REALIMENTAÇÃO DE RELEVÂNCIA PARA BUSCA VISUAL EM BASES DE DADOS MULTIMODAIS

PATRÍCIA CORREIA SARAIVA

REALIMENTAÇÃO DE RELEVÂNCIA PARA BUSCA VISUAL EM BASES DE DADOS MULTIMODAIS

Proposta de 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

Dezembro de 2011

Lista de Figuras

2.1 Modelo RGB 9
2.2 Modelo HSV 10
2.3 Precisão *versus* Revocação 12
2.4 Abordagens de fusão precoce e fusão tardia. 13
2.5 Exemplo de um indivíduo representado como uma árvore em PG. 15
2.6 Exemplo de uma operação de cruzamento em PG. 16
2.7 Exemplo de uma operação de mutação em PG. 17

3.1 Distribuição de tamanho dos documentos. 33
3.2 Curva de precisão-revocação para todas as passagens de texto. 35
3.3 Curva de precisão-revocação para as fontes de evidência isoladas no arcabouço Bayesiano. 36
3.4 Curva de precisão para a combinação de múltiplas fontes de evidência no arcabouço Bayesiano. 37
3.5 Curva de precisão-revocação para a comparação do PG e baselines. 38

4.1 Amostra de um objeto da base de dados de produto. 42
4.2 Imagem original à esquerda e imagem segmentada à direita. 43
4.3 Imagem original à esquerda e imagem quantizada à direita. 43
4.4 Planejamento das atividades. 44

Lista de Tabelas

3.1 Terminais usados no arcabouço de PG. 26
3.2 Configuração experimental para o projeto fatorial completo em dois níveis. 29
3.3 Abordagens de ranking para o modelo Bayesiano. 30
3.4 Estatísticas da coleção usada nos experimentos. 32
3.5 Estatísticas da distribuição de tamanho dos documentos. 34
3.6 Medidas de MAP para as passagens de texto. 34
3.7 Medida de MAP para as fontes de evidência isoladas no arcabouço Bayesiano. 36
3.8 Medidas de MAP para as combinações de múltiplas fontes de evidência no arcabouço Bayesiano. 37
3.9 Medidas de MAP e P@N para os arcabouços de PG, Bayesiano e BM25. . 38

Sumário

Lista de Figuras **v**

Lista de Tabelas **vii**

1 Introdução **1**

1.1 Proposta de Tese 3

1.2 Contribuições Esperadas 5

2 Conceitos e Trabalhos Relacionados **7**

2.1 Recuperação de Imagens Baseada em Texto 7

2.2 Recuperação de Imagens Baseada em Conteúdo 8

2.2.1 Cor 9

2.3 Medidas de Precisão 11

2.4 Fusão Multimodal 13

2.5 Realimentação de Relevância 14

2.6 Programação Genética 15

2.7 Trabalhos Relacionados 16

2.7.1 Recuperação de Imagens 17

2.7.2 Fusão Multimodal 19

2.7.3 Realimentação de Relevância 20

3 Programação Genética para Fusão de Evidências **25**

3.1 Recuperação de Imagens Baseada em Programação Genética 25

3.1.1 Fontes de Evidência Textual 26

3.2 Experimentos 27

3.2.1 Parametrização do Arcabouço de PG 27

3.2.2 Baselines 30

3.2.3 Coleção 32

3.2.4 Experimentos com o Modelo Bayesiano 33

3.2.5 Resultados Experimentais com o Arcabouço de PG 37
3.3 Conclusões 38

4 Próximos Passos **39**
4.0.1 Planejamento das Atividades 44

Referências Bibliográficas **45**

Capítulo 1

Introdução

O uso de imagens para recuperar informação tem se tornado cada vez mais frequente nos últimos anos. A popularização de dispositivos portáteis que permitem conjuntamente a captura de imagens e acesso à Internet tem motivado o surgimento de novos cenários de busca que utilizam imagens para recuperar informação relevante. Exemplos de tais dispositivos são os smartphones, tablets e celulares. Embora a forma mais convencional para buscar informação ainda seja através de uma consulta textual, a popularização destas tecnologias aliada à grande quantidade de imagens disponíveis em sistemas oferecidos via Web, tem demandado a criação de novas aplicações e o aprimoramento de aplicações já existentes, que fazem uso de imagens como consulta para identificar e atender uma necessidade de informação do usuário.

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]. Nesta abordagem, conhecida como *(TBIR - Text-based Image Retrieval)*, anotações textuais fornecidas manualmente, ou ainda extraídas automaticamente dos documentos que contem as imagens são utilizadas para indexar e recuperar tais imagens. Problemas como anotações incompletas e imprecisas, ou o esforço necessário para anotar manualmente grandes bases de imagens são apontados como as principais desvantagens desta abordagem.

Numa tentativa de transpor as dificuldades e desvantagens da abordagem anterior, pesquisadores buscaram desenvolver técnicas de recuperação baseadas exclusivamente no conteúdo visual das imagens, mais conhecidas como *(CBIR - Content-Based Image Retrieval)*. Diversas técnicas baseadas no conteúdo visual foram desenvolvidas e publicadas na literatura nos últimos anos [Smeulders et al., 2000; Datta et al., 2008; Vani & Raju, 2010]. Embora alguns avanços tenham sido obtidos nesta área, a tarefa de recuperar informação relevante a partir de uma imagem fornecida pelo usuário segue

sendo um grande desafio. Um dos motivos é a baixa efetividade de recuperação apresentada pelos sistemas de CBIR que comprovam a dificuldade em capturar a semântica das imagens digitais utilizando apenas propriedades visuais.

Embora as abordagens de recuperação de imagens baseada em informação textual apresentem resultados mais efetivos se comparados às abordagens puramente visuais, a busca baseada somente em texto nem sempre apresenta resultados satisfatórios, ou nem sempre é considerada o cenário de busca ideal. Neste último caso, podemos citar como exemplo, aplicações onde um usuário deseja obter informação sobre um objeto, pessoa ou lugar a partir de uma foto tirada com seu dispositivo móvel e muitas vezes 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 é considerado inconveniente. Mesmo quando há a possibilidade de entrada de uma consulta textual, existem ainda problemas relacionados à semântica capturada através da informação textual. Muitas vezes a ocorrência de palavras ambíguas ou mesmo irrelevantes para a imagem atrapalham o processo de recuperação de imagens baseada exclusivamente em texto.

Por outro lado, as abordagens de busca que incorporam apenas propriedades de baixo-nível como cor, forma ou textura, são capazes de recuperar imagens visualmente similares mas que nem sempre estão semanticamente relacionadas com a consulta (*semantic gap*). Isto acontece porque diferentes usuários podem ter diferentes percepções sobre uma mesma imagem e nem sempre são capazes de descrever sua necessidade de informação através de propriedades de baixo-nível. Neste cenário, o envolvimento do usuário no processo de recuperação passou a ser incorporado em sistemas de CBIR através de técnicas de realimentação de relevância (RF) numa tentativa de transpor esta barreira semântica.

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 bastante estudada nos últimos anos [Snoek et al., 2005; Zhang & Guan, 2009; Clinchant et al., 2011; Cheng et al., 2011; Arampatzis et al., 2011a,b; Depeursinge & Müller, 2010]. Esta combinação de duas ou mais modalidades de dados tem sido referenciada na literatura como fusão multimodal. Diversos 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 estas 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.

1.1 Proposta de Tese

O objetivo principal desta proposta é estudar e propor novas estratégias de realimentação de relevância e fusão multimodal com o intuito de melhorar a qualidade das respostas providas por sistemas de busca por imagens. Nesta proposta, o termo multimodal faz referência à capacidade de representar, processar e analisar duas modalidades de dados simultaneamente: texto e imagens. A ideia é estudar mecanismos de expansão de consultas de forma automática ou guiada por usuários, através de realimentação de relevância, com o objetivo de tornar mais efetivos os sistemas de busca por imagens em grandes bases dados. Além disso, explorar técnicas de fusão de informação multimodal utilizando um método de aprendizagem indutivo, a Programação Genética [Koza, 1992], para a descoberta de novas formas de fusão.

Mais especificamente, será estudado o problema onde deseja-se recuperar informação a partir de uma imagem fornecida pelo usuário. Nesse contexto, a imagem fornecida pode ser de baixa qualidade, muitas vezes capturada sob condições adversas de iluminação e ângulos que nem sempre ajudam na identificação de objetos dentro de uma base de dados.

Como aplicação alvo, será estudada a busca visual por produtos. A busca visual por produtos tem surgido como uma nova possibilidade de pesquisa na área de recuperação de imagens. Isto porque os sites de venda on-line de produtos lidam com uma grande quantidade de informação visual e estão cada vez mais populares motivados principalmente pela comodidade nas compras e pela diversidade dos produtos oferecidos. A escolha do problema de busca por produtos se deu pelo fácil acesso que temos a bases de dados e consultas a um sistema de busca por produtos, devido a uma parceria do nosso grupo de pesquisa com uma empresa que atua no segmento de busca para lojas on-line.

Neste tipo de aplicação, o usuário fornece a imagem de um objeto, geralmente uma foto tirada a partir de um celular ou de outro dispositivo móvel, e deseja obter informações sobre o objeto capturado. Por exemplo, no caso de busca por produtos, o usuário fornece uma foto de um produto e espera receber informações tais como o preço, características técnicas ou configurações alternativas. Muitas vezes, a imagem inicial pode apresentar problemas diversos de qualidade ou conter objetos que não ocorrem

na base de dados onde será realizada a busca. Neste caso, a chance do processo inicial da busca baseada apenas no conteúdo ter pouca precisão é grande, o que dificultaria a aplicação de técnicas de fusão multimodal que têm sido propostas na literatura, as quais dependem muito da qualidade inicial obtida a partir da consulta original.

Uma das alternativas que pretendemos estudar é a criação de mecanismos de realimentação de relevância multimodal que considerem as características específicas da aplicação e da base de dados. Acreditamos que a realimentação de relevância possa desempenhar papel importante em aplicações de identificação e recuperação de informação de objetos a partir de imagens fornecidas por usuários. Até onde sabemos, a realimentação de relevância multimodal a partir de uma imagem de consulta tem sido pouco explorada, abrindo assim um amplo leque de possibilidades de pesquisa na área.

Nossa proposta para contornar o problema é realizar um estudo considerando a hipótese do usuário poder realimentar o sistema a partir do conjunto resposta inicial recuperado a partir da imagem fornecida na busca. Essa realimentação serviria como mecanismo para definir melhor a consulta e aumentar as chances do sistema encontrar a informação desejada pelo usuário. A realimentação fornecida pelo usuário traria não só uma melhor definição visual da consulta, mas também informações sobre o contexto semântico da busca que podem estar associadas às imagens fornecidas para realimentação. No caso da busca de produtos, a informação associada à categoria das imagens fornecidas para realimentação seriam utilizadas para refinar a consulta, o que ajudaria a eliminar opções provenientes de outras categorias. Com esse trabalho de pesquisa desejamos saber se é possível obter bons resultados utilizando estratégias de fusão multimodal e realimentação de relevância neste cenário.

Uma pergunta importante a ser respondida é saber quais as melhores técnicas de realimentação de relevantes para o cenário de busca a ser estudado e se há a necessidade de se propor novos métodos de fusão multimodal para atender à especificidades da aplicação que temos como alvo.

Apesar do foco principal desta pesquisa estar baseado na busca visual de produtos, espera-se que os resultados de pesquisa obtidos possam ser extrapolados para outros tipos de coleções multimídia que possuam características similares. Por exemplo, além da busca por produtos, usuários podem demandar outras aplicações que exijam a identificação de objetos a partir de uma imagem fornecida pelo usuário. É provável que parte dos resultados a serem obtidos no estudo com a aplicação alvo possam ser generalizados e utilizados em outras aplicações. Pretendemos realizar experimentos dentro desse trabalho para verificar essa possibilidade sempre que possível.

Finalmente, como ponto de partida para o trabalho, desenvolvemos até o momento um estudo sobre o uso de programação genética na combinação de evidências

textuais para a busca por imagens. A programação genética deve desempenhar um papel central em nossa proposta, visto que será o mecanismo utilizado ao longo do trabalho para combinar informações visuais e não-visuais para a geração de respostas ordenadas. Por essa razão, o estudo sobre o uso de programação genética em busca por imagens será de grande importância para os próximos passos a serem desenvolvidos no trabalho.

