UMA ABORDAGEM EVOLUTIVA PARA RECUPERAÇÃO DE IMAGENS DA WEB

KATIA CRISTINA LAGE DOS SANTOS

UMA ABORDAGEM EVOLUTIVA PARA RECUPERAÇÃO DE IMAGENS DA WEB

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Marcos André Gonçalves
Co-orientador: Ricardo da Silva Torres

Belo Horizonte
29 de maio de 2009

Santos, Katia Cristina Lage dos
S237a Uma Abordagem Evolutiva para Recuperação de Imagens da Web / Katia Cristina Lage dos Santos. — Belo Horizonte, 2009
xx, 58 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Marcos André Gonçalves
Co-orientador: Ricardo da Silva Torres

1. Processamento de imagens - Tese. 2. Recuperação da informação - Tese. 3. Programação genética (computação) - Tese. I. Título.

CDU 519.6*84

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma Abordagem Evolutiva para Recuperação de Imagens da Web

KATIA CRISTINA LAGE DOS SANTOS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARCOS ANDRÉ GONÇALVES - Orientador
Departamento de Ciência da Computação - UFMG

PROF. RICARDO DA SILVA TORRES - Co-orientador
Instituto de Computação - UNICAMP

PROF. JOÃO MARCOS BASTOS CAVALCANTI
Universidade Federal do Amazonas - UFAM

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 29 de maio de 2009.

Dedico este trabalho à minha família.

Agradecimentos

Agradeço a Deus pela família que me deu e pela força, amor e misericórdia em todos os momentos de minha vida.

Aos meus pais, Maria das Graças e Francisco Xavier, pelo apoio, dedicação e amor incondicional. Aos meus avós, Rosa e Diordede, pelo exemplo e carinho de sempre. Certamente essa conquista é de todos vocês que sempre acreditaram em mim e se abdicaram de muitas coisas para me proporcionar uma vida feliz e abençoada.

Ao meu namorado Daniel e sua família por todo amor, carinho e torcida. Certamente são parte essencial em minha vida, muito obrigada.

Ao meu orientador Professor Marcos André e ao co-orientador Ricardo Torres, por todo empenho em me orientar na realização de um trabalho de qualidade e por poder contar com a imensa compreensão e paciência de vocês. Também agradeço aos demais Professores do DCC e ao pessoal da Secretaria que foram muito compreensíveis e solidários. Em especial, agradeço aos Professores Alberto Laender, Geraldo Robson e Mario Campos e às Secretarias Renata, Sheila e Tulia.

Aos meus amigos que sempre me incentivaram. Assim o meu muito obrigada à Fabíola, Brenda, Sônia, Terezinha, aos amigos do Serpro, Lapo, LBD e do mestrado.

Resumo

Os avanços no armazenamento de dados e nas tecnologias para aquisição de imagens tornaram possível a criação de grandes bases ou coleções de imagens. Além disso, o enorme sucesso da Web tem proporcionado baixo custo e acessibilidade em larga escala deste material. Aliado a esses fatores, uma variedade de atividades lucrativas demandam transferência de informação baseada em imagem. Exemplos típicos são projetos arquitetônicos, desenhos de engenharia e de moda, perfumes, novos carros, campanhas de marketing, páginas da Internet, etc. Todas essas necessidades têm levado a um crescimento do interesse em organizar, indexar e consultar imagens digitais na última década.

O problema conhecido na literatura relacionada como o Problema de Recuperação de Imagem pode ser enunciado da seguinte maneira: a partir de uma coleção formada por imagens, é realizada uma consulta por um usuário, sendo esta formada por evidências textuais e/ou de conteúdo relacionados a uma imagem, defina o conjunto-resposta mais relevante para aquela consulta. Esta tarefa é desafiadora, dada a dificuldade em se definir quais as evidências são as mais relevantes para a indexação de informações sobre as imagens da coleção. Uma dificuldade imediata consiste em ordenar o conjunto-resposta com relação a uma consulta do usuário.

A partir da contextualização anterior, o presente trabalho apresenta um arcabouço evolucionário para recuperação de imagens. Este opera a partir da combinação de múltiplas evidências textuais e utiliza os preceitos da Programação Genética (PG), técnica da Inteligência Artificial. A PG é baseada na idéia de melhoramento contínuo da qualidade de uma solução para o problema em análise, a partir de soluções anteriores. Nossa motivação, portanto, é contribuir com o desenvolvimento de mecanismos de busca capazes de recuperar de uma maneira mais eficaz as imagens de uma coleção, no caso da World Wide Web. Como resultado prático, uma consulta realizada pelo usuário desse sistema de recuperação de imagens terá como resultado imagens mais relevantes sob a ótica do próprio usuário, aumentando assim o grau de satisfação e confiança do usuário no sistema de busca. Dentre os desafios advindos com a utiliza-

ção dessa coleção podem ser mecionados: a heterogeneidade dos tipos de imagens, a inexistência de padronização no preenchimento das *tags* HTML onde as imagens são inseridas, a inclusão de referências incorretas para outro documento HTML, dentre outros.

Experimentos realizados com uma coleção extraída da Web mostraram que, comparado com o Modelo Bayesiano apresentado em [Coelho et al., 2004], o arcabouço evolucionário apresentou um desempenho duas vezes maior, considerando as medidas de precisão, revocação e MAP. Estas são métricas importantes para avaliação da qualidade das imagens retornadas a partir de uma consulta. Além disso, a solução computacional desenvolvida neste trabalho apresenta uma grande flexibilidade, uma vez que em trabalhos futuros poderão ser adicionadas novas evidências com um custo mínimo de manutenção, além da possibilidade de ser empregada para a recuperação de imagens pertencentes a coleções existentes em outros contextos como uma coleção formada por imagens médicas.

Palavras-chave: Recuperação de Imagens da Web, Evidência Textual, Programação Genética.

Abstract

The developments in data storage and image acquisition technologies made the creation of huge databases or image collections possible. Moreover, the Web has been providing low cost and large scale acessibility to this material. Along with these factors, a variety of profitable activities demand image-based information. Typical examples are architectural projects, engineering and fashion designs, perfumes, new car models, marketing campaings and Internet sites, among others. All these needs explain the increasing in the interest in organizing, indexing, and retrieving digital images in the last decade.

This problem can be simply described as follows: given a textual or a image-based query, supplied by the user to a system which contains a image database, define the most relevant answer-set to the query. This is a challenging task due to the difficulty in extracting from the image sources the 'best' information needed for their representation and indexing.

This work presents a evolutionary framework for image retrieval based on the combination of multiple textual sources of evidence. It explores the Genetic Programming concepts, based on the continuous improvement of the solution quality for a given problem. Therefore, our motivation is to contribute with the development of search mecanisms capable of representing, in a more reliable manner, images in the the WWW. As a pratical result, a query made to this system will have more relevant images as result, increasing the user satisfaction and confidence on the search system.

Experiments performed with a collection extracted from the Web showed that, compared to the Bayesian Model presented in [Coelho et al., 2004] to solve the same problem, the evolutionary framework presents a performance twice as good, under precision, recall and MAP. These are important metrics to evaluate the quality of a image set returned from a query. Besides, the computational solution developed in this work presents great flexibility since the adition of new sources of evidences will be allowed with a minimum cost, besides the possibility to use the framework to retrieve images from other collections.

Keywords: Web Image Retrieval, Textual evidence, Genetic Programming.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Os avanços no armazenamento de dados e nas tecnologias para aquisição de imagens tornaram possível a criação de grandes bases ou coleções de imagens [da S. Torres et al., 2005]. Além disso, o enorme sucesso da Web tem proporcionado baixo custo e acessibilidade em larga escala deste material [Coelho et al., 2004]. Segundo os autores de [Veloso et al., 2000], conforme dados extraídos no ano de 2000, 24% das páginas da Web brasileira não possuem imagens, enquanto que 53% possuem até três imagens. Em média, cada página da Web brasileira contém 4,6 imagens. Em [Modesto et al., 2005], constatou-se conforme dados extraídos no ano de 2005, que o tipo de arquivo multimídia mais encontrado na Web brasileira corresponde a imagens. Aliado a esses fatores, uma variedade de atividades lucrativas demandam transferência de informação baseada em imagem. Exemplos típicos são projetos arquitetônicos, desenhos de engenharia e de moda, perfumes, novos carros, campanhas de marketing, páginas da Internet, etc. Todas essas necessidades têm levado a um crescimento do interesse em organizar, indexar e consultar imagens digitais na última década [Datta et al., 2008].

No caso de uma coleção formada por imagens, o serviço de consulta é de extrema relevância para o usuário. A consulta, em geral, tem como entrada alguma evidência textual ou uma imagem-exemplo. O resultado da consulta consiste em um conjunto-resposta que o sistema identificou como formado pelas imagens da coleção que são mais similares ao texto da consulta ou à imagem-exemplo. A apresentação desse conjunto, em geral, segue uma ordem tal que, dentre todas as imagens do conjunto-resposta, a primeira imagem resposta corresponde àquela mais similar à consulta, a segunda imagem resposta corresponde a segunda mais similar à consulta e assim por diante.

O problema conhecido na literatura relacionada como o Problema de Recuperação de Imagem pode ser enunciado da seguinte maneira: a partir de uma coleção formada

por imagens, é realizada uma consulta por um usuário, sendo esta formada por evidências textuais e/ou de conteúdo relacionados a uma imagem, defina o conjunto-resposta mais relevante para aquela consulta. Esta tarefa é desafiadora, dada a dificuldade em se definir quais as evidências são as mais relevantes para a indexação de informações sobre as imagens da coleção. Uma dificuldade imediata consiste em ordenar o conjunto-resposta com relação a uma consulta do usuário [Coelho, 2001].

A recuperação de imagem é geralmente baseada em uma das duas abordagens básicas: recuperação baseada em conteúdo ou recuperação baseada em texto. Na recuperação baseada em conteúdo, características da imagem, como cor, forma ou textura, são utilizadas para indexação e busca. Para fazer uma consulta, o usuário provê um esboço de um desenho do que deseja recuperar, uma imagem-exemplo (que pode ou não pertencer à coleção) do que deseja ou simplesmente uma combinação de ambos. Já na recuperação baseada em texto, alguma forma de descrição textual associada ao conteúdo da imagem é utilizada, tal como alguma anotação explícita sobre o conteúdo da imagem, texto nas redondezas da imagem, o próprio nome do arquivo da imagem ou alternativas. Por exemplo, em páginas HTML, o conteúdo da *tag* ALT, é utilizado para definir um texto que substitui a imagem correspondente, caso esta não esteja disponível ou não seja suportada pelo navegador ou outro mecanismo empregado para exibição do conteúdo de uma página HTML).

1.1 Motivação

Em linhas gerais, para uma ampla coleção, a adoção de indexação com posterior recuperação de imagem somente baseada no conteúdo apresenta como desvantagens o alto custo computacional para se extrair características de cada uma das imagens e uma dificuldade no processo de busca relacionado à necessidade intrínseca ao processo de se definir uma imagem de consulta similar à imagem objetivo para expressar a consulta do usuário. Já para a realização dos processos de indexação e recuperação baseada em texto identifica-se como problema a definição de quais evidências textuais relacionadas às imagens são mais relevantes para representar o conteúdo da imagem.

O presente trabalho apresenta um arcabouço evolucionário para recuperação de imagens. Este opera a partir da combinação de múltiplas evidências textuais e utiliza os preceitos da Programação Genética (PG), baseada na idéia de melhoramento contínuo da qualidade da solução de um dado problema a partir de soluções anteriores. Dentre os fatores que motivam a utilização da PG para o referido problema a ser tratado podem ser citados: a automatização do processo de melhoria do resultado; o fato de

não requerer que a função de avaliação a ser usada seja contínua, desde que ela possa diferenciar boas soluções (mais próximas do ótimo global) de soluções razoáveis (ótimos locais), proporcionando comumente bons resultados, independentemente da consulta realizada; o paralelismo intrínseco à técnica que proporciona um melhor desempenho no processo de busca; a capacidade de escolher automaticamente as melhores evidências e combiná-las de forma efetiva e a obtenção de bons resultados quando utilizada a Programação Genética para solução de outros problemas de ranqueamento Lacerda et al. [2006], de Almeida et al. [2007], de Carvalho et al. [2008] e Ferreira et al. [2008]. Em Lacerda et al. [2006] foi proposta uma abordagem para solucionar a difícil tarefa de associar automaticamente anúncions publicitários com páginas Web. Uma estratégia evolucionária para definir o ranking de coleções distintas foi proposta e avaliada em de Almeida et al. [2007]. Já o problema conhecido na literatura relacionado como Identificação de Réplicas foi tratado em de Carvalho et al. [2008]. Arcabouços para recuperação de imagens por conteúdo com realimentação de relevância foram tratados em Ferreira et al. [2008].

Nossa motivação, portanto, é contribuir com o desenvolvimento de mecanismos de busca capazes de recuperar de uma maneira mais eficaz as imagens de uma coleção, no caso da World Wide Web. Como resultado prático, uma consulta realizada pelo usuário desse sistema de recuperação de imagens terá como resultado imagens mais relevantes sob a ótica do próprio usuário, aumentando assim o grau de satisfação e confiança do usuário no sistema de busca. Dentre os desafios advindos com a utilização dessa coleção podem ser mecionados: a heterogeneidade dos tipos de imagens, a inexistência de padronização no preenchimento das *tags* HTML onde as imagens são inseridas, a inclusão de referências incorretas para outro documento HTML, dentre outros.

1.2 Objetivos

O objetivo deste trabalho é desenvolver um arcabouço para recuperação de imagens da Web, capaz de retornar conjuntos-resposta mais relevantes sob a ótica do usuário que realizou uma determinada consulta textual. Particularmente, estaremos interessados em:

- Obter um melhor desempenho sob os aspectos de precisão e revocação, métricas importantes para avaliação da qualidade dos resultados da consulta retornados, em comparação com o baseline que utiliza o Modelo Bayesiano para solução do Problema de Recuperação de Imagem. A escolha do Modelo Bayesiano como baseline decorre do fato dele servir como uma ferramenta teórica e formal de

avaliação e estudo tanto de evidências textuais em isolado quanto de forma de combiná-las;

- Estudar o comportamento do arcabouço para diferentes configurações dos parâmetros aderentes à técnica da Programação Genética;
- Contemplar uma alta flexibilidade na arquitetura do arcabouço, de forma a viabilizar a adição de novas evidências textuais à solução computacional em trabalhos futuros com custo mínimo de manutenção do mesmo.

1.3 Trabalhos Relacionados

Em decorrência do amplo contexto em que o presente trabalho está inserido, a revisão de importantes referências foi separada por tópicos concernentes aos assuntos relacionados ao Problema da Recuperação de Imagens. Essa separação compreende as duas abordagens básicas, a combinação de informações textuais e visuais (ou de conteúdo) da imagem e à técnica da Programação Genética.

1.3.1 Recuperação de Imagens

Grandes sistemas de recuperação de imagens, como Google [Google, 2009], Yahoo [Yahoo, 2009] e MSN [MSN, 2009], permitem que os usuários da Web realizem consultas a imagens a partir de palavras-chave, de forma similar a busca por páginas web. Esses sistemas utilizam anotação automática de imagem e técnicas de índice invertido para atender as consultas dos usuários e retornar as imagens relevantes com base nas palavras-chave informadas na consulta.

De forma similar a estas, em [Chang, 2008], é proposto um sistema para recuperação de imagens da web que contempla técnicas de anotação automática das imagens da coleção. Nesse cenário, são armazenadas somente as imagens reduzidas, denominadas *thumbnails* e informações da imagem como largura, altura e formato do arquivo, além de dados extraídos do próprio documento HTML, como o conteúdo das *tags img*, *table* e *a*. Estas informações são armazenadas em um índice invertido, por meio de técnicas tradicionais de recuperação de informação, e são a base para a geração de anotações automáticas das imagens da coleção. Experimentos realizados com uma coleção de 20.219.113 documentos HTML e 352.295.297 imagens consideraram tipos distintos de anotações. Foram avaliadas dez consultas, por vinte usuários, com uma técnica denominada pelos autores de teste cego. Como resultado, constatou-se que a anotação gerada a partir da geração de uma assinatura digital da imagem por meio do método

Message-Digest algorithm 5 pode melhorar a qualidade das imagens recuperadas. O Message-Digest algorithm 5 (MD5) é um algoritmo de hash de 128 bits unidirecional desenvolvido pela RSA Data Security, Inc., descrito na RFC 1321, usado por softwares com protocolo ponto-a-ponto (P2P), verificação de integridade e logins.

Em [Coelho et al., 2004] é empregada uma técnica da recuperação de imagens da Web a partir de evidências textuais. São consideradas quatro fontes de evidência contidas em documentos Web: *tags* de descrição, meta *tags*, texto completo e passagens textuais. O objetivo do trabalho é avaliar o resultado das consultas realizadas utilizando a cada evidência textual isoladamente e a combinação das referidas evidências. Para a recuperação das imagens a partir de múltiplas evidências textuais, é empregado o conceito de Redes Bayesianas. Os experimentos mostraram que a combinação de evidências múltiplas produz melhores resultados que a utilização de uma única fonte de evidência. Além disso, experimentalmente comprovou-se que a utilização de passagens de texto com 40 termos produz melhores resultados que a utilização de todo o texto do documento HTML. As duas principais fontes de evidências destacadas pelos experimentos foram as *tags* de descrição e as passagens textuais com 40 termos. Ainda neste trabalho os autores sugerem a inclusão de evidências relacionadas ao conteúdo da imagem em si como uma possibilidade de melhoria da qualidade dos resultados obtidos.

