. Os sistemas de recuperação por conteúdo suprimem a

necessidade de anotação utilizando uma imagem de exemplo como padrão de consulta e ordenando as imagens da base de acordo com sua similaridade visual. Entretanto, esta segunda abordagem possui duas desvantagens principais: o usuário nem sempre consegue definir uma imagem adequada para a consulta desejada e a representação da imagem por suas características visuais não é tão flexível quanto as descrições textuais [71].

Com o objetivo de aprimorar a eficácia dos sistemas de recuperação de imagens, diversos trabalhos propõem e aplicam abordagens multimodais. Neste sentido, mecanismos de recuperação utilizam cálculos de similaridades baseados no conteúdo visual da imagem e em descrições textuais associadas. A utilização de múltiplas modalidades permite ao sistema se beneficiar das vantagens de cada uma delas. Além disso, acredita-se que as abordagens textuais e visuais são suplementares e, desta forma, a abordagem multimodal demonstra-se promissora à medida que consegue equilibrar os benefícios e as dificuldades da utilização de cada modalidade em separado.

Sistemas que utilizam da multimodalidade buscam o chamado efeito Chorus. Segundo [115], "o efeito Chorus ocorre quando várias abordagens de recuperação sugerem que um item é relevante para uma consulta (...) e isto tende a ser uma evidência mais forte da relevância do que se a relevância fosse indicada por apenas uma abordagem". Segundo [121], a combinação de resultados de múltiplas modalidades pode consistentemente melhorar o desempenho da recuperação quando comparada com a utilização isolada de cada modalidade.

Alguns trabalhos já exploraram a multimodalidade na combinação de evidências textuais e visuais de imagens. Estes trabalhos demonstraram uma significativa melhoria nos resultados quando foi utilizada a combinação das modalidades e são descritos a seguir.

O projeto *Chabot* [85], um dos primeiros a usufruir de múltiplas modalidades, busca integrar as técnicas de análise de imagem baseada em cor com sistemas de recuperação baseado em informações textuais. Ele inclui uma interface que permite ao usuário consultar e atualizar um banco de imagens. O mecanismo de consulta recupera a imagem com base nos dados textuais armazenados e em relações complexas entre estes dados. Estas relações são caracterizadas pelas cores encontradas nas imagens. Este sistema não investiga técnicas de análise de conteúdo como textura e forma.

No trabalho apresentado em [2] é proposta uma técnica de recuperação multimodal de imagens por similaridade que utiliza *keywords* visuais construídas a partir de descritores de cor MPEG-7 e *keywords* textuais. Estas *keywords* são combinadas e a coleção de imagens é representada como uma matriz, similar àquela que representa termos e documentos. Nesta abordagem, as imagens são segmentadas em uma malha pré-definida e os *tiles* (quadros) resultantes são agrupados. *Tiles* pertencentes ao mesmo grupo (*cluster*) são considerados idênticos. Cada um dos *clusters* é tratado como um termo. A matriz das imagens e seus termos é criada a partir da concatenação dos *clusters* (termos visuais)

com os termos obtidos das descrições textuais. Assim, cada imagem é representada com um vetor cujos valores indicam a quantidade de *tiles* pertencentes a cada *cluster* ou o número de ocorrência de uma *keyword textual*. Esta é uma representação esparsa, pois uma imagem normalmente possui poucas *keywords* quando comparada com a quantidade total de *keywords* da coleção. A matriz resultante é utilizanda para o cálculo dos pesos dos termos usando o algoritmo tf-idf (term frequency - inverse document frequency) [9].

Como uma variação de [2], em [3] é proposta a utilização de uma abordagem não linear de difusão baseada em *kernel* para fusão das diferentes modalidades de *keywords* apresentada em [29]. Tanto em [2] quanto em [3], experimentos mostraram que os melhores resultados são obtidos quando a recuperação considera tanto *keywords* visuais quanto *keywords* textuais.

Em [105] é proposto um sistema de recuperação de imagens que combina recuperação por conteúdo visual com a utilização de informações da estrutura e do conteúdo textual de documentos XML. Assim, são utilizadas máquinas de busca independentes para cada modalidade e seus resultados são combinados para produção do resultado final. Neste sentido, as imagens da base são ordenadas por similaridade e em seguida para cada um dos seus elementos é atribuída heuristicamente uma pseudo-frequência assim como é feito com os termos em uma busca textual. Assim, o mecanismo de *ranking* textual pode ser aplicado para ambas as modalidades.

Em [71] é apresentada a proposta de um sistema para recuperação multimodal de imagens que, assim como [105], utiliza duas máquinas de buscas concorrentes, uma baseada no conteúdo visual da imagem e outra baseada na anotação textual (utilizando uma variante do tf-idf). Assim, é feita uma soma ponderada das similaridades calculadas com base em cada descrição visual. O resultado da busca por conteúdo é utilizado para refinar os resultados obtidos com a busca textual. A idéia desta abordagem é que se uma imagem aparece em ambos os resultados, seu *ranking* deve ser elevado no resultado final.

Em [60] é proposta uma solução que utiliza uma combinação linear de evidências de um sistema de recuperação de imagens por conteúdo (GIFT[1]) e um sistema de recuperação orientado a conteúdo para documentos XML. Nesta abordagem os *ranks* obtidos para cada imagem nas diferentes modalidades de recuperação são somados de maneira ponderada.

Em [114] é apresentado um estudo de diferentes estratégias de fusão para métodos de combinação de técnicas de recuperação visual e textual. O sistema proposto é composto de três módulos: módulo de recuperação textual, módulo de recuperação visual, e módulo de fusão. Este último é responsável por fundir os resultados obtidos pelos primeiros. A fusão é realizada em duas etapas. Na primeira, as imagens do resultado final são definidas por meio de um operador de combinação que pode ser: união, intersecção, junção externa (esquerda ou direita). As imagens resultantes da fusão são então re-ordenadas com base em um novo valor de relevância. Este valor é obtido com base nos valores de relevância

calculados pelos módulos de recuperação visual e textual e é definido por operadores como: máximo, mínimo, média, max-min ($mm = max(a,b) + min(a,b) \times min(a,b)/(max(a,b) + min(a,b))$). Resultados experimentais mostram maior eficácia para a combinação de resultados usando a junção externa esquerda e re-ordenação com o operador max-min.

Early Fusion versus Late Fusion

Existem duas técnicas básicas para fusão de modalidades: *early fusion* e *late fusion*. Estas duas abordagens diferem no modo em que são combinadas as características obtidas a partir de cada modalidade. Como descrito em [99], abordagens que se baseiam em *early fusion* primeiramente extraem as características referentes a cada uma das modalidades. Depois da análise de cada componente das modalidades, estas são então combinadas para formar uma única representação. Por sua vez, abordagens baseadas em *late fusion* também realizam a extração das características para cada uma das modalidades, entretanto, ao contrário da *early fusion*, os algoritmos de aprendizado são aplicados a cada uma das modalidades em separado. Os resultados obtidos a partir de cada modalidade são então combinados para produzir a classificação final dos itens da coleção.

A *early fusion* permite a criação de uma representação verdadeiramente multimodal, visto que as características dos itens da coleção são combinados desde o início do processo. Contudo, a combinação das diferentes modalidades em uma representação única mostra-se complexa e ainda se configura como campo aberto de pesquisa [71].

A *late fusion* baseia-se na combinação dos resultados obtidos de maneira independente em cada modalidade e possui a desvantagem de exigir a execução em separado de algoritmos de aprendizado e algoritmos de combinação dos resultados obtidos com cada uma delas.

Para técnicas baseadas em *late fusion* existem diversas estratégias para realizar a combinação dos resultados obtidos em cada modalidade. Em [66], são apresentadas várias estratégias incluindo a combinação utilizando-se o produto, soma ponderada, votação e agregação min-max. Segundo [119], dentre estes métodos, os mais populares são a combinação utilizando-se do produto e a soma ponderada, sendo que esta segunda é mais tolerante a ruídos do que a primeira. Em [32], os autores mostram que a soma ponderada está entre as abordagens mais eficazes para a recuperação baseada na combinação de texto e imagem.

No trabalho em [51], são comparados os desempenhos da utilização de características visuais e da combinação com texto. Foram utilizadas as agregações de características, soma ponderada da posição no *ranking* e soma ponderada da posição inversa do *ranking*. O estudo concluiu que a combinação de evidências gera melhores resultados do que a recuperação apenas visual e que a soma ponderada da posição inversa do *ranking* é mais eficiente.

Alguns trabalhos como [61, 99] já realizaram estudos comparativos entre *early fusion* e *late fusion*. Entretanto, ainda não existe um consenso sobre qual abordagem é a mais eficiente. Em [61], os autores afirmam que resultados obtidos indicam que sistemas baseados em *late fusion* proporcionam poucos ganhos frente aos sistemas unimodais. Já em [99], os autores concluem que esquemas baseados em *late fusion*, na maioria dos casos, tendem a obter melhor desempenho, entretanto, quando os resultados obtidos com a *early fusion* são melhores, a diferença é ainda mais significativa. Os experimentos tanto de [61] quanto [99] foram realizados com base na coleção TRECVID [84].

Um resumo das características de trabalhos correlatos em recuperação multimodal de imagens e suas respectivas coleções de teste pode visto nas tabelas 2.1 e 2.2. N/D corresponde a itens não definidos.

Apesar dos trabalhos presentes nas tabelas 2.1 e 2.2 estarem todos relacionados à recuperação multimodal, alguns apresentam características especiais. O trabalhos em [2] e [3] foram os únicos a utilizar a segmentação das imagens em forma de malha. Além disso, os trabalhos [7,102] diferenciam-se dos demais por não utilizarem imagens, mas sim vídeos e páginas *web*.

Tabela 2.1: Resumo das características de trabalhos correlatos em recuperação multimodal.

Artigo	Características Visuais	Características Textuais	Fusão	Consulta	Medidas de validação
Agrawal et al. (2006) [2]	MPEG-7: SCD, CLD e CSD	Controle morfológico [112]. Geração da matriz de termos e documentos [123]	Early Fusion	Multimodal	Precision/Recall
Agrawal et al. (2007) [3]	MPEG-7: SCD, CLD e CSD	Keywords textuais são unidas às keywords visuais	Early Fusion	Multimodal	Precision/Recall
Tjondronegoro et al. (2006) [105]	Histograma de cor, textura, linhas detectáveis e extração de objetos	Lista invertida de termos em um banco Access. Variante do TF-IDF	Late Fusion	Multimodal	Médias de Precision/-Recall interpoladas
Tjondronegoro et al. (2007) [71]	Histograma de cor, histograma de cor de objetos, hough transform, textura.	Engine de busca em XML: GPX [126]	Late Fusion	Multimodal	Precision/Recall
Iskandar et al. (2006) [60]	GIFT[1]	Engines: Zettair (engine para busca textos) e eXist (banco de dados para XML)	Late Fusion	Multimodal	Precision/Recall, MAP, Average Interpolated Precision, TRECeval, HiXEval
Amir et al. (2005) [7]	cor, textura e localizações [51]	Semantic conceps [6], video OCR [53], reconhecimento de fala [6, 48] e closed captions	Early Fusion	Multimodal	Mean Average Precision (MAP)
Fernández et al. (2006) [80]	GIFT[1]	Lucene[2] e Ksite[3]	Late Fusion	Multimodal	MAP
Villena-Román et al. (2008) [114]	GIFT[1] e Fire[5]	Lucene[2] e Xapian[4]	Late Fusion	Multimodal	MAP, P10, P20, P30, número de imagens relevantes retornadas
Jing et al. (2007) [63]	Auto-correlograma, RGB 64 bins com 4 distâncias 1, 3, 5 e 7.	Propagação de keywords baseada em características visuais	Não realiza fusão.	Textual ou Visual	Precision
Sznadjer et al. (2008) [102]	MPEG-7: ScalableColor, EdgeHistogram e ColorLayout	Filtragem boleana com base em keywords	Late Fusion	Multimodal	Recall

[1]GNU Image Finding Tool. http://www.gnu.org/software/gift/
[2]http://lucene.apache.org/
[3]http://www.daedalus.es/productos/k-site
[4]Xapian. http://www.xapian.org
[5]FIRE: Flexible Image Retrieval System. http://www-i6.informatik.rwth-aachen.de/ deselaers/fire/

Tabela 2.2: Resumo das coleções utilizadas em trabalhos correlatos em recuperação multimodal.

Artigo	Coleção	Dados	Tamanho	Categorias	Anotação
Agrawal et al. (2006) [2]	Labelme	Imagens	658	15	Sim
Agrawal et al. (2007) [3]	Corel e Labelme	Imagens	999 (Corel) e 658 (Labelme)	10 (Corel) e 15 (Labelme)	Sim (keywords)
Tjondronegoro et al. (2006) [105]	Lonely Planet XML document collection	Imagens	463 documentos XML e 1947 imagens	N/D	Sim (documentos XML)
Tjondronegoro et al. (2007) [71]	Inex 2006 MM	Imagens	166559	N/D	Sim (documentos XML)
Iskandar et al. (2006) [60]	Inex 2005 MM e Lonely Planet Collection	Imagens	N/D	N/D	Sim (documentos XML)
Amir et al. (2005) [7]	TRECVID 2004	Videos	N/D	N/D	Sim
Fernández et al. (2006) [80]	ImageCLEFmed 2005	Imagens	N/D	Não	Sim (estruturada em campos)
Villena-Román et al. (2008) [114]	ImageCLEFphoto 2007	Imagens	20.000	Não	Sim (estruturada em campos)
Jing et al. (2007) [63]	Corel	Imagens	10000	Sim	Sim (apenas 10% da coleção com 1 keyword correspondente à categoria)
Sznadjer et al. (2008) [102]	Páginas do Flickr[6]	Web pages	160000 páginas	N/D	Sim (título da página, tags e comentários)

2.4 Programação Genética

Programação genética (PG) [69] constitui um conjunto de técnicas da inteligência artificial para a solução de problemas baseadas nos princípios da herança biológica, seleção natural e evolução. Nesse contexto, cada solução potencial é chamada de indivíduo em uma população. Sobre essa população são aplicadas iterativamente transformações genéticas, como cruzamentos e mutações, com o intuito de criar indivíduos mais aptos (melhores soluções) em gerações subseqüentes. Uma função de adequação (*fitness*) é utilizada para atribuir valores para cada indivíduo com o intuito de definir o seu grau de adequação, ou evolução, perante os demais membros da população.

Em PG, a representação dos indivíduos pode ser realizada utilizando-se de estruturas, como árvores, listas encadeadas ou pilhas [70]. Além disso, o tamanho dos indivíduos em PG não é fixo, embora se possam restringir certos limites na implementação. Em virtude do paralelismo intrínseco no mecanismo de busca e de sua capacidade de exploração global em espaços de dimensões mais elevadas, PG é utilizada para resolver uma ampla gama de problemas de otimização.

Os métodos de PG são formados por componentes, como os mostrados na Tabela 2.3. Essa tabela define alguns componentes (e seus significados) quando se utiliza árvore como estrutura de representação dos indivíduos.

Em [24, 25], é proposta a utilização de programação genética para combinar descritores de imagens e caracterizam o processo de recuperação de imagens com PG como segue. Para um dado banco de imagens e um padrão de consulta fornecido pelo usuário, como uma imagem, o sistema retorna uma lista das imagens mais similares ao padrão de consulta, de acordo com um conjunto de propriedades da imagem. Essas propriedades

[6] www.flickr.com

Tabela 2.3: Componentes essenciais de PG.

Componentes	Significado
Terminais	Nós folhas na estrutura da árvore.
Funções	Nós não-folhas utilizados para combinar os nodos folha. Operações numéricas comuns: +, -, *, /, log e $\sqrt{\ }$
Função de Adequação	A função que PG busca otimizar.
Reprodução	Um operador genético que copia os indivíduos com os melhores valores de adequação diretamente para a próxima geração, sem passar pela operação de *crossover*.
Crossover	Um operador genético que troca sub-árvores de dois pais para formar dois novos descendentes.
Mutação	Um operador genético que troca uma sub-árvore de um determinado indivíduo, cuja raiz é um ponto de mutação escolhido, com uma sub-árvore gerada aleatoriamente.

são representadas por descritores simples (ver Definição 2 na Seção 2.2). Esses descritores são combinados utilizando um descritor composto $\mathcal{D}_{PG}$, onde $\delta_{\mathcal{D}_{PG}}$ é uma expressão matemática representada como uma árvore de expressão, em que os nós internos são operadores numéricos (veja Tabela 2.3) e os nós folha são um conjunto composto de valores de similaridade d_i, $i = 1, 2, \ldots, k$. A Figura 2.4 mostra uma possível combinação (obtida a partir do uso do arcabouço de PG) dos valores de similaridade d_1, d_2 e d_3 de três descritores simples.

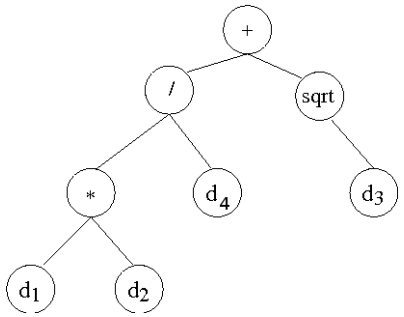


Figura 2.4: Exemplo de uma função de similaridade baseada em PG representada em uma árvore.

