IComp/UFAM

RE-RANKING DE BUSCA VISUAL DE PRODUTOS USANDO INFORMAÇÃO MULTIMODAL

Joyce Miranda dos Santos

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Computação - IComp, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientador: João Marcos Bastos Cavalcanti

Março de 2013

Manaus - AM

RE-RANKING DE BUSCA VISUAL DE PRODUTOS USANDO INFORMAÇÃO MULTIMODAL

Joyce Miranda dos Santos

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO DO INSTITUTO DE COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DO AMAZONAS COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM INFORMÁTICA.

Aprovado por:

Prof. João Marcos Bastos Cavalcanti, D.Sc.

Prof. Edleno Silva de Moura, D.Sc.

Prof. Ricardo da Silva Torres, D.Sc.

MARÇO DE 2013

MANAUS, AM – BRASIL

A minha mãe Cássia e a minha irmã Jéssica que me deram o suporte emocional necessário para concluir mais essa etapa da minha vida. Em especial, dedico este trabalho ao meu pai Jazon (in memorian), meu amigo, apoiador incondicional, de quem sinto uma saudade imensurável.

Agradecimentos

A Deus, que me deu sabedoria nos momentos difíceis desta jornada.
Aos meus pais, pela dedicação e esforço de toda uma vida para que eu pudesse concluir meus estudos e alcançar meus objetivos.
Ao professor João Cavalcanti pela orientação e ao professor Edleno Moura pelo direcionamento em cada etapa do mestrado.
Aos meus colegas de mestrado, que compartilharam comigo os momentos de desespero, mas que também dividiram comigo muitos momentos de diversão, que renderão histórias para a vida toda.
Por fim, a todos que, de alguma forma, contribuíram para a conclusão deste trabalho.

"não temas, porque eu sou contigo;

não te assombres, porque eu sou o teu Deus;

eu te fortaleço, te ajudo, e te sustento com a minha destra fiel"

(Isaías 41:10)

Resumo

Com o rápido desenvolvimento da Internet, a popularização de dispositivos móveis e de sites de comércio eletrônico, procurar um produto específico a partir de uma imagem tem se tornado uma área de pesquisa promissora. Nesse contexto, técnicas de CBIR (*Content-Based Image Retrieval*) vêm sendo exploradas para apoiar e melhorar a experiência de compra dos consumidores. Neste trabalho, abordamos o problema de busca visual de produtos usando uma imagem como consulta, no lugar da mais popular abordagem de busca que é baseada em palavras-chave. Nós propomos uma estratégia de re-ranking que faz uso de informações multimídia normalmente disponíveis nas bases de dados de produtos. Nossa estratégia faz uso de informações de categoria e descrição textual associadas às imagens melhor posicionadas de um ranking inicial gerado por técnicas puramente de CBIR. Experimentos foram realizados considerando o julgamento de usuários em duas coleções de imagens coletadas a partir de sites de comércio eletrônico. Nossos resultados mostram que nossa estratégia alcança ganhos significativos quando comparada à busca puramente visual.

PALAVRAS-CHAVE: Busca visual de produtos, Re-ranking de imagens, Comércio eletrônico

Abstract

With the fast development of the Internet and the popularization of mobile devices, searching for a specific product in e-commerce Web sites through a query image has become a very promising area of research. In this context, CBIR (Content-Based Image Retrieval) techniques have been exploited to support and improve the shopping experience of consumers. In this dissertation, we address the problem of product visual search using an image as a query, instead of the more popular approach of search based on keywords. We propose a strategy for re-ranking based on multimedia information usually available in database of products. Our strategy makes use of category information and textual description associated with the top-k images of an initial ranking generated by CBIR techniques only. Experiments were performed considering the judgment of users on two collections of images collected from popular e-commerce Web sites. Our results show that our strategy achieves significant gains compared to an approach based only on CBIR techniques.

KEY-WORDS: Products visual search, Image re-ranking, E-commerce

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Capítulo 1

Introdução

Desde a criação da Web, uma grande quantidade de dados digitais vem sendo acumulada. Com a popularização de dispositivos capazes de capturar imagens e armazená-las em meio digital, o volume desse tipo de conteúdo só aumenta. Freqüentemente, imagens digitais estão presentes em quase todos Web sites e em aplicações específicas. Essa realidade faz surgir uma demanda cada vez maior por ferramentas capazes de recuperar imagens de forma rápida e eficiente.

Marcações textuais vêm sendo utilizadas com sucesso para organizar e buscar informações no meio digital. Entretanto, quando o objetivo da busca é o de recuperar imagens, utilizar apenas texto para este fim pode diminuir de forma considerável sua eficiência. Ao usar texto, é necessário que haja anotações textuais associadas às imagens da coleção. Normalmente, essas anotações são feitas de forma manual. Uma limitação gerada por isso, é que a anotação gerada está sujeita à interpretação e subjetividade da pessoa que descreveu a imagem. Outra limitação, é o fato de que gerar descrição manual para bases com milhares ou milhões de imagens torna-se praticamente inviável.

Uma alternativa para isso é utilizar técnicas de CBIR (*Content-Based Image Retrieval*) que fazem uso de características de baixo nível como cor, forma e textura para recuperar imagens a partir de uma imagem de consulta. Técnicas de CBIR vêm sendo aplicadas com sucesso em diversas aplicações, tais como: identificação digital, sistemas de informação de biodiversidade, prevenção à criminalidade e investigação médica. Uma das limitações dessa abordagem está no fato de o usuário ao buscar por uma imagem, além de características visuais semelhantes, está interessado também em sua semântica, ou seja, no significado associado à imagem. Identificar a semântica de uma imagem a partir de

características de baixo nível ainda é um desafio a ser superado na área de CBIR [20].

Dentro do contexto de recuperação de imagens, quando existe uma base com evidências textuais e visuais disponíveis, o grande desafio está em definir a melhor maneira de combinar as abordagens de recuperação textual e recuperação visual. Nesse sentido, a utilização de uma estratégia multimodal é apontada como uma direção promissora. O propósito desse tipo de estratégia é tentar tirar vantagem da riqueza de informação presente nas características visuais da imagem e da semântica oferecida pelo conteúdo textual. A ideia é usar diferentes modalidades para expandir e complementar a informação sobre o item consultado.

Quanto mais genérico for o domínio da aplicação, mais difícil é definir uma estratégia multimodal adequada. Um exemplo disso é a busca na Web, em que as diversas possibilidades de busca do usuário torna muito complexa a tarefa de definir uma solução eficaz para todos os cenários. Em aplicações de domínio específico, a semântica da busca está presente em um contexto limitado, permitindo assim mais opções para que características visuais e textuais sejam exploradas de forma eficiente. Neste trabalho, estratégias multimodais são exploradas no domínio específico de sites de comércio eletrônico. Essa escolha foi motivada pelo fato desses sites apresentarem como característica uma base multimodal composta por imagens de produtos e suas respectivas informações textuais.

O setor de comércio eletrônico é apontado como uma das áreas que apresentam grandes perspectivas de crescimento para os próximos anos[1]. Pesquisas recentes apontam que a categoria de produtos ligados à moda e acessórios está entre os cinco maiores mercados por número de pedidos desse setor [2]. Considerando que o apelo visual desses produtos é muito grande para a decisão de compra do consumidor, a maioria dos sites limitam seus consumidores a procurar produtos utilizando apenas descrições textuais.

Utilizar recursos visuais para encontrar produtos tornou-se uma característica importante para apoiar o processo de compras on-line. Neste sentido, aplicações de comércio eletrônico podem se beneficiar com a utilização de técnicas de CBIR. A submissão de uma imagem que possua aspectos visuais semelhantes pode ajudar o usuário na tarefa de encontrar produtos mais similares ao que está procurando.

A aplicação de uma estratégia multimodal para o problema de busca visual de pro-

[1]Estudo divulgado em abril de 2011 pela Forrester Research, uma empresa especializada na elaboração de pesquisas voltadas para a internet. http://www.forrester.com

[2]Matéria divulgada em março de 2012 pela Isto É Dinheiro, uma revista especializada em tendências na Economia e nos negócios. http://www.istoedinheiro.com.br

dutos mostra ser uma solução viável. Comumente, esses sites classificam seus produtos em categorias como por exemplo: eletrônicos, eletrodomésticos, brinquedos e vestuário. No setor de vestuário, algumas das categorias definidas são roupas femininas/masculinas, calçados femininos/masculinos, bolsas e acessórios. Além disso, todos sites de comércio eletrônico apresentam dados sobre seus produtos, tais como nome, descrição e preço. Isto proporciona uma importante e rica fonte de informação que pode ser utilizada para melhorar os resultados da busca.