O trabalho [Torres et al., 2009] pode ser reconhecido como o estado da arte em termos de recuperação de imagens utilizando informações acerca do conteúdo em si. Neste trabalho, os autores definem um framework baseado na técnica da Programação Genética para combinar diferentes descritores de forma. Um descritor de forma é um tipo de descritor de imagem que considera a característica forma de uma imagem. Descritores de imagens podem ser definidos como um par composto por um vetor de características extraídas e uma função de distância. O vetor de características representa um conjunto de propriedades de uma determinada imagem (cor, textura e forma, por exemplo) enquanto que a função de distância diz respeito a dissimilaridades entre duas imagens tendo como base características das mesmas. Estas distâncias podem ser mensuradas através de métricas como a Euclidiana [**?**]. Em [Torres et al., 2009] são realizados experimentos sobre duas coleções de imagens distintas e comparados os resultados de consultas utilizando cinco descritores de forma isoladamente e a combinação de todos os descritores por meio de duas técnicas: Algoritmos Genéticos e Programação Genética. Os resultados mostraram que, para ambas as coleções, a solução que emprega a Programação Genética produz melhores resultados. Além da avaliação das técnicas em si, em [Torres et al., 2009] são avaliadas conjuntamente as funções de avaliação da qualidade dos indivíduos, soluções do problema avaliado, durante o processo de evolução. Essas funções são descritas em [Fan et al., 2004a]. Os experimentos real-

izados também foram capazes de indicar quais dessas funções são as mais apropriadas para a aplicação proposta.

Expressivas contribuições da década atual relacionadas ao tema de recuperação de imagem são apresentadas e discutidas em [Datta et al., 2008]. Foram analisados cerca de 300 publicações, recuperados com a máquina de busca *Google Scholar* e os valores de pontuação das citações presente neste mecanismo. Neste cenário, foram contrastados os primeiros anos da recuperação de imagem com o progresso do campo na década corrente. Além disso, foram especificadas futuras direções da área. Os autores de [Datta et al., 2008] acreditam que a área de recuperação de imagens experimentará uma mudança de paradigma no futuro breve, com um maior foco em trabalhos de domínio específico, o que tende a gerar impacto considerável no dia-a-dia. Dentre as conclusões do trabalho estão: a imensa diversidade de trabalhos da área, a presença da recuperação de imagens como elemento catalisador da associação mais próxima de campos de pesquisa sem qualquer correlação *a priori* e a constatação de que a maior parte dos trabalhos relacionados às áreas de aprendizado de máquina, visão computacional e inteligência artificial, com alguma relação com a recuperação de imagens, foram citados muito frequentemente, além de terem sido publicados em eventos de recuperação de imagem. De uma forma geral, as análises dos dados coletados indicam ainda que enquanto aspectos relacionados a sistemas, extração de características e *relevante feedback* tenham recebido muita atenção, aspectos relacionados a domínios específicos como interface, visualização, escalabilidade e avaliação têm tradicionalmente recebido menor consideração. Devido a natureza dinâmica da informação,os autores decidiram armazenar as informações coletadas em uma página acessível pela internet e atualizar periodicamente[1].

Com o objetivo de viabilizar consultas a uma base de dados mista formada por imagens e texto, além de melhorar a qualidade das respostas obtidas, em [Torres & Falcão, 2006] é apresentado um framework para Sistemas de Informação sobre Biodiversidade. Utilizando conceitos relacionados com a modularização e maior extensibilidade do sistema desenvolvido, os autores apresentam uma arquitetura para o framework de forma que o usuário final possa realizar buscas a partir de texto, imagem ou ambos. Os experimentos realizados mostraram que a combinação de metadados de ambos os tipos de evidências (imagem e texto) apresentam melhores resultados que a utilização de apenas uma fonte de informação. Outros experimentos envolvendo também a recuperação baseada no conteúdo da imagem, além dos metadados de ambas as fontes (imagens e texto), apresentaram melhores resultados. Os resultados vão ao encontro do esperado

[1]As estatísticas estão disponíveis no endereço eletrônico http://wang.ist.psu.edu/survey/analysis.

pelos próprios autores, pois segundos eles, a combinação de várias fontes de evidência heterogêneas contribuem em algum grau para o resultado final.

Em [Iskandar et al., 2005] é apresentado um sistema que combina evidências e define o *rank* de consultas com base no texto e no conteúdo das imagens. Esta solução é resultado da fusão de dois subsistemas: a ferramenta de recuperação de imagem (GNU Image Finding Tool - GIFT) e o sistema de recuperação de XML híbrido. No subsistema GIFT foi utilizado o HSV (Hue-Saturation-Value), os Filtros de Gabor para descrever a textura e o clássico algoritmo IDF para avaliar as caraterísticas da imagem. Para o segundo subsistema foram combinadas as características retornadas pela Zettair[2], uma máquina de busca de texto completo, e uma base de dados XML nativa, denominada eXist[3]. Cada consulta foi avaliada separadamente por cada subsistema. Para combinar ambos os valores de rank foi utilizada uma equação linear simples, envolvendo a parametrização de pesos a serem atribuídos a cada natureza de valor. A avaliação do sistema quanto a relevância das imagens e dos componentes de texto foi realizada sobre a base de dados do INEX 2005, em seis avaliações distintas. O *Initiative for the Evaluation of XML Retrieval* (INEX) fornece uma plataforma para que os participantes avaliem a eficácia de suas técnicas de recuperação de informação, utilizando procedimentos de avaliação uniforme e um fórum para comparar os resultados. O segmento multimídia do INEX contempla consultas que utilizam estrutura e conteúdo para representar um tópico. No total, foram formulados vinte e três tópicos multimídia considerados relevantes para ponderação. A métrica de avaliação da TREC (TRECeval) foi utilizada como o método de avaliação oficial, para as seguintes medidas: P@1, P@5, P@10, MAP e R-Precision. As métricas P@1, P@5 e P@10 representam o valor da precisão obtida, respectivamente, pelo primeiro documento, cinco primeiros documentos e dez primeiros documentos retornados. A métrica *Mean Average Precision* (MAP) representa o valor médio das precisões calculadas sobre o valor médio de um conjunto de consultas. Sendo R o conjunto de documentos relevantes de uma consulta e |R| o número de documentos relevantes contidos neste conjunto, a métrica R-precision consiste na precisão dos |R| primeiros documentos retornados pelo método avaliado. Os resultados mostraram que o sistema de recuperação XML orientado a conteúdo se beneficia da utilização de alguma evidência advinda do sistema CBIR, fato este observado em especial quando utilizadas as métricas MAP e R-precision.

[2]http://www.seg.rmit.edu.au/zettair/
[3]http://exist-db.org/

1.3.2 Programação Genética Aplicada à Recuperação de Informação

Diferentes abordagens para descoberta de funções de ordenação baseadas em técnicas de aprendizado de máquina, tais como algoritmos genéticos e programação genética, têm sido propostas na literatura de recuperação de informação.

Em Fan et al. [2000] é proposta uma nova abordagem para gerar automaticamente estratégias de ponderação de termos para diferentes contextos baseada em Programação Genética. Como argumento foi apresentado que cada contexto específico exige um estratégia diferente de ponderação de termos, isto é, uma função de ordenação deve ser adaptada para diferentes coleções de documentos.

Ao longo dos anos de pesquisa, os trabalhos apresentados em Fan et al. [2004b], Fan et al. [2004c], Fan et al. [2004d] e Fan et al. [2005] têm demonstrado que a Programação Genética é uma técnica eficaz para melhorar a eficácia da tarefa de recuperação de informação. Em Fan et al. [2004a], é estudado também o efeito de diferentes funções de utilidade, isto é, funções que, quando usadas como funções de aptidão, privilegiam a recuperação de documentos relevantes no topo da lista de documentos retornados. Uma função de aptidão é aquela utilizada na Programação Genética para avaliar a qualidade de uma solução do problema ou indivíduo. Em todos esses trabalhos, os terminais, elementos constituintes de um indivíduo definidos no arcabouço evolucionário, são baseados fundamentalmente em informações estatísticas brutas da coleção, documentos, termos e consultas, tais como freqüência do termo (tf), número total de documentos, tamanho em bytes de um documento, número de termos distintos em um documento etc.

Em Fan et al. [2004c], os autores comparam sua abordagem (referenciada de agora em diante por FAN-GP) com uma abordagem baseada em aprendizado de máquina que utiliza redes neurais. Nesse trabalho, os resultados mostram que FAN-GP supera substancialmente a estratégia baseada em redes neurais. Em Fan et al. [2005], FAN-GP descobre funções de ordenação para busca na Web, explorando a informação de estrutura dos documentos Web, superando uma outra abordagem baseada em máquinas de vetor de suporte (na tarefa de recuperação de informação).

Em [de Almeida et al., 2007] foram investigadas estratégias para definir o ranking para coleções específicas. O método elaborado e descrito é capaz de considerar as únicas e importantes características de cada coleção, assim a função descoberta é mais efetiva que qualquer solução genérica. Para isso, foi utilizada a técnica de Programação Genética. Porém, diferentemente de outras abordagens baseadas nessa técnica evolucionária, que utilizam somente informações estatísticas básicas advindas de termos e

documentos, no referido trabalho foram adotados componentes ricos, significativos e comprovadamente efetivos que estão presente em fórmulas de ranking bem conhecidas, tais como Okapi BM25 e Pivoted TF-IDF. Em termos de Programação Genética, o framework apresentado em [de Almeida et al., 2007] é basicamente um processo iterativo com duas fases: treinamento e validação. Para cada fase, é selecionado um conjunto de consultas e um conjunto de documentos da coleção, denominados conjunto de treinamento e conjunto de validação. Os terminais contêm informações extraídas das principais fórmulas de ranking publicadas na literatura relacionada à recuperação de informação. Partiu-se do princípio de que ao prover as fórmulas descobertas como terminais do programa, o processo genético pode levar vantagem de todo conhecimento que foi aplicado para produzi-lo. Essa abordagem foi denominada de Abordagem de Componentes Combinados - ACC. Como consequência, é capaz de explorar melhor o espaço de busca. Para avaliar a abordagem genética proposta, foram realizados experimentos com as coleções TREC-8 e WBR99. Quase todos os parâmetros e configurações utilizados na abordagem apresentada baseiam-se no trabalho [Fan et al., 2004b]. Os resultados indicam que o uso de componentes significativos em um framework baseado em Programação Genética, conduz a funções de ranking efetivas que têm melhor desempenho que baselines (TF-IDF padrão, BM25 e outra abordagem GP apresentada em [Fan et al., 2004b]). Comparando a abordagem ACC com a anterior, que também utiliza a técnica de Programação Genética, contatou-se que as funções de ranking da abordagem ACC converge para bons resultados de forma mais rápida que a abordagem GP utilizada como baseline. Além disso, a especialização na geração de melhores indivíduos em relação ao conjunto de consultas consideradas durante a fase de treinamento (*overtraining*) também foi reduzida.

1.4 Organização do documento

Os capítulos da dissertação estão organizados da seguinte forma:

Capítulo 2: Apresenta os conceitos básicos relacionados com a técnica da programação genética e as métricas utilizadas para avaliação da qualidade dos resultados obtidos em sistemas de recuperação de informação.

Capítulo 3: Contém um estudo quantitativo e qualitativo da coleção Web indexada e utilizada neste trabalho, um detalhamento das evidências textuais consideradas e das ferramentas empregadas para a indexação da base de dados e o processo de escolha das consultas textuais empregadas e de definição do conjunto de imagens relevantes.

Capítulo 4: Descreve dois métodos de recuperação de imagens: o baseline que empregada o Modelo de Redes Bayesianas e o Arcabouço Evolucionário desenvolvido no presente trabalho.

Capítulo 5: Apresenta e analisa os resultados obtidos pelos métodos de recuperação de imagens apresentados no capítulo 5.

Capítulo 6: Faz um resumo dos objetivos definidos com os resultados alcançados, além de projetar futuras abordagens para a evolução deste trabalho.

Capítulo 2

Conceitos Básicos

2.1 Programação Genética

Algoritmos Genéticos (AG) e Programação Genética (PG) formam um conjunto de técnicas de Inteligência Artificial para solução de problemas baseados no princípio de herança biológica e evolução. A principal diferença entre essas duas técnicas refere-se à representação interna, ou estrutura de dados, de um indivíduo Zhang [2006]. Em geral, aplicações em AG representam um indivíduo como uma string de bits de tamanho fixo, como (11011110...) ou uma sequência de números reais (1,2, 2,4,...). Por outro lado, na PG são utilizadas estruturas de dados mais complexas como árvores, listas encadeadas ou pilhas. As estruturas de dados não têm tamanho fixo, entretanto pode ser restringido o seu tamanho por um parâmetro denominado profundidade. A profundidade corresponde ao número de níveis, ou distância, entre o nodo raiz e os nodos folhas ou terminais da árvore. Cada solução potencial, na PG, é chamada de um indivíduo (ou seja, um cromossomo) em uma população Zhang [2006].

A figura 2.1 apresenta um esquema de funcionamento da Programação Genética. Inicialmente é criada uma população de indivíduos, soluções potenciais do problema, geralmente por meio de um método aleatório. A partir deste ponto, enquanto o critério de parada (que pode ser por exemplo um número máximo de gerações) para o algoritmo não for atingido, são realizadas avaliações sobre a qualidade de cada indivíduo da população atual. Para isso, é atribuído um valor para uma medida de aptidão (também conhecido como *fitness*) para cada indivíduo, com base em uma lei de avaliação comum a todos os indivíduos de todas as iterações subsequentes. Após a avaliação da população, são aplicadas operações genéticas, tais como cruzamento, mutação e reprodução, com o objetivo de criar indivíduos mais diversos e com melhor eficácia nas gerações subsequentes. Em seguida, a população é novamente avaliada e o teste de parada é

realizado. Esse ciclo repete-se até que o critério de parada seja satisfeito.

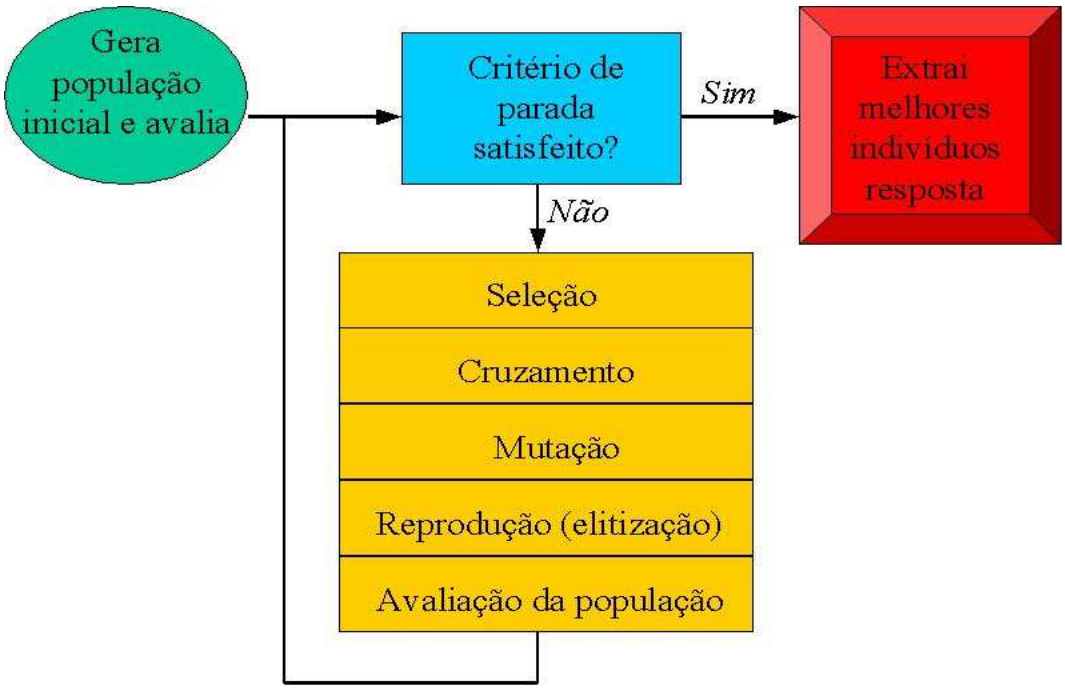

Figura 2.1. Esquema de funcionamento da Programação Genética.

2.1.1 Indivíduos

Um indivíduo é formado por funções e terminais. Os terminais são entradas do programa a serem evoluídos, ou seja, representam informações relacionados com o problema em estudo. As funções são operadores utilizados para combinar os terminais. Exemplos de funções comumente utilizadas são os operadores aritméticos (soma, subtração, divisão, etc) e as constantes.

2.1.2 Tamanho dos Indivíduos

As árvores envolvidas na Programação Genética podem crescer enormemente. Para evitar o despendimento de tempo na avaliação de um pequeno número de árvores, é possível estabelecer limites no número de nodos e/ou na profundadidade de um indivíduo. Nos experimentos de Koza, apresentados em [Koza, 1992] e [Koza, 1994], a profundidade máxima de um indivíduo foi definido como sendo 17, sem qualquer restrição do número de nodos.