Em [24, 25], vários experimentos foram realizados visando avaliar o uso dessa abordagem para a combinação de descritores de forma. Com base nesses experimentos, conclui-se que os descritores combinados com PG apresentaram melhores resultados do que os obtidos quando os descritores eram utilizados separadamente.

Algoritmo 1 Algoritmo básico de recuperação de imagens com PG.

1 Gere a população inicial de árvores "aleatórias"
2 Para N_{gen} gerações, sobre conjunto de imagens de treinamento, faça:
3 Calcule a adequação de cada árvore de similaridade.
4 Armazene as melhores $N_{melhores}$ árvores de similaridade.
5 Crie uma nova população por:
6 Reprodução
7 *Crossover*
8 Mutação
9 Aplique a "melhor árvore de similaridade"(a melhor da última geração) em um conjunto de imagens de teste (consulta).

O processo geral de recuperação é mostrado pelo Algoritmo 1.

O trabalho apresentado em [27] propõe a *Abordagem de Componentes Combinados (ACC)*, um mecanismo de recuperação textual que visa realizar a descoberta de funções de ordenação que sejam adequadas para as caraterísticas de uma coleção específica de documentos. Para isso, são utilizadas como terminais para os indivíduos da programação genética, funções de similaridade presentes em diferentes modelos e sistemas de recuperação disponíveis na literatura e reconhecidamente eficazes [4, 13, 91, 97]. Um exemplo pode ser visto na Figura 2.5, onde em uma estrutura de árvore representando o modelo tf-idf.

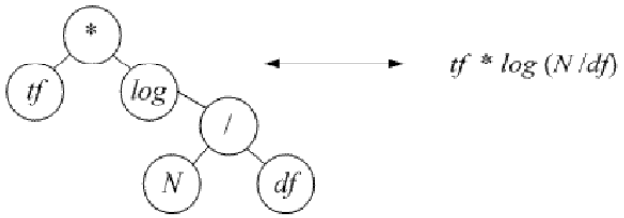


Figura 2.5: Um exemplo de uma árvore para um indivíduo *tf-idf* baseado em informações estatísticas da coleção. Fonte [27].

Ao contrário de trabalhos anteriores [38–42, 109], a ACC utiliza terminais mais significativos ao invés de apenas informações estatísticas básicas da coleção. Como visto na Figura 2.6, o idf é por si só um terminal e não uma combinação de outros terminais menos significativos como a quantidade de documentos na coleção. Em trabalhos anteriores o idf seria tratado como uma subárvore que deveria ser explicitamente descoberta pelo processo

Figura 2.6: Um exemplo de uma árvore para um indivíduo *tf-idf* baseado na Abordagem de Componentes Combinados, onde tf e idf são componentes. Fonte [27].

evolutivo, pondendo até mesmo não ser descoberta. O conjunto de terminais definidos pela ACC e suas descrições são apresentados na Tabela 2.4.

Na ACC, "para combinar os terminais na formação do indivíduo, foram utilizadas as seguintes funções: adição (+), multiplicação (×), divisão protegida (/) e logaritmo protegido (log). Logaritmo protegido é uma modificação da função de logaritmo natural para previnir o retorno de valores negativos. O logaritmo protegido irá retornar 0 (zero) se o argumento for menor do que 1 (um) ou o logartimo natural do valor absoluto do argumento passado para a função" [27]. Experimentos demonstraram que a ACC possui desempenho médio superior ao esquema de ponderação tf-idf, ao BM25 [90] e ao FAN-GAP [40].

Tabela 2.4: Terminais utilizados pela abordagem ACC. Fonte: [27]

Id	Terminal	Descrição
t_{01}	tf	Freqüência bruta do termo (número de vezes que um termo ocorre em um documento) [94]
t_{02}	$1 + \log(tf)$	Logaritmo natural da freqüência bruta do termo como apresentado em [117], usado para suavizar a influência da freqüência do termo
t_{03}	$0.5 + \frac{0.5+tf}{maxtf}$	Freqüência do termo em um documento normalizada pela freqüência máxima de um termo em um documento, e mais adiante ajustada para ficar situada entre 0,5 e 1,0 [8, 94]
t_{04}	$\frac{1+log(tf)}{1+log(avgtf)}$	Freqüência do termo em um documento normalizada pela freqüência média de termos em um documento, como definido em [97]. Parte da fórmula do esquema de ponderação do sistema SMART como definido em [13]
t_{05}	$\frac{(k_1+1)tf}{(k_1((1-b)+b\times dl/avgdl)+dl)+tf}$	Parte da função de ordenação do sistema Okapi BM25 com os componentes de freqüência do termo *tfc* e de normalização *nc* [90]
t_{06}	$\log(\frac{N}{df})$	Inverso da freqüência do documento (*idf*) [94]
t_{07}	$\log(\frac{N}{df}+1)$	Uma alternativa para o inverso da freqüência do documento (*idf*) como apresentado em [117]
t_{08}	$\log(\frac{N-df+0.5}{0.5})$	Uma variação para o peso definido por Robertson-Sparck Jones [31], que suaviza as diferenças entre os valores de *df*
t_{09}	$w^{(1)} = \log(\frac{N-df+0.5}{df+0.5})$	Peso definido por Robertson-Sparck Jones [89–91]
t_{10}	$\log(\frac{N-df}{df})$	Uma alternativa probabilística para o inverso da freqüência do documento [89, 94]
t_{11}	$\frac{\log\frac{N+0.5}{df}}{\log N+1}$	Parte da função de ordenação do sistema INQUERY para calcular a crença de um termo dentro de um documento como definido em [4]
t_{12}	$\frac{1}{\sqrt{\sum_{i=0}^{2} w_{i,j}^2}}$	Normalização do cosseno onde $w_{i,j}^2 = t_{01} \times t_{07}$ [94, 117]
t_{13}	$\frac{1}{\sqrt{\sum_{i=0}^{2} w_{i,j}^2}}$	Normalização do cosseno onde $w_{i,j}^2 = t_{02} \times t_{07}$ [94, 117]
t_{14}	dl	Normalização do tamanho do documento (em *bytes*). Técnica de normalização usada em coleções de documentos antigas tais como os repositórios baseados em OCR (digitalizados) [97]
t_{15}	$\frac{1}{(1-slope)+slope\times\frac{avgt_{13}}{t_{13}}}$	Normalização pivoteada do cosseno com o terminal t_{13} como definido em [97]
t_{16}	$\frac{1}{(1-slope)\times avgdl+slope\times dl}$	Normalização pivoteada do tamanho do documento como definido em [97]
t_{17}	$\frac{1}{(1-slope)\times pivot+slope\times \#of uniqueterms}$	Normalização pivoteada baseada nos termos distintos de um documento [97], onde *pivot* é a média do número de termos distintos de um documento em toda a coleção
t_{18}	$\frac{1}{(k_1\times(1-b)+b\times\frac{dl}{avgdl})+tf}$	Fator de normalização presente na função de ordenação do sistema Okapi BM25, como definido em [90, 91]
t_{19}	$\frac{(k_3+1)\times qtf}{k_3+qtf}$	Fator relativo à consulta presente na função de ordenação do sistema Okapi BM25, como definido em [90, 91]
t_{20}	$0.5 + \frac{0.5+qtf}{maxqtf}$	Freqüência do termo na consulta normalizada, como definido em [8, 94]

2.5 Realimentação de Relevância

No processo de recuperação de imagens, um ponto importante é a forma de representação das propriedades visuais/textuais da imagem. Dessa maneira, *descritores* [23, 87, 101] são utilizados para a codificação dos *vetores de características* e comparação das imagens. Com isso, a comparação entre duas imagens é realizada utilizando-se métricas de similaridade (funções de distância) aplicadas a esses vetores e métricas de similaridade textual aplicadas às anotações das imagens.

Assim, uma maneira comum de se realizar uma consulta em um sistema de recuperação de imagens consiste na definição de um padrão de consulta pelo usuário [54], uma imagem por exemplo. O sistema deve retornar as imagens mais similares à consulta, segundo os descritores de conteúdo visual utilizados. Porém, o conceito de similaridade é subjetivo. Por isso, é necessário prover um meio para que o usuário possa exprimir qual a sua necessidade. Uma forma de realizar essa tarefa consiste no ajuste de parâmetros dos descritores utilizados, como a atribuição de pesos para cada elemento do vetor de características. Entretanto, isso exige que o usuário conheça detalhes do processo de descrição

de baixo nível das imagens, por exemplo o valor semântico relacionado aos elementos do vetor de características.

Com o objetivo de reduzir este problemas, técnicas de realimentação de relevância [15, 18,29,30,35,49,56,65,74,88,93,107,108,130] têm sido propostas. . Em geral, estas técnicas têm obtido sucesso no aumento da qualidade e precisão dos resultados de recuperação. Esse mecanismo tem por objetivo possibilitar que o usuário expresse sua necessidade na especificação de uma consulta, sem recorrer a propriedades de baixo nível utilizadas na representação de imagens. Para isso, o usuário apenas precisa indicar quais imagens considera relevantes ou irrelevantes dentre um conjunto retornado pelo sistema de busca. A cada iteração, o algoritmo de realimentação de relevância tenta identificar quais propriedades visuais melhor definem as imagens relevantes, a partir das informações fornecidas pelo usuário. Dessa forma, a consulta é reformulada automaticamente e submetida novamente ao sistema. Após um determinado número de iterações, o sistema retorna as imagens mais similares à imagem de consulta.

Um estudo amplo sobre técnicas de realimentação de relevância pode ser encontrado em [43], onde são apresentadas informações detalhadas sobre as características de mecanismos de realimentação de relevância para recuperação de imagens por conteúdo, modelagem desses sistemas, métodos de seleção e aprendizado usualmente utilizados em trabalhos correlatos. Neste texto será dada enfase à descrição de trabalhos correlatos em realimentação de relevância empregada à recuperação multimodal de imagens como apresentado na Seção 2.6.

2.6 Métodos de Recuperação de Imagens com Realimentação de Relevância

A recuperação de imagens com realimentação de relevância é amplamente estudada e tem obtido bastante êxito, quando comparada com técnicas de recuperação de imagens tradicionais. Em [43] foram estudados e comparados diversas propostas de realimentação de relevância para recuperação de imagens por conteúdo. A seguir, serão caracterizados alguns trabalhos que utilizam realimentação de relevância para recuperação de imagens a partir de evidências visuais e textuais.

O trabalho apresentado em [55] utiliza *pseudo-relevance feedback* em experimentos de recuperação multimodal de imagens. Assim, dada uma coleção de imagens com textos associados, inicialmente as imagens são ordenadas com base em uma técnica de modelagem de linguagem (tf-idf, okapi ou *kl-divergence*) e posteriormente é calculada a similaridade entre as imagens de consulta e as imagens melhor ranqueadas. As similaridades textuais e visuais são então combinadas de maneira ponderada e o resultado é re-ordenado.

Em [108], resultados de busca multimodais são construídos por meio da combinação linear de resultados em máquinas de busca textual e visual e *pseudo-relevance feedback* inter-modalidade. A recuperação textual é realizada utilizando o XFIRM [96] e a recuperação visual é processada com o sistema FIRE [29]. Na recuperação visual os documentos associados às k imagens mais similares são utilizados como *pseudo-relevance feedback*. Os termos mais frequentes nesses documentos são utilizados para contrução de uma nova consulta textual que é então processada também com o XFIRM. Os valores de relevância das imagens das duas listas ordenadas são combinados para construção do resultado final. Experimentos mostram melhores resultados para a busca combinada frente aos resultados das buscas independentes nas duas modalidades.

O trabalho em [88] apresenta duas técnicas de recuperação multimodal com realimentação de relevância e expansão de consulta para reajuste de pesos intra e inter-modalidade e atualização dos vetores de característica da consulta. A primeira abordagem de recuperação realiza busca sequencial entre as modalidades. Nesta técnica a realimentação de relevância é utilizada para refinar os resultados obtidos na recuperação puramente textual. Em seguida, sobre a 1000 imagens mais relevantes é executada uma busca baseada nas características visuais e o resultado também é refinado com realimentação de relevância. A segunda abordagem realiza recuperação concorrente nas duas modalidades e refina os resultados usando realimentação de relevância. Nas duas abordagens a cada iteração os pesos intra e inter-modalidade são re-definidos bem como os vetores de característica das consultas em cada uma das modalidades.

Em [15] são propostas duas estratégias para combinação e similaridades entre objetos multimodais. A primeira estratégia, chamada *Complementary Feedback (CF)*, utiliza como algoritmo de realimentação de relevância uma variação do método proposto em [124]. Nesta estratégia, por exemplo, dada uma consulta multimodal M, composta de um componente visual V e um componente textual T, os textos associados aos N melhores resultados são utilizados para compor a parte T' de uma nova consulta textual. Essa nova consulta textual é contruída por meio da combinação linear entre T e T'. Alternativamente, este processo pode ser iniciado por uma consulta multimodal e uma nova consulta visual seria criada a partir do componente visual do melhores N resultados. A segunda estratégia, chamada *Transmedia Document Reranking (TR)*, utiliza a criação de uma representação visual para uma consulta textual ou vice-versa. Assim, dada uma consulta textual, o conjunto das N melhores imagens do resultado é utilizado como representação visual (RV) da consulta textual. A partir deste ponto, o valor da distância entre RV e uma determinada imagem da base I é dada pelo somatório das distâncias entre a I e cada imagem componete de RV. Para gerar a ordenação final, as distâncias geradas para cada uma das novas representações são combindas por meio de uma soma ponderada. Os resultados obtidos com CF foram superiores ao TR e também aos melhores dentre todas

as submissões para a *ImageCLEFphoto 2007 photographic retrieval track* na categoria de realimentação de relevância. Além disso, o CF alcançou eficácia (MAP) apenas 0,22% menor do que a melhor submissão dentre todas as categorias.

O trabalho apresentado em [30] propõe um mecanismo de recuperação baseado na combinação de classificadores que geram resultados baseados em realimentação de relevância com exemplos positivos e negativos. Além disso, apresenta uma abordagem para determinação de pesos que são utilizados para cálculo de distância entre as imagens com base na distância L1. Esta abordagem basicamente busca descobrir pesos de forma que a distância entre imagens relevantes seja minimizada e a distância entre imagens relevantes e irrelevantes seja maximizada. Experimentos mostram que a utilização de realimentação de relevância para combinação de classificadores e o aprendizado de funções de distância ponderadas resultam em maior eficácia do que métodos tradicionais como algoritmo *Rocchio*.

A Tabela 2.5 resume as principais características dos trabalhos correlatos em recuperação de imagens com realimentação de relevância. As características exibidas são: a abordagem utilizada para o aprendizado; o critério de seleção; a forma como o problema é abordado; o tipo de *feedback* fornecido pelo usuário; e o tipo de informação provido. A marcação MP indica que as imagens mais positivas, ou seja, as mais similares à consulta são aquelas selecionadas para serem exibidas ao usuário. A marcação N/D significa que o item não foi definido. Na última linha da tabela são mostradas as características do projeto proposto neste trabalho. Essas características serão descritas em detalhes nos capítulos 4 e 5.

Tabela 2.5: Resumo das características dos trabalhos correlatos em realimentação de relevância.

Trabalhos	**Aprendizado**	**Seleção**	**Problema**	***Feedback***	**Informação**
Rui et al. [93]	Atualização de pesos	MP	Ordenação	Positivo/Negativo com grau de relevância	Intra-consulta
Cox et al. [18–20]	Bayesiano	MP	*Target Search*	Positivo/Negativo	Intra-consulta
Duan et al. [35]	Bayesiano dependente	MP	*Target Search*	Positivo/Negativo	Intra-consulta
Li et al. [74]	Atualização de pesos/ *query vector movement*	MP	Ordenação	Positivo/Negativo	Intra-consulta
Hong et al. [56]	SVM	N/D	Classificação	Positivo/Negativo	Intra-consulta
Tong et al. [107]	SVM	MI	Classificação	Positivo/Negativo	Intra-consulta
Gondra et al. [49]	SVM	MP	Classificação	Positivo	Inter-consulta
S. C. H. Hoi [55]	N/D	MP	Ordenação	Positivo	Intra-consulta
M. Torjmen et al [108]	N/D	MP	Ordenação	Positivo	Intra-consulta
M. M. Rahman [88]	Atualização de Pesos	MP	Ordenação	Positivo/Negativo	Intra-consulta
E. Clinchant et al. [15]	Redefinição de consulta	MP	Ordenação	Positivo	Intra-consulta
T. Deselaers et al. [30]	Atualização de pesos	MP	Ordenação	Positivo Negativo	Intra-consulta
Projeto proposto	**Programação Genética**	**MP**	**Ordenação**	**Positivo**	**Intra-consulta**

A Tabela 2.6 resume as características dos experimentos realizados nos trabalhos citados. As características presentes na tabela são: número de imagens da base e como foram obtidas; as propriedades visuais (*features*) utilizadas; o número de imagens retornadas por iteração com o usuário; o número máximo de iterações realizadas; e o tipo de usuário. Em algumas posições dessa tabela aparece a expressão N/D. Essa marcação significa *não definido*. Por exemplo, em [35] não fica claro quais propriedades foram codificadas. É importante salientar que, apesar de importante, o tempo de execução das consultas não foi exibido em nenhum dos experimentos. Na última linha da tabela são mostradas as características que devem ser exploradas nos experimentos a serem realizados no projeto proposto neste trabalho.

