# PROGRAMAÇÃO GENÉTICA APLICADA À BUSCA DE IMAGENS

PATRÍCIA CORREIA SARAIVA

# PROGRAMAÇÃO GENÉTICA APLICADA À BUSCA DE IMAGENS

Tese apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Doutor em Informática.

ORIENTADOR: JOÃO MARCOS BASTOS CAVALCANTI

Manaus - Amazonas

Fevereiro de 2014

Saraiva, Patrícia Correia
D1234p Programação Genética Aplicada à Busca de Imagens / Patrícia Correia Saraiva. — Manaus - Amazonas, 2014
xxii, 90 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal do Amazonas
Orientador: João Marcos Bastos Cavalcanti

1. Recuperação de Imagens. 2. Programação Genética. I. Título.

CDU 519.6*82.10

# [Folha de Aprovação]

Quando a secretaria do Curso fornecer esta folha,
ela deve ser digitalizada e armazenada no disco em formato gráfico.

Se você estiver usando o `pdflatex`,
armazene o arquivo preferencialmente em formato PNG
(o formato JPEG é pior neste caso).

Se você estiver usando o `latex` (não o `pdflatex`),
terá que converter o arquivo gráfico para o formato EPS.

Em seguida, acrescente a opção `approval={`*nome do arquivo*`}`
ao comando `\ppgccufmg`.

*Para Edleno, Lucas e Júlia, as pessoas mais importantes da minha vida.*

# Agradecimentos

Agradeço ao professor João Marcos, meu orientador, por ter me aceitado como aluna de doutorado e por ter me conduzido até o fim desta caminhada.

Aos professores Marcos Gonçalves, da Universidade Federal de Minas Gerais e, Ricardo Torres, da Universidade Estadual de Campinas, pela valiosa colaboração durante o todo o doutorado.

Ao meu esposo Edleno, pela paciência, amor e incentivo constante.

Aos meus filhos, Lucas e Júlia, por colaborarem brincando um pouco mais quietos, mesmo sem entender o porquê, quando precisei de tranquilidade na reta final.

Ao professor Nivio Ziviani, da Universidade Federal de Minas Gerais e meu orientador no mestrado, por ter cedido as máquinas do LATIN (Laboratório de Tratamento da Informação) para realização de experimentos durante esta tese.

À professora Jussara Almeida, também da Universidade Federal de Minas Gerais, pela ajuda no entendimento e configuração dos projetos experimentais.

À minha mãe, Valquíria, pelo apoio incondicional durante toda a vida para que eu continuasse estudando.

À Suanny, por ter cuidado tão bem dos meus filhos, meus maiores tesouros, durante todos esses anos.

À FAPEAM, SUFRAMA, projeto TTDSW (Técnicas para Tratamento de Documentos Semi-estruturados na Web) PRONEM/FAPEAM (Protocolo 7985.UNI301.4630.29102012) e projeto INCT para a Web (processo CNPq nº 573871/2008-6) pelo apoio financeiro durante o doutorado.

A todos da secretaria do Programa de Pós-Graduação em Informática, especialmente à Elienai, secretária da pós, pela amizade e carinho.

A todos do laboratório de Banco de Dados e Recuperação de Informação por compartilharem esta experiência comigo.

*“A verdadeira riqueza não está nas coisas, mas no coração.”*
(*Papa Francisco.*)

# Resumo

O volume de informação codificada sob a forma de imagens tem aumentado de forma significativa nas últimas décadas. O uso cada vez mais frequente de *tablets*, *smartphones*, câmeras digitais e *notebooks* com suporte à aquisição de imagens e a facilidade para tornar essas imagens disponíveis publicamente em repositórios compartilhados, são fatores que contribuem ainda mais para este cenário. Atualmente, imagens são usadas nas mais diversas aplicações, seja para registrar momentos e ações em jornais e revistas eletrônicas, ou redes sociais, ou ainda para divulgar produtos em aplicações de comércio eletrônico. Na medida em que cresce o volume de imagens, cresce também o interesse por sistemas capazes de realizar busca em bases de dados de imagem.

O objetivo principal desta tese é investigar o impacto do uso de programação genética (GP - *Genetic Progamming*) como ferramenta para combinar diferentes fontes de informação disponíveis durante a busca de imagens. Mais especificamente, foram abordados dois contextos distintos como estudos de caso: a busca de imagens na Web utilizando informação textual extraída automaticamente das páginas Web e, a busca visual por meio da expansão da imagem de consulta utilizando informação derivadas de diferentes modalidades de dados, como texto e conteúdo visual. Para avaliar as estratégias propostas para o contexto de busca visual, escolheu-se como estudo de caso a busca visual de produtos em lojas de comércio eletrônico voltadas para o segmento de moda.

Os experimentos realizados no contexto de busca de imagens na Web mostraram que a abordagem evolucionária superou a melhor abordagem utilizada como *baseline*, com ganhos de 22,36% em termos de MAP. No cenário de busca visual de produtos em lojas de comércio eletrônico, os resultados experimentais mostraram que a expansão automática baseada em GP é uma alternativa efetiva para melhorar a qualidade dos resultados de um sistema de busca de imagens. Quando comparado a uma abordagem baseada somente em propriedades visuais, a expansão multimodal obteve ganhos de pelo menos 19% em todos os cenários de busca considerados. Quando comparado a uma abordagem similar, mas completamente *ad hoc*, o arcabouço baseado em GP obteve ganhos de até 54% em termos de MAP.

# Abstract

The volume of information encoded in the form of images has increased significantly in the last decades. Contributing to this scenario, the wide-spread use of mobile devices, such as tablets and smartphones, and even notebooks, which not only can take photos, but also easily send them to connected applications, such as web services and social networks. Nowadays, images are used in several applications, such as to record personal moments of people's life or showing products in e-commerce online stores. As a consequence, not only does the volume of images increase, but also the interest in solutions able to retrieve these images.

The main goal of this thesis is to investigate the impact of using genetic programming (GP) as a tool for combining different sources of evidence available when retrieving images. As case studies, we considered the application of GP in two different contexts: image retrieval on the Web using textual information automatically extracted from Web pages, and visual search by expanding the image query using information derived from different types of data, such as text and visual content. We evaluate the proposed expansion strategies in an application of visual search for products focused on e-commerce stores for the fashion domain.

Experiments performed in the context of image retrieval on the Web showed that the evolutionary approach outperformed the best baseline with gains of 22.36% in terms of MAP. In the context of visual search for e-commerce applications, experimental results indicated that automatic expansion based on genetic programming is an effective alternative for improving the quality of image search results. When compared to a genetic programming system based only on visual information, the multimodal expansion achieved gains of at least 19% in all scenarios considered. When compared to a similar approach, but completely *ad hoc*, the GP framework achieved gains of up to 54% in terms of MAP.

# Lista de Figuras

# Lista de Tabelas

# Sumário

# Capítulo 1

# Introdução

O volume de informação codificada sob a forma de imagens tem aumentado de forma significativa nas últimas décadas. O uso cada vez mais frequente de *tablets*, *smartphones*, câmeras digitais e *notebooks* com suporte à aquisição de imagens e a facilidade para tornar essas imagens disponíveis publicamente em repositórios compartilhados, são fatores que contribuem ainda mais para este cenário. Atualmente, imagens são usadas nas mais diversas aplicações, seja para registrar momentos e ações em jornais e revistas eletrônicas, ou redes sociais, ou ainda para divulgar produtos em aplicações de comércio eletrônico. Na medida em que cresce o volume de imagens, cresce também o interesse por sistemas capazes de recuperar essas imagens.

Historicamente, os primeiros sistemas de recuperação de imagens eram baseados exclusivamente em técnicas tradicionais de recuperação de informação textual (Chang & Fu, 1980; Chang & Kunii, 1981). Nessa abordagem, conhecida como TBIR (*Text-based Image Retrieval)*, anotações textuais fornecidas manualmente ou extraídas automaticamente dos documentos que contêm as imagens são utilizadas para indexá-las e recuperá-las. Problemas como anotações incompletas, imprecisas e não padronizadas, e o esforço necessário para anotar manualmente grandes bases de imagens são apontados como as principais desvantagens desta abordagem.

Na tentativa de transpor as dificuldades e desvantagens da abordagem anterior, diversas técnicas de recuperação baseadas no conteúdo visual das imagens, também conhecidas como CBIR (*Content-based Image Retrieval*), foram desenvolvidas e publicadas na literatura nos últimos anos (Smeulders et al., 2000; Datta et al., 2008; Vani & Raju, 2010). Essas abordagens são fundamentadas exclusivamente no uso de descritores, que extraem automaticamente as propriedades visuais das imagens, como cor, forma ou textura. Apesar dos avanços obtidos nesta área, a tarefa de recuperar informação relevante a partir de uma imagem de consulta segue sendo um grande desafio. Isso porque apesar desse tipo de abordagem ser

capaz de recuperar imagens visualmente similares, as imagens retornadas nem sempre estão semanticamente relacionadas com a imagem de consulta. Isso acontece porque diferentes usuários podem ter diferentes percepções sobre uma mesma imagem e nem sempre suas necessidades de informação podem ser traduzidas ou capturadas por meio de propriedades de baixo nível.

Embora ambas as abordagens possam ser aplicadas para recuperação de imagens em geral, a escolha entre TBIR e CBIR deve sempre considerar as características da aplicação alvo. No cenário da Web, por exemplo, a adoção de técnicas de CBIR apresenta algumas desvantagens. Primeiro, o custo de processamento para extrair as características visuais e calcular a similaridade pode afetar o desempenho do sistema negativamente e se tornar proibitivo. Segundo, para uma coleção altamente dinâmica e heterogênea, como é o caso das imagens disponíveis na Web, fica difícil decidir quais características visuais melhor representam o conteúdo das imagens. E terceiro, prover uma imagem de consulta para descrever uma necessidade de informação, nem sempre é uma tarefa trivial.

Por outro lado, apesar das abordagens de TBIR apresentarem, em geral, resultados mais efetivos se comparados às abordagens puramente visuais, esse tipo de abordagem nem sempre é considerada a escolha ideal. Nesse caso, podemos citar como exemplo, aplicações nas quais um usuário deseja obter informação sobre um objeto, pessoa ou lugar a partir de uma foto tirada com seu dispositivo móvel. Muitas vezes o usuário não é capaz de descrever a sua consulta textualmente. E mesmo quando isto é possível, devido às restrições e dificuldades inerentes à entrada de dados em tais dispositivos, o uso de tal modalidade de busca pode ser considerado inconveniente. Há ainda problemas relacionados à semântica capturada através da informação textual, como a ocorrência de palavras ambíguas, ou mesmo irrelevantes para a imagem que dificultam o processo de recuperação de imagens baseada exclusivamente em texto.

Com o objetivo de aproveitar as vantagens de ambas as técnicas para aumentar a efetividade dos sistemas de recuperação de imagens, a combinação destas duas modalidades de dados, texto e conteúdo visual, tem sido estudada nos últimos anos (Snoek et al., 2005; Zhang & Guan, 2009; Clinchant et al., 2011; Cheng et al., 2011; Arampatzis et al., 2011a,b). Essa combinação de duas ou mais modalidades de dados tem sido referenciada na literatura como fusão multimodal. Trabalhos de pesquisa demonstram que técnicas de combinação multimodal apresentam resultados melhores que as abordagens isoladas mesmo utilizando estratégias simples de fusão. Isto demonstra que essas duas modalidades são complementares entre si, apesar das diferenças de desempenho de cada abordagem individualmente. No entanto, a fusão de tipos de informação normalmente expressos em domínios diferentes, como é o caso de texto e imagens, inclui algumas questões importantes, como por exemplo: quais propriedades devem ser consideradas em um sistema com abordagem de recuperação

multimodal e de que forma estes dados podem ser combinados.

O objetivo principal desta tese é investigar o impacto do uso de Programação Genética (GP - *Genetic Progamming*) como ferramenta para combinar diferentes fontes de informação disponíveis durante a busca de imagens. Mais especificamente, foram abordados dois contextos distintos como estudos de caso: a busca de imagens na Web utilizando informação textual extraída automaticamente das páginas Web e, a busca visual por meio da expansão da imagem de consulta utilizando informações derivadas de diferentes modalidades de dados, como texto e conteúdo visual. Para avaliar as estratégias propostas para o contexto de busca visual, escolheu-se como estudo de caso a busca visual de produtos em lojas de comércio eletrônico voltadas para o segmento de moda.

A escolha pelo uso de Programação Genética deve-se ao sucesso obtido com o uso dessa técnica nos últimos anos na área de recuperação de informação (Oren, 2002; Fan et al., 2000, 2004, 2005; Trotman, 2005; de Almeida et al., 2007; Fan & Zhou, 2009) e na área de recuperação de imagens (Torres et al., 2009; dos Santos et al., 2009). Outras razões para o uso de GP incluem ainda a sua habilidade em explorar de forma otimizada o espaço de busca das soluções possíveis para um determinado problema e o fato de GP conseguir lidar com múltiplos objetivos.

Outros autores publicaram recentemente estudos sobre formas de utilizar GP em aplicações de busca por imagens (Torres et al., 2009; dos Santos et al., 2009; Calumby et al., 2012; Faria et al., 2010). A maior parte desses trabalhos foi publicada durante o desenvolvimento desta tese. Comparada aos trabalhos desenvolvidos em (Torres et al., 2009) e (Faria et al., 2010), esta tese também utilizou a programação genética como ferramenta para combinar, de forma efetiva, diversas evidências disponíveis para busca de imagens. No entanto, aqui a programação genética também foi aplicada para enriquecer a imagem de consulta com informação multimodal extraída dos vários resultados de descritores de imagens individuais, com o objetivo de melhorar o desempenho dos sistemas de busca de imagens por conteúdo. Nesta tese, a programação genética é ainda utilizada de forma *offline* para realizar a expansão automática da imagem de consulta, ao invés de ser aplicada durante o processamento da consulta, como feito em (Calumby et al., 2012). Comparada ao trabalho de dos Santos et al. (2009), esta tese apresenta uma extensão desse trabalho, com uma análise mais aprofundada das evidências textuais utilizadas, inclusão de novas evidências e estudo sobre o impacto da escolha de alguns parâmetros no desempenho do arcabouço de GP. Os resultados aqui apresentados contribuem para um melhor entendimento do potencial que GP tem na evolução de sistemas para busca por informação em imagens.

Esta tese de doutorado é baseada na hipótese de pesquisa de que a Programação Genética pode ser aplicada como ferramenta para combinar diferentes fontes de evidências disponíveis para a busca de imagens, bem como para enriquecer também a imagem de consulta

com informação multimodal, com a finalidade de melhorar os resultados dos sistemas de busca por imagens. Em particular, buscou-se durante esta tese a resposta aos seguintes questionamentos: (i) Programação Genética pode ser aplicada para derivar boas funções de ranking em problemas de busca por imagens, assim como feito em aplicações de busca para a Web? (ii) Programação Genética pode ser utilizada como solução para problemas de expansão multimodal, principalmente nos casos mais difíceis em que a consulta original é constituída apenas de uma imagem? Esta segunda questão é importante principalmente em função do novo cenário trazido pelas recentes evoluções tecnológicas, onde a busca visual deve ter um papel cada vez mais importante em aplicações de grande valor estratégico e comercial.

## 1.1 Contribuições

Esta tese produziu descobertas importantes no que diz respeito à aplicação de programação genética em abordagens para recuperação de imagens. Como principais contribuições desta tese podemos citar:

- Um estudo sobre a aplicação de GP na recuperação de imagens na Web utilizando diversas fontes de evidência textuais. Foram incluídas novas características textuais, além das evidências já exploradas em outros trabalhos (Piji & Jun, 2009; dos Santos et al., 2009). O arcabouço foi avaliado em uma coleção de imagens de domínio genérico e comparado com outras abordagens de recuperação de imagens.

- Um estudo sobre a aplicação de GP na recuperação de imagens através da expansão da consulta utilizando informação multimodal. São propostas quatro estratégias para expandir automaticamente uma imagem de consulta utilizando informações derivadas de outras modalidades de dados disponíveis na coleção. Para avaliar o desempenho das abordagens propostas, foram criadas duas coleções contendo imagens e informações de produtos de diferentes lojas do segmento de moda.

- Uso do projeto fatorial em dois níveis na avaliação experimental para melhor investigar o efeito de alguns parâmetros no desempenho dos arcabouços de GP nos cenários de busca considerados. Esta técnica provê um suporte necessário à tarefa de parametrização de abordagens evolucionárias.

- Descoberta de quais as evidências que mais contribuem para a efetividade de recuperação através da análise dos resultados produzidos pelas abordagens baseadas em GP.

### Publicações

A seguir é apresentada uma lista de publicações produzidas como resultado de pesquisa ao longo desta tese:

1. Saraiva, P. C.; Cavalcanti, J. M. B.; Gonçalves, M. A.; Santos, K. C. L.; Moura, E. S. and da S. Torres, R. (2013). Evaluation of parameters for combining multiple textual sources of evidence for web image retrieval using genetic programming. *Journal of the Brazilian Computer Society*, 19(2):147–160.

2. Saraiva, P. C.; Cavalcanti, J. M. B.; Moura, E. S.; Gonçalves, M. A. and da S. Torres, R. A Multimodal query expansion based on genetic programming for visually-oriented e-commerce applications. Submetido para o *Expert Systems With Applications International Journal*.

   Além dos artigos citados acima, outros foram publicados em co-autoria com alunos de mestrado, sendo contribuições periféricas de cooperação no decorrer desta tese:

3. Kimura, P.; Cavalcanti, J.; Saraiva, P.; Torres, R. and Gonçalves, M. (2011). Evaluating retrieval effectiveness of descriptors for searching in large image databases. *Journal of Information and Data Management*, 2(3):305–321.

4. dos Santos, J. M.; Cavalcanti, J. M. B.; Saraiva, P. C. and de Moura, E. S. (2013). Multimodal re-ranking of product image search results. In *Proceedings of the European Conference in Information Retrieval*, pp. 62–73.

## 1.2 Organização da Tese

Esta tese está organizada em 5 capítulos. O Capítulo 2 descreve os conceitos teóricos e principais trabalhos relacionados à área de abrangência desta tese. O Capítulo 3 apresenta um arcabouço para busca de imagens na Web que utiliza Programação Genética para a combinação de diversas fontes de evidências textuais. O Capítulo 4 apresenta um arcabouço para busca de imagens que utiliza a Programação Genética para a expansão automática de imagens de consulta por meio do uso de informação multimodal presente na coleção. Finalmente, o Capítulo 5 apresenta as conclusões e possibilidades de trabalhos futuros.

# Capítulo 2

# Conceitos Básicos e Trabalhos Relacionados

Este capítulo apresenta os conceitos básicos e trabalhos relacionados às áreas de abrangência desta tese. A Seção 2.1 apresenta conceitos e trabalhos relacionados à recuperação de imagens baseada em informação textual. A Seção 2.2 apresenta conceitos e trabalhos relacionados à abordagem de recuperação de imagens baseada em conteúdo visual. A Seção 2.3 apresenta a abordagem de recuperação de imagens baseada em informação multimodal e alguns trabalhos relacionados a esta abordagem. A Seção 2.4 descreve o paradigma de programação genética e seus componentes principais. Por fim, a Seção 2.5 descreve as medidas de avaliação de qualidade utilizadas ao longo desta tese.

## 2.1 Recuperação de Imagens Baseada em Texto

A recuperação de imagens baseada em texto foi a abordagem pioneira empregada para realizar a busca em coleções de imagens digitais. Nessa abordagem, cada imagem da coleção possui uma informação textual associada obtida a partir de diversas fontes, como, por exemplo, metadados, texto de páginas web, descrições manuais e até mesmo reconhecimento ótico de caracteres. A informação textual utilizada durante a fase de indexação das imagens é também utilizada para o cálculo de similaridade no processamento da consulta.

Nessa abordagem, o sistema de recuperação de imagens funciona como um sistema de recuperação de informação tradicional, no qual um modelo de recuperação de informação é empregado para representar documentos e consultas. O modelo de espaço vetorial (McGill & Salton, 1983; Baeza-Yates & Ribeiro-Neto, 2011) é, sem dúvida, o modelo mais conhecido em recuperação de informação. A ideia central do modelo vetorial é representar algebricamente, o conjunto dos termos de uma consulta $q$ e dos documentos de uma coleção $D$, como vetores em

um espaço euclidiano t-dimensional, onde $t$ é o número de palavras distintas da coleção (vocabulário). Ou seja, um documento $d_j$ é associado a um vetor $\overrightarrow{d_j} = (w_{1,j}, ..., w_{i,j}, ..., w_{t,j})$ e uma consulta $q$ é associada a um vetor $\overrightarrow{d_q} = (w_{1,q}, ..., w_{i,q}, ..., w_{t,q})$, onde $w_{i,j}$ e $w_{i,q}$ representam os pesos dos termo $i$ no documento $d_j$ e na consulta $q$.

Existem diferentes estratégias para se calcular os pesos $w_{i,j}$ e $w_{i,q}$, mas a estratégia mais usual é conhecida como $tf \times idf$. Aqui $tf$ (*term frequency*) é a frequência que um termo $i$ ocorre em um documento $d_j$. O $idf$ (*inverse document frequency*) representa o grau de importância de um termo na coleção e determina o quanto esse termo discrimina o documento. O $idf$ é definido pela seguinte equação:

$$idf = log\frac{N}{df_i} \tag{2.1}$$

sendo $N$ o número total de documentos na coleção e $df_i$ o número de documentos em que o termo $i$ ocorre.

Uma vez representados os documentos e a consulta em um espaço vetorial, é possível calcular o grau de similaridade de um documento $d_j \in D$ em relação a uma consulta $q$ como sendo a similaridade entre seus vetores. A similaridade pode ser calculada, por exemplo, utilizando-se o cosseno do ângulo $\theta$ formado entre estes dois vetores, fórmula conhecida por medida do cosseno (Baeza-Yates & Ribeiro-Neto, 2011) definida a seguir:

$$sim(d_j, q) = \cos\theta = \frac{\overrightarrow{d} \bullet \overrightarrow{q}}{|\overrightarrow{d}| \times |\overrightarrow{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{2.2}$$

Outra medida bastante utilizada para o cálculo de similaridade é a medida Okapi-BM25 (Robertson & Walker, 1999) definida pela Equação 2.3:

$$Okapi(d_j, q) = \sum_{i=1}^{n} \frac{(k_1+1) \times tf}{tf + k_1 \times \left((1-b) + b \times \frac{|d_j|}{avgdl}\right)} \times \log\frac{N - df + 0.5}{df + 0.5} \tag{2.3}$$

onde $tf$ é a frequência do termo $i$ no documento $d_j$, $N$ é o total de documentos na coleção, $df_i$ é o número de documentos no qual o termo $i$ ocorre, $| d_j |$ é o tamanho do documento $d_j$ (em palavras), $avgdl$ é o tamanho médio dos documentos na coleção (em palavras), $k_1$ e $b$ são parâmetros de ajuste de desempenho.

Os primeiros sistemas de recuperação de imagens baseados em texto são descritos em (Chang & Fu, 1980; Chang & Kunii, 1981). Nesses sistemas, as imagens eram rotuladas por meio de anotações manuais e recuperadas utilizando sistemas de gerenciamento de banco de dados.

Com o surgimento da World Wide Web, a informação textual associada às imagens da coleção passou a ser extraída das páginas web e usado para descrever a semântica das imagens embutidas na página. O trabalho de Tsymbalenko & Munson (2001) utilizou atributos de *tag* da imagem, o texto ao seu redor e o título da página web para indexar e recuperar as imagens de uma coleção.

O trabalho de Coelho et al. (2004), analisou quatro fontes de evidências textuais extraídas das páginas Web: *tags* de descrição, *tags* de metadados, texto completo e passagens de texto ao redor da imagem. O objetivo do trabalho era avaliar o resultado da busca utilizando cada evidência textual considerada e comparar com o resultado da busca utilizando diferentes combinações dessas evidências. Para compor a busca com diversas fontes de evidência, os autores empregaram o conceito de Redes Bayesianas e propuseram um novo modelo de recuperação de imagens na Web baseado no modelo de rede de crenças. Os autores concluíram que a combinação de diversas evidências textuais produz melhores resultados em relação às abordagens isoladas. Além disso, os autores também concluíram que o uso de passagens de texto ao redor da imagem apresenta melhores resultados quando comparados com o uso de texto completo. As duas principais fontes de evidência textual destacadas pelos experimentos foram as *tags* de descrição e passagens com 40 termos. O método de Coelho et al. (2004) foi utilizado como *baseline* nos experimentos realizados no Capítulo 3.

## 2.2 Recuperação de Imagens Baseada em Conteúdo

Sistemas de recuperação de imagens baseados em conteúdo (CBIR - *Content-based Image Retrieval*) são fundamentados no uso de descritores de imagens. Para que a busca por conteúdo seja viável em tais sistemas, é necessário que as imagens sejam descritas pelas suas propriedades intrínsecas, tais como cor, forma ou textura, normalmente representadas através de vetores de características. Nesse caso, os descritores de imagens são utilizados para extrair tais características das imagens, viabilizando assim as fases de indexação e busca.

Em Torres et al. (2009), um descritor $d$ é formalmente definido por um par $(\varepsilon_d, \delta_d)$, onde:

- $\varepsilon_d : I \rightarrow \mathbb{R}^n$ é uma função que extrai um vetor de características $\overrightarrow{v_I}$ de uma imagem $I$.
- $\delta_d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ é uma função que computa a similaridade entre duas imagens a partir de um cálculo de distância entre seus vetores de características correspondentes.

A Figura 2.1 mostra o fluxo típico de uma sistema de recuperação de imagens baseado em conteúdo. Um processo de extração de características é aplicado sobre cada imagem de uma coleção de imagens, por meio da função $\varepsilon_d$. O resultado desse processo é a geração de vetores que codificam características visuais das imagens, tais como cor, forma, textura ou uma combinação dessas propriedades. O tamanho do vetor vai depender da quantidade de características usada para representar as imagens. Uma vez que uma imagem de consulta é submetida pelo usuário, o mesmo processo de extração é realizado sobre a imagem e um vetor de característica também é obtido. A partir desse momento, o vetor de características da imagem de consulta é comparado com os vetores de características que foram gerados a partir da coleção de imagens, por meio da função $\delta_d$. Baseado nos valores de similaridade, um resultado final é produzido com as imagens da coleção ordenadas de acordo com seus respectivos valores de similaridade em relação à imagem de consulta.