Neste trabalho, abordamos o problema de busca visual de produtos, permitindo que usuários desses tipos de Web sites possam apresentar a imagem de um produto como consulta, a fim de obter produtos similares disponíveis para compra. O que torna esse problema desafiador é a falta de informação textual associada à consulta. Uma solução imediata para este problema consiste em aplicar técnicas existentes de CBIR para recuperar imagens semelhantes à imagem de consulta. No entanto, depois de experimentar algumas soluções encontradas na literatura, percebemos que o resultado obtido com esta abordagem geralmente alcança taxas de precisão inferiores ou equivalentes às abordagens puramente textuais. Assim, concluímos que usar apenas a abordagem CBIR não é uma solução eficaz para o problema abordado neste trabalho.

Recentemente, muitos trabalhos de pesquisa têm explorado o *re-ranking visual* como estratégia para melhorar a relevância dos resultados da busca por imagens [18, 32, 34, 1, 24, 48]. Re-ranking visual pode ser definido como a reordenação do resultado inicial de uma busca por imagem com base em informação multimodal. Nossa estratégia se baseia na extração de informações de categoria e de descrição associadas às imagens dos produtos presentes no topo-k de um ranking inicial gerado por técnicas de CBIR. Foram realizados experimentos considerando o julgamento de relevância de usuários em duas coleções extraídas de sites de comércio eletrônico. Os resultados alcançados mostraram que nossa estratégia alcança ganhos significativos quando comparada à busca puramente visual.

Este trabalho tem como objetivo definir um modelo eficiente de busca visual capaz de obter resultados relevantes em uma base multimodal de produtos. Existem várias coleções nas quais tal modelo pode ser aplicado. Neste trabalho, o modelo proposto será avaliado sobre uma coleção multimodal de produtos de comércio eletrônico, limitando o escopo a imagens de roupas, calçados e acessórios. O visual desses produtos afeta diretamente a

decisão de compra do usuário.

As contribuições deste trabalho são: (i) um método automatizado para descobrir a categoria de um produto dada uma imagem de consulta e (ii) um novo método para busca visual de produtos que combina informação de categoria para reordenar o ranking visual.

1.1 Organização do Trabalho

A presente dissertação está estruturada da seguinte forma. No Capítulo 2, são apresentados conceitos importantes para o entendimento do trabalho, além de incluir trabalhos relacionados a re-ranking e busca visual de produtos. No Capítulo 3, explicamos a solução proposta para o problema apresentado. No Capítulo 4, apresentamos os experimentos realizados e os resultados obtidos. Por fim, no Capítulo 5, discutimos as conclusões e direcionamento para trabalhos futuros.

Capítulo 2

Fundamentação Teórica e Trabalhos Relacionados

O avanço tecnológico tornou possível o armazenamento e a manipulação de dados multimídia como texto, imagem, áudio e vídeo nas mais diversas aplicações. Sistemas de recuperação multimídia surgiram a partir da necessidade de gerenciar o grande volume desse tipo de conteúdo presente em diferentes domínios. Nesse contexto, diversos métodos vêm sendo propostos com o intuito de recuperar imagens de forma rápida e eficiente. Basicamente, existem duas abordagens para recuperar imagens, que são: TBIR (*Text-Based Image Retrieval*) e CBIR (*Content-Based Image Retrieval*).

A abordagem TBIR parte de uma busca textual para recuperar imagens e é baseada em técnicas tradicionais de recuperação de texto. Nesse caso, para que a recuperação seja realizada, é necessário que exista alguma anotação textual associada às imagens da coleção consultada. Normalmente, essa anotação é feita de forma manual, estando sujeita à interpretação e subjetividade da pessoa que descreveu a imagem. Para a tarefa de busca por imagens, na maioria das vezes, essa anotação é insuficiente para descrever as características visuais presentes em uma imagem.

A abordagem CBIR foi proposta para ser uma alternativa para a busca por imagens. Nela, a busca é feita a partir de uma imagem de consulta, e não a partir de um texto como na abordagem TBIR. Sistemas CBIR usam como agente principal descritores que são responsáveis por reconhecer o conteúdo visual de uma imagem e retornar imagens com conteúdo semelhante. Esse conteúdo é representado por características de baixo nível como cor, forma e textura. Uma limitação encontrada na área de CBIR é a falta de

semântica associada à consulta. Isto porque, os usuários ao buscarem por uma imagem além de características visuais semelhantes, também estão interessados no significado associado à imagem, que dificilmente é obtido apenas por meio de características de baixo nível.

Na tentativa de suprir as limitações das abordagens apresentadas, existem linhas de pesquisa que buscam combinar estas abordagens por meio de estratégias multimodais. Este trabalho explora a busca visual de produtos a partir da aplicação de técnicas de CBIR, recuperação multimodal e estratégias de re-ranking. Sendo assim, nesse capítulo são apresentados conceitos e trabalhos relacionados a esses assuntos.

2.1 Fundamentos de CBIR

O descritor de imagem é um dos componentes mais importantes em um sistema CBIR. Sua função é a de quantificar quão similar são duas imagens. Em [41], um descritor é apresentado como uma tupla $(\varepsilon_d, \delta_d)$, onde:

- ε_d: é a função responsável por extrair o conteúdo visual de uma imagem e armazená-lo em um vetor de características. Esse vetor será formado por informações visuais que consideram aspectos como cor, forma e textura. Essas informações visuais são obtidas a partir de técnicas e algoritmos de processamento de imagens.

- δ_d: é a função responsável por comparar dois vetores de características. Dados dois vetores, essa função calcula a similaridade entre duas imagens.

A Figura 2.1 mostra o fluxo típico de uma sistema CBIR. Um processo de extração de características é aplicado sobre cada imagem de uma coleção de imagens, por meio da função ε_d. O resultado desse processo é a geração de vetores que codificam características visuais das imagens, tais como cor, forma e textura. O tamanho do vetor vai depender da quantidade de características usada para representar as imagens.

Uma vez que uma imagem de consulta é submetida, o mesmo processo de extração é realizado sobre a imagem e um vetor de característica também é obtido. A partir desse momento, esse vetor é comparado com os vetores de características que foram gerados a partir da coleção de imagens, por meio da função δ_d. Baseado nos valores de similaridade,

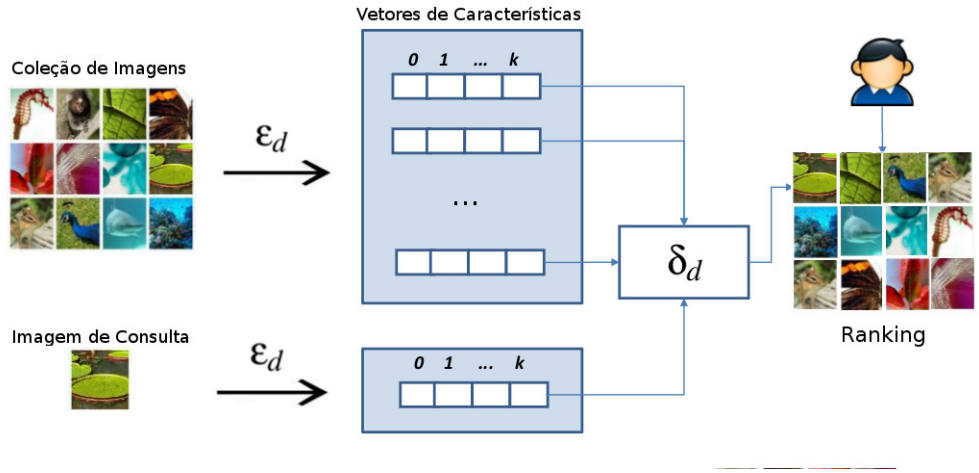

Figura 2.1: Fluxo de uma solução CBIR típica.

um ranking final é gerado com as imagens da coleção ordenadas a partir dos seus valores de similaridade com a imagem de consulta.

Durante a definição de descritores de imagens é preciso que sejam levadas em consideração algumas propriedades para que a eficácia seja garantida no processo de indexação e recuperação de imagens. Dentre as propriedades desejáveis para um descritor estão: insensibilidade a ruídos, invariância a algumas classes de transformação (rotação e translação), geração de vetores de características compactos que exijam pouco espaço de armazenamento e utilização de uma função de extração computacionalmente eficiente [43].

É importante ressaltar que a eficácia de um descritor não depende somente do algoritmo de geração do vetor de características, mas depende também da função de similaridade usada. O uso adequado de medidas de similaridade ajuda a melhorar o resultado das consultas. Um estudo comparativo entre funções de similaridade é realizado em [5]. As funções da família Minkowski (L_p), são normalmente utilizadas em espaços vetoriais, por esse motivo são amplamente utilizadas em CBIR. Um espaço vetorial existe se os objetos de um determinado domínio correspondem a valores numéricos estruturados em um vetor. As funções mais conhecidas dessa família são: L_1 (*City Block*), L_2 (*Euclidian*) e L_∞ (*Chebychev*).