1.2 Contribuições Esperadas

Algumas das contribuições esperadas com este trabalho de pesquisa são:

- Estudo e proposta de novos mecanismos de realimentação de relevância em bases de dados multimodais;
- Implementação de um sistema de busca visual para produtos com mecanismo de realimentação de relevantes para bases de dados multimodais;
- Uso de Programação Genética para encontrar novas formas de fusão multimodal;
- Descoberta de quais evidências (visuais e textuais) contribuem para melhorar a recuperação de imagens em uma base multimodal;
- Avaliação de desempenho das abordagens propostas em termos de eficácia e eficiência utilizando usuários reais;

Capítulo 2

Conceitos e Trabalhos Relacionados

Este capítulo apresenta os principais conceitos teóricos utilizados ao longo desta proposta de tese além dos trabalhos relacionados à área de abrangência deste trabalho.

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 [Coelho et al., 2004]. Nesta abordagem, as imagens da coleção eram inicialmente rotuladas manualmente e a busca era realizada utilizando sistemas de gerenciamento de banco de dados [Chang & Fu, 1980; Chang & Kunii, 1981]. Com o crescimento do tamanho das bases de imagens motivado principalmente pelos avanços nas tecnologias para aquisição de imagens, este processo de anotação manual tornou-se uma tarefa onerosa e inviável.

Com o surgimento da Web, a recuperação de imagens baseada em texto passou a utilizar conteúdo das páginas baseada na suposição de que o texto associado pode descrever a semântica das imagens embutidas na página. Este texto extraído diretamente do HTML é utilizado para indexar e recuperar as imagens. Grandes sistemas de busca de imagens, como Google correlacionam palavras-chave com as imagens embutidas baseando-se na sua importância e posições relativas às imagens.

Estas evidências são associadas às imagens embutidas na página Web e utilizadas na indexação da base de imagens. Nesta abordagem, uma consulta textual é fornecida pelo usuário e utilizada pelo sistema de busca para recuperar as imagens da coleção que possuem as mesmas palavras da consulta. O sistema funciona como uma máquina de busca tradicional e os resultados são apresentados em ordem de relevância da consulta com a informação armazenada na base de imagens. A similaridade entre a consulta e as imagens retornadas é computada usando medidas clássicas de ponderação como

similaridade do cosseno, Okapi-BM25 e *bag of words* [Baeza-Yates & Ribeiro-Neto, 2011].

2.2 Recuperação de Imagens Baseada em Conteúdo

Sistemas de recuperação de imagens baseados em conteúdo 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. Neste caso, os descritores de imagens são utilizados para extrair tais características das imagens, viabilizando assim as fases de indexação e busca. Formalmente, um descritor D é constituído por um par (ϵ_D, δ_D), onde:

- $\epsilon_D : I \rightarrow \mathbb{R}^n$ 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.

Nesta abordagem, o usuário geralmente provê uma imagem-consulta cujas propriedades visuais são comparadas com as propriedades visuais das imagens da coleção. A similaridade entre duas imagens é computada através de uma função de distância entre seus vetores de características.

Uma questão fundamental na recuperação de imagens baseada em conteúdo é decidir quais propriedades visuais que melhor se adequam ao domínio da aplicação, uma vez que isso pode melhorar consideravelmente a eficiência na recuperação. Para algumas aplicações de busca de imagens de domínio específico, a escolha das propriedades pode-se basear 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.

As propriedades visuais comumente utilizadas na recuperação de imagens baseada em conteúdo são cor, forma e textura. Dentre elas, a cor é a propriedade mais utilizada com diversos descritores publicados na literatura. Alguns descritores baseados em cor

bastante populares são o histograma global/local de cor [Swain & Ballard, 1991], vetor de coerência de cor [Pass et al., 1996] e momentos de cor [Stricker & Orengo, 1995].

2.2.1 Cor

A informação de cor pode ser representada em diferentes espaços de cores, como RGB, HSV ou YIQ. Tipicamente, um espaço de cor é um modelo matemático utilizado para representar a cor em termos de níveis de intensidade [Wang, 2001], cuja dimensionalidade pode variar de uma a quatro dimensões. Cada componente de cor, ou canal de cor, representa uma das dimensões.

Os espaços de cor mais utilizados em sistemas de CBIR são o modelo RGB (*Red, Green, Blue*) e o HSV (*Hue, Saturation, Value*). O modelo RGB é um modelo aditivo no qual a cor desejada é produzida a partir da soma das cores primárias vermelho, verde e azul. O modelo RGB utiliza o sistema de coordenadas cartesianas como apresentado na Figura 2.1.

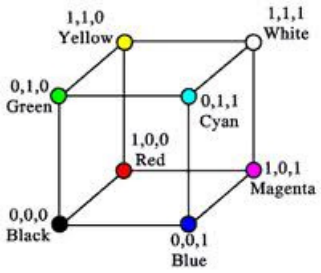

Figura 2.1. Modelo RGB

O modelo HSV utiliza um sistema de coordenadas na forma de um cone como apresentado na Figura 2.2. No HSV, o componente matiz (H) varia de 0 a 360 e as cores correspondentes variam de vermelho, amarelo, verde, ciano, azul, magenta e novamente vermelho. No componente de saturação (S), as cores variam de totalmente dessaturadas (tons de cinza) a totalmente saturadas (sem influência da cor branca). O componente valor (V) varia de 0 a 1.0 e as cores se tornam mais brilhantes na medida em que o valor de V aumenta. O componente de valor (V) representa a intensidade de uma cor, e é desassociado da informação de cor na imagem representada. Por sua vez, os componentes de matiz (H) e saturação (S) estão intimamente relacionados à forma com que o olho humano percebe as cores e por isto é bastante utilizado na recuperação de imagens.

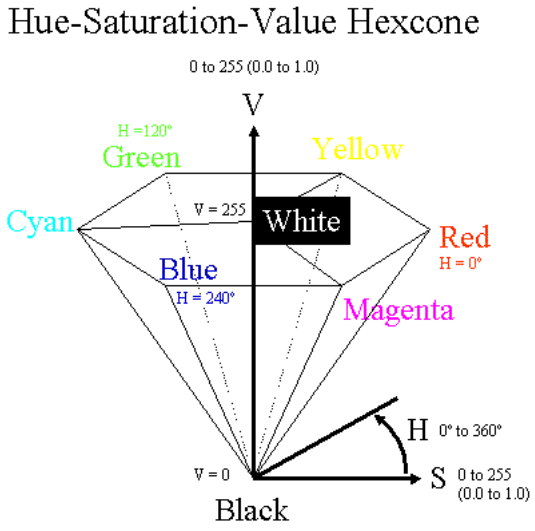

Figura 2.2. Modelo HSV

A utilização do espaço de cor HSV requer a conversão dos componentes de cor RGB entre os dois espaços. Segundo [Gonzalez & Woods, 2008], a conversão pode ser obtida de forma não-linear através das seguintes fórmulas:

$$H = cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right\} \tag{2.1}$$

$$S = 1 - \frac{3}{R+G+B}[min(R,G,B)] \tag{2.2}$$

$$V = \frac{1}{3}(R+G+B) \tag{2.3}$$

O desempenho de diferentes espaços de cor em sistemas de CBIR foram comparados em Wang [2001]; Sural et al. [2002]; Yu et al. [2003]. Dentre os espaços de cor analisados, o modelo HSV apresenta, em média, melhor eficácia quando comparados com outros espaços de cor como RGB, CIELab e CIELuv.

Quantização do Espaço de Cor

Dada uma imagem digital I, formada por um conjunto finito de pixels P, onde $P \subset \mathbb{Z}^2$, cada pixel $p_i \in P$ possui um valor numérico que representa o seu nível de intensidade ou cor. Cada cor é geralmente codificada utilizando-se 8 bits por cada canal de cor. Considerando um espaço de cor de 3 canais, como é o caso do RGB, uma imagem pode conter até $2^{8\times3} = 16.777.216$ cores distintas. Esta alta dimensionalidade do espaço de cor resulta em um processo demorado durante as fases de extração e comparação das

propriedades visuais, além de demandar mais espaço para armazenamento do vetor de características. Portanto, a redução do espaço de cor deve ser realizada para diminuir os custos computacionais envolvidos em todas as fases de um sistema de CBIR.

O processo de reduzir o número de cores distintas em uma imagem digital é conhecido como quantização do espaço de cores. O resultado esperado da quantização é que a imagem resultante seja visualmente tão similar quanto possível à imagem original, sem perda significativa de qualidade e de informação. O efeito da quantização de cores na efetividade da recuperação de imagens é reportado em diversos trabalhos com diferentes esquemas de quantização Wang [2001].

A quantização escalar (ou uniforme) é o processo de redução do espaço de cor mais simples e rápido e é totalmente independente da distribuição das cores na imagem. Cada componente de cor original $c_i = (R_i, G_i, B_i)$ na faixa de $[0 \cdots m-1]$ é independentemente convertido para uma novo componente de cor c'_i na faixa de $[0 \cdots n-1]$ aplicando-se a seguinte transformação:

$$c'_i = \left\lfloor c \times \frac{n}{m} \right\rfloor \tag{2.4}$$

O processo de quantização é fundamental em sistemas de CBIR pois permite reduzir a dimensão do espaço de cor, e consequentemente, reduzir o tamanho dos vetores de características, com o objetivo de minimizar os custos envolvidos no armazenamento e nos cálculos de similaridade entre os vetores.

2.3 Medidas de Precisão

A eficácia de um sistema de recuperação é uma medida relacionada à satisfação do usuário com o resultado obtido pelo sistema de recuperação. Para estabelecer uma medida de eficácia, a primeira decisão a ser tomada é definir quais julgamentos são permitidos ao usuário no momento da sua avaliação. A escolha básica é entre uma medida binária e uma medida de múltipla escolha, onde o usuário pode definir graus de relevância para um determinado resultado. A medida binária é a escolha mais comum de ser implementada, na qual cada imagem pode ser aceita ou rejeitada. Estas condições estão normalmente ligadas à relevância de uma imagem para um usuário. Entretanto, estas condições apresentam um alto grau de subjetividade, uma vez que diferentes usuários, ou os mesmos usuários em diferentes circunstâncias, podem perceber o conteúdo visual de uma imagem de diferentes maneiras.

As medidas de avaliação utilizadas nos sistemas de recuperação de imagens em geral, foram originalmente desenvolvidas para avaliar sistemas de recuperação de do-

cumentos. A utilização destas medidas é 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.

Entre as medidas de avaliação de desempenho existentes, a curva de Precisão *versus* Revocação (P x R) [Baeza-Yates & Ribeiro-Neto, 2011] é uma das mais conhecidas e utilizadas na prática.

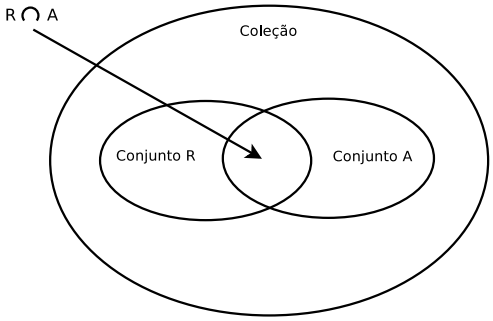

Figura 2.3. Precisão *versus* Revocação

De acordo com a Figura 2.3, a precisão P é definida pela fração dos documentos recuperados (o conjunto A) que é relevante:

$$P = \frac{\mid R \cap A \mid}{\mid A \mid} \tag{2.5}$$

A revocação R é a fração dos documentos relevantes (o conjunto R) que foi recuperado:

$$R = \frac{\mid R \cap A \mid}{\mid R \mid} \tag{2.6}$$

Outra medida de avaliação importante é o MAP *Mean Average Precision* Baeza-Yates & Ribeiro-Neto [2011], que representa a média das precisões sobre um conjunto de consultas:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{2.7}$$

onde Q é o número de consultas. $AveP(q)$ é dado por:

$$AveP(q) = \frac{\sum_{r=1}^{N}(P(r) * rel(r))}{R} \tag{2.8}$$

onde N é o número de documentos recuperados, R é o número de documentos relevantes para a consulta q, $rel(r)$ é uma função binária da relevância do documento

r e $P(r)$ é a precisão de r em um determinado ponto.

Neste trabalho usaremos estas medidas para avaliação da eficácia dos métodos propostos. Na medida do possível, outras métricas poderão ser utilizadas para comparação com outras abordagens propostas.

2.4 Fusão 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. Os resultados também demonstram que sistemas de recuperação alcançam melhor desempenho caso a fusão de ambas modalidades de dados sejam exploradas 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*). Estas 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 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 4.1 ilustra estas duas abordagens de fusão.