2.1.3 Inicialização da População

A população inicial é comumente gerada aleatoriamente. Entretanto, durante esse processo é possível adotar estratégias distintas para criação dos indivíduos. Em [Zongker & Punch, 1996] são citadas:

- *Full*: este método gera somente árvores cheias, ou seja, árvores cujos terminais estão no mesmo nível. Outra forma de dizer isso é que o tamanho do caminho da árvore, de qualquer terminal até o nodo raiz, é o mesmo.
- *Grow*: nesta estratégia, é escolhido um nodo qualquer como sendo raiz. Este nodo, recursivamente chama a si próprio para gerar árvores filhas. É um método restritivo, pois cada árvore tem uma profundidade máxima.
- *Half-and-half*: este método simplesmente escolhe o método *full* para 50% dos casos e o método *grow* para os 50% restantes.

2.1.4 Operadores Genéticos

2.1.4.1 Seleção

O método da seleção é uma rotina utilizada para selecionar um indivíduo de uma população. Dentre as abordagens existentes, as seleções são utilizadas para dois propósitos: selecionar os indivíduos que serão utilizados por um operador genético de cruzamento ou mutação e na escolha de indivíduos que serão trocados entre populações, em problemas que empreguem mais de uma população de indivíduos. Três métodos de seleção comumente utilizados são:

- Seleção proporcional ao valor da função de aptidão ou roleta russa: este seleciona um indivíduo com base no valor proporcional à função de aptidão, em comparação com o valor da função de aptidão total da população. Quando há n indivíduos na população, um indivíduo i será escolhido com probabilidade:

$$p_i = \frac{f_a(i)}{\sum_{j=1}^{n} f_a(j)}$$

  onde $f_a(j)$ é o valor da função de aptidão para o indivíduo j.

  Ou seja, cada indivíduo tem uma probabilidade definida com base na fórmula acima, de forma que a soma das porções é igual a 1. Um número aletório entre 0 e 1 é gerado, simulando a idéia de que a roleta é então girada, para determinar qual indivíduo é o próximo a ser selecionado. Dessa forma, indivíduos com um

maior valor da função de aptidão tem uma chance maior de ser escolhido, mas todo indivíduo tem alguma chance. Maiores detalhes sobre o funcionamento deste método pode ser encontrado em [Goldberg, 1989].

- Seleção Gulosa ou *overselection*: apesar da seleção proporcional ao valor da função de aptidão ser considerado boa para a maior parte das aplicações, algumas vezes é desejável acelerar o processo. Assim, para grandes populações (definidas como maiores que 500 em [Koza, 1992]), deve ser utilizada a seleção gulosa. Este método separa a população em dois grupos, denominados grupo I e grupo II. Na formulação padrão de Koza, 80% do tempo, os indivíduos são selecionados do grupo I (formado a partir da seleção proporcional ao valor da função de aptidão) e 20% do tempo a partir dos indivíduos que formam o grupo II. A formação dos grupos pode ser arbitrária, mas Koza adota o critério da proporcionalidade com relação ao valor da função de aptidão do indivíduo.

- Seleção por Torneio ou *Tournament Selection*: está abordagem surgiu da técnica dos Algoritmos Genéticos, como uma tentativa de se evitar a seleção constante dos mesmos indivíduos (*overselection*), o que por sua vez poderia resultar em uma convergência prematura da solução. Na seleção por torneio n indivíduos são escolhidos de uma forma aleatória, segundo uma distribuição uniforme. Em seguida, o melhor dentre os n indivíduos, tendo como critério o valor da função de aptidão, é selecionado. O valor de n refere-se ao tamanho do torneio. Koza utiliza este tipo de seleção, com um valor de n igual a 7, na maior parte das execuções apresentadas em [Koza, 1994].

2.1.4.2 Cruzamento

O cruzamento é efetuado a partir de dois indivíduos escolhidos na fase de seleção. Depois de escolhidos os indivíduos, a sua informação genética é partilhada, dando origem a dois novos indivíduos diferentes dos anteriores. Para tal, efetua-se um corte em cada uma das árvores que representam um indivíduo e troca-se depois os ramos cortados de cada um deles. Como resultado, são formados dois novos indivíduos.

A figura 2.2 indica o modo como são efetuados os cortes em cada um dos indivíduos selecionados para cruzamento. De uma forma aleatória, são escolhidos os ramos que vão sofrer o corte. Após terem sido efetuados os cortes em cada um dos indivíduos, a informação genética de ambos é partilhada dando origem a dois novos indivíduos com características genéticas diferentes dos primeiros, tal como indica a figura 2.3.

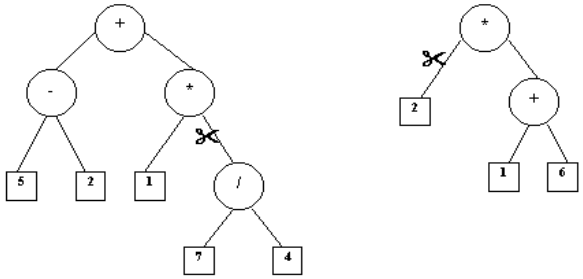

Figura 2.2. Definição do ponto de corte de dois indivíduos na operação de Cruzamento.

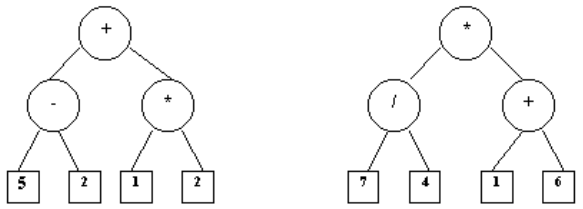

Figura 2.3. Dois novos indivíduos gerados com a operação de Cruzamento.

2.1.4.3 Mutação

A mutação nem sempre é efetuada sobre uma determinada população. A existência dessa operação genética depende de um valor que indica a probabilidade de existir mutação numa determinada geração.

Quando a mutação é efetuada, é escolhido aleatoriamente um indivíduo que sofrerá a mutação. Em um ponto do indivíduo também escolhido de forma aleatória, ele sofrerá uma mutação que poderá ser a nível local de um nó da árvore, ou modificando um ramo do mesmo, escolhido de forma aleatória.

A figura 2.4, exemplifica uma mutação efetuada apenas no nível de um nó isolado da árvore. Neste caso, o operador de multiplicação é substituído pelo de adição, mantendo-se igual a restante estrutura da árvore.

Um outro tipo de mutação consiste em escolher aleatoriamente um ramo da árvore, retirá-lo e substituí-lo por um outro ramo gerado aleatoriamente. Este tipo de mutação é capaz de modificar a estrutura da árvore, dado que não obriga que o novo ramo tenha o mesmo tamanho do que foi retirado. A figura 2.5 ilustra o método descrito.

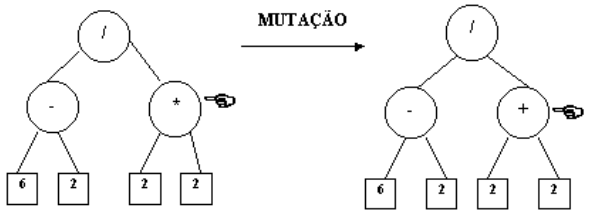

Figura 2.4. Definição do ponto de mutação.

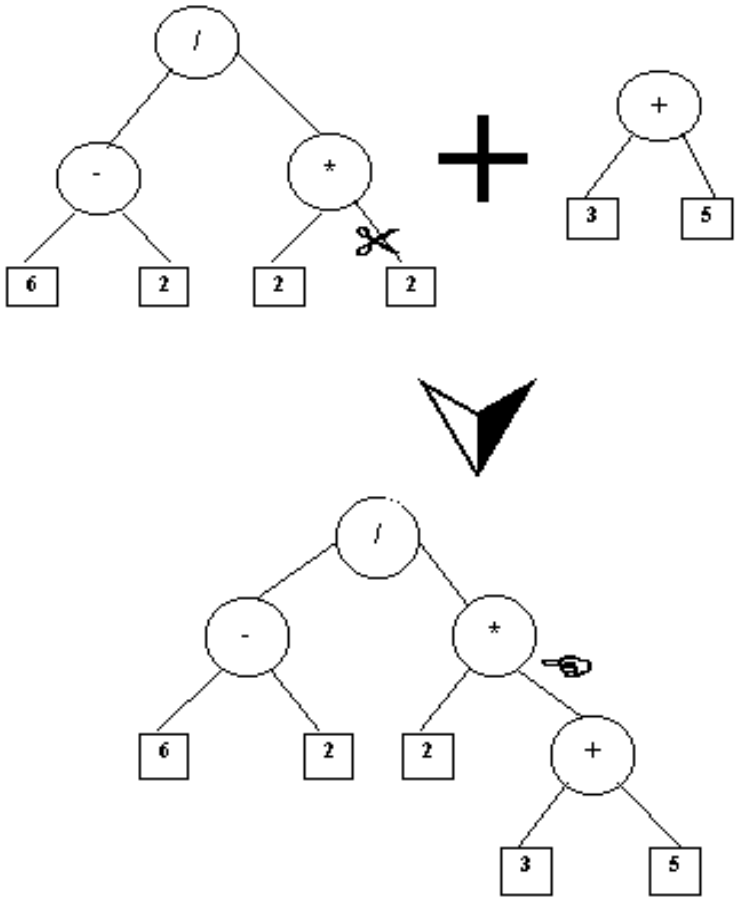

Figura 2.5. Método que emprega a geração de parte do novo indivíduo com a Mutação.

2.2 Métricas para Avaliação

Uma vez que a consulta realizada pelo usuário é por natureza vaga, os documentos recuperados não são respostas exatas e têm que ser ordenados de acordo com a sua relevância para a consulta. Assim, os sistemas de recuperação de informação necessitam avaliar o quão bom é o conjunto resposta (conjunto recuperado). Este tipo de avaliação

é conhecida como uma avaliação de eficácia de recuperação.

2.2.1 Precisão e Revocação

Considere uma requisição de informação I , de uma coleção de referência de teste, e seu conjunto R de documentos relevantes. Seja $|R|$ o número de documentos neste conjunto. Assuma que uma dada estratégia de recuperação (que está sendo avaliada) processa a requisição de informação I e gera um conjunto de documentos de resposta A. Seja $|A|$ o número de documentos neste conjunto. Além disso, seja $|R_a|$ o número de documentos na interseção dos conjuntos R e A. A Figura 2.6 ilustra estes conjuntos.

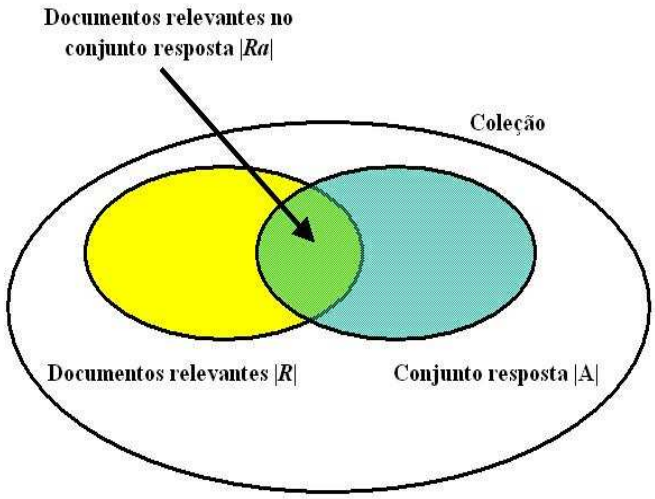

Figura 2.6. Representação esquemática do cálculo da precisão e da revocação.

As medidas de revocação e precisão são definidas a seguir:

- Revocação (recall): é a fração dos documentos relevantes (o conjunto R) que foram recuperados, ou seja,

$$\textbf{Revocação} = \frac{|R_a|}{|R|}$$

- Precisão: é a fração dos documentos recuperados (o conjunto A) que é relevante, ou seja,

$$\textbf{Precisão} = \frac{|R_a|}{|A|}$$

Revocação e precisão, como definidas anteriormente, assumem que todos os documentos no conjunto resposta A foram examinados (ou vistos). Porém, geralmente

não são apresentados ao usuário todos os documentos do conjunto resposta A de uma só vez. Ao invés disto, os documentos em A são primeiramente ordenados de acordo com seu grau de relevância (ou seja, um ranking é gerado). O usuário então examina esta lista ordenada começando dos elementos do topo. Nesta situação, as medidas de revocação e precisão variam enquanto o usuário continua com a análise do conjunto resposta A. Assim, a própria avaliação sugere a geração da curva de precisão versus revocação, como definida na seqüência [Baeza-Yates & Ribeiro-Neto, 1999].

Como anteriormente, considere uma coleção de referência e seu conjunto de requisições de informação tomado como exemplo. Vamos focar em uma dada requisição de informação relacionada com uma consulta q formulada. Assuma que um conjunto R_q contendo os documentos relevantes para q foi definido. Sem perda de generalidade, assuma também que o conjunto Rq é composto dos seguintes documentos

$$R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$$

Assim, de acordo com um grupo de especialistas, há dez documentos que são relevantes para a consulta q.

Considere agora um novo algoritmo de recuperação que acabou de ser desenvolvido. Assuma que este algoritmo retorna, para cada consulta q um ranking de documentos no conjunto resposta como a seguir:

Ranking para a consulta q:

$1.d_{123}$•	$6.d_9$•	$11.d_{38}$
$2.d_{84}$	$7.d_{511}$	$12.d_{48}$
$3.d_{56}$•	$8.d_{129}$	$13.d_{250}$
$4.d_6$	$9.d_{187}$	$14.d_{113}$
$5.d_8$	$10.d_{25}$•	$15.d_3$•

Os documentos que são relevantes para a consulta q são marcados com o símbolo (•) após o identificador do documento. Se examinarmos este ranking, começando do documento de topo, observamos os seguintes pontos. Primeiro, o documento d_{123} que ocupa a posição 1 é relevante. Por isso, este documento corresponde a 10% de todos os documentos relevantes no conjunto R_q. Por isso, dizemos que temos uma precisão de 100% com 10% de revocação. Segundo, o documento d_{56} que é ordenado como o número 3 é o próximo documento relevante. Neste ponto, nós dizemos que temos uma precisão de aproximadamente 66% (dois documentos dos três avaliados são relevantes) com 20% de revocação (dois dos dez documentos relevantes foram vistos). Terceiro, se continuarmos com nossa avaliação do ranking gerado, podemos gerar uma curva de precisão versus revocação como ilustrado na Figura 2.7.

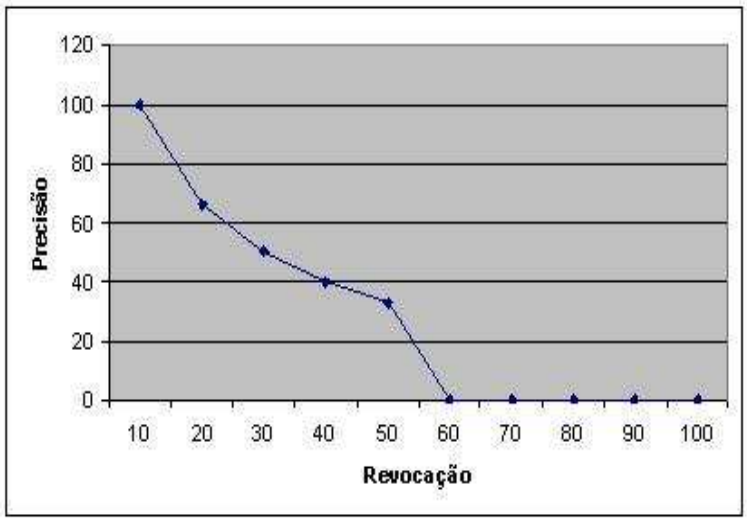

Figura 2.7. Curva de Precisão versus revocação nos 11 níveis.

A precisão em níveis de revocação maiores que 50% cai para 0 pois nem todos os documentos relevantes foram recuperados. Esta curva de precisão versus revocação é geralmente baseada em em uma revocação padrão de 11 (ao invés de dez) níveis, que são 0%, 10%, 20%, ..., 100%. Para o nível de revocação de 0%, a precisão é obtida por meio de um procedimento de interpolação como detalhado a seguir.

Na Figura 2.7, os valores de precisão e revocação correspondem a uma única consulta. Geralmente, entretanto, algoritmos de recuperação são desenvolvidos para avaliar diversas consultas distintas. Para avaliar a eficácia de recuperação de um algoritmo sobre todas as consultas de teste, calculamos a precisão média em cada nível de revocação, como a seguir:

$$\boxed{\overline{P}(r) = \sum_{i=1}^{Nq} Pi(r)/Nq}$$

onde $\overline{P}(r)$ é a precisão média no nível de revocação r, Nq é o número de consultas utilizadas e $Pi(r)$ é a precisão no nível de revocação r para a í-ésima consulta.

Uma vez que os níveis de revocação para cada consulta podem ser distintos dos 11 níveis de revocação padrão, a utilização de um mecanismo de interpolação se faz necessário. Por exemplo, considere novamente o conjunto de 15 documentos apresentados anteriormente. Assunma que o conjunto de documentos relevantes para a consulta q mudou e agora é dado por:

$$\text{Rq} = d_3,\, d_{56},\, d_{129}$$

Novamente, os documentos que são relevantes para a consulta q são marcados com o símbolo (•) após o número do documento:

1.d_{123}	6.d_9	11.d_{38}
2.d_{84}	7.d_{511}	12.d_{48}
3.d_{56}•	8.d_{129}•	13.d_{250}
4.d_6	9.d_{187}	14.d_{113}
5.d_8	10.d_{25}	15.d_3•

Neste caso, o primeiro documento relevante no ranking para a consulta q é d_{56}, que provê um nível de revocação de 33,3% (com precisão também igual a 33,3% - 1 documento dos recuperados é relevante), pois, neste ponto, um terço de todos os documentos relevantes foram identificados. O segundo documento relevante é d_{129} que provê um nível de revocação de 66,6% (com precisão igual a 25% - 2 documentos dos 8 recuperados são relevantes). O terceiro documento é d_3 que provê um nível de revocação de 100% (com precisão igual a 20% - 3 documentos dos 15 recuperados são relevantes). A precisão nos 11 níveis padrões de revocação é interpolada como descrito a seguir.