Tabela 2.6: Resumo das características dos experimentos realizados nos trabalhos apresentados.

Trabalhos	**Base**	**Propriedades**	**Imagens retornadas**	**Num. iterações**	**Usuário**
Rui et al. [93]	286 Museum Educational Site Licensing Project / 70000 Corel	Cor, forma e textura	12-1000	3	Simulado
Cox et al. [20]	4522	Cor	4	55 e 75	Real
Duan et al. [35]	800/1000/ 1500 Corel	N/D	25/100	30	N/D
Li et al. [74]	10000 Corel	Cor e textura	50	10	N/D
Hong et al. [56]	17000 Corel	Cor e textura	20	N/D	N/D
Tong et al. [107]	602/1277/ 1920 Corel	Cor e textura	20-150	5	Simulado
Gondra et al. [49]	20000	N/D	20	N/D	N/D
S. C. H. Hoi [55]	20000 ImageCLEFphoto 2007	cor, forma, textura e texto	N/D	N/D	Simulado
M. Torjmen et al. [108]	20000 ImageCLEFphoto 2007	Sistemas FIRE [29] e XFIRM [96]	6 e 15	N/D	Simulado
M. M. Rahman [88]	20000 ImageCLEFphoto 2007	Cor, textura e texto	30	N/D	Simulado/Real
E. Clinchant et al. [15]	20000 ImageCLEFphoto 2007	Cor, textura e texto	15	N/D	Simulado
T. Deselaers et al. [30]	20000 ImageCLEFphoto 2007	Cor, textura e texto	20	5	Simulado
Projeto proposto	**ImageCLEFphoto 2008 e Universidade de Washington**	**Cor, textura e texto**	**20**	**10**	**Simulado**

Pode ser observado na Tabela 2.6 que alguns trabalhos utilizaram mais de um conjunto de imagens em seus experimentos. Por exemplo, em [35] foram utilizados três conjuntos de testes contendo imagens selecionadas aleatoriamente da galeria Corel. O primeiro conjunto é composto por 800 imagens divididas em 32 categorias. O segundo contém 100 categorias com 100 imagens em cada categoria. E o último contém 1500 imagens didividas em 15 categorias.

Também pode ser observado que o número de imagens retornadas variam em alguns trabalhos. Em [107], o número de imagens retornadas varia entre 20 e 150. Já em [35] foram feitos experimentos considerando-se 25 e 100 imagens retornadas. Um outro fato que merece ser comentado é a utilização de usuários reais nos experimentos realizados em [20], ao contrário dos demais trabalhos que utilizaram usuários simulados por computador.

Capítulo 3

RFCore - Arcabouço Para Manipulação de Dados Com Realimentação de Relevância

Este capítulo descreve o arcabouço de realimentação de relevância proposto para manipulação de objetos digitais. Sobre este arcabouço, a seção 3.1 apresenta um conjunto de características de implementação comuns a sistemas que utilizam técnicas de realimentação de relevância e os aspectos motivadores do desenvolvimento. Além disso, apresenta uma implementação de referência do arcabouço especificado, apresentando classes e interfaces que podem ser utilizadas como infra-estrutura padronizada para construção de sistemas que usam realimentação de relevância para manipulação de dados.

3.1 Motivação e Características comuns

Com o avanço das tecnologias de captura, armazenamento e processamento de dados, a quantidade de informação disponível é cada vez maior. Neste cenário, verifica-se a necessidade de técnicas eficazes para manipulação desses dados. Além disso, dado o *gap* semântico entre as representações de baixo nível utilizadas para representação de características e a subjetividade da interpretação dos usuários sobre um determinado dado, diversas pesquisas têm apontado a utilização de realimentação de relevância como técnica para obtenção de resultados mais precisos e adaptados às necessidades de diferentes usuários [15, 18, 29, 30, 35, 49, 65, 74, 88, 108, 130].

A realimentação de relevância tem sido utilizada com sucesso no desenvolvimento de sistemas em diversas áreas de pesquisa, como: recuperação da informação, recomendação, classificação e anotação de dados. Pesquisas que utilizam a realimentação de relevância aplicam diferentes técnicas, como redes semânticas [78], combinação linear [74, 93] e não

linear [83] de características, agrupamento de dados [65], inferência Bayesiana [18, 35], redes neurais [67], máquinas de vetores de suporte [49,56] e programação genética [34,44].

O desenvolvimento de projetos e experimentos precisa ser feito em curto espaço de tempo e por isso uma atividade que torna este processo mais rápido é a reutilização de código, possivelmente em diferentes linguagens de programação. Soma-se a isso a necessidade de execução de experimentos comparativos de maneira adequada e consequentemente a reutilização de dados com diferentes formatos e originários de diferentes fontes. Ainda no que diz respeito à pesquisa científica existe a necessidade de avaliação de diferentes técnicas, métricas e algoritmos. Especificamente, em se tratando das diversas etapas do processo de realimentação de relevância, diferentes abordagens podem ser utilizadas.

Abordagens que utilizam realimentação de relevância seguem um fluxo de execução tradicional padrão, que é mostrado no Algoritmo 2.

Algoritmo 2 Algoritmo básico de realimentação de relevância.

```
1 Indicação pelo usuário do(s) objeto(s) de entrada/consulta q
2 Mostre o conjunto inicial de objetos retornados
3 Enquanto o usuário não estiver satisfeito faça
4     Indicação pelo usuário da relevância dos objetos apresentados
5     Atualize o padrão de consulta Q
6     Reorganize os objetos da coleção
7     Selecione novo conjunto de objetos a serem exibidos
8     Mostre o novo conjunto de objetos ao usuário
9 fim enquanto
```

O usuário inicia o processo indicando um ou mais objetos como padrão de consulta. O sistema utiliza o padrão de consulta para gerar o conjunto inicial de objetos a serem exibidos para o usuário (linha 2). Enquanto o usuário não estiver satisfeito com o resultado, ele pode interagir com o sistema indicando a relevância dos objetos apresentados (linha 3). Com essa informação, o sistema atualiza o padrão de consulta (linha 5) e realiza uma nova organização da coleção – ordenação, agrupamento, etc. (linha 6). O sistema então seleciona o novo conjunto de objetos a serem exibidos ao usuário (linha 7). O usuário verifica o novo conjunto de resultados, caso ainda não esteja satisfeito, ele avalia a relevância dos objetos do resultado e inicia uma nova iteração de realimentação de relevância.

3.2 Arcabouço RFCore

Dadas as necessidades de pesquisa aprensentadas e considerando-se o estudo dos trabalhos citados e o algoritmo tradicional de realimentação de relevância (Algoritmo 2), foi

proposto o desenvolvimento do arcabouço *RFCore*. Trata-se de um arcabouço de realimentação de relevância que pode ser utilizado como infra-estrutura básica para o desenvolvimento e experimentação de sistemas e técnicas de manipulação de objetos digitais. Além disso, o *RFCore* visa permitir a reutilização de código e construção de plataformas para execução de experimentos de maneira dinâmica, bem como o desenvolvimento de experimentos comparativos de modo rápido e com alto grau de compatibilidade de resultados.

Na implementação do arcabouço foi construída uma máquina para realimentação de relevância (*RFEngine*) na qual o Algoritmo 2 foi mapeado para a arquitetura apresentada na Figura 3.1. Assim a *RFEngine* funciona como um elo de comunicação entre os diferentes módulos acoplados.

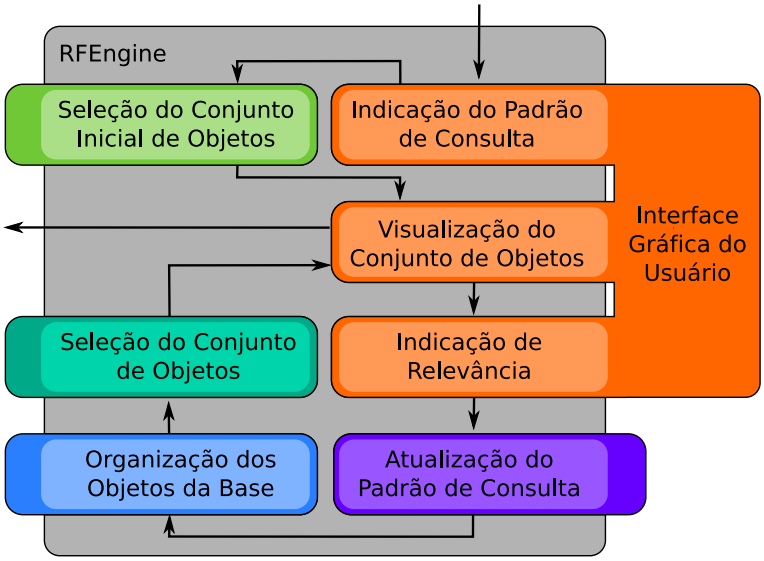


Figura 3.1: Arquitetura do arcabouço proposto.

O arcabouço desenvolvido utiliza um conceito simples de objetos digitais e cada etapa do processo foi mapeada para um método (procedimento) específico. Os métodos e tipos de objetos a serem utilizados para uma aplicação do arcabouço podem ser criados e estendidos de acordo com as necessidades de cada projeto.

O módulo de *interface gráfica* encapsula toda a comunicação da *engine* com o usuário. Por meio deste módulo, o usuário consegue definir o conjunto de dados de entrada (por exemplo, objetos da consulta, a serem classificados ou anotados), visualizar os resultados parciais e final e definir valores de relevância para objetos presentes nestes resultados.

O módulo de *seleção do conjunto inicial* de objetos utiliza a informação de entrada

fornecida pelo usuário para construir o conjunto inicial de resultados.

O módulo de *atualização do padrão de consulta* é responsável por coletar as informações de relevância fornecidas pelo usuário, e com base nela, atualizar o padrão de consulta, por exemplo, reformulando a consulta a partir da adição ou remoção de objetos ao padrão ou do ajuste parâmetros do método.

Após a atualização do padrão de consulta, esta informação pode ser utilizada para realizar uma nova organização dos objetos da coleção, utilizando, por exemplo, técnicas de aprendizado de máquina.

Dada a nova organização dos objetos da coleção, o módulo de seleção do conjunto de objetos é responsável por selecionar os objetos a serem exibidos ao usuário na próxima iteração. Caso o usuário esteja satisfeito ele pode encerrar o processo ou iniciar uma nova iteração julgando e submetendo o resultado apresentado.

Dada a diversidade de técnicas que podem ser utilizadas em cada etapa de um processo de realimentação de relevância e as características da *RFEngine*, foi desenvolvida uma implementação de referência com objetivo de permitir o acoplamento de diferentes implementações para cada um dos módulos de maneira dinâmica.

Desse modo, a *RFEngine* foi estendida e para cada um dos módulos foi definida uma interface de entrada e saída padronizada. O arcabouço foi desenvolvido na linguagem Java e utilizando a API de reflexão da biblioteca padrão. Essa estrutura padronizada e dinâmica permite o aproveitamente de grande parte de um sistema que utilize o arcabouço, dado que novas implementações de um determinado módulo podem ser utilizadas com todo o resto sendo mantido.

Como cada módulo pode utilizar algoritmos tão complexos quanto necessário para uma determinada aplicação, definiu-se que as interfaces seriam denominadas *"interfaces de estratégias"*. Cara uma das interfaces possui métodos que são invocados para a execução da estratégia. As interfaces definidas, suas assinaturas e respectivos módulos são apresentados na tabela 3.1.

Tabela 3.1: Lista de interfaces de programação definidas na implementação de referência.

Módulo	Interface	Método
Seleção do conjunto inicial	IInitialSetOfObjectsRetrievalStrategy	run
Atualização do padrão de consulta	IQueryPatternUpdatingStrategy	run
Organização dos objetos da base	ICollectionOrganizationStrategy	run
Seleção do conjunto de objetos	IObjectsSelectiontrategy	run
Interface com o usuário	IUserActionListener	indicationOfQueryPattern
Interface com o usuário	IUserActionListener	showSetOfObjects
Interface com o usuário	IUserActionListener	userIndicationOfObjectsRelevance

Capítulo 4

MMRF-GP - Realimentação de Relevância Multimodal Baseada em Programação Genética

Este capítulo apresenta o arcabouço para recuperação de objetos multimodais com realimentação de relevância baseada em programação genética (*MMRF-GP - Multimodal Relevance Feedback Based on Genetic Programming*) que foi desenvolvido sobre o arcabouço descrito no capítulo 3. A seção 4.1 apresenta uma visão geral do arcabouço MMRF-GP. A seção 4.1.1 descreve a construção do conjunto inicial de objetos a serem apresentados ao usuário. A seção 4.1.2 descreve a descoberta de novas funções de combinação de similaridades por meio da aplicação da programação genética. A seção 4.1.3 descreve como o conjunto de objetos é ordenado a partir das funções de combinação de similaridades descobertas e como os objetos a serem exibidos ao usuário são selecionados.

4.1 Realimentação de Relevância Baseada em Programação Genética

Dado um conceito simples de objetos digitais, pressupõe-se a possibilidade do cálculo de similaridades entre dois objetos a partir das diferentes modalidades de informação que os compõem. Um objeto pode possuir diferentes tipos de evidências para um mesmo tipo de modalidade e para cada evidência, valores de similaridade podem ser calculados a partir de diferentes abordagens. Por exemplo, supondo um objeto multimodal composto de uma imagem e uma passagem textual associada, tem-se então evidências textuais e visuais. Dentre as evidências visuais, podem-se citar, por exemplo, cores, texturas e formas. Para cada evidência, diferentes valores de similaridade podem ser calculados

por meio de diferentes medidas, por exemplo, diferentes descritores de conteúdo visual baseados em cor.

Partindo destes conceitos, o Algoritmo 3 apresenta uma visão geral do arcabouço de recuperação imagens por conteúdo proposto em [34, 44] e adapatado neste trabalho para recuperação de objetos multimodais. No início do processo, o usuário indica o conjunto Q de objetos de consulta (linha 1). Em seguida, um conjunto inicial de objetos recuperados é exibido. Esse conjunto é selecionado considerando a valor médio, segundo todas as fontes de similaridade utilizadas, entre o(s) objeto(s) de consulta e todos os objetos da coleção (linha 2). Com isso, é possível que o usuário inicie as iterações do processo de realimentação de relevância indicando os objetos relevantes contidos no conjunto inicial. Cada iteração consiste nos seguintes passos: indicação dos objetos relevantes pelo usuário (linha 4); atualização do padrão de consulta (linha 5); aprendizado da percepção de similaridade do usuário utilizando PG (linha 6); ordenação dos objetos da coleção (linha 7); e exibição dos objetos mais similares para o usuário (linha 8). No Algoritmo 3, as interações do usuário estão indicadas em itálico.

Algoritmo 3 O processo proposto de Realimentação de Relevância baseado em PG.

```
1 Indicação pelo usuário do(s) objeto(s) de consulta Q
2 Exiba o conjunto inicial de objetos
3 Enquanto o usuário não estiver satisfeito faça
4     Indicação dos objetos relevantes pelo usuário
5     Atualize o padrão de consulta Q
6     Utilize PG para encontar os melhores indivíduos (funções de combinação de
      similaridades)
7     Ordene os objetos da base
8     Exiba os L objetos mais similares ao padrão de consulta
9 fim
```

De maneira análoga ao formalismo apresentado na seção 2.2, pode-se definir a coleção de objetos $\mathcal{C} = \{o_1, o_2, \ldots, o_N\}$ e a função de similaridade composta $S = (\mathcal{MS}, \delta_{\mathcal{S}})$. A função S é utilizada para combinar os valores de similaridades obtidos a partir das diferentes medidas presentes no conjunto $\mathcal{MS} = \{MS_1, MS_2, \ldots, MS_K\}$. A similaridade entre dois objetos O_k e O_j, calculada por MS_i é representada por $ms_{iO_kO_j}$. O valor de similaridade combinado, obtido por meio de δ_S, pode ser então utilizado para ordenar os objetos da coleção.

Como definido no Algoritmo 3, o usuário informa ao sistema o conjunto de objetos de consulta $Q = \{q_1, q_2, \ldots, q_M\}$. Estes objetos são utilizados para iniciar o processo de recuperação e, à medida que as interações de realimentação de relevância acontecem o usuário informa quais objetos são relevantes para sua consulta. Estes objetos relevantes

são então adicionados ao padrão de consulta Q, configurando uma abordagem de múltiplos pontos de consulta [43].