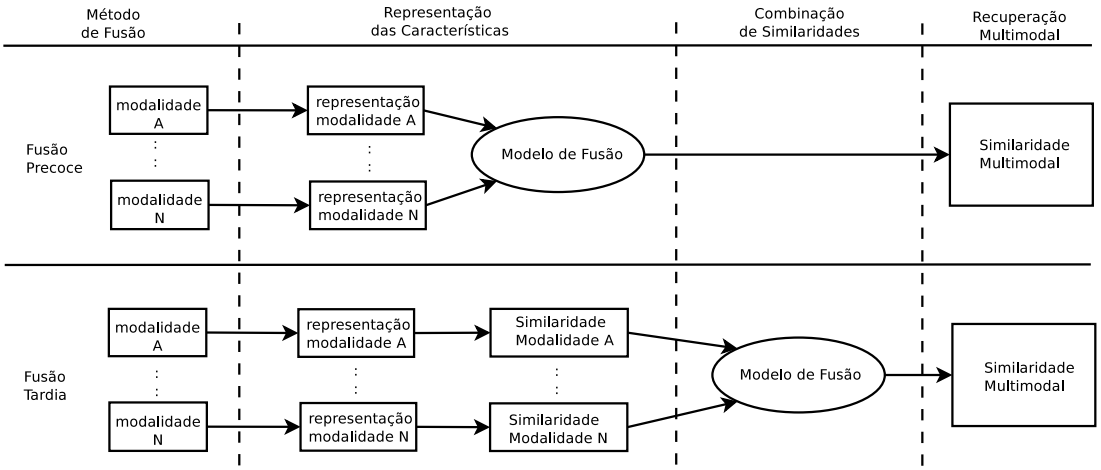

Figura 2.4. Abordagens de fusão precoce e fusão tardia.

O método de fusão precoce mais simples consiste em concatenar as representações textuais e visuais das imagens. Abordagens mais elaboradas utilizam esquemas de ponderação das características [van Zaanen & de Croon, 2004; Deselaers et al., 2005, 2007].

Trabalhos de fusão tardia incluem estratégias de fusão baseada na ordenação ou baseada na similaridade das respostas [Arampatzis et al., 2011a,b; Clinchant et al., 2011]. Outras abordagens se baseiam em interseção dos resultados das diferentes modalidades [Villena-Román et al., 2007a,b; Müller et al., 2005] ou na combinação linear das similaridades das respostas Depeursinge & Müller [2010]. Outras abordagens utilizam regras para reordenação onde os documentos recuperados textualmente são reordenados baseados na similaridade visual, ou vice-versa [Zhou et al., 2009; chen Chang & hsi Chen, 2007]. Mais detalhes sobre técnicas de fusão multimodal podem ser encontradas em [Depeursinge & Müller, 2010].

2.5 Realimentação de Relevância

Uma forma de identificar uma necessidade de informação do usuário é utilizar mecanismos de realimentação de relevância. Em geral, realimentação de relevância é qualquer informação fornecida iterativamente pelo usuário (*Relevance Feedback* - *RF*) sobre os resultados recuperados.

A realimentação de relevância tem sido utilizada com sucesso no desenvolvimento de sistemas em diversas áreas de pesquisa, como: recuperação de informação, recomendação, classificação e anotação de dados. Pesquisas que utilizam a realimentação de relevância aplicam diferentes técnicas, como redes semânticas [Lu et al., 2000], combinação linear [Li & Yuan, 2004; Rui et al., 1998] e não-linear de características, agrupamento de dados [Kim et al., 2005], redes neurais [Ko & Byun, 2002], máquinas de vetores de suporte [Gondra & Heisterkamp, 2004; Hong et al., 2000] e programação genética [dos Santos et al., 2008; Ferreira et al., 2008].

Algorithm 1

```
1: O usuário fornece uma consulta inicial q;
2: Apresente um conjunto inicial de respostas mais similares à consulta q;
3: while o usuário não estiver satisfeito do
4:    Usuário indica seu julgamento de relevância no conjunto recuperado;
5:    Atualize a consulta q;
6:    Reordene os resultados da coleção baseada na nova consulta q;
7:    Apresente o novo conjunto resposta;
8: end while
```

Abordagens que utilizam realimentação de relevância seguem um fluxo de execução tradicional como apresentado no Algoritmo 1.

2.6 Programação Genética

Programação Genética (PG) Koza [1992] constitui um conjunto de técnicas da inteligência artificial para a solução de problemas baseadas nos princípios da herança biológica, seleção natural e evolução. Neste contexto, cada solução potencial é chamada de indivíduo em uma população. Sobre essa população são aplicadas sucessivas transformações genéticas, como cruzamento, mutação e reprodução com o intuito de criar novos indivíduos mais aptos à solução do problema em gerações subsequentes. Uma função de adequação (*fitness function*) é utilizada para atribuir valores para cada indivíduo com o intuito de definir o seu grau de adequação, ou evolução, perante os demais membros da população.

Em essência, PG evolui um número de soluções candidatas, os indivíduos, representados em memória através de estruturas de árvores binárias. Cada nó interno da árvore é uma função, como por exemplo, operações numéricas $+, -, /, *, \sqrt{}, \log$. E cada nó folha, conhecido como terminal, representa uma variável ou uma constante. O número máximo de nodos em um indivíduo é determinado pela profundidade da árvore, que é definida antes do início do processo de evolução. Um exemplo de indivíduo representado por uma estrutura de árvore é provido na Figura 2.5.

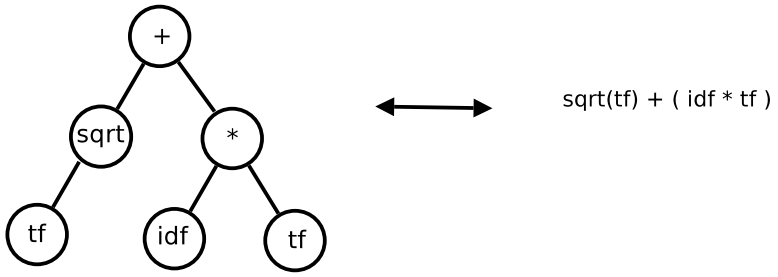

Figura 2.5. Exemplo de um indivíduo representado como uma árvore em PG.

O Algoritmo 2 apresenta um esquema de funcionamento da Programação Genética. Inicialmente é criada aleatoriamente uma população de indivíduos como soluções potenciais do problema. A partir deste ponto, enquanto o critério de parada para o algoritmo não for atingido, são realizadas avaliações sobre a qualidade de cada individuo na população atual. Para isto, é atribuído um valor para cada indivíduo através de uma medida de aptidão (*fitness*). Após a avaliação da população, sucessivas transformações, como cruzamento, mutação e reprodução, são aplicadas com o objetivo

de criar indivíduos mais diversos e com melhor eficácia nas gerações subsequentes. A nova população é avaliada e o ciclo se repete até que o critério de parada seja satisfeito.

Algorithm 2

1: Geração aleatória de uma população inicial de indivíduos;
2: **while** O critério de parada não for satisfeito **do**
3: Avalie cada indivíduo da população atual;
4: Selecione os melhores indivíduos para aplicar as transformações genéticas;
5: Aplique reprodução;
6: Aplique cruzamento;
7: Aplique mutação;
8: **end while**

A operação de reprodução sobre os indivíduos selecionados consiste em mantê-los inalterados, garantindo a sua permanência na próxima geração. A operação de cruzamento pode ser visualizada na Figura 2.6. Esta operação permita a troca de conteúdo genético entre dois outros indivíduos, os pais. Em um processo de PG, dois indivíduos pais são selecionados de acordo com a política de seleção. Em seguida, uma sub-árvore de cada pai é selecionada aleatoriamente. Os indivíduos filhos deste transformação genética são o resultado da troca das sub-árvores entre os indivíduos pais.

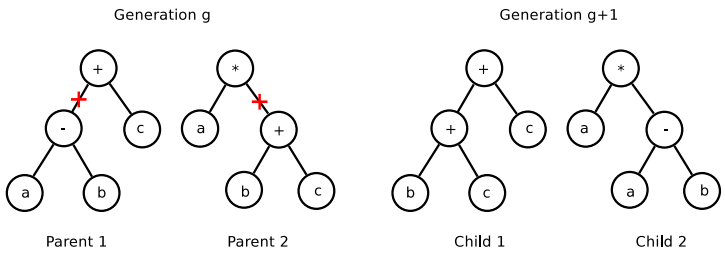

Figura 2.6. Exemplo de uma operação de cruzamento em PG.

O operador de mutação tem o papel de manter um nível de diversidade mínima dos indivíduos em uma população. Nesta operação, uma sub-árvore escolhida aleatoriamente em um indivíduo é trocada por uma nova sub-árvore também criada aleatoriamente. A operação de mutação é ilustrada na Figura 2.7.

2.7 Trabalhos Relacionados

A recuperação de imagens tem sido estudada em diversos contextos ao longo dos últimos anos. Resumos detalhados da área podem ser encontrados em [Kherfi et al., 2004; Datta

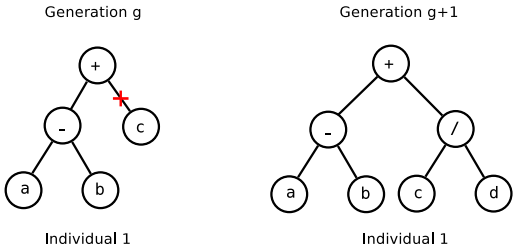

Figura 2.7. Exemplo de uma operação de mutação em PG.

et al., 2008]. Com o simples objetivo de facilitar a leitura, alguns trabalhos foram agrupados de acordo com o tema mais relacionado. Na seção 2.7.1 descrevemos alguns trabalhos sobre busca de imagens de forma mais geral, na seção 2.7.2 descrevemos alguns trabalhos com fusão multimodal e na seção 2.7.3 descrevemos alguns trabalhos com realimentação de relevância.

2.7.1 Recuperação de Imagens

Um sistema de recuperação de imagens na Web chamado Cortina[1] foi 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. No sistema Cortina, as imagens são associadas ao texto extraído das páginas Web (*tag* ALT e passagens de 50 termos ao redor da imagem). Esta informação textual junto com dados adicionais sobre o tamanho da imagem, são armazenados em um banco de dados relacional e as imagens propriamente ditas são armazenadas em um servidor de armazenamento. As características visuais são extraídas utilizando quatro descritores do padrão MPEG-7: descritor de textura *Texture Descriptor* - *HTD*, descritor de histograma de arestas *Edge Histogram Descriptor* - *EHD*, descritor de cor escalável *Scalable Color Descriptor* - *SCD* e descritor de cor dominante *Dominant Color Descriptor* - *DCD*. A busca no espaço visual é feita utilizando o algoritmo *K-NN*.

Jing & Baluja [2008] propuseram um novo algoritmo para busca de imagens chamado *VisualRank* que ordena as imagens baseada nas suas similaridades visuais. A ideia dos autores era encontrar temas visuais em comum em um determinado conjunto de imagens e então encontrar aquelas imagens que melhor representassem os temas. O objetivo principal do *VisualRank* é identificar os nós autoridades em um grafo inferido a partir da similaridade visual entre as imagens e aquelas identificadas como autoridades são consideradas as melhores respostas para uma determinada consulta. No *VisualRank*, as características locais das imagens são extraídas por meio do método

[1] http://cortina.ece.ucsb.edu/

SIFT *(Scale Invariant Feature Transform)* e a similaridade entre duas imagens é medida pelo número de características locais compartilhadas entre estas duas imagens normalizado pelo número médio dos pontos de interesse. Estas similaridades são consideradas as ligações visuais probabilísticas entre as imagens e utilizadas na construção do grafo de similaridade visual. Este algoritmo é similar ao algoritmo *PageRank* Brin & Page [1998] utilizado em páginas Web. O algoritmo VisualRank pode ser usado em nossos trabalhos futuros como um possível ponto de partida para a busca a partir de uma imagem, sendo usado para criar uma ordenação inicial de respostas para servir de entrada no processo de realimentação de relevância.

O uso de Programação Genética Koza [1992] tem sido explorado com sucesso em diversos trabalhos da área de RI [Oren, 2002; Fan et al., 2000, 2004, 2005; Trotman, 2005; de Almeida et al., 2007; Fan et al., 2009]. Na área de recuperação por imagens, o uso de PG também tem se mostrado eficiente no sentido de encontrar novas formas de combinação de evidências e geração de funções de ordenação mais efetivas. O trabalho de [da S. Torres et al., 2009] introduziu o uso de Programação Genética para recuperação de imagens baseada em conteúdo. O arcabouço proposto utilizou diversos descritores de imagens e usou os princípios de PG para descobrir funções de combinação mais efetivas para recuperar imagens baseada em características de forma.

Em Li & Ma [2009], uma abordagem baseada em programação genética foi utilizada para recuperação de imagens na Web. Os autores propuseram um arcabouço evolucionário baseado em programação genética (PG) [Koza, 1992] que combina diversas evidências tais como, texto, conteúdo das imagens, análise de *links* e informação temporal das imagens e das páginas. Os autores utilizaram o mesmo algoritmo *VisualRank* apresentado em [Jing & Baluja, 2008] para representar visualmente as imagens. O trabalho constatou que o uso de programação genética é capaz de gerar funções de ordenação mais eficazes para busca de imagens quando comparados com outros métodos propostos na literatura.