A função L_1 consiste na soma das diferenças entre as coordenadas de um vetor. Sendo V_q o vetor de características de uma imagem de consulta, V_b o vetor de características de uma imagem da coleção, e v_s o tamanho dos vetores. A distância L_1 usada para calcular a

distância entre dois vetores, é definida formalmente na Equação 2.1.

$$d_{L_1}(V_q, V_b) = \sum_{i=1}^{v_s} \mid V_q(i) - V_b(i) \mid \tag{2.1}$$

A função L_2 corresponde à função comumente utilizada para calcular a distância entre dois vetores. Sua distância é definida formalmente na Equação 2.2.

$$d_{L_2}(V_q, V_b) = \sqrt{\sum_{i=1}^{v_s} (V_q(i) - V_b(i))^2} \tag{2.2}$$

A função L_∞, recebe o máximo da diferença entre suas coordenadas e é definida formalmente na Equação 2.3.

$$d_{L_\infty}(V_q, V_b) = max_{1 \leq i \leq v_s} \mid V_q(i) - V_b(i) \mid \tag{2.3}$$

Escolher o descritor mais adequado para uma determinada aplicação é crucial para o sucesso de um sistema CBIR. Por isso, é importante que sejam conduzidos experimentos comparativos utilizando diferentes descritores com o intuito de utilizar o que alcance o melhor desempenho. Definir que tipos de evidência (cor, forma, textura) irão compor os vetores de características usados para representar as imagens depende diretamente do contexto onde o descritor será aplicado. Cada evidência representa aspectos específicos da imagem que devem ser considerados de acordo com a necessidade da aplicação.

2.1.1 Descritores de Cor

A cor é uma das principais evidências utilizadas por sistemas CBIR devido a sua simplicidade e por exigir um menor custo computacional de extração quando comparada à representação de outras evidências. Ao utilizar essa característica o foco é representar a distribuição de cores da imagem de forma a recuperar imagens que possuam uma composição de cor similar, mesmo que elas pertençam a contextos diferentes.

Uma imagem digital pode ser representada como uma matriz *n x m* em que cada elemento da matriz (pixel) está associado a uma intensidade que representa uma cor em um determinado ponto da imagem. A escolha do espaço de cor pelo qual as imagens serão representadas, analisadas e comparadas é o primeiro passo na definição de um sistema de recuperação de imagens baseado em cor [43]. Entre os modelos de espaços de cor

mais conhecidos estão: RGB (vermelho, verde e azul), CMY (ciano, magenta e amarelo), HSV (tonalidade, saturação e valor), HSI (tonalidade, saturação e intensidade) e YIQ (luminância, interpolação e quadratura) [12]. A cor de um pixel é, em geral, representada por três valores, um para cada canal do espaço de cor utilizado.

Ao codificar cores é comum usar 8 bits para representar um canal de cor. Isto equivale a $2^8 = 256$ níveis de cor para cada canal. Considerando o RGB que possui três canais de cores, isso resultaria em aproximadamente 17 milhões ($256[R]$ *x* $256[G]$ *x* $256[B]$) de cores distintas. Uma imagem com resolução de 300 *x* 300 equivale a 90.000 pontos que deveriam ser considerados em uma análise comparativa pixel a pixel de duas imagens. Esses valores, em termos de quantidade de cores e dimensão espacial, tornaria inviável o processamento de sistemas de recuperação de imagens.

Um sistema de recuperação de imagens necessita de uma representação compacta da distribuição de cores [39]. Para isso ser possível, é feito um processo de quantização que consiste em reduzir a quantidade de bits usada por pixel. Na prática, normalmente, são utilizados no máximo 8 bits por pixel que equivalem a 256 cores diferentes. Assim, torna-se necessário definir um índice com as 256 cores mais significativas e determinar a equivalência entre as cores da imagem e as cores da tabela.

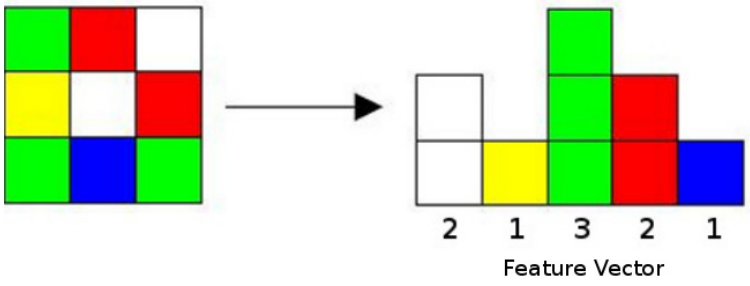

Figura 2.2: Exemplo da construção de um histograma.

A implementação mais comum de um descritor de cor consiste na criação de histogramas que representam a distribuição de cores de uma imagem. O histograma é produzido primeiramente a partir da amostragem das cores de uma imagem em um número de índice de cores. Em seguida, é realizada a contagem do número de pixels da imagem para cada índice. Um exemplo pode ser observado na Figura 2.2, que ilustra a formação de um histograma de cores baseado em um conjunto de nove pixels. O processo de formação de um histograma se baseia na construção de diversas pilhas, uma para cada cor da imagem.

Em seguida, é feito um somatório das ocorrências de uma cor, incrementando, assim, a pilha correspondente.

Uma forma simples de montar o vetor de características a partir de um histograma é inserindo seqüencialmente, em cada índice do vetor, a frequência de cada cor. A partir da representação do conteúdo da imagem por meio do histograma, é possível utilizar as distâncias L1, L2 ou L_∞, como métricas para verificar a similaridade entre dois histogramas.

Vários descritores de cor têm sido propostos na literatura (GCH[40], LCH[40], CCV[31], CBC[37] e BIC[38]). De uma forma geral, eles são agrupados em três classes levando-se em consideração se codificam ou não codificam a informação relacionada à distribuição espacial. As abordagens existentes podem ser classificadas em: (i) *globais*: descrevem a distribuição de cores das imagens como um todo, desprezando a sua distribuição espacial, (ii) *baseadas em particionamento*: decompõem espacialmente as imagens utilizando uma estratégia de particionamento simples e comum a toda imagem, sem levar em consideração o seu conteúdo visual, e (iii) *regionais*: utilizam técnicas automáticas de segmentação para decompor as imagens de acordo com o seu conteúdo visual.

2.1.2 Descritores de Textura

Existem contextos em que apenas as características de cor ou sua intensidade são insuficientes para realizar a descrição das imagens. Algumas imagens se caracterizam pela repetição de um padrão visual sobre uma região, sendo esse padrão repetido de forma exata ou em pequenas variações. Os descritores de textura são usados quando existe nas imagens a serem buscadas um padrão visual com algumas propriedades de homogeneidade que não são resultados simplesmente de uma cor ou intensidade. Para isso, esses descritores buscam representar aspectos de superfície de um objeto, verificando o relacionamento dos pixels com seus vizinhos. Isso permite a representação de atributos como: rugosidade, contraste, aspereza e semelhança com linhas. Entre os descritores de textura existentes na literatura podemos citar: LBP[**?**], HTD[35] e CCOM[22].

2.1.3 Descritores de Forma

Grande parte da informação semântica de uma imagem está associada aos objetos nela presentes. Por exemplo, se uma pessoa é requisitada para escolher imagens semelhantes

a uma imagem que possui uma forte semântica associada a objetos, esta pessoa provavelmente irá ignorar ou colocar em segundo plano características como cor e textura.

Descritores de forma usam técnicas que concedem a descrição total da borda de objetos ou a descrição das características morfológicas das regiões presentes na imagem. Estes descritores são usados quando o usuário deseja executar pesquisas com base no perfil e na estrutura física de um objeto. Contextos bastante comuns para este tipo de aplicação são a busca de informações em bancos de dados de medicina, nos quais as imagens têm características de cor e textura muito semelhantes. Outras aplicações possíveis são: o reconhecimento de caracteres alfanuméricos em documentos, rastreamento de objetos em vídeos e reconhecimento de pessoas em sistemas de segurança. Entre os descritores de forma existentes estão: CSS[29], BAS[2] e SS[42].

Descritores de forma são muito dependentes de um bom processo de segmentação, que consiste em encontrar objetos presentes nas imagens. Em bases cujo o conteúdo é conhecido e controlado é possível ajustar parâmetros e obter bons resultados de segmentação. Em bases heterogêneas, como é o caso da Web, ajustar parâmetros que satisfaçam todas categorias de objetos possíveis é quase impraticável. Outro problema associado aos descritores de forma é o alto custo computacional exigido para o seu processamento.

No cenário de comércio eletrônico, foco deste trabalho, é importante que a solução aplicada não exija um alto custo computacional e seja rápida no retorno de respostas para o usuário. Por esse motivo, nossos experimentos se limitaram na aplicação de descritores de cor, para representar a distribuição de cores, e descritores de textura, para representar o padrão visual das imagens dos produtos.