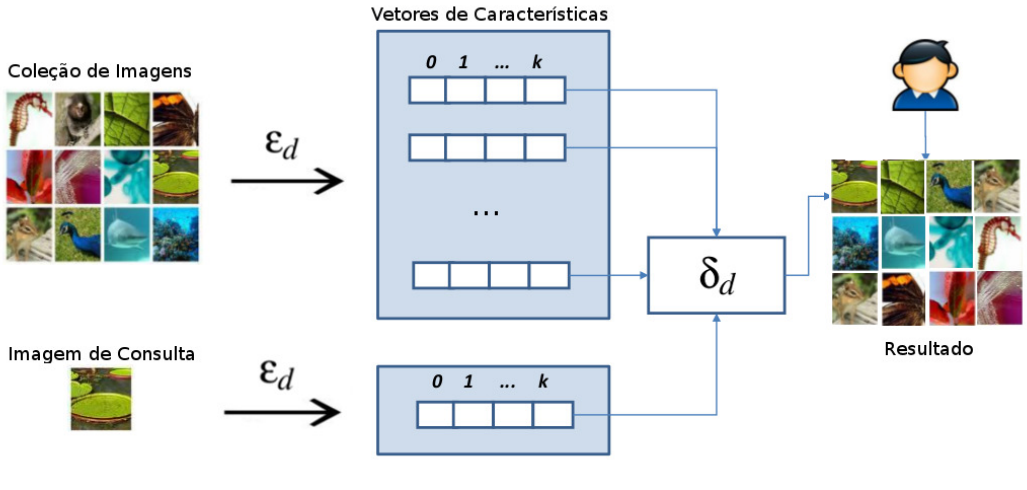


**Figura 2.1.** Fluxo básico de um sistema de recuperação de imagens baseado em conteúdo.

Uma questão fundamental na recuperação de imagens baseada em conteúdo é decidir quais as propriedades visuais que melhor se adequam ao domínio da aplicação, uma vez que tal escolha pode interferir consideravelmente a eficiência na recuperação. Para algumas

aplicações de busca de imagens de domínio específico, a escolha das propriedades pode ser baseada na homogeneidade das imagens contidas na coleção. Em uma aplicação de reconhecimento de impressões digitais, por exemplo, propriedades de textura dos objetos seria suficiente para alcançar bons resultados (Kherfi et al., 2004). No entanto, a escolha das propriedades visuais mais adequadas em coleções genéricas torna-se mais difícil devido à heterogeneidade das imagens a serem processadas.

Diversos sistemas de CBIR foram publicados na literatura ao longo dos últimos anos (Niblack et al., 1993; Sclaroff et al., 1997; Del Bimbo, 1999; Kherfi et al., 2002; Quack et al., 2004). Uma descrição mais completa e detalhada pode ser encontrada em (Shandilya & Singhai, 2010). Dentre os mais referenciados encontra-se o sistema Cortina[1], apresentado por Quack et al. (2004) como sendo o primeiro sistema de busca de imagens na Web baseada em conteúdo em larga escala que indexava mais de 3 milhões de imagens.

Chum et al. (2007) propuseram um método para expansão automática de consultas baseada apenas no conteúdo visual. O método, batizado de *Total Recall*, adota o modelo de recuperação de *bag-of-words* e realiza a expansão da consulta sobre o resultado inicial. Uma imagem é submetida e as top-k imagens retornadas passam por uma etapa de verificação geométrica com a imagem de consulta para suprimir falsos-positivo do resultado. As imagens com maior pontuação na verificação geométrica são utilizadas para expandir automaticamente a imagem de consulta. Esse método foi experimentado em coleções clássicas de imagens para detecção de versões (*near-duplicate*) e foi utilizado como *baseline* nos experimentos realizados no Capítulo 4.

## 2.3 Recuperação Multimodal

Nos últimos anos, pesquisas têm demonstrado que a recuperação de imagens baseada somente em informação textual ou visual sofre de limitações inerentes às próprias abordagens, como os problemas com anotações incompletas, imprecisas e não padronizadas e as dificuldades em capturar a semântica das imagens utilizando apenas propriedades visuais. Os resultados também demonstram que sistemas de recuperação alcançam desempenho melhor caso a fusão de ambas modalidades de dados seja explorada para compensar suas limitações (Smeulders et al., 2000).

Existem duas técnicas básicas para fusão de modalidades de dados: fusão precoce (*early fusion*) e fusão tardia (*late fusion*). Essas duas abordagens diferem no modo em que são combinadas as características obtidas a partir de cada modalidade. Como descrito em (Snoek et al., 2005), as abordagens que se baseiam em fusão precoce primeiramente

[1] `http://vision.ece.ucsb.edu/multimedia/cortina.shtml` Em 01/01/2014.

extraem as informações de cada modalidade e as combinam para produzir uma representação única do objeto analisado. Esta abordagem permite uma representação verdadeiramente multimodal, uma vez que as modalidades de dados são combinadas desde o início do processo de busca. Por sua vez, as abordagens baseadas em fusão tardia também realizam a extração das modalidades em separado. Ao contrário da abordagem de fusão precoce, na fusão tardia os resultados de cada modalidade são obtidos independentemente e depois combinados para produzir um resultado final. A Figura 2.2 ilustra estas duas abordagens de fusão.

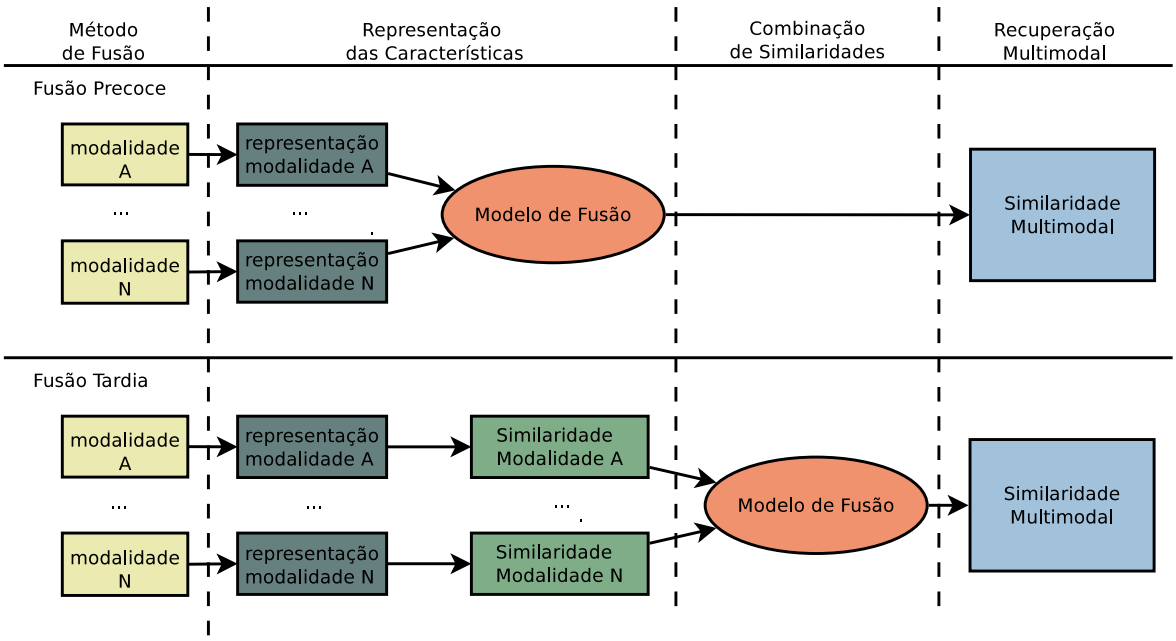


**Figura 2.2.** Abordagens de fusão precoce e fusão tardia.

O método de fusão precoce mais simples consiste em concatenar as representações textuais e visuais das imagens. No trabalho de Ferecatu & Sahbi (2008), a fusão precoce foi utilizada para combinar informação textual e visual em tarefas de anotação automática de imagens. Neste trabalho, as duas modalidades de dados foram simplesmente normalizadas antes de serem concatenadas. Uma comparação com uma abordagem de fusão tardia mostrou que fusão precoce obteve desempenho ligeiramente melhor, mas sem ganhos estatísticos. Abordagens mais elaboradas utilizam esquemas de ponderação das características (van Zaanen & de Croon, 2004; Deselaers et al., 2005, 2007).

Em Kittler et al. (1996), são apresentadas várias estratégias de fusão tardia incluindo a combinação por meio da soma ponderada, produto, votação e agregação min-max. Em McDonald & Smeaton (2005), os autores mostram que a soma ponderada está entre as abordagens mais eficazes para a recuperação baseada na combinação de texto e imagem. Estudos comparativos sobre abordagens de fusão precoce e fusão tardia foram realizados em Iyengar et al.

(2005) e Snoek et al. (2005). Entretanto, ainda não há um consenso sobre qual abordagem é a mais eficiente. Em Iyengar et al. (2005), os resultados obtidos demonstram que os sistemas baseados em fusão tardia proporcionam poucos ganhos em relação aos sistemas unimodais. Já em Snoek et al. (2005), os autores concluem que esquemas baseados em fusão tardia tendem a obter melhor desempenho para a maioria dos experimentos realizados. No entanto, para aqueles resultados onde a fusão precoce apresentou melhores resultados, a diferença entre as abordagens foi mais significativa.

Trabalhos de fusão tardia incluem estratégias de fusão baseadas na reordenação dos resultados, onde os documentos recuperados textualmente são reordenados baseados na similaridade visual (Zhou et al., 2009; Chang & Chen, 2007). Ou ainda, estratégias baseadas na similaridade das respostas realizando a interseção dos resultados das diferentes modalidades ou na combinação linear das similaridades das respostas (Villena-Román et al., 2007a,b; Müller et al., 2005; Depeursinge & Müller, 2010). Mais detalhes sobre técnicas de fusão multimodal podem ser encontradas em (Depeursinge & Müller, 2010).

Os trabalhos apresentados em (Arampatzis et al., 2011a; Chen et al., 2010; Clinchant et al., 2011; Cui et al., 2008; Liu et al., 2009; Yao et al., 2010) exploram a relação entre fontes de evidência textuais e visuais para melhorar os resultados de recuperação de imagens. Todas as abordagens compartilham a ideia de dividir o processo de busca em dois estágios principais. No primeiro estágio, a informação textual é usada para obter um resultado inicial. Em seguida, técnicas de CBIR são aplicadas para reordenar as imagens retornadas no passo anterior.

A abordagem de recuperação de imagens apresentada no Capítulo 4 segue a direção inversa dos trabalhos citados acima. Ao invés da busca ser realizada no domínio textual no primeiro estágio, evidências visuais são utilizadas para obter um resultado inicial. A partir de então, evidências textuais associadas às imagens recuperadas no passo anterior são utilizadas para realizar uma reordenação automática. Esta abordagem foi adotada devido às restrições do problema alvo estudado no Capítulo 4 e traz novos desafios uma vez que iniciar a busca com informação visual pode introduzir ruído e imprecisão ao processo devido ao problema do *gap* semântico, quando comparado com a recuperação baseada em texto. Para lidar com estes problemas, GP foi empregada para ajudar a encontrar as melhores possibilidades de expansão de consulta e reordenação de resultados em um grande espaço de possíveis soluções. Nesta linha de pesquisa, podemos citar o trabalho de (dos Santos et al., 2013), o qual foi desenvolvido durante esta tese, em cooperação com uma aluna de mestrado e foi utilizado como *baseline* nos experimentos do Capítulo 4. Esse trabalho apresenta o método TCatBR (*Term and Category-Based Reranking*), uma estratégia *ad hoc* para busca visual de produtos do segmento de moda que reordena automaticamente os resultados iniciais utilizando informação multimodal.

## 2.4 Programação Genética

Programação Genética (GP - *Genetic Programming*), uma extensão de Algoritmos Genéticos (GA - *Genetic Algoritm*), é uma técnica de aprendizagem baseada nos princípios descritos por Darwin de herança biológica, seleção natural e evolução, nos quais ocorre o fenômeno de adaptação das espécies na luta pela sobrevivência (Koza, 1992). O paradigma de Programação Genética pode ser definindo como um método automatizado para gerar geneticamente um programa de computador para resolver um dado problema, através da exploração otimizada do espaço de busca de todas as possíveis soluções para o problema. É a evolução direta de um conjunto de programas ou algoritmos para a finalidade de aprendizagem por indução.

No paradigma de Programação Genética, populações de programas de computadores, representados por indivíduos, vão evoluindo ao longo de gerações, a partir dos princípios de Darwin de sobrevivência dos indivíduos mais aptos e do cruzamento de espécies. As estruturas utilizadas em Programação Genética são hierárquicas com forma e tamanho variáveis. O processo de solução de problemas usando Programação Genética pode ser definido como sendo a busca pelo indivíduo mais apto, ou melhor indivíduo, no espaço de possibilidade das soluções para um dado problema. Em particular, este espaço de busca é o hiperespaço dos programas de computadores composto de funções e terminais apropriados para o domínio do problema. A Programação Genética tem sido bastante utilizada para solucionar problemas complexos diversos, onde os espaços de busca de soluções são amplos, e para os quais métodos convencionais não são capazes de encontrar uma boa resposta facilmente.

A população de um algoritmo de Programação Genética é representada por um conjunto de indivíduos. Cada indivíduo representa uma solução em potencial para o problema alvo e possui um conjunto de características próprias que o distingue dentre os demais indivíduos da população. Normalmente, os indivíduos são armazenados por meio de estruturas de dados como árvores binárias, listas encadeadas ou pilhas. Um indivíduo é formado pela combinação de funções e terminais adequados ao domínio do problema. Um exemplo de indivíduo é ilustrado na Figura 2.3.

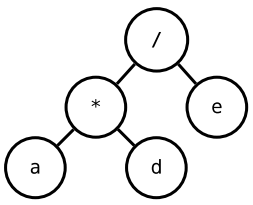


**Figura 2.3.** Exemplo de representação de um indivíduo com árvore binária.

O conjunto de indivíduos que forma a população inicial é gerado aleatoriamente a partir de um conjunto de funções e de terminais. O processo é realizado escolhendo-se funções e terminais para a composição da árvore, até que a profundidade máxima para a árvore seja atingida. Geralmente a profundidade máxima de um indivíduo é um parâmetro do arcabouço GP. A geração aleatória dos indivíduos da população inicial pode ser feita utilizando os métodos *Full*, *Grow* e *Ramped-half-and-half*.

**Método *Full*** O método *Full* de geração aleatória de indivíduos para a população inicial cria árvores completas, nas quais todos os nós-folha possuem a mesma distância até o nó-raiz. Esta distância é igual à profundidade máxima previamente definida para os indivíduos.

**Método *Grow*** O método *Grow* cria árvores de profundidade variável. A escolha dos nós é feita de forma aleatória entre funções e terminais; porém, a profundidade de um nó qualquer até o nó-raiz está restrita à profundidade máxima definida para um indivíduo.

**Método *Ramped-half-and-half*** Combinar os métodos *Full* e *Grow* com objetivo de gerar um número igual de árvores para cada profundidade, entre 2 e a profundidade máxima, é a base do método *Ramped-half-and-half*. Por exemplo, supondo que a profundidade máxima seja 6, então serão geradas árvores com profundidades de 2, 3, 4, 5 e 6 equitativamente. Isto significa que 20% dos indivíduos gerados por este método terão profundidade 2, 20% terão profundidade 3 e assim sucessivamente. Para cada profundidade, 50% são geradas pelo método *Full* e 50% pelo método *Grow*.

O conjunto de todos os indivíduos forma um espaço $\Sigma$, no qual uma função de aptidão $(f(\cdot) :\rightarrow \mathbb{R})$ toma este espaço de soluções, $\Sigma$, como seu domínio, e retorna um número real para cada indivíduo no espaço. Neste caso, possíveis soluções para o problema alvo, representadas pelos indivíduos, passam a ser avaliadas e ordenadas de acordo com seus valores de aptidão, com o intuito de definir o seu grau de adequação, ou evolução, perante os demais membros da população. Os melhores indivíduos, ou soluções, terão maiores valores de aptidão e, por conseguinte, terão mais chance de se reproduzirem. A função de aptidão deve medir apropriadamente quão bem um indivíduo pode solucionar o problema em questão e sua definição depende diretamente do domínio do problema.

O objetivo do GP é buscar por uma solução mais próxima do ótimo, evoluindo a população de indivíduos geração após geração. O processo evolutivo ocorre através de transformações genéticas com o intuito de criar uma população mais diversa e com melhores soluções nas gerações subsequentes. Estas transformações genéticas são realizadas pelos operadores de Reprodução, Mutação, e Cruzamento.

**Reprodução** A operação de reprodução simplesmente copia ou, usando um termo mais apropriado, clona alguns indivíduos da geração $g$ para a geração $g+1$ sem modificar sua estrutura. A probabilidade de um indivíduo ser selecionado para reprodução deve ser proporcional ao seu valor retornado pela função de aptidão. Quanto melhor o indivíduo soluciona o problema, mais alta é a probabilidade de este ser clonado para a próxima geração.

Enquanto a operação de reprodução mantém os melhores indivíduos nas gerações futuras, as operações de cruzamento e mutação são responsáveis por introduzir mudanças genéticas na população, e portanto, prover novos indivíduos para a próxima geração.

**Cruzamento** Na operação de cruzamento, dois grupos de indivíduos são formados aleatoriamente. Os melhores indivíduos em cada um dos grupos são selecionados como pais, de acordo com a função de aptidão. Ocorre então a troca de sub-árvores entre os pais para produzir dois novos indivíduos, os filhos. Esta operação é ilustrada na Figura 2.4. Assim, os novos indivíduos podem obter os melhores fragmentos dos seus pais e portanto, podem superá-los, provendo uma melhor solução para o problema.

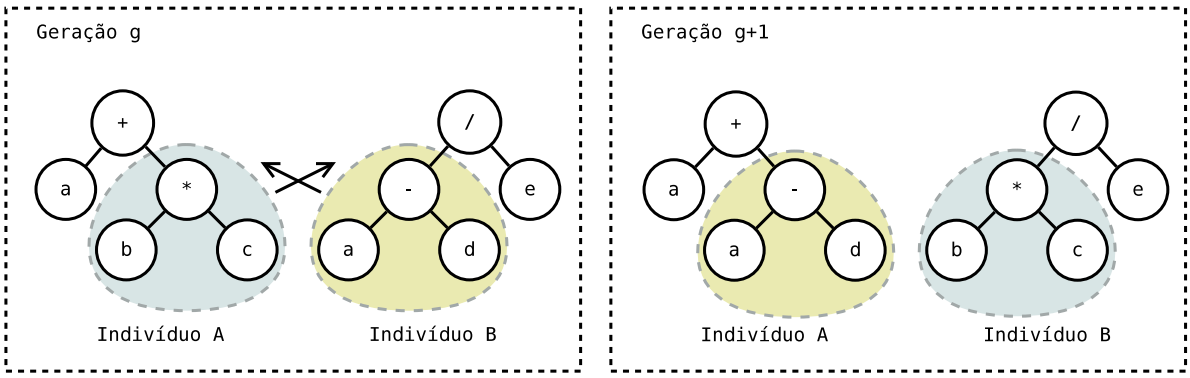


**Figura 2.4.** Operação de cruzamento entre indivíduos de uma população.

**Mutação** A mutação opera em um único indivíduo da população. As mudanças realizadas neste indivíduo representam a inclusão de um novo material genético. Um ponto de mutação é escolhido de forma aleatória no indivíduo e a sub-árvore que contém este ponto de mutação é substituída por uma nova sub-árvore gerada aleatoriamente. A operação de mutação é ilustrada na Figura 2.5. Devido ao fato de que o material genético recém-adicionado ao indivíduo não foi testado ainda, é provável que ocorra uma diminuição do valor de aptidão deste indivíduo. Por esta razão, embora a operação

de mutação seja importante na programação genética, ela tende a ser realizada com menor frequência do que a operação de cruzamento.

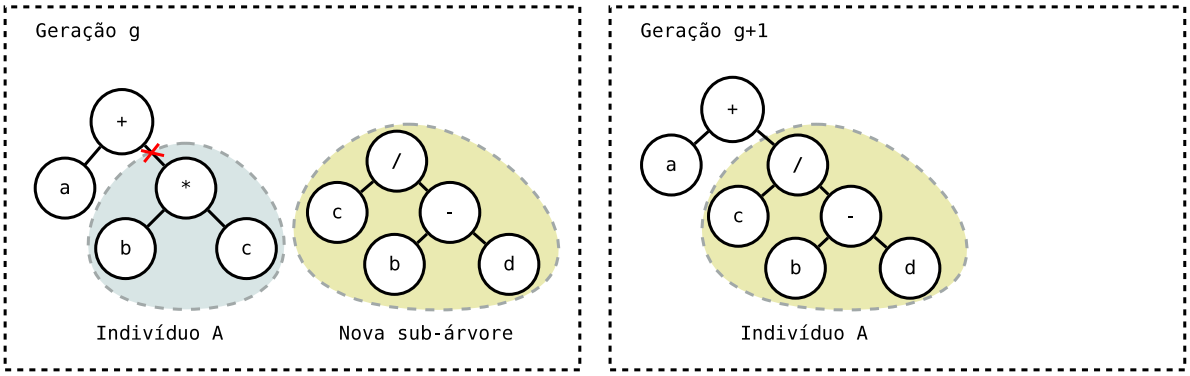


**Figura 2.5.** Operação de mutação em um indivíduo de uma população.

Com o uso desses operadores genéticos, gerações subsequentes mantém os indivíduos com os melhores valores de aptidão da última geração e renovam a população com novas soluções para o problema alvo, simulando os princípios da evolução Darwiniana. Um critério de parada, que normalmente é um número máximo de gerações ou um valor pretendido de aptidão, é responsável por interromper o processo evolutivo.

Para a aplicação de GP na resolução de problemas, alguns componentes chaves devem ser definidos no sistema. Estes componentes são apresentados na Tabela 2.1.

| COMPONENTES | DESCRIÇÃO |
|---|---|
| Terminais | Nós folhas na estrutura da árvore, como por exemplo, os nodos $a$, $d$, e $e$ na Figura 2.3. |
| Funções | Nós não-folhas utilizados para combinar os nós folha. Operações numéricas comuns são: $+, -, \times, \div$ |
| Operadores genéticos | Transformações genéticas aplicadas aos indivíduos: mutação, cruzamento e reprodução. |
| Função de aptidão | A função a ser otimizada pelo GP. |

**Tabela 2.1.** Principais componentes da Programação Genética.

A Figura 2.6 mostra o fluxograma básico do funcionamento da Programação Genética. O processo todo começa com a criação de uma população de indivíduos gerados aleatoriamente, chamada de população inicial (Passo 1). A partir daí, todos os indivíduos da população são avaliados conforme a função de aptidão definida no problema alvo (Passo 2). Caso uma

condição de parada previamente estabelecida seja alcançada, o melhor indivíduo daquela geração é retornado como solução para o problema. Caso contrário, indivíduos da população atual são selecionados baseados nos seus valores de aptidão (Passo 3). Então, operadores genéticos de mutação, reprodução e cruzamento são aplicados aos indivíduos selecionados na etapa anterior resultando em uma nova população (Passos 4 e 5). O ciclo volta a se repetir até que o critério de parada seja satisfeito. Esse critério de parada pode ser definido como o número máximo de gerações ou um determinado valor a ser atingido durante o processo evolucionário.

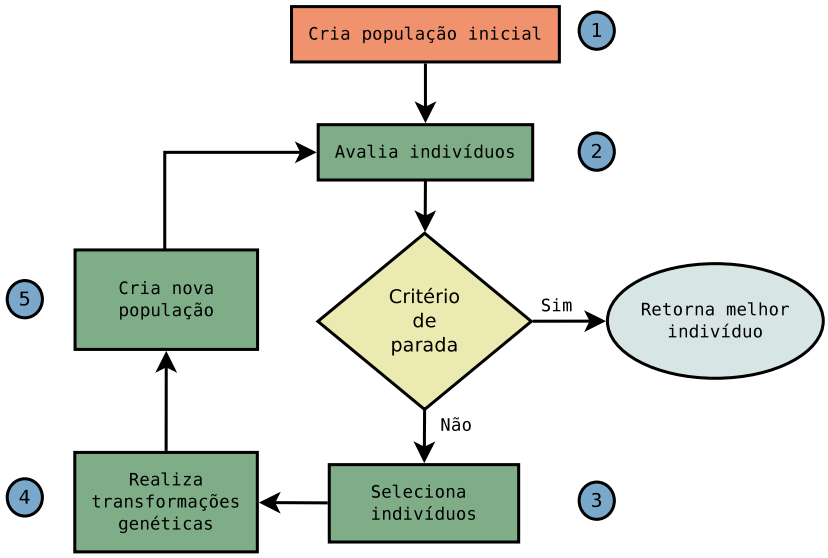


**Figura 2.6.** Funcionamento básico da Programação Genética.

Programação Genética tem sido explorada com bastante sucesso em diversas pesquisas na área de recuperação de informação (Oren, 2002; Fan et al., 2000, 2004, 2005; Trotman, 2005; de Almeida et al., 2007; Fan & Zhou, 2009). Sua aplicação em alguns trabalhos relacionados à recuperação de imagens também se mostrou bastante eficaz.

Torres et al. (2009) foram os primeiros a aplicar GP em recuperação de imagens baseada em conteúdo. O arcabouço proposto pelos autores utiliza descritores de forma e explora os princípios de GP para encontrar uma função de combinação de descritores mais efetiva, superando os resultados dos descritores individuais e do arcabouço baseado em Algoritmos Genéticos.

Em Faria et al. (2010), três abordagens de aprendizagem de máquina, CBIR-SVM, CBIR-GP, e CBIR-AR, foram propostas para combinar múltiplos descritores de imagens, utilizando Máquina de Vetores de Suporte (SVM - *Support Vector Machine*), Programação Genética (GP - *Genetic Programming*) e Regras de Associação (AR - *Association Rules*),

respectivamente. Os experimentos mostraram que CBIR-GP e CBIR-AR tiveram desempenho similar e ambos superaram o CBIR-SVM gerando uma melhor função de ordenação. Em um contexto diferente, GP foi usado por (Andrade et al., 2012) para combinar descritores locais e globais de imagens com o objetivo de recuperar vídeos.

Piji & Jun (2009) propuseram um modelo de recuperação de imagens baseado em GP para recuperação de imagens na Web, que combinava diferentes tipos de evidências como metadados, características visuais e análise de *links*. Informação temporal também foi utilizada baseada na hipótese de que usuários buscam informação mais recente. Nesse trabalho, a consulta é composta por uma imagem e uma descrição textual.

O trabalho de dos Santos et al. (2009) propôs um arcabouço de Programação Genética para combinar diversas evidências textuais extraídas automaticamente de páginas Web para gerar novas funções de ordenação. Os resultados foram comparados com uma abordagem de combinação baseada em Rede Bayesianas. O Capítulo 3 desta tese apresenta uma extensão ao trabalho proposto por dos Santos et al. (2009), com uma análise mais aprofundada das evidências utilizadas, inclusão de novas evidências textuais e estudo do impacto da escolha de parâmetros no desempenho do arcabouço de GP.

Em Ferreira et al. (2008), um novo arcabouço é proposto para recuperação de imagens baseada em conteúdo. O arcabouço emprega Programação Genética com Realimentação de Relevância para descobrir uma combinação efetiva de descritores que melhor caracteriza a percepção de similaridade do usuário. Calumby et al. (2012) utilizou o mesmo arcabouço para recuperação de imagens agregando informação textual e visual.