O trabalho de [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 e se mostraram superiores na qualidade das respostas obtidas. Da mesma forma que em Li & Ma [2009] e Santos et al. [2009], nós também utilizaremos GP como técnica para a combinação de evidências. Por esta razão, o Capítulo 3 desta proposta apresenta uma extensão ao trabalho proposto por Santos et al. [2009], com um estudo prévio que fizemos utilizando GP para combinar evidências em um problema de busca por imagens na Web. Esse estudo serviu para termos um melhor entendimento a respeito do funcionamento e dos parâ-

metros da programação genética, servindo portanto de base para nosso trabalho futuro na combinação de evidência dentro da nossa proposta de realimentação de relevância em ambientes multimodais.

2.7.2 Fusão Multimodal

Diversas estratégias de fusão multimodal têm sido estudadas para realizar a combinação dos resultados obtidos em cada modalidade na recuperação de 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.

Em [Kittler et al., 1996] são apresentadas várias estratégias de fusão tardia incluindo a combinação através 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]; 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.

Apesar de interessantes, os trabalhos apresentados acima sobre fusão multimodal têm pouca chance de funcionar bem em um ambiente de recuperação de imagens no qual a consulta inicial é formada apenas por uma imagem fornecida pelo usuário, principal cenário a ser estudado nesta tese. Isso ocorre porque nos trabalhos de fusão multimodal existe um passo que visa expandir automaticamente uma resposta inicial obtida com a consulta fornecida pelo usuário. A expansão automática é realizada utilizando-se as primeiras respostas fornecidas inicialmente para a consulta. Para que esse tipo de técnica funcione, há uma premissa implícita de que o topo da resposta inicial, as primeiras posições, contém um conjunto de respostas relevantes. Entretanto, no caso em que a consulta inicial é uma imagem, esta premissa nem sempre é válida, o que inviabiliza o uso de técnicas de expansão automática.

2.7.3 Realimentação de Relevância

Realimentação de Relevância (RF) Datta et al. [2008] é uma técnica proposta inicialmente para recuperação de documentos que tem sido usada com grande sucesso em CBIR. RF aborda duas questões em relação à recuperação de imagens baseada em conteúdo. O primeiro é o problema do *semantic gap* entre as propriedades de alto-nível das imagens e propriedades de baixo-nível usadas para representá-las. Outra questão está relacionada com a questão da subjetividade de percepção da imagem, pois diferentes usuários podem ter percepções visuais distintas sobre uma mesma imagem.

Uma dos primeiros métodos de realimentação de relevância em CBIR foi proposto em [Rui et al., 1998]. Neste trabalho, o processo de aprendizagem é baseado em atribuir pesos para cada descritor e também para cada elemento do vetor de características. O algoritmo de aprendizagem estima de forma heurística pesos que melhor interpretam a necessidade de informação do usuário no processo de busca. Estes mesmos autores empregaram novamente a estratégia de atribuição de pesos em [Rui & Huang, 2000]. Entretanto, uma otimização no arcabouço é aplicada para estimar os pesos baseada na minimização da distância Euclidiana Generalizada. Além disso, esta técnica utiliza uma abordagem conhecida com *query point movement*, que tenta estimar o vetor de característica do padrão de consulta que melhor representa a percepção do usuário.

O trabalho de Chen et al. [2001] utiliza uma abordagem de recuperação mista, que combina o uso de características visuais e textuais para construir um modelo chamado de *Document Space Model*, que é uma representação das imagens usando um conjunto de vetores das características extraídas. Adicionalmente os autores utilizam técnicas de mineração de dados e realimentação de relevância para construir um segundo modelo chamado de *User Space Model* com o objetivo de melhorar a performance de recuperação do sistema. O resultado final da aplicação é a construção de outro espaço vetorial, chamado de modelo de espaço do usuário *(User Space Model)* que é gerado a partir da informação minerada dos logs de dados fornecidos pelo usuário. Em Chen et al. [2001], os resultados iniciais da combinação das características de texto e conteúdo das imagens são apresentados para o usuário, que define quais imagens da resposta são relevantes ou irrelevantes de acordo com seu critério de subjetividade. A partir desta iteração inicial com o usuário, um método de realimentação de relevância é aplicado para modificar a consulta para adivinhar a real intenção do usuário usando a fórmula tradicional de Rocchio definida como:

$$Q' = Q + \beta\frac{\Sigma Q^+}{n^+} - \gamma\frac{\Sigma Q^-}{n^-} \tag{2.9}$$

onde Q é a consulta original, Q^+ é o conjunto de exemplos positivos (relevantes),

Q^- é o conjunto de exemplos negativos (irrelevantes), n^+ é o tamanho de Q^+, n^- é o tamanho de Q^-, Q' é a consulta modificada e β e γ são constantes de ajuste. As ideias propostas por Chen et al. [2001] vão servir de ponto de partida para o nosso trabalho futuro com técnicas de realimentação de relevância.

O trabalho de [Doulamis & Doulamis, 2006] também utilizaram uma otimização no arcabouço para atribuir pesos internamente nos vetores de características das imagens. Entretanto, ao invés de usar a distância Euclidiana Generalizada, o método deles objetivava maximizar a correlação cruzada normalizada entre a imagem de consulta e cada imagem relevante encontrada durante o processo de recuperação. Os autores argumentam que a correlação cruzada normalizada é mais confiável que a distância Euclidiana devido a invariância em relação à escala e translação do vetor de características.

Outro trabalho pioneiro na área é o sistema PicHunter apresentado por [Cox et al., 2000]. O sistema PicHunter utiliza um arcabouço Bayesiano no processo de aprendizagem. Este mecanismo atribui uma probabilidade para cada imagem da base de dados e tenta prever aquelas imagens que são de interesse para o usuário. [Duan et al., 2005] propuseram outra abordagem para realimentação de relevância baseada em inferência Bayesiana, no qual consideram a consistência entre sucessivas interações do usuário no processo de aprendizagem. Em [León et al., 2007], análise de regressão logística foi utilizada para tentar avaliar a probabilidade de relevância de uma imagem. Esta abordagem constrói modelos de regressão logística considerando conjuntos de características que são semanticamente relacionados uns com os outros. A probabilidade de relevância produzida pelo modelo é agrupada através do uso de operadores especiais.

[Arevalillo-Herráez et al., 2008] tenta aprender a percepção do usuário definindo um conjunto *fuzzy* de imagens no qual o usuário está interessado. A realimentação é empregada por uma função que atribui o grau de pertinência de cada imagem da base de dados neste conjunto.

Outra técnica de aprendizagem comumente utilizada em realimentação de relevância é Máquina de Vetores de Suporte (SVM). Basicamente, o objetivo dos métodos baseados em SVM é encontrar o hiperplano que separa as imagens relevantes das não-relevantes no espaço de características. [Tong & Chang, 2001] propuseram o uso de SVM ativo para separar imagens relevantes das imagens restantes.

[Liu et al., 2008] propuseram uma abordagem diferente para selecionar as imagens que serão apresentadas para o usuário. Os autores argumentam que existe uma certa redundância entre as imagens mais ambíguas e que isso pode afetar negativamente o processo de recuperação. Neste trabalho foi utilizado um método de aprendizado não-supervisionado para agrupar as imagens e então selecionam uma imagem de cada

grupo para ser rotulada pelo usuário no processo de realimentação.

Um método de realimentação de relevância baseado em Algorítimos Genéticos (AG) foi proposto por Stejic et al. [2003]. Nesta abordagem, as imagens da base de dados são particionadas uniformemente em grid. Um arcabouço baseado em Algoritmo Genético e realimentação de relevância é utilizado para determinar as características que melhor descrevem cada imagem.

Recentemente [Ferreira et al., 2011] propuseram dois arcabouços de CBIR com realimentação de relevância baseada em Programação Genética. O primeiro arcabouço considera apenas imagens rotuladas como relevantes e o segundo considera tanto as imagens rotuladas como relevantes, quanto as rotuladas como não-relevantes. Experimentos foram realizados em três coleções de imagens utilizando descritores de cor, forma e textura, para comparar a abordagem de PG com outros métodos de aprendizagem propostos na literatura. Os resultados apresentados mostram que o arcabouço de GP obteve ganhos significativos em termos de eficácia e eficiência sobre baselines do estado da arte.

Os trabalhos acima mencionados utilizam mecanismos de realimentação de relevância em CBIR para melhor identificar as necessidades de informação dos usuários e melhorar a percepção do sistema durante o processo de busca. No entanto, todos os trabalhos citados somente aplicam realimentação de relevância em bases de imagens contendo apenas uma modalidade de dado, o conteúdo visual das imagens. Neste trabalho de pesquisa iremos explorar o mecanismo de realimentação de relevância em bases de dados multimodais, utilizando tanto o conteúdo visual quanto a informação textual associada às imagens. À esta abordagem daremos o nome de Realimentação de Relevância Multimodal. Nesta pesquisa será abordado o problema onde o processo de busca se inicia por meio de uma imagem fornecida pelo usuário, sem qualquer informação do contexto semântico da busca.

Realimentação de relevância multimodal foi estudada em [Yang et al., 2007] para sistemas de recomendação de vídeos on-line e em [Aksoy & Cavus, 2005] para recuperação de vídeos de notícias. Em [Calumby, 2010] foi proposto um arcabouço de realimentação de relevância baseada em PG para recuperar imagens em bases de dados multimodais. A programação genética foi adicionada no processo de realimentação de relevância para aprender novas funções de combinação de evidências a partir das informações julgadas relevantes pelo usuário. Embora este trabalho esteja na mesma linha de pesquisa desta proposta, muitas questões importantes ainda não foram investigadas sobre o tema. Como exemplos de questões a serem estudadas, estão a avaliação de eficácia do processo de realimentação com usuários reais e avaliação e estudo de diversos parâmetros envolvidos no processo, tais como o número de interações

necessárias para encontrar informação relevante; estudos sobre o tamanho do conjunto de treinamento; sobre o número de imagens apresentadas a cada interação; inclusão de objetos não-relevantes no processo de realimentação; e a busca por técnicas para uma melhor definição do conjunto inicial de respostas mostradas durante o processo de realimentação. O estudo destas questões e a aplicação de realimentação de relevância no contexto de busca visual de produtos, bem como suas particularidades, servirão de suporte para a proposta de melhorias ou novas abordagens de realimentação de relevância que se adequem melhor ao ambiente em estudo.

Capítulo 3

Programação Genética para Fusão de Evidências

Neste capítulo será apresentado um trabalho de pesquisa preliminar no qual a Programação Genética (PG) foi utilizada como método para combinação de evidências. Este trabalho foi realizado como expansão de um trabalho anterior publicado em [Santos et al., 2009] e serviu como base para aprimorar os conhecimentos de Programação Genética, que vão permitir o uso desta abordagem em passos futuros desta pesquisa.

O foco deste trabalho preliminar foi a utilização de PG para combinar fontes de evidência textual com o objetivo de gerar melhores funções de similaridade para busca de imagens na Web.

3.1 Recuperação de Imagens Baseada em Programação Genética

O arcabouço de PG é basicamente um processo evolutivo separado em duas fases: treinamento e validação. Cada fase seleciona um conjunto de consultas e documentos da base de dados, chamados de conjunto de treinamento e conjunto de validação.

O arcabouço inicia com a criação de uma população inicial de indivíduos gerados aleatoriamente que evolui geração após geração através de operações genéticas de reprodução, cruzamento e mutação. O processo evolucionário continua até que um critério de parada seja alcançado. Na fase de treinamento, a cada vez que uma nova geração de indivíduos é criada, uma função de aptidão definida pelo usuário é aplicada a cada novo indivíduo para selecionar somente os mais aptos para as gerações futuras. Desde que cada indivíduo é modelado como uma função de similaridade, a aplicação

da função de aptidão corresponde a ordenar os documentos de acordo com as consultas do conjunto de treinamento. O valor de aptidão obtido é simplesmente uma avaliação da qualidade do *ranking* gerado.

Durante a fase de treinamento, 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. Na fase de validação, os melhores indivíduos na fase anterior são avaliados usando um segundo conjunto de dados. A ideia aqui é evitar uma super-especialização dos indivíduos baseado nas características do conjunto de treinamento (*overfitting*).