2.2 Recuperação de Informação Multimodal

Recuperação de informação multimodal consiste em combinar mais de uma modalidade de evidência para recuperar informação. No caso da recuperação de imagens, essa abordagem pode ser aplicada quando existe uma base multimodal composta por imagens e alguma anotação textual associada. O desafio está em definir a melhor forma de combinar tais evidências de forma a maximizar as vantagens e minimizar as limitações de ambas.

O primeiro passo dessa estratégia consiste em definir como representar e processar cada evidência de forma individual. No caso das evidências textuais, existem estratégias

bastante conhecidas, como é o caso do Modelo Vetorial [28] que é eficaz e tem sido amplamente estendido para resolver problemas de busca em bases textuais. No caso do CBIR, ainda não existe uma concordância geral sobre o tipo de representação ou o modelo de recuperação que deve ser aplicado.

Uma vez definida a forma como as evidências serão representadas e processadas, o próximo passo é escolher qual modelo será utilizado para combinar as evidências. De uma forma geral, existem três modelos que podem ser classificados de acordo com o nível escolhido para combinar evidências [13]: *early fusion*, *late fusion* e *intermedia fusion* [1]. A estratégia de fusão de cada modelo pode ser observada na Figura 2.3.

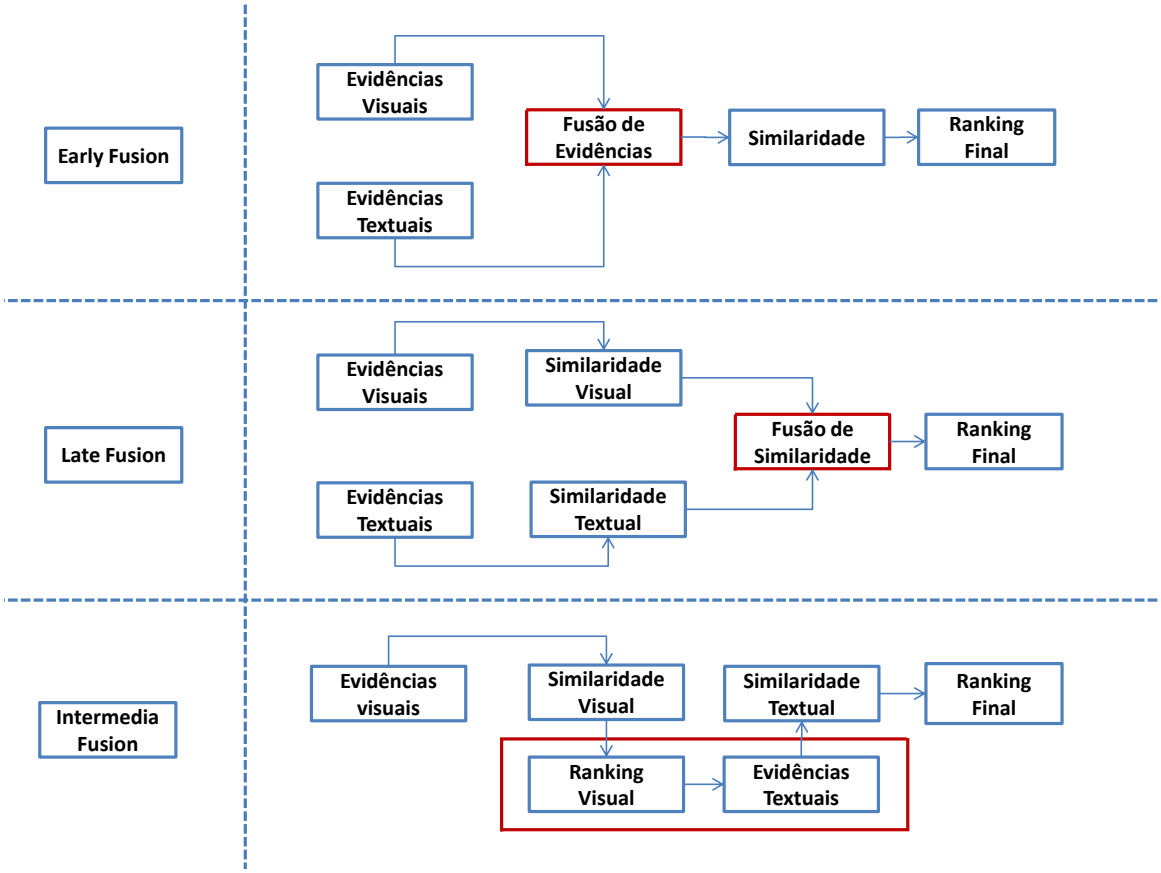

Figura 2.3: Abordagens de fusão evidências.

A abordagem early fusion consiste em concatenar as representações textuais e visuais em um único vetor de características. A vantagem desse modelo é que este permite uma verdadeira representação multimídia em que uma regra de decisão está baseada em todas as fontes de evidência. Uma das desvantagens desse modelo é a alta dimensão do vetor

[1]Esses termos podem ser traduzidos respectivamente para fusão precoce, fusão tardia e fusão entre mídias. Usaremos estes termos em inglês devido a sua ampla utilização na literatura

de características resultante. Espaços de alta dimensão tendem a dispersar agrupamentos de instâncias pertencentes aos mesmos conceitos [13].

As abordagens late fusion e transmedia/intermedia fusion não agem no nível das características, mas sim no nível da similaridade das modalidades. Assume-se que existem sistemas de recuperação unimodal eficientes e busca-se combinar suas respectivas decisões no lugar de tentar preencher o gap semântico no nível de característica.

A abordagem late fusion se refere à técnica de combinação da saída resultante de diferentes sistemas de recuperação. As características de cada modalidade são armazenadas em sua própria estrutura de indexação. Assim, para cada modelo utilizado calcula-se a similaridade entre os documentos e a consulta. Essas similaridades são combinadas de forma a encontrar um valor de distância único e recuperar imagens mais semelhantes.

No que diz respeito aos métodos de intermedia fusion, o processo age como um mecanismo de pseudo-relevance feedback no lugar de uma simples combinação se comparado ao late fusion. Nesta abordagem a consulta original do usuário é modificada e uma modalidade provê feedback para a outra por meio de expansão de consulta. A ideia principal é inicialmente usar uma das modalidades, por exemplo, evidências visuais, para conseguir imagens semelhantes, e depois usar a outra modalidade, por exemplo, as informações textuais associadas às imagens retornadas com a primeira modalidade, para expandir a consulta.

O método proposto neste trabalho se baseia no modelo intermedia fusion, em que a informação de categoria e descrição textual são extraídas de um ranking visual. Desta forma, um novo ranking baseado em informações textuais é obtido. Em seguida, o modelo late fusion é aplicado para combinar as evidências do ranking visual e do ranking textual para se obter o ranking final.

2.3 Trabalhos Relacionados

Nessa seção, são apresentados trabalhos relacionados a re-ranking visual e ao uso de métodos CBIR no contexto da busca visual de produtos.

2.3.1 Re-ranking Visual

O processo de re-ranking visual tem como foco melhorar a precisão da busca por meio da reordenação de documentos visuais baseando-se em evidências multimodais extraídas de um ranking inicial e de informações auxiliares disponíveis. As informações auxiliares podem ser obtidas a partir de características extraídas de cada documento visual ou da similaridade multimodal entre eles [16]. As características extraídas podem ser informações textuais, propriedades visuais ou a combinação de ambas.

Pesquisas nessa área seguem duas direções distintas: (i) *self-reranking* [32, 34, 1, 24], que extrai informação do resultado inicial da busca para refinar automaticamente o ranking e reordenar os resultados e (ii) *example-reranking* [11, 18], que usa o feedback fornecido por usuários para reordenar os resultados.

Um método não supervisionado de re-ranking que explora a relação entre documentos retornados por sistemas CBIR é apresentado em [32]. A proposta é analisar informação de contexto considerando os k-vizinhos mais próximos para redefinir as distâncias entre esses vizinhos e as outras imagens da coleção. Baseado nas novas distâncias, um novo ranking é calculado de uma forma interativa.

Em [34], é proposto o método *lightweight re-ranking*, que se baseia na similaridade entre resultados de uma busca visual e a dissimilaridade entre os resultados e uma classe externa com imagens variadas. Algumas imagens da classe externa são adicionadas aos resultados da consulta de forma a encontrar o quanto uma imagem está próxima à própria classe e distante da classe externa. A intuição usada considera que resultados relevantes estão visualmente relacionados a outras respostas de uma mesma consulta e resultados irrelevantes estão próximos às imagens da classe externa.

Uma abordagem para recuperar imagens em duas fases é apresentada em [1]. Primeiro, uma consulta é processada a partir de informações textuais apenas, então uma estratégia CBIR é aplicada para realizar o re-ranking dos topo-k resultados. O valor de k é computado dinamicamente por consulta, assegurando que o CBIR será aplicado sobre o melhor subconjunto do primeiro ranking.