Os resultados de pesquisa apresentados nesta tese soma-se aos esforços de aplicação de GP a problemas relacionados à busca por imagens. São apresentados estudos que ampliam o escopo de utilização da técnica de GP e melhoram o entendimento de suas potencialidades.

## 2.5 Métricas de Avaliação de Qualidade

Métricas de avaliação de qualidade são medidas utilizadas para avaliar o quão preciso é o conjunto resposta de um sistema de busca. Segundo (Baeza-Yates & Ribeiro-Neto, 2011), este tipo de avaliação é conhecido por Avaliação de Desempenho de Recuperação. As medidas de avaliação mais conhecidas são Precisão e Revocação, que avaliam a qualidade de um conjunto de documentos retornados para uma determinada consulta. Estas métricas foram originalmente desenvolvidas para avaliar sistemas tradicionais de recuperação de documentos. No entanto, sua utilização em sistemas de recuperação de imagens é perfeitamente aceitável, uma vez que o propósito principal em ambos os sistemas é avaliar a informação recuperada de acordo com o julgamento de relevância, que traduz o nível de satisfação do usuário em

relação aos resultados obtidos.

Para se definir formalmente os conceitos de precisão e revocação é necessário considerar uma coleção de documentos $D$, uma consulta $q$ que expressa uma necessidade de informação do usuário, o conjunto $A$ dos documentos retornados pelo sistema em resposta a esta consulta $q$, o conjunto $R$ de documentos relevantes da coleção $D$ para a consulta $q$ e o conjunto $R_a$ dos documentos relevantes retornados pelo sistema de recuperação, formado pela interseção dos conjuntos $R$ e $A$. A Figura 2.7 ilustra estes conjuntos. A partir desta definição, seja $|A|$ o número de documentos em $A$, $|R|$ o número de documentos em $R$ e $|R_a|$ o número de documentos em $R_a$.

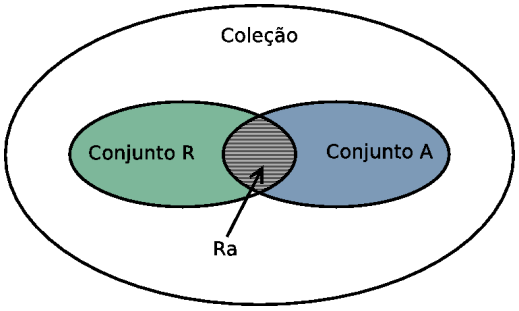


**Figura 2.7.** Conjunto de documentos relevantes e documentos retornados para uma consulta.

**Precisão** É a medida da habilidade de um sistema de retornar somente documentos relevantes. Isto é, precisão é a razão entre o número de documentos retornados que são relevantes $|R_a|$ e o número total de documentos retornados $|A|$.

$$\text{Precisão} = \frac{R_a}{|\,A\,|} \tag{2.4}$$

**Revocação** É a medida da habilidade de um sistema de retornar todos os documentos relevantes. Isto é, revocação é a razão entre o número de documentos retornados que são relevantes $|R_a|$ e o número total de documentos relevantes $|R|$.

$$\text{Revocação} = \frac{R_a}{|\,R\,|} \tag{2.5}$$

**Precisão na N-ésima Posição - (P@N)** É uma medida que baseia-se na definição de um limite $N$ de documentos que serão avaliados e assim a precisão é obtida até a N-ésima posição do resultado.

$$P@N = \frac{|\ relN\ |}{|\ N\ |} \tag{2.6}$$

sendo *relN* o número de documentos relevantes retornados até a posição *N* do resultado.

Para melhor analisar e visualizar o desempenho de um sistema, usualmente são utilizadas curvas de precisão em vários níveis de revocação. Ou seja, calcula-se a precisão para determinados valores de revocação. No entanto, para verificar se um sistema supera outro, é conveniente utilizar um único valor médio de precisão para cada consulta.

**Precisão Média (AP - *Average Precision*)** É uma medida que reflete o desempenho do sistema sobre todos os documentos relevantes. Esta medida é a média dos valores de precisão para uma determinada consulta obtidos depois que cada documento relevante é recuperado. Quando um documento relevante não é recuperado, assume-se que sua precisão é igual a 0.

$$AP(q) = \frac{\sum_{i=1}^{N}(P(d_i) \times rel(d_i))}{|R|} \tag{2.7}$$

onde $rel(d)$ assume o valor 1 se o documento $d$ é relevante e 0 caso contrário, $|R|$ é o número total de relevantes na coleção e $N$ é o número total de documentos retornados.

**Média dos Valores de Precisão Média (MAP - *Mean Average Precision*)** É uma medida do desempenho médio de um sistema para um conjunto de consultas $Q$. Isto é, calcula-se a média dos valores de precisão média obtidos para cada consulta $q$.

$$MAP = \frac{\sum_{i=1}^{|Q|} AP(q_i)}{|Q|} \tag{2.8}$$

onde $AP(q)$ é a precisão média para uma consulta $q$ e $|Q|$ é o número total de consultas utilizadas.

# Capítulo 3

# Busca de Imagens na Web

Este capítulo apresenta um arcabouço para recuperação de imagens na Web baseado em programação genética. O arcabouço assume que o texto e metadados presentes nas páginas podem ser usados como potenciais fontes de evidência para descrever as imagens e utiliza os princípios da programação genética para derivar boas funções de combinação não-lineares de evidências para melhorar a efetividade de sistemas de recuperação de imagens na Web.

Este arcabouço foi apresentado primeiro em (dos Santos et al., 2009) e é parte da dissertação de mestrado apresentada em (dos Santos, 2009). Como parte do trabalho apresentado neste capítulo, uma nova versão do mecanismo de extração foi desenvolvida para permitir a inclusão de novas evidências textuais e a inclusão de novos terminais no arcabouço de GP, além da inclusão de mais um *baseline* nos experimentos.

Em resumo, as principais contribuições deste capítulo são: (i) uma análise sobre quais fontes de evidência textuais são mais importantes para representar o conteúdo das imagens na Web; (ii) inclusão de novas características textuais como terminais do GP, além das evidências já exploradas em outros trabalhos (Piji & Jun, 2009; dos Santos et al., 2009); (iii) O uso do projeto fatorial em dois níveis na avaliação experimental para melhor investigar o efeito de alguns parâmetros na configuração do arcabouço de GP; (iv) uma avaliação experimental do arcabouço de GP em uma coleção de imagens de domínio genérico coletada da Web; (v) produção do artigo *Evaluation of parameters for combining multiple textual sources of evidence for web image retrieval using genetic programming* publicado no *Journal of the Brazilian Computer Society* em 2013.

## 3.1 Motivação

A World Wide Web é sem dúvida o maior e mais diverso repositório público de imagens já criado pela humanidade. Na Web, uma enorme quantidade de imagens é disponibilizada

diariamente nos mais diversos contextos, como viagens, notícias, comércio eletrônico, mapas, etc. Toda essa diversidade dificulta a adoção de uma abordagem de recuperação baseada em conteúdo visual (CBIR) que seja satisfatória para todos os contextos de busca.

Na Web, por outro lado, o texto das páginas está prontamente disponível e pode ser utilizado como fonte para gerar descrições das imagens presentes nessas páginas. Além disso, processar termos textuais comparado a processar características visuais é muito mais rápido e, portanto, mais adequado para o cenário da Web. Como a consulta do usuário também pode ser formulada em forma de texto para descrever a imagem de interesse, isto torna esta abordagem uma boa alternativa para recuperação de imagens na Web. De fato, máquinas de busca comerciais como Google[1] e Bing[2] utilizam descrições textuais como primeira opção para representar as imagens e utilizam técnicas tradicionais de recuperação textual no mecanismo de busca. Mesmo quando a abordagem de CBIR é aplicada no cenário da Web, o texto presente nas páginas não pode ser ignorado, uma vez que pode oferecer alguma forma de descrição do conteúdo semântico das imagens.

A próxima seção apresenta um visão geral do arcabouço de GP para busca de imagens na Web. O arcabouço utiliza diversas fontes de evidência textuais extraídas automaticamente das páginas e aplica os princípios da programação genética com o objetivo de descobrir funções de ordenação mais efetivas. A programação genética foi escolhida em razão do bom desempenho obtido com a utilização dessa técnica em outros trabalhos (Fan et al., 2000, 2004, 2005; de Almeida et al., 2007; dos Santos et al., 2009) e pela sua capacidade exploratória em grandes espaços de busca.

## 3.2 Arcabouço de GP para Busca de Imagens Baseada em Texto

O arcabouço de GP estudado é basicamente um processo iterativo de duas fases: treino e validação. Para cada fase, são selecionados um conjunto de consultas e documentos da coleção, que são chamados conjunto de treino, para a fase de treino, e conjunto de validação, para a fase de validação.

O arcabouço começa com a criação de uma população inicial de indivíduos, gerada aleatoriamente, que evolui geração após geração. Na primeira fase, o sistema é treinado com um conjunto de dados com o objetivo de aprender quais são as características que definem um indivíduo como uma boa solução. No final desta fase, são escolhidos os melhores indivíduos que serão avaliados mais adiante na fase de validação. São escolhidos também, dentre os

[1] `http://images.google.com` Em 01/01/2014.
[2] `http://br.bing.com` Em 01/01/2014.

melhores indivíduos, aqueles nos quais são aplicados os operadores genéticos de cruzamento, reprodução e mutação, a fim de que seja gerada a nova população de indivíduos da próxima geração. Durante a fase de validação, os indivíduos escolhidos são avaliados utilizando-se um segundo conjunto de dados. A ideia aqui é evitar uma super-especialização dos indivíduos baseada nas características do conjunto de treino, problema conhecido como *overfitting*.

Cada indivíduo é avaliado por meio da aplicação de uma função de aptidão (*fitness*) previamente definida. Uma vez que cada indivíduo é modelado como uma função de similaridade, a aplicação da função de aptidão significa calcular a função de ordenação para um indivíduo de acordo com o conjunto de documentos e consultas. O valor obtido para a função de aptidão é uma medida de qualidade do resultado obtido em relação à relevância dos documentos retornados.

O processo evolutivo continua até ser atingido um critério de parada. Ao final do processo evolutivo, de acordo com o método de seleção utilizado, o melhor indivíduo é escolhido como solução final para ser aplicado na fase de teste.

### 3.2.1 Fontes de Evidência Textuais

Documentos Web normalmente incluem uma diversidade de dados textuais que podem ser utilizados para descrever as imagens embutidas na página. Entretanto, no cenário da Web, o conteúdo textual de uma página nem sempre contém uma descrição apropriada das imagens presentes. Algumas vezes, o texto ao redor da imagem é apenas de caráter navegacional, como "próximo"ou "clique aqui". Muitas vezes, os nomes das imagens são gerados automaticamente e não carregam qualquer informação semântica do seu conteúdo como, por exemplo, imagens nomeadas como "imagem01.jpg". Estes problemas impedem a escolha de uma única fonte de evidência textual para descrever as imagens.

Uma solução possível para resolver este problema é considerar cada parte do documento HTML como uma fonte de evidência individual. A proposta é utilizar o arcabouço de GP para melhorar a qualidade do conjunto de imagens recuperadas para uma dada consulta textual, inferindo formas de combinação não-lineares dessas evidências extraídas automaticamente das páginas Web, descartando inclusive aquelas que não contribuem com a precisão do resultado da busca.

As evidências textuais utilizadas neste arcabouço são apresentadas na tabela 3.1. Tais evidências também foram utilizadas no trabalho de (Coelho et al., 2004) e no trabalho preliminar apresentado em (dos Santos et al., 2009). No entanto, em ambos os trabalhos citados, os conteúdos da *tag* de âncora, do nome do arquivo da imagem e do atributo ALT da *tag* <IMG> foram concatenados em uma única fonte de evidência denominada *tags* de descrição. Da mesma forma, os conteúdos extraídos do título da página Web e dos atributos

autor, descrição e palavras-chave foram concatenados em uma única fonte de evidência denominada *tags* de metadados. Nos experimentos descritos neste capítulo, essas evidências foram consideradas de maneira isolada, sem concatenação, para avaliar a utilidade de cada evidência individualmente. Outros tamanhos de passagens de texto ao redor da imagem foram incluídos, como passagens de 60, 80 e 100 termos, além das passagens de 10, 20 e 40 termos também utilizadas nos trabalhos de (Coelho et al., 2004) e (dos Santos et al., 2009). Estes novos tamanhos foram incluídos em razão de uma análise inicial da distribuição do tamanhos dos documentos presentes na coleção.

| EVIDÊNCIA TEXTUAL | DESCRIÇÃO |
|---|---|
| Texto completo | Texto completo da página Web contido na tag BODY (sem a marcação HTML) |
| Passagens de 10 termos | Texto ao redor da imagem (5 termos antes e 5 termos depois) |
| Passagens de 20 termos | Texto ao redor da imagem (10 termos antes e 10 termos depois) |
| Passagens de 40 termos | Texto ao redor da imagem (20 termos antes e 20 termos depois) |
| Passagens de 60 termos | Texto ao redor da imagem (30 termos antes e 30 termos depois) |
| Passagens de 80 termos | Texto ao redor da imagem (40 termos antes e 40 termos depois) |
| Passagens de 100 termos | Texto ao redor da imagem (50 termos antes e 50 termos depois) |
| Autor | Nome do autor da página extraído do atributo AUTHOR da tag META |
| Palavras-chave | Texto extraído do atributo KEYWORDS da tag META |
| Descrição | Descrição da página extraída do atributo DESCRIPTION da tag META |
| Tag SRC | Nome do arquivo da imagem encontrado no atributo SRC da tag IMG |
| Tag de âncora | Texto extraído das tags de âncora <A> |
| Tag ALT | Texto extraído do atributo ALT da tag IMG |
| Título | Título da página Web contido na tag TITLE |

**Tabela 3.1.** Evidências textuais utilizadas no arcabouço de recuperação de imagens baseado em GP.

### 3.2.2 Indivíduos

Indivíduos são utilizados para computar um *ranking* de imagens para cada consulta submetida ao arcabouço de recuperação de imagens. Para cada termo da consulta, o indivíduo é usado para computar um valor de similaridade para as imagens da coleção. Então, o valor de similaridade final de uma imagem em relação à consulta é resultado da soma dos valores obtidos para cada termo. O resultado final é obtido ordenando as imagens em ordem decrescente de seus respectivos valores de similaridade.

Os indivíduos são compostos pela combinação entre terminais e funções, representados na forma de uma estrutura de árvore como apresentado na Figura 2.3. Os valores dos terminais refletem algumas informações estatísticas derivadas diretamente da coleção, tais como a frequência do termo (tf), a frequência inversa do documento (idf), ou algum tipo de informação comprovadamente efetiva, como a similaridade Okapi-BM25, permitindo uma exploração efetivamente orientada do espaço de busca. Uma descrição de todos os terminais utilizados no processo evolucionário é apresentada na Tabela 3.2. Para cada uma das 14 evidências textuais apresentadas na seção anterior, estas informações foram aplicadas como terminais no GP. Além destes, foram também utilizados valores constantes na faixa de [0..100]. No trabalho preliminar apresentado em (dos Santos et al., 2009) apenas o terminal $tf\_idf$ foi utilizado no arcabouço de GP.

| TERMINAL | DESCRIÇÃO |
|---|---|
| tf | Frequência crua do termo |
| idf | Frequência inversa do documento definida pela equação 2.1 |
| tf_idf | Esquema de ponderação tf-idf |
| avgdl | Tamanho médio dos documentos da coleção |
| bm25 | Similaridade Okapi BM25 definida pela equação 2.3 |
| norma | Norma do documento |

**Tabela 3.2.** Terminais utilizados no arcabouço de recuperação de imagens baseado em GP.

Para combinar os terminais na formação dos indivíduos, foram utilizadas as funções de adição (+), subtração (-), multiplicação (*), divisão (/), logaritmo natural (log), logaritmo na base-10 (log10) e raiz quadrada (sqrt).

### 3.2.3 Função de Aptidão

A função de aptidão adotada no arcabouço evolucionário deve ser capaz de avaliar o resultado produzido por cada indivíduo em uma população de indivíduos. O resultado da função

de aptidão será usado no arcabouço como parâmetro para seleção dos indivíduos. Como os indivíduos representam esquemas de ponderação a serem utilizados em uma função de ordenação de documentos (imagens), a função de aptidão deve medir a qualidade do *ranking* produzido por um determinado indivíduo. A medida MAP apresentada na Equação 2.8 foi definida como função de aptidão no arcabouço de GP. MAP foi escolhida por ser capaz de expressar a qualidade do resultado produzido pelo indivíduo por meio de valor único de precisão em todos os valores de revocação.

### 3.2.4 Critério para Escolha do Melhor Indivíduo

A escolha de um indivíduo como solução final do arcabouço evolucionário deve levar em consideração o resultado obtido nas fases de treino e de validação, de modo a escolher um indivíduo capaz de generalizar para consultas não conhecidas e não escolher um indivíduo muito especializado para o conjunto de treino. Dessa forma, o critério utilizado para escolha do melhor indivíduo é baseado nos valores de aptidão obtidos em ambas as fases e no desvio padrão entre esses dois valores, como proposto em (de Almeida et al., 2007). A escolha é feita através do cálculo de $SUM_\sigma$ para cada indivíduo $i$ da população, de acordo com a Equação 3.1.

$$SUM_\sigma = (t_i + v_i) - \sigma_i \tag{3.1}$$

onde $t_i$ é o desempenho de um indivíduo $i$ na fase de treino, $v_i$ é o desempenho de um indivíduo $i$ na fase de validação e $\sigma_i$ o desvio padrão entre os valores de $t_i$ e $v_i$.

O indivíduo que possuir o maior valor de $SUM_\sigma$ é escolhido como melhor indivíduo da população. Vale ressaltar que um valor menor de $\sigma_i$ contribui para a seleção de indivíduos que tiveram desempenho mais regular nas fases de treino e validação, enquanto $(t_i + v_i)$ também dá preferência para aqueles que um bom desempenho em ambas as fases.

## 3.3 Experimentos

Nesta seção são apresentados os detalhes sobre os experimentos realizados para avaliação da abordagem de recuperação de imagens na Web baseada em GP. São apresentadas a coleção de imagens utilizada, a configuração e os parâmetros de GP utilizados nos experimentos, assim como os resultados obtidos a partir dos experimentos realizados.

O arcabouço de recuperação de imagens baseado em GP foi implementado com su-

porte da lil-gp (v1.1) [3], uma biblioteca muito utilizada para o desenvolvimento eficiente de aplicações baseada em programação genética em linguagem C.

Os experimentos realizados com o arcabouço evolucionário foram comparados com outras duas abordagens de recuperação: (i) Okapi BM25 (Robertson & Walker, 1999), como descrito na Equação 2.3, com os valores de $k_1 = 2$ e $b = 0,75$; e (ii) o arcabouço de recuperação Bayesiano apresentado em (Coelho et al., 2004).

No arcabouço Bayesiano, sete estratégias de recuperação com diversas evidências textuais foram utilizadas. Estas estratégias estão definidas na Tabela 3.3. A evidência textual denominada *tags* de descrição é formada pela concatenação dos textos extraídos da *tag* de âncora, do nome do arquivo da imagem e do conteúdo do atributo ALT da *tag* <IMG>. Da mesma forma, a evidência textual denominada *tags* de metadados é formada pela concatenação dos textos extraídos do título da página Web, e do conteúdo textual dos atributos autor, descrição e palavras-chave na definição da página.

| ESTRATÉGIAS DE RECUPERAÇÃO |
|---|
| *Tags* de Descrição |
| *Tags* de Metadados |
| Passagem de Texto |
| Descrição + Metadados |
| Descrição + Passagem de Texto |
| Passagem de Texto + Metadados |
| Descrição + Metadados + Passagem de Texto |

**Tabela 3.3.** Estratégias de recuperação para o arcabouço Bayesiano.

### 3.3.1 Coleção de Imagens

Com o objetivo de avaliar a abordagem de recuperação de imagens baseada em GP, foram realizados vários experimentos utilizando uma coleção de páginas coletadas do diretório Yahoo[4]. Todas as páginas coletadas foram armazenadas em conjunto com suas respectivas imagens para extrair as evidências textuais mencionadas anteriormente na seção 3.2.1.

A Tabela 3.4 apresenta alguns dados sobre a coleção utilizada nos experimentos. A coleção de imagens é bastante heterogênea, sem nenhuma categorização ou subdivisão em classes, e as imagens foram armazenadas da mesma forma que elas foram coletadas, sem

[3] `http://garage.cse.msu.edu/software/lil-gp/` Em 01/01/2014
[4] `http://br.yahoo.com` Em 01/01/2014.

nenhum processamento ou redução do tamanho. Foram consideradas imagens distintas àquelas que apresentaram URLs absolutas distintas.

Os experimentos foram conduzidos utilizando um total de 50 consultas textuais extraídas de um log de consultas de uma máquina de busca de imagens [5]. As consultas usadas nos experimentos foram: *Praia Rio de Janeiro, Pôr do sol, Fernando de Noronha, Mapa do Brasil, Igreja, Bola de futebol, Serra da Canastra, Jesus, Fotos de carnaval, Vaso de flores, Turma da Mônica, Hotel Glória, Cavalo mangalarga, Marisa Monte, Tubarão, Linux, Cerveja skol, Coca-cola, Carrefour, Corcovado, Basset, Pirenópolis, Machado de Assis, Brasil império, Hotel fazenda, Ditadura militar, Crianças desaparecidas, Maconha, Backstreet Boys, Pokemon, Índio, Coríntias, Barbie, Rosa, Palmeiras, Paisagem, Frutas, Halloween, Formatura, Aquecimento global, Praias, Natal, Cachorro, Natal, Bebê, Desenhos para colorir, Flamengo, Papai Noel, Animais, Jornal e Vírus.*

| **DADOS** | |
|---|---|
| Tamanho da coleção | 21 GB |
| Número de páginas HTML | 89.568 |
| Número de imagens distintas | 195.794 |
| Número de consultas | 50 |
| Número médio de imagens candidatas por consulta | 62 |
| Número médio de imagens relevantes por consulta | 28 |

**Tabela 3.4.** Dados sobre a coleção de imagens utilizada nos experimentos.

Para cada uma das 50 consultas, foram executadas as abordagens de recuperação baseadas no Okapi-BM25 e no arcabouço Bayesiano. No arcabouço Bayesiano, todas as sete estratégias apresentadas na Tabela 3.3 foram consideradas. As 30 imagens do topo dos resultados de cada abordagem foram agrupadas em um conjunto único (*pool*) de imagens candidatas para cada consulta. Desta forma, não é possível saber qual abordagem recuperou qual imagem. Cada *pool* foi analisado por um grupo de voluntários para avaliar as imagens como relevante ou irrelevante em relação à respectiva consulta. Ao final da avaliação, todas as imagens do *pool* foram rotuladas como relevantes ou irrelevantes, independente de como foram recuperadas.

O conjunto de imagens formou um *pool* de 62 imagens na média, por consulta. Ao final da avaliação, cada consulta continha uma média de 28 imagens consideradas relevantes. Esta técnica de *pooling* é bastante utilizada em coleções da TREC (Voorhees & Harman, 1999). Ela evita a necessidade de avaliar a coleção inteira e garante que o usuário avaliando as

[5] `http://busca.uol.com.br/` Em 01/01/2014.

imagens não tem nenhum conhecimento sobre a estratégia usada para recuperá-la, provendo assim uma avaliação imparcial de relevância das imagens recuperadas.

### 3.3.2 Parametrização do Arcabouço de GP

Um sistema baseado em GP possui um grande número de parâmetros que devem ser configurados antes do início do processo evolutivo. Esta configuração inicial cria uma explosão combinatória de possibilidades no espaço de parâmetros e torna a busca por uma configuração ótima, ou próxima do ótimo, uma tarefa difícil. Geralmente os trabalhos que utilizam GP configuram os parâmetros empiricamente baseados em poucos experimentos, ou utilizando valores padrões. Para facilitar a configuração do GP foi utilizada uma técnica de projeto experimental para avaliar o efeito de alguns parâmetros de GP e suas combinações para determinar seus efeitos quantitativos no resultado final.

O trabalho de Feldt & Nordin (2000) foi o primeiro trabalho a propor o uso de projeto experimental como uma metologia sólida e sistemática para estudar o efeito dos parâmetros de GP. Esta técnica pode ser utilizada para aumentar o desempenho de um sistema baseado em GP guiando a escolha de bons parâmetros de configuração. Adicionalmente, esta técnica também ajuda a investigar o impacto de usar altos valores para alguns parâmetros que possam impactar negativamente no tempo de treinamento. Caso um parâmetro não cause muito impacto nos resultados em termo de eficácia, pode-se reduzir seu valor de configuração para ganhar em eficiência.

Baseado nos resultados obtidos em (Feldt & Nordin, 2000), foi realizado um projeto fatorial completo em dois níveis (Box et al., 1978) para investigar o impacto de três parâmetros importantes: tamanho da população, número de gerações e profundidade máxima dos indivíduos no sistema de GP. Os dois primeiros parâmetros foram selecionados porque foram os que apresentaram maior impacto nos experimentos realizados por (Feldt & Nordin, 2000), e geralmente são os principais parâmetros configurados empiricamente em trabalhos usando GP. O último parâmetro foi adicionado ao projeto para investigar se o tamanho da árvore iria apresentar alguma influência significativa na resposta final.

Em um projeto fatorial completo em dois níveis, cada parâmetro a ser investigado é chamado de fator e possui dois níveis discretos, um nível mínimo (-) e um nível máximo (+). A saída do projeto fatorial é chamada de variável resposta. O experimento é realizado variando os níveis de cada fator, resultando em $2^k$ execuções diferentes, onde $k$ é o número de fatores no projeto. Os três fatores e seus respectivos valores mínimos e máximos são apresentados na Tabela 3.5. Os níveis dos parâmetros foram escolhidos para representar níveis qualitativamente distintos baseados em experiências prévias com o sistema de GP em uso.