Seja r_j, $j \in 0, 1, 2, ..., 10$ ser uma referência ao j-ésimo nível de revocação padrão (ou seja, r_5 é uma referência ao nível de 50% de revocação). Então:

$$\boxed{P_j(r_j) = max_{r_j \leq r \leq r_{j+1}} P(r)}$$

A precisão interpolada no j-ésimo nível de revocação padrão é igual à máxima precisão conhecida em qualquer nível de revocação entre o j-ésimo e o (j+1)-ésimo níveis de revocação [Baeza-Yates & Ribeiro-Neto, 1999].

No exemplo anterior, esta regra de interpolação resulta na precisão e revocação ilustradas na Figura 2.8.

Nos níveis de revocação de 0%, 10%, 20% e 30%, a precisão interpolada é igual a 33,3% (que é a precisão conhecida no nível de revocação de 33,3%). Nos níveis de revocação de 40%, 50% e 60%, a precisão interpolada é de 25% (que é a precisão no nível de revocação de 66,6%). Nos níveis de revocação de 70%, 80% e 90%, a precisão interpolada é de 20% (que é a precisão no nível de revocação de 100%).

A curva de precisão versus revocação que resulta do cálculo do valor médio dessas métricas para várias consultas é geralmente referenciada como figuras de precisão versus revocação. Tais figuras são geralmente utilizadas para comparar a eficácia de recuperação de algoritmos distintos desenvolvidos para recuperação de informação.

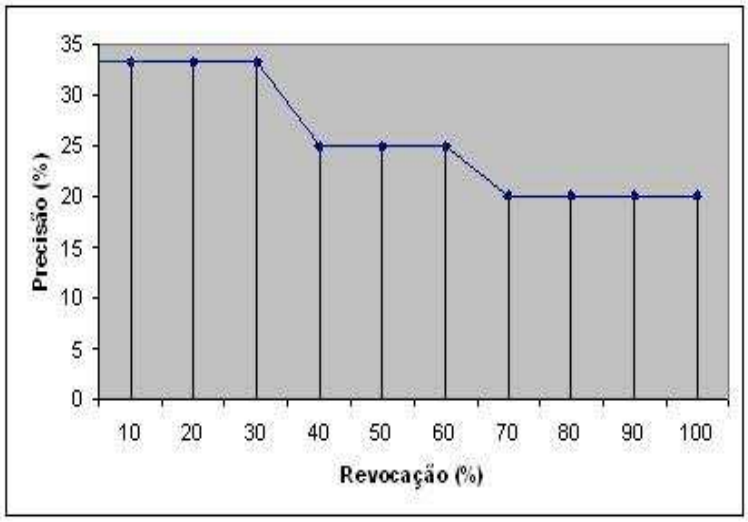

Figura 2.8. Curva de Precisão interpolada.

2.2.2 Mean Average Precision

A métrica *Mean Average Precision* (MAP) representa o valor médio das precisões calculadas sobre o valor médio de um conjunto de consultas e definida como:

$$MAP = \frac{1}{k}\sum_{q=1}^{k} P_q$$

onde k é o número total de consultas e P_q é a precisão média nos documentos relevantes recuperados para a consulta q, sendo definida como:

$$P_q = \frac{1}{n}(\sum_{i=1}^{m} r_{q_i} * \frac{1}{i}\sum_{j=1}^{i} r_{q_j})$$

onde m é o número de documentos retornados como resultado da consulta q , n é o número total de documentos relevantes para a consulta q e r_{q_i} é 1 se o i-ésimo documento retornado é relevante e 0 zero caso contrário. Esta métrica intuitivamente atribui um peso maior aos documentos relevantes que aparecem nas posições do topo da lista do *rank* [Silva et al., 2009].

2.2.3 R-Precision

Sendo R o conjunto de documentos relevantes de uma consulta e |R| o número de documentos relevantes contidos neste conjunto, a métrica R-precision consiste na precisão dos |R| primeiros documentos retornados pelo método avaliado. Pode ser avaliado por

meio da seguinte relação:

$$R - precision = \frac{nr}{|R|}$$

onde nr é o número de documentos relevantes encontrados dentre os |R| primeiros retornados pelo método avaliado.

Capítulo 3

Coleção Web

A coleção Web utilizada nos experimentos deste trabalho consiste de parte do diretório Yahoo. A partir de URLs extraídas do referido diretório, foram coletadas as páginas HTML e as imagens associadas, utilizando o comando *wget* do Linux. As coletas foram realizadas em duas fases: a primeira coleta foi realizada em dezembro de 2005 e a segunda em março de 2006. A coleção foi cedida pelo Grupo de Tecnologia da Informação da Universidade Federal do Amazonas (UFAM). Neste trabalho, a referida coleção foi indexada utilizando a API Java da máquina de busca Lucene, alida a API HTMLParser.

Nas seções seguintes são apresentadas: informações quantitativas sobre a coleção analisada; o processo de indexação da coleção Yahoo; as evidências textuais indexadas; as consultas textuais avaliadas e o critério adotado para definir o conjunto de imagens relevantes.

3.1 Caracterização da Coleção Web

Alguns conceitos são importantes para o entendimento das informações coletadas. *Documento* é o arquivo resultante de uma requisição HTTP correta (exemplos: HTML, PDF, Doc). Uma *página* é um documento no formato *HTML*. Um *domínio* é qualquer nome da forma $x.y.z$, onde y é o domínio do primeiro nível regulamentado pelo Registro.br[1]. Um *Web site* ou *site* representa uma coleção de documentos referenciados por URLs que dividem o mesmo endereço de domínio. *Níveis* são contados por meio de estruturas de diretórios encontrada dentro dos servidores. O diretório raiz constitui o nível zero, os sub-diretórios do diretório raiz constituem o nível um e assim por diante.

[1] Órgão que regulamenta os domínios da Internet no Brasil (http://www.registro.br)

3.1.1 Tamanho

A coleção compactada tem cerca de 20 GB e descompactada contabiliza cerca de 39 GB. É formada por 1.478.698 arquivos, destes 254.495 correspondem a imagens grandes (foram eliminados banners e ícones) e 1.222.050 correspondem a documentos no formato HTML. Todas as imagens têm uma página associada, mas as páginas que não têm imagem não foram retiradas da coleção[2].

3.1.2 Tipos de arquivos

O levantamento sobre os tipos de arquivos tem por objetivo determinar quais arquivos presentes na coleção que correspondem a páginas Web e os quais referem-se a imagens. Foram verificadas as sequências de caracteres seguidos do operador ponto (.) de cada nome de arquivo contido na coleção. Dessa forma, não foi considerada a existência ou não de um link para o referido arquivo. Foram encontradas trinta e oito extensões distintas. As extensões encontradas e o número de arquivos na coleção foram apresentadas na Tabela 3.1. Os arquivos sem extensão ou que individualmente representam um percentual inferior a 0,0001% do número de arquivos na coleção estão representados pela entrada *Outros* da referida tabela:

Extensão do Arquivo	Percentual
HTML	50,46
HTM	32,18
JPG/JPeg	17,21
TXT	0,12
Outros	0,03

Tabela 3.1. Sumário dos tipos de arquivos da coleção web.

3.1.3 Documentos

Encontraram-se 1.223.829 arquivos com extensões do TXT, HTM ou HTML. Os arquivos em maior número seguem o formato HTML, seguido pelo txt. Não foram encontrados arquivos com as extensões pdf ou Microsoft Word.

Com relação aos documentos no formato HTML, foram encontrados arquivos com extensões htm e HTML. Em termos práticos, não existe diferença entre um documento com extensão HTM e HTML. Entretanto, o htm recebe esta extensão porque foi feito

[2]Informação fornecida pelo Grupo de Tecnologia da Informação da UFAM, que nos cedeu a coleção

em um ambiente antigo, como, por exemplo o DOS, que só permite três caracteres de extensão.

3.1.4 Hierarquia da coleção

A hierarquia da coleção refere-se à forma em que as páginas estão organizadas em termos de diretórios e subdiretórios. A partir da análise da hierarquia, pode-se contatar a profundidade que a coleta alcançou, dentre outras informações gerais que estão apresentadas na Tabela 3.2.

Páginas	Valor
Quantidade total	1.223.829
Tamanho médio (kB)	14,60
Web sites	Valor
Quantidade total	31.015
Profundidade máxima	21
Tamanho médio (mB)	1,14

Tabela 3.2. Análise da hierarquia da coleção web.

3.2 Indexação da Coleção Yahoo

3.2.1 Fontes de Evidências Textuais

As fontes de evidências textuais que serão consideradas, são as mesmas analisadas no trabalho Coelho et al. [2004]. Foram consideradas três possíveis fontes de evidência:

3.2.1.1 *Tags* de descrição

Composta de termos, isto é palavras, encontradas ou correspontes ao:

- Nome do arquivo da imagem;
- Conteúdo do atributo ALT da *tag* IMG;
- Conteúdo entre as *tags* $< A >$ e $< /A >$;

Esses termos são geralmente utilizados para descrever a imagem com a qual estão associados. O caminho completo do arquivo de imagem (nome absoluto), que inclui o host e o diretório das imagens, não foi utilizado, uma vez que geralmente é irrelevante.

Segue uma definição mais detalhada e exemplo de uso das *tags* mencionadas anteriormente:

- *Tag* A
  Exemplo de link para o site html.net:

  ```
  <a href="http://www.html.net/">Aqui um link para html.net</a>
  ```

  O elemento *a* refere-se a *tags* denominadas *tags* âncora (*anchor tag*). O atributo href é abreviação para referência a hypertexto (*hypertext reference*) e especifica o destino do link que normalmente é um endereço na Internet ou um arquivo.

  Observações:

  – links internos como cabeçalho com apontadores para os tópicos;

  – atributo title, uma breve descrição do link. Exemplo:

  ```
  <a href = "http://www.html.net/"
  title=" Visite HTML.net e aprenda HTM"'>HTML.net
  </a>'
  ```

- *Tag* IMG
  Para adicionar uma imagem, deve ser utilizada a *tag img*. Associoado a essa *tag* deve ser informado o endereço físico do arquivo que contém a imagem. Para isso, deve ser utilizado o atributo src, abreviatura para *source*. O conteúdo do atributo src pode ser o caminho no sistema de arquivos ou um link.

  Exemplo:

  ```
  <img src="tim.jpg" alt = "Tim Berners-Lee"/>
  ```

  Note que:

  – a *tag img* é do tipo comando isolado, isto é, uma só *tag* de abertura e fechamento;

  – "tim.jpg"é o nome do arquivo da imagem que você quer inserir na página;

  – ".jpg"é a extensão do tipo de imagem.

  FALAR DO ALT

3.2.1.2 *Tags* de Metadados

- Conteúdo entre as *tags* $< TITLE >$ e $< /TITLE >$;
- Metadado *Author*: relacionada com o autor do documento HTML;
- Metadados *Keywords*: relacionada com as palavras chaves associadas ao conteúdo do documento HTML;
- Metadado *Description*: relacionada com a descrição da página Web.

Esses termos são utilizados para descrever o conteúdo do documento. Quando os utilizamos para descrever uma imagem, assuminos que o tópico da imagem está relacionado com o do documento, ou seja, palavras que descrevem o conteúdo do documento também descrevem o conteúdo da imagem.

A seguir, apresenta-se uma definição mais detalhada e exemplo de uso das *tags* mencionadas anteriormente:

- *Tag* TITLE
  Para dar um título do documento, título este que aparece no topo da barra do navegador você deverá usar a seção "head". A *tag* para acrescentar um título é $< TITLE >$.

  O título não aparece na página propriamente dita. Tudo o que aparece na página é conteúdo e deverá ser colocado entre as *tags* "body".

- *Tags* de Metadados *Author*, *Keywords* e *Description*
  São etiquetas que servem para orientar um mecanismo automatizado que vai a uma página web e recolhe as informações contidas nessas etiquetas. Uma *tag description* completa tem a seguinte sintaxe:

  ```
  <meta name="description" content = "Aqui dentro vai a
  descrição da página ou do site">
  ```

  Outras *tags* são utilizadas na classificação das páginas. Por exemplo:

  – a *tag keywords* com as palavras-chave. Exemplo:

  ```
  <meta name = "keywords" content = "descrição da página,
  descrição do site, site">
  ```

  Os metadados ficam no cabeçalho página web. Dentro da *tag* $< HEAD >$, pela ordem, há o título da página ou do site e as *tags* de metadados. Exemplo:

```
<html>

<head>

<title>Metadados de descrição e palavra-chave<\title>

<meta name="description" content="Aqui vai a descrição do site.">

<meta name="keywords" content="descrição da página, site">

</head>

</html>
```

3.2.1.3 Passagens de Texto

As denominadas passagens de texto são compostas de palavras localizadas próximas às imagens. Espera-se que o texto ao redor de uma imagem seja relacionado com o conteúdo da imagem. Estes pequenos pedaços de texto são geralmente chamados de passagens. Quatro tipos de passagens de texto são consideradas neste trabalho:

- Passagem de texto de tamanho igual a 10 termos ao redor da imagem (5 termos antes da imagem e 5 depois);
- Passagem de texto de tamanho igual a 20 termos ao redor da imagem (10 termos antes da imagem e 10 depois);
- Passagem de texto de tamanho igual a 30 termos ao redor da imagem (20 termos antes da imagem e 20 depois);
- Texto Completo.

O texto completo é composto de todas as palavras na página Web. Aqui, assume-se novamente que o tópico do documento está relacionado com o tópico das imagens contidas no documento. Isso difere do uso de metadados, uma vez que provê o conjunto de palavras mais rico para descrever o tópico do documento. É importante notar que a abordagem com o texto completo pode ser vista como um caso especial da abordagem de passagem, onde consideramos passagens grandes o suficiente para conter todo o documento. Por esta razão, em Coelho et al. [2004] e no presente trabalho,

passagens textuais de 10, 20 e 40 termos não foram combinados entre si, nem com o texto completo.

Nos experimentos realizados em Coelho et al. [2004], foi encontrado que uma passagem formada por 40 termos (20 termos antes e 20 depois da imagem) geram os melhores resultados.

3.2.2 A Máquina de Busca

A máquina de busca utilizada neste trabalho para as tarefas de indexação da coleção e definição do conjunto-resposta das consultas avaliadas foi construída com o auxílio da API Lucene desenvolvida na linguagem Java. As seções seguintes apresentam os principais conceitos relacionados com a referida API.

3.2.2.1 Definições

Os conceitos fundamentais na Lucene são *índice*, *documento*, *campo* e *termo*:

- um índice contém uma sequência de documentos;
- um documento é uma sequência de campos;
- um campo é uma sequência determinada de termos;
- um termo é uma string.

A mesma string em dois campos diferentes é considerada como um termo diferente. Portanto, termos são representados como um par de strings, o primeiro denomina o campo e o segundo denomina o texto dentro do campo.

3.2.2.2 Índice Invertido

O índice armazena estatísticas sobre termos com o intuito de tornar a busca baseada em termo mais eficiente. O índice da Lucene se enquadra na família dos índices conhecidos como "índices invertidos". Isso porque ele pode listar, para um termo, os documentos que o contém. Este é o inverso da relação natural em que os documentos listam os termos.

3.2.2.3 O Método de Ponderação

O método de ponderação da Lucene utiliza uma combinação do *Modelo do Espaço Vetorial (VSM)*, da Recuperação de Informação e o *Modelo Booleano* para determinar quão relevante um documento é para uma consulta do usuário.

De forma geral, a idéia por trás do VSM é que quanto mais vezes um termo da consulta aparece em um documento, em relação ao número de vezes que o termo aparece em todos os documentos na coleção, mais relevante aquele documento é para a consulta.

A Lucene utiliza o Modelo Booleano para primeiro restringir (estreitar) os documentos que precisam ser pontuados com base na lógica booleana na especificação da consulta. Nesse mecanismo, a Lucene também adiciona algumas capacidades e refinamentos sobre este modelo para suportar busca booleana e fuzzy.

Na Lucene, os objetos que estamos pontuando são *documentos*. Um documento é uma coleção de campos. Cada campo tem uma semântica sobre como ele é criado e armazenado. Isto é importante porque dois documentos com o mesmo conteúdo, mas um tendo o conteúdo em dois campos e o outro em um campo apenas, retornarão uma pontuação diferente para a mesma consulta, devido à normalização do tamanho.