Em [48], assume-se que existe uma relação de reforço mútuo entre evidências visuais e textuais que podem ser refletidas no processamento do re-ranking. Dois grafos conectados são gerados respectivamente a partir dos scores visuais e textuais obtidos inicialmente. A partir disso, é feito um passeio randômico no grafo, assumindo que os padrões visuais e

textuais consistentes deverão receber maiores scores.

Um método chamado *crowd-reranking* [24] busca minerar padrões visuais que são relevantes para uma consulta a partir dos resultados de múltiplas máquinas de busca. Dada uma consulta textual, um ranking inicial de documentos visuais é obtido. Ao mesmo tempo, essa consulta alimenta várias máquinas de busca de imagens e vídeo. A partir dos resultados obtidos, é detectado um conjunto de palavras visuais significativas a partir do agrupamento de características visuais locais. Esses padrões são utilizados para realizar a reordenação do ranking inicial.

Um re-ranking interativo é apresentado em [11]. A partir de resultados retornados a partir de uma primeira modalidade (texto), o usuário seleciona uma imagem que esteja próxima ao que procura. A imagem selecionada é categorizada em uma das várias categorias definidas. Para cada categoria, é definido um esquema de ponderação que define um peso específico para combinar características e executar o re-ranking. Este esquema de ponderação é obtido minimizando a perda de posição para todas as imagens de consulta em um conjunto de treinamento.

Em [18], é usada uma estratégia de treinamento para prever a quantidade normalizada de clicks para o topo-k de milhares de imagens retornadas para consultas. As imagens do ranking são reordenadas a partir de uma combinação linear entre a quantidade de clicks previstos e os scores originais.

Assim como apresentado em [1, 48, 11, 24], neste trabalho é explorada a relação entre evidências visuais e textuais para melhorar os resultados da busca por imagens. Estes trabalhos apresentam métodos de re-ranking que dividem a recuperação de imagens em duas fases. Primeiro é usada uma evidência textual para obter um ranking inicial e então padrões visuais obtidos a partir do ranking inicial são usados para realizar o re-ranking. A nossa solução segue a direção inversa. Uma imagem de consulta é usada para a obtenção do ranking inicial. Então, as evidências textuais associadas às imagens retornadas são usadas para realizar a reordenação dos resultados. Nós adotamos a estratégia de self-reranking, uma vez que consideramos uma parte do ranking inicial para detectar padrões para utilizar no re-ranking. Neste trabalho, nós consideramos informação contextual analisando o topo-k dos resultados em relação às imagens de mesma categoria. Nós trabalhamos com a ideia de usar informação de categoria para prever a intenção de busca do usuário, mas diferente do que foi feito em [11], a categoria é detectada automaticamente

sem necessidade de treinamento e interação com o usuário.

2.3.2 Busca Visual de Produtos

O aumento crescente do número de sites de comércio eletrônico gerou a necessidade de criação de sistemas de busca de produtos eficientes. Para melhorar a experiência de compra dos usuários alguns sites incorporam estratégias de CBIR na tarefa de busca de produtos. Sites como Ebay[2] e Amazon[3] disponibilizam alguns recursos de busca visual. Entretanto, a maioria dos sites que fornecem esses recursos se baseia em imagens da própria coleção. Isto é, o usuário de alguma forma indica uma imagem do mesmo site como uma consulta para encontrar produtos relacionados. Isso torna a tarefa de busca mais fácil pois todos os dados relacionados ao produto como nome, descrição e categoria estão disponíveis e podem ser usados para refinar a busca e melhorar os resultados.

A pesquisa feita em [23] apresenta técnicas para a construção de um sistema de busca visual de produtos que inclui: combinação de múltiplas evidências, estratégia de busca em multi-estágios, mecanismo de feedback com o usuário e um método dinâmico de ponderação para melhorar a busca de imagens de baixa qualidade. Para a extração de características, um subconjunto dos descritores visuais do MPEG-7 foram testados e os melhores resultados foram obtidos com o descritor Edge Histogram Shape (EHD) [27].

Em [19], alguns descritores de forma foram usados para recuperação de produtos. Uma estratégia baseada em detecção de arestas é proposta para eliminar a interferência do fundo da imagem e de informações não pertencentes ao produto.

A integração entre evidências visuais e textuais para melhorar o desempenho da busca por imagens de roupas e acessórios é explorada em [10]. O método representa termos textuais em um vetor de características visuais e faz uso de um esquema de ponderação de pesos guiado por texto. Esse esquema infere a intenção do usuário a partir dos termos da consulta e melhora as características visuais que são relevantes para tal intenção.

O foco do trabalho apresentado em [45] está na recuperação de imagens de vestuário em uma coleção de comércio eletrônico. O método proposto suporta recuperação baseada em características por meio de categorias de forma e estilos. A solução faz uso do Shape Context Descriptor [4] cujo processo se divide em segmentação, extração de caracterísi-

[2] *www.ebay.com*. Último acesso em março de 2013.
[3] *www.amazon.com*. Último acesso em março de 2013.

cas e associação de forma.

Em [30], é proposta uma abordagem de recuperação de imagens de roupas. Uma vez construída uma estrutura semântica para conceitos de roupas, as características são extraídas usando o descritor SIFT [25] que representa características que são invariantes à escala, rotação e parcialmente invariantes à mudanças de iluminação.

Nosso objetivo com este trabalho é propor um método para a tarefa de busca visual em uma coleção de imagens de comércio eletrônico, assim como realizado em [15, 45]. Nossa coleção de experimentos consiste em imagens de roupas [10, 15, 44, 30], calçados e acessórios. Nós trabalhamos com múltiplas características por meio da combinação de evidências textuais e visuais, como feito em [10, 23]. Nosso método analisa informações associadas às imagens retornadas para uma consulta de forma a inferir a intenção do usuário. Nós também propomos uma estratégia para eliminar a interferência do fundo da imagem, mas diferente do que foi proposto em [19], adotamos partições fixas adaptadas às características das imagens das coleções. O descritor EHD foi testado como feito em [23], mas o melhor resultado foi obtido pelo descritor CEDD [8].

Capítulo 3

Re-ranking multimodal de busca visual

Neste capítulo, são descritos dois métodos definidos durante o desenvolvimento deste trabalho com o intuito de melhorar a relevância das respostas de uma busca visual de produtos. Ambos os métodos se baseiam em informações de categoria com o objetivo de eliminar ambigüidades da consulta e reordenar o ranking visual original. As principais diferenças entre os dois métodos são: (i) a maneira como a categoria da consulta é estimada e (ii) a forma como a fusão de evidências é utilizada na geração do ranking final.

Inicialmente, apresentamos o *Term and Category-Based Re-ranking (TCat-BR)* [36]. Este método faz um re-ranking utilizando a categoria mais freqüente do topo das respostas de um ranking gerado a partir de técnicas de CBIR e utiliza o resultado da fusão de evidências visuais e textuais extraídas deste ranking para reordenar seus resultados.

Devido algumas limitações observadas no *TCat-BR*, propusemos algumas modificações e as implementamos no método chamado *Term and Category-Weight-Based Re-ranking (TCatW-BR)*, que faz um re-ranking utilizando um peso gerado para cada categoria do topo das respostas do ranking visual e utiliza a fusão de evidências para expandir e complementar este ranking. Estas modificações resultaram no aumento de relevância dos resultados de uma forma geral. Mas, como pode ser verificado na Seção 4.4.3, ainda houve situações em que o primeiro método (*TCat-BR*) obteve melhor desempenho.

3.1 Term and Category-Based Re-ranking (*TCat-BR*)

Nesta seção, apresentamos o método de re-ranking de imagens denominado *Term and Category-Based Re-ranking (TCat-BR)*. As etapas do método são apresentadas no Algo-

ritmo 1. Inicialmente, uma busca visual é realizada a partir de uma imagem de consulta (Q_v). Nesta etapa, um descritor é utilizado para recuperar imagens visualmente similares à imagem de consulta, obtendo assim o ranking inicial denominado *VisualRank* (Algoritmo 1, linha 1). O descritor possui um papel fundamental, pois seu resultado afeta diretamente o desempenho de outras etapas do método, como a estimativa da categoria da consulta e a extração de informações textuais. Os experimentos realizados para escolher o descritor são detalhados na Seção 4.3.

Uma vez que as imagens visualmente similares foram retornadas, o próximo passo é deduzir a categoria da consulta por meio da análise das respostas do topo do ranking da busca visual (Algoritmo 1, linha 2). Uma solução simples utilizada para estimar a categoria da imagem da consulta foi utilizar um algoritmo de classificação k-NN. Foi definido que a categoria da imagem de consulta seria a categoria mais freqüente do topo-k do ranking visual. Os experimentos realizados para definir o valor de k são apresentados na Seção 4.4.1.