A medida MAP foi utilizada como variável resposta. Como são analisados três fatores

| FATOR | PARÂMETRO | DESCRIÇÃO | MÍN. | MÁX. |
|---|---|---|---|---|
| A | *pop_size* | Número de indivíduos na população | 50 | 300 |
| B | *max_gen* | Número máximo de gerações no arcabouço de GP | 5 | 30 |
| C | *max_depth* | Profundidade máxima do indivíduo | 4 | 12 |

**Tabela 3.5.** Fatores e seus respectivos valores mínimo e máximo no projeto fatorial.

(A, B e C) e dois níveis para cada fator ($a_1$, $a_2$, $b_1$, $b_2$, $c_1$, e $c_2$), o projeto fatorial resultou em ($2^3$) experimentos diferentes como mostrado na Tabela 3.6. Para cada um dos oito experimentos, três replicações foram executadas para nos permitir avaliar o erro experimental, resultando em 24 execuções. No projeto fatorial, cada replicação é um experimento repetido com uma nova semente randômica configurada no arcabouço de GP para gerar uma população inicial de indivíduos completamente diferente das outras replicações.

| FATORES | RUNS | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $a_1b_1c_1$ | $a_2b_1c_1$ | $a_1b_2c_1$ | $a_2b_2c_1$ | $a_1b_1c_2$ | $a_2b_1c_2$ | $a_1b_2c_2$ | $a_2b_2c_2$ |
| Fator A | - | + | - | + | - | + | - | + |
| Fator B | - | - | + | + | - | - | + | + |
| Fator C | - | - | - | - | + | + | + | + |
| Interação AB | + | - | - | + | + | - | - | + |
| Interação AC | + | - | + | - | - | + | - | + |
| Interação BC | + | + | - | - | - | - | + | + |
| Interação ABC | - | + | + | - | + | - | - | + |

**Tabela 3.6.** Configuração experimental para o projeto fatorial completo em dois níveis.

Os efeitos de cada fator são calculados subtraindo a média das respostas nas quais o fator estava no seu nível mínimo pela média das respostas quando este mesmo fator estava no seu nível máximo. Maiores detalhes sobre projeto fatorial podem ser encontrados em (Box et al., 1978). Os efeitos de cada fator e suas respectivas interações são mostrados na Tabela 3.7 em ordem decrescente de efeito.

Pode-se observar que o tamanho da população (fator A) tem o maior efeito e explica $34,27\%$ da variação na resposta. O fator A foi cerca de $113\%$ maior que o efeito do fator B e foi cerca da $120\%$ maior que o fator C. Este resultado indica que a escolha de uma tamanho de população grande é importante para obter bons resultados neste cenário. Erros experimentais ou não-observados foram responsáveis por cerca de $9\%$ da variação na resposta.

Os valores utilizados na parametrização do arcabouço de GP são apresentados na Tabela 3.8. A configuração dos parâmetros tamanho da população, número de gerações e

| FATOR | EFEITO(%) |
|---|---|
| A | 34,27 |
| B | 16,05 |
| C | 15,55 |
| BC | 9,16 |
| ABC | 7,8 |
| AB | 4,5 |
| AC | 3,6 |

**Tabela 3.7.** Resultado do projeto fatorial completo em dois níveis para o arcabouço de GP.

profundidade máxima do indivíduo foi guiada pelos resultados obtidos com o projeto fatorial. Os demais parâmetros foram configurados para valores padrões como estabelecido em (Koza, 1992).

| PARÂMETRO | VALOR |
|---|---|
| Tamanho da população | 300 |
| Número máximo de gerações | 30 |
| Profundidade máxima do indivíduo | 7 |
| Operador de Cruzamento | 90% |
| Reprodução | 5% |
| Mutação | 5% |

**Tabela 3.8.** Configuração dos parâmetros no arcabouço de GP.

A população inicial foi gerada aleatoriamente utilizando o método *ramped half-and-half* descrito no Capítulo 2. A profundidade inicial dos indivíduos variou entre 2 a 6. Devido à estabilidade dos resultados obtidos no projeto fatorial, o número máximo de gerações, definido como critério de parada, foi 30. O parâmetro de profundidade máxima para as árvores geradas foi configurado para 7, por ser grande o suficiente para conter todas as características textuais definidas na Seção 3.1. Ao final de cada geração, a fase de validação foi executada para os 20 melhores indivíduos retornados pela fase de treinamento naquela geração.

### 3.3.3 Resultados Experimentais

Inicialmente, foram realizados experimentos com as diferentes estratégias de recuperação do arcabouço Bayesiano apresentadas na Tabela 3.3. Em seguida, experimentos com o arcabouço evolucionário são realizados e comparados com as duas abordagens de recuperação: Okapi-

BM25 e a estratégia de recuperação do arcabouço Bayesiano que apresentou melhor resultado em função da métrica adotada.

Em todos os experimentos realizados foi utilizada o método de validação cruzada denominado *k-fold*. A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. A ideia central deste método consiste em dividir o conjunto total de dados em $k$ subconjuntos mutualmente exclusivos, e a partir disto, um subconjunto é utilizado para teste e os $k-1$ subconjuntos restantes são utilizados para treino e validação do modelo. Este processo é realizado $k$ vezes alternando de forma circular o subconjunto de teste. Nos experimentos realizados foi utilizado $k = 5$. Desta forma, as 50 consultas definidas na Seção 3.3.1 foram divididas em cinco subconjuntos de 10 consultas cada. Em cada execução, três subconjuntos foram utilizados para treino, 1 subconjunto para validação e 1 subconjunto para teste.

#### 3.3.3.1 Resultados Experimentais com o Arcabouço Bayesiano

##### Texto Completo *versus* Passagens de Texto

O primeiro experimento foi realizado para determinar o melhor tamanho para as passagens de texto ao redor das imagens. Inicialmente, foi investigado o tamanho do texto completo dos documentos sem as *tags* HTML, para escolher bons tamanhos de passagens de texto a serem usadas nos experimentos.

A Figura 3.1 mostra a distribuição de tamanho dos documentos, em escala logarítmica, onde os documentos são visualizados em ordem decrescente de acordo com os seus tamanhos e o tamanho do documento é expresso em número de termos. Observa-se que a distribuição é de cauda pesada, onde uma pequena fração dos documentos tem um grande número de termos, e a grande maioria, cerca de 76% dos documentos, tem menos de 100 termos. A Tabela 3.9 apresenta algumas informações estatísticas sobre a distribuição de tamanho dos documentos da coleção de imagens.

| DADOS ESTATÍSTICOS | NÚMERO DE TERMOS |
|---|---|
| Tamanho médio dos documentos | 288 |
| Tamanho da mediana dos documentos | 90 |
| Tamanho do maior documento | 126.712 |
| Tamanho do menor documento | 1 |

**Tabela 3.9.** Informações estatísticas da distribuição de tamanho dos documentos.

O tamanho médio é bem maior que a mediana, o que confirma que a distribuição é tendenciosa. Além disso, existe uma grande variabilidade no tamanho dos documentos da

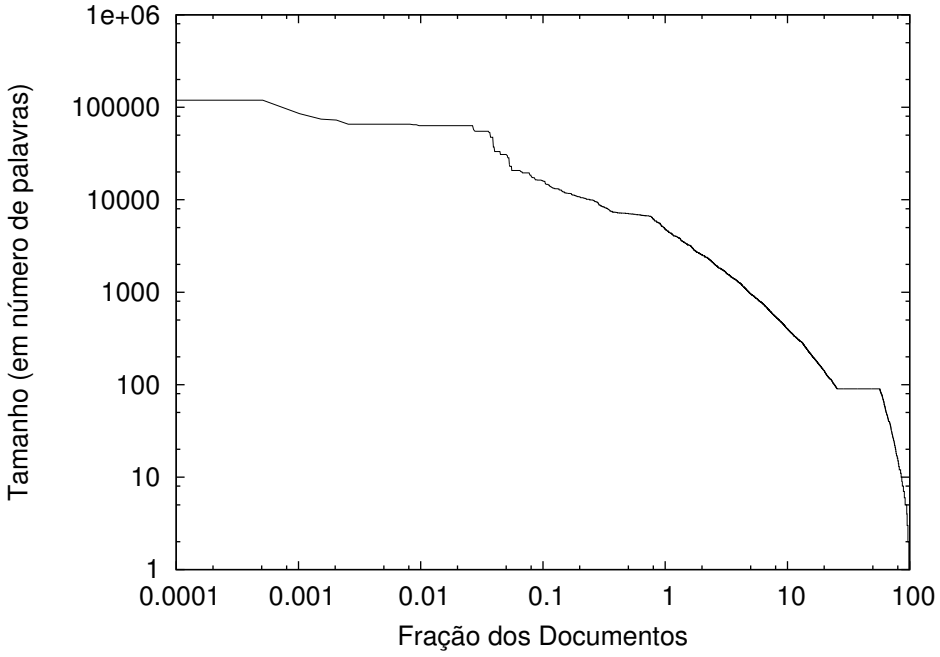


**Figura 3.1.** Distribuição de tamanho dos documentos da coleção de imagens.

coleção. Baseado nestas observações, foram utilizados diversos tamanhos de passagens de texto nos experimentos. Os tamanhos de passagens de texto escolhidos inicialmente foram 10, 20, 40, 60, 80 e 100 termos.

A Tabela 3.10 mostra os valores de MAP obtidos para cada tamanho de passagem de texto e para o texto completo utilizando a medida de similaridade Okapi-BM25. As passagens de 60 termos produziram o melhor resultado, enquanto que as passagens menores, de 10 e 20 termos, obtiveram os valores mais baixos de MAP. Uma conclusão evidente é que passagens de texto podem ser muito mais informativas sobre o conteúdo das imagens do que o texto completo da página. Uma razão para isto é que texto completo da página Web pode ser muito ambíguo, lidando com vários tópicos que podem não estar relacionados com o conteúdo das imagens contidas no documento. Por outro lado, passagens de texto com poucos termos podem ser insuficientes para prover boas descrições para as imagens.

A Figura 3.2 apresenta a curva de Precisão$\times$Revocação para todas as passagens de texto e o texto completo. Observa-se que as passagens de 10 e 20 termos foram as que obtiveram os piores resultados. As curvas para as passagens de texto de 40, 60, 80 e 100 termos apresentaram comportamento muito similares entre si, seguidas pela evidência textual de texto completo.

Para avaliar se as passagens de texto analisadas são estatisticamente diferentes uma das outras, foi aplicado o teste de significância Wilcoxon nos resultados para guiar a escolha pela melhor evidência. Embora a passagem de 60 termos tenha alcançado o melhor resultado

| TAMANHO DA PASSAGEM | MAP(%) |
|---|---|
| Texto completo | 24,859 |
| 10 termos (10T) | 21,536 |
| 20 termos (20T) | 19,981 |
| 40 termos (40T) | 26,791 |
| 60 termos (60T) | 28,341 |
| 80 termos (80T) | 27,945 |
| 100 termos (100T) | 25,428 |

**Tabela 3.10.** Resultados de MAP obtidos para as passagens de texto e texto completo no arcabouço Bayesiano.

em termos de medida MAP, este tamanho de passagem foi considerado estatisticamente equivalente às passagens de 40, 80, 100 e texto completo de acordo com o resultado do teste estatístico. Em razão do bom compromisso entre desempenho de recuperação e o menor uso de recursos computacionais exigidos para processar a passagem de texto, a passagem de 40 termos foi escolhida como melhor opção entre todas as passagens de texto analisadas, pois ela representa um ótimo custo-benefício na coleção em uso.

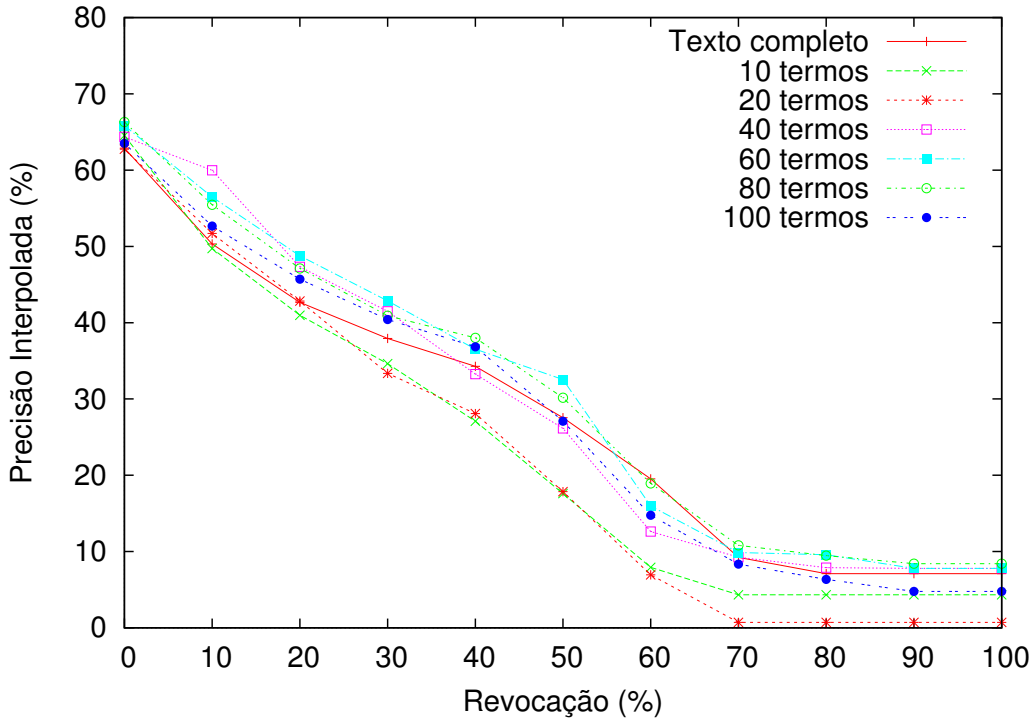


**Figura 3.2.** Curvas de Precisão×Revocação para todas as passagens de texto e texto completo no arcabouço Bayesiano.

### Evidências Isoladas

Com objetivo de reproduzir o arcabouço Bayesiano apresentado em (Coelho et al., 2004) nos experimentos, as evidências textuais da *tag* de âncora, nome do arquivo da imagem e *tag* ALT foram concatenadas em uma única evidência, denominada de *tags* de descrição. Da mesma forma, as evidências textuais extraídas do título da página Web, autor, descrição e palavras-chave também foram concatenadas em uma única evidência, denominada de *tags* de metadados.

Para avaliar o desempenho de cada evidência isoladamente, as *tags* de descrição e *tags* de metadados foram comparadas com a evidência de passagem de texto de 40 termos escolhida no experimento anterior. A Tabela 3.11 apresenta os valores de MAP obtidos para as três evidências analisadas. Observa-se que as passagens de texto tem maior contribuição na recuperação das imagens, seguido pela evidência de metadados. *Tags* de descrição foi a evidência menos informativa. Para avaliar se as evidências analisadas são estatisticamente diferentes uma das outras, foi aplicado o teste de significância Wilcoxon nos resultados. As passagens de texto foram estatisticamente melhores que as outras duas abordagens com nível de confiança superior a 95%.

| EVIDÊNCIA TEXTUAL | MAP(%) |
|---|---|
| Passagem de texto (40T) | 26,793 |
| *Tags* de Metadados | 18,172 |
| *Tags* de Descrição | 9,127 |

**Tabela 3.11.** Resultados de MAP obtidos para as fontes de evidência isoladas no arcabouço Bayesiano.

A Figura 3.3 mostra a curva de Precisão×Revocação para as três fontes de evidência analisadas. Nota-se que as passagens de texto são muito melhores que as evidências de metadados e de descrição para descrever as imagens na coleção em uso.

### Combinação de Evidências

No arcabouço Bayesiano, diversas fontes de evidências são combinadas como estratégias para recuperar imagens conforme apresentadas na Tabela 3.3. A Tabela 3.12 apresenta os valores de MAP obtidos para as quatro combinações analisadas: descrição+passagem de texto, descrição+metadados, metadados+passagem de texto, e descrição+metadados+passagem de texto. Observa-se que as combinações de descrição+metadados e descrição+passagem de texto obtiveram os piores resultados devido ao baixo desempenho da abordagem de *tags* de descrição obtido isoladamente. O teste de significância Wilcoxon foi aplicado para

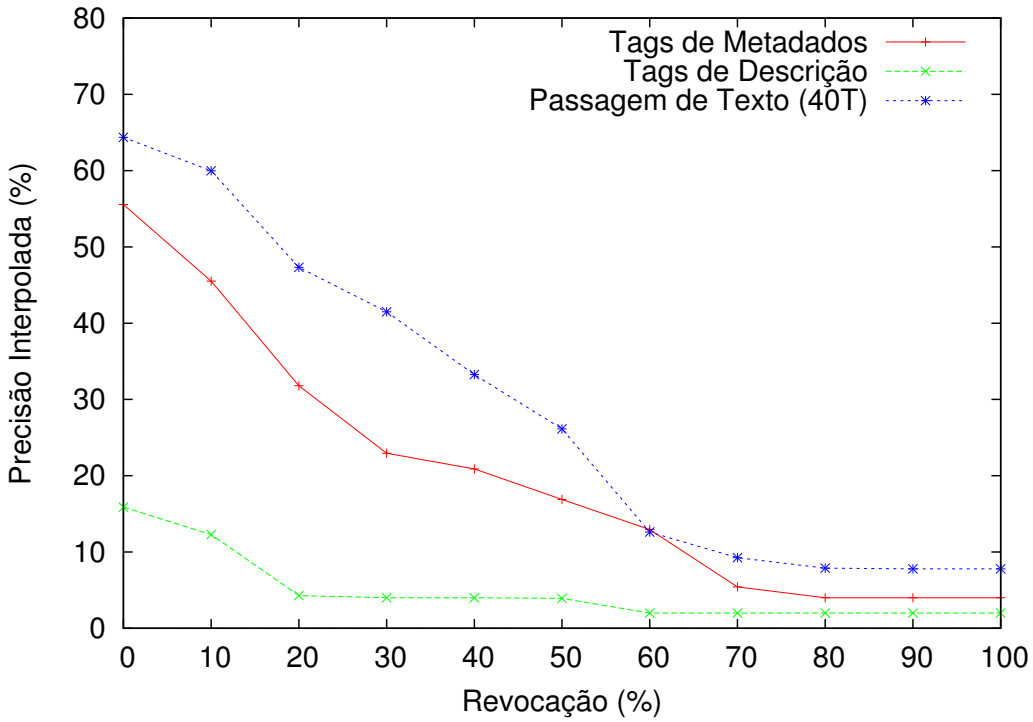


**Figura 3.3.** Curvas de Precisão×Revocação para as fontes de evidência isoladas no arcabouço Bayesiano.

analisar os resultados estatisticamente. A combinação metadados+passagem de texto foi significantemente superior à combinação descrição+metadados com nível de confiança de 99% e 90% superior quando comparada às combinações de descrição+metadados+passagem de texto e descrição+passagem de texto.

| EVIDÊNCIAS TEXTUAIS | MAP(%) |
|---|---|
| Metadados + Passagem de texto (40T) | 29,275 |
| Descrição + Metadados + Passagem de texto (40T) | 27,398 |
| Descrição + Passagem de texto (40T) | 24,687 |
| Descrição + Metadados | 19,367 |

**Tabela 3.12.** Resultados de MAP obtidos para as combinações de diversas fontes de evidência no arcabouço Bayesiano.

A Figura 3.4 apresenta as curvas de Precisão×Revocação para as quatro combinações analisadas. As abordagens de metadados+passagem de texto e metadados+descrição+passagem de texto apresentaram comportamentos similares, no entanto, a abordagem de metadados+passagem de texto apresentou valores de precisão mais altos até o nível de revocação de 80%. Acima dos 80%, a combinação de evidências de

metadados+descrição+passagem de texto foi ligeiramente melhor. Entretanto, é necessário dizer que valores altos de precisão são mais importantes em baixos níveis de revocação. Devido aos bons resultados obtidos pela abordagem de metadados+passagem de texto, esta abordagem será usada na comparações com o arcabouço de GP.

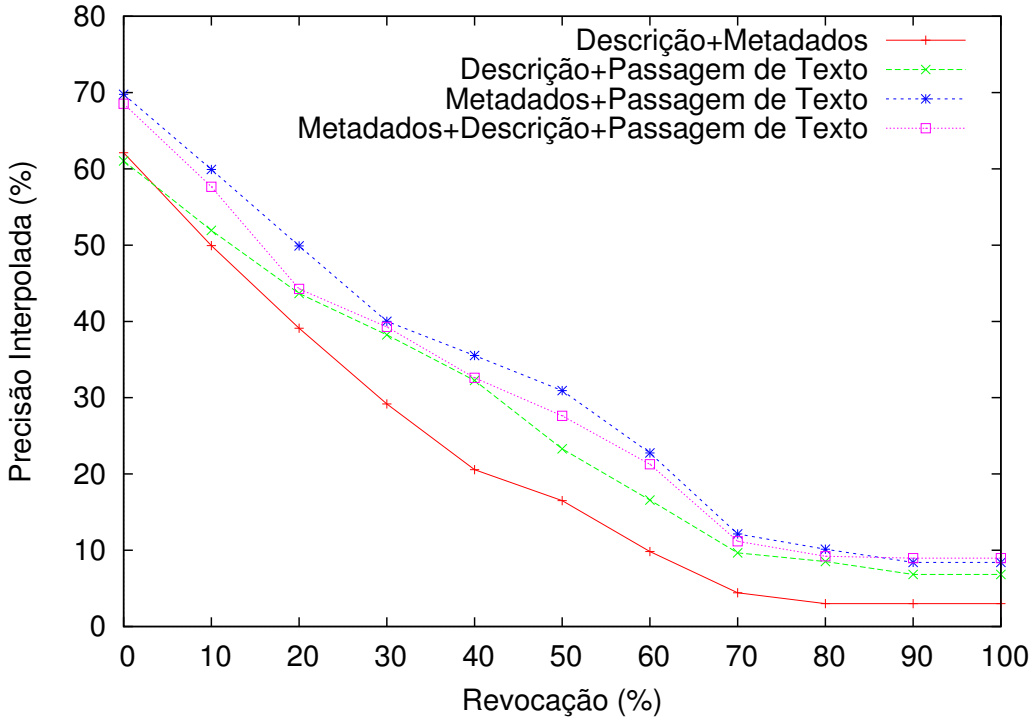


**Figura 3.4.** Curvas de Precisão×Revocação para as diversas fontes de evidência no arcabouço Bayesiano.

#### 3.3.3.2 Resultados Experimentais com o Arcabouço de GP

A Figura 3.5 apresenta a evolução do arcabouço de GP nas fases de treinamento, validação e teste ao longo de 30 gerações. Para cada geração, foram plotados os 20 melhores indivíduos de acordo com a função de aptidão utilizada. A figura mostra que apesar do fato dos conjuntos de treino, validação e teste apresentarem diferentes valores de aptidão, as curvas de validação e teste tendem a seguir o comportamento da curva de treino.

A Tabela 3.13 apresenta os valores de MAP obtidos pelo arcabouço de GP, Okapi-BM25 e a melhor estratégia de recuperação do arcabouço Bayesiano (metadados+passagem de texto). Pode-se observar que o arcabouço evolucionário foi capaz de melhorar os resultados de MAP obtidos pelo Okapi-BM25 de $34,79$ para $42,57$, um ganho de $22,36\%$. De acordo com o teste de significância Wilcoxon, os resultados obtidos pelo arcabouço de GP foram

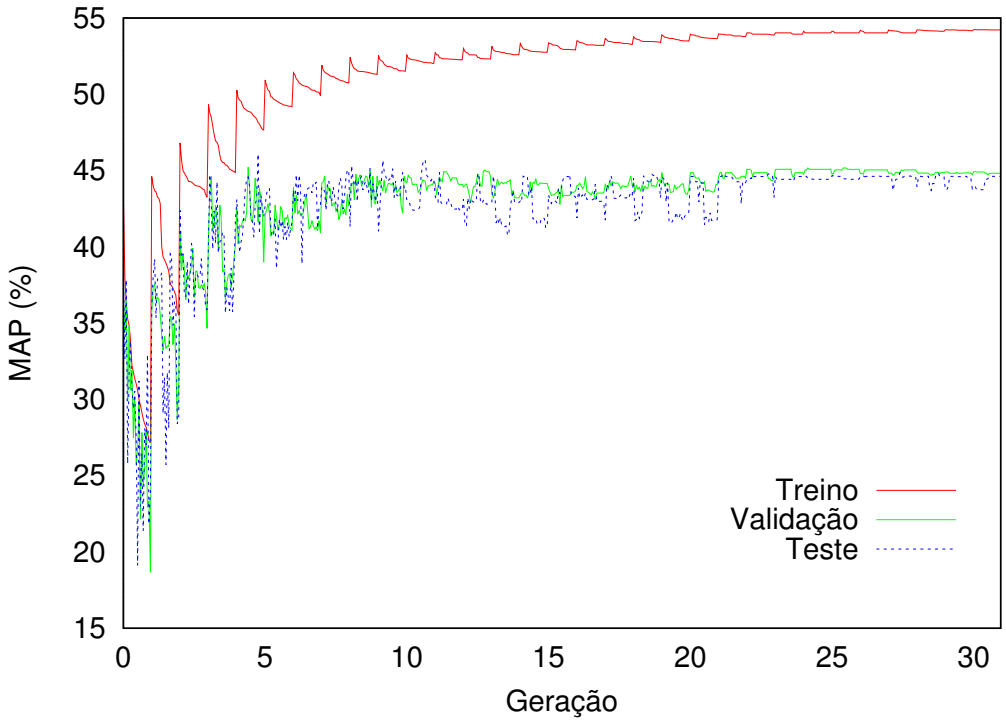


**Figura 3.5.** Curvas de evolução do arcabouço de GP para os melhores indivíduos em 30 gerações.

considerados estatisticamente significantes com um nível de confiança de 99% em relação ao arcabouço Bayesiano e 98% em relação ao Okapi-BM25.

| MEDIDA | BASELINES | | GP | | |
|---|---|---|---|---|---|
| | BM25 | Bayesiano | | Ganho sobre BM25 | Ganho sobre Bayesiano |
| P@10 | 45,00 | 36,60 | 48,00 | +6,67% | +31,15% |
| P@20 | 38,30 | 31,60 | 40,50 | +5,74% | +28,16% |
| P@30 | 35,00 | 27,33 | 37,00 | +5,71% | +35,37% |
| MAP | 34,79 | 29,28 | 42,57 | +22,36% | +45,39% |

**Tabela 3.13.** Resultados de MAP obtidos para os arcabouços baseados em GP, Okapi-BM25 e Modelo Bayesiano.

A Figura 3.6 apresenta as curvas de Precisão×Revocação obtidas pelo arcabouço de GP, Okapi-BM25 e a melhor estratégia de recuperação do arcabouço Bayesiano (metadados+passagem de texto). Observa-se que a abordagem de GP foi superior em todos os níveis de revocação.