No final do processo evolutivo, somente os indivíduos que obtiveram melhor desempenho nas duas fases são selecionados como soluções finais. O método de seleção utilizado para escolher os melhores indivíduos é baseado na soma dos valores de aptidão dos indivíduos em ambas as fases menos o desvio padrão destes valores como sugerido em [de Almeida et al., 2007]. Nesse trabalho de PG nós adotamos a medida MAP *Mean Average Precision*) [Baeza-Yates & Ribeiro-Neto, 2011] como função de aptidão para avaliar a qualidade do *ranking* para um conjunto de consultas.

A Tabela 3.1 apresenta uma descrição dos terminais utilizados neste trabalho. Como funções para combinar terminais e sub-árvores de um indivíduo, foram utilizadas a soma (+), subtração (-), multiplicação (*), divisão (/), logaritmo natural (*log*), logaritmo na base-10 (*log10*), exponenciação (*exp*) e raiz quadrada (*sqrt*).

Terminal	**Descrição**
tf	Frequência crua do termo.
idf	Frequência Inversa do Documento dado por $log(1 + \frac{N}{df})$, N é o número de documentos na coleção e df é o número de documentos onde o termo aparece.
$tf * idf$	esquema de ponderação tf-idf.
$avgdl$	Tamanho médio do documento.
$bm25$	Fórmula Okapi BM25 definida na equação 3.4.
$norm$	Norma do documento.

Tabela 3.1. Terminais usados no arcabouço de PG.

3.1.1 Fontes de Evidência Textual

Neste trabalho várias evidências textuais extraídas automaticamente de páginas Web foram associadas às imagens contidas nestas páginas. O arcabouço de PG foi utilizado para descobrir novas formas de combinação destas evidências com o intuito de melhorar

a qualidade do conjunto de imagens recuperado para uma dada consulta textual. As evidências utilizadas foram:

i. Texto completo o texto completo da página Web sem a marcação HTML.

ii. Passagens de texto texto ao redor da imagem na página Web. Neste trabalho diversos tamanhos de passagens foram utilizados.

iii. Texto da tag âncora o conjunto de palavras encontradas entre as tags âncoras

iv. Nome do arquivo da imagem o nome do arquivo da imagem encontrado no atributo SRC da tag IMG.

v. Texto da tag ALT o texto extraído do atributo ALT da tag IMG.

vi. Título da página o título da página Web.

vii. Autor o texto extraído do atributo AUTHOR da tag META.

viii. Palavras-chave o texto extraído do atributo KEYWORDS da tag META.

ix. Descrição o texto extraído do atributo DESCRIPTION da tag META.

3.2 Experimentos

Nesta seção, nós descrevemos os detalhes sobre os experimentos realizados, descrevendo os parâmetros adotados no arcabouço de PG, os métodos utilizados como *baselines* e a coleção de dados utilizada.

3.2.1 Parametrização do Arcabouço de PG

Um sistema baseado em PG 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 PG configuram os parâmetros empiricamente baseados em poucos experimentos, ou utilizando valores padrões. Para facilitar a configuração do PG foi utilizada uma técnica de projeto experimental para avaliar o efeito de alguns parâmetros de PG e suas combinações para determinar seus efeitos quantitativos no resultado final.

[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 PG. Esta técnica pode ser utilizada para aumentar o desempenho de um sistema baseado em PG guiando a escolha de bons parâmetros de configuração. Finalmente, 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, o seu valor valor de configuração pode ser reduzido 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 de configuração do PG: tamanho da população, número de gerações e profundidade máxima dos indivíduos no sistema de PG. Os dois primeiros parâmetros foram selecionados porque foram aqueles que apresentaram maior impacto nos experimentos realizados por [Feldt & Nordin, 2000] e geralmente são os parâmetros configurados empiricamente em trabalhos usando PG. O último parâmetro foi adicionado ao projeto para investigar se o tamanho da árvore iria apresentar 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 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 a seguir. Os níveis dos parâmetros foram escolhidos para representar níveis qualitativamente distintos baseados em experiências prévias com o sistema de PG em uso.

A. pop_size: o número de indivíduos na população. No nível mínimo é 50 e no nível máximo é 300.

B. max_gen: o número máximo de gerações no arcabouço de PG. No nível mínimo é 5 e no nível máximo é 30.

C. max_depth: o tamanho máximo do indivíduo em uma população. No nível mínimo é 4 e no nível máximo é 12.

Neste trabalho, nós utilizamos a medida MAP como variável resposta. Como nós temos 3 fatores (A, B e C) e 2 níveis para cada fator ($a_1, a_2, b_1, b_2, c_1, ec_2$), nosso projeto fatorial resultou em (2^3) experimentos diferentes como mostrado na Tabela 3.2. Para

cada um dos oito experimentos, três replicações foram executadas para nos permitir avaliar o erro experimental, resultando em 24 execuções. Em nosso projeto fatorial, cada replicação é um experimento repetido com uma nova semente randômica no arcabouço de PG para gerar uma nova população inicial de indivíduos.

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].

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

Tabela 3.2. Configuração experimental para o projeto fatorial completo em dois níveis.

Os efeitos de cada fator e suas respectivas interações são mostrados a seguir em ordem decrescente de efeito:

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

Nós podemos 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 cerce da 120% maior que o fator C. Este resultado indica que o a escolha do tamanho de população é importante para obter bons resultados neste cenário. Erros experimentais ou não-observados foram responsáveis por cerca de 9% da variação na resposta. Os resultados destes experimentos serviram de guia na parametrização do arcabouço de PG.

Em nossos experimentos nós configuramos o tamanho da população para 300 indivíduos. A população inicial foi gerada randomicamente usando o método *ramped half-and-half* [Koza, 1992] e a profundidade inicial das árvores variando entre 2 a 6. Devido à estabilidade dos resultados obtidos com 30 gerações, nós escolhemos este valor como critério de parada.

Para as operações genéticas nós utilizamos as taxas de 90%, 5%, e 5% para cruzamento, reprodução e mutação, respectivamente. Ao final de cada geração, a fase de validação foi executada para os 20 melhores indivíduos retornados pela fase de treinamento naquela geração. A profundidade máxima para as árvores geradas foi configurado para 7, que é grande o suficiente para conter todas as características textuais utilizadas neste trabalho.

3.2.2 Baselines

Esta seção apresentas as duas estratégias de *ranking* utilizadas como baselines neste trabalho.

Modelo Bayesiano

Para avaliar o desempenho do nosso arcabouço de PG nós comparamos com o trabalho realizado em [Coelho et al., 2004] que apresentou uma estratégia de ranking baseada no modelo de rede de crenças Bayesiana.

Em [Coelho et al., 2004] foram utilizadas as mesmas evidências textuais descritas na seção 3.1.1. No entanto, em [Coelho et al., 2004], as fontes de evidência **iii**, **iv** e **v** são agrupadas para compor a evidência a qual os autores chamaram de tags de descrição. As fontes de evidência **vi**, **vii**, **viii** e **ix** são agrupadas para compor a evidência a qual os autores chamaram de tags de metadados.

Do modelo Bayesiano apresentado em [Coelho et al., 2004], sete estratégias de *ranking* foram derivadas conforme mostra a Tabela 3.3, onde cada fórmula define uma expressão $P(i_j|q)$ para ordenar uma imagem i_j em relação à uma consulta textual q e a fonte de evidência textual sendo considerada.

Abordagem de Ranking	$P(i_j\|q)$
Tags de desdrição	$\eta \times RD_{j,q}$
Tags de matadados	$\eta \times RM_{j,q}$
Passagem/Texto completo	$\eta \times RP_{j,q}$
Descrição + Metadados	$\eta \times [1 - (1 - RD_{j,q}) \times (1 - RM_{j,q})]$
Descrição + Passagem/Texto completo	$\eta \times [1 - (1 - RD_{j,q}) \times (1 - RP_{j,q})]$
Passagem/Texto completo + Metadados	$\eta \times [1 - (1 - RP_{j,q}) \times (1 - RM_{j,q})]$
Descrição + Metadados + Passagem/Texto completo	$\eta \times [1 - (1 - RD_{j,q}) \times (1 - RM_{j,q}) \times (1 - RP_{j,q})]$

Tabela 3.3. Abordagens de ranking para o modelo Bayesiano.

$RD_{j,q}$, $RM_{j,q}$, e $RP_{j,q}$ são probabilidades da evidência textual sendo observada dada uma consulta q em relação às tags de descrição, metadados e passagens de texto,

respectivamente. η é uma constante de normalização [Pearl, 1988] introduzida para que a soma de todas as probabilidades seja igual a 1.

De acordo com [Coelho et al., 2004], a probabilidade de cada evidência e_j sendo observada dado k pode ser estimada pela função de similaridade provida pelo modelo de espaço vetorial [McGill & Salton, 1983] cuja fórmula é definida a seguir:

$$P(e_j|k) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \tag{3.1}$$

onde k é o estado (1 se o termo $i \in q$, 0 caso contrário) para cada termo i. $w_{i,j}$ é o peso do termo i na respectiva fonte de evidência sendo observada e associada com a imagem I_j, $w_{i,q}$ é o peso do termo i na consulta. Os pesos foram definidos como em [Witten et al., 1999]:

$$w_{i,j} = (1 + ln(f_{i,j})) \times (ln(1 + \frac{N}{n_i})) \tag{3.2}$$

$$w_{i,q} = (1 + ln(f_{i,q})) \times (ln(1 + \frac{N}{n_i})) \tag{3.3}$$

$f_{i,j}$ é a frequência crua do termo i no documento contendo a imagem I_j, $f_{i,q}$ é a frequência crua do termo i na consulta q, N é o total de imagens na coleção, n_i é o número de evidências textuais sendo considerada que contém o termo i. Maiores detalhes sobre o arcabouço baseado no modelo Bayesiano podem ser obtidos em [Coelho et al., 2004].

Okapi BM25

Nós também comparamos o nosso arcabouço de PG com a estratégia de *ranking* Okapi BM25, como descrito em [Robertson et al., 1996]. Esta estratégia tem sido utilizado com sucesso em competições da TREC. Mais formalmente, dada uma consulta q contendo os termos $q_1, ..., q_n$, a similaridade de um documento D é definida pelo BM25 como:

$$BM25(q,d) = \sum_{i=1}^{n} \frac{(k_1 + 1) \times tf}{tf + k_1 \times \left((1 - b) + b \times \frac{|D|}{avgdl}\right)} \times \log \frac{N - df + 0.5}{df + 0.5} \tag{3.4}$$

onde tf é a frequência do termo no documento, N é o número total de documentos na coleção, df é o número de documentos na coleção que contém o termo da consulta, $\mid D \mid$

é o tamanho do documento (em palavras), *avgdl* é o tamanho médio do documento na coleção (em palavras). k_1 e b são parâmetros usados para ajustar o desempenho da busca. Nós utilizamos os mesmos valores como em [Robertson et al., 1996] para k_1 e b: $k1 = 2$ e $b = 0.75$.

3.2.3 Coleção

Com o objetivo de avaliar a abordagem de PG, nós realizamos vários experimentos utilizando uma coleção de páginas coletadas do diretório Yahoo[1]. Todas as páginas coletadas foram armazenadas com suas respectivas imagens para extrair as evidências textuais já descritas na seção 3.1.1.

A Tabela 3.4 apresenta algumas estatísticas 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 nenhum processamento ou redução do tamanho. Nós consideramos como imagens distintas aquelas que apresentaram URLs absolutas distintas. Portanto, imagens que apareceram em páginas distintas, foram consideradas distintas. Nossos experimentos foram conduzidos usando 50 consultas textuais extraídas de um log de consultas de uma máquina de busca de imagens [2].

Tamanho da coleção	$21GB$
Número de páginas HTML	89,568
Número de imagens distintas	195,794
Número de consultas	50
Número médio de imagens candidatas por consulta	62,08
Número médio de imagens relevantes por consulta	28,04

Tabela 3.4. Estatísticas da coleção usada nos experimentos.

Para cada consulta, nós rodamos os baselines e pegamos as 30 imagens do topo recuperados por cada estratégia de *ranking*. Estas imagens foram agrupadas em um único conjunto de imagens candidatas para cada consulta. Desta forma, não é possível dizer qual método recuperou qual imagem. Cada conjunto de imagens foi então analisado por um grupo de voluntários para avaliar as imagens como relevante ou irrelevante em relação à respectiva consulta. Ao final da avaliação, nós temos um conjunto de imagens para cada consulta rotuladas como relevantes ou irrelevantes independente de como foram recuperadas. Esta técnica de *pooling* é bastante utilizada em coleções

[1] www.yahoo.com
[2] http://busca.uol.com.br/imagem/index.html

da TREC [Voorhees & Harman, 1999]. Ela evita a necessidade de avaliar a coleção inteira e garante que o usuário avaliando as imagens não tem nenhum conhecimento sobre a estratégia usada para recuperá-la, provendo assim uma avaliação imparcial de relevância das imagens recuperadas.