A categoria definida é usada para realizar um primeiro re-ranking (Algoritmo 1, linha 3). Este re-ranking é realizado com o objetivo de mover para o topo do ranking as imagens que pertencem à categoria mais freqüente das respostas retornadas. O ranking produzido nessa etapa é referenciado nesse trabalho como *Category-Based Ranking (Cat-BR)*.

Algoritmo 1 Pseudocódigo do método TCatBR

1: $VisualRank \leftarrow visualSearch(Q_v)$;
2: $cat \leftarrow estimateCategoryOfQuery(VisualRank,\ k)$;
3: $CatBR \leftarrow reRankByCategory(VisualRank,\ cat)$;
4: $q_t \leftarrow$ **buildTextualQuery**$(CatBR,\ k)$;
5: $TextualRank \leftarrow textualSearch(q_t)$;
6: $CombinedRank \leftarrow combineRanks(VisualRank,\ TextualRank)$;
7: $TCatBR \leftarrow reRankTCatBR(VisualRank,\ CombinedRank)$;

Uma vez obtido o ranking baseado na categoria mais freqüente, temos como objetivo refinar os resultados em termos de subcategoria do produto e utilizar essa informação para melhorar ainda mais os resultados. Por exemplo, uma vez definido que uma consulta é da categoria "roupas femininas", pretendemos identificar sua subcategoria, ou seja, definir se o produto é uma blusa, um vestido ou uma saia. Outro exemplo, se uma consulta pertencer à categoria "calçados masculinos", identificar se o produto é um sapato social ou um tênis.

Para alcançar esse objetivo, partimos do princípio de que a informação de subcategoria está presente nos termos mais freqüentes da descrição associada às imagens do ranking visual. Então, montamos uma consulta textual formada pela categoria inferida e os termos associados às imagens do topo-k do ranking obtido pelo *Cat-BR* (Algoritmo 1, linha 4). Um esquema detalhado de como funciona o processo de extração das evidências textuais é apresentado no Algoritmo 2. Foram estudadas várias alternativas para extrair palavras-chave da descrição dos produtos. O resultado desse estudo pode ser verificado na Seção 4.4.1. As palavras-chave extraídas da descrição são concatenadas para criar uma consulta textual (q_t).

A consulta textual formada é submetida a um sistema que indexa as descrições textuais dos produtos presentes na coleção. Nessa fase, o modelo adotado para computar a similaridade entre as imagens foi o Modelo de Espaço Vetorial [28]. Como resultado, obtemos um ranking que representa a similaridade textual entre os termos associados à imagem de consulta e os produtos da coleção. Esse ranking é referenciado nesse trabalho como *TextualRank* (Algoritmo 1, linha 5).

Algoritmo 2 Pseudocódigo da função de extração de informações textuais.

```
1: function BUILDTEXTUALQUERY(Rank, k)
2:     TopRank ← top_k of Rank;
3:     q_t ← '';
4:     for (i = 0; i < TopRank.size ; i++)
5:         q_t ← q_t + extractDescription(TopRank[i]);
6:     end for
7:     return q_t;
8: end function
```

A próxima etapa do método é responsável por realizar a fusão dos scores encontrados no *TextualRank* e *VisualRank* (Algoritmo 1, linha 6). Neste trabalho, definimos score como sendo a medida de similaridade atribuída à cada resposta retornada como parte do ranking gerado por um método. Como resultado desta etapa, é gerado um ranking denominado *CombinedRank*. A fusão é feita por meio da função *CombSum* [14], que é um caso particular de fusão de evidências em que os scores de cada modalidade são somados para obter o score final. Sendo i, a imagem, N_j, o número de evidências da imagem a serem combinadas e S_j, o score associado à imagem para a evidência j. Sua

Figura 3.1: Exemplo das etapas geradas pelo *TCat-BR*.

fórmula é definida na Equação 3.1.

$$S_{mixed}(i) = \sum_{j=1}^{N_j} S_j(i) \tag{3.1}$$

A etapa final do método consiste em gerar o ranking final denominado *TCat-BR*. Este ranking é obtido por meio da função *reRankTCatBR* (Algoritmo 1, linha 7) que é responsável por reordenar as imagens presentes no *VisualRank* a partir da posição destas imagens no *CombinedRank*. Nesse caso, os resultados gerados pelo *TextualRank* e *CombinedRank* não são agregados ao resultado final, servindo apenas para reposicionar as imagens relevantes do *VisualRank* nas melhores posições do ranking.

Um exemplo do resultado de cada etapa do método é mostrado na Figura 3.1. Nesse exemplo, a imagem de uma mulher com um vestido vermelho (Figura 3.1-a) é submetida como consulta. O resultado da busca visual, representado pelo *VisualRank*, é apresentado na Figura 3.1-b.

Nesse exemplo, a categoria inferida foi "roupa feminina", devido à sua maior frequência no ranking inicial. Com a informação da categoria estimada, o próximo passo consistiu em reordenar os resultados de forma a colocar as respostas pertencentes à categoria estimada nas posições iniciais do ranking (Figura 3.1-c). Dessa forma, obtivemos o ranking *Cat-BR*.

Seguindo com a execução do método, a etapa de extração de informações textuais

gerou uma consulta onde a palavra “vestido” combinada com a categoria “roupas femininas” teve um peso significativo no resultado do *TextualRank*, que pode ser visualizado na Figura 3.1-d. Assim, a combinação do ranking visual com o textual, fez refletir um posicionamento melhor para imagens visualmente e semanticamente similares à imagem de consulta, como pode ser visualizado na Figura 3.1-e. Vale ressaltar que o *Visual-Rank* (Figura 3.1-b), o *Cat-BR* (Figura 3.1-c) e o *TextualRank* (Figura 3.1-d) são resultados intermediários, não sendo apresentados ao usuário final.

A partir dos resultados apresentados na Seção 4 é possível perceber que o método *TCat-BR* alcança o objetivo proposto neste trabalho, que é o de definir uma solução capaz de obter resultados relevantes para a busca visual de produtos. Quando comparado a outros métodos de busca visual, os resultados são significativamente melhores. Entretanto, após uma avaliação crítica sobre o método TCat-BR foram feitas algumas observações com o objetivo de melhorar ainda mais os resultados obtidos pelo método.

A primeira observação foi com relação a forma como a categoria é estimada. Usar a categoria mais freqüente alcança resultados muito bons quando o método acerta a categoria. Mas se o método erra a categoria, o resultado tende a piorar de forma considerável. Ao usar a categoria mais freqüente, a distribuição das frequências das categorias no resultado não é considerada. Por exemplo, um usuário submete a imagem de uma “camisa feminina”, que resulta em um ranking inicial formado por 12 imagens da categoria “roupas femininas” e 13 imagens da categoria “roupas masculinas”. Nesse caso, a categoria estimada seria “roupas masculinas” e todas as imagens dessa categoria subiriam no ranking. Isso influenciaria diretamente de forma negativa no primeiro re-ranking (*Cat-BR*). E também refletiria no desempenho geral do método, uma vez que subiriam no ranking todas as imagens associadas à categoria de “roupas masculinas”.

Figura 3.2: Exemplo de estimativa incorreta de categoria do *TCat-BR*.

A Figura 3.2, mostra o pior caso do *Cat-BR*, em que o método erra a categoria esti-

mada. A consulta submetida é de uma "bolsa" (Figura 3.2-a). O ranking inicial retornado para a consulta pode ser visualizado na Figura3.2-b. Como a maioria das imagens retornadas pertencem à categoria "calçados femininos", todas as imagens dessa categoria subiram no ranking. Sendo assim, o re-ranking como pode ser visto na Figura 3.2-c, foi bastante prejudicado, resultando em basicamente nenhuma imagem relevante nas posições iniciais do ranking. Por consequência, isso afeta a extração de termos para a busca textual e o resultado final do método de uma forma geral.

Outra observação é o fato que o ranking inicial (*VisualRank*) não tem seus scores modificados em momento algum, nem quando é gerado o primeiro re-ranking (*Cat-BR*). Isso afeta a fusão de evidências, pois a função de combinação é baseada nos scores e não na posição das imagens. Assim, o (*Cat-BR*) não possui papel ativo no aumento da relevância dos resultados. Ele beneficia apenas a extração de boas evidências textuais.

3.2 Term and Category-Weight-Based Re-ranking (*TCatW-BR*)

O método *Term and Category-Weight-Based Re-ranking* (TCatW-BR) foi definido com base na proposta do método *TCat-BR*, visando superar algumas limitações nele observadas. Assim, as modificações realizadas foram feitas a partir da maneira como a categoria da consulta é estimada e da forma como a fusão de evidências é utilizada na geração do ranking final. O *TCatW-BR* deixa de usar a categoria mais freqüente e passa a gerar pesos para cada categoria presente no ranking. Esses pesos são aplicados às imagens do ranking, resultando assim na modificação dos scores de acordo com o peso aplicado. Com relação à fusão de evidências, os resultados da busca visual e da busca textual são combinados e apresentados ao usuário final. Diferente do *TCat-BR* que utiliza a fusão de evidências apenas para reordenar o ranking visual.