4.1.1 Seleção do Conjunto Inicial de Objetos

Dados os M objetos de consulta fornecidos pelo usuário, o conjunto inicial de objetos é definido pela ordenação dos objetos da coleção a partir do maior valor médio de similaridade entre o objeto o_i da coleção e os objetos q_k do padrão de consulta. Inicialmente cada medida MS_j é utilizada para calcular a similaridade $ms_{jo_iq_k}$. Em seguida é calculado o valor médio das medidas de similaridade entre o_i e q_k, ou seja,

$$\delta_{media}(o_i, q_k) = \frac{\sum_{j=1}^{J} ms_{jo_iq_k}}{J} \tag{4.1}$$

sendo J o número de medidas de similaridade utilizadas.

Para se obter o valor de similaridade final entre o objeto o_i e o padrão de consulta Q, obtem-se o maior valor de similaridade entre o_i e cada objeto do padrão de consulta Q.

$$\delta_{final}(o_i, Q) = \max(\delta_{media}(o_i, q_k)), q_k \in Q \tag{4.2}$$

A partir da ordenação inicial, os L primeiros objetos são então exibidos para o usuário. O usuário inicia o processo de realimentação de relevância indicando quais objetos são relevantes para sua consulta. Os objetos marcados como relevantes são adicionados ao padrão de consulta, e o processo de aprendizado é iniciado como descrito na seção 4.1.2.

Na abordagem proposta neste trabalho, opcionalmente, apenas um subconjunto de $\mathcal{MS}$ poderia ser utilizado no processo de ordenação do conjunto inicial, delimitando as evidências ou modalidades a serem utilizadas.

4.1.2 Busca de Funções de Combinação de Similaridades

Como definido na proposta do arcabouço, o aprendizado consiste na descoberta de funções de combinação de similaridades que melhor se adequem às necessidades do usuário. Para tanto, utilizando a informação da realimentação de relevância, a técnica de programação genética é empregada na busca de novas funções de combinação. Partindo dos conceitos apresentados na seção 2.4, precisamos definir a representação das funções de similaridade e como estas são avaliadas. O processo evolutivo do MMRF-GP foi desenvolvido utilizando a biblioteca de programação genética *Java Genetic Algorithms Programming* (JGAP) [81].

Representação dos Indivíduos

Como proposto em [25, 34, 44], os indíviduos da programação genética representam as funções de combinação de similaridades S e são codificados numa estrutura de árvore. Assim, os nós folhas de uma árvore são compostos de valores de similaridades $ms_{iO_jO_k}$ ou valores constantes e os nós internos são compostos por operadores aritméticos, como +, *, / e $\sqrt{\ }$. Para a função $f(N, df) = \log \frac{N-df+0,5}{df+0,5} \times tf_{d1}$, o respectivo indivíduo é mostrado na Figura 4.1.

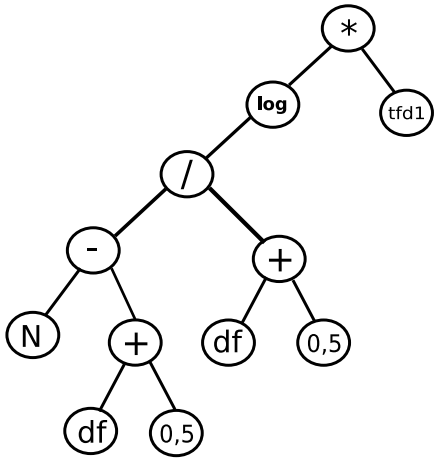


Figura 4.1: Exemplo de indivíduo. A função f é um dos componentes da medida Okapi (seção 2.1).

Função de Adequação

A função de avaliação de indivíduos é utilizada para definir os indivíduos que melhor se adequam às necessidades do usuário. Assim, a função de avaliação fornece uma medida de qualidade (*fitness*) para as funções descobertas. Essa qualidade baseia-se na capacidade de uma dada função melhor ordenar a coleção de objetos, ou seja, posicionar os objetos relevantes nas primeiras posições do *ranking*. Dadas as restrições de tempo e espaço, a avaliação das funções não é realizada sobre toda a coleção de objetos, mas sim sobre um conjunto de treinamento composto como na Definição 5.

Definição 5. *O conjunto de treinamento é dado pela tupla $\mathcal{T} = (T, r)$, onde T é um conjunto de N_t imagens distintas e r é uma função que indica a relevância definida pelo usuário para cada objeto do conjunto de treinamento. A função r devolve valor 1 caso o objeto tenha sido marcado como relevante ou 0 caso contrário.*

O tamanho do conjunto de treinamento precisa ser definido com cautela para evitar que o tempo do processo de aprendizado atrapalhe a experiência do usuário na utilização do sistema. Ao mesmo tempo este conjunto deve possuir informação suficiente sobre as necessidades do usuário e as características da coleção. Sendo M o número de objetos marcados como relevantes, L o número de objetos exibidos ao usuário a cada iteração e N_t o tamanho do conjunto de treinamento, na abordagem proposta os objetos que compõem o conjunto de treinamento são escolhidos da seguintes maneira:

Caso $M < L$, $\mathcal{T}$ é composto por:

- M objetos marcadados como relevantes;
- $L - M$ objetos exibidos para o usuário;
- $N_t - L$ objetos escolhidos aleatoriamente da coleção.

Caso contrário, $\mathcal{T}$ é composto por:

- L objetos escolhidos aleatoriamente dentre os relevantes;
- $N_t - L$ objetos escolhidos aleatoriamente na coleção.

É valido citar que a construção do conjunto de treinamento deste modo permite que as funções sejam avaliadas sobre um conjunto de dados que incluem características tanto das necessidades do usuário (objetos marcados como relevantes e objetos não marcados), quanto da base (objetos selecionados aleatoriamente).

Com o conjunto de treinamento criado, o valor de adequação de uma função pode ser calculado com base na qualidade da ordenação dos objetos do conjunto de treinamento pela função, considerando um objeto de consulta. Este processo é realizado como descrito a seguir.

Dada uma função de similaridade δ_i (indivíduo que está sendo avaliado), para cada imagem do padrão de consulta, é gerada uma ordenação do conjunto de treinamento. Para cada uma dessas ordenações, que definem listas ordenadas $rk_{j\delta_i}$, é calculado um valor de qualidade com base nos L primeiros objetos. Este valor é calculado de acordo com uma função fq que indica a qualidade da ordenação com base no posicionamento dos objetos relevantes. Tanto neste trabalho quanto em [43], esta função de qualidade é dada por

$$fq(rk_{j\delta_i}) = \lambda \times \sum_{l=1}^{L} r(rk_{j\delta_i}[l]) \times g(l) \tag{4.3}$$

onde λ é uma constante, r é a função que indica a relevância de um objeto, $rk_{j\delta_i}[l]$ é o l-ésimo objeto da lista ordenada $rk_{j\delta_i}$ e g uma função descrescente.

A função $fq(rk_{j\delta_i})$ define a qualidade de uma ordenação de acordo com as posições dos objetos que foram considerados relevantes pelo usuário. Quanto mais próximos os objetos relevantes estiverem do topo, maior a qualidade da ordenação. A contribuição de um objeto para a qualidade da ordenação diminui à medida que estes são posicionados mais próximos ao final da lista ordenada.

O valor adequação F do indivíduo δ_i é dado pela média aritmética entre valores de qualidade de cada uma das listas geradas com base em cada objeto do padrão de consulta.

$$F(f_{q_1\delta_i}, f_{q_2\delta_i}, \ldots, f_{q_M\delta_i}) = \frac{\sum_{j=1}^{M} f_{q_j\delta_i}}{M} \tag{4.4}$$

4.1.3 Ordenação da Base

Ao final do processo de aprendizado (evolução dos indivíduos) é gerado um conjunto ordenado de soluções (funções de combinação). A partir dos valores de adequação é possível definir a função que melhor se ajusta às necessidades de usuário. Aqui vale ressaltar que dadas as configurações do processo evolutivo e do padrão de consulta, é possível que vários indivíduos alcancem valores de adequação altos e sejam candidatos à solução para ordenação da coleção.

Dada a possibilidade de existência de vários indivíduos com bons valores de adequação, neste arcabouço é proposta a utilização de mais de um indivíduo para a ordenação da coleção por meio de um processo de votação. A seção 4.1.3 descreve como os indivíduos votantes são selecionados e a seção 4.1.3 descreve como a votaçao é realizada para ordenação dos objetos da coleção é realizada por meio da votação.

Seleção de indivíduos

Sendo δ_{melhor} o indivíduo de melhor adequação da iteração corrente, o conjunto de indivíduos votantes selecionados é dado por

$$V = \{\delta_i | \frac{F_{\delta_i}}{F_{\delta_{melhor}}} \geq \alpha\} \tag{4.5}$$

com $\alpha \in [0, 1]$ constante. No arcabouço proposto, opcionalmente, o número de indivíduos votantes pode ser limitado por um número máximo.

Votação

Os objetos a serem exibidos para o usuário são selecionados a partir de um conjunto ordenado gerado por um processo de votação que é dado como segue. A partir do conjunto V de funções de combinação, cada uma das funções é utilizada para ordenar toda a coleção

de objetos, gerando v listas ordenadas. Cada lista gera um voto para cada objeto da coleção. O voto atribuído a um determinado objeto é inversamente proporcional à sua posição na lista. Desse modo, para cada objeto é calculada a soma dos votos atribuídos por cada um dos v indivíduos. Os objetos da coleção são então ordenados de acordo com a soma de votos de cada um. A partir desta ordenação, o L primeiros objetos são selecionados e exibidos ao usuário. Este pode finalizar o processo de recuperação ou novamente julgar a relevância dos objetos do resultado, iniciando uma nova iteração de realimentação de relevância.

A ordenação da coleção que é feita por meio de cada indivíduo é realizada utilizando-se a similaridade de cada objeto da coleção em relação ao padrão de consulta, do mesmo modo que foi realizado para seleção do conjunto inicial (seção 4.1.1). Entretanto, neste caso, a medida de similaridade é dada pela função do indivíduo votante. Assim, a similaridade entre o objeto o_i da coleção e o padrão de consulta Q é dada por $\delta_i = max(\delta_i(o_i, q_k))$, $q_k \in Q$.

Neste arcabouço, a ordenação da coleção é feita com base na similaridade em relação à imagem fornecida pelo usuário e em relação a todas as imagens marcadas como relevantes. Assim, considera-se a possibilidade de que imagens relevantes estejam agrupadas no espaço de características não apenas em torno da imagem de consulta, mas também em torno das imagens marcadas como relevantes.

Capítulo 5

Aspectos de Validação

Este capítulo apresenta os resultados da utilização do arcabouço apresentado no capítulo 4 em experimentos de realimentação de relevância para combinação de evidências visuais e textuais para recuperação de imagens. A seção 5.1 apresenta o projeto dos experimentos realizados e a seção 5.2 apresenta configurações, comparativos e discussão sobre os resultados experimentais.

5.1 Projeto dos Experimentos

Para validação do arcabouço proposto no Capítulo 4 foram realizados experimentos de recuperação de imagens utilizando-se evidências textuais e visuais. Nesta implementação buscou-se, por meio das iterações de realimentação de relevâncias, descobrir novas funções de combinação para similaridades baseadas em anotações textuais e conteúdo visual de imagens. A seção 5.1.1 descreve as coleções de imagens utilizadas nos experimentos. A seção 5.1.2 apresenta os descritores de conteúdo visual empregados. Na seção 5.1.3 são descritas medidas de similariades textual. Finalmente, a seção 5.1.4 descreva as medidas de eficácia utilizadas para avaliação e comparativos dos experimentos.

5.1.1 Bases de Imagens

Nos experimentos executados foram utilizadas duas coleções de imagens. A primeira foi a base desevolvida pela Universidade de Washington, daqui para a frente referida como UW[8]. A segunda coleção de validação foi a ImageCLEFphoto 2008, originalmente utilizada na *ImageCLEF Photographic Retrieval Task 2008* [1]. Detalhes de cada uma das coleções são apresentados a seguir.

[8]Disponível em: `http://www.cs.washington.edu/research/imagedatabase/groundtruth/`

Coleção UW

Esta coleção foi desenvolvida na Universidade de Washington e contém 1109 imagens divididas em 20 categorias. Nesta coleção existe um alto grau de heterogeneidade visual e textual entre as imagens de uma mesma categoria. A menor categoria tem 22 e a maior 255 imagens. O tamanho médio é de 55 imagens por categoria. Para cada imagem da base existe um conjunto de palavras-chave associadas. Esta coleção possui imagens de diferentes tamanhos e contém principalmente fotografias capturadas durante um período de férias, tomadas em diferentes locais, por exemplo, Austrália, Indonésia e Iran.

Nos experimentos realizados, foram aleatoriamente selecionadas 110 imagens de consulta dentre as existentes na coleção. A quantidade de consultas selecionadas em cada categoria foi proporcional ao número de imagens da categoria. Exemplos de imagens e anotações desta coleção podem ser vistos na Figura 5.1. Nos experimentos, esta coleção foi utilizada para validação da maior eficácia da recuperação multimodal frente às abordagens baseadas apenas na modalidade visual ou textual.

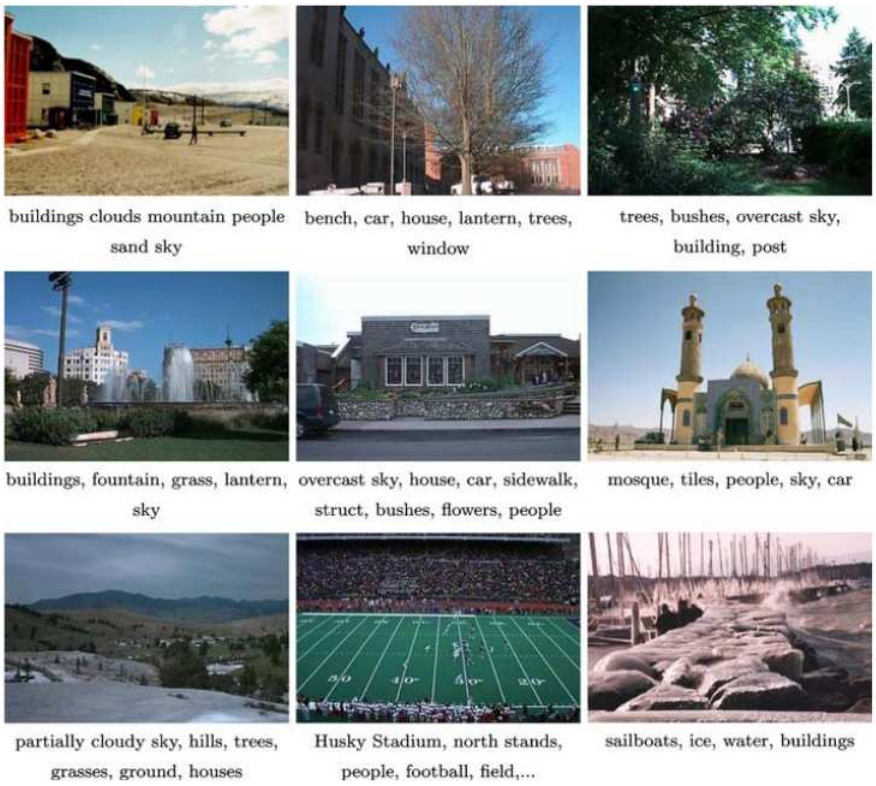


Figura 5.1: Exemplos de imagens da coleção UW com anotações. Fonte [31].

Coleção ImageCLEFphoto 2008

A coleção de imagens utilizada na *ImageCLEF Photographic Retrieval Task 2008* (ICphoto) foi desenvolvida a partir do *IAPR TC-12 Benchmark* criado pelo comitê técnico 12 da Associação Internacional de Reconhecimento de Padrões (*IAPR, International Association of Pattern Recognition*). Esta coleção é composta de 20.000 imagens coloridas tomadas em diversas partes do mundo. Cada imagem da coleção é acompanhada de um documento que contém os seguintes campos: identificador único, título, descrição em texto livre do conteúdo semântico e visual, notas adicionais, fornecedor, local e data.

As consultas utilizadas nos experimentos foram as mesmas especificadas pela ICphoto. Ao todo são 60 consultas (2007) e 39 consultas (2008, subconjunto das de 2007) que representam atividades de busca comuns para uma coleção genérica de imagens. Detalhes sobre como as consultas foram definidas podem ser encontrados em [1, 50]. Cada uma das consultas é composta de um título (um sequência pequena de palavras descrevendo a busca) e um conjunto de 3 imagens consideradas relevantes para a busca. Exemplos de uma imagem e respectiva anotação desta coleção são apresentados na Figura 5.2. Nos experimentos, esta coleção foi utilizada para validação da maior eficácia da recuperação multimodal frente às abordagens baseadas apenas na modalidade visual ou textual e para comparativo com os resultados das submissões à ICphoto, como apresentado na seção 5.2. Os grupos participantes do ImageCLEF 2008 utilizaram técnicas como *Local Content Analysis*, pseudo-realimentação de relevância, expansão de consultas, Wordnet, algoritmos de agrupamento, etc [1].

```
<DOC>
 <DOCNO>annotations/16/16392.eng</DOCNO>
 <TITLE>Sunset in Salvador</TITLE>
 <DESCRIPTION>a sandy beach at the sea with dark rocks behind
 it; the setting sun in an orange sky in the background;
 </DESCRIPTION>
 <NOTES></NOTES>
 <LOCATION>Salvador, Brazil</LOCATION>
 <DATE>10 October 2004</DATE>
 <IMAGE>images/16/16392.jpg</IMAGE>
 <THUMBNAIL>thumbnails/16/16392.jpg</THUMBNAIL>
</DOC>
```

Figura 5.2: Exemplo de imagem da coleção ImageCLEF 2008 com anotação. Fonte [1].