3.2.4 Experimentos com o Modelo Bayesiano

Nesta seção nós apresentamos os resultados obtidos com o modelo de rede de crenças Bayesiano apresentado em [Coelho et al., 2004].

3.2.4.1 Texto Completo *versus* Passagens de Texto

Um primeiro experimento foi realizado para determinar o melhor tamanho para as passagens ao redor das imagens. Inicialmente, nós decidimos investigar o tamanho dos documentos, que é o texto completo sem as tags HTML, para escolher bons tamanhos de passagens a serem usadas em nossos 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. O tamanho dos documentos é expresso em número de termos.

Nós observamos que esta 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 3.5 apresenta algumas estatísticas sobre a distribuição de tamanho dos documentos.

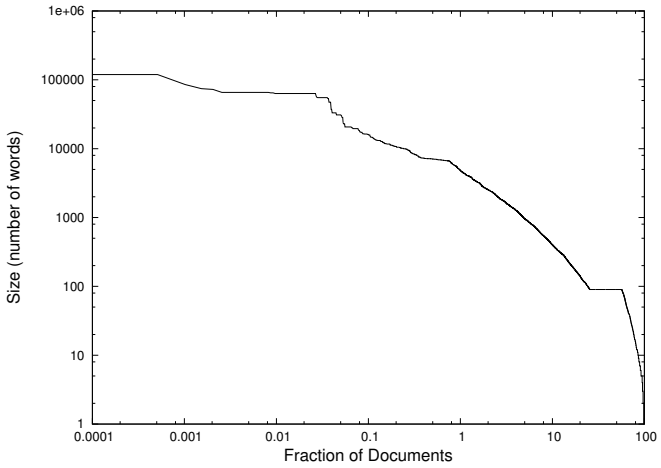

Figura 3.1. Distribuição de tamanho dos documentos.

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

Tamanho médio dos documentos	288
Tamanho da mediana dos documentos	90
Tamanho do maior documento	126.712
Tamanho do menor documento	1

Tabela 3.5. Estatísticas da distribuição de tamanho dos documentos.

da coleção. Baseada nestas observações, nós decidimos experimentar diversos tamanhos de passagens para determinar quais os tamanhos de passagens para utilizarmos nos experimentos. Os tamanhos de passagens escolhidos inicialmente foram 10, 20, 40, 60, 80 e 100 termos.

A Tabela 3.6 mostra os resultados de MAP para cada tamanho de passagem, considerando o conjunto de imagens relevantes. 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 ao 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.

Passagens de Texto						
Texto completo	10T	20T	40T	60T	80T	100T
24.859	21.536	19.981	26.791	28.341	27.945	25.428

Tabela 3.6. Medidas de MAP para as passagens de texto.

A Figura 3.2 mostra a curva de precisão-revocação para todas as passagens. Nós podemos confirmar que as passagens de 10 e 20 termos foram as que obtiveram os piores resultados. As curvas para as passagens de 60, 80 e 100 termos tiveram comportamento muito similares entre si, seguidas pela passagem de 40 termos. O texto completo somente supera as demais abordagens somente em níveis de revocação acima de 50%. Para avaliar se as passagens de texto que nós testamos são estatisticamente diferentes uma das outras, nós aplicamos o teste Wilcoxon nos resultados para guiar nossa escolha pela melhor abordagem. Embora a passagem de 60 termos tenha alcançado o melhor resultado 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. Devido ao bom compromisso entre desempenho de recuperação e o menor uso de recursos computacionais exigido pela passagem de 40 termos em comparação aos outros tamanhos estatisticamente equivalentes, nós consideramos

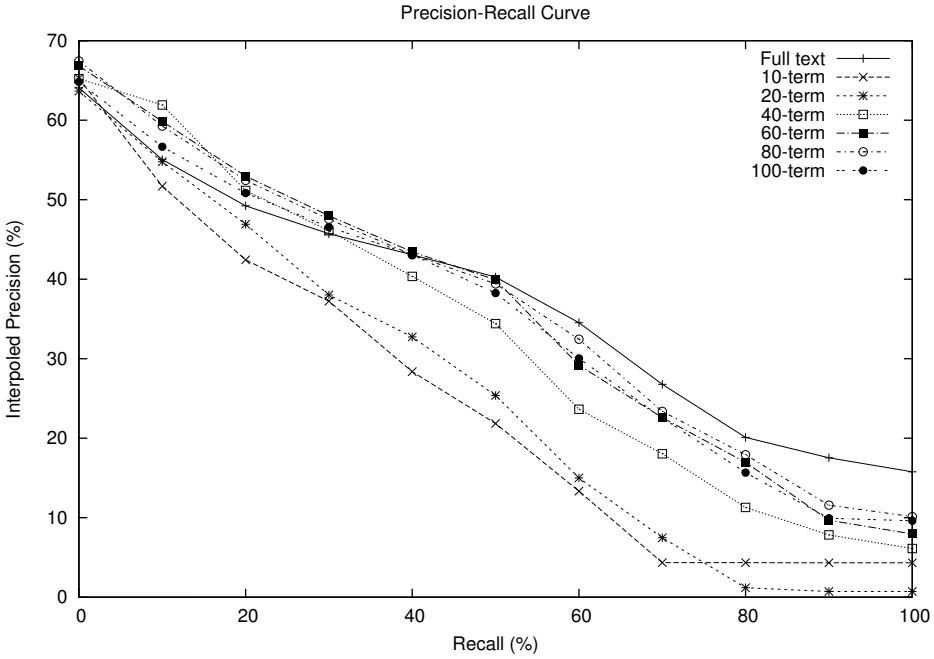

Figura 3.2. Curva de precisão-revocação para todas as passagens de texto.

a passagem de 40 termos na comparação com as outras fontes de evidência isoladas no arcabouço Bayesiano, pois ela representa o melhor custo-benefício na coleção em uso.

3.2.4.2 Evidências Isoladas no Modelo Bayesiano

Para investigar como passagens de 40 termos contribuem para a recuperação das imagens, nós comparamos esta evidência com as evidências de descrição e metadados para avaliar o desempenho de cada abordagem. A Figura 3.3 mostra a curva de precisão-revocação para as três fontes de evidência. Nós observamos que as passagens de texto são muito melhores que as evidências de metadados e de descrição para descrever as imagens na coleção em uso.

A Tabela 3.7 apresenta as medidas de MAP obtidas para as três evidências isoladas. Nós observamos que as passagens de texto tem maior contribuição na recuperação das imagens, seguido pela evidência de metadados. Nós aplicamos o teste Wilcoxon nos resultados e as passagens de texto foram estatisticamente melhores que as outras abordagens com nível de confiança acima de 95%.

3.2.4.3 Combinação de Evidências no Modelo Bayesiano

Nesta seção nos apresentamos os resultados obtidos quando combinamos múltiplas fontes de evidência no arcabouço Bayesiano. A Figura 3.4 apresenta os resultados para

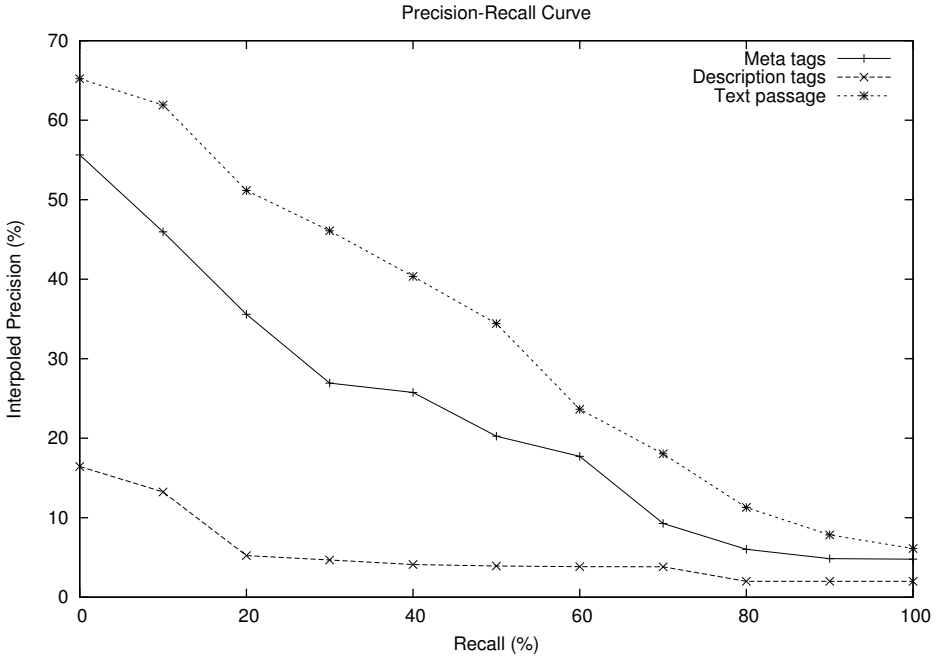

Figura 3.3. Curva de precisão-revocação para as fontes de evidência isoladas no arcabouço Bayesiano.

Fontes de Evidência		
Tags de metadados	Tags de Descrição	Passagens de Texto (40T)
18.172	9.1275	26.793

Tabela 3.7. Medida de MAP para as fontes de evidência isoladas no arcabouço Bayesiano.

as quatro combinações apresentadas em [Coelho et al., 2004]: *descrição+passagem, descrição+metadados, metadados+passagem, e descrição+metadados+passagem.*

Nós podemos observar que as combinações de *descrição+metadados* e *descrição+passagem* obtiveram os piores resultados devido ao baixo desempenho da abordagem de tags de descrição. As abordagens de *metadados+passagem* e *metadados+descrição+passagem* obtiveram comportamentos similares embora a abordagem de *metadados+passagem* apresentou valores de precisão mais altos até 60% de revocação.

A Tabela 3.8 apresenta os resultados de MAP obtidos para as quatro combinações testadas.

Devido aos bons resultados obtidos pela abordagem *metadados+passagem*, nós escolhemos esta abordagem para ser usada na comparações com o arcabouço de PG.

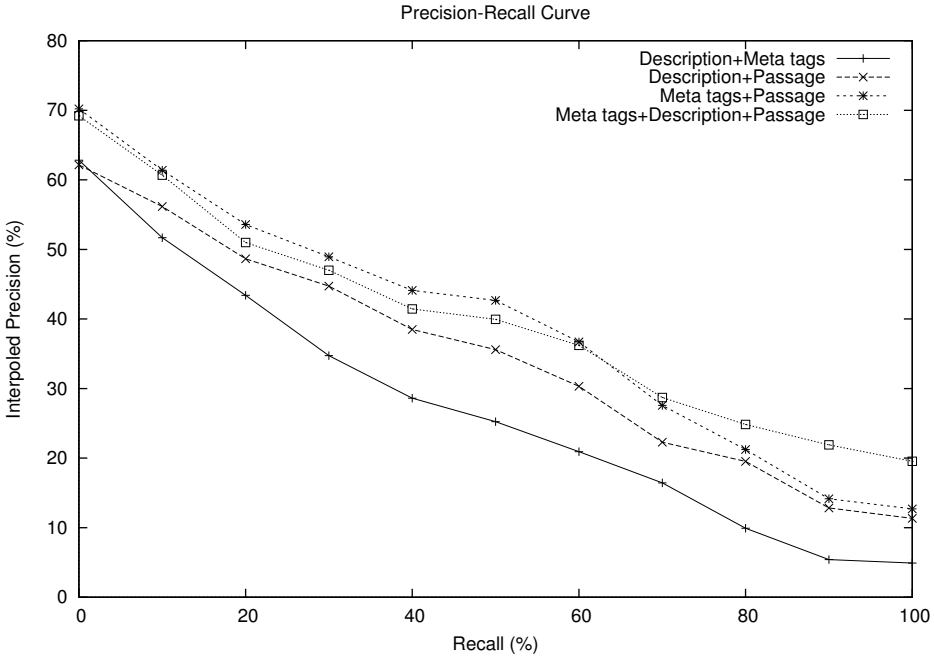

Figura 3.4. Curva de precisão para a combinação de múltiplas fontes de evidência no arcabouço Bayesiano.

Múltiplas Fontes of Evidência			
Descrição+Metadados	Descrição+Passagem	Metadados+Passagem	Descrição+Metadados+Passagem
19.367	24.687	29.275	27.398

Tabela 3.8. Medidas de MAP para as combinações de múltiplas fontes de evidência no arcabouço Bayesiano.