3.3 Consultas Textuais Avaliadas

Para definição das consultas textuais que seriam avaliadas e do conjunto de imagens relevantes associadas foi adotado o método *pooling*, comumente adotado nas coleções TRECs, para reunir as avaliações de relevância [Voorhees & Harman, 2000]. Neste método, um *pool* de possíveis documentos relevantes é criado ao obter uma amostra de documentos selecionados pelos diversos sistemas que serão avaliados. Este *pool* é então apresentado para usuários da funcionalidade de recuperação de informação, que realizam julgamento binário (sim/não) quanto à relevância de cada documento do *pool*. Os documentos não avaliados são assumidos como não-relevantes. O método particular de amostragem utilizado na TREC consiste em selecionar os 100 documentos do topo, recuperados por cada método que será avaliado para uma dada consulta, seguida pela junção de todos os documentos recuperados em um pool para avaliação. Esta é uma técnica de amostragem válida, uma vez que todos os sistemas utilizam métodos de rank para recuperação, sendo retornados os documentos mais relevantes nas primeiras posições. Cada *pool* é ordenado pelo identificador do documento, assim os usuários não conseguem dizer se um documento tem um valor de rank elevado para algum sistema avaliado ou quantos sistemas (ou quais sistemas) recuperaram aquele dado documento.

3.3.1 Consultas Consideradas

Com base na metodologia anteriormente apresentada, foram avaliadas 100 consultas. Parte destas consistem em consultas populares retiradas do *log* da máquina de busca

TodoBr e que se encontram disponíveis na coleção WBR99. Outra parte das consultas consiste naquelas utilizadas no trabalho Coelho et al. [2004]. Cada uma das 100 consultas foi submetida à máquina de busca construída a partir da Lucene utilizando separadamente cada tipo de evidência textual indexada. Cada conjunto-resposta retornado pelas diversas evidências textuais foram posteriormente unidos, gerando assim o conjunto-resposta final. Como critério para a seleção das consultas mais expressivas desse grupo foi considerado o número de imagens-resposta retornadas. Assim, das 100 imagens submetidas ao método de *pooling*, foram selecionadas as 35 com maior conjunto-resposta, em número de imagens constituintes, para posterior definição do conjunto de imagens relevantes. Após a geração do conjunto-resposta de cada consulta, foi verificado que o número médio de imagens que compõem este conjunto é 146,86.

As consultas consideradas estão apresentadas na tabela 3.3:

1. por do sol	2. igreja
3. coca cola	4. Corcovado
5. Serra da Canastra	6. Linux
7. Jesus	8. Pirenópolis
9. Fernando de Noronha	10. fotos carnaval
11. tubarão	12. alimentos transgênicos
13. animais	14. backstreet boys
15. cartões de natal	16. cartões virtuais
17. filmes	18. gifs animados
19. hotel fazenda	20. índios
21. jornal	22. machado de assis
23. maconha	24. mp3
25. municípios	26. origem da vida
27. papel de parede	28. piadas
29. poesias	30. pokemon
31. poluição	32. receitas
33. turismo	34. vestibular
35. vírus	

Tabela 3.3. *Consultas Consideradas.*

3.3.2 Definição das imagens relevantes

Avaliações de relevância são de importância crítica para coleções de teste. Nesse sentido, para a avaliação da eficácia de um dado sistema é necessário que, para cada tópico ou consulta, seja gerado uma lista dos documentos relevantes.

Neste trabalho em específico, o julgamento de relevância do conjunto-resposta formado para cada uma das 35 consultas consideradas foi realizado por dois voluntários, seguindo a definição de relevância ou não-relevância, independentemente da forma como as imagens foram recuperadas. Nos casos de discordância quanto ao julgamento de relevância, foram realizadas novas avaliações em conjunto para determinar a relevância ou a não-relevância. Após a conclusão deste julgamento de relevância das imagens, foi verificado que o número médio de imagens relevantes por pool é de 35,7.

Capítulo 4

Modelos de Recuperação de Imagem

4.1 Programação Genética

Conforme apresentado no capítulo 2, a Programação Genética é baseada na idéia de melhoramento contínuo da qualidade da solução para um problema-alvo, a partir de soluções anteriores. Para o emprego dessa metodologia, inicialmente é necessário definir claramente qual é o objetivo da solução computacional que se deseja modelar, a partir das fontes de informação existentes para solução do problema-alvo. Neste último caso, conforme apresentado no capítulo 3 as fontes de informação são três tipos de evidências textuais extraídas dos documentos HTML onde as imagens estão inseridas: metadados, *tags* de descrição e as passagens textuais (essas comportam as quantias de 10, 20 e 40 termos ao redor da imagem e o texto completo). Dessa forma, o objetivo do arcabouço evolucionário pode ser definido como: definir uma função que combine as evidências textuais de tal forma que seja utilizada para determinar um conjunto-resposta ordenado (*rank*) mais preciso quanto à relevância da imagem para uma dada consulta.

Para atingir tal objetivo, utilizamos a biblioteca lilgp [Zongker & Punch, 1996]. Foram separados três subconjuntos: um para treino, um para validação e outro para teste. Na modelagem do arcabouço a partir desta biblioteca, definiu-se o seguinte conjunto de funções e terminais:

- Funções: os operadores aritméticos de soma, multiplicação, produto e logaritmo;
- Terminais: O valor de tf-idf de cada evidência textual e alguns valores constantes. O valor de tf-idf consiste na multiplicação de duas medidas calculadas de forma

independente: a frequência do termo na evidência textual (tf) e o inverso da frequência do termo na evidência textual (idf). No caso de consultas com mais de um termo, o valor do terminal consiste na soma aritmética dos valores tf-idf de cada termo da consulta existente na evidência textual correspondente, que está associada à imagem. A frequência do termo em uma dada evidência textual é simplesmente o número de vezes que um dado termo aparece na evidência textual considerada. Essa contagem geralmente é normalizada para previnir a priorização de evidências textuais maiores, que podem ter uma maior frequência do termo em detrimento da real importância do termo na evidência textual, para prover uma medida da importância do termo t_i dentro da evidência d_j, ou seja:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{n,k}}$$

onde $n_{i,j}$ é o número de ocorrências do termo considerado na evidência textual d_j e o denominador é a soma do número de ocorrências de todos os termos na evidência textual d_j.

O idf (*inverse document frequency*) é a medida da importância geral do termo, obtida a partir da divisão do número total de todas as imagens da coleção pelo número de evidências textuais daquele tipo que contêm o termo e então obtendo o logaritmo daquele quociente. Ou seja:

$$idf = \log \frac{|D|}{|d_j : t_i \in d_j|}$$

com:

- $|D|$: número total de imagens na coleção;
- $|d_j : t_i \in d_j|$: número de imagens onde o termo t_i aparece (ou seja, $n_{i,j} \neq 0$

Para viabilizar o processo de evolução da população, foi necessário ainda definir uma métrica de cálculo da função de aptidão de um dado indivíduo. Em função dos bons resultados obtidos em [de Almeida et al., 2007], foi utilizada a métrica MAP. Essa informação é de extrema importância, pois ela avalia o quão bom um indivíduo é na população.

As configurações relacionadas com os operadores genéticos a serem utilizados, o número de indivíduos da população inicial e o número de gerações que o arcabouço contruído irá evoluir serão avaliadas no capítulo 5. A partir da avaliação das possíveis combinações, será determinado o cenário mais pertinente para a solução do Problema de Recuperação de Informação.

4.2 O Modelo de Rede de Crenças

Esse modelo foi originalmente apresentado em [Coelho, 2001]. Ele combina as mesmas evidências textuais empregadas no presente trabalho como fonte de informação sobre imagens da Web: *tags* de descrição, metadados, texto completo e passagens textuais com 10, 20 e 40 termos. Do Modelo Bayesiano apresentado em Coelho et al. [2004] podem ser extraídas diferentes equações que combinam múltiplas evidências textuais, viabilizando a identificação das evidências textuais mais adequadas para relacionar com o conteúdo da imagem. Esta é uma característica similar ao arcabouço evolucionário, pois este tem como resultado de sua execução uma função que combina as diversas fontes de evidências. Além disso, o Modelo Bayesiamo serve como uma ferramenta teórica e formal de avaliação e estudo tanto das evidências em isolado quanto de forma de combiná-las. Esses fatos motivaram a seleção do Modelo Bayesiano como baseline.

O Modelo de Rede de Crenças é caracterizado por interpretar probabilidades como graus de "crença", sem que experimentações sejam realizadas (escola de pensamento da probabilidade conhecida como visão epistemológica). Esse modelo adota o arcabouço das Redes Bayesianas, que é bastante útil por fornecer um formalismo gráfico que modela as dependências entre as variáveis de uma distribuição de probabilidades.

As Redes Bayesianas são representadas por grafos acíclicos direcionais, nos quais nós representam variáveis aleatórias e arestas indicam o relacionamento existente entre essas variáveis. As probabilidades condicionais associadas aos pares de nós fornecem a "intensidade" de cada relacionamento. Os nós-pai de um nó (chamado nó-filho) são a causa direta direta dele, ou seja, se um nó A é nó-pai de um nó B, então a probabilidade do nó B depende da probabilidade de A. Essa dependência é representada por uma seta direcional de A para B. Pode-se dizer ainda que A tem influência direta em B. Os nós-raiz não são influenciados por nenhum outro nó da rede e, por isso, suas probabilidades possuem valor independentemente dos valores de probabilidades associados aos demais nós da rede.

De acordo com esse modelo, os termos de indexação extraídos dos documentos da coleção formam o universo de discurso, denotado por U (espaço conceitual). Um conceito u é um subconjunto de U e pode representar um documento na coleção ou uma consulta do usuário. Os nós K_i modelam os termos na coleção e a eles estão associadas variáveis binárias também chamadas K_i. Se essa variável possui valor 1, então o termo K_i correspondente é um membro do conceito u. Essas variáveis apenas assumem valores booleanos porque eles são expressivos o bastante em termos semânticos e não envolvem complexidade de cálculo. Termos de indexação são considerados independentes entre si.

Documentos e consultas são modelados de forma idêntica, como nós da rede, aos quais estão associados termos da coleção. Assim, um documento d_j pode ser representado por um conceito $d_j = k_1, k_2, ..., k_N$. O mesmo ocorre com uma consulta q, modelada na rede pelo nó rotulado Q. Se um termo k_i é utilizado para descrever um documento d_j, então o nó que representa esse termo está ligado ao nó que representa aquele documento e, portanto, o termo k_i está ativo nesse documento. O mesmo ocorre com uma consulta q: o nó que a representa é apontado pelos nós dos termos de indexação que compõem o conceito associado a essa consulta.

A figura 4.1 modela as dependências entre as variáveis de acordo com o modelo apresentado. N é o número de documentos na coleção e t é o número total de termos distintos na coleção.

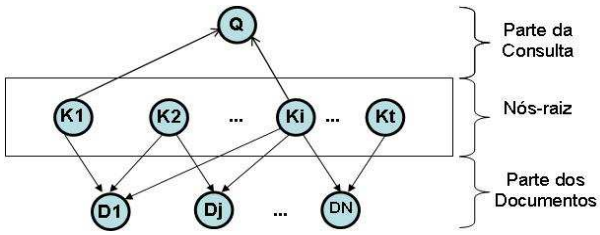

Figura 4.1. Esquema da Redes de Crença para uma consulta q.

Com base na explanação anterior, a similaridade de um documento d_j com relação à consulta q pode ser calculada pela probabilidade $P(d_j|k)$:

$$P(d_j|k) = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

onde:

- k é o estado em que figuram exatamente as palavras-chaves ativas que estão na consulta q;
- w_{ij} é o peso do termo k_i na evidência textual considerada, associada a uma imagem presente no documento j, ou seja, Ij;
- w_{iq} é o peso do termo i na consulta do usuário.

Esses pesos são definidos como:

$$w_{ij} = (1 + \ln f_{ij})$$

$$w_{iq} = \ln(1 + \frac{N}{n_i})$$

onde:

- f_{ij} é a freqência do termo k_i na evidência textual considerada da imagem I_j;
- N é o número total de imagens;
- n_i é o número de evidências textuais, do tipo considerado, que contêm o termo k_i.

Essa equação pode ser utilizada para o cálculo da similaridade de uma imagem com relação a uma consulta para as diversas evidências textuais consideradas. Em cada caso, os pesos dos termos são calculados de acordo com sua ocorrência na respectiva fonte de evidência.

Para representar a probabilidade de cada possível fonte de evidência ser observada, com relação à consulta q, será adotada a seguinte notação:

- RD_{jq}: *tags* de descrição;
- RM_{jq}: metadados;
- $RPjq$: passagens textuais; consistem em uma das passagens de 10, 20 ou de 40 termos ou o texto completo.

Sendo k o estado das variáveis K_i em que todas as variáveis que estão ligadas correspondem àquelas correspondentes à consulta q e η uma constante de normalização [Coelho, 2001], a probabilidade $P(i_j|q)$ da imagem ser observada em uma combinação dessas fontes de evidência pode ser calculada a partir da seguinte relação:

$$P(i_j|q) = \eta * [1 - (1 - RD_{jq}) * (1 - RM_{jq}) * (1 - RP_{jq})]$$

Para testar as distintas combinações de evidências, podemos setar para zero o fator associado à evidência que queremos desconsiderar. Por exemplo, para avaliar as *tags* de descrição isoladamente, setamos $RM_{jq} = 0$ e $RP_{jq} = 0$. Substituindo os valores na equação anterior, obtemos:

$$P(i_j|q) = \eta * RD_{jq}$$

De forma análoga, $P(i_j|q)$ pode ser calculada para as outras possíveis evidências ou combinação de evidências. A tabela 4.1 mostra as várias equações de recuperação de imagem que podem ser geradas.

Abordagem	$P(i_j\|q)$
Tags de descrição	$\eta * RD_{jq}$
Metadados	$\eta * RM_{jq}$
Passagens/Texto Completo	$\eta * RP_{jq}$
Descrição + Metadados	$\eta * [1 - (1 - RD_{jq}) * (1 - RM_{jq})]$
Descrição + Passagens/Texto Completo	$\eta * [1 - (1 - RD_{jq}) * (1RP_{jq})]$
Passagens/Texto Completo + Metadados	$\eta * [1 - 1 - RP_{jq}) * (1 - RM_{jq})]$
Descrição + Metadados + Passagens	$\eta * [1 - (1 - RD_{jq}) * (1 - RM_{jq}) * (1 - RP_{jq})]$

Tabela 4.1. *Equações de similaridades obtidas do Modelo Bayesiano.*

Capítulo 5

Resultados Experimentais

Os experimentos realizados tiveram por objetivo avaliar a eficácia dos métodos de recuperação de imagens da Web: o arcabouço evolucionário e o Modelo Bayesiano. Para isso, foi utilizada a coleção Web descrita no capítulo 3. A tabela 5.1 apresenta um resumo de algumas estatísticas levantadas sobre a referida coleção:

Tamanho da coleção (GB)	20
Número de páginas	1.223.829
Número de Imagens Distintas	254.495
Número de Consultas	35
Número de Imagens Por Pool de Consulta	146,86
Número de Imagens Relevantes Por Pool de Consulta	35,7

Tabela 5.1. *Características da coleção Web usada nos experimentos.*

5.1 Metodologia

Os experimentos iniciais contemplaram o estudo do arcabouço evolucionário, no que se refere à sua configuração em termos de número de indivíduos e de gerações necessários para se obter um resultado significativamente melhor que aquele alcançado com a abordagem bayesiana. Os resultados desta análise estão apresentados na seção 5.2. Em seguida, foram realizados experimentos com as distintas equações de recuperação de imagem do Modelo Bayesiano, apresentadas na tabela 4.1. Os resultados obtidos estão apresentados na seção 5.3, onde também foram comparados com aqueles obtidos em [Coelho et al., 2004] com uma coleção Web distinta. Por fim, na seção 5.4 estão apresentados os experimentos comparativos do Arcabouço Evolucionário e das equações

de recuperação de imagem do Modelo Bayesiano que apresentaram melhores resultados em termos das métricas adotadas.

Nos experimentos apresentados nas seções 5.3 e 5.4, foi adotado o método de avaliação denominado *K-fold cross validation*. Segundo este método, os dados de entrada do problema em análise são divididos em k subconjuntos e o método para solução é repetido k vezes. Para cada vez, um dos k subconjuntos é utilizado como o conjunto de teste e os outros k-1 subconjuntos são utilizados no conjunto de treinamento. Cada um dos dados de entrada estará presente em um conjunto de teste exatamente uma vez e em um conjunto de treinamento k-1 vezes. A variância do resultado é reduzida tanto quanto k é aumentado. A desvantagem deste método é que o algoritmo de treinamento tem que ser executado novamente k vezes, o que siginifica que leva um tempo computacional k vezes maior para avaliar o método em questão. Um estudo mais detalhado sobre as diversas abordagens do *cross validation*, bem como a razão e a forma de serem aplicadas, podem ser obtidas em [Duda et al., 2000].

A partir desta descrição do *K-fold cross validation*, realizamos experimentos com o valor de k igual a 5. Dessa forma, as trinta e cinco consultas avaliadas, apresentadas na seção 3.3, foram divididas em cinco grupos de sete consultas cada. Para cada uma das cinco execuções dos métodos de recuperação de imagens da Web, foi adotado um conjunto de consultas distinto. Ainda, para o arcabouço evolucionário que possui um caráter de aprendizado incorporada à Programação Genética, para cada execução distinta do método, dos quatro grupos restantes de consultas, três foram tomados para a fase de treinamento do método. O quarto grupo de consultas foi utilizado para uma fase de validação do algoritmo, com o intuito de evitar uma possível especialização dos indivíduos formados para o grupo de consultas utilizado na fase de treinamento. Este problema tratado é comumente conhecido na literatura relacionada como *overfitting*.