5.1.2 Descritores de Imagens

Os descritores de imagens utilizados nos experimentos são apresentados na Tabela 5.1. Detalhes sobre as funções de distância utilizadas por cada um deles podem ser encontrados

em [86].

Descritor	Tipo de Evidência	Função de distância
GCH [101]	Cor	L1
BIC [100]	Cor	dLog
ACC [58]	Cor	L1
JAC [116]	Cor	L1
CCOM [68]	Textura	Somatório das diferenças entre matrizes de co-ocorrência
LAS [104]	Textura	L1
QCCH [57]	Textura	L1
HTD [118]	Textura	Diferença entre componentes de média e desvio padrão

Tabela 5.1: Descritores de imagem usados nos experimentos.

5.1.3 Métricas de Similaridade Textual

Além das métricas de similaridade textual descritas na seção 2.1, outras métricas foram utilizadas nos experimentos, dentre elas as medidas tfidf-sum [95], dice e jackard [73].

Dada a consulta c composta de n termos t_i, a medida tfidf-sum é dada pelo somatório do valor de tf-idf para cada termo da consulta com relação ao documento d da coleção. Esta medida define a similaridade do documento com relação à consulta simplesmente com base na relevância dos termos no documento e na coleção inteira.

$$tdidf - sum(c, d) = \sum_{i}^{n} tf(t_i, d) \times idf(t_i) \tag{5.1}$$

A medida dice define a similaridade de um documento com relação a uma consulta de acordo com a quantidade de termos que os dois documentos têm em comum em proporção à quantidade total de termos existentes nos dois documentos, é dada por

$$dice(c, d) = \frac{2 \times |c \cap d|}{|c| + |d|} \tag{5.2}$$

A medida jackard define a relevância de uma consulta com relação a um documento de acordo com a quantidade de termos que dois documentos têm em comum em proporção à quantidade de termos incomuns entre eles.

$$jackard(c, d) = \frac{|c \cap d|}{|c| + |d| - 2 \times |c \cap d|} \tag{5.3}$$

5.1.4 Medidas de Avaliação

A avaliação dos experimentos de recuperação de imagens foi realizada com várias medidas de eficácia, sendo elas: precisão X revocação, P20, MAP (*Mean Average Precision*), GMAP (*geometric MAP*) e bpref (*binary preference*). Estas medidas são amplamente utilizadas para avaliação de sistemas de recuperação de imagens e são as medidas utilizadas para avaliações de submissões ao *ImageCLEF*.

A primeira medida utilizada foi a P20, uma instância da medida P@N que indica a precisão para N primeiras imagens retornadas e é dada por

$$pN = \frac{rel(N)}{N} \tag{5.4}$$

sendo $rel(N)$ o número de imagens relevantes nas primeiras N posições.

A segunda medida é a curva precisão x revocação na qual valores de precisão são calculados para um conjunto de valores de revocação. Estes valores variam no intervalo [0,1]. Comumente são calculados valores de precisão para valores de revocação em intervalos de tamanho 0,1. O valor de precisão é dado como na Equação 5.4 e o valor de revocação(r) é dado por

$$r(N) = \frac{rel(N)}{nrel} \tag{5.5}$$

com $rel(N)$ sendo o número de imagens relevantes dentre as N primeiras retornadas e $nrel$ sendo o número de imagens relevantes existentes na coleção. Consequentemente, o valor de precisão para $r(N)$ é dado por pN.

A terceira medida utilizada foi a MAP. Para uma dada consulta o valor de precisão média (AP) é dado pela média dos valores de $P@N$ para todos os documentos relevantes.

$$AP(q) = \frac{\sum_{n=1}^{N}(P@n \times r(n))}{nrel} \tag{5.6}$$

com q uma dada consulta, N o número de documentos retornados e $r(n)$ uma função que retorna 1 caso a n-ésima imagem seja relevante ou 0 caso contrário. O valor de MAP é dado pela média dos valores de AP para um conjunto de consultas distintas. Assim, sendo $Q = q_1, q_2, \ldots, q_K$ um conjunto de K consultas, o valor de MAP é dado por

$$MAP(Q) = \frac{\sum_{k=1}^{K} AP(q_k)}{K} \tag{5.7}$$

A quarta medida utilizada foi a GMAP, originalmente desenvolvida para enfatizar bons resultados em consultas difíceis. Ao contrário do MAP que usa a média aritmética

dos valores de precisão média, a GMAP é calculada a partir da média geométrica. Por exemplo, para uma dada consulta A, o primeiro resultado obtem $AP = 0.02$ e segundo $AP = 0.04$, enquanto que para uma consulta B o primeiro resultado obtem $AP = 0.4$ e o segundo $AP = 0.38$. Para estes resultados, a média aritmética permanece a mesma, entretanto a média geométrica vai indicar um aumento de eficácia. O valor de GMAP é dado por

$$GMAP(Q) = \sqrt[K]{\prod_{k=1}^{K} AP(q_k)} \tag{5.8}$$

A última medida utilizada foi a bpref que é calculada com base na fração de imagens julgadas irrelevantes que aparecem no resultado à frente de imagens julgadas relevantes. O valor de bpref é dado por

$$bpref = \frac{1}{R} \sum_{r} (1 - \frac{nrel}{min(R, N)}) \tag{5.9}$$

sendo R o número de imagens julgadas relevantes, N o número de imagens julgadas irrelevantes, r um dado documento relevante retornado e $nrel$ a quantidade R de imagens irrelevantes posicionadas à frente de r.

5.2 Experimentos

Nesta seção serão apresentadas as configurações da instância do arcabouço MMRFGP (Capítulo 4) utilizadas nos experimentos de recuperação de imagens a partir de evidências textuais e visuais.

5.2.1 Parâmetros do Método

A Tabela 5.2 apresenta os parâmetros utilizados no método de realimentação de relevância e a Tabela 5.3 os parâmetros utilizados no processo evolutivo da programação genética. Os valores dos parâmetros foram escolhidos com base nos melhores valores apresentados em [44].

Os descritores de imagens utilizados como terminais para os indivíduos para a coleção UW foram: BIC, GCH, JAC, HTD, LAS e QCCH. Já para a coleção ICphoto, foram: ACC, BIC, GCH, JAC, CCOM, LAS e QCCH. Este descritores foram escolhidos de acordo com o estudo realizado em [86] em bases de imagens com características semelhantes às utilizadas nos experimentos aqui apresentados. Dada a grande quantidade de tempo necessária para extração dos vetores de características e cálculo de distâncias usando o

Realimentação de Relevâcia	
Parâmetro	Valor
Número de iterações	10
Número de imagens exibidas em cada iteração	20
Número máximo de indivíduos votantes	4

Tabela 5.2: Parâmetros de realimentação de relevância utilizados nos experimentos.

Programação Genética	
Parâmetro	Valor
Tamanho da população	60
Número de gerações	20
Probabilidade de crossover	0,8
Probabilidade de reprodução	0,0
Probabilidade de mutação	0,2
Profundidade mínima inicial dos indivíduos	2 [37]
Profundidade máxima inicial dos indivíduos	5
Limitante de seleção de indivíduos (α)	0,999
Tamanho do conjunto de treinamento	55
Constante da função de utilidade	2
Operadores	+, *, /, e $\sqrt{\ }$

Tabela 5.3: Parâmentros da programação genética utilizados nos experimentos.

descritor HTD, este foi então substituído nos experimentos da ICphoto pelo descritor CCOM que possui eficácia média similar de acordo com os estudos em [86].

Ainda nos terminais, para ambas as coleções, foram utilizadas todas as medidas de similaridade textual apresentadas nas seções 2.1 e 5.1.3 (*Bag-of-words*, Cosseno, Okapi, tfidf-sum, Dice e Jackard). Todos os valores de similaridade foram normalizados no intervalo [0,1] utilizando normalização gaussiana [93].

Na função de avaliação dos indivíduos a função decrescente g utilizada foi

$$g(x) = \log_{10}(\frac{1000}{x}) \tag{5.10}$$

5.2.2 Técnicas de Realimentação de Relevância Implementadas

Nos experimentos foram utilizadas as 5 técnicas de realimentação de relevância mostradas na Tabela 5.4. Na tabela 5.4, *mm* indica uso de informação multimodal (textual e visual), *txt* indica uso apenas de similaridade textual e *vis* indica uso apenas de similaridade visual. Por exemplo, na técnica mm-txt o conjunto incial de objetos é construído levando-se em

consideração apenas as similaridades textuais entre as anotações das imagens, simulando uma consulta textual, em seguida o processo de realimentação de relevância utiliza tanto a informação textual quanto a visual para composição dos resultados.

Tipo	**Conjunto Inicial**	**Realimentação de Relevância**
Recuperação Multimodal (mm-mm)	txt+vis	mm
Recuperação Multimodal com início textual (mm-txt)	txt	mm
Recuperação Multimodal com início visual (mm-vis)	vis	mm
Recuperação Textual (txt)	txt	txt
Recuperação Visual (vis)	vis	vis

Tabela 5.4: Tipos de execução utilizados nos experimentos.

Para simulação do *feedback* do usuário e avaliação de resultados na coleção UW, imagens pertencentes à mesma classe da imagem de consulta foram consideradas como relevantes. Para simulação do *feedback* do usuário e avaliação de resultados na coleção ICphoto, foram usados os jugamentos de relevância (*qrels*) divulgados junto à coleção (detalhes em [1, 50]). No *qrels*, para cada consulta, são listadas as imagens relevantes esperadas.

5.2.3 Resultados e Discussão

Nas próximas seções são apresentados os resultados de eficácia para as consultas executadas sobre as coleções UW e ICphoto. Todas as medidas de avaliação foram calculadas com o aplicativo trec_eval[9] que é largamente utilizado para avaliação de sistemas de recuperação da informação. O trec_eval é utilizado como ferramenta padrão para avaliação das submissões ao ImageCLEF.

Resultados Coleção UW

As Figuras 5.3 e 5.4, apresentam os resultados comparativos para as consultas realizadas sobre a coleção UW nas diversas modalidades apresentadas na Tabela 5.4. Pela Figura 5.3 observam-se desempenhos semelhantes entre as abordagens multimodais. Além disso, verifica-se que a melhor abordagem multimodal foi a mm-mm, que supera a abordagem puramente visual em aproximadamente 9% e a textual em 2,5%, considerando a medida MAP.

[9] `http://trec.nist.gov/trec_eval/index.html`

Apesar da proximidade entre valores de MAP das diferentes abordagens, por meio das curvas de precisão x revocação (Figura 5.4) é possível notar que a partir 20% de revocação as abordagens multimodais apresentam maior número de imagens relevantes do que a abordagem visual. Aqui vale ressaltar que, apesar da grande homogeneidade entre as imagens pertencentes à mesma categoria, existem em cada categoria imagens que, mesmo com alto grau de similaridade semântica, são consideradas *outliers* em se tratando de características puramente visuais. Além disso, como as imagens são anotadas com palavras-chave é possível que imagens visualmente distintas possuam características semânticas semelhantes ou que imagens visualmente semelhantes possuam características semânticas ligeiramente distintas, o que justifica a queda de desempenho da abordagem textual para valores de revocação superiores a 70%.

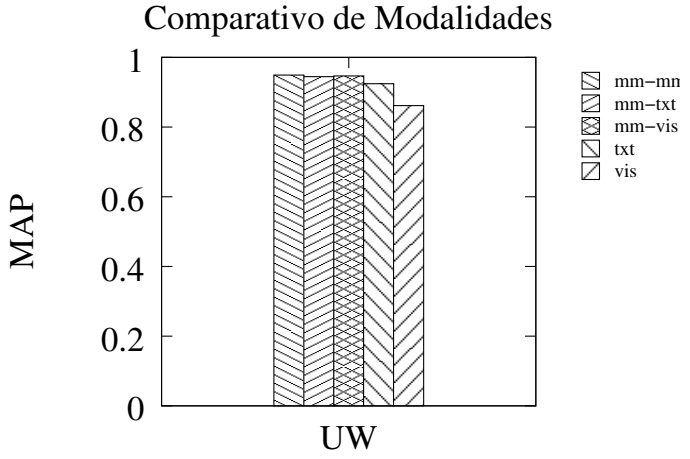


Figura 5.3: Comparativo entre as diferentes modalidades de recuperação para a coleção UW. Resultado médio para 110 consultas.

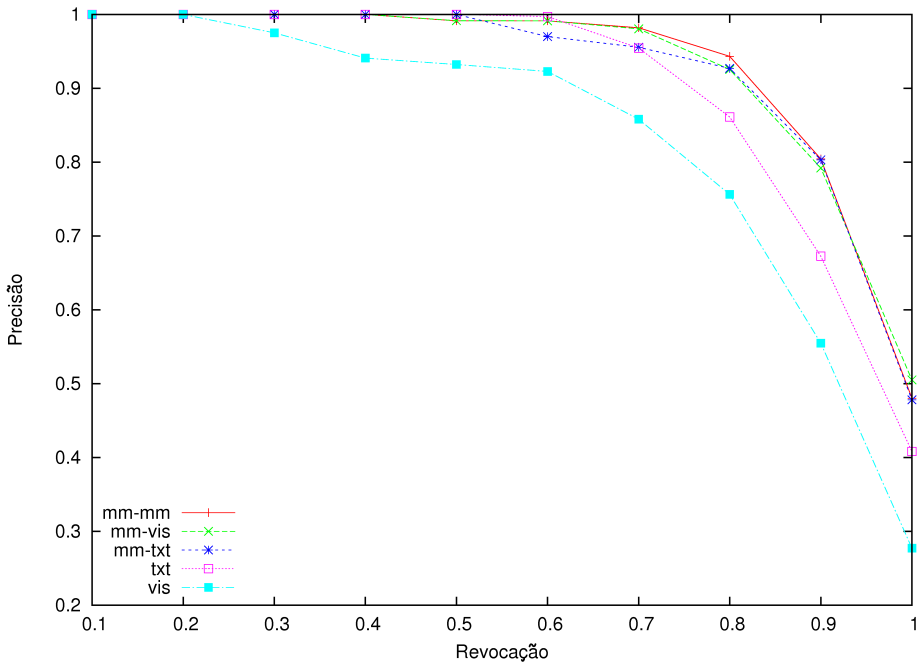


Figura 5.4: Comparativo de precisão x revocação entre as diferentes modalidades de recuperação para a coleção UW. Resultado médio para 110 consultas.

É importante notar também que a heterogeneidade visual e conceitual entre as imagens da classe faz com que sejam obtidos valores de precisão elevados. Isso é verificado dado que a média entre as 5 modalidades de execução para a medida P20 é de 99% com desvio padrão de 1%. Todas as execuções obtiveram valores de bpref igual a 100%.

Neste ponto, pode-se notar a importância da utilização de múltiplas fontes de evidência, visto que a abordagem multimodal consegue superar o desempenho das adordagens isoladas fazendo com que características visuais e conceituais sejam balanceadas e permitam a geração de resultados mais precisos.

Resultados Coleção ICphoto

As Figuras 5.5, 5.6, 5.7, 5.8, 5.9, 5.10 e 5.11 apresentam os resultados comparativos para as consultas realizadas sobre a coleção ICphoto nas diversas modalidades apresentadas na Tabela 5.4.

A figura 5.5 mostra um comparativo de eficácia entre as diferentes adordagens de execução para 39 consultas. Fica claro que as abordagens multimodais superam em 25% de MAP com relação à abordagem visual, o que em valor relativo equivale a um ganho de 110%. Frente a abordagem textual a multimodalidade alcança valor de MAP 22% maior, o que equivale a 85% de ganho. Vale ressaltar que apesar da abordagem mm-vis ter sido apenas 1% inferior à abordagem mm-mm, ela foi 3% mais eficaz do que a abordagem mm-txt, em se tratando de MAP absoluto, o que equivale a um ganho 6%.

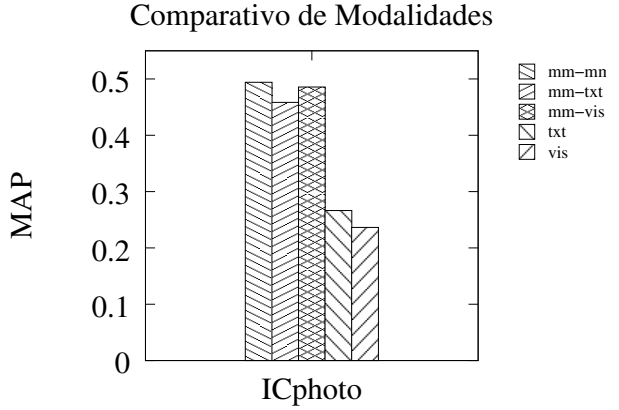


Figura 5.5: Comparativo entre os diferentes tipos de execução para 39 consultas na coleção ICphoto.