As etapas da execução do método são apresentadas no Algoritmo 3. Inicialmente, uma imagem de consulta (Q_v) é submetida e um ranking visual (*VisualRank*) é obtido (Algoritmo 3, linha 1). Nessa etapa, um descritor é utilizado para recuperar imagens visualmente similares à imagem de consulta. Os experimentos realizados para escolher o descritor são detalhados na Seção 4.3.

O próximo passo consiste em estimar um peso para as categorias presentes no topo-*m*

do ranking visual (Algoritmo 3, linha 2). Os experimentos realizados para definir o valor de m são apresentados na Seção 4.4.2.

Algoritmo 3 Pseudocódigo do método TCatW-BR

1: $VisualRank \leftarrow visualSearch(Q_v)$;
2: $ListOfWeights \leftarrow$ **generateWeightByCategory**$(VisualRank,\ m)$;
3: $CatWBR \leftarrow reRankByCategoryWeight(VisualRank,\ ListOfWeights)$;
4: $Q_t \leftarrow$ **buildTextualQuery**$(CatWBR,\ n)$;
5: $TextualRank \leftarrow textualSearch(Q_t)$;
6: $TCatWBR \leftarrow combineRanks(CatWBR,\ TextualRank)$;

Sendo R, o ranking com as imagens (i) e suas respectivas informações de categoria (c) e scores, a fórmula do peso gerado para uma categoria (C) é especificada na Equação 3.2.

$$P(C) = \frac{\sum_{\substack{\forall i \in R, \\ c(i) = C}} score(i)}{\sum_{\forall i \in R} score(i)} \tag{3.2}$$

A função de geração de pesos consiste em percorrer o topo-m do *VisualRank* e calcular o total de scores das imagens por categoria. Em seguida, esse valor é normalizado a partir da razão entre o total de score por categoria e o total de score das imagens do topo-m do ranking. No Algoritmo 4, é apresentada a estratégia de geração de pesos para as categorias.

Uma vez calculados os pesos por categoria, o próximo passo consiste em reordenar o ranking visual usando essas informações como base (Algoritmo 3, linha 3). No Algoritmo 5, é apresentada a estratégia de reordenação do ranking visual baseada no peso gerado para cada categoria. Nesta etapa, os pesos são aplicados às imagens do *Visualrank* de acordo com a categoria a qual elas pertencem. Feito isso, as imagens são reordenadas de acordo com os novos scores e o ranking *CatW-BR* é gerado.

A partir deste momento, as evidências textuais são extraídas de um topo-n dos resultados do *CatW-BR* para montar uma consulta textual (Algoritmo 3, linha 4). Os experimentos realizados para definir o valor de n são apresentados na Seção 4.4.2. Nessa fase, o processo de extração de informações textuais é semelhante ao apresentado no Algoritmo 2. Na busca textual, utilizamos o Modelo de Espaço Vetorial [28] para computar a similaridade entre as imagens. Nesse caso, são consideradas respostas que pertencem às categorias presentes no topo-n do ranking *CatW-BR*. Como resultado, obtemos um ranking que representa a similaridade textual entre os termos associados à imagem de consulta e os produtos da coleção. Esse ranking é referenciado nesse trabalho como *Tex-*

Algoritmo 4 Pseudocódigo da função de geração de pesos por categoria.

```
1: function GENERATEWEIGHTBYCATEGORY(Rank, m, ListOfWeights)
2:     TopRank ← top_m of Rank;
3:     scoreGeral ← 0;
4:     ListOfWeights ← null;
5:     for (i = 0; i < TopRank.size ; i++)
6:         scoreGeral ← scoreGeral + TopRank[i].score;
7:         foundCategory ← false;
8:         for (j = 0; j < ListOfWeights.size ; j++)
9:             if (ListOfWeights[j].category == TopRank[i].category) then
10:                ListOfWeights[j].weight = ListOfWeights[j].weight +
   TopRank[i].score;
11:                foundCategory ← true;
12:                break;
13:            end if
14:        end for
15:        if (foundCategory == false) then
16:            addWeight(TopRank[i].category, TopRank[i].score, ListOfWeights);
17:        end if
18:    end for
19:    ListOfWeights ← normalizeWeight(scoreGeral, ListOfWeights);
20:    return ListOfWeights;
21: end function
```

Algoritmo 5 Pseudocódigo da função de re-ranking a partir dos pesos por categoria.

```
1: function RERANKBYCATEGORYWEIGHT(Rank, ListOfWeights)
2:     for (i = 0; i < Rank.size ; i++)
3:         foundCategory ← false;
4:         for (j = 0; j < ListOfWeights.size ; j++)
5:             if (Rank[i].category == ListOfWeights[j].category) then
6:                 Rank[i].score = Rank[i].score * ListOfWeights[j].weigth;
7:                 foundCategory ← true;
8:                 break;
9:             end if
10:        end for
11:        if (foundCategory == false) then
12:            Rank[i].score = 0;
13:        end if
14:    end for
15:    Rank ← reorder(Rank);
16:    return Rank;
17: end function
```

tualRank (Algoritmo 3, linha 5).

A última etapa do método consiste em expandir e complementar o resultado obtido pelo *CatW-BR* baseando-se na combinação desse ranking com o resultado obtido pelo *TextualRank*. Para isso, uma função de combinação linear é utilizada para realizar a fusão de evidências. Sendo S_t e S_v respectivamente o score textual e o score visual de uma imagem (i), a função de combinação utilizada é definida na Equação 3.3:

$$S_{mixed}(i) = \alpha\, S_t(i) + (1-\alpha)\, S_v(i) \tag{3.3}$$

O valor de α determina o peso que será atribuído para cada evidência na combinação. Os experimentos realizados para verificar a influência do peso definido para cada evidência são apresentados na Seção 4.4.2. Assim, os rankings *CatW-BR* e *TextualRank* são combinados de forma a obter o ranking final denominado *TCatW-BR* (Algoritmo 3, linha 6).

A Figura 3.3 apresenta o resultado da aplicação do método *TCatW-BR*. Nesse exemplo, a imagem de uma mulher com um vestido vermelho (Figura 3.1-a) é submetida como consulta. Um ranking inicial (Figura 3.3-b) é gerado com base nas características visuais da imagem. A próxima etapa consiste na geração de pesos para as categorias presentes nesse ranking. No exemplo, existem duas categorias: "roupas femininas" e "roupas masculinas". Nesse caso, um peso maior é atribuído à categoria "roupas femininas", tendo em vista a posição e a quantidade das imagens dessa categoria no ranking inicial, e um peso menor é atribuído à categoria "roupas masculinas".

Assim, obtemos o ranking *CatW-BR* (Figura 3.3-c), gerado a partir da aplicação dos pesos sobre os scores das imagens. Nele, a maioria das imagens pertencem à categoria "roupas femininas", mas podemos observar que existe uma pequena parcela de imagens que pertencem à categoria "roupas masculinas". Prosseguindo com a execução do método, os termos textuais são extraídos da descrição das imagens pertencentes ao topo do ranking *CatW-BR* e uma consulta textual é gerada. Nesse momento, é aplicada a recuperação textual que obtém o *TextualRank* (Figura 3.3-d). Nesse exemplo, o termo "vestido" obteve um peso maior no resultado. O ranking final gerado pelo método, consiste em combinar os scores dos rankings *CatW-BR* (Figura 3.3-c) e *TextualRank* (Figura 3.3-d). Assim, o ranking final é formado pela combinação desses dois rankings, posicionando as imagens mais relevantes nas melhores posições, como pode ser observado na Figura 3.3-e.

Figura 3.3: Exemplo das etapas geradas pelo *TCatW-BR*.

Figura 3.4: Aplicação do *CatW-BR* no pior caso do *TCat-BR*.

A Figura 3.4-c apresenta o resultado da obtenção do *CatW-BR* para o pior caso do *Cat-BR*, como foi apresentado na Figura 3.2-c. É possível observar que o *CatW-BR* gera um resultado mais justo e equilibrado. Isso acontece pois seu re-ranking leva em consideração a distribuição das categorias e os scores dos resultados do ranking.

Capítulo 4

Experimentos

Este trabalho tem como foco resolver o problema da busca visual em bases de produtos. Sendo assim, o objetivo da realização dos experimentos foi comparar o desempenho de descritores visuais sobre coleções de produtos e depois compará-los as nossas propostas de solução. As avaliações foram realizadas utilizando avaliadores reais, simulando o mais próximo da realidade qual seria o comportamento dos descritores caso eles fossem inseridos em um ambiente de busca real. Para comprovar o desempenho do nosso método utilizamos algumas medidas de avaliação comumente utilizadas na área de recuperação de imagens e testes estatísticos para demonstrar que os ganhos obtidos foram realmente significativos.