5.2 Experimentos com o Arcabouço Evolucionário

Nos experimentos com o arcabouço evolucionário foi adotada a seguinte configuração: população inicial gerada a partir do método *half-and-half*[1], profundidade máxima do indivíduo igual a 4, critério de parada igual ao número de gerações, seleção por torneio[2] com taxa de crossover de 90% e taxas de mutação e de reprodução iguais a 5%.

[1] A população inicial é gerada de forma que 50% dos indivíduos são árvores cujos terminais estão no mesmo nível e os demais 50% são gerados a partir de um nodo raiz, que é chamado recursivamente para gerar árvores filhas, com uma profundidade máxima.

[2] Os indivíduos são escolhidos de uma forma aleatória, segundo uma distribuição uniforme. Em seguida, o melhor dentre os n indivíduos, tendo como critério o valor da função de aptidão, é selecionado. O valor de n refere-se ao tamanho do torneio.

O primeiro parâmetro adotado para análise do comportamento do arcabouço evolucionário foi o tamanho da população. Para isso, o arcabouço foi executado uma vez para cada tamanho de população, utilizando variados tamanhos de população e mantendo fixos todas as demais variáveis do problema. Experimentos iniciais mostraram que para os tamanhos de população avaliados, ocorre uma estabilização da curva de evolução do método evolucionário a partir de quinze gerações. Por esse motivo, cada execução do método evolucionário de recuperação de imagem foi analisado após as quinze iterações. Além disso, foi adotada a proporção de 14, 7 e 14 consultas, respectivamente, para as fases de treinamento, validação e teste. A formação de cada conjunto de imagens foi realizada de forma aleatória, dentre o conjunto de trinta e cinco consultas avaliadas.

Na tabela 5.2 são apresentados os resultados da precisão em níveis fixos de resultados recuperados, da média desta última métrica, o R-precision e o MAP, para cada tamanho de população considerado. As informações na coluna Nível indica o valor da precisão alcançada para no N primeiros documentos retornados. Por exemplo, os valores contidos na linha da tabela 5.2 relativa ao nível *Em 5 docs* indicam os valores de precisão obtidos para os cinco primeiros documentos, considerando cada um dos tamanhos de população avaliados.

Nível	Número de Indivíduos da População							
	4 ind	8 ind	12ind	16 ind	20 ind	30 ind	40 ind	50 ind
Em 5 docs	33,846	33,846	18,461	44,615	23,076	21,538	29,230	33,846
Em 10 docs	26,923	33,846	10,769	30,000	13,846	16,153	19,230	26,153
Em 15 docs	20,512	20,512	8,717	22,051	16,410	12,820	16,923	21,025
Em 20 docs	17,307	17,307	7,307	20,000	17,692	11,923	16,153	18,076
Em 30 docs	14,102	14,102	6,667	17,179	16,410	14,359	18,205	17,692
Em 100 docs	6,615	6,615	6,538	6,538	7,692	12,462	13,230	12,384
Em 200 docs	7,242	7,242	6,588	4,588	5,857	8,319	8,818	8,780
Em 500 docs	3,548	3,548	6,194	3,594	3,594	4,041	4,040	4,040
Em 1000 docs	2,137	2,137	3,337	2,144	2,129	3,467	2,198	2,198
Média	14,692	15,462	8,287	**16,746**	11,857	11,676	**14,226**	**16,022**
R-precision	25,473	25,473	16,661	**28,463**	21,105	27,526	**33,706**	**32,627**
MAP	26,443	26,443	19,278	**30,469**	21,295	27,468	**30,168**	**30,274**

Tabela 5.2. *Resultados obtidos para diferentes tamanhos de população, em 15 gerações, do arcabouço evolucionário.*

Analisando inicialmente os valores obtidos da precisão em níveis fixos de resultados, para níveis de confiança iguais ou superiores a 10%, pode-se afirmar que todos os tamanhos de população são estatisticamente iguais. Já a partir da verificação dos valores obtidos pelas demais métricas, foi identificado que as populações com 16, 40 e 50

indivíduos apresentaram resultados mais expressivos, dentre as demais configurações do grupo. Para resolver esse impasse e determinar qual a configuração em termos do tamanho da população deveria ser adotada, foram traçadas as curvas de Precisão *versus* Revocação para 16, 40 e 50 indivíduos em 15 gerações. A figura 5.1 apresenta as curvas obtidas.

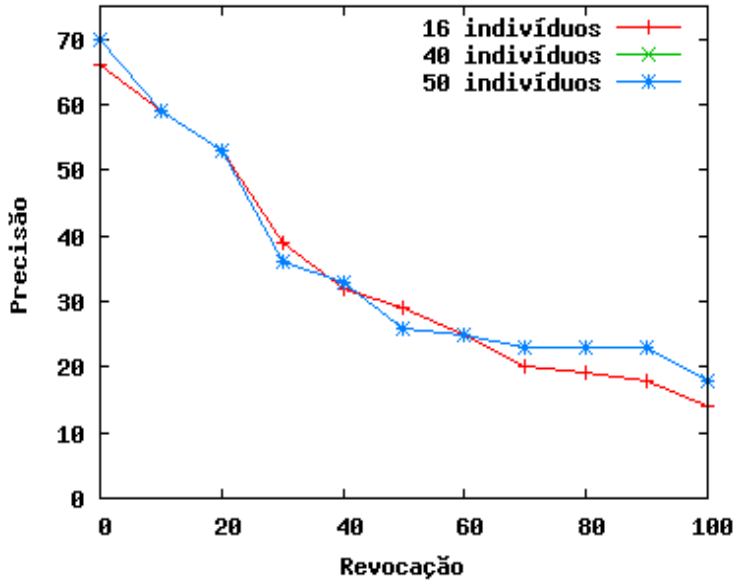

Figura 5.1. Curvas de Precisão versus Revocação para populações com 16, 40 e 50 indivíduos, em 15 gerações, do arcabouço evolucionário.

A partir das curvas geradas, pode-se constatar que os melhores valores de precisão, para cada nível de revocação, foi obtido com as populações de 40 e 50 indivíduos. Analisando visualmente, as curvas com 40 e 50 indivíduos se sobrepõem. Para definir qual das duas populações seria a escolhida, analisamos novamente os valores das métricas da tabela 5.2. Para as métricas relativas a Média e ao MAP a população de 50 indivíduos apresentou melhores resultados que a população de 40 indivíduos, tornando portanto esse tamanho de população utilizado nos experimentos comparativos com o Modelo Bayesiano.

O próximo parâmetro estudado foi o número de gerações necessárias para se obter melhores resultados das métricas avaliadas, quando utilizada uma população igual a 50 indivíduos. Para isso, o arcabouço evolucionário foi executado para diferentes números de gerações, mantendo constante todos os demais parâmetros do método evolucionário. A partir de cada execução realizada, foi calculado o valor de MAP em cada uma das fases: treinamento, validação e teste. As curvas de evolução obtidas para 50 indivíduos em 30 gerações estão apresentadas na figura 5.2.

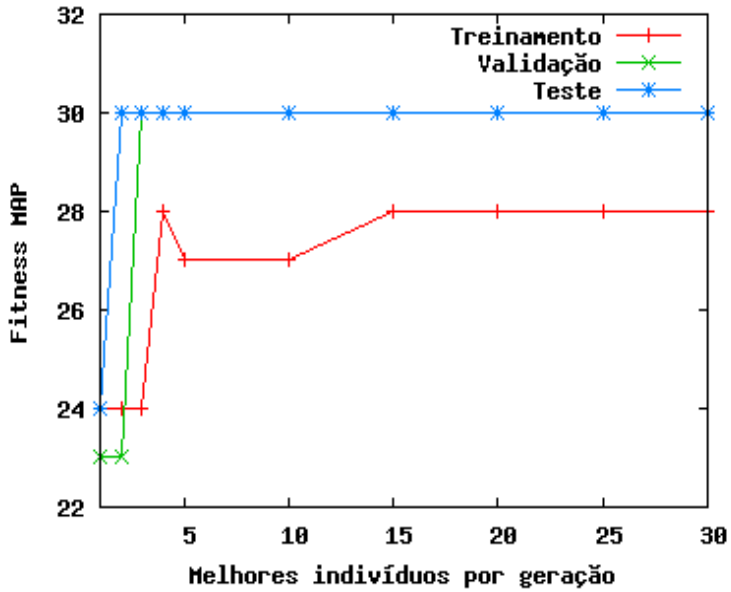

Figura 5.2. Curvas de evolução do arcabouço evolucionário.

Analisando os resultados obtidos constata-se que a estabilização da curva para 15 gerações foi alcaçada em todas as fases.

5.3 Experimentos com o Modelo Bayesiano

5.3.1 Texto Completo *versus* Passagens Textuais

Os primeiros experimentos com o Modelo Bayesiano consistiram em determinar o melhor tamanho de passagens ao redor das imagens. Para cada grupo de consultas consideradas para teste, foram testadas passagens de 10, 20 e 40 termos e o texto completo. A figura 5.3 apresenta os valores de precisão média para cada nível de revocação.

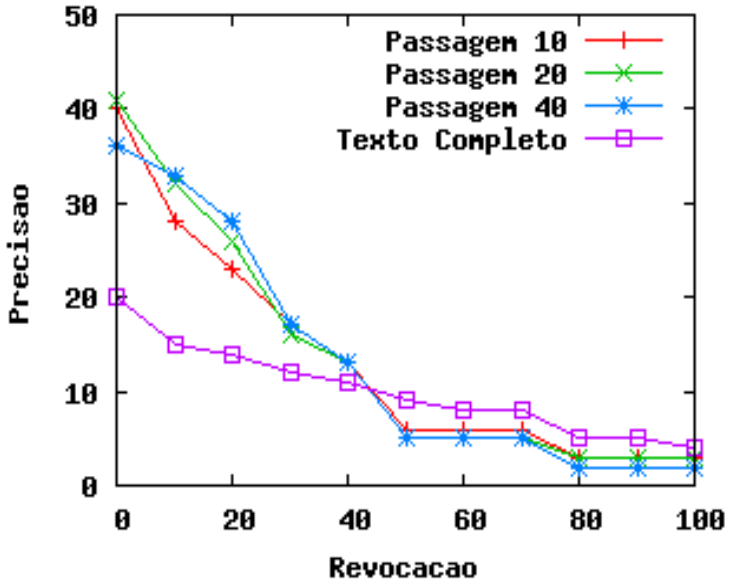

Figura 5.3. Curvas de precisão versus revocação das passagens textuais.

A tabela 5.3 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP. As denominações 10 T, 20 T, 40 T e TC referem-se, respectivamente, às passagens textuais de 10, 20, 40 termos e ao texto completo.

Revocação	Abordagem			
	10 T	20 T	40 T	TC
0	40,590(15,407)	41,777(18,070)	36,460(18,311)	20,876(18,726)
10	28,326(22,111)	32,641(16,819)	33,989(18,423)	15,216(10,677)
20	23,564(21,281)	26,837(15,850)	28,821(17,949)	14,963(10,967)
30	17,117(17,632)	16,501(17,484)	17,020(18,595)	12,773(9,591)
40	13,974(15,742)	13,635(15,807)	13,430(15,563)	11,979(9,292)
50	6,260(8,420)	5,821(7,742)	5,675(7,541)	9,786(6,673)
60	6,2602(8,420)	5,805(7,722)	5,675(7,541)	8,894(6,199)
70	6,254(8,425)	5,805(7,722)	5,675(7,541)	8,751(6,276)
80	3,450(7,473)	3,050(6,534)	2,966(6,345)	5,368(3,404)
90	3,363(7,520)	3,039(6,539)	2,953(6,351)	5,015(3,343)
100	3,363(7,520)	3,039(6,539)	2,953(6,351)	4,564(3,193)
MAP	11,272(9,813)	**11,823(8,836)**	11,810(8,810)	8,639(6,586)

Tabela 5.3. *Precisão média para cada nível de revocação e MAP das passagens textuais.*

Observa-se que as passagens textuais de 10, 20 e 40 termos apresentam comportamento bastante similar, com um desempenho bastante inferior a estas identificado para o texto completo. Em função do maior valor média de precisão para o primeiro nível de revocação e de MAP, utilizamos a passagem textual de 20 termos na comparação com os demais resultados obtidos pelas outras evidências isoladas.

5.3.2 Evidências Isoladas

O próximo experimento contemplou a comparação dos resultados obtidos pelas passagens textuais de 20 termos com aqueles alcançados pelas demais evidências isoladas: *tags* de descrição e metadados. A figura 5.4 apresenta os valores de precisão média para cada nível de revocação.

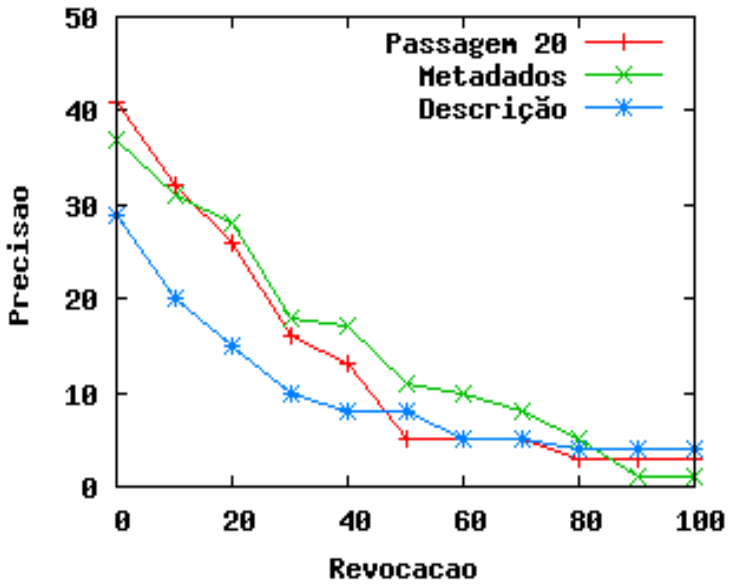

Figura 5.4. Curvas de Precisão versus Revocação das evidências isoladas.

A tabela 5.4 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP. As denominações META e DESC referem-se, respectivamente, às *tags* de metadados e de descrição.

Revocação	Abordagem		
	META	DESC	20 T
0	37,097(12,788)	29,475(15,545)	41,777(18,070)
10	31,745(15,569)	20,455(10,599)	32,641(16,819)
20	28,024(13,500)	15,141(10,579)	26,837(15,850)
30	18,861(12,715)	10,356(5,775)	16,501(17,484)
40	17,406(11,770)	8,493(6,449)	13,635(7,742)
50	11,386(9,823)	8,425(6,378)	5,821(7,742)
60	10,439(9,060)	5,671(6,476)	5,805(7,722)
70	8,620(6,494)	5,0219(5,906)	5,805(7,722)
80	5,167(5,729)	4,029(5,679)	3,050(6,534)
90	1,270(2,625)	4,029(5,679)	3,039(6,539)
100	1,270(2,625)	4,029(5,679)	3,039(6,539)
MAP	**13,091(7,067)**	8,071(5,295)	11,823(8,836)

Tabela 5.4. *Precisão média para cada nível de revocação e MAP das das evidências isoladas.*

Comparando os valores da precisão média apresentados pelas curvas da figura 5.4, constata-se que para um nível de revocação inferior a 15%, as passagens textuais de 20 termos apresentaram melhores valores de precisão média. A partir desse nível de revocação até o valor de 82%, os melhores resultados já são apresentados pelas *tags* de metadados. Para níveis superiores a 82%, os maiores valores de precisão média são obtidos pelas *tags* de descrição. Ao analisar os valores de MAP, ver tabela 5.4, as *tags* de metadados apresentam um valor superior em comparação com as demais fontes de evidência isolada. Por esse motivo, pode-se concluir que dentre as fontes de evidência isoladas as *tags* de metadados são aquelas que apresentam melhor resultado.

5.3.3 Múltiplas Fontes de Evidência

Examinando agora os resultados obtidos quando combinadas múltiplas fontes de evidência no Modelo Bayesiano, inicialmente serão analisados os resultados obtidos com as combinações que envolvem as *tags* descrição e uma outra fonte de evidência textual: *tags* de descrição + metadados e *tags* de descrição + passagens textuais. São apresentados na figura 5.5 os valores de precisão média para cada nível de revocação. A tabela 5.5 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP. O sinal indicativo de adição representa a combinação das evidências textuais envolvidas, assim *DESC + META* representa a combinação das *tags* de descrição com as de metadados.

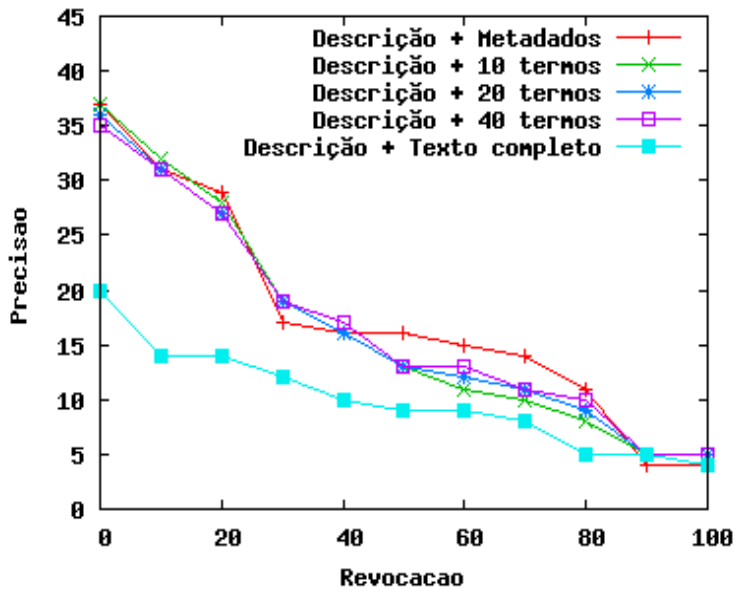

Figura 5.5. Curvas de Precisão versus Revocação das tags de descrição.