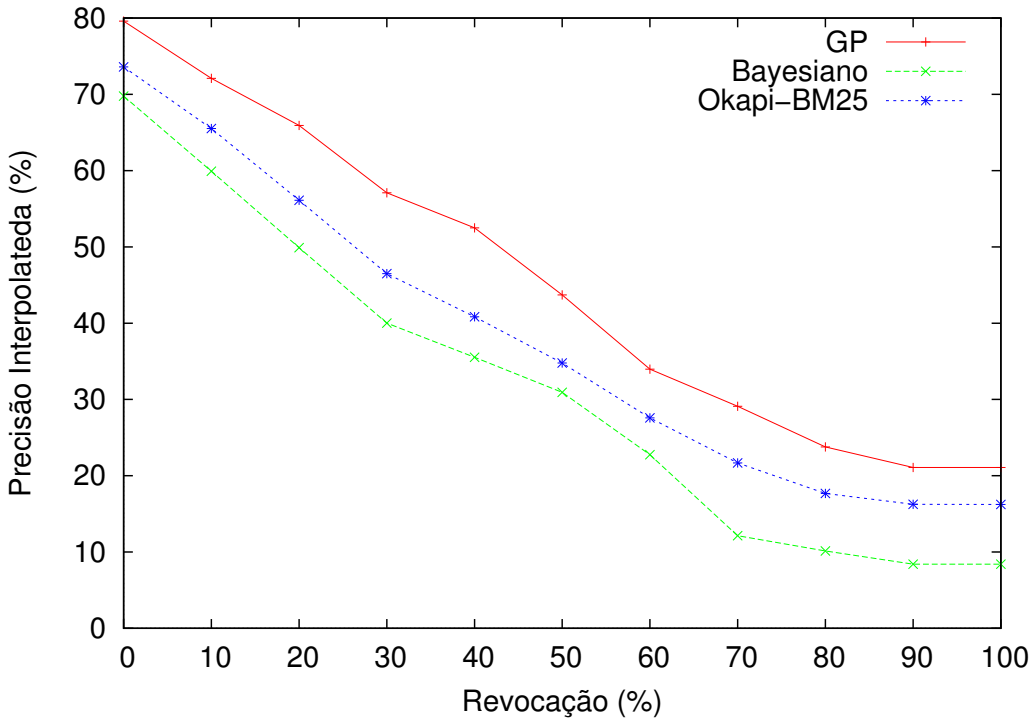


**Figura 3.6.** Curvas de Precisão×Revocação obtidas pelo arcabouço de GP, Okapi-BM25 e arcabouço Bayesiano.

### 3.3.4 Análise das Funções

Em cada execução do arcabouço de GP um único indivíduo é escolhido dentre todos os indivíduos de uma população ao final do processo evolucionário. Este indivíduo representa a melhor solução encontrada para o problema alvo no espaço de busca considerado durante aquela execução. As Equações 3.2 e 3.3 mostram exemplos de duas funções escolhidas como boas soluções pelo arcabouço de GP. Pode-se observar que as funções produzidas são, em geral, grandes e complexas.

$$\begin{aligned}&((((avgld\_psg20 * avgld\_psg10) + (tf\_psg40 * bm25\_psg40)) + \\ &((norma\_alt + bm25\_title) + ((avgld\_psg20 * avgld\_psg10) + (tf\_psg40 * bm25\_psg40)))) + \\ &((\sqrt{(bm25\_title + ((avgld\_psg20 * avgld\_psg10) + (bm25\_text + bm25\_title)))}) * \\ &(avgld\_psg20 * avgld\_psg10)) + (bm25\_text + (avgld\_psg20 * avgld\_psg10))))\end{aligned} \tag{3.2}$$

$$\begin{aligned}&((bm25\_src + bm25\_text) + (idf\_title/tf\_title)) + \\ &((((bm25\_src + (bm25\_src + avgld\_psg20)) + ((bm25\_src + bm25\_text) + bm25\_psg40)) + \\ &\sqrt{norma\_title}) + ((bm25\_src + bm25\_text) + avgld\_psg20))\end{aligned} \tag{3.3}$$

A análise das funções geradas pelo arcabouço de GP permite que se verifique a contribuição de cada evidência nas soluções descobertas. Pode-se saber, por exemplo, quais terminais estão sendo utilizados nas funções geradas e se alguma evidência foi descartada pelo arcabouço. A Tabela 3.14 apresenta estatísticas de ocorrência de terminais em 25 funções geradas pelo arcabouço de GP em execuções distintas do sistema. Pode-se observar que os terminais correspondentes ao texto completo, título da página e passagem de 40 termos foram as evidências que ocorreram com maior frequência, com participação em 72%, 84% e 52% das funções analisadas, respectivamente. O fato de ser uma coleção muito diversa, com páginas provenientes de diversos *sites* e com estruturas muito diferentes pode ter contribuído para que a evidência de texto completo prevalecesse em relação às passagens de texto. O fato do texto completo também incorporar todos os tamanhos de passagens contribuiu para que esta evidência tenha sido utilizada com mais frequência nas funções analisadas.

Os terminais correspondentes às tags SRC, texto de âncora e ALT figuraram em 36%, 24% e 12% das funções analisadas, respectivamente. Isto indica que, embora a participação dessas evidências tenha ocorrido em menor proporção, elas ainda contribuíram de alguma forma no processo de recuperação de imagens.

| EVIDÊNCIAS TEXTUAIS | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| Texto completo | 34 | 18 |
| Título | 31 | 21 |
| Passagem de texto (40T) | 18 | 13 |
| Passagem de texto (20T) | 13 | 8 |
| Tag SRC | 10 | 9 |
| Tag de âncora | 7 | 6 |
| Tag ALT | 5 | 3 |
| Passagem de texto (10T) | 5 | 4 |
| Passagem de texto (60T) | 3 | 2 |
| Passagem de texto (100T) | 3 | 1 |
| Passagem de texto (80T) | 2 | 1 |
| Autor | 0 | 0 |
| Keywords | 0 | 0 |
| Descrição | 0 | 0 |

**Tabela 3.14.** Total de ocorrências das evidências textuais nas funções analisadas.

A Figura 3.7 ilustra o número de ocorrências dos terminais referentes à cada evidência textual nas funções analisadas.

Nota-se que os terminais correspondentes às evidências de autor, descrição e palavras-chave foram descartados pelo arcabouço de GP, indicando que estes atributos não colaboraram

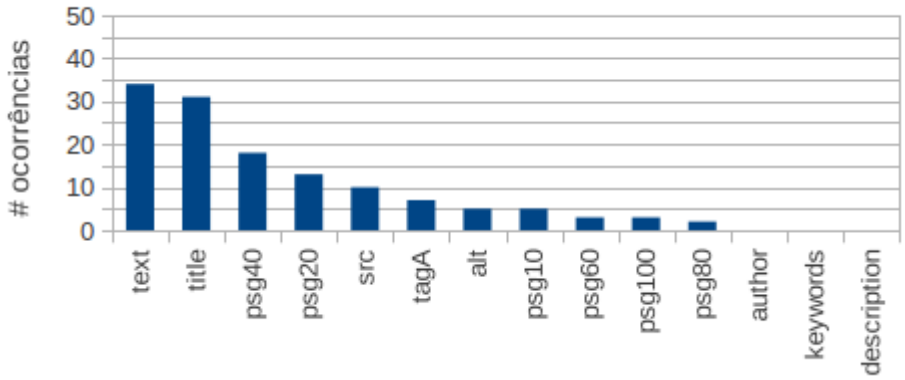


**Figura 3.7.** Frequência de ocorrência das evidências textuais.

para o processo de recuperação de imagens. O fato destes campos serem raramente preenchidos pode explicar tal comportamento. Além disso, estes atributos são utilizados para prover informações sobre a página Web, e portanto, estão mais relacionados à semântica da página e não necessariamente relacionados às imagens contidas nessa página.

# Capítulo 4

# Busca Visual em Comércio Eletrônico

O capítulo anterior apresenta um exemplo de como programação genética pode ser aplicada com sucesso na combinação de múltiplas fontes de evidências textuais para o desenvolvimento de sistemas de busca de imagens baseados em texto. Este capítulo discute a aplicação de GP sob um novo contexto. Mais especificamente, apresenta um arcabouço para expansão multimodal de consultas baseado em programação genética. O arcabouço proposto expande automaticamente uma imagem de consulta utilizando informação multimodal e computa um novo *ranking* de resultados baseado na consulta expandida.

Como aplicação alvo foi estudado o problema de busca visual de produtos em lojas de comércio eletrônico. São propostas quatro alternativas para expandir automaticamente a imagem de consulta. Para avaliar o desempenho das abordagens propostas, foram criadas duas coleções contendo imagens e informações de produtos de diferentes lojas do segmento de moda. Os produtos são peças de vestuário, como roupas, sapatos e acessórios. Experimentos realizados indicam que os métodos de expansão apresentados são uma alternativa efetiva para melhorar a qualidade dos resultados de busca de imagens.

Em resumo, as principais contribuições deste capítulo são: (i) um arcabouço para expansão multimodal de consultas baseado em em programação genética. São apresentadas quatro alternativas para expandir automaticamente uma imagem de consulta utilizando informação multimodal disponível nas coleções; (ii) uma avaliação experimental das abordagens propostas em uma aplicação de busca visual em lojas de comércio eletrônico do segmento de moda. (iii) produção do artigo *A Multimodal query expansion based on genetic programming for visually-oriented e-commerce applications* submetido para o *Expert Systems with Applications Journal*.

## 4.1 Motivação

Os recentes avanços tecnológicos têm contribuído para o surgimento de novas oportunidades para aplicações de CBIR. Este é o caso dos sistemas de busca de produtos em sites de comércio eletrônico. Tais sistemas tornaram-se um tópico de pesquisa importante em razão do grande crescimento no setor e das constantes demandas por melhorias na tecnologia de busca, com o objetivo de melhorar a experiência de compra dos usuários.

Estudos recentes revelam que o setor de comércio eletrônico apresenta grandes perspectivas de crescimento para os próximos anos. Particularmente, a venda de produtos no segmento de moda está entre os mercados que mais cresceram no último ano, com um crescimento registrado de 25%, segundo informações divulgadas pelo site *E-commerce News* [1]. No entanto, apesar das lojas *online* desse segmento comercializarem produtos com grande apelo visual para a decisão de compra, elas geralmente limitam seus consumidores a buscar produtos utilizando apenas descrições textuais. Neste contexto, valer-se de recursos visuais para buscar produtos tornou-se uma característica importante para apoiar o processo de compras *online*. Isso porque, neste domínio específico, utilizar uma imagem para buscar um produto é muito mais descritivo que uma consulta textual e pode ajudar o usuário a encontrar produtos mais similares ao que está procurando. Dessa forma, técnicas de CBIR podem ser aplicadas para dar suporte à busca de produtos em tais aplicações.

Nos sites de comércio eletrônico em geral, cada produto é apresentado por meio de uma imagem, uma descrição textual, um preço e está classificado em uma categoria, como, por exemplo, eletrônicos, eletrodomésticos, brinquedos, etc. No segmento de moda, algumas das categorias definidas são roupas, calçados, bolsas e acessórios. Este conjunto de atributos proporciona uma importante e rica fonte de informação que pode ser utilizada para melhorar os resultados da busca.

Dentro do contexto de recuperação de imagens, quando existe uma base com evidências textuais e visuais disponíveis, a utilização de uma estratégia de recuperação multimodal é apontada como uma direção promissora. O propósito desse tipo de estratégia é tentar tirar vantagem da riqueza de informação presente nas características visuais da imagem e da semântica oferecida pelo conteúdo textual. Porém, o grande desafio está em definir uma maneira de combinar as abordagens de recuperação textual e recuperação visual.

Quanto mais genérico for o domínio da aplicação, mais difícil é definir uma estratégia multimodal adequada que funcione bem para todos os cenários. Por outro lado, em aplicações de domínio específico, a semântica da busca está presente em um contexto limitado, permitindo assim que as características visuais e textuais sejam exploradas de forma mais eficiente.

[1] `http://http://ecommercenews.com.br/tag/moda`. Em 01/01/2014.

Este capítulo aborda o problema da busca de produtos em sites de comércio eletrônico voltados para o segmento de moda. Mais especificamente, a busca visual de produtos na qual o usuário submete apenas uma imagem como consulta ao sistema. Um cenário comum para este tipo de aplicação é, por exemplo, quando um usuário tira uma foto de um produto de seu interesse e quer buscar por produtos similares ao da foto em lojas de vendas *online*. Este tipo de consulta é bastante útil quando o usuário está buscando por produtos de vestuário, onde a apresentação visual é essencial para a decisão de compra.

Uma característica importante que torna esta aplicação alvo diferente de aplicações de busca de imagens tradicionais é o objetivo final da tarefa de busca. Aqui, a consulta é uma imagem que representa um produto, mas o que torna um resultado relevante ou não para o usuário, pode estar codificado em outros atributos que não estão representados visualmente. Exemplos deste problema incluem diferenças na maneira em como aos fotos dos produtos são produzidas, tais como camisetas dobradas ou desdobradas, ou ainda quando produtos com cores e texturas diferentes podem ser considerados relevantes, em alguns casos, devido ao estilo.

Outro ponto que torna este problema desafiador é a ausência de informação textual na consulta inicial. Diferentemente de outros trabalhos, a consulta aqui é apenas uma imagem. Apesar das restrições impostas, a solução proposta permite encontrar produtos relevantes em relação a uma imagem de consulta mesmo quando a apresentação do produto não é similar à consulta. Isto porque a abordagem expande a consulta inicial com informações sobre a categoria inferida da imagem de consulta e o conteúdo textual associado a outras imagens similares.

A ideia principal é usar a informação visual da imagem de consulta para produzir um *ranking* inicial e então extrair informação adicional dos resultados para expandir automaticamente a consulta inicial. O uso de uma abordagem de expansão totalmente automática foi escolhido por não exigir esforço adicional dos usuários para realimentar o sistema de busca com informação de relevância. Apesar de não descartar a possibilidade do uso de técnicas de realimentação de relevância como um trabalho futuro, esta abordagem pode ser considerada inconveniente pelos usuários devido ao processo de refinamento da consulta através de sucessivas interações, o que pode desestimular o usuário de uma aplicação de comércio eletrônico.

O arcabouço proposto aplica os princípios da programação genética para realizar tanto a expansão da consulta inicial quanto a produção de um novo *ranking* baseado na consulta expandida. São analisadas quatro alternativas de usar GP para derivar métodos de expansão multimodal a partir de imagens de consultas. GP é utilizado para encontrar a melhor combinação multimodal possível através das fontes de evidências disponíveis. Até onde se sabe, GP nunca foi aplicado no cenário de busca abordado neste capítulo, no qual somente

a informação visual está disponível para expansão multimodal de imagens de consulta. O desafio de explorar somente aspectos visuais advém da dificuldade de mapear características de baixo-nível para conceitos de alto-nível encontrados nas imagens, um problema conhecido como *gap* semântico (Liu et al., 2007).

A próxima seção apresenta um visão geral do arcabouço de GP para busca visual através da expansão automática da imagem de consulta utilizando informação adicional extraída do topo do resultado inicial.

## 4.2 Arcabouço de GP para Busca Visual em Comércio Eletrônico

O impacto causado pela aplicação de técnicas de CBIR na busca de produtos em sistemas de comércio eletrônico, tais como os voltados para o segmento de moda, é potencialmente grande. Nesses domínios, o uso exclusivo de consultas textuais para buscar um produto geralmente leva a resultados ruins ao produzir muitos falsos positivos. Além disso, uma imagem de exemplo é geralmente muito mais descritiva que uma consulta textual, pois através da imagem é possível fornecer detalhes de estilo visual que são difíceis de descrever através de uma consulta textual.

Assim como no trabalho sobre busca de imagens baseada em informação textual apresentado no Capítulo 3, o arcabouço de GP é basicamente um processo iterativo de duas fases: treino e validação. Para cada fase, são selecionados um conjunto de consultas e documentos da coleção, que são chamados conjunto de treino, para a fase de treino, e conjunto de validação, para a fase de validação.

A função de aptidão neste arcabouço de GP para expansão multimodal é modelada para medir a qualidade no topo do *ranking* que se deseja otimizar. Os indivíduos evoluem, geração após geração, através de operações genéticas de reprodução, cruzamento e mutação, criando novas populações de indivíduos até que um critério de terminação seja satisfeito.

Cada indivíduo representa uma função de *ranking* que atribui um valor de similaridade para cada imagem da coleção para uma determinada consulta. A programação genética é aplicada para realizar tanto a expansão da imagem de consulta inicial quanto a produção de um novo *ranking* baseado na consulta expandida.

São apresentadas a seguir quatro alternativas de usar GP para derivar métodos de expansão multimodal a partir de imagens de consultas. A Figura 4.1 ilustra duas destas alternativas.

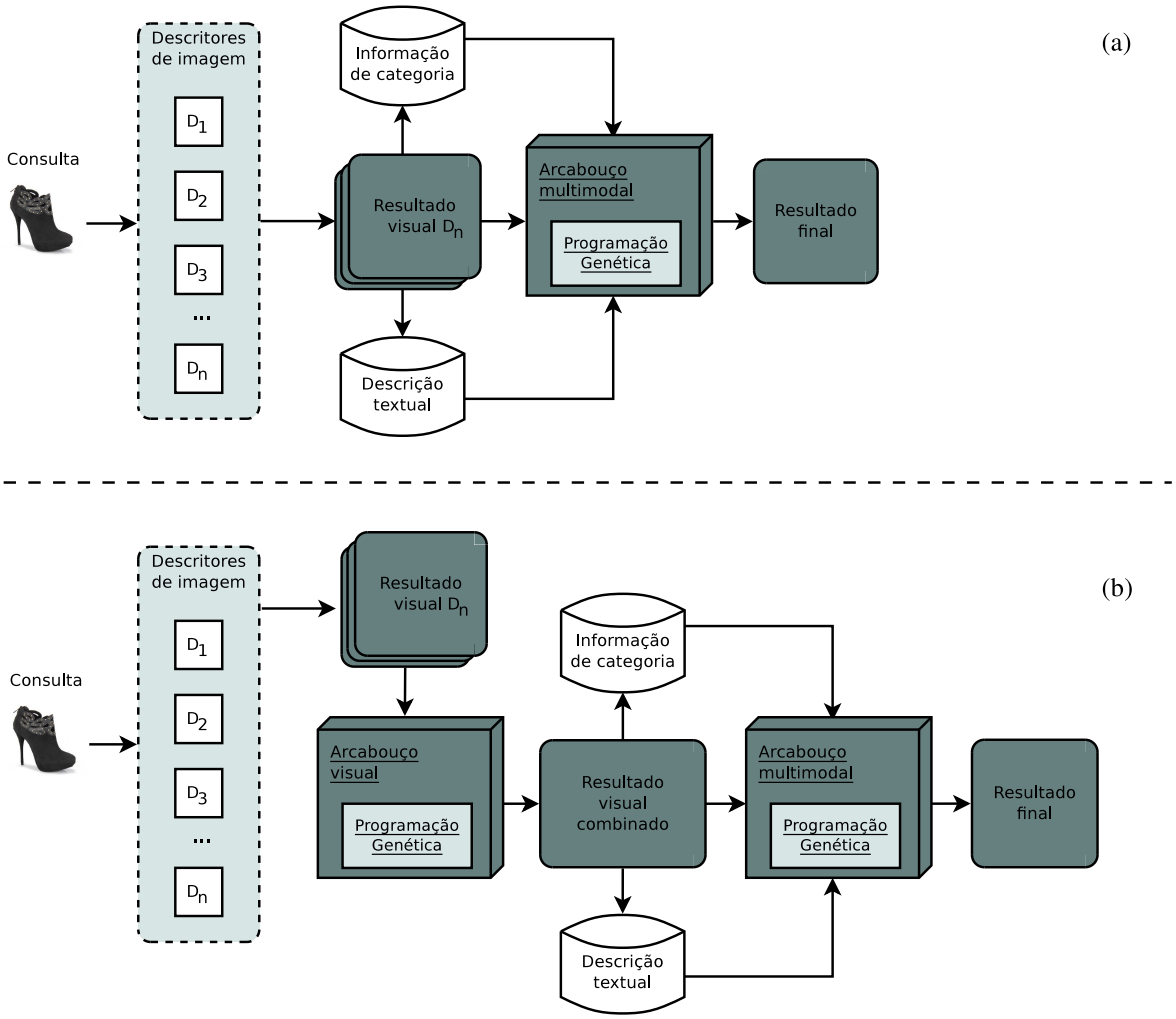


**Figura 4.1.** Visão geral de dois métodos de expansão denominados *Expansão-GPI* (a) e *Expansão-GPC* (b).

***Expansão-GPI*** : este método expande a imagem de consulta com informação multimodal adicionando características textuais e de categoria obtidas dos resultados de cada descritor visual adotado. A Figura 4.1(a) apresenta os principais passos do método *Expansão-GPI*, onde utiliza-se como terminais não somente as características visuais da imagens, mas também a informação adicional extraídas a partir delas.

***Expansão-GPC*** : este método de expansão extrai informação multimodal dos resultados obtidos a partir da combinação de todas as propriedades visuais da imagem. A combinação é derivada utilizando-se um arcabouço de GP apenas com propriedades visuais, denominado aqui de *Visual-GP*. Ao invés de se ter diversos *rankings* iniciais, um para cada descritor visual, tem-se um único *ranking* de onde se extrai as características multimodais. Espera-se que o *Visual-GP* forneça um resultado melhor que o das ca-

racterísticas visuais individuais. Com base em tal hipótese, o método *Expansão-GPC* foi considerado para verificar se o *ranking* inicial provido pelo *Visual-GP* pode resultar em uma expansão multimodal de melhor qualidade. A Figura 4.1(b) ilustra os passos principais do método *Expansão-GPC*. Pode-se observar que na abordagem *Expansão-GPC*, a programação genética é utilizada duas vezes, uma para combinar as características visuais e prover um *ranking* inicial, e outra para realizar a expansão multimodal propriamente dita. O objetivo aqui é investigar se a expansão a partir de um *ranking* inicial mais robusto pode gerar resultados melhores do que uma expansão feita a partir de vários *rankings* individuais, menos efetivos, mas que representam diferentes formas de medir a similaridade entre imagens.

***Rerank-GPI*** : este método é similar à alternativa *Expansão-GPI*. No entanto, esta abordagem utiliza a expansão multimodal apenas para reordenar os resultados do topo do *ranking* provido pelas características visuais individuais. Considera-se a união dos resultados dos k resultados do topo como resposta inicial a ser reordenada. Nos experimentos foi utilizado k=100. Observa-se que resultados que não estão presentes no topo das respostas de acordo com as características visuais individuais, não são introduzidos nesta abordagem, enquanto que este fenômeno pode ocorrer no método *Expansão-GPC*. A comparação entre alternativas que permitem a entrada de novos resultados na resposta e alternativas que apenas reordenam os resultados iniciais, *Expansão-GPI versus Rerank-GPI*, é útil para verificar se a expansão multimodal pode trazer novas respostas relevantes, mesmo que a imagem associada a essas respostas não seja em princípio considerada próxima à imagem de consulta.

***Rerank-GPC*** : Este método é similar à abordagem *Expansão-GPC*. No entanto, utiliza apenas a informação multimodal para reordenar os k resultados do topo do resultado provido pelo *Visual-GP*. Nos experimentos também foi utilizado k=100. Novamente, deseja-se saber com as duas alternativas, *Rerank-GPC versus Expansão-GPC*, se há vantagens na inclusão de novos resultados na resposta inicial além das obtidas com características visuais, ou se é melhor apenas reordenar as respostas obtidas inicialmente.

### 4.2.1 Descritores de Imagens

Existem diversos descritores de imagens disponíveis na literatura, cada um com seus respectivos pontos fortes e fracos. Descritores de imagens são utilizados para caracterizar propriedades visuais das imagens, tais como cor, forma, textura ou uma combinação destas propriedades, e representar tal informação por meio de vetores de características. Na medida

em que estas propriedades são extraídas, elas podem ser utilizadas por sistemas de CBIR para recuperar imagens similares a uma dada imagem de consulta de acordo com as propriedades representadas pelos vetores de características.

Dado que o problema estudado neste capítulo é, inicialmente, uma tarefa de CBIR, o desempenho de diversos descritores de imagens foi avaliado nas coleções de imagens adotadas. Os descritores de imagens foram utilizados para extrair as propriedades visuais e calcular a distância entre a imagem de consulta e as imagens da coleção. Foram utilizados os seguintes descritores visuais: CEDD (Chatzichristofis & Boutalis, 2008a), FCTH (Chatzichristofis & Boutalis, 2008b), CLD (Kasutani & Yamada, 2001), JCD (Chatzichristofis et al., 2009), ACC (Huang et al., 1997), GCH (Lux, 2011), BIC (Stehling et al., 2002), SDLC (Vidal et al., 2012), PHOG (Bosch et al., 2007), SIFT Lowe (1999) e CSIFT (Abdel-Hakim & Farag, 2006). Esses descritores foram escolhidos por figurarem entre os descritores mais utilizados na literatura, por estarem disponíveis em bibliotecas de software para o processamento de imagens e por explorarem diferentes características visuais para representar imagens. A Tabela 4.1 apresenta as propriedades visuais exploradas por cada descritor utilizado nos experimentos.

| Descritor de Imagem | Propriedade Visual |
|---|---|
| GCH (*Global Color Histogram*) | Cor |
| ACC (*Auto Color Correlogram*) | Cor |
| CLD (*Color Layout Descriptor*) | Cor |
| PHOG (*Pyramid Histogram of Oriented Gradients*) | Forma |
| CEDD (*Color and Edge Directivity Descriptor*) | Cor e Textura |
| FCTH (*Fuzzy Color and Texture Histogram*) | Cor e Textura |
| JCD (*Joint Composite Descriptor*) | Cor e Textura |
| BIC (*Border/Interior pixel Classification*) | Cor |
| SDLC (*Sorted Dominant Local Color*) | Cor |
| SIFT (*Scale Invariant Feature Transform*) | Pontos de Interesse |
| CSIFT (*Colored Scale Invariant Feature Transform*) | Cor e Pontos de Interesse |

**Tabela 4.1.** Descritores de imagens utilizados nos experimentos.

### 4.2.2 Indivíduos

Um indivíduo GP é representado por uma combinação de funções e terminais, organizados em uma estrutura de árvore binária. Para combinar os terminais na formação dos indivíduos, foram utilizadas as funções de adição (+), subtração (-), multiplicação (*), divisão (/), logaritmo natural (log), logaritmo na base-10 (log10), máximo (max), mínimo (min), e raiz quadrada

($\sqrt{}$). Terminais, ou nodos folhas, contêm informações derivadas de evidências visuais, textuais e de categoria dos produtos.

Geralmente as bases de produtos disponíveis em lojas de comércio contêm outros dados complementares à imagem do produto, tais como uma descrição textual sobre o produto e uma categoria a qual ele pertence. A questão chave aqui é investigar se o processo de aprendizagem pode tirar vantagem destas informações complementares para melhorar a qualidade do resultado se comparado a sistemas que utilizam somente características visuais para computar o *ranking* final.

O primeiro passo para responder esta questão é determinar uma maneira de associar informação multimodal com a imagem de consulta. O objetivo aqui é propor e avaliar estratégias que não requerem intervenção do usuário para fornecer qualquer informação extra sobre sua necessidade de informação além da imagem de consulta. Portanto, a associação deve ser feita de modo automático, sem qualquer intervenção explícita como realimentação de relevância.

Para abordar este problema, são investigadas 88 características extraídas das descrições textuais e das categorias dos produtos e 22 terminais associados às propriedades visuais. Além destes terminais foram também utilizados valores constantes na faixa de [0..100] como terminais no GP.