3.2.5 Resultados Experimentais com o Arcabouço de PG

Nesta seção nós apresentamos os resultados dos nossos experimentos com o arcabouço de PG. Os experimentos foram realizados utilizando a estratégia de validação-cruzada de 5 *folds*. A cada *fold* uma nova semente era fornecida para gerar populações inicias distintas dos demais *folds*. A Figura 3.5 apresenta as curvas de precisão-revocação obtida pelo nosso arcabouço de PG, pela melhor abordagem obtida no modelo Bayesiano ($metadados + passagem$) e pelo BM25 sobre o texto completo. Nós observamos que a abordagem de PG obteve melhores valores de precisão em todos os níveis de revocação.

A Tabela 3.9 descreve as medidas de P@10, P@20, P@30 e MAP obtidas para o arcabouço de PG e os baselines. A arcabouço de PG obteve um ganho de 22.36% sobre o BM25. De acordo com os resultados do teste estatístico, o sistema de PG foi estatisticamente melhor que o arcabouço Bayesiano com 99% de confiança e melhor que o BM25 com 98% de confiança.

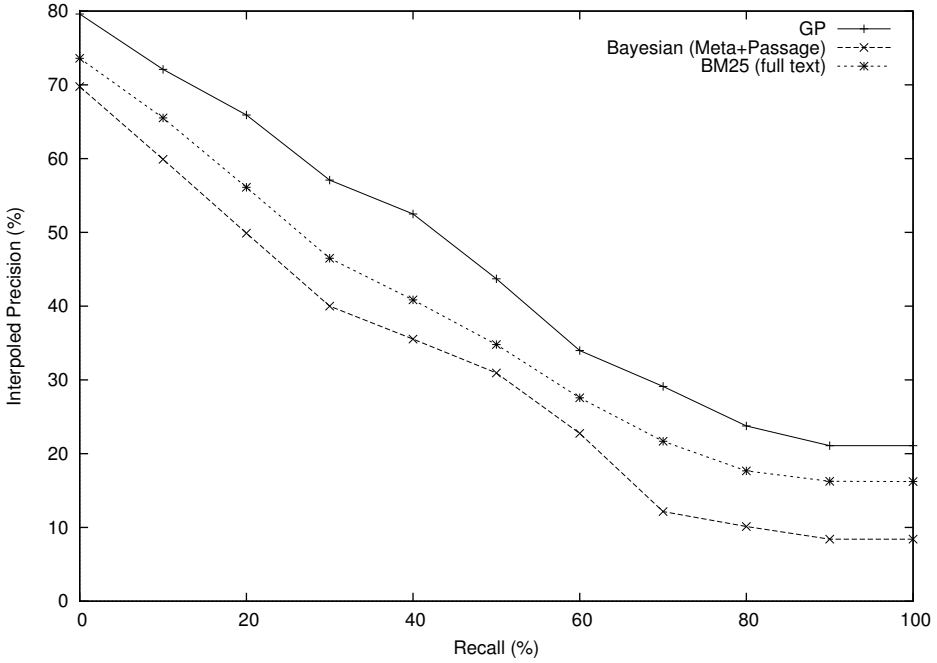

Figura 3.5. Curva de precisão-revocação para a comparação do PG e baselines.

	PG	BM25 (texto completo)	Bayesiano (*metadados* + *passagem*)
MAP	42.57	34.79	29.28
P@10	48.0	45.0	14.8
P@20	40.5	38.3	23.6
P@30	37.0	13.9	0.9

Tabela 3.9. Medidas de MAP e P@N para os arcabouços de PG, Bayesiano e BM25.

3.3 Conclusões

Neste trabalho de pesquisa preliminar nós utilizamos uma abordagem baseada em PG para combinar diversas fontes de evidência textual para compor estratégias de *ranking* mais eficazes. Entre estas fontes de evidência, aquelas que mais apareceram nas fórmulas geradas pelo arcabouço de PG foram o título da página, texto completo, tamanho médio do documento (avgdl) e passagem de 10 termos.

De acordo com os nosso experimentos, a abordagem de PG adotada obteve ganhos significativos em relação aos baselines testados. A abordagem de PG apresentou ganhos de 22.36% sobre o BM25 e ganhos de 43% em cima do arcabouço Bayesiano proposto por [Coelho et al., 2004].

Outra contribuição deste trabalho preliminar foi o uso do projeto fatorial como ferramenta para parametrizar o arcabouço de PG.

Capítulo 4

Próximos Passos

Esta proposta de tese terá como aplicação alvo a implementação de um sistema para busca visual de produtos em uma base de dados multimodal, onde cada objeto contém informação visual (imagens) e informação textual (descrição). Neste sistema, o usuário fornece inicialmente uma imagem para servir de consulta no processo de recuperação das imagens. Um conjunto inicial de imagens é retornado e apresentado para o usuário que interage com o sistema fornecendo um julgamento de relevância sobre o conjunto recuperado. O sistema então modifica a consulta inicial baseado no julgamento fornecido para recuperar um novo conjunto de respostas mais adequado à necessidade de informação do usuário.

Neste contexto, identificamos dois tipos de busca em um sistema de busca visual de produtos. No primeiro tipo, que chamaremos de busca por conceito, o usuário está interessado em obter informação de produtos similares ao fornecido na consulta. Por exemplo, o usuário pode fornecer uma imagem de um vestido e está interessado em vestidos similares. Neste caso, o usuário tem como alvo uma classe de produtos. No segundo tipo, que chamaremos de busca por objeto alvo, o usuário está interessado em obter informação específica sobre um produto. É o caso por exemplo, da imagem fornecida ser a imagem de um celular e o usuário deseja recuperar informação como modelo, preço e outras características do aparelho.

Em ambos os casos, a utilização de informação multimodal pode ser fundamental para recuperar produtos relevantes. Por exemplo, em uma busca por uma camisa polo com determinada estampa, o usuário pode ter interesse em outras peças de roupa com estampa parecida. No entanto, questões que dificilmente são capturadas na imagem podem ser fundamentais para encontrar outros tipos de resposta relevante. Por exemplo, pode ser interessante determinar se a consulta trata-se de uma blusa masculina ou feminina, a própria informação da blusa ser polo, e assim por diante. Com

a realimentação de relevantes, o usuário pode em poucas interações fornecer de forma implícita este e outros tipos de informação. Já no exemplo da consulta por um celular, a realimentação de relevantes pode guiar o sistema para a categoria correta, evitando que o sistema recupere objetos de categorias completamente distintas da desejada pelo usuário.

Estes exemplos servem para ilustrar a importância do processo de realimentação de relevantes multimodal no processamento de consultas visuais. Neste trabalho, vamos estudar os diversos problemas relacionados com a aplicação de busca visual em produtos e o potencial uso de técnicas de realimentação de relevantes no desenvolvimento de soluções para esse ambiente.

Como passo fundamental para o andamento do trabalho nos próximos meses, será implementado um sistema de busca por imagens com mecanismo de realimentação de relevância multimodal. A ideia é utilizar a biblioteca OPENCV [Bradski & Kaehler, 2008] para suporte nas implementações de processamento de imagens. As evidências visuais devem ser combinadas utilizando-se PG como abordagem para fusão de evidências. Pretendemos utilizar programação genética sempre que possível na tarefa de combinar evidências, dado que nosso trabalho preliminar e outros trabalhos apontados na literatura mostram que PG representa uma ótima opção para a combinação de evidências em sistemas de busca. Para permitir o uso de PG, é necessário a criação de uma coleção de referência para busca por imagens de produtos. A criação desta coleção de referência será uma das tarefas a ser realizada nesta tese.

O sistema básico de busca por imagens a ser implementado deve utilizar informações visuais como cor, forma e textura. O uso de uma abordagem baseada em pontos de interesse da imagem como o descritor SURF [Bay et al., 2006] ou SIFT [Lowe, 1999] também será investigada. Com a implementação deste sistema, pretende-se experimentar alternativas de realimentação de relevantes para que tenhamos um melhor entendimento do problema e possamos propor novas soluções, além de avaliar as alternativas existentes que tenham sido utilizadas em outros contextos. Experimentos com usuários reais deverão ser realizados para avaliar a eficácia e eficiência das abordagens propostas.

Dentre as diversas questões importantes a serem estudadas com esse trabalho de pesquisa, pode-se citar:

- Verificar se o tipo de busca pode influenciar na escolha do conjunto inicial e sua apresentação para o usuário. Se a busca for por conceito, uma alternativa é apresentar produtos visualmente similares mas que pertencem à diferentes categorias. Caso a busca seja por objeto específico, é provável que objetos visualmente si-

milares à consulta e que pertencem à uma mesma categoria sejam mais eficazes para o processo de realimentação.

- Verificar se é possível derivar algoritmos de ordenação por relevância que acelerem a convergência para informações relevantes durante o processo e realimentação. Uma ideia que pretendemos experimentar pelo fato de usarmos uma base de dados multimodal é procurar selecionar respostas com o objetivo de maximizar a chance de convergência. Por exemplo, adotar abordagens que tragam diversidade no conjunto inicial (imagens de categorias diferentes) para ajudar o usuário a descrever melhor o que está procurando.

- Uma ideia que pretendemos explorar é a utilização de estratégias diferentes para a recuperação do conjunto inicial de respostas e para a recuperação do conjunto de respostas a partir da realimentação. No primeiro estágio, técnicas mais eficientes de análise global das imagens podem ser utilizadas para restringir o espaço de busca e técnicas mais sofisticadas como a análise de pontos de interesse podem ser utilizadas nas fases de realimentação.

- Em geral, sistemas com realimentação de relevância são baseados em múltiplos julgamentos de relevância no conjunto de respostas a cada interação. Neste trabalho pretendemos estudar o impacto do tamanho do conjunto de respostas e como selecionar o melhor conjunto de imagens a cada interação com o usuário para que o número total de interações necessários para alcançar o objetivo seja mínimo. Outra questão a ser estudada é saber se é possível obter uma convergência rápida adotando uma abordagem minimalista com poucas imagens no conjunto resposta para que o usuário indique no máximo uma imagem a cada interação.

Base de Dados

Inicialmente, a base de dados a ser utilizada nesta proposta de tese é uma base de produtos que contém informação multimodal: imagens dos produtos e descrições textuais. Em relação às descrições textuais cada objeto da base contém o nome do produto, uma descrição detalhada do produto, o nome da categoria principal a qual o produto pertence, o grupo mais específico dentro da categoria na qual o produto pertence, a marca, o preço, autor (quando houver), isbn (quando houver), editora (quando houver). Um exemplo de um objeto da base de dados de produto é apresentado a seguir:


```
<item>
      <g:id>164064</g:id>
      <title><![CDATA[Meninos do Mangue]]></title>
      <description><![CDATA[Quando foram pescar siri no mangue, a Sorte e a Preguiça fizeram uma aposta:
      ganharia quem pescasse o siri com mais patas. A Sorte venceu, claro, e a Preguiça teve que contar oito histórias, uma pa... - <b>
      <g:em></g:em>
      <g:brand><![CDATA[]]></g:brand>
      <g:price>R$ 26,90</g:price>
      <g:condition>new</g:condition>
      <link><![CDATA[http://www.submarino.com.br/produto/1/164064]]></link>
      <g:image_link><![CDATA[http://i.s8.com.br/images/books/cover/img4/164064.jpg]]></g:image_link>
      <g:product_type><![CDATA[Livros > Literatura Infanto-Juvenil > Infantil - 4 a 8 anos]]></g:product_type>
      <g:category><![CDATA[Livros]]></g:category>
      <g:category_id>1</g:category_id>
      <g:group><![CDATA[Literatura Infanto-Juvenil]]></g:group>
      <g:autart><![CDATA[ROGER MELLO]]></g:autart>
      <g:isbn>8574061034</g:isbn>
      <g:publisher><![CDATA[Companhia das Letras]]></g:publisher>
</item>
```

Figura 4.1. Amostra de um objeto da base de dados de produto.

Nesta base os produtos estão separados em categorias mais específicas de produtos como por exemplo, estojos escolares ou réguas e compassos. Cada objeto pertence ainda a uma categoria (categoria-pai) que oferece uma definição mais genérica do contexto semântico do objeto, como por exemplo, papelaria para os exemplos já citados. A coleção que iremos trabalhar inicialmente é composta de 42.851 imagens de produtos e suas respectivas informações textuais distribuídas em 1.083 categorias específicas. O tamanho das categorias varia entre 1 e 772 itens de produto.

Uma etapa de pré-processamento das imagens da base e da informação textual será realizada para extração das informações em cada modalidade de dados.

Pré-processamento das Imagens

O pré-processamento das imagens da base consiste de etapas de segmentação, transformação do espaço de cor e quantização uniforme para redução do tamanho dos vetores

de características. A etapa de segmentação é realizada para eliminar informação de fundo (*background*) da imagem gerando uma mascara binária para os pixels do objeto em destaque (*foreground*). Esta máscara binária resultante da etapa de segmentação indica quais os pixels da imagem devem ser considerados nas fases posteriores e quais pixels devem ser ignorados.