Os resultados apresentados na Figura 5.6 equivalem a execução de 60 consultas e possuem comportamento bastante alinhado aos resultados apresentados na Figura 5.5 cujos resultados foram gerados a partir das 39 consultas.

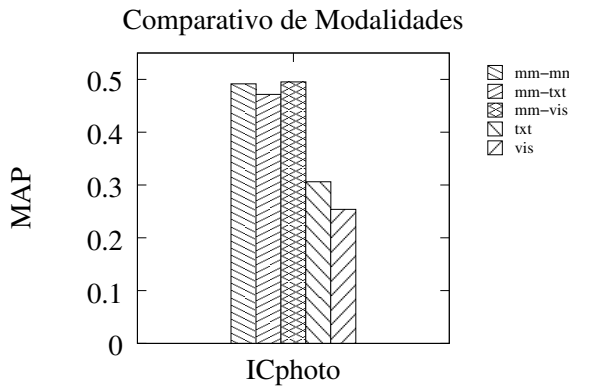


Figura 5.6: Comparativo entre os diferentes tipos de execução para 60 consultas na coleção ICphoto.

Um comparativo de MAP para as diferentes abordagens para os dois conjuntos de consultas é mostrado na Figura 5.7. Os resultados mostrados nas Figuras 5.5, 5.6 e 5.7 mostram claramente o comportamento similar entre as diferentes abordagens de execução e consequentemente a estabilidade do método frente a diferentes quantidades e conjuntos de consultas.

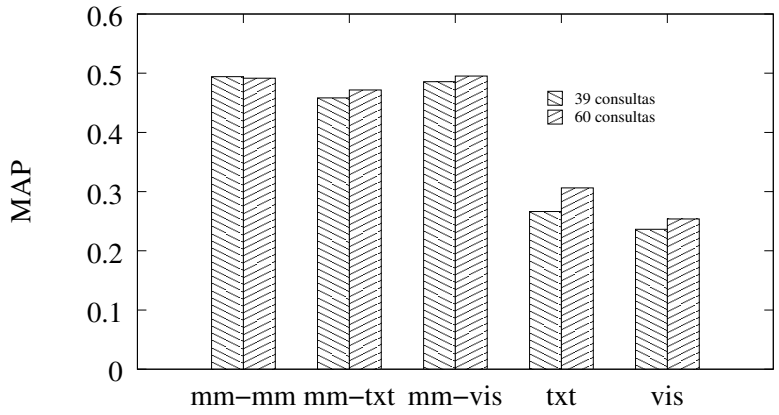


Figura 5.7: Comparativo entre as diferentes modalidades de recuperação para a coleção ICphoto para as 39 consultas do ImageCLEFphoto 2008 e 60 consultas do ImageCLEFphoto 2007.

A Figura 5.8 apresenta os resultados de precisão x revocação para as diferentes abordagens para o conjunto de 60 consultas. Nota-se que para valor de revocação inferiores a 10% a abordagem visual possui eficácia ligeiramente superior à abordagem textual, o que se inverte a partir de valores de revocação superiores a 20%. Isso pode ser justificado dado que para uma determinada consulta existe um conjunto de imagens com características visuais altamente semelhantes às da consulta, mas este conjunto é relativamente pequeno frente à quantidade de imagens relevantes esperadas para a consulta.

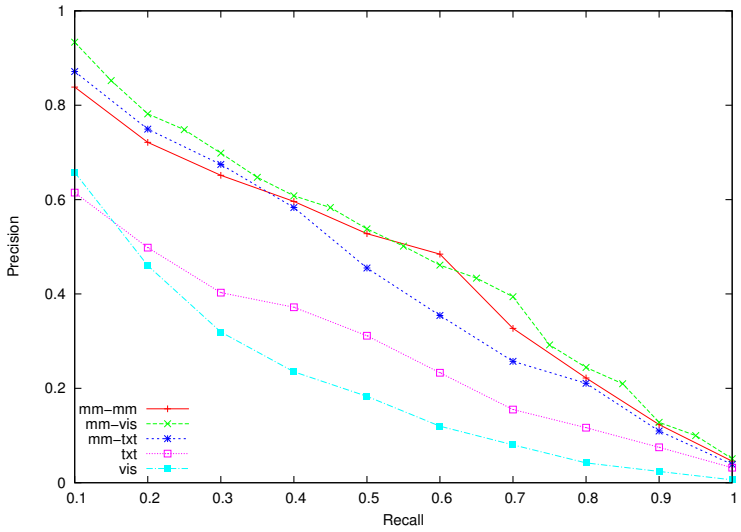


Figura 5.8: Comparativo de precisão x revocação entre diferentes modalidades de execução do MMRF-GP. Resultado médio para 60 consultas na coleção ICphoto.

A superioridade da abordagem textual frente a visual pode ser justificada dadas as características da coleção que possui 20.000 imagens. Para cada consulta desta coleção a quantidade média de imagens relevantes esperada é de aproximadamente 60, o que equivale a apenas 0,3% da coleção. Dada a variedade visual das imagens da coleção, é esperado que a maioria das imagens relevantes sejam aquelas cujos valores conceituais são os mais próximos da consulta, o que favorece a abordagem textual. Isso é indicado também pela melhor eficácia dos resultados textuais sobre os visuais dentre as submissões ao ImageCLEF 2006, 2007 e 2008 [1, 16, 50].

Pelos resultados apresentados nas Figuras 5.5, 5.6 e 5.8 é possível notar novamente a superioridade da abordagem multimodal frente as abordagens visual e textual, ratificando a capacidade de redução do bem conhecido *gap* semântico.

Comparação com outras técnicas multimodais

A partir deste ponto, serão apresentados resultados comparativos entre as melhores execuções submetidas à *ImageCLEF Photographic Retrieval Task 2008* [1] nos tipos de execução apenas textual, apenas visual e multimodal (textual e visual). Os valores apresentados foram calculados de maneira relativa indicando o ganho de uma abordagem sobre a outra.

A Figura 5.9 mostra os resultados comparativos de eficácia entre a melhor submissão textual e o melhor resultado textual obtido com o MMRF-GP. Assim, nota-se que em todas as medidas de eficácia, o MMRF-GP obteve desempenho inferior. Estima-se que a baixa eficácia para recuperação textual com o MMRF-GP deve-se ao fato de que apesar das anotações das imagens disponibilizarem diversos campos, como título, descrição, local, data, fornecedor, etc., apenas o campo descrição foi utilizado nos experimentos. Isso foi feito para evitar diferentes níveis de anotação textual entre as imagens da coleção, dado que todas as imagens da coleção possuem descrição textual. Assim, acredita-se que resultados mais eficazes podem ser alcançados com a utilização do maior número de informação disponível. Dessa forma, experimentos adicionais fazem-se necessários para verificação desta hipótese.

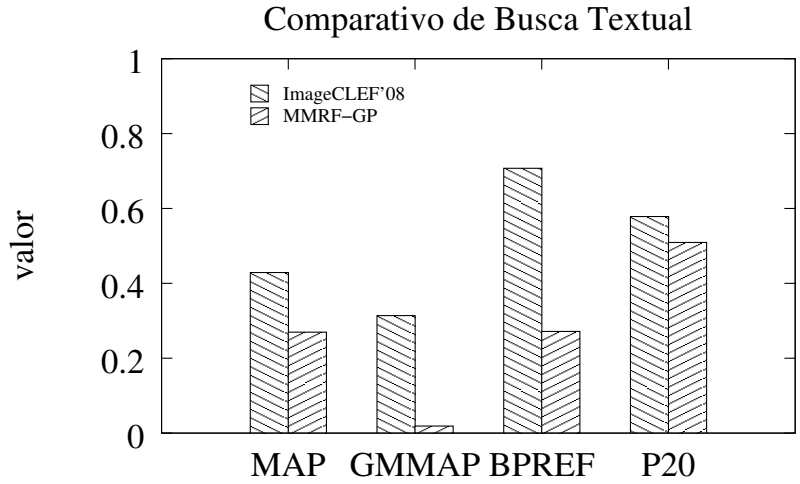


Figura 5.9: Comparativo entre o melhor resultado textual do ImageCLEFphoto 2008 e o melhor resultado textual do MMRF-GP. Resultado médio para 39 consultas.

A Figura 5.10 mostra os resultados comparativos de eficácia entre a melhor submissão visual e o melhor resultado da abordagem visual obtido com o MMRF-GP. O MMRF-GP obteve resultado superior em todas as medidas com exceção da bpref (discutido posteriormente). Os resultados apresentados na Figura 5.10 mostram que o MMRF-GP obteve ganho de 57%, 70% e 54% para MAP, GMAP e P20, respectivamente.

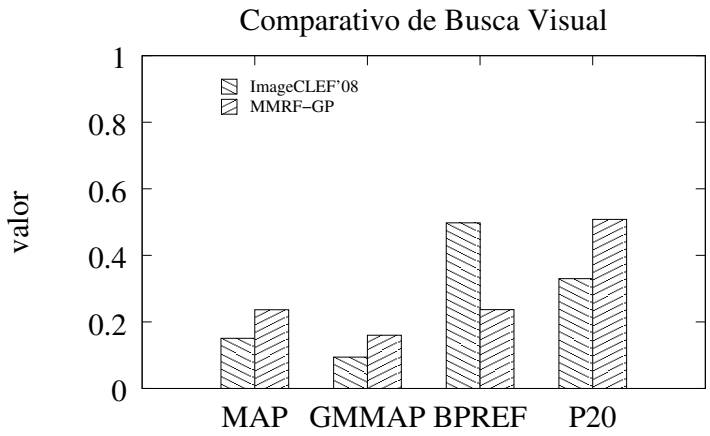


Figura 5.10: Comparativo entre o melhor resultado visual do ImageCLEFphoto 2008 e o melhor resultado visual do MMRF-GP. Resultado médio para 39 consultas.

A Figura 5.11 mostra os resultados comparativos de eficácia entre a melhor submissão multimodal e o melhor resultado multimodal obtido com o MMRF-GP. Assim como na abordagem puramente visual, o MMRF-GP obteve resultado superior em todas as medidas com exceção da bpref.

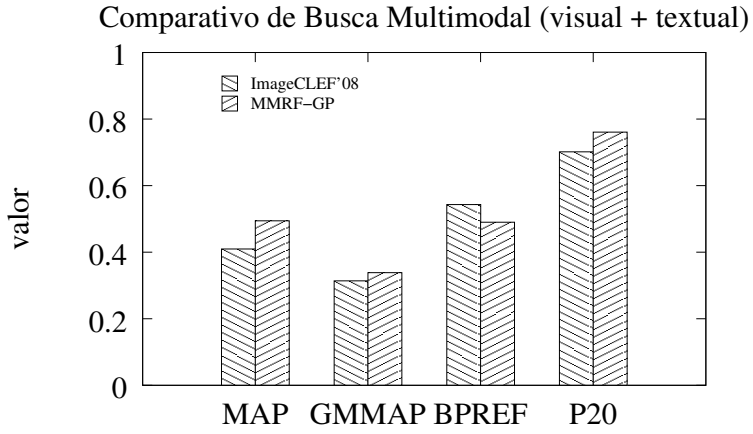


Figura 5.11: Comparativo entre o melhor resultado multimodal do ImageCLEFphoto 2008 e o melhor resultado multimodal do MMRF-GP. Resultado médio para 39 consultas.

Os resultados com a medida bpref, reportados nas Figuras 5.10 e 5.11, podem ser explicados dado o poder de distintividade dos descritores utilizados que, apesar de serem úteis para o posicionamento de muitas imagens relevantes no início do resultado (vide valor de P20), acabam não sendo eficientes para imagens que não se adequam visualmente ao padrão de consulta, mesmo com elevado grau de similaridade conceitual. Isso faz com que imagens irrelevantes sejam erroneamente posicionadas à frente de algumas das imagens relevantes.

Os resultados apresentados na Figura 5.11 mostram que o MMRF-GP obteve ganho, de 21%, 8% e 8% para MAP, GMAP e P20, respectivamente, em relação ao método de referência.

Melhores Funções de Combinação de Similaridades Multimodais

A seguir são apresentadadas δ_{1-5} que correspondem a exemplos das melhores funções de combinação de similaridade encontradas no processo evolutivo após a décima iteração de realimentação de relevância para algumas consultas na coleção ICphoto. Verifica-se a construção de funções para combinação de similaridades provenientes de diferentes modalidades e evidências. Nestas funções operadores são utilizados para combinar os valores de similaridade obtidos a partir das diferentes modalidades, por exemplo, *cos_descricao* indica a similaridade usando a função cosseno entre a descrição textual das imagens comparadas. Além disso, termos como *las*, *acc* e *bic* indicam os valores de similaridade obtidos usando os respectivos descritores visuais entre as imagens que são comparadas.

Em δ_1, pode-se notar a ocorrência de diversos descritores de imagens, incluíndo cor e textura, bem como as medidas de tfidf-sum e cosseno para a descrição textual da imagem.

$$\delta_1 = \text{cos\_descricao} + (((\text{sqrt}((\text{acc} + \text{las}))) \times (\text{sqrt}((\text{jac} + (\text{cos\_descricao} + (((\text{tfidf\_descricao} \times (\text{sqrt}(\text{tfidf\_descricao}))) \times (\text{sqrt}((\text{qcch} + (\text{sqrt}(\text{tfidf\_descricao})))))) + (\text{acc} \times (\text{sqrt}((\text{acc} + \text{cos\_descricao}))))))))) + ((\text{cos\_descricao} + (\text{las} + \text{acc})) + \text{bic}))$$

Na função δ_2, pode-se notar uma maior variedade de medidas de similaridade textual, incluindo okapi, dice e tfidf-sum, além dos descritores de cor jac e acc e dos descritores textura qcch e ccom. Já em δ_3, os descritores de cor que ocorrem são o gch, o acc e o jac e os descritores de textura são qcch e las.

$$\delta_2 = (((\text{sqrt}(\text{qcch})) \times \text{okapi\_descricao}) \times ((\text{sqrt}(\text{acc})) \times ((((\text{jac} \times \text{jac}) \,/\, (\text{sqrt}(\text{dice\_descricao}))) + ((\text{sqrt}(\text{las})) \times (\text{dice\_descricao} + \text{jac}))) \times \text{tfidf\_descricao}))) \times (\text{acc} \times (\text{sqrt}((\text{acc} + \text{ccom}))))$$

$$\delta_3 = ((\text{sqrt}(((\text{las} + \text{gch}) \times (\text{acc} \times \text{acc})))) + ((((\text{tfidf\_descricao} + \text{acc}) + \text{ccom}) \,/\, \text{dice\_descricao}) \times ((\text{sqrt}((\text{sqrt}(\text{tfidf\_descricao})))) \times (\text{las} \times \text{cos\_descricao})))) +$$

((sqrt(((bic + gch) / (gch + bow_descricao)))) + (sqrt((dice_descricao / (dice_descricao × (sqrt((las × bic))))))))

Em δ_4, nota-se a ocorrência apenas do descritor de cor jac e das medidas de similaridade textual okapi e tfidf. Já em δ_5, ocorrem três descritores de cor (bic, jac e acc) e um de textura (las) combinados com 3 medidas de similaridade textual (bow, cosseno e tfidf-sum).

$$\delta_4 = (\text{okapi\_descricao} + ((((\text{jac} + \text{tfidf\_descricao}) + \text{okapi\_descricao}) \times \text{jac}) + (\text{jac} + ((\text{jac} \times (\text{jac} + \text{tfidf\_descricao})) \times \text{tfidf\_descricao})))) \times (\text{sqrt}((((\text{jac} + \text{tfidf\_descricao}) + \text{okapi\_descricao}) \times \text{jac})))$$

$$\delta_5 = (\text{sqrt}(\text{bic})) + (\text{bow\_descricao} \times (\text{sqrt}(((\text{jac} + ((\text{cos\_descricao} + \text{acc}) \times (\text{sqrt}(\text{las})))) + \text{tfidf\_descricao}))))$$

Exemplo Visual de Consultas

A seguir será apresentado um exemplo de consulta realizada na base ICphoto utilizando-se apenas a modalidade textual, apenas a modalidade visual e combinação das duas modalidades. Serão mostradas as imagens retornadas como conjunto inicial, o julgamento do usuário sobre elas e as imagens obtidas ao final do processo de realimentação de relevância. A Figura 5.12 mostra as 3 imagens e o texto utilizado na consulta.

Figura 5.12: Imagens e texto da consulta de exemplo.

A Figura 5.13 apresenta as 6 primeiras imagens retornadas como conjunto inicial, construído com base na abordagem de realimentação de relevância que considera apenas a similaridade visual. As imagens marcadas em verde são as que foram consideradas relevantes pelo usuário. A Figura 5.14 mostra as 6 primeiras imagens retornadas após o processo de realimentação de relevância utilizando a abordagem puramente visual (vis), apresentada na seção 5.2.2. Esta consulta obteve valor de MAP igual a 0.0794.