### Informação Visual das Imagens mais Similares

Para cada um dos 11 descritores de imagens apresentados na na Tabela 4.1, um valor associado ao descritor é utilizado como terminal no arcabouço de GP. O valor do terminal gerado por cada descritor visual é dado pela função $\delta_d$, que computa a distância entre a imagem de consulta e uma determinada imagem da coleção. Além disso, o valor de distância mínima obtido por cada descritor também é usada como terminal no GP, somando assim mais 11 terminais extras. Apesar da informação de distância mínima atribuir, dada uma consulta, um valor igual para todas as imagens da coleção, tal constante por ser útil no processo de aprendizagem. Esse valor pode, por exemplo, ser usado como fator de normalização de valores entre as consultas, dado que as distâncias podem variar muito entre consultas, facilitando assim a seleção de boas fórmulas genéricas.

### Informação de Categoria das Imagens mais Similares

As categorias dos produtos encontrados no topo dos resultados visuais obtidos por cada descritor apresentado na Seção 4.2.1 são associadas à imagem de consulta. Para cada descritor, é computada a frequência de todas as categorias de produtos encontradas no topo dos seus

respectivos resultados. Desta forma é atribuído, a cada produto da coleção, a frequência da sua categoria como um terminal do processo de aprendizagem.

A Figura 4.2 apresenta um exemplo de uma imagem de consulta e os 5 resultados mais similares na coleção *Amazon* de acordo com um descritor visual. Como se pode observar, quatro dos produtos mais similares pertencem à categoria de roupas femininas, e um pertence à categoria de roupas masculinas. Neste exemplo, os resultados produzem um terminal de frequência da categoria que atribui os valores 4 e 1 para estas duas categorias, e 0 para as demais categorias. Dessa forma, produtos pertencentes àquelas categorias que aparecem no topo têm maior chance de serem promovidos no *ranking*.

| Consulta | 5 resultados do topo | |
|---|---|---|
|  |  | descrição: guess by marciano shauna mesh halter **dress**<br>categoria: clothing and accessories - women |
| |  | descrição: kenneth cole women fluttery squares **dress**<br>categoria: clothing and accessories - women |
| |  | descrição: tiana women fun pick up **dress**<br>categoria: clothing and accessories - women |
| |  | descrição: london times women matte jersey mesh shutter tuck **dress**<br>categoria: clothing and accessories - women |
| |  | descrição: rochester big and tall non iron pleated **shorts**<br>categoria: clothing and accessories - men |

**Figura 4.2.** Exemplo de uma imagem de consulta e os cinco resultados mais similares de na coleção *Amazon* acordo com um descritor visual.

A frequência da categoria é computada levando-se em consideração o topo com k=1, 5, 10, 20 elementos. Como são considerados 11 descritores de imagens e 4 formas alternativas de considerar o topo dos resultados, um total de 44 terminais com informação de categoria são utilizados nos experimentos.

### Informação Textual das Imagens mais Similares

Para cada descritor de imagem adotado, uma consulta textual é formada com a concatenação dos termos presentes na descrição textual dos produtos mais similares. De fato, somente um segmento da descrição do produto é considerado, baseado nos resultados apresentados em (dos Santos et al., 2013) que indicam que considerar até três termos da descrição é melhor

que considerar a descrição toda. Nos experimentos realizados, o tamanho do segmento de texto adotado foi de 1 termo apenas. Este tamanho foi escolhido devido aos bons resultados obtidos em (dos Santos et al., 2013) e também por diminuir o custo computacional requerido pelo método. Isto significa que no caso da coleção *DafitiPosthaus*, cujas descrições são apresentadas em Português, o primeiro termo presente na descrição do produto foi utilizado para gerar a consulta textual. Enquanto que na coleção *Amazon*, cujas descrições são apresentadas em Inglês, foi considerado o último termo. Esta diferença é devido à estrutura gramatical de cada idioma.

A Figura 4.2 pode ser novamente usada para ilustrar este processo. Os produtos mais similares à imagem de consulta são utilizados para criar uma consulta textual. Neste exemplo, a consulta textual é formada por quatro ocorrências da palavra "*dress*"e uma ocorrência da palavra "*shorts*". A consulta textual gerada é comparada às descrições dos produtos obtendo um valor de similaridade entre a consulta e a descrição textual de cada produto. A similaridade textual é computada utilizando o modelo de espaço vetorial (McGill & Salton, 1983) e usada como terminal no GP.

De maneira similar à abordagem utilizada no cálculo da frequência da categoria, foi levado em consideração o topo com k=1, 5, 10, 20 elementos. Como são considerados 11 descritores de imagens e 4 formas alternativas de considerar o topo dos resultados, um total de 44 terminais com informação textual são utilizados nos experimentos.

### 4.2.3 Função de Aptidão

A função de aptidão adotada no arcabouço evolucionário deve ser capaz de avaliar o resultado produzido por cada indivíduo em uma população de indivíduos. O resultado da função de aptidão será usado no arcabouço como parâmetro para seleção dos indivíduos. Como os indivíduos representam esquemas de ponderação a serem utilizados em uma função de ordenação de documentos (imagens), a função de aptidão deve medir a qualidade do *ranking* produzido por um determinado indivíduo. A medida P@10 apresentada na Equação 2.6 foi definida como função de aptidão no arcabouço de GP.

### 4.2.4 Seleção do Melhor Indivíduo

A escolha de um indivíduo como solução final do arcabouço de GP deve levar em consideração o desempenho do indivíduo obtido nas fases de treino e de validação, de modo a escolher um indivíduo capaz de ser genérico o suficiente para obter bons resultados em consultas não conhecidas e não escolher um indivíduo muito especializado para o conjunto de treino. Desta forma, o critério utilizado para escolha dos melhores indivíduos é baseado nos valores de

aptidão obtidos nas duas fases e no desvio padrão correspondente a esses dois valores, como proposto em (de Almeida et al., 2007). Este método é conhecido como $\text{SUM}_\sigma$, já apresentado na Equação 3.1. O indivíduo com maior valor de $\text{SUM}_\sigma$ é escolhido como melhor.

## 4.3 Experimentos

Esta seção apresenta os experimentos realizados para avaliar as estratégias de busca visual propostas neste capítulo, bem como as coleções adotadas e a configuração dos parâmetros do GP. Para extrair as propriedades visuais das imagens das coleções, foram utilizadas as bibliotecas LIRE (Lux, 2011), JFeature (Graf, 2012) e VLFeat (Vedaldi & Fulkerson, 2010). Os experimentos foram realizados utilizando-se a estratégia de validação cruzada de *5-folds*. Cada conjunto de consulta foi dividido na proporção de 40% para treino, 40% para validação e 20% para teste em cada *fold*.

O arcabouço de recuperação de imagens baseado em GP foi implementado com suporte da lil-gp (v1.1) [2], uma biblioteca muito utilizada para o desenvolvimento eficiente de aplicações baseada em programação genética em linguagem C.

O processo de execução do GP depende da seleção de uma semente aleatória para gerar a população inicial. Para reduzir o efeito desta aleatoriedade o processo completo foi executado 10 vezes e foi escolhido o melhor indivíduo no treino e na validação dentre todas as execuções para ser aplicado na fase de teste. Esta estratégia de repetição de execuções do GP foi proposta inicialmente em (da Costa Carvalho et al., 2012) e diminui as chances de uma única semente levar a um desempenho abaixo do desempenho médio. Em da Costa Carvalho et al. (2012) também foram utilizadas 10 repetições e tal número foi suficiente para reduzir as chances de se obter um resultado muito positivo, ou muito negativo, devido à escolha de uma boa, ou de uma má, semente para gerar a população inicial do processo evolutivo.

Finalmente, P@10 e MAP foram as métricas adotadas para avaliar a efetividade dos métodos nos experimentos. Foi aplicado o teste de significância estatística *t-test* para verificar se as diferenças nos resultados obtidos são significativas. Foram considerados ganhos significativos de um método sobre outro, aqueles cujos valores foram iguais ou superiores a 95%.

### 4.3.1 Coleções de Imagens

As coleções de imagens disponíveis na literatura para realizar experimentos de busca visual de produtos são compostas unicamente de imagens, sem qualquer informação adicional associada. É o caso das coleções *Stanford Mobile Visual* (Chandrasekhar et al., 2011), que

[2] `http://garage.cse.msu.edu/software/lil-gp/` Em 01/01/2014

contém basicamente imagens de capas de livros e CDs/DVDs e PI100 (Xie et al., 2008), que contém imagens de vários produtos, tais como instrumentos musicais, brinquedos, utensílios, etc.

Em razão da ausência de uma coleção com informação multimodal para ser utilizada no escopo deste trabalho, foram montadas duas coleções contendo produtos de vestuário comercializados em três lojas de comércio eletrônico. Estas coleções foram apresentadas inicialmente em (dos Santos et al., 2013). A primeira coleção, denominada *DafitiPosthaus*, contém dados de 23.154 produtos coletados de duas lojas de venda *online* de vestuário no Brasil, Dafiti[3] e Posthaus[4]. A segunda coleção, denominada *Amazon*[5], contém dados de 12.807 produtos de uma loja internacional de comércio eletrônico. Em ambas as coleções, cada produto é representado por uma imagem, uma descrição textual e uma categoria. A informação de categoria foi extraída diretamente das lojas de comércio eletrônico coletadas para criar as coleções. Os produtos pertencem a uma das seis categorias: roupas femininas, roupas masculinas, calçados femininos, calçados masculinos, bolsas e acessórios. A Tabela 4.2 apresenta alguns dados sobre a porcentagem de produtos em cada categoria nas coleções adotadas.

| Categorias | *DafitiPosthaus* (%) | *Amazon* (%) |
|---|---|---|
| Roupas masculinas | 18,31 | 6,42 |
| Roupas femininas | 39,37 | 34,78 |
| Bolsas | 5,47 | 12,42 |
| Calçados femininos | 27,34 | 44,12 |
| Calçados masculinos | 7,34 | 1,89 |
| Acessórios | 2,17 | 0,37 |

**Tabela 4.2.** Informações sobre as categorias nas coleções *DafitiPosthaus* e *Amazon*.

Para a realização dos experimentos foi utilizado um total de 100 imagens de consulta, divididas em dois conjuntos descritos a seguir:

[3] `http://www.dafiti.com.br` Em 01/03/2014
[4] `http://www.posthaus.com.br` Em 01/03/2014
[5] `http://www.amazon.com` Em 01/03/2014

**Conjunto 1 (Q1):** composto por 50 imagens selecionadas de outros sites de comércio eletrônico diferentes daqueles coletados na montagem das duas coleções adotadas. As imagens nesse caso possuem características visuais similares às imagens contidas nas coleções, ou seja, imagens de produtos de moda com fundo homogêneo. A Figura 4.3 ilustra as imagens de consulta do conjunto Q1.

**Figura 4.3.** Imagens de consulta do conjunto Q1.

**Conjunto 2 (Q2):** composto por 50 imagens selecionadas de diferentes sites como blogs, revistas e jornais eletrônicos. As imagens são, em geral, fotos de pessoas famosas e representam uma classe de consultas difíceis, porém relevantes para usuários que estão em busca de um produto similar. O que torna este conjunto diferente do conjunto Q1 é a grande presença de ruído no fundo da imagem. A Figura 4.4 ilustra as imagens de consulta do conjunto Q2.

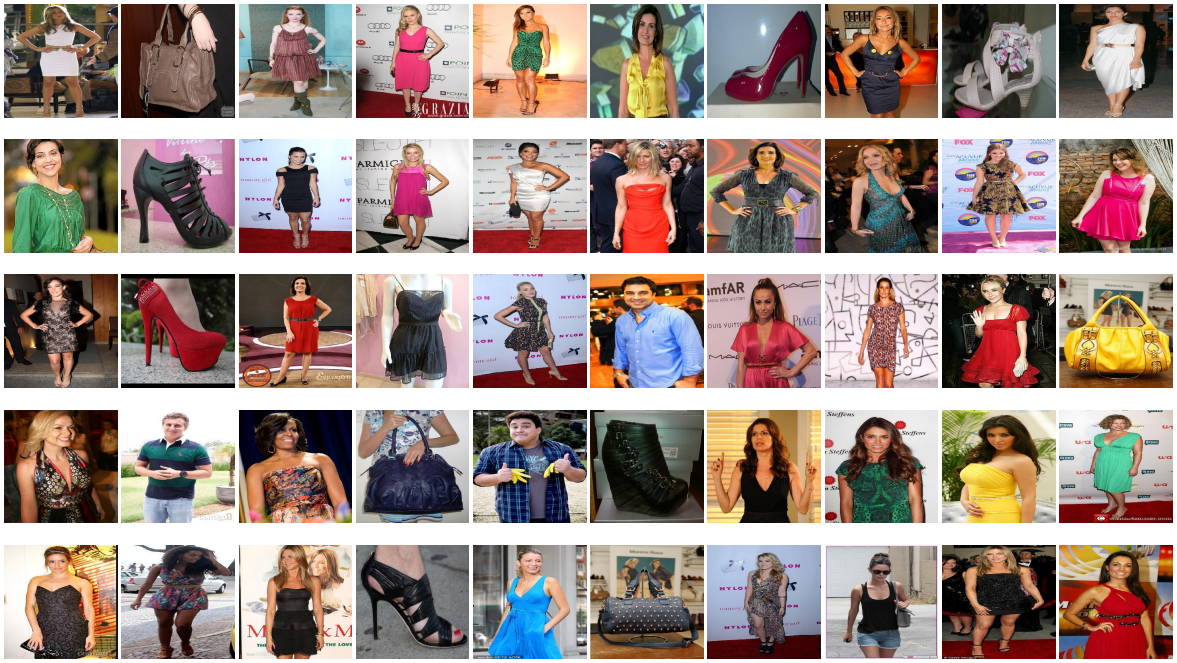

**Figura 4.4.** Imagens de consulta do conjunto Q2.

De forma a avaliar a relevância das respostas retornadas por cada método, solicitamos a 30 voluntários, divididos em 10 grupos de três pessoas, para fornecer um julgamento binário de relevância (relevante ou irrelevante) para cada resultado de cada imagem de consulta. Foram consideradas como relevantes, as respostas que receberam classificação relevante por, pelo menos, 2 avaliadores.

A técnica de *pooling* foi utilizada para avaliar as 25 imagens do topo retornadas por uma série de descritores de imagens para cada consulta. Para os experimentos realizados, foram incluídos novos julgamentos de relevância, de forma a avaliar todos os descritores de imagem apresentados na Tabela 4.1, uma vez que alguns deles não estavam incluídos no trabalho de (dos Santos et al., 2013). O número médio de produtos considerados relevantes por consulta é de $55,5$ e $35,5$ para o conjunto Q1 para as coleções *DafitiPosthaus* e *Amazon*, respectivamente. No conjunto Q2, o número médio de produtos considerados relevantes é de $15,2$ e $22,8$ para as coleções *DafitiPosthaus* e *Amazon*, respectivamente.

### 4.3.2 *Baselines*

Três abordagens distintas foram incluídas como *baselines* nos experimentos realizados para avaliar a efetividade das estratégias propostas neste capítulo.

A primeira abordagem, denominada *Visual-GP*, é uma versão do arcabouço de GP que considera apenas as propriedades visuais da imagem utilizando os descritores já mencionados na Seção 4.2.1. Os valores dos parâmetros utilizados na configuração do arcabouço são os mesmos apresentados na Seção 4.3.3.

A segunda abordagem, denominada *Total Recall*, é um método de expansão proposto em (Chum et al., 2007) que adota o modelo de *bag-of-visual-words* e realiza a expansão da imagem de consulta no resultado inicial. A imagem de consulta é submetida e o resultado inicial é reordenado através de um passo de verificação geométrica para suprimir resultados falso positivos do topo do resultado. Uma nova consulta é construída por uma média dos resultados verificados da consulta original. O método *Total Recall* foi implementado aplicando o descritor SIFT, como proposto em (Chum et al., 2007) e também aplicando o descritor CSIFT. Como esperado, a melhor alternativa para o problema abordado neste capítulo foi aquela utilizando o CSIFT, e portanto apenas estes resultados são reportados nos experimentos. O tamanho do vocabulário utilizado no modelo de *bag-of-visual-words* foi 50.000 palavras. Para o passo de verificação geométrica foi utilizado o algoritmo LO-RANSAC (Lebeda et al., 2012), o qual é o que apresenta os melhores resultados de acordo com a literatura.

Finalmente, a terceira abordagem, denominada TCatBR, é um método *ad-hoc* proposto em (dos Santos et al., 2013) que realiza uma reordenação dos $k_1$ resultados do topo de um descritor visual (CEDD). O método associa a categoria mais frequente encontrada no topo

$k_2$ do resultado à imagem de consulta e considera a descrição textual dos $k_3$ resultados que pertencem a esta categoria para realizar uma consulta textual. Os resultados da consulta textual são então combinados linearmente com o resultado visual para reordenar os $k_4$ resultados do topo. Os valores das variáveis $k_1$, $k_2$, $k_3$, e $k_4$ são os mesmos utilizados no trabalho de (dos Santos et al., 2013), ou seja, $k_1 = 100$, $k_2 = 20$, $k_3 = 20$, e $k_4 = 100$.

### 4.3.3 Configuração dos Parâmetros de GP

Para ajudar na configuração dos parâmetros de GP, foi adotada a mesma técnica de projeto experimental aplicada no Capítulo 3. Um projeto fatorial completo em dois níveis foi realizado para investigar o impacto dos três parâmetros principais e suas interações: tamanho da população (Fator A), número máximo de gerações (Fator B) e a profundidade máxima da árvore para representação dos indivíduos (Fator C). Os fatores A, B e C foram analisados nos níveis mínimos de 50, 5 e 4, e nos níveis máximos de 500, 50, e 12, respectivamente. Os efeitos de cada fator e suas respectivas interações são apresentados na Tabela 4.3.

| FATOR | EFEITO(%) |
|---|---|
| A | 31,51 |
| B | 15,80 |
| C | 15,38 |
| BC | 10,08 |
| ABC | 8,55 |
| AB | 5,68 |
| AC | 4,72 |

**Tabela 4.3.** Resultado do projeto fatorial completo em dois níveis para o arcabouço de GP (em ordem decrescente de efeito).

Novamente pode-se observar que o fator A, que representa o tamanho da população, tem maior influência no desempenho do arcabouço de GP e explica $31,51$ na variação da resposta. Os fatores B e C tiveram praticamente o mesmo efeito, em torno de 15%. O fator A foi cerca de 99,5% maior que o fator B e 104,8% maior que o fator C. Erros experimentais ou não-observados foram responsáveis por cerca de $8,3\%$ da variação na resposta. Este resultado indica que a escolha de uma tamanho de população grande é importante para obter bons resultados neste cenário.

Como resultado, adotou-se um tamanho de população de 500 indivíduos, 40 gerações e uma profundidade máxima de 7 para as árvores geradas. A população inicial foi gerada aleatoriamente utilizando o método *ramped half-and-half* descrito no Capítulo 2 com profun-

didade inicial dos indivíduos variando entre 2 a 6. Para as operações genéticas de cruzamento, reprodução e mutação, foram adotadas as taxas de 90%, 5%, and 5%, respectivamente.

### 4.3.4 Resultados Experimentais

#### 4.3.4.1 Experimentos com Descritores de Imagens

O desempenho de cada descritor de imagem apresentado na Tabela 4.1 foi avaliado isoladamente nas coleções adotadas. Em ambas as coleções, o descritor de imagem que obteve o melhor desempenho médio foi o descritor CEDD, conforme os resultados apresentados nas Tabelas 4.4 e 4.5. É importante ressaltar que estes valores são apenas uma referência de qualidade dos resultados entre os descritores individuais que serão utilizados no arcabouço de GP. O foco aqui, entretanto, não é comparar os descritores entre si.

| Descritor de | Q1 | | Q2 | |
|---|---|---|---|---|
| imagem | P@10 | MAP | P@10 | MAP |
| CEDD | 0.496 | **0.300** | **0.224** | **0.186** |
| FCTH | 0.380 | 0.185 | 0.136 | 0.105 |
| JCD | **0.516** | 0.290 | **0.224** | 0.176 |
| BIC | 0.174 | 0.143 | 0.096 | 0.083 |
| SDLC | 0.088 | 0.070 | 0.020 | 0.015 |
| PHOG | 0.220 | 0.065 | 0.034 | 0.016 |
| GCH | 0.210 | 0.177 | 0.110 | 0.063 |
| CLD | 0.204 | 0.159 | 0.064 | 0.021 |
| ACC | 0.140 | 0.040 | 0.040 | 0.021 |
| SIFT | 0.156 | 0.073 | 0.020 | 0.006 |
| CSIFT | 0.300 | 0.105 | 0.042 | 0.022 |

**Tabela 4.4.** Desempenho dos descritores de imagens na coleção *DafitiPosthaus*. Melhores resultados são apresentados em negrito.

Analisando-se os resultados obtidos nos experimentos, observa-se que os descritores de imagens que tiveram melhor desempenho na aplicação alvo são aqueles que combinam informação visual de cor e textura em uma única representação, como CEDD, JCD e FCTH. Isto indica que a informação de textura é uma característica importante a ser considerada, além da informação de cor, neste tipo de aplicação. Apesar dos descritores de imagens baseados apenas em cor, como GCH, CLD, SDLC, BIC e ACC apresentarem desempenho inferior aos três descritores compostos citados acima, observa-se que a informação de cor é bastante utilizada como seletor de relevância pelos usuários. Os descritores de imagem PHOG, CSIFT e SIFT obtiveram os piores resultados nas coleções adotadas. O fato do descritor PHOG considerar apenas a informação de forma na representação da imagem explica o seu baixo

| Descritor de | Q1 | | Q2 | |
|---|---|---|---|---|
| imagem | P@10 | MAP | P@10 | MAP |
| CEDD | **0.327** | **0.165** | **0.241** | 0.125 |
| FCTH | 0.224 | 0.103 | 0.182 | 0.116 |
| JCD | 0.308 | 0.160 | 0.206 | **0.144** |
| BIC | 0.123 | 0.065 | 0.199 | 0.096 |
| SDLC | 0.197 | 0.053 | 0.053 | 0.019 |
| PHOG | 0.047 | 0.018 | 0.030 | 0.026 |
| GCH | 0.223 | 0.077 | 0.219 | 0.105 |
| CLD | 0.214 | 0.071 | 0.126 | 0.065 |
| ACC | 0.205 | 0.061 | 0.072 | 0.059 |
| SIFT | 0.132 | 0.044 | 0.020 | 0.018 |
| CSIFT | 0.148 | 0.050 | 0.076 | 0.025 |

**Tabela 4.5.** Desempenho dos descritores de imagens na coleção *Amazon*. Melhores resultados são apresentados em negrito.

desempenho. No caso dos descritores SIFT e CSIFT, apesar de serem considerados excelentes descritores de imagens em outras coleções e contextos, sua baixa efetividade já foi reportada anteriormente quando aplicados ao problema de busca de produtos (Shen et al., 2012).

Em razão dos resultados obtidos com os descritores de imagens individualmente, uma alternativa para melhorar os resultados em sistemas de busca visual de produtos é combinar os vários descritores para produzir uma função única de *ranking* que tire proveito das especialidades de cada descritor. Adicionalmente, a informação complementar disponível em bases de dados de produtos pode ser utilizada como evidência extra para melhorar a ordenação do resultado final.

#### 4.3.4.2 Experimentos com o Arcabouço de GP

As Tabelas 4.6 e 4.7 apresentam os resultados obtidos pelas quatro alternativas de expansão propostas neste capítulo quando aplicadas às coleções adotadas. Os resultados são comparados com os *baselines Visual-GP* and *Total Recall*. A comparação com o arcabouço *Visual-GP* é importante para avaliar se os ganhos obtidos com o uso de GP são decorrentes apenas da combinação dos múltiplos descritores de imagem utilizados, ou se há de fato ganho com a expansão multimodal.

Observa-se claramente os ganhos obtidos com o uso de GP sobre os resultados dos descritores de imagens apresentados na Tabelas 4.4 e 4.5. O método *Visual-GP* obteve resultados superiores a todos os descritores individuais. Por exemplo, na coleção *DafitiPosthaus* o melhor descritor foi o JCD para os dois conjuntos de consultas, considerando a medida P@10. Em ambos os casos, *Visual-GP* apresentou ganhos estatisticamente significantes sobre

| | Q1 | | Q2 | |
|---|---|---|---|---|
| | P@10 | MAP | P@10 | MAP |
| Total Recall | 0,300 | 0,117 | 0,042 | 0,023 |
| Visual-GP | 0.604 | 0.330 | 0.250 | 0.220 |
| Expansão-GPI | **0.722**† | 0.405 | **0.344**† | 0.263 |
| Expansão-GPC | 0.678 | 0.367 | 0.285 | 0.232 |
| Rerank-GPI | 0.704 | **0.418**† | 0.332 | **0.276**† |
| Rerank-GPC | 0.694 | 0.391 | 0.313 | 0.258 |

**Tabela 4.6.** Desempenho dos métodos *Visual-GP*, *Total Recall* e das estratégias de expansão e reordenação baseadas em GP na coleção *DafitiPosthaus*. Valores mais altos são apresentados em negrito. Diferenças estatisticamente significantes entre *Visual-GP* e os métodos de expansão são marcados com (†).

| | Q1 | | Q2 | |
|---|---|---|---|---|
| | P@10 | MAP | P@10 | MAP |
| Total Recall | 0.150 | 0.062 | 0.080 | 0.030 |
| Visual-GP | 0.342 | 0.168 | 0.259 | 0.147 |
| Expansão-GPI | **0.419**† | **0.243**† | **0.367**† | **0.222**† |
| Expansão-GPC | 0.394 | 0.221 | 0.327 | 0.196 |
| Rerank-GPI | 0.391 | 0.226 | 0.315 | 0.187 |
| Rerank-GPC | 0.371 | 0.209 | 0.293 | 0.173 |

**Tabela 4.7.** Desempenho dos métodos *Visual-GP*, *Total Recall* e das estratégias de expansão e reordenação baseadas em GP na coleção *Amazon*. Valores mais altos são apresentados em negrito. Diferenças estatisticamente significantes entre *Visual-GP* e os métodos de expansão são marcados com (†).

o descritor de imagem JCD. Na mesma coleção, considerando a medida MAP, o melhor descritor foi o CEDD para os dois conjuntos de consultas. O método *Visual-GP* apresentou ganhos de 10% para o conjunto Q1 e de 22% para o conjunto Q2 sobre o CEDD. Em resumo, o uso de GP para combinar múltiplos descritores de imagem pode melhorar significativamente os resultados se comparado ao uso de cada descritor individualmente.

O melhor desempenho entre as quatro alternativas apresentadas neste capítulo foi obtido pelo método *Expansão-GPI*, o qual expande os resultados de cada descritor individualmente e permite a inclusão no *ranking* de novos resultados trazidos pela expansão multimodal.

As alternativas de expansão estudadas variam de acordo com duas propriedades. A primeira é permitir ou não a inclusão de novos resultados a partir da expansão. A segunda é expandir os resultados sobre os descritores individuais ou sobre um *ranking* resultante da aplicação de GP para combinar os descritores individuais.