Revocação	Abordagem				
	DESC + META	DESC + 10 T	DESC + 20 T	DESC + 40 T	DESC + TC
0	37,145(14,423)	37,513(15,431)	36,368(14,749)	35,086(16,358)	20,489(19,013)
10	31,591(10,634)	32,221(17,273)	31,850(16,426)	31,064(16,210)	14,996(11,263)
20	29,449(10,843)	28,968(20,542)	27,845(20,079)	27,014(19,884)	14,769(11,391)
30	17,443(5,627)	19,829(15,896)	19,987(15,454)	19,592(14,900)	12,294(9,216)
40	16,580(4,709)	16,734(15,074)	16,997(14,711)	17,288(15,515)	10,361(6,127)
50	16,371(4,547)	13,876(9,770)	13,747(9,352)	13,607(9,216)	9,666(5,898)
60	15,544(4,340)	11,873(7,554)	12,108(7,469)	13,250(9,249)	9,120(6,050)
70	14,899(3,788)	10,540(7,147)	11,517(6,978)	11,283(6,882)	8,694(6,290)
80	11,643(4,796)	8,0786(6,187)	9,214(5,568)	10,422(6,725)	5,489(2,693)
90	4,652(5,857)	5,314(6,224)	5,410(6,378)	5,440(6,439)	5,000(2,893)
100	4,636(5,852)	5,314(6,224)	5,410(6,378)	5,440(6,439)	4,504(2,858)
MAP	**14,898(4,501)**	14,071(9,653)	14,214(8,841)	14,151(8,960)	8,415(6,329)

Tabela 5.5. *Precisão média para cada nível de revocação e MAP das tags de descrição.*

Em decorrência da baixa eficácia das *tags* de descrição e do texto completo, quando tomados isoladamente, a combinação dessas duas fontes de evidência resultou nos piores resultados. Utilizando esse mesmo critério, as combinações que envolvem as *tags* de descrição e as demais passagens textuais apresentam comportamento bastante similar. O desempenho da combinação das *tags* de descrição com as *tags* de metadados, apesar de bastante similar àquele obtido pelas passagens textuais de 10, 20 e 40 termos, apresenta melhores valores de precisão média entre 42% e 80% de revocação. Estendendo a comparação para os valores médios de MAP, esta última combinação confirma o melhor desempenho.

O próximo grupo de combinações de múltiplas fontes de evidência contempla os resultados obtidos com as combinações que envolvem as *tags* de metadados e uma outra fonte de evidência textual: metadados + *tags* de descrição e metadados + passagens textuais. São apresentados na figura 5.6 os valores de precisão média para cada nível de revocação.

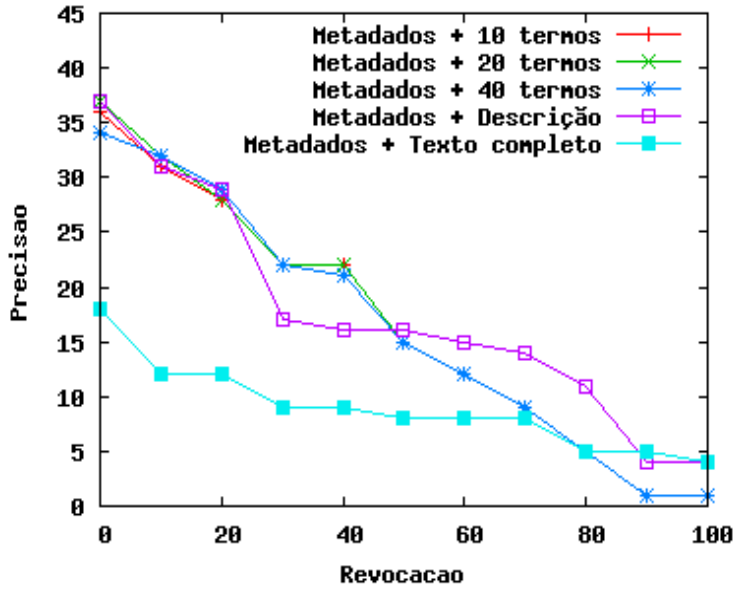

Figura 5.6. Curvas de Precisão versus Revocação dos metadados.

A tabela 5.6 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP.

Revocação	Abordagem				
	META + DESC	META + 10 T	META + 20 T	META + 40 T	META + TC
0	37,145(14,423)	36,683(10,745)	37,155(22,148)	34,147(16,636)	18,632(20,079)
10	31,591(10,634)	31,153(14,441)	32,069(30,459)	32,116(15,281)	12,877(12,014)
20	29,449(10,843)	28,846(13,717)	28,843(27,430)	29,805(14,786)	12,337(12,012)
30	17,443(5,627)	22,410(15,257)	22,407(30,515)	22,283(15,102)	9,920(11,020)
40	16,580(4,709)	22,410(15,257)	22,407(30,515)	21,934(14,673)	9,920(11,020)
50	16,371(4,547)	15,064(13,851)	15,187(28,095)	15,157(14,006)	8,609(8,321)
60	15,544(4,340)	12,941(11,816)	12,938(23,632)	12,937(11,818)	8,563(8,370)
70	14,899(3,788)	9,960(9,204)	9,957(18,405)	9,956(9,204)	8,200(7,938)
80	11,643(4,796)	5,124(5,897)	5,120(11,786)	5,120(5,893)	5,344(4,032)
90	4,652(5,857)	1,299(2,692)	1,295(5,368)	1,295(2,684)	5,228(4,070)
100	4,636(5,852)	1,299(2,692)	1,295(5,368)	1,295(2,684)	4,500(3,471)
MAP	**14,898(4,501)**	14,498(8,143)	14,618(8,355)	14,558(8,495)	7,527(7,491)

Tabela 5.6. *Precisão média para cada nível de revocação e MAP dos metadados.*

Assim como foi observado nos resultados das *tags* de descrição e suas combinações, em decorrência da baixa eficácia do texto completo, a combinação resultante desta evidência textual com as tags de metadados resultou nos piores resultados. As demais combinações de passagens textuais e das tags de metadados apresentaram resultados similares. Comparando os valores de MAP, a combinação com as tags de descrição foi aquela que apresentou maior valor.

Para finalizar a análise dos resultados obtidos com as diversas equações do Modelo Bayesiano, são apresentados na figura 5.7 os valores de precisão média para cada nível de revocação, das múltiplas evidências textuais consideradas.

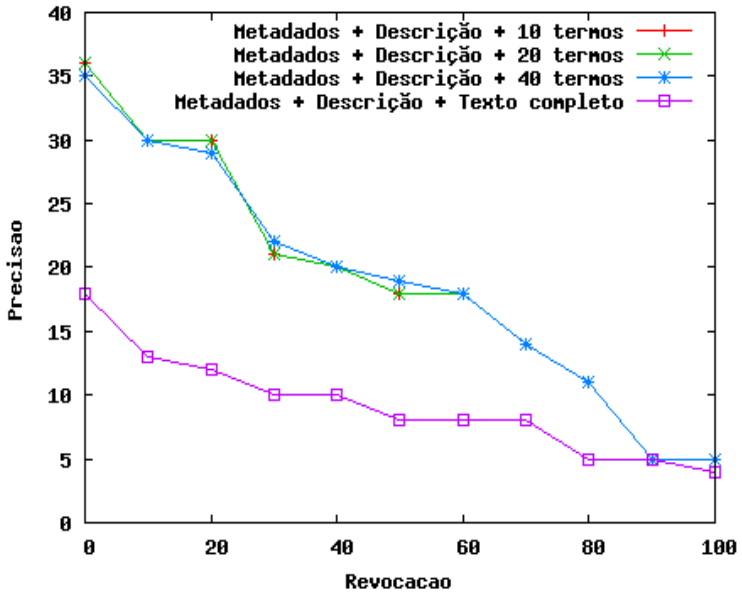

Figura 5.7. Curvas de Precisão versus Revocação da combinação de todas as evidências textuais.

A tabela 5.7 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP.

Revocação	Abordagem			
	META+DESC+10T	META+DESC+20T	META+DESC+40T	META+DESC+TC
0	36,143(14,919)	36,143(14,927)	35,838(15,450)	18,302(19,763)
10	30,812(11,527)	30,812(11,536)	30,507(12,006)	13,093(12,014)
20	30,097(11,386)	30,098(11,396)	29,792(11,836)	12,724(11,871)
30	21,939(11,464)	21,939(11,473)	22,265(12,049)	10,296(10,893)
40	20,552(10,250)	20,826(10,787)	20,586(10,281)	10,288(10,898)
50	18,970(9,964)	18,975(9,984)	19,000(9,968)	8,638(8,069)
60	18,200(9,392)	18,197(9,395)	18,225(9,388)	8,638(8,069)
70	14,901(3,825)	14,898(3,828)	14,916(3,844)	8,265(7,482)
80	11,882(5,078)	11,884(5,093)	11,905(5,124)	5,752(3,346)
90	5,564(5,430)	5,564(5,428)	5,589(5,503)	5,493(3,560)
100	45,564(5,430)	5,564(5,428)	5,589(5,503)	4,833(3,057)
MAP	16,281(7,149)	16,307(7,204)	**16,340(7,416)**	7,651(7,237)

Tabela 5.7. *Precisão média para cada nível de revocação e MAP da combinação de todas as evidências textuais.*

A análise dos resultados obtidos nos mostra que as combinações que envolvem as tags de descrição, os metadados e as passagens textuais de 10, 20 ou 40 termos apresentam comportamento similar, havendo visivilmente uma melhor eficácia da combinação que envolvem as pessagens de 40 termos. Assim como nas demais combinações, novamente constata-se o pior desempenho da combinação que envolve o texto completo.

5.4 Experimentos Comparativos dos Métodos de Recuperação de Imagens da Web

Os experimentos finais contemplam a comparação dos resultados obtidos pelo arcabouço evolucionário e os melhores resultados obtidos pelas diversas abordagens do Modelo Bayesiano, apresentados na seção anterior. Na figura 5.8 são apresentadas as curvas geradas a partir dos valores de precisão média, para cada nível de revocação e na tabela 5.8 contém os valores médio de precisão para cada nível de revocação, com o desvio médio apresentado entre parênteses, e os valores médio de MAP.

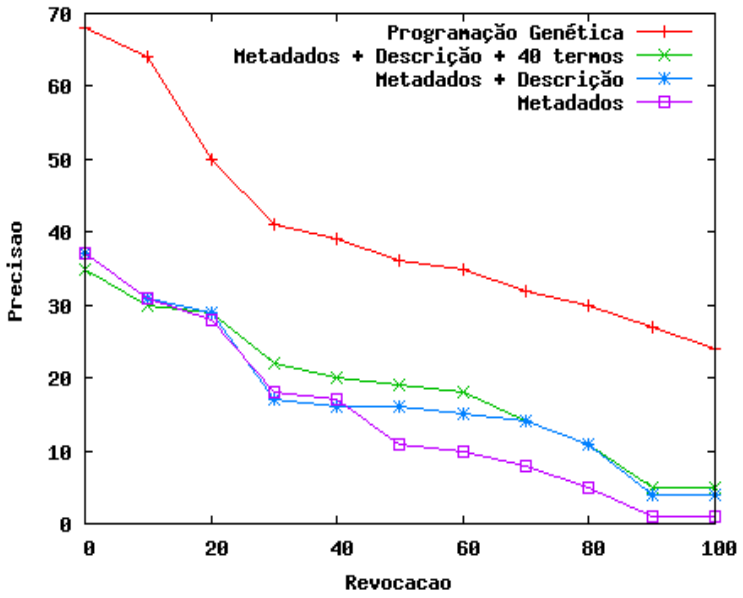

Figura 5.8. Curvas de Precisão versus Revocação do arcabouço e das melhores abordagens do bayesiano.

Revocação	Abordagem			
	META	META + DESC	META + DESC + 40 T	PG
0	37,097(12,788)	37,145(14,423)	35,838(15,450)	68,144(5,392)
10	31,745(15,569)	31,591(10,634)	30,507(12,006)	64,014(6,671)
20	28,024(13,500)	29,449(10,843)	29,792(11,836)	50,618(7,471)
30	18,861(12,715)	17,443(5,627)	22,265(12,049)	41,586(10,327)
40	17,406(11,770)	16,580(4,709)	20,586(10,281)	39,957(9,276)
50	11,386(9,823)	16,371(4,547)	19,000(9,968)	36,141(12,708)
60	10,439(9,060)	15,544(4,340)	18,225(9,388)	35,529(12,619)
70	8,620(6,494)	14,899(3,788)	14,916(3,844)	32,608(10,315)
80	5,167(5,729)	11,643(4,796)	11,905(5,124)	30,604(7,703)
90	1,270(2,625)	4,652(5,857)	5,589(5,503)	27,837(8,436)
100	1,270(2,625)	4,636(5,852)	5,589(5,503)	24,527(7,933)
MAP	13,091(7,067)	14,898(4,501)	16,340(7,416)	**35,313(8,861)**

Tabela 5.8. *Precisão média para cada nível de revocação e MAP do arcabouço e das melhores abordagens do bayesiano.*

Comparando os resultados apresentados é possível constatar o melhor desempenho do arcabouço evolucionário para todos os níveis de revocação. Esse mesmo cenário pode ser confirmado pelo valor do MAP médio. Para constatar estatisticamente os melhores resultados do arcabouço evolucionário, na tabela 5.9 são apresentados os valores dos intervalos de confiança do PG e das melhores abordagens do bayesiano, para um nível de confiança de 95%. Como não há sobreposição dos intervalos de confiança, para qualquer nível de revocação, pode-se afirmar que o modelo evolucionário apresenta melhores resultados que o Modelo Bayesiano.

Revocação	Abordagem			
	META	META + DESC	META + DESC + 40 T	PG
0	$37,0979 \pm 11,209$	$37,145 \pm 12,642$	$35,838 \pm 13,543$	$68,144 \pm 4,726$
10	$31,745 \pm 13,647$	$31,591 \pm 9,321$	$30,507 \pm 10,524$	$64,014 \pm 5,848$
20	$28,024 \pm 11,833$	$29,449 \pm 9,504$	$29,792 \pm 10,374$	$50,618 \pm 6,549$
30	$18,861 \pm 11,145$	$17,443 \pm 4,932$	$22,265 \pm 10,562$	$41,586 \pm 9,052$
40	$17,406 \pm 10,317$	$16,580 \pm 4,128$	$20,586 \pm 9,012$	$39,957 \pm 8,131$
50	$11,386 \pm 8,611$	$16,371 \pm 3,985$	$19,000 \pm 8,737$	$36,141 \pm 11,139$
60	$10,439 \pm 7,941$	$15,544 \pm 3,804$	$18,225 \pm 8,229$	$35,529 \pm 11,061$
70	$8,620 \pm 5,692$	$14,899 \pm 3,320$	$14,916 \pm 3,369$	$32,608 \pm 9,041$
80	$5,167 \pm 5,022$	$11,643 \pm 4,203$	$11,905 \pm 4,491$	$30,604 \pm 6,752$
90	$1,270 \pm 2,301$	$4,652 \pm 5,134$	$5,589 \pm 4,823$	$27,837 \pm 7,394$
100	$1,270 \pm 2,301$	$4,636 \pm 5,130$	$5,589 \pm 4,823$	$24,527 \pm 6,953$

Tabela 5.9. *Intervalos de confiança a 95% para a PG e as melhores abordagens do bayesiano*

Conforme apresentado no capítulo 4, o objetivo do arcabouço evolucionário é definir uma função que combine as evidências textuais de tal forma que esta função seja utilizada para determinar um conjunto-resposta ordenado (*rank*) mais preciso quanto a relevância para uma dada consulta. Dessa forma, para cada uma das cinco execuções do arcabouço evolucionário, foi obtido um melhor indivíduo dentre aqueles avaliados para o correspondente conjunto de imagens da fase de teste. Esses indivíduos correspondem às funções de combinação mencionadas e estão apresentados na tabela 5.10. Nos operadores apresentados, *meta* refere-se as evidências de metadados, *passa40* e *passa20* são, respectivamente, passagens textuais de 20 e 40 termos, *textoCompleto* é o texto completo do documento HTML, *descricao* são as tags de descrição e *rlog* representa o logaritmo na base dez.