Figura 5.13: 6 primeiras imagens do conjunto inicial da consulta de exemplo usando apenas informação visual.

Figura 5.14: 6 primeiras imagens do resultado após realimentação de relevância usando apenas informação visual.

A Figura 5.15 apresenta 6 primeiras imagens retornadas como conjunto inicial com suas respectivas descrições usadas no cálculo de similariade entre a consulta e a coleção. As imagens marcadas em verde são as que foram consideradas relevantes pelo usuário. A Figura 5.16 mostra as 6 primeiras imagens retornadas após o processo de realimentação de relevância utilizando a abordagem puramente textual (txt), apresentada na seção 5.2.2. Esta consulta obteve valor de MAP igual a 0.5476.

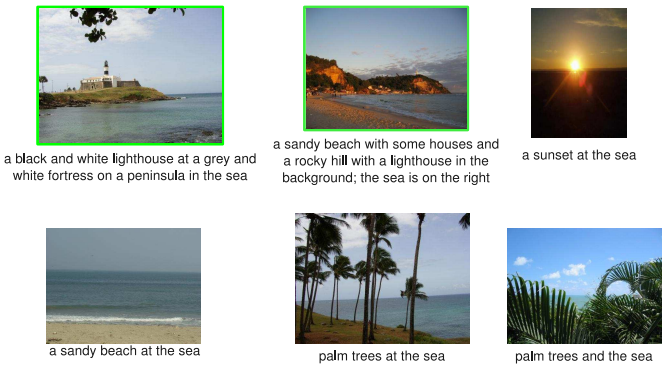


Figura 5.15: 6 primeiras imagens do conjunto inicial da consulta de exemplo usando apenas informação textual

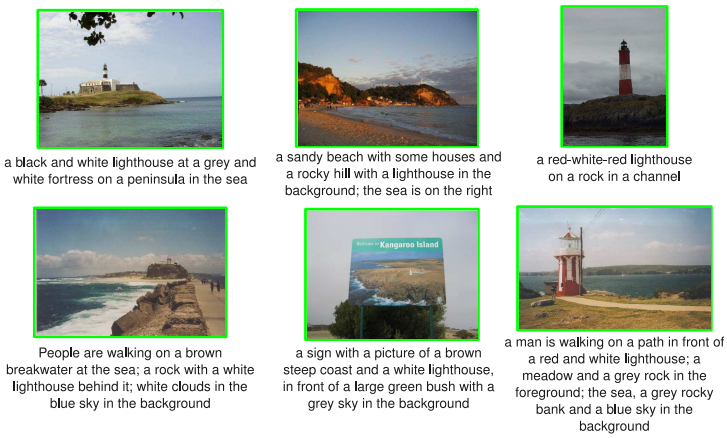


Figura 5.16: 6 primeiras imagens do resultado após realimentação de relevância usando apenas informação textual.

A Figura 5.17 mostra as 6 primeiras imagens retornadas como conjunto inicial, construído com base tanto na similaridade visual quanto textual entre as imagens de consulta e a coleção. As imagens marcadas em verde são aquelas consideradas relevantes pelo usuário. A Figura 5.18 mostra as 6 primeiras imagens retornas após os processo de realimentação de relevância utilizando a abordagem multimodal (mm-mm) apresentada na seção 5.2.2. Esta consulta obteve valor de MAP igual a 0.7245.

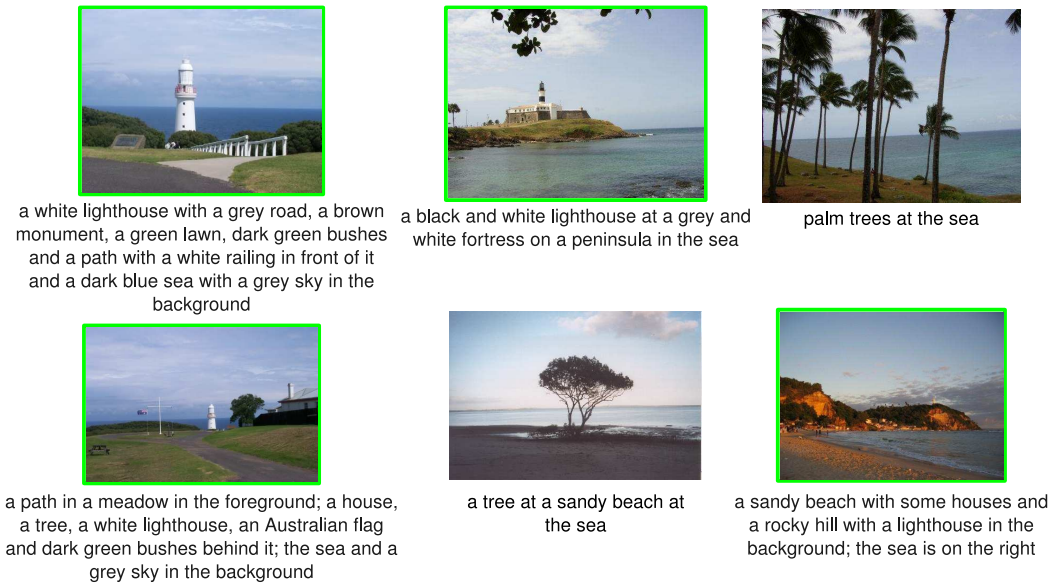


Figura 5.17: 6 primeiras imagens do conjunto inicial da consulta de exemplo usando informações visual e textual

a white lighthouse with a grey road, a brown monument, a green lawn, dark green bushes and a path with a white railing in front of it and a dark blue sea with a grey sky in the background

a black and white lighthouse at a grey and white fortress on a peninsula in the sea

a path in a meadow in the foreground; a house, a tree, a white lighthouse, an Australian flag and dark green bushes behind it; the sea and a grey sky in the background

a sandy beach with some houses and a rocky hill with a lighthouse in the background; the sea is on the right

a red-white-red lighthouse on a rock in a channel

a red-white-red lighthouse on a greyish-brown island surrounded by water in the foreground; snow-covered mountains behind it and grey clouds in the background

Figura 5.18: 6 primeiras imagens do resultado após realimentação de relevância usando informações textual e visual.

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Conclusões

Este trabalho propôs a utilização de diferentes modalidades e fontes de evidências no processo de recuperação de objetos digitais e a utilização de realimentação de relevância baseada em programação genética. Para tanto, foram apresentados e implementados dois arcabouços.

O primeiro arcabouço, RFCore, incorpora características comuns de aplicações de realimentação de relevância e foi desenvolvido com o objetivo de atuar como infra-estrutura para implementação, avaliação e comparação de técnicas que apliquem a realimentação de relevância em atividades de manipulação de dados. Este arcabouço foi construído com base no algoritmo clássico de realimentação de relevância e é adequado para utilização de diferentes técnicas em cada uma de suas etapas. Assim, para cada etapa foram definidas interfaces padronizadas que servem como modelo para implementação de componentes que podem ser acoplados a uma estrutura central responsável pela consistência do processo de realimentação de relevância.

O segundo arcabouço, MultimodalRFGP, foi desenvolvido sobre o RFCore, e fornece funcionalidades para aplicação da realimentação de relevância no processo de recuperação de objetos multimodais. No MMRF-GP o processo de aprendizado sobre a informação realimentada é realizado utilizando a programação genética. Neste contexto, a programação genética foi usada para a descoberta de funções de combinação de similaridades que consequem capturar às necessidades do usuário e consequentemente geram ordenações para objetos da coleção de maneira eficaz. É importante destacar que a programação genética tem papel fundamental neste processo dado que é capaz de, durante o processo de evolução, selecionar as características que mais se adequam às consultas e consequentemente geram resultados mais precisos.

Estes arcabouços foram validados por meio da construção de uma aplicação que utiliza

realimentação de relevância para recuperação de imagens a partir de evidências textuais e visuais. Os experimentos foram realizados utilizando um conjunto de descritores visuais e métricas de similaridade textual. Foram realizados experimentos sobre duas coleções de imagens. Resultados indicam a superioridade da combinação de diferentes fontes de evidências (multimodalidade) frente a abordagens puramente textuais ou visuais. Além disso, experimentos comparativos indicaram desempenho superior da utilização do MMRF-GP para recuperação de imagens quando comparados com os melhores resultados das submissões para o ImageCLEF 2008.

As principais contribuições deste trabalho são:

- Proposta e implementação de um arcabouço genérico para manipulação de dados com realimentação de relevância;
- Proposta e implementação de um arcabouço para recuperação de objetos multimodais com realimentação de relevância baseada em programação genética;
- Proposta e implementação parcial de uma máquina de busca multimodal de imagens, construída sobre os arcabouços propostos e que utiliza evidência textuais e visuais;
- Validação do sistema implementado por meio de experimentos e comparativos entre diferentes técnicas de recuperação e com outros metódos de comprovada eficácia.

6.2 Trabalhos Futuros

Esta seção apresenta sugestões de trabalhos de extensão e aprimoramento dos arcabouços e aplicações propostos neste trabalho.

Arcabouços:

- Aplicação dos arcabouços propostos para recuperação de outros tipos de objetos digitais;
- Utilização, no arcabouço MMRF-GP, de realimentação de relevância com exemplos negativos, com níveis de relevância variados ou até mesmo níveis *fuzzy*;
- Implementação de outras técnicas de recuperação de objetos digitais com separação total de evidências a serem utilizadas na construção do conjunto inicial e no processos de realimentação de relevância.

Recuperação de imagens:

- Combinação de métricas de similaridades textuais aplicadas à recuperação multimodal de imagens com dicionários visuais [122];

- Utilização de técnicas híbridas de aprendizado, por exemplo, utilização de máquinas de vetores de suporte para definição do conjunto inicial de objetos, incluisndo os mais informativos e posterior aplicação da programação genética nas iterações de realimentação de relevância. Experimentos com técnicas de aprendizado apropriadas para conjuntos de treinamento reduzidos;

- Utilização de técnicas de agrupamento para definição de um conjunto inicial de objetos que sejam relevantes, mas que exibem bom grau de diversidade com relação aos possíveis resultados relevantes para a consulta;

- Aplicação e avaliação da técnica de recuperação multimodal de imagens proposta, para coleções de imagens médicas, por exemplo sobre a coleção da *CLEF 2009 Medical Image Retrieval Track* [82];

- Realização de experimentos e avaliação da proposta implementada de recuperação multimodal de imagens com usuários reais;

- Aplicação e avaliação da técnica de recuperação multimodal de imagens proposta, com diferentes tipos de interfaces para definição de consulta e visualização de resultados;

- Utilização de descritores de conteúdo visual de diferentes naturezas, por exemplo: descritores globais e locais; descritores de cor, forma, textura e localização espacial; e descritores baseados em pontos de interesse;

- Utilização de diferentes técnicas de pré-processamento de texto e expansão de consultas, por exemplo, recuperação multilingual e expansão de consultas.

- Aplicação do arcabouço proposto pra busca multimodal de imagens em páginas *Web*;

- Utilização dos terminais propostos pela ACC [27] para cálculo de similaridades baseado em evidências textuais.

Programação Genética:

- Análise estatística da importância dos diferentes parâmetros da programação genética bem como estudo do espaço paramétrico para as diferentes coleções;

- Utilização de diferentes operadores aritméticos e até mesmo operadores mais complexos na programação genética, por exemplo operadores condicionais e de laço;

- Utilização de valores constantes nos terminais dos indivíduos da programação genética.

Referências Bibliográficas

[1] Thomas A., Paul C., M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access*, volume 5706 of *Lecture Notes in Computer Science*, pages 500–511. Springer Berlin / Heidelberg, 2009.

[2] R. Agrawal, W. Grosky, and F. Fotouhi. Image retrieval using multimodal keywords. In *ISM '06: Proceedings of the Eighth IEEE International Symposium on Multimedia*, pages 817–822, Washington, DC, USA, 2006.

[3] R. Agrawal, W. Grosky, F. Fotouhi, and C. Wu. Application of diffusion kernel in multimodal image retrieval. In *ISMW '07: Proceedings of the Ninth IEEE International Symposium on Multimedia Workshops*, pages 271–276, Washington, DC, USA, 2007.

[4] J. Allan, J. Callan, F. Feng, and D. Malin. Inquery and trec-8. In *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, volume 500, pages 637–644, 1999.

[5] P. Alshuth, T. Hermes, J. Kreyb, and M. Roper. *Image and Multi-media Search*, volume 8, chapter Intelligent Retrieval for Images and Videos. A. W. M. Smeulders and R. Jain, World Scientific, Farrer Road, Singapore, 1997.

[6] A. Amir, J. Argillander, M. Berg, S-F. Chang, M. Franz, W. Hsu, G. Iyengar, J. R Kender, L. Kennedy, C-Y. Lin, M. Naphade, A. Natsev, J. R. Smith, J. Tesic, G. Wu, R. Yan, and D. Zhang. IBM Research TRECVID-2004 Video Retrieval System. In *Proceedings of TRECVID 2004*, 2004.

[7] A. Amir, M. Berg, and H. Permuter. Mutual relevance feedback for multimodal query formulation in video retrieval. In *MIR '05: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 17–24, New York, NY, USA, 2005.

[8] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley-Longman, Boston, MA, USA, 1999.

[9] R. A. Baeza-Yates, R. Baeza-Yates, and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 1999.

[10] W. Banzhaf, P. Nordin, R. Keller, and F. Francone. *Genetic Programming - An Introduction*. Morgan Kaufmann Publishers, Inc, San Francisco, CA, 1998.

[11] B. Bhanu and Y. Lin. Object Detection in Multi-Modal Images Using Genetic Programming. *Applied Soft Computing*, 4(2):175–201, May 2004.

[12] E. Bruno, J. Kludas, and S. Marchand-Maillet. Combining multimodal preferences for multimedia information retrieval. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 71–78, New York, NY, USA, 2007.

[13] C. Buckley. New retrieval approaches using smart: Trec 4. In *Proceedings of the Fourth Text REtrieval Conference (TREC-4)*, pages 25–48, 1996.

[14] D. Cai, X. He, Z. Li, W. Ma, and J. Hierarchical clustering of www image search results using visual, textual and link information. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 952–959, New York, NY, USA, 2004.

[15] S. Clinchant, J. Renders, and G. Csurka. Trans-media pseudo-relevance feedback methods in multimedia retrieval. In *CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 569–576. Springer, 2008.

[16] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Müller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 579–594. Springer Berlin / Heidelberg, 2007.

[17] T. A. S. Coelho, P. P. Calado, L. V. Souza, B. Ribeiro-Neto, and R. Muntz. Image retrieval using multiple evidence ranking. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):408–417, 2004.

[18] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The Bayesian Image Retrieval System, PicHunter: Theory, Implementation, and Psychophysical Experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

[19] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target Testing and the PicHunter Bayesian Multimedia Retrieval System. In *Proceedings of the Third Forum on Research and Technology Advances in Digital Libraries*, pages 66–75, 1996.

[20] I.J. Cox, M.L. Miller, S.M. Omohundro, and P.N. Yianilos. PicHunter: Bayesian relevance feedback for image retrieval. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 3, pages 361–369, Vienna, August 1996.

[21] R. da S. Torres. *Integrating Image and Spatial Data for Biodiversity Information Management*. PhD thesis, Institute of Computing, University of Campinas, 2004.

[22] R. da S. Torres and A. X. Falcão. Content-Based Image Retrieval: Theory and Applications. *Revista de Informática Teórica e Aplicada*, 13(2):161–185, 2006.

[23] R. da S. Torres, A. X. Falcão, and L. da F. Costa. A Graph-based Approach for Multiscale Shape Analysis. *Pattern Recognition*, 37(6):1163–1174, June 2004.

[24] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, B. Zhang, W. Fan, and E. A. Fox. A New Framework to Combine Descriptors for Content-based Image Retrieval. Technical Report IC-05-21, Institute of Computing, State University of Campinas, Campinas, Brazil, 2005.

[25] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, J. P. Papa, B. Zhang, W. Fan, and E. A. Fox. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42(2):283–292, 2009.

[26] R. da S. Torres, J. A. M. Zegarra, J. A. Santos, C. D. Ferreira, O. A. B. Penatti, F. A. Andaló, and J. G. Almeida Jr. Recuperação de Imagens: Desafios e Novos Rumos. In *XXXV Seminário Integrado de Software e Hardware (SEMISH)*, Belém, Jul 2008.

[27] H. M. de Almeira. Uma abordagem de componentes combinados para a geração de funções de ordenação usando programação genética. Master's thesis, Universidade Federal de Minas Gerais, 2007.

[28] S. Deb and Y. Zhang. An overview of content-based image retrieval techniques. In *Proceedings of the 18th International Conference on Advanced Information Networking and Applications*, volume 1, pages 59–64, 2004.

[29] T. Deselaers, D. Keysers, and H. Ney. FIRE — flexible image retrieval engine: ImageCLEF 2004 evaluation. In *Multilingual Information Access for Text, Speech*

and Images, volume 3491 of *Lecture Notes in Computer Science*, pages 688–698. Springer Berlin / Heidelberg, 2005.