As abordagens *Rerank-GPI* e *Rerank-GPC*, que realizam apenas a reordenação dos

resultados obtidos pelos descritores visuais, resultaram em ganhos quando comparadas ao uso de GP sem expansão realizado pelo *Visual-GP*. No entanto, *Rerank-GPI* e *Rerank-GPC* obtiveram resultados inferiores aos das alternativas *Expansão-GPI* e *Expansão-GPC*, que permitem que a expansão multimodal inclua novos resultados no *ranking*.

As estratégias *Expansão-GPC* e *Rerank-GPC*, que expandem os resultados a partir do *ranking* do *Visual-GP*, obtiveram inferiores aos obtidos pelas alternativas *Expansão-GPI* e *Rerank-GPI*, que realizam a expansão dos resultados a partir das respostas de cada descritor individual. Isto significa que o arcabouço de GP apresentado tira vantagem da informação provida individualmente por cada descritor para computar o *ranking* de respostas com expansão multimodal. Os descritores individuais apresentam maneiras distintas de calcular a distância entre a imagem de consulta e as imagens dos produtos presentes na coleção. Esta diversidade permite que imagens similares com a imagem de consulta sejam recuperadas sobre o ponto de vista de vários aspectos, tais como cor, textura e forma. Ao expandir a consulta a partir dos vários descritores individuais, esta diversidade é também provida por meio das categorias e descrições textuais obtidas.

O método *Expansão-GPI* obteve ganhos em todos os cenários nas coleções adotadas quando comparado ao método *Visual-GP*, com diferenças estatisticamente significantes em todas as comparações de acordo com o *t-test*. Por exemplo, o ganho de MAP obtido pelo método *Expansão-GPI* em relação ao método *Visual-GP* foi de cerca de $22,7\%$ no conjunto de consultas Q1 e de $19,5\%$ no conjunto de consultas Q2 na coleção *DafitiPosthaus*. Considerando a medida P@10 na mesma coleção, os ganhos foram de cerca de $19,5\%$ no conjunto de consultas Q1 e cerca de $37,6\%$ no conjunto de consultas Q2.

Na coleção *Amazon*, os ganhos de MAP do método *Expansão-GPI* em relação ao *Visual-GP* foram de $44,6\%$ no conjunto de consultas Q1 e de $51\%$ no conjunto de consultas Q2. Considerando a medida P@10 na mesma coleção, os ganhos foram de $22,5\%$ no conjunto de consultas Q1 e de $41,7\%$ no conjunto de consultas Q2. Isto mostra que apesar da diferença de idioma entre as duas coleções, os ganhos potenciais obtidos pelas estratégias de expansão multimodal não são afetados.

Quando comparadas ao método de expansão *Total Recall*, todas as quatro alternativas de expansão propostas, assim como o *Visual-GP*, obtiveram resultados superiores. É importante ressaltar que o método *Total Recall* é um método de expansão que não é baseado em nenhuma técnica de aprendizagem de máquina e não permite a expansão dos resultados a partir da combinação de múltiplas características visuais. Além disso, esse método não foi projetado para a aplicação específica de busca por imagem de produtos estudada neste capítulo e sim para o problema de encontrar imagens muito similares em coleções (*near-duplicate*). A inclusão do método *Total Recall* nos experimentos se deu para evitar eventuais dúvidas quanto ao seu desempenho em relação aos métodos propostos, dado que esse método *Total Recall* é

considerado uma referência importante entre os métodos de expansão para busca de imagens.

As Tabelas 4.8 e 4.9 apresentam a comparação entre os resultados obtidos pelo *baseline TCatBR* e pelo método *Expansão-GPI*, a melhor alternativa de expansão segundo os resultados apresentados.

| | Q1 | | Q2 | |
|---|---|---|---|---|
| | P@10 | MAP | P@10 | MAP |
| Expansão-GPI | **0.722**$^{\dagger}$ | **0.405**$^{\dagger}$ | **0.344**$^{\dagger}$ | **0.263**$^{\dagger}$ |
| TCatBR | 0.640 | 0.262 | 0.288 | 0.205 |

**Tabela 4.8.** Desempenho dos métodos *Expansão-GPI* e *TCatBR* na coleção *Dafiti-Posthaus*. Valores mais altos são apresentados em negrito. Diferenças estatisticamente significantes entre o método *Expansão-GPI* e o método *TCatBR* são marcados com (†).

| | Q1 | | Q2 | |
|---|---|---|---|---|
| | P@10 | MAP | P@10 | MAP |
| Expansão-GPI | **0.419** | **0.243**$^{\dagger}$ | **0.367**$^{\dagger}$ | **0.222**$^{\dagger}$ |
| TCatBR | 0.402 | 0.192 | 0.324 | 0.182 |

**Tabela 4.9.** Desempenho dos métodos *Expansão-GPI* e *TCatBR* na coleção *Amazon*. Valores mais altos são apresentados em negrito. Diferenças estatisticamente significantes entre o método *Expansão-GPI* e o método *TCatBR* são marcados com (†).

O método *Expansão-GPI* obteve ganhos superiores ao *TCatBR* em todas as coleções e cenários de consulta analisados. Na coleção *DafitiPosthaus*, os ganhos de MAP obtidos pelo método *Expansão-GPI* em relação ao método *TCatBR* foram de $54,6\%$ para o conjunto Q1 e de $28,3\%$ para o conjunto Q2. Considerando a medida P@10 na mesma coleção, os ganhos foram $12,8\%$ e $19,4\%$ para os conjuntos Q1 e Q2, respectivamente. Todos os ganhos nesta coleção foram considerados estatisticamente significantes de acordo com o resultado do *t-test*.

Na coleção *Amazon*, os ganhos de MAP foram considerados estatisticamente significantes para os dois conjuntos de consultas. Considerando a medida P@10 na mesma coleção, apenas os ganhos obtidos no conjunto Q2 foram considerados estatisticamente significantes, apesar do método *Expansão-GPI* também ter apresentado resultados melhores que o *TCatBR* no conjunto Q1.

Os resultados gerais indicam que GP é uma alternativa viável para expandir automaticamente uma imagem de consulta. Além dos resultados competitivos obtidos, a expansão baseada em GP oferece um arcabouço que pode ser usado em outras aplicações de busca de imagem onde a informação multimodal está disponível, enquanto que o método *TCatBR*, é

um método de reordenação específico, proposto exclusivamente para ser adotado na busca de produtos de moda.

As Figuras 4.5 e 4.6 ilustram as diferenças de valores de P@10 entre o método de expansão *Expansão-GPI* e o método sem expansão *Visual-GP* para os conjuntos Q1 e Q2, respectivamente, na coleção *DafitiPosthaus*. Esta comparação é útil para entender as vantagens do método de expansão *Expansão-GPI* em relação ao método *Visual-GP*. Ao analisar os resultados das consultas individualmente, o método de expansão *Expansão-GPI* teve melhor desempenho em 64% das consultas, desempenho equivalente em 14% e, desempenho inferior em apenas 22% das consultas do conjunto Q1. No conjunto Q2, o método de expansão *Expansão-GPI* obteve melhor desempenho em 44% das consultas, desempenho equivalente em 38% e, desempenho inferior em apenas 18% das consultas.

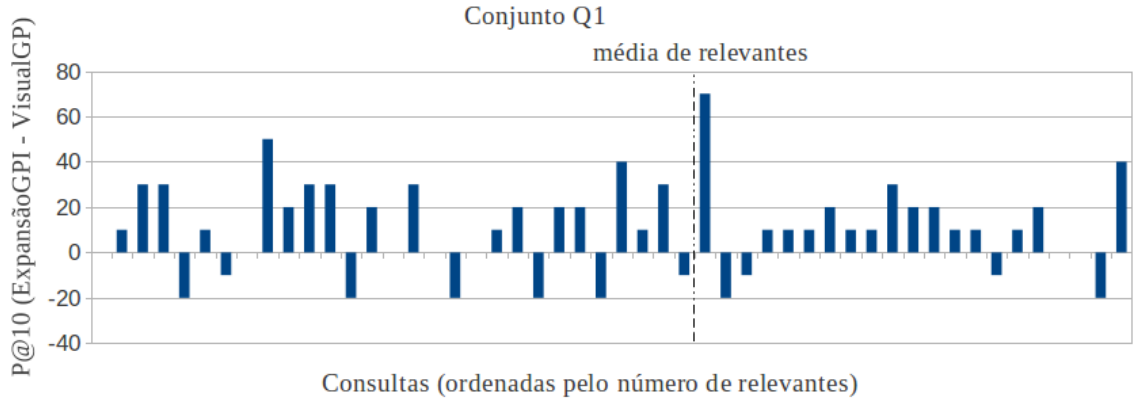


**Figura 4.5.** Diferença nos valores de P@10 entre o método *Expansão-GPI* e *Visual-GP* para cada consulta no conjunto Q1 na coleção *DafitiPosthaus*.

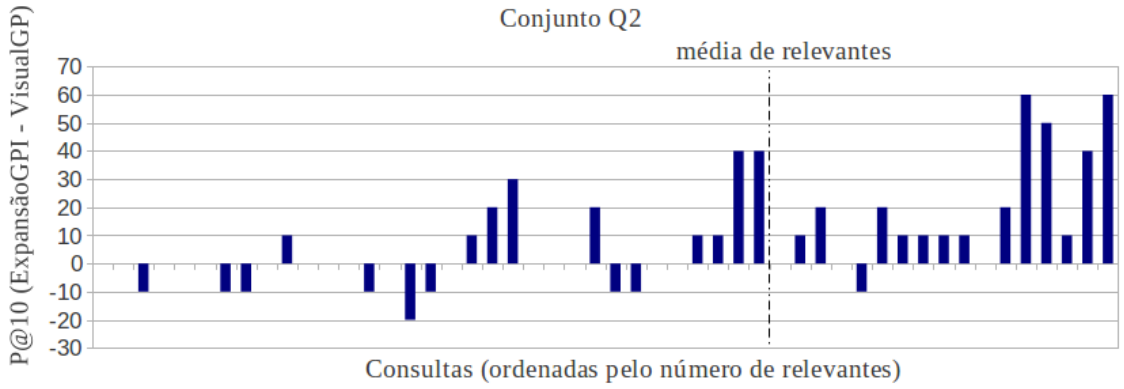


**Figura 4.6.** Diferença nos valores de P@10 entre o método *Expansão-GPI* e *Visual-GP* para cada consulta no conjunto Q2 na coleção *DafitiPosthaus*.

Ao analisar as 28 consultas no conjunto Q2 nas quais o método *Expansão-GPI* teve desempenho equivalente ou inferior ao método *Visual-GP*, foi constatado que estes resultados se referem às consultas que contêm poucos resultados relevantes, com 24 delas contendo um número de imagens relevantes abaixo da média apresentada para o conjunto. O mesmo fenômeno ocorre quando se analisa os casos de perdas no conjunto Q1. Das 11 consultas nas quais o método de expansão teve desempenho pior, sete delas têm um número de relevantes abaixo da média de relevantes para o conjunto.

A Figura 4.7 apresenta o exemplo de consulta na coleção *DafitiPosthaus*, onde o método de expansão *Expansão-GPI* obteve desempenho melhor em relação ao método *Visual-GP*, de acordo com com o julgamento relevância dos usuários. A imagem de consulta ilustra uma mulher carregando uma bolsa, o que parece ser o produto procurado de acordo com o julgamento relevância dos usuários. No resultado apresentado pode-se observar que o método *Expansão-GPI* foi capaz de filtrar resultados errados do topo da resposta.


**Figura 4.7.** Resultado da busca visual pela imagem de consulta (a), onde o método *Expansão-GPI* (c) obteve melhor desempenho em relação ao método *Visual-GP* (b).

A Figura 4.8 apresenta um exemplo de consulta na coleção *DafitiPosthaus*, onde o método de expansão *Expansão-GPI* obteve desempenho pior em relação ao método *Visual-GP*. Neste caso, a imagem de consulta ilustra um homem vestindo uma camiseta, o que parece ser o produto procurado de acordo com o julgamento relevância dos usuários. O método *Visual-GP* recuperou imagens similares à consulta, mas apresentou uma camiseta feminina nos 10 primeiros resultados. Embora o método *Expansão-GPI* não tenha recuperado nenhuma camiseta feminina no topo da resposta, este método incluiu três camisetas no topo do resultado que não foram consideradas relevantes pelos usuários. Este exemplo mostra que uma qualidade mínima no *ranking* inicial é necessária para fazer a expansão funcionar de maneira adequada. Esta evidência é reforçada pelos gráficos comparativos das Figuras 4.5 e 4.6, que mostram que os ganhos são maiores nas consultas que contém um número maior de relevantes.


**Figura 4.8.** Resultado da busca visual pela imagem de consulta (a), onde o método *Expansão-GPI* (c) obteve pior desempenho em relação ao método *Visual-GP* (b).

### 4.3.5 Custos Computationais

Muitos dos custos computacionais relacionados com as abordagens propostas neste capítulo são adicionados em um processo *offline* para derivar as funções de combinação com o uso de GP. Ao processar as consultas, a principal sobrecarga é recuperar as características multimodais para os primeiros resultados. Nos experimentos realizados, estas informações multimodais foram derivadas dos 20 resultados do topo. Isto significa que é necessário recuperar apenas a descrição textual dos 20 primeiros resultados, criar uma consulta composta pela concatenação dos primeiros termos presentes nestas descrições e submeter esta consulta textual. O processamento de cada consulta textual é, de fato, muito mais rápido do que o processamento da imagem de consulta inicial.

Nos experimentos realizados, este custo adicional foi próximo do tempo para processar a imagem de consulta sem expansão. Com a expansão, o tempo médio para processar uma imagem de consulta subiu de 1182$ms$ para 1506$ms$. A função de combinação representa um custo muito pequeno e o custo extra para processar as informações textuais não chega a dobrar o tempo de execução.

Vale ressaltar que os experimentos foram realizados utilizando a biblioteca LIRE para processar as consultas visuais e a biblioteca Lucene para processar as consultas textuais, com todas as tarefas sendo executadas sequencialmente. É importante salientar que o foco deste trabalho não era fazer uma análise de desempenho, nem propor uma solução otimizada sob o ponto de vista de custo de processamento. No entanto, em um sistema real projetado para considerar questões de eficiência, esta sobrecarga pode ser reduzida. Por exemplo, a expansão multimodal para cada descritor visual poderia ser feita em paralelo, reduzindo-se o tempo para processar as consultas.

### 4.3.6 Análise das Funções

Assim como no capítulo anterior, o indivíduo escolhido no final de cada execução do arcabouço evolucionário representa a melhor solução encontrada para o problema alvo no espaço de busca considerado durante aquela execução. As Equações 4.1 e 4.2 mostram exemplos de duas funções geradas pelo método de expansão *Expansão-GPI* para o conjunto Q1 nas coleções *DafitiPosthaus* e *Amazon*, respectivamente.

$$csift_{text20} + fcth_{text10} - rlog10(max(fcth_{mindist}, gch_{text1}) - (jcd_{mindist} + cedd)) + min(phog, jcd_{text20}) + phog_{text20} + gch_{text1} + (cld_{text10} * cedd_{cat5}) \quad (4.1)$$

$$cedd_{cat20} - (rlog10((fcth_{text5} + (fcth_{cat5} - jcd)) - rlog10(cSift))) \quad (4.2)$$

onde os nomes dos descritores sozinhos representam a distância entre a imagem de consulta e uma imagem da coleção, de acordo com aquele descritor. Um descritor $d$, seguido por text$n$ ($d_{text\{n\}}$), representa a similaridade textual entre um documento e a descrição textual das top $N$ imagens no ranking inicial dado por $d$. Um descritor $d$, seguido por cat$n$ ($d_{cat\{n\}}$), representa a frequência da categoria do documento nos top $N$ resultados dados por $d$. A função *min* retorna o valor mínimo entre dois valores fornecidos como parâmetros. Um descritor $d$, seguido por $mindist(d_{mindist})$, representa a distância mínima de qualquer documento para a imagem de consulta, conforme o descritor $d$.

Uma análise das funções geradas pelo arcabouço de GP permite saber quais terminais estão sendo utilizados nas funções geradas e quais foram descartados pelo arcabouço. A Tabela 4.10 apresenta estatísticas de ocorrência dos terminais derivados dos 11 descritores individuais, em 50 funções geradas pelo pelo método *Expansão-GPI* para o conjunto Q1 na coleção *DafitiPosthaus*.

Pode-se observar na Figura 4.9 que os terminais derivados dos descritores CEDD e JCD foram os que ocorreram com maior frequência nas funções analisadas. Esses dois descritores foram os que apresentaram os melhores desempenhos individualmente, conforme os resultados mostrados na Tabela 4.4. Os terminais derivados dos descritores ACC e SIFT foram os que ocorreram com menor frequência, figurando em apenas 10% das funções. Esses dois descritores também foram os que apresentaram os piores desempenhos individualmente, considerando a medida P@10. Os resultados indicam portanto uma interessante correlação entre o desempenho dos descritores sobre a coleção e sua frequência de ocorrência nas

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| CEDD | 166 | 42 |
| JCD | 143 | 43 |
| CSIFT | 127 | 42 |
| CLD | 106 | 38 |
| FCTH | 102 | 39 |
| GCH | 86 | 38 |
| BIC | 86 | 30 |
| PHOG | 64 | 28 |
| SDLC | 48 | 10 |
| ACC | 20 | 5 |
| SIFT | 15 | 5 |

**Tabela 4.10.** Estatísticas de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *DafitiPosthaus*.

fórmulas produzidas pelo arcabouço. Tal correlação é um indício de que o arcabouço é capaz de identificar os terminais que produzem bons resultados.

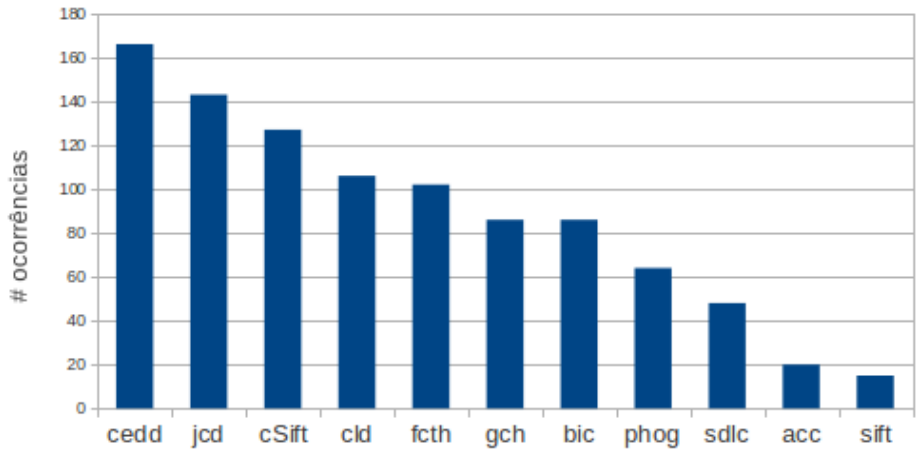


**Figura 4.9.** Frequência de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *DafitiPosthaus*.

A Tabela 4.11 apresenta os 15 terminais mais frequentes nas funções produzidas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *DafitiPosthaus*. Pode-se observar que o uso do texto extraído a partir do CEDD e do JCD estão entre os terminais mais usados. Além disso, oito dos dez terminais mais utilizados estão relacionados ao texto da descrição dos

produtos da resposta inicial, um indício de que a expansão a partir do texto é uma característica com papel importante no arcabouço de expansão apresentado.

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| $cedd_{text20}$ | 60 | 29 |
| $jcd_{text20}$ | 40 | 24 |
| $cedd_{freq20}$ | 33 | 18 |
| $cld_{text20}$ | 30 | 12 |
| $cld_{text10}$ | 29 | 20 |
| $jcd_{text10}$ | 24 | 16 |
| $cedd$ | 23 | 14 |
| $cSift_{text20}$ | 22 | 13 |
| $cSift_{text10}$ | 21 | 13 |
| $cSift_{text5}$ | 20 | 12 |
| $cedd_{freq5}$ | 19 | 14 |
| $fcth_{text5}$ | 19 | 9 |
| $jcd_{freq20}$ | 19 | 15 |
| $cld_{mindist}$ | 18 | 6 |
| $cld_{text1}$ | 18 | 10 |

**Tabela 4.11.** Estatísticas de ocorrência dos 15 terminais mais frequentes nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *DafitiPosthaus*.

A Tabela 4.12 apresenta estatísticas de ocorrência dos terminais derivados dos 11 descritores individuais, em 50 funções geradas pelo pelo método *Expansão-GPI* para o conjunto Q2 na coleção *DafitiPosthaus*. Novamente percebe-se uma certa correlação entre o desempenho de cada descritor individualmente e sua frequência de uso pelo arcabouço de GP. Resultado que também se repete nos demais cenários.

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| CEDD | 284 | 45 |
| JCD | 122 | 36 |
| FCTH | 77 | 28 |
| CLD | 64 | 29 |
| BIC | 57 | 24 |
| GCH | 56 | 25 |
| CSIFT | 49 | 22 |
| PHOG | 30 | 16 |
| ACC | 30 | 15 |
| SDLC | 20 | 6 |
| SIFT | 15 | 5 |

**Tabela 4.12.** Estatísticas de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *DafitiPosthaus*.

Na Figura 4.10 pode-se observar que novamente os terminais derivados do CEDD e JCD foram os que ocorreram com maior frequência nas funções analisadas. Os terminais derivados do SDLC e SIFT foram os que ocorreram com menor frequência, figurando em apenas 12% e 10% das funções, respectivamente. Esse dois descritores também foram os que apresentaram os piores desempenhos individualmente, considerando a medida P@10.

A Tabela 4.13 apresenta os 15 terminais mais frequentes nas funções produzidas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *DafitiPosthaus*. Para as consultas do conjunto Q2, a informação de categoria passa a ter uma frequência mais próxima a de texto, com as duas aparecendo de maneira mais homogênea nas fórmulas produzidas. Tal resultado pode ser consequência do fato das consultas produzirem rankings iniciais piores, o que prejudica a informação mais específica sobre os produtos gerada pela expansão textual e favorece a expansão a partir de informação mais genérica produzida pelas categorias.

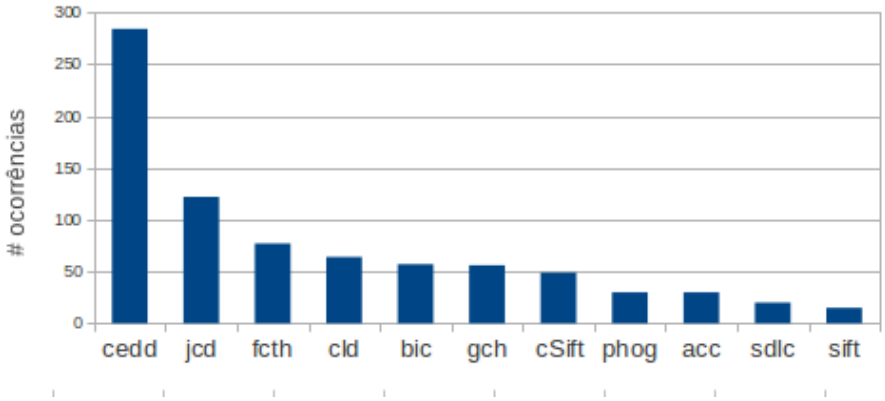


**Figura 4.10.** Frequência de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *DafitiPosthaus*.

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| $cedd_{cat20}$ | 94 | 37 |
| $cedd$ | 59 | 26 |
| $cedd_{text20}$ | 55 | 27 |
| $jcd$ | 39 | 19 |
| $jcd_{text10}$ | 26 | 12 |
| $cld_{mindist}$ | 21 | 9 |
| $cedd_{cat5}$ | 20 | 13 |
| $cedd_{text5}$ | 19 | 10 |
| $fcth_{text5}$ | 17 | 7 |
| $cedd_{cat10}$ | 15 | 11 |
| $jcd_{cat10}$ | 14 | 7 |
| $jcd_{text20}$ | 13 | 8 |
| $fcth_{text1}$ | 12 | 5 |
| $fcth_{cat10}$ | 12 | 5 |
| $cSift_{text10}$ | 12 | 5 |

**Tabela 4.13.** Estatísticas de ocorrência dos 15 terminais mais frequentes nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *DafitiPosthaus*.

A Tabela 4.14 apresenta estatísticas de ocorrência dos terminais derivados dos 11 descritores individuais, em 50 funções geradas pelo pelo método *Expansão-GPI* para o conjunto Q1 na coleção *Amazon*. Novamente observa-se boa correlação entre desempenho dos descritores individuais e seu uso por parte do arcabouço de GP proposto.

Na Figura 4.11 observa-se que os terminais derivados dos descritores CEDD e JCD

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| CEDD | 199 | 30 |
| JCD | 118 | 28 |
| GCH | 50 | 19 |
| FCTH | 49 | 19 |
| CLD | 40 | 19 |
| ACC | 39 | 19 |
| SDLC | 21 | 10 |
| CSIFT | 20 | 11 |
| BIC | 18 | 7 |
| SIFT | 17 | 8 |
| PHOG | 14 | 7 |

**Tabela 4.14.** Estatísticas de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *Amazon*.

ocorreram com maior frequência em 60% e 56% das funções, respectivamente. Os terminais derivados do descritor PHOG foram os que menos ocorreram. Estes resultados também refletem os desempenhos obtidos por esses descritores individualmente, conforme apresentado na Tabela 4.5.

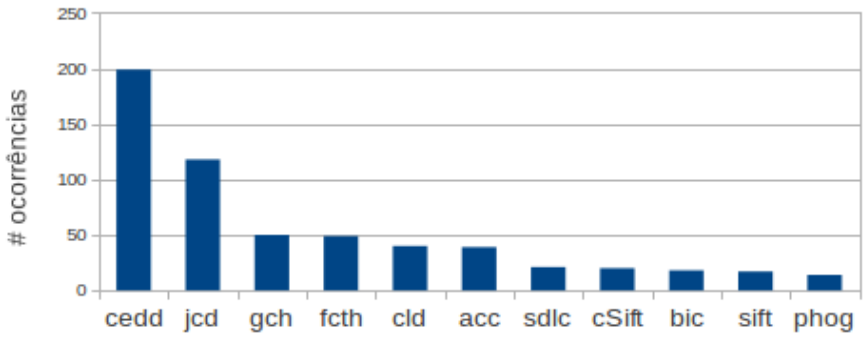


**Figura 4.11.** Frequência de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *Amazon*.

A Tabela 4.15 apresenta os 15 terminais mais frequentes nas funções produzidas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *Amazon*. Neste caso não há o domínio de terminais relacionados à expansão textual, como ocorre na coleção *DafitiPosthaus*. O fato dos descritores individuais na coleção *Amazon* terem apresentado resultados ligeiramente piores do que os obtidos na coleção *DafitiPosthaus* pode explicar a ausência do predomínio de

características textuais nos terminais gerados para a coleção *Amazon*. Ao observar a coleção mais detalhadamente, percebeu-se que parte das descrições textuais presentes na coleção *Amazon* possuem códigos de produto inseridos ao final da descrição, o que prejudicou também o desempenho da expansão textual. No entanto, percebe-se que o arcabouço de GP foi capaz de contornar de certa forma tal problema, fazendo menos uso da informação textual do que o fez na coleção *DafitiPosthaus*.