A partir dos dados apresentados na Tabela 5.10 pode-se constatar que as evidências correspondentes às *tags* de metadados, às passagens textuais de 20 e 40 termos e às *tags* de descrição, estão presentes em todos os melhores indivíduos obtidos pelo método evolucionário. Deve ser destacado que a presença dessas evidências ocorre por

Fold	Melhor Indivíduo PG
1	(rlog (+ (+ passa40 passa40) (rlog (+ meta descricao))))
2	(rlog (+ (* (+ passa40 textoCompleto) (* meta meta)) (+ descricao passa20)))
3	(+ (+ 48.48 (rlog (rlog passa40))) (rlog passa40))
4	(+ (rlog passa40) 48.48)
5	(rlog (+ meta (+ descricao (+ passa40 passa40))))

Tabela 5.10. *Melhores indivíduos obtidos pela PG, para cada uma das execuções do 5-fold crossvalidation*

mais de uma vez, como nos *folds* 1, 2 e 5. Essas constatações reafirmam a expressiva contribuição das informações armazenadas nessas fontes de evidência no processo de recuperação de imagens da Web e indicam sua maior contribuição no processo (e.g., passa40 tem duas vezes mais peso nos *folds* 1 e 5). Interessante observar a presença do texto completo na formação do indivíduo correspondente à segunda execução (*fold* igual a 2). Este resultado vai de encontro àquele obtido com as diversas equações do Modelo Bayesiano que consideram essa fonte de evidência, pois nestes casos a consideração do texto completo do documento HTML contribui de forma pouco expressiva ou até mesmo negativa para a qualidade dos resultados. Além disso, também indica que esta fonte de evidência pode ser interessante para algumas consultas.

5.5 Exemplo de Imagens Retornadas pelos Métodos de Recuperação de Imagens

Para ilustrar a aplicação dos métodos de recuperação de imagens analisados neste trabalho, foram recuperadas as dez primeiras imagens retornadas pelos métodos que empregam a PG e o Modelo Bayesiano.

As figuras 5.9 e 5.10 contêm as dez primeiras imagens retornadas,respectivamente, pelo Arcabouço Evolucionário e pelo Modelo Bayesiano, considerando neste caso a equação resultante da cvidência isolada correspondente aos metadados.

Figura 5.9. Imagens retornadas pelo arcabouço evolucionário para a consulta por *animais*.

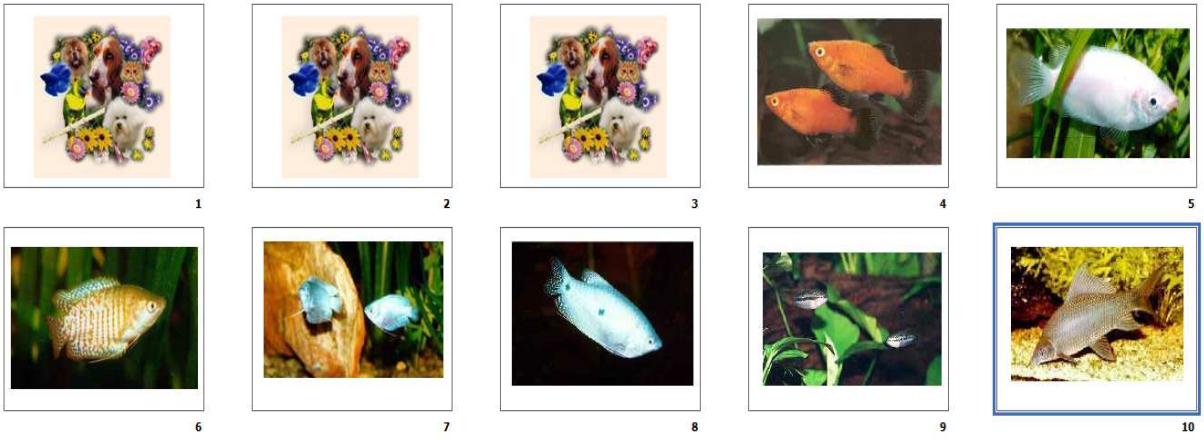

Figura 5.10. Imagens retornadas pela evidência textual metadados do Modelo Bayesiano para a consulta por *animais*.

As figuras 5.11 e 5.12 contêm as dez primeiras imagens retornadas,respectivamente, pelo Arcabouço Evolucionário e pelo Modelo Bayesiano, considerando neste caso a equação resultante da combinação dos metadados e *tags* de descrição.

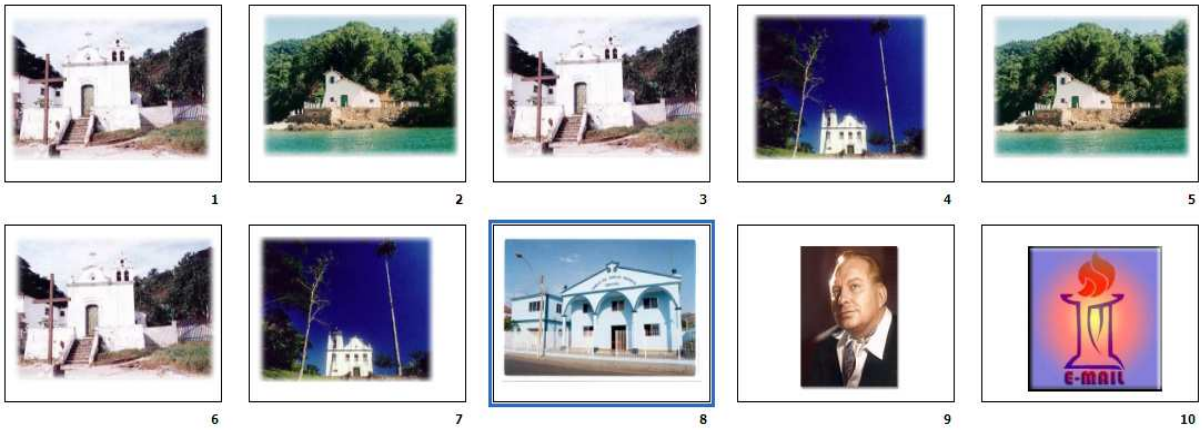

Figura 5.11. Imagens retornadas pelo arcabouço evolucionário para a consulta por *igreja*.

Figura 5.12. Imagens retornadas pela combinação de metadados e *tags* de descrição do Modelo Bayesiano para a consulta por *igreja*.

Nas figuras 5.13 e 5.14 estão apresentadas as dez primeiras imagens retornadas,respectivamente, pelo Arcabouço Evolucionário e pelo Modelo Bayesiano, considerando neste caso a equação resultante da combinação dos metadados, *tags* de descrição e passagens de texto com 40 termos.

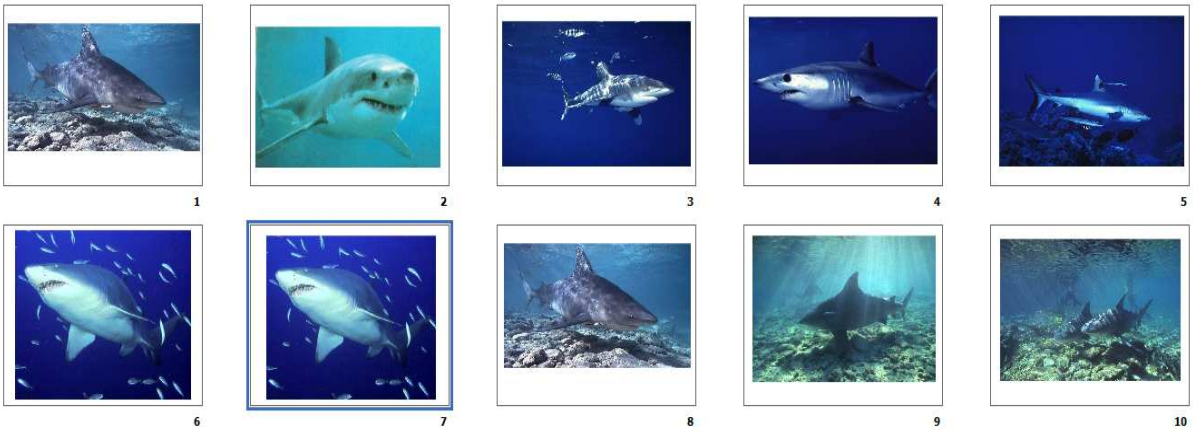

Figura 5.13. Imagens retornadas pelo arcabouço evolucionário para a consulta por *tubarão*.

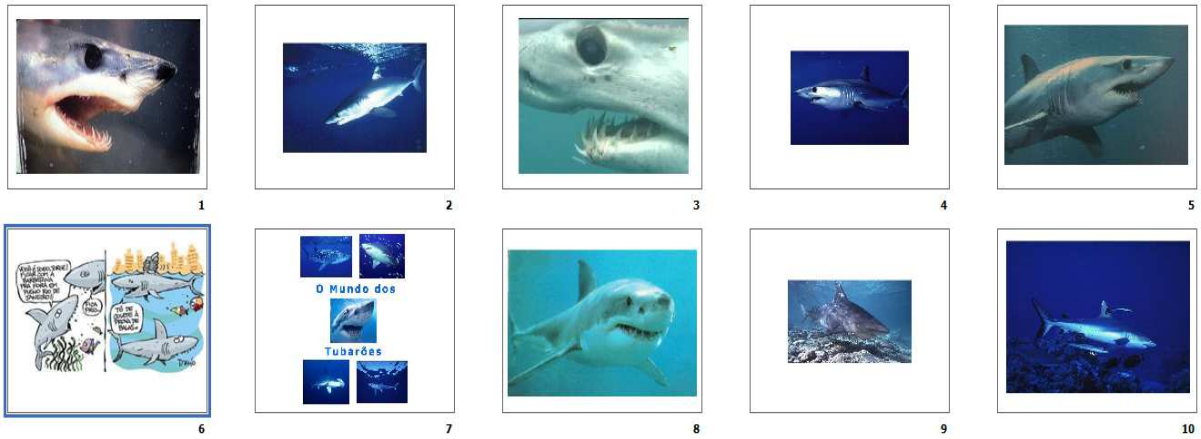

Figura 5.14. Imagens retornadas pela combinação de metadados, *tags* de descrição e passagem de 40 termos do Modelo Bayesiano para a consulta por *tubarão*.

Capítulo 6

Conclusões e Trabalhos Futuros

Neste trabalho foi proposto um modelo de recuperação de imagem, baseado na Programação Genética. O objetivo deste é combinar informação de distintas fontes de evidências textuais, ocorridas em documentos HTML, para melhorar a qualidade das imagens recuperadas a partir de uma coleção Web. A solução do método evolucionário então consiste em uma função resultante da combinação de distintas evidências textuais associadas às imagens de um documento HTML. Para comparar o desempenho do método proposto, implementamos o Modelo Bayesiano apresentado em [Coelho et al., 2004] e realizamos testes com as diferentes equações derivadas deste. Para a avaliação dos resultados, foram utilizadas as métricas de precisão, revocação, R-precision e MAP.

Um dos objetivos alcançados por este trabalho corresponde ao estudo do comportamento do arcabouço para diferentes configurações dos parâmetros aderentes à técnica da Programação Genética. Foi analisada a melhor configuração do método evolucionário quanto a dois quesitos: tamanho da população e número de gerações. Os testes realizados mostraram que a configuração correspondente a uma população de 50 indivíduos, com um número de 15 gerações é aquela que gera melhores valores, dentre as métricas avaliadas.

Em seguida, por meio de experimentos do tipo *5-fold cross validation*, foi avaliado o impacto de recuperação no Modelo Bayesiano, a partir de três fontes distintas associadas às imagens de uma coleção Web: *tags de descrição*, os *metadados* do documento que contém a imagem e as *passagens textuais*, sendo que nestas estão incluídas as passagens textuais com 10, 20 ou 40 termos ao redor da imagem e o texto completo do documento. Concluímos que três são as abordagens que apresentam melhores resultados: *metadados*, *metadados + tags de descrição* e *metadados + tags de descrição + passagens de 40 termos*. Este resultado difere daquele obtido em [Coelho et al., 2004], o qual demonstrou por meio dos resultados obtidos que a utilização das passagens tex-

tuais de 40 termos, isoladamente, seria suficiente para obter resultados melhores. Essa diferença dos resultados pode ser justificada pelo fato da coleção web empregada no presente trabalho conter um maior número de documentos HTML e de imagens, em comparação com a coleção empregada em [Coelho et al., 2004]. Outra razão pode versar sobre uma melhor geração dos documentos HTML presentes na web atual, em que podemos destacar o preenchimento mais adequado das informações sobre o documento e as informações relacionadas à descrição das imagens, ou seja, os *metadados* e as *tags de descrição*.

Em atendimento a outro objetivo deste trabalho, ainda empregando o método *5-fold cross validation*, foram realizados experimentos comparativos do arcabouço evolucionário e das melhores abordagens do Modelo Bayesiano. Os resultados mostraram que o método que emprega a Programação Genética apresenta melhor desempenho em termos de precisão média, para todos os níveis de revocação, e de MAP, considerando um nível de confiança igual a 95%.

Como trabalhos futuros podem ser citados: utilizar o arcabouço na recuperação de imagens pertencentes a coleções existentes em outros contextos como uma coleção formada por imagens médicas, incorporar novas evidências textuais e funções de similaridade como *OKapi* e *BM25*, além de comparar os resultados obtidos com outras soluções para o problema, em especial com aquelas que envolvam métodos de aprendizagem.

Referências Bibliográficas

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Chang, H.-T. (2008). Web image retrieval systems with automatic web image annotating techniques. *WSEAS Transactions on Information Science and Applications*, 5(8):1313--1322.

Coelho, T. A. S. (2001). Recuperação de imagens na web baseada em múltiplas evidências textuais. Master's thesis, UFMG.

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

da S. Torres, R.; Falcão, A. X.; Zhang, B.; Fan, W.; Fox, E. A.; Gonçalves, M. A. & Calado, P. (2005). A new framework to combine descriptors for content-based image retrieval. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, number IC-05-21, pp. 335--336, New York, NY, USA.

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

de Almeida, H.; Gonçalves, M.; Cristo, M. & Calado, P. (2007). A combined component approach for finding collection-adapted ranking functions based on genetic programming. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 399--406.

de Carvalho, M. G.; Laender, A. H. F.; Gonçalves, M. A. & Porto, T. C. (2008). The impact of parameter setup on a genetic programming approach to record deduplication. In *SBBD '08: Proceedings of the 23rd Brazilian symposium on Databases*, pp. 91--105, Porto Alegre, Brazil, Brazil. Sociedade Brasileira de Computação.

Duda, R. O.; Hart, P. E. & Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience.

Fan, W.; Fox, E. A.; Pathak, P. & Wu, H. (2004a). The effects of fitness functions on genetic programming-based ranking discovery for web search: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 55(7):628--636.

Fan, W.; Gordon, M. & Pathak, P. (2004b). A generic ranking function discovery framework by genetic programming for information retrieval. *Information Processing and Management*, 40(4):587--602.

Fan, W.; Gordon, M. D. & Pathak, P. (2000). Personalization of search engine services for effective retrieval and knowledge management. In *ICIS '00: Proceedings of the twenty first international conference on Information systems*, pp. 20--34, Atlanta, GA, USA. Association for Information Systems.

Fan, W.; Gordon, M. D. & Pathak, P. (2004c). Discovery of context-specific ranking functions for effective information retrieval using genetic programming. *IEEE Trans. on Knowl. and 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. *J. Manage. Inf. Syst.*, 21(4):37--56.

Fan, W.; Gordon, M. D.; Pathak, P.; Xi, W. & Fox, E. A. (2004d). Ranking function optimization for effective web search by genetic programming: An empirical study. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4*, p. 40105, Washington, DC, USA. IEEE Computer Society.

Ferreira, C. D.; S. Torres, R.; Gonçalves, M. A. & Fan, W. (2008). Image retrieval with relevance feedback based on genetic programming. In *SBBD*, pp. 120--134, Campinas, Brazil.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Google (2009). Google image search. `http://images.google.com/`.

Iskandar, D.; Pehcevski, J.; Thom, J. & Tahaghoghi, S. (2005). Combining image and structured text retrieval. *Advances in XML Information Retrieval and Evaluation: 4th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX*, 3977:525--539.

Koza, J. R. (1992). *Genetic Programming*. MIT Press.

Koza, J. R. (1994). *Genetic Programming II*. MIT Press.

Lacerda, A.; Cristo, M.; Gonçalves, M. A.; Fan, W.; Ziviani, N. & Ribeiro-Neto, B. (2006). Learning to advertise. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 549--556, New York, NY, USA. ACM.

Modesto, M.; Álvaro R. Pereira Jr.; Castillo, C.; Ziviani, N. & Baeza-Yales, R. (2005). Um novo retrato da web brasileira. *Anais do XXXII Seminário Integrado de Hardware e Software (SEMISH 32)*, pp. 2005--2017.

MSN (2009). Msn image search. `http://search.msn.com/images/`.

Silva, A.; Gonçalves, M.; Laender, A.; Modesto, M.; Cristo, M. & Ziviani, N. (2009). Finding what is missing from a digital library: A case study in the computer science field. *Information Processing and Management*.

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

Torres, R. D. S. & Falcão, A. X. (2006). Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada*, 13:161--185.

Veloso, E. A.; Moura, E. S.; Golgher, P. B.; Silva, A. S.; Laender, A. H.; Ribeiro-Neto, B. & Ziviani, N. (2000). Um retrato da web brasileira. *Anais do XXI Seminário Integrado de Hardware e Software (SEMISH 00)*.

Voorhees, E. & Harman, D. (2000). Overview of the eighth text retrieval conference (trec-8). *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, pp. 1--24.

Yahoo (2009). Yahoo image search. `http://images.search.yahoo.com/images`.

Zhang, B. (2006). *Intelligent Fusion of Evidence from Multiple Sources for Text Classification*. Phd thesis, Virginia Polytechnic Institute and State University, USA.

Zongker, D. & Punch, B. (1996). lilgp 1.01 user's manual. Technical report, Michigan State University, USA.