[30] T. Deselaers, R. Paredes, E. Vidal, and H. Ney. Learning weighted distances for relevance feedback in image retrieval. In *19th International Conference on Pattern Recognition, 2008. ICPR 2008*, pages 1–4, Dec. 2008.

[31] Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, April 2008.

[32] K. Donald and A. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. *International Conference on Image and Video Retrieval*, pages 61–70, 2005.

[33] R. Dorairaj and K.R. Namuduri. Compact combination of MPEG-7 color and texture descriptors for image retrieval. In *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 387–391, November 2004.

[34] J. A. dos Santos, C. D. Ferreira, and R. da S. Torres. A genetic programming approach for relevance feedback in region-based image retrieval systems. In *Brazilian Symposium on Computer Graphics and Image Processing*, Campo Grande, Oct. 12–15, 2008. IEEE Computer Society.

[35] L. Duan, W. Gao, W. Zeng, and D. Zhao. Adaptive relevance feedback based on Bayesian inference for image retrieval. *Signal Processing*, 85(2):395–399, February 2005.

[36] W. Equitz and W. Niblack. Retrieving images from a database using texture-algorithms from the QBIC system. IBM Research Report Technical Report RJ 9805, IBM, May 1994.

[37] W. Fan, E. A. Fox, P. Pathak, and H. Wu. The Effects of Fitness Functions on Genetic Programming-Based Ranking Discovery for Web Search. *Journal of the American Society for Information Science and Technology*, 55(7):628–636, 2004.

[38] W. Fan, M. D. Gordon, and P. Pathak. Personalization of search engine services for effective retrieval and knowledge management. In *ICIS '00: Proceedings of the twenty first international conference on Information systems*, pages 20–34, Atlanta, GA, USA, 2000.

[39] W. Fan, M. D. Gordon, and P. Pathak. Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *IEEE Transactions on Knowledge and Data Engineering*, 16(4):523–527, 2004.

[40] W. Fan, M. D. Gordon, and P. Pathak. A generic ranking function discovery framework by genetic programming for information retrieval. *International Journal of Information Processing and Management*, 40(4):587–602, 2004.

[41] W. Fan, M. D. Gordon, and P. Pathak. Genetic programming-based discovery of ranking functions for effective web search. *Journal of Management Information Systems*, 21(4):37–56, 2005.

[42] W. Fan, M. D. Gordon, P. Pathak, W. Xi, and E. A. Fox. Ranking function optimization for effective web search by genetic programming: An empirical study. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4*, pages 105–112, Washington, DC, USA, 2004.

[43] C. D. Ferreira. Recuperação de imagens com realimentação de relevância baseada em programação genética. Master's thesis, Universidade Estadual de Campinas, 2007.

[44] C. D. Ferreira, R. da S. Torres, M. Gonçalves, and W. Fan. Image retrieval with relevance feedback based on genetic programming. In *Brazilian Symposium on Databases*, pages 120–134, 2008.

[45] M. Flickner, H. Sawhney, W. Niblack, Q. Huang J. Ashley, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and Video Content: the QBIC System. *IEEE Computer*, 28(9):23–32, Sep 1995.

[46] R. B. Freitas and R. da S. Torres. OntoSAIA: Um ambiente Baseado em Ontologias para Recuperação e Anotação Semi-Automática de Imagens. In *Proceedings of Primeiro Workshop de Bibliotecas Digitais, Simpósio Brasileiro de Banco de Dados*, pages 60–79, Uberlandia, MG, Brazil, October 2005.

[47] I. Gagliardi G. Ciocca and R. Schettini. Quicklook^2: An Integrated Multimedia System. *Journal of Visual Languages & Computing*, 12(1):81–103, February 2001.

[48] J. Gauvain, L. Lamel, and G. Adda. The limsi broadcast news transcription system. *Speech Communications*, 37(1-2):89–108, 2002.

[49] I. Gondra and D. Heisterkamp. Adaptive and Efficient Image Retrieval with One-Class Support Vector Machines for Inter-Query Learning. *WSEAS Transactions on Circuits and Systems*, 3(2):324–329, April 2004.

[50] M. Grubinger, P. Clough, A. Hanbury, and H. Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *Advances in Multilingual and Multimodal Information Retrieval*, volume 5152 of *Lecture Notes in Computer Science*, pages 433–444. Springer Berlin / Heidelberg, 2008.

[51] N. Haque. *Image Ranking for Multimedia Retrieval*. PhD thesis, School of Computer Science and Information Technology - Royal Melbourne Institute of Technology, 2003.

[52] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, Copenhagen, Denmark, 1992.

[53] A. Hauptmann, M. Chen, M. Christel, C. Huang, W. Lin, T. Ng, N. Papernick, A. Velivelli, J. Yang, R. Yan, H. Yang, , and H.D. Wactlar. Confounded expectations: Informedia at trecvid 2004. In *In Proceedings of TRECVID 2004*, 2004.

[54] K. Hirata and T. Kato. Query by visual example - content based image retrieval. In *Proceedings of the 3rd International Conference on Extending Database Technology*, pages 56–71, London, UK, 1992. Springer-Verlag.

[55] S. C. H. Hoi. Cross-language and cross-media image retrieval: An empirical study at imageclef2007. In *CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 538–545. Springer, 2008.

[56] P. Hong, Q. Tian, and T. S. Huang. Incorporate support vector machines to content-based image retrieval with relevant feedback. In *Proceedings of the 7th IEEE International Conference on Image Processing*, pages 750–753, 2000.

[57] C. Huang and Q. Liu. An orientation independent texture descriptor for image retireval. In *International Conference on Computational Science*, pages 772–776, 2007.

[58] J. Huang, R. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[59] P. Huang, J. Bu, C. Chen, and G. Qiu. *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, chapter Improving Web Image Retrieval Using Image Annotations and Inference Network, pages 617–621. Springer Berlin / Heidelberg, 2008.

[60] D. N. F. Awang Iskandar, Jovan Pehcevski, James A. Thom, and S. M. M. Tahaghoghi. *Advances in XML Information Retrieval and Evaluation*, chapter Combining Image and Structured Text Retrieval, pages 525–539. 2006.

[61] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. P. Khudanpur, D. Klakow, M. R. Krause, R. Manmatha, H. J. Nock, D. Petkova, B. Pytlik, and P. Virga. Joint visual-text modeling for automatic retrieval of multimedia documents. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 21–30, New York, NY, USA, 2005.

[62] W. Jiang, G. Er, Q. Dai, and J. Gu. Hidden annotation for image retrieval with long-term relevance feedback learning. *Pattern Recognition*, 38(11):2007–2021, November 2005.

[63] F. Jing, H.-J. Zhang M. Li, and B. Zhang. Relevance feedback for keyword and visual feature-based image retrieval. *Lecture notes in computer science*, 3115:438–447, 2004.

[64] A. Kak and C. Pavlopoulou. Content-based image retrieval from large medical databases. *First International Symposium on 3D Data Processing Visualization and Transmission*, 10(1):138–147, June 2002.

[65] D.-H. Kim, C.-W. Chung, and K. Barnard. Relevance feedback using adaptive clustering for image similarity retrieval. *Journal of Systems and Software*, 78(1):9–23, October 2005.

[66] J. Kittler, M. Hatef, and R. P. W. Duin. Combining classifiers. In *Proceedings of 1996 International Conference on Pattern Recognition*, volume 1015, pages 897–901, 1996.

[67] B. Ko and H. Byun. Probabilistic neural networks supporting multi-class relevance feedback in region-based image retrieval. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 138–141 vol.4, 2002.

[68] V. Kovalev and S. Volmer. Color co-occurence descriptors for querying-by-example. In *Proceedings of the 1998 Conference on MultiMedia Modeling*, pages 32–38, 1998.

[69] J. R. Koza. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA, USA, 1992.

[70] W. B. Langdon. *Data Structures and Genetic Programming: Genetic Programming + Data Structures = Automatic Programming!* Kluwer, 1998.

[71] C. Lau, D. Tjondronegoro, J. Zhang, S. Geva, and Y. Liu. *Comparative Evaluation of XML Information Retrieval Systems*, chapter Fusing Visual and Textual Retrieval Techniques to Effectively Search Large Collections of Wikipedia Images, pages 345–357. 2007.

[72] M. S. Lew, editor. *Principles of Visual Information Retrieval – Advances in Pattern Recognition*. Springer-Verlag, London Berlin Heidelberg, 2001.

[73] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H.R. Garner. Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, (18):2298.

[74] B. Li and S. Yuan. A novel relevance feedback method in content-based image retrieval. In *Proceedings of International Conference on Information Technology: Coding an Computing*, pages 120–123, 2004.

[75] H. Lieberman, E. Rosenzweig, and P. Singh. Aria: An Agent for Annotating and Retrieving Images. *IEEE Computer*, 34(7):57–62, 2001.

[76] S. Loncaric. A Survey of Shape Analysis Techniques. *Pattern Recognition*, 31(8):983–1190, August 1998.

[77] K. Lu and X. He. Image retrieval based on incremental subspace learning. *Pattern Recognition*, 38(11):2047–2054, November 2005.

[78] Y. Lu, C. Hu, X. Zhu, H. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 31–37, New York, NY, USA, 2000. ACM.

[79] W. Y. Ma and B. S. Manjunath. Netra: A Toolbox for Navigating Large Image Databases. In *IEEE International Conference on Image Processing*, pages 256–268, 1997.

[80] J. L. Martínez-Fernández, J. Villena-Román, A. Garcia-Serrano, and J. C. G. Cristbal. Combining textual and visual features for image retrieval. In *In Proceedings of CLEF'2005*, pages 680–391, 2006.

[81] K. Meffert. Jgap - java genetic algorithms and genetic programming package. Disponível em `http://jgap.sf.net`. Acessado em 05/02/2010.

[82] H. Mller, J. K. Cramer, I. Eggel, S. Bedrick, S. Radhouani, B. Bakke, C. E. K. Jr., and W. Hersh. Overview of the clef 2009 medical image retrieval track. In *Working Notes for the CLEF 2009 Workshop*, Corfu, Greece, 2009.

[83] P. Muneesawang and Ling Guan. An interactive approach for CBIR using a network of radial basis functions. *IEEE Transactions on Multimedia*, 6(5):703–716, November 2004.

[84] NIST. Trec video retrieval evaluation. Disponível em `http://www-nlpir.nist.gov/projects/trecvid/`. Acessado em 08/02/2010.

[85] V. E. Ogle and M. Stonebraker. Chabot: Retrieval from Relational Database of Images. IEEE *Computer*, 28(9):40–48, Sep 1995.

[86] O. B. Penatti and R. da S. Torres. Color Descriptors for Web Image Retrieval: a Comparative Study. In *XXI Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens*, 2008.

[87] E. Persoon and K. Fu. Shape Discrimination Using Fourier Descriptors. *IEEE Transanctions on Systems, Man, and Cybernetics*, 7(3):170– 178, 1977.

[88] M. M. Rahman, B. C. Desai, and P. Bhattacharya. Multi-modal interactive approach to imageclef 2007 photographic and medical retrieval tasks by cindi. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary, September 2007.

[89] S. E. Robertson and K. Sparck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[90] S. E. Robertson and S. Walker. Okapi/keenbow at trec-8. In *Proceedings of the Eighth text Retrieval Conference (TREC-8)*, volume 500, pages 151–162, 1999.

[91] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*.

[92] Y. Rui, T. S. Huang, and S. F. Chang. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Communications and Image Representation*, 10(1):39–62, March 1999.

[93] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655, 1998.

[94] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[95] K. L. Santos, H. Almeida, R. da S. Torres, and M. A. Gonçalves. Recuperação de imagens da web utilizando múltiplas evidências textuais e programação genética. In *Brazilian Symposium on Databases*, pages 91–105, Fortaleza, Brazil, 2009.

[96] K. Sauvagnat. *Modèle flexible pour la recherche d'information dans des corpus de documents semi-structurés*. PhD thesis, Université Paul Sabatier, 2005.

[97] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29, New York, NY, USA, 1996.

[98] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[99] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders. Early versus late fusion in semantic video analysis. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402, New York, NY, USA, 2005.

[100] R. Stehling, M. Nascimento, and A. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 102–109, 2002.

[101] M. Swain and D. Ballard. Color Indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[102] B. Sznadjer, J. Mamou, Y. Mass, and M. Shmueli-Scheuer. Metric inverted - an efficient inverted indexing method for metric spaces. In *Proceedings of the Efficiency Issues in Information Retrieval Workshop*, pages 55–61, 2008.

[103] H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978.

[104] B. Tao and B. Dickinson. Texture recognition and image retrieval using gradient indexing. *Journal of Visual Communication and Image Representation*, 11(3):327–342, 2000.

[105] D. Tjondronegoro, J. Zhang, J. Gu, A. Nguyen, and S. Geva. *Advances in XML Information Retrieval and Evaluation*, chapter Integrating Text Retrieval and Image Retrieval in XML Document Searching, pages 511–524. 2006.

[106] H. Tong, J. He, M. Li, C. Zhang, and W. Ma. Graph based multi-modality learning. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 862–871, New York, NY, USA, 2005.

[107] S. Tong and E. Y. Chang. Support vector machine active learning for image retrieval. In *Proceedings of 9th ACM international conference on Multimedia*, pages 107–118, New York, NY, USA, 2001.

[108] M. Torjmen, K. Pinel-Sauvagnat, and M. Boughanem. Using pseudo-relevance feedback to improve image retrieval results. In *CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 665–673. Springer, 2008.

[109] A. Trotman. Learning to rank. *Information Retrieval*, 8(3):359–381, 2005.

[110] T. Tsikrika and M. Lalmas. Merging techniques for performing data fusion on the web. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 127–134, New York, NY, USA, 2001.

[111] A. Vadivel, A.K. Majumdar, and S. Sural. Characteristics of weighted feature vector in content-based image retrieval applications. *International Conference Intelligent Sensing and Information Processing*, pages 127–132, 2004.

[112] C. J. van Rijsbergen, S. E. Robertson, and M. F. Porter. New models in probabilistic information retrieval. Technical Report 5587, British Library Research and Development, 1980.

[113] R. C. Veltkamp and M. Tanase. *Content-Based Image and Video Retrieval*, chapter A Survey of Content-Based Image Retrieval Systems, pages 47–101. Kluwer, 2002.

[114] J. Villena-Román, S.Lana-Serrano, J. L. Martínez-Fernández, and J. C. González-Cristóbal. Miracle at imageclefphoto 2007: Evaluation of merging strategies for multilingual and multimedia information retrieval. In *CLEF 2007*, volume 5152 of *Lecture Notes in Computer Science*, pages 500–503. Springer, 2008.

[115] C. C. Vogt and G. W. Cottrell. Predicting the performance of linearly combined ir systems. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 190–196, New York, NY, USA, 1998.

[116] A. Williams and P. Yoon. Content-based image retrieval using joint correlograms. *Multimedia Tools Applications*, 34(2):239–248, 2007.

[117] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[118] P. Wu, B. S. Manjunanth, S. D. Newsam, and H. D. Shin. A texture descriptor for image retrieval and browsing. In *CBAIVL '99: Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, page 3, Washington, DC, USA, 1999. IEEE Computer Society.

[119] Y. Wu, E. Y. Chang, K. C. Chang, and J. R. Smith. Optimal multimodal fusion for multimedia data analysis. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM International conference on Multimedia*, pages 572–579, New York, NY, USA, 2004.

[120] Z. Xu, X. Xu, K. Yu, and V. Tresp. A Hybrid Relevance-Feedback Approach to Text Retrieval. *Proceedings of the 25th European Conference on Information Retrieval Research, Lecture Notes in Computer Science*, 2633:81–293, April 2003.

[121] R. Yan and A. G. Hauptmann. A review of text and image retrieval approaches for broadcast news video. *Information Retrieval*, 10(4-5):445–484, October 2007.

[122] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206, 2007.

[123] D. Zeimpekis and E. Gallopoulos. Tmg: A matlab toolbox for generating term-document matrices from text collections. Technical Report 01, Computer Engineering and Informatics Department of the University of Patras, 2005.

[124] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *In Proceedings of Tenth International Conference on Information and Knowledge Management*, pages 403–410, 2001.

[125] B. Zhang, M. A. Gonçalves, W. Fan, Y. Chen, E. A. Fox, P. Calado, and M. Cristo. Combining structural and citation-based evidence for text classification. In *Proceedings of the 13th ACM Conference on Information and Knowledge Management*, pages 162–163, 2004.

[126] D. Zhang and G. Lu. Evaluation of similarity measurement for image retrieval. *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, 2:928–931 Vol.2, Dec. 2003.

[127] D. Zhang and G. Lu. Review of Shape Representation and Description. *Pattern Recognition*, 37(1):1–19, Jan 2004.

[128] R. Zhang, Z. Zhang, M. Li, W. Ma, and H. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. *Multimedia Systems*, 12(1):27–33, August 2006.

[129] R. Zhao and W.I. Grosky. Narrowing the semantic gap - improved text-based web document retrieval using visual features. *IEEE Transactions on Multimedia*, 4(2):189–200, Jun 2002.

[130] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia System*, 8(6):536–544, 2003.