Neste trabalho utilizaremos um algoritmo de segmentação conhecido como grabcut Rother et al. [2004]. Um exemplo do resultado da etapa de segmentação é apresentado a seguir:

Figura 4.2. Imagem original à esquerda e imagem segmentada à direita.

Após a etapa de segmentação, a mascara resultante é utilizada nas etapas de transformação do espaço de cor RGB para o HSV e quantização uniforme. Para a etapa de quantização, utilizamos um esquema de 8x3x3 (8 níveis para o canal H, 3 níveis para o canal S, e 3 para o canal V) que resulta em 72 níveis de intensidade de cor. Outros esquemas de quantização podem ser experimentados na coleção utilizada. Um exemplo do resultado da etapa de quantização é apresentado a seguir:

Figura 4.3. Imagem original à esquerda e imagem quantizada à direita.

Pré-processamento do Texto

Nesta proposta, o texto associado às imagens será representado utilizando o modelo de espaço vetorial [McGill & Salton, 1983], tradicionalmente utilizado em recuperação de informação textual. Uma etapa de pré-processamento das informações textuais será realizada com operações de remoção de *stop-words* e *stemming* para a obtenção

do vocabulário da coleção. A biblioteca Lucene [1] será utilizada para indexação da informação textual.

4.0.1 Planejamento das Atividades

Um planejamento de atividades para os próximos meses é apresentado a seguir:

Atividades planejadas	2011	2012								2013		
	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago a Dez	Jan	Fev	Mar
Criação de coleção de referência												
Indexação da base de produtos												
Inclusão de prop. visuais no PG												
Experimentos com fusão multimodal												
Implementação de técnicas de RF												
Desenvolvimento de novos Métodos de RF e experimentos												
Escrita de artigo												
Escrita da Tese												

Figura 4.4. Planejamento das atividades.

[1] http://lucene.apache.org/

Referências Bibliográficas

Aksoy, S. & Cavus, O. (2005). A relevance feedback technique for multimodal retrieval of news videos. *EUROCON 2005 The International Conference on Computer as a Tool*, pp. 139--142.

Arampatzis, A.; Zagoris, K. & Chatzichristofis, S. A. (2011a). Dynamic two-stage image retrieval from large multimodal databases. In *ECIR*, pp. 326–337.

Arampatzis, A.; Zagoris, K. & Chatzichristofis, S. A. (2011b). Fusion vs. two-stage for multimodal retrieval. In *Proceedings of the 33rd European conference on Advances in information retrieval*, pp. 759--762.

Arevalillo-Herráez, M.; Zacarés, M.; Benavent, X. & de Ves, E. (2008). A relevance feedback cbir algorithm based on fuzzy sets. *Sig. Proc.: Image Comm.*, 23(7):490–504.

Baeza-Yates, R. A. & Ribeiro-Neto, B. (2011). *Modern Information Retrieval - the concepts and technology behind search.* Addison-Wesley, New York, second edition edição.

Bay, H.; Tuytelaars, T. & Gool, L. J. V. (2006). Surf: Speeded up robust features. In Leonardis, A.; Bischof, H. & Pinz, A., editores, *ECCV (1)*, volume 3951 of *Lecture Notes in Computer Science*, pp. 404–417. Springer.

Box, G. E. P.; Hunter, W. G. & Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* John Wiley & Sons, New York, USA, 1 edição.

Bradski, G. & Kaehler, A. (2008). *Learning OpenCV: Computer Vision with the OpenCV Library.* O'Reilly, Cambridge, MA.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Journal of Computer Networks and ISDN Systems*, 30:107--117.

Calumby, R. T. (2010). Recuperaćão multimodal de imagens com realimentaćão de relevância baseada em programaćão genética. Master's thesis, Instituto de Computaćão da Universidade Federal de Campinas.

Chang, N. & Fu, K. (1980). Query-by-pictorial-example. *tse*, SE-6:519--524.

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

Chen, Z.; Wenyin, L.; Zhang, F. & Li, M. (2001). Web mining for web image retrieval. *JASIST*, 52:831--839.

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

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

Clinchant, S.; Ah-Pine, J. & Csurka, G. (2011). Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, ICMR '11, pp. 44:1--44:8.

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

Cox, I. J.; Miller, M. L.; Minka, T. P.; Papathomas, T. V. & Yianilos, P. N. (2000). The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37.

da S. Torres, R.; ao, A. X. F.; 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.*, 42:283--292.

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.; GonÃ§alves, M. A.; Cristo, M. & Calado, P. (2007). A combined component approach for finding collection-adapted ranking functions based on genetic programming. In *SIGIR*, pp. 399–406.

Depeursinge, A. & Müller, H. (2010). Fusion techniques for combining textual and visual information retrieval. In *ImageCLEF*, volume 32 of *The Information Retrieval Series*, pp. 95--114. Springer Berlin Heidelberg.

Deselaers, T.; Keysers, D. & Ney, H. (2005). Fire – flexible image retrieval engine: Imageclef 2004 evaluation. In *Multilingual Information Access for Text, Speech and Images – Fifth Workshop of the Cross-Language Evaluation Forum, CLEF 2004*, pp. 688–698.

Deselaers, T.; Weyand, T. & Ney, H. (2007). Image retrieval and annotation using maximum entropy. In Peters, C.; Clough, P.; Gey, F.; Karlgren, J.; Magnini, B.; Oard, D.; de Rijke, M. & Stempfhuber, M., editores, *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, pp. 725–734.

dos Santos, J. A.; Ferreira, C. D. & da Silva Torres, R. (2008). A genetic programming approach for relevance feedback in region-based image retrieval systems. In *SIBGRAPI*, pp. 155–162. IEEE Computer Society.

Doulamis, N. D. & Doulamis, A. D. (2006). Evaluation of relevance feedback schemes in content-based in retrieval systems. *Signal Processing: Image Communication*, 21(4):334–357.

Duan, L.; Gao, W.; Zeng, W. & Zhao, D. (2005). Adaptive relevance feedback based on bayesian inference for image retrieval. *Signal Processing*, 85(2):395–399.

Fan, W.; Gordon, M. D. & Pathak, P. (2000). Personalization of search engine services for effective retrieval and knowledge management. In *ICIS*, pp. 20–34.

Fan, W.; Gordon, M. D. & Pathak, P. (2004). Discovery of context-specific ranking functions for effective iformation retrieval using genetic programming. *IEEE Trans. Knowl. Data Eng.*, 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.

Fan, W.; 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.

Feldt, R. & Nordin, P. (2000). Using factorial experiments to evaluate the effect of genetic programming parameters. In *Genetic Programming, Proceedings of EuroGP'2000*, volume 1802, 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.; Santos, J.; da S. Torres, R.; Gonçalves, M.; Rezende, R. & Fan, W. (2011). Relevance feedback based on genetic programming for image retrieval. *Pattern Recognition Letters*, 32(1):27 – 37.

Ferreira, C. D.; da Silva Torres, R.; GonÃ§alves, M. A. & Fan, W. (2008). Image retrieval with relevance feedback based on genetic programming. In de Amo, S., editor, *SBBD*, pp. 120–134. SBC.

Gondra, I. & Heisterkamp, D. R. (2004). Adaptive and efficient image retrieval with one-class support vector machines for inter-query learning. *WSEAS Transactions on Circuits and Systems*, 3.

Gonzalez, R. C. & Woods, R. E. (2008). *Digital image processing*. Prentice-Hall, Upper Saddle River, NJ, third edição.

Hong, P.; Tian, Q. & Huang, T. S. (2000). Incorporate support vector machines to content-based image retrieval with relevant feedback. In *ICIP*.

Iyengar, G.; Duygulu, P.; Feng, S.; Ircing, P.; Khudanpur, S.; 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 Multimedia*, pp. 21–30.

Jing, Y. & Baluja, S. (2008). Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analisys and Machine Intelligence*, 30:1877--1890.

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

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

Kittler, J.; Hatef, M. & Duin, R. P. W. (1996). Combining classifiers. In *Proceedings of the Sixth International Conference on Pattern Recognition*, pp. 897--901. IEEE Computer Society Press.

Ko, B. & Byun, H. (2002). Probabilistic neural networks supporting multi-class relevance feedback in region-based image retrieval. In *ICPR (4)*, pp. 138–141.

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

León, T.; Zuccarello, P.; Ayala, G.; de Ves, E. & Domingo, J. (2007). Applying logistic regression to relevance feedback in image retrieval systems. *Pattern Recognition*, 40(10):2621–2632.

Li, B. & Yuan, S. (2004). A novel relevance feedback method in content-based image retrieval. In *ITCC (2)*, pp. 120–123. IEEE Computer Society.

Li, P. & Ma, J. (2009). Learning to rank for web image retrieval based on genetic programming. In *IEEE IC-BNMT*, pp. 137--142.

Liu, H.; Song, D.; Rüger, S.; Hu, R. & Uren, V. (2008). Comparing dissimilarity measures for content-based image retrieval. In *Proceedings of the 4th Asia information retrieval conference on Information retrieval technology*, pp. 44--50. Springer-Verlag.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pp. 1150 –1157 vol.2.

Lu, Y.; Hu, C.; Zhu, X.; Zhang, H. & Yang, Q. (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *ACM Multimedia*, pp. 31–37.

McDonald, K. & Smeaton, A. F. (2005). A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *CIVR*, 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 *CLEF*, pp. 724–732.

Oren, N. (2002). Reexamining tf.idf based information retrieval with genetic programming. *SAICSIT '02: Proceedings of the 2002 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology*, pp. 224--234.

Pass, G.; Zabih, R. & Miller, J. (1996). Comparing images using color coherence vectors. In *ACM Multimedia*, pp. 65–73.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

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

Robertson, S.; Walker, S.; Beaulieu, M.; Gatford, M. & Payne, A. (1996). Okapi at trec-4. In *In Proceedings of the 4th Text REtrieval Conference (TREC-4*, pp. 73--96.

Rother, C.; Kolmogorov, V. & Blake, A. (2004). "grabcut": interactive foreground extraction using iterated graph cuts. In *ACM SIGGRAPH 2004 Papers*, SIGGRAPH '04, pp. 309--314.

Rui, Y.; Huang, T.; Ortega, M. & Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):644–655.

Rui, Y. & Huang, T. S. (2000). Optimizing learning in image retrieval. In *CVPR*, pp. 1236–. IEEE Computer Society.

Santos, K. C. L.; Almeida, H. M.; Goncalves, M. A. & da S. Torres, R. (2009). RecuperaÃ§Ã£o de imagens textuais da web utilizando mÃºltiplas evidÃªncias textuais e programaÃ§Ã£o genÃ©tica. In *SBBD*, pp. 91--105.

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*, 22:1349--1380.

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

Stejic, Z.; Takama, Y. & Hirota, K. (2003). Genetic algorithm-based relevance feedback for image retrieval using local similarity patterns. *Inf. Process. Manage.*, 39(1):1–23.

Stricker, M. & Orengo, M. (1995). Similarity of color images. In *Proc. SPIE Storage and Retrieval for Image and Video Databases*.

Sural, S.; Qian, G. & Pramanik, S. (2002). Segmentation and histogram generation using the HSV color space for image retrieval. In *Proceedigs of International Conference on Image Processing*, pp. 589–592.

Swain, M. J. & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7:11--32.

Tong, S. & Chang, E. (2001). Support vector machine active learning for image retrieval. In *Proc. of ACM Int. Conf. on Multimedia*, pp. 107–118.

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

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

Vani, V. & Raju, S. (2010). A detailed survey on query by image content techniques. In *Proceedings of the 12th international conference on Networking, VLSI and signal processing*, pp. 204--209.

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 *CLEF*, 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 *CLEF*, pp. 500–503.

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

Wang, J. Z. (2001). *Integrated region-based image retrieval*. The Kluwer international Series on information retrieval. Kluwer Academic Publishers.

Witten, I. H.; Moffat, A. & Bell, T. C. (1999). *Managing Gigabytes : Compressing and Indexing Documents and Images*. Morgan Kaufmann, San Francisco, CA, 2. edição.

Yang, B.; Mei, T.; Hua, X.-S.; Yang, L.; Yang, S.-Q. & Li, M. (2007). Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pp. 73--80.

Yu, H.; Li, M.; Zhang, H.-J. & Feng, J. (2003). Color texture moments for content-based image retrieval. In *Proceedings of the International Conference on Image Processing*, pp. 24--28.

Zhang, R. & Guan, L. (2009). Multimodal image retrieval via bayesian information fusion. In *IEEE ICME*, pp. 830–833.

Zhou, X.; Gobeill, J. & Müller, H. (2009). The medgift group at imageclef 2008. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pp. 712--718, Berlin, Heidelberg. Springer-Verlag.