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| $cedd$ | 60 | 24 |
| $cedd_{cat20}$ | 30 | 12 |
| $jcd$ | 27 | 16 |
| $jcd_{cat10}$ | 23 | 7 |
| $cedd_{cat1}$ | 20 | 8 |
| $cedd_{text10}$ | 17 | 10 |
| $cedd_{text1}$ | 16 | 8 |
| $jcd_{text20}$ | 16 | 9 |
| $cedd_{text20}$ | 15 | 7 |
| $jcd_{cat20}$ | 15 | 6 |
| $cedd_{cat5}$ | 14 | 8 |
| $fcth_{text1}$ | 13 | 8 |
| $cld_{mindist}$ | 13 | 6 |
| $cedd_{cat10}$ | 12 | 7 |
| $jcd_{mindist}$ | 11 | 7 |

**Tabela 4.15.** Estatísticas de ocorrência dos 15 terminais mais frequentes nas funções geradas pelo método *Expansão-GPI* para o conjunto Q1 na coleção *Amazon*.

A Tabela 4.16 apresenta estatísticas de ocorrência dos terminais derivados dos 11 descritores individuais, em 50 funções geradas pelo pelo método *Expansão-GPI* para o conjunto Q2 na coleção *Amazon*.

Na Figura 4.12 observa-se que os terminais derivados dos descritores CEDD e FCTH ocorreram com maior frequência em 56% e 60% das funções, respectivamente. Os terminais derivados do descritor SIFT e PHOG foram os que menos ocorreram, ocorrendo em apenas 16% e 10% das funções, respectivamente. Estes resultados também refletem os resultados obtidos pelos descritores individuais, conforme apresentado na Tabela 4.5.

A Tabela 4.17 apresenta os 15 terminais mais frequentes nas funções produzidas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *Amazon*. Assim como na coleção *DafitiPosthaus*, a informação de categoria foi mais utilizada na expansão das imagens de consulta do conjunto Q2 do que no conjunto Q1. Dos dez terminais mais utilizados, seis

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| CEDD | 119 | 28 |
| FCTH | 113 | 30 |
| JCD | 107 | 27 |
| ACC | 98 | 27 |
| BIC | 84 | 24 |
| GCH | 73 | 26 |
| CLD | 37 | 17 |
| CSIFT | 17 | 7 |
| SDLC | 17 | 6 |
| SIFT | 15 | 8 |
| PHOG | 11 | 5 |

**Tabela 4.16.** Estatísticas de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *Amazon*.

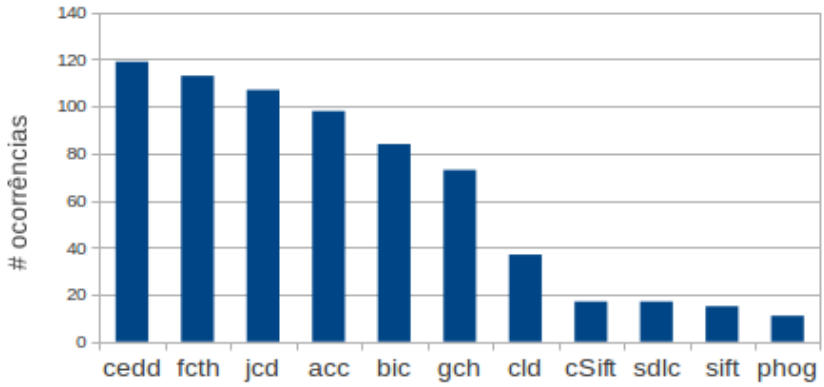


**Figura 4.12.** Frequência de ocorrência dos terminais derivados dos descritores individuais nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *Amazon*.

são derivados da informação de categoria no conjunto Q2, enquanto que no conjunto Q1, apenas quatro dos dez terminais mais utilizados foram derivados da informação de categoria. Contudo, a diferença é menor do que a observada na coleção *DafitiPosthaus*, assim como a diferença de desempenho dos descritores entre os conjuntos Q1 e Q2 também é menor na coleção *Amazon*.

No geral, pode-se observar que os terminais derivados dos descritores visuais compostos CEDD e JCD, que exploram propriedades de cor e textura, estão entre os mais frequentes tanto no conjunto Q1, quanto no conjunto Q2, em ambas as coleções. Terminais derivados de outros descritores, como FCTH, CSIFT, CLD, GCH, ACC e BIC também figuram entre os mais frequentes de acordo com as especificidades de cada conjunto de consulta em cada

| DESCRITOR | # OCORRÊNCIAS | # FUNÇÕES |
|---|---|---|
| $jcd$ | 29 | 22 |
| $acc_{text20}$ | 27 | 11 |
| $cedd_{cat10}$ | 25 | 11 |
| $cedd_{cat20}$ | 23 | 9 |
| $fcth_{cat5}$ | 19 | 8 |
| $acc_{cat5}$ | 19 | 12 |
| $acc_{text5}$ | 19 | 8 |
| $cedd$ | 17 | 9 |
| $fcth_{cat20}$ | 16 | 12 |
| $jcd_{cat1}$ | 15 | 7 |
| $cedd_{cat5}$ | 14 | 7 |
| $fcth_{cat10}$ | 14 | 11 |
| $gch$ | 14 | 11 |
| $cld_{mindist}$ | 14 | 6 |
| $bic$ | 14 | 11 |

**Tabela 4.17.** Estatísticas de ocorrência dos 15 terminais mais frequentes nas funções geradas pelo método *Expansão-GPI* para o conjunto Q2 na coleção *Amazon*.

coleção. É possível observar também que cada descritor contribui no processo de recuperação através de diversos terminais em diferentes topos do *ranking* inicial. Vale salientar que o arcabouço implementado é flexível o bastante para ser utilizado com outros descritores visuais diferentes daqueles que foram considerados nos experimentos. Esses outros descritores podem ser incluídos para explorar novas características visuais das imagens.

# Capítulo 5

# Conclusão e Trabalhos Futuros

Este capítulo discute as conclusões e algumas possibilidades de trabalhos de pesquisa futuros relacionados à esta tese.

## 5.1 Conclusões

Nesta tese foi abordada a aplicação de programação genética (GP - *Genetic Progamming*) em problemas de busca de imagens sob dois contextos distintos: a busca de imagens na Web utilizando informação textual extraída automaticamente das páginas Web e, a busca visual através da expansão da imagem de consulta utilizando informação multimodal. Para avaliar as estratégias propostas para o contexto de busca visual, foi escolhido como estudo de caso a busca visual de produtos em lojas de comércio eletrônico voltadas para o segmento de moda. Os dois contextos foram utilizados como estudo de caso para a aplicação de GP em problemas de busca de imagens, visando validar as hipóteses de que há espaço para novas soluções que utilizem GP e também de que GP é uma técnica que pode ser utilizada com sucesso em diferentes cenários de busca de imagem.

No primeiro contexto, foi estudado um arcabouço para recuperação de imagens no cenário da Web. O arcabouço utilizou o texto e metadados presentes nas páginas como fontes de evidências para descrever as imagens e, aplicou os princípios da programação genética para derivar formas de combinação não-lineares dessas evidências, descartando inclusive àquelas que não contribuem com a melhoria da qualidade do conjunto de imagens recuperadas para uma dada consulta textual.

Este primeiro estudo foi desenvolvido como uma extensão ao trabalho apresentado inicialmente em (dos Santos et al., 2009). Em resumo, as principais contribuições em relação ao trabalho preliminar foram: (i) uma análise mais detalhada sobre quais fontes de evidência textuais são mais importantes para representar o conteúdo das imagens na Web; (ii) inclusão

de novas características textuais como terminais do GP; (iii) uso do projeto fatorial em dois níveis na avaliação experimental para melhor investigar o efeito de alguns parâmetros na configuração do arcabouço de GP; (iv) uma avaliação experimental do arcabouço de GP em uma coleção de imagens de domínio genérico coletada da Web, comparando-o com duas abordagens de recuperação; (v) produção do artigo *Evaluation of parameters for combining multiple textual sources of evidence for web image retrieval using genetic programming* publicado no *Journal of the Brazilian Computer Society* em 2013.

Entre as fontes de evidência estudadas, pôde-se observar que os terminais correspondentes ao texto completo, título da página e passagem de 40 termos foram as evidências que mais contribuíram para a melhoria da qualidade nos resultados. Os terminais derivados das tags SRC, texto de âncora e ALT, embora tenham ocorrido com menor frequência, também contribuíram de alguma forma no processo de recuperação de imagens. Já as evidências de autor, descrição e palavras-chave foram completamente descartadas pelo arcabouço de GP, indicando que estes atributos não colaboraram para o processo de recuperação de imagens.

De acordo com os experimentos realizados, a abordagem de GP adotada para o cenário da Web foi estatisticamente superior em relação às abordagens Okapi-BM25 e arcabouço Bayesiano, com ganhos de 22,36% sobre o BM25 e ganhos de 45,39% sobre o arcabouço Bayesiano. Os resultados reforçam a hipótese de que GP é uma boa alternativa para o desenvolvimento de soluções de busca por imagem baseada em texto.

No segundo contexto, foi estudado o problema da busca visual através da expansão da imagem de consulta utilizando informação multimodal. Para avaliar as estratégias propostas para o contexto de busca visual, foi escolhido como estudo de caso a busca visual de produtos em lojas de comércio eletrônico voltadas para o segmento de moda. Mais especificamente, a busca de produtos na qual o usuário submete apenas uma imagem como consulta ao sistema. O arcabouço proposto aplicou os princípios da programação genética para realizar tanto a expansão automática da consulta inicial quanto a produção de um novo *ranking* baseado na consulta expandida. Até onde se sabe, GP nunca foi aplicado no cenário de busca abordado neste contexto, no qual somente a informação visual está disponível para expansão multimodal de imagens de consulta.

Foram propostas quatro estratégias baseadas em programação genética para derivar métodos de expansão multimodal a partir de imagens de consultas. As estratégias, denominadas de *Expansão-GPI*, *Expansão-GPC*, *Rerank-GPI* e *Rerank-GPC* foram avaliadas e comparadas com outras abordagens encontradas na literatura. Os resultados apresentados indicam que ganhos expressivos na qualidade dos resultados podem ser obtidos neste cenário quando se realiza a expansão multimodal da imagem de consulta. Comparado com o *ranking* gerado pelo método *Visual-GP* na coleção *DafitiPosthaus*, o método *Expansão-GPI* obteve ganhos de $22,7\%$ e de $19,5\%$ em termos de MAP para conjuntos de consulta Q1 e Q2, respectivamente.

Considerando a medida P@10, os ganhos foram de 19,5% no conjunto de consultas Q1 e de 37,6% no conjunto de consultas Q2. Na coleção *Amazon*, os ganhos de MAP do método *Expansão-GPI* em relação ao *Visual-GP* foram de 44,6% no conjunto de consultas Q1 e de 51% no conjunto de consultas Q2. Considerando a medida P@10 na mesma coleção, os ganhos foram de 22,5% no conjunto de consultas Q1 e de 41,7% no conjunto de consultas Q2.

Em resumo, as principais contribuições neste segundo contexto estudado são: (i) um arcabouço para expansão multimodal de consultas baseado em em programação genética. São apresentadas quatro alternativas para expandir automaticamente uma imagem de consulta utilizando informação multimodal disponível nas coleções; (ii) uma avaliação experimental das abordagens propostas em uma aplicação de busca visual em lojas de comércio eletrônico do segmento de moda. (iii) produção do artigo *A Multimodal query expansion based on genetic programming for visually-oriented e-commerce applications* submetido para o *Expert Systems with Applications Journal*.

## 5.2 Trabalhos futuros

Os estudos e métodos propostos nesta tese abrem um leque de possibilidades para trabalhos de pesquisa futuros, tais como:

- Introduzir evidências relacionadas ao conteúdo visual das imagens no arcabouço de GP apresentado no Capítulo 3 para recuperação de imagens na Web utilizando informação multimodal.

- Estudar a possibilidade de derivar funções de combinação para tipos específicos de consulta. A ideia é verificar se estas funções específicas podem superar os resultados obtidos com uma função de uso geral como apresentado nesta tese. As consultas poderiam ser agrupadas em função de propriedades específicas, tais como a popularidade, localização geográfica do usuário, etc. Trabalhos anteriores indicam que a criação de funções específicas para cada categoria de consultas pode resultar em melhorias significativas no desempenho de sistemas de busca para a Web (Berlt et al., 2010).

- Outra possibilidade é estudar o impacto dos parâmetros de GP nestas funções específicas.

- Incluir técnicas de realimentação de relevância na expansão multimodal. Técnicas de realimentação de relevância, incluindo interações diretas com usuários ou o uso de informação de clique em consultas passadas, podem ser empregadas como uma

nova forma de aplicação de GP para a expansão de consultas. Tais estratégias podem aumentar a precisão da descrição inicial obtida sobre a imagem de consulta, dado que o usuário apontaria quais as imagens de respostas são relevantes para a imagem de consulta. Constatou-se nos experimentos que uma maior concentração de relevantes na resposta inicial dada a imagem de consulta resulta em melhorias na expansão multimodal.

- Estudar as estratégias de expansão utilizando outros métodos de aprendizagem e em outros cenários. Apesar de terem sido experimentadas em um único cenário nesta tese, as ideias podem ser aplicadas em outros cenários, tais como geo-codificação de imagens, recuperação de vídeos e recuperação de imagens médicas para ajudar no diagnóstico e tratamento de pacientes.

- Estudar formas de detectar automaticamente situações nas quais a expansão pode não resultar em melhoria, evitando assim custos computacionais desnecessários. Poderia-se estudar, por exemplo, se formas de mensurar a divergência entre as respostas do topo do *ranking* inicial e a imagem de consulta fornecida pelo usuário não seriam bons indícios de que a consulta não deve ser expandida.

- Estudar alternativas para selecionar dinamicamente as imagens do topo dos resultados das quais são extraídas as informações adicionais para realizar a expansão multimodal.

# Referências Bibliográficas


Abdel-Hakim, A. E. & Farag, A. A. (2006). Csift: A sift descriptor with color invariant characteristics. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1978--1983.

Andrade, F. S. P.; Almeida, J.; Pedrini, H. & da S. Torres, R. (2012). Fusion of local and global descriptors for content-based image and video retrieval. In *Iberoamerican Congress on Pattern Recognition*, volume 7441, pp. 845--853.

Arampatzis, A.; Zagoris, K. & Chatzichristofis, S. A. (2011a). Dynamic two-stage image retrieval from large multimodal databases. In *European Conference on Advances in Information Retrieval*, pp. 326--337.

Arampatzis, A.; Zagoris, K. & Chatzichristofis, S. A. (2011b). Fusion vs. two-stage for multimodal retrieval. In *European Conference on Advances in Information Retrieval*, pp. 759--762.

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval - the concepts and technology behind search*. Pearson Education, England, 2 edição.

Berlt, K.; de Moura, E. S.; Carvalho, A.; Cristo, M.; Ziviani, N. & Couto, T. (2010). Modeling the web as a hypergraph to compute page reputation. *Journal of Information Systems*, 35(5):530--543.

Bosch, A.; Zisserman, A. & Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, pp. 401--408.

Box, G. E. P.; Hunter, J. S. & Hunter, W. G. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons, New York, USA.

Calumby, R. T.; da S. Torres, R. & Gonçalves, M. A. (2012). Multimodal retrieval with relevance feedback based on genetic programming. *Multimedia Tools and Applications*, pp. 1--29.

Chandrasekhar, V. R.; Chen, D. M.; Tsai, S. S.; Cheung, N.-M.; Chen, H.; Takacs, G.; Reznik, Y.; Vedantham, R.; Grzeszczuk, R.; Bach, J. & Girod, B. (2011). The stanford mobile visual search data set. In *ACM Conference on Multimedia Systems*, pp. 117--122.

Chang, N.-S. & Fu, K.-S. (1980). Query-by-pictorial-example. *IEEE Transactions on Software Engineering*, 6(6):519--524.

Chang, S.-K. & Kunii, T. L. (1981). Pictorial data-base systems. *IEEE Computer*, 14(11):13–21.

Chang, Y.-c. & Chen, H.-h. (2007). Experiment for using web information to do query and document expansion. In *In Working Notes of the 2007 CLEF Workshop*.

Chatzichristofis, S. A. & Boutalis, Y. S. (2008a). Cedd (color and edge directivity descriptor): A compact descriptor for image indexing and retrieval. In *International Conference on Computer Vision Systems*, pp. 312--322.

Chatzichristofis, S. A. & Boutalis, Y. S. (2008b). Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 191--196.

Chatzichristofis, S. A.; Boutalis, Y. S. & Lux, M. (2009). Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and Applications*, pp. 134--140.

Chen, Y.; Yu, N.; Luo, B. & Chen, X.-w. (2010). ilike: Integrating visual and textual features for vertical search. In *International Conference on Multimedia*, pp. 221--230.

Cheng, Z.; Ren, J.; Shen, J. & Miao, H. (2011). The effects of heterogeneous information combination on large scale social image search. In *International Conference on Internet Multimedia Computing and Service*, pp. 39--42.

Chum, O.; Philbin, J.; Sivic, J.; Isard, M. & Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision*, pp. 1--8.

Clinchant, S.; Ah-Pine, J. & Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *ACM International Conference on Multimedia Retrieval*, pp. 1--8.

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

Cui, J.; Wen, F. & Tang, X. (2008). Real time google and live image search re-ranking. In *ACM International Conference on Multimedia*, pp. 729--732.

da Costa Carvalho, A. L.; Rossi, C.; de Moura, E. S.; da Silva, A. S. & Fernandes, D. (2012). Lepref: Learn to precompute evidence fusion for efficient query evaluation. *Journal of the American Society for Information Science and Technology*, 63(7):1383--1397.

Datta, R.; Joshi, D.; Li, J. & Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.

de Almeida, H. M.; Gonçalves, M. A.; Cristo, M. & Calado, P. (2007). A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 399--406.

Del Bimbo, A. (1999). *Visual information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Depeursinge, A. & Müller, H. (2010). Fusion techniques for combining textual and visual information retrieval. In *Working Notes of ImageCLEF Workshop*, pp. 95--114.

Deselaers, T.; Keysers, D. & Ney, H. (2005). Fire – flexible image retrieval engine: Imageclef 2004 evaluation. In *Cross-language Evaluation Forum Conference*, pp. 688--698.

Deselaers, T.; Weyand, T. & Ney, H. (2007). Image retrieval and annotation using maximum entropy. In *Cross-language Evaluation Forum Conference*, pp. 725--734.

dos Santos, J. M.; Cavalcanti, J. M. B.; Saraiva, P. C. & Moura, E. S. d. (2013). Multimodal re-ranking of product image search results. In *European Conference on Advances in Information Retrieval*, pp. 62--73.

dos Santos, K. C. L. (2009). Uma abordagem evolutiva para recuperação de imagens da web. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte.

dos Santos, K. C. L.; de Almeida, H. M.; A., M. G. & da S. Torres, R. (2009). Recuperação de imagens da web utilizando múltiplas evidências textuais e programação genética. In *Simpósio Brasileiro de Banco de Dados*, pp. 91--105.

Fan, Weiguoand Pathak, P. & Zhou, M. (2009). Genetic-based approaches in ranking function discovery and optimization in information retrieval - a framework. *Decision Support Systems*, 47(4):398--407.

Fan, W.; Gordon, M. D. & Pathak, P. (2000). Personalization of search engine services for effective retrieval and knowledge management. In *International Conference on Information Systems*, pp. 20--34.

Fan, W.; Gordon, M. D. & Pathak, P. (2004). Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *IEEE TKDE*, 16(4):523--527.

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

Faria, F. F.; Veloso, A.; Almeida, H. M.; Valle, E.; Torres, R. d. S.; Gonçalves, M. A. & Meira, Jr., W. (2010). Learning to rank for content-based image retrieval. In *International Conference on Multimedia Information Retrieval*, pp. 285--294.

Feldt, R. & Nordin, P. (2000). Using factorial experiments to evaluate the effect of genetic programming parameters. In *European Conference on Genetic Programming*, pp. 271--282.

Ferecatu, M. & Sahbi, H. (2008). Telecom paristech at imageclefphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In *Working Notes for the CLEF 2008 Workshop*.

Ferreira, C. D.; da S. Torres Ricardo; Gonçalves, M. A. & Fan, W. (2008). Image retrieval with relevance feedback based on genetic programming. In *Simpósio Brasileiro de Banco de Dados*, pp. 120--134.

Graf, F. (2012). Jfeaturelib - A free java library containing feature descriptors and detectors. https://JFeatureLib.googlecode.com; Version 1.3.1.

Huang, J.; Kumar, S. R.; Mitra, M.; Zhu, W.-J. & Zabih, R. (1997). Image indexing using color correlograms. In *IEEE Computer Vision and Pattern Recognition*, pp. 762--768.

Iyengar, G.; Duygulu, P.; Feng, S.; Ircing, P.; Khudanpur, S. P.; Klakow, D.; Krause, M. R.; Manmatha, R.; Nock, H. J.; Petkova, D.; Pytlik, B. & Virga, P. (2005). Joint visual-text modeling for automatic retrieval of multimedia documents. In *ACM International Conference on Multimedia*, pp. 21--30.

Kasutani, E. & Yamada, A. (2001). The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *International Conference Image Processing*, pp. 674--677.

Kherfi, M. L.; Ziou, D. & Bernardi, A. (2002). Learning from negative example in relevance feedback for content-based image retrieval. In *IEEE International Conference on Pattern Recognition*, pp. 933--936.

Kherfi, M. L.; Ziou, D. & Bernardi, A. (2004). Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Computing Surveys*, 36(1):35--67.

Kittler, J.; Hatef, M. & Duin, R. P. W. (1996). Combining classifiers. In *International Conference on Pattern Recognition*, pp. 897--901.

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

Lebeda, K.; Matas, J. & Chum, O. (2012). Fixing the locally optimized ransac. In *British Machine Vision Conference*, pp. 1--11.

Liu, Y.; Mei, T. & Hua, X. (2009). Crowdreranking: exploring multiple search engines for visual search reranking. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 500--507.

Liu, Y.; Zhang, D.; Lu, G. & Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262--282.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, pp. 1150--1157.

Lux, M. (2011). Content based image retrieval with LIRE. In *ACM International Conference on Multimedia*, pp. 735--738.

McDonald, K. & Smeaton, A. F. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *International Conference on Image and Video Retrieval*, pp. 61--70.

McGill, M. & Salton, G. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Müller, H.; Geissbühler, A.; Marty, J.; Lovis, C. & Ruch, P. (2005). The use of medgift and easyir for imageclef 2005. In *Cross-language Evaluation Forum Conference*, pp. 724--732.

Niblack, C. W.; Barber, R.; Equitz, W.; Flickner, M. D.; Glasman, E. H.; Petkovic, D.; Yanker, P.; Faloutsos, C. & Taubin, G. (1993). The qbic project: Querying images by content, using color, texture, and shape. In *Storage and Retrieval for Image and Video Database*, pp. 173–187.

Oren, N. (2002). Reexamining tf.idf based information retrieval with genetic programming. In *Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology*, pp. 224--234.

Piji, L. & Jun, M. (2009). Learning to rank for web image retrieval based on genetic programming. In *IEEE International Conference on Broadband Network & Multimedia Technology*, pp. 137--142.

Quack, T.; Mönich, U.; Thiele, L. & Manjunath, B. S. (2004). Cortina: a system for large-scale, content-based web image retrieval. In *ACM International Conference on Multimedia*, pp. 508--511.

Robertson, S. E. & Walker, S. (1999). Okapi/keenbow at trec-8. In *Text REtrieval Conference (TREC-8)*.

Sclaroff, S.; Taycher, L. & Cascia, M. L. (1997). Imagerover: A content-based image browser for the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pp. 2--9.

Shandilya, S. K. & Singhai, N. (2010). Article: A survey on: Content based image retrieval systems. *International Journal of Computer Applications*, 4(2):22--26.

Shen, X.; Lin, Z.; Brandt, J. & Wu, Y. (2012). Mobile product image search by automatic query object extraction. In *European Conference on Computer Vision*, pp. 114--127.

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

Snoek, C. G. M.; Worring, M. & Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In *ACM International Conference on Multimedia*, pp. 399--402.

Stehling, R. O.; Nascimento, M. A. & Falcão, A. X. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. In *International Conference on Information and Knowledge Management*, pp. 102--109.

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

Trotman, A. (2005). Learning to rank. *Information Retrieval*, 8(3):359--381.

Tsymbalenko, Y. & Munson, E. V. (2001). Using html metadata to find relevant images on the world wide web. In *Internet Computing*, pp. 842--848.

van Zaanen, M. & de Croon, G. (2004). FINT: Find images and text. In *Working Notes for the CLEF 2004 Workshop*.

Vani, V. & Raju, S. (2010). A detailed survey on query by image content techniques. In *International Conference on Networking, VLSI and Signal Processing*, pp. 204--209.

Vedaldi, A. & Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms.

Vidal, M.; Cavalcanti, J. M. B.; de Moura, E. S.; da Silva, A. S. & da S. Torres, R. (2012). Sorted dominant local color for searching large and heterogeneous image databases. In *International Conference on Pattern Recognition*, pp. 1960--1963.

Villena-Román, J.; Lana-Serrano, S. & Cristóbal, J. C. G. (2007a). Miracle at imageclefmed 2007: Merging textual and visual strategies to improve medical image retrieval. In *Cross-language Evaluation Forum Conference*, pp. 593--596.

Villena-Román, J.; Lana-Serrano, S.; Martínez-Fernández, J. L. & Cristóbal, J. C. G. (2007b). Miracle at imageclefphoto 2007: Evaluation of merging strategies for multilingual and multimedia information retrieval. In *Cross-language Evaluation Forum Conference*, pp. 500--503.

Voorhees, E. M. & Harman, D. (1999). Overview of the eighth text retrieval conference (trec-8). In *Text REtrieval Conference (TREC-8)*.

Xie, X.; Lu, L.; Jia, M.; Li, H.; Seide, F. & Ma, W.-Y. (2008). Mobile search with multimodal queries. *Proceedings of the IEEE*, 96(4):589--601.

Yao, T.; Mei, T. & Ngo, C. (2010). Co-reranking by mutual reinforcement for image search. In *International Conference on Image and Video Retrieval*, pp. 34--41.

Zhang, R. & Guan, L. (2009). Multimodal image retrieval via bayesian information fusion. In *IEEE International Conference on Multimedia and Expo*, pp. 830--833.

Zhou, X.; Gobeill, J. & Müller, H. (2009). The medgift group at imageclef 2008. In *Cross-language Evaluation Forum Conference*, pp. 712--718.