Começamos este capítulo com a apresentação das métricas de avaliação utilizadas. Em seguida, apresentamos as coleções sobre as quais os experimentos foram realizados e mostramos como o baseline foi escolhido. Por fim, apresentamos e discutimos os resultados obtidos a partir da comparação entre o baseline e as soluções propostas.

4.1 Métricas de Avaliação

Métodos de recuperação de imagens são avaliados em termos de eficácia com o objetivo de medir a capacidade de recuperar respostas relevantes. Os resultados dos experimentos realizados neste trabalho foram obtidos por meio de métricas como: Precisão na *n*-ésima posição (P@N) [3] e MAP (*Mean Average Precision*) [3]. Para a validação dos resultados obtidos, utilizamos o teste de validação Wilcoxon Matched-Pairs Signed-Ranks [46].

4.1.1 Precisão na N-ésima posição do ranking - P@N

Precisão é uma métrica de avaliação bastante utilizada, cujo cálculo consiste na razão entre o número de documentos relevantes retornados e o número total de documentos retornados para uma consulta. A precisão pode ser obtida quando existe um conjunto conhecido de relevantes para cada consulta avaliada. Nos cenários nos quais não existe um conjunto de relevantes previamente definido, é necessário obtê-lo por meio da avaliação de usuários reais. Esse tipo de estratégia tem a vantagem de conseguir avaliações que se aproximam da opinião de usuários potenciais do sistema.

Algumas bases possuem milhares de imagens, o que torna inviável que usuários consigam classificar todos os resultados de uma consulta. Para esses casos, a eficácia é calculada por meio da medida P@N [3]. Esta medida baseia-se na definição de um limite N de resultados que serão avaliados e assim a precisão é obtida até uma posição N do ranking.

Para uma determinada consulta, a precisão dos resultados do topo N de um ranking pode ser calculada por meio da Equação 4.1, sendo $\mid rel_N \mid$ o número de imagens relevantes no topo N dos resultados.

$$P@N = \frac{\mid rel_N \mid}{N} \tag{4.1}$$

Nesta dissertação, foram utilizadas as medidas P@10 e P@20. Ou seja, foram avaliadas respectivamente 10 e 20 imagens do topo do ranking de imagens retornadas. Considerando o ambiente de sites de comércio eletrônico, apenas uma parcela dos resultados são apresentadas ao usuário. E neste caso, é fundamental que as imagens relevantes estejam nas primeiras posições do ranking.

4.1.2 Mean Average Precision - MAP

MAP [3] é uma métrica popular de avaliação que resulta em uma medida única usada para avaliar sistemas de recuperação de informação. Seu cálculo é baseado na média das precisões obtidas para cada documento relevante recuperado para uma consulta. Esta métrica possui como denominador o número total de documentos relevantes por consulta, ou seja, evita que seja feito o corte de apenas os documentos relevantes que foram recuperados.

A fórmula de MAP é apresentada na Equação 4.2, sendo k o número total de consultas

e P_q a precisão média para a consulta q.

$$MAP = \frac{1}{k}\sum_{q=1}^{k} P_k \tag{4.2}$$

P_q é definida pela Equação 4.3, onde m é o número de documentos recuperados para a consulta q, n é o número de documentos relevantes para a consulta q e r_{q_i} é a função binária que indica quando o documento da posição i é relevante ou não para a consulta q.

$$P_q = \frac{1}{n}\left(\sum_{i=1}^{m} r_{q_i} \times \frac{1}{i}\sum_{j=1}^{i} r_{q_j}\right) \tag{4.3}$$

4.1.3 Teste de validação Wilcoxon Matched-Pairs Signed-Ranks

O teste de validação Wilcoxon [46] é utilizado com o propósito de garantir que os resultados obtidos de uma comparação possuam diferenças estatisticamente significativas. Esse teste é sugerido quando: (i) os dados comparados possuem algum relacionamento entre si; (ii) não existe certeza sobre a distribuição de probabilidade dos valores obtidos; e (iii) os avaliadores são selecionados de forma aleatória.

Por atender a esses requisitos, o teste Wilcoxon foi aplicado sobre os resultados dos experimentos realizados neste trabalho. Foram considerados ganhos significativos de um método sobre outro, aqueles cujos valores foram iguais ou superiores a 95%.

4.2 Base de Imagens

As coleções de imagens de produtos encontradas na literatura para realizar experimentos com métodos de busca visual, como *Stanford Mobile Visual* [6] e *PI 100* [47], normalmente são compostas em sua maioria por imagens de capa de livros, CD/DVD ou imagens sem qualquer informação textual associada.

Devido à falta de coleções disponíveis para serem aplicadas no escopo desse trabalho, foram montadas duas coleções de imagens extraídas de três sites de comércio eletrônico. As imagens extraídas pertencem à seis categorias, definidas como: *roupas femininas*, *roupas masculinas*, *calçados femininos*, *calçados masculinos*, *bolsas* e *acessórios*. Cada coleção contém as imagens com suas respectivas informações de categoria e descrição textual do produto. A primeira coleção denominada *DafitiPosthaus* inclui 23.154 ima-

gens coletadas dos sites Dafiti[1] e Posthaus[2], de duas lojas de moda populares no Brasil. A segunda coleção, denominada *Amazon* contém 12.807 imagens coletadas do site da Amazon[3], uma loja on-line de compras mundialmente conhecida.

Para a realização dos experimentos foi utilizado um total de 200 consultas divididas em três conjuntos como a seguir:

Conjunto 1 (Q1): composto por 50 imagens selecionadas da coleção *DafitiPosthaus* e 50 imagens selecionadas da coleção *Amazon*. Nesse caso, as imagens de consulta estão presentes na base. Consequentemente, a categoria da imagem já é conhecida.

Conjunto 2 (Q2): composto por 50 imagens selecionadas de sites de comércio eletrônico que não estão presentes nas coleções utilizadas nesse trabalho. As imagens nesse caso possuem características visuais similares às imagens contidas nas coleções, ou seja, imagens de produtos de moda com fundo homogêneo.

Conjunto 3 (Q3): composto por 50 imagens selecionadas de diferentes sites como blogs, revistas e jornais. As imagens são, em geral, fotos de pessoas famosas e representam uma classe de consultas difíceis, porém relevantes para usuários que estão em busca de um produto similar. Nesse conjunto de consultas, as imagens apresentam o fundo com bastante ruído.

De forma a avaliar a relevância das respostas retornadas por cada método, selecionamos 30 voluntários, divididos em 10 grupos de três pessoas, para fornecer um julgamento binário de relevância (relevante ou não relevante) para cada resposta da consulta. Foram consideradas como relevantes, as respostas que receberam classificação relevante por, pelo menos, dois utilizadores.

4.3 Definição do Descritor

Sistemas CBIR são utilizados para oferecer suporte à recuperação de imagens levando em consideração características de baixo nível como cor, forma e textura. Seu principal objetivo é recuperar imagens similares à uma imagem de consulta. Internamente, um sistema CBIR baseia-se no conceito de descritor de imagem que é responsável por extrair características visuais e codificá-las dentro de um vetor de características. Muitos descritores

[1] http://www.dafiti.com.br. Último acesso em março de 2013.
[2] http://www.posthaus.com.br. Último acesso em março de 2013.
[3] http://www.amazon.com. Último acesso em março de 2013.

são apresentados na literatura com seus pontos fortes e fracos. A escolha de um descritor afeta de forma crítica o desempenho geral de um sistema CBIR.

Tabela 4.1: Desempenho de descritores na coleção *DafitiPosthaus* sem partição. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,466*	0,434*	0,158*	0,416*	0,358*	0,128*	0,040*	0,034*	0,013*
BIC	0,388	0,356	0,117	0,266	0,243	0,074	0,006	0,007	0,003
FCTH	0,366	0,330	0,111	0,332	0,277	0,084	0,014	0,013	0,004
EHD	0,282	0,282	0,059	0,178	0,161	0,033	0,010	0,013	0,003
ACC	0,336	0,306	0,098	0,242	0,221	0,060	0,034	0,032	0,012

Tabela 4.2: Desempenho de descritores na coleção *Amazon* sem partição. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,484*	0,455*	0,175*	0,258*	0,218*	0,135*	0,054	0,055*	0,032*
BIC	0,444	0,393	0,146	0,174	0,141	0,081	0,046	0,042	0,025
FCTH	0,434	0,408	0,153	0,182	0,173	0,082	0,050	0,040	0,020
EHD	0,250	0,237	0,050	0,048	0,042	0,019	0,022	0,021	0,007
ACC	0,360	0,333	0,109	0,138	0,121	0,044	0,064*	0,055*	0,031

Devido à influência do descritor de imagens na qualidade do resultado da busca visual, neste trabalho foram feitos experimentos com vários descritores de forma a escolher o mais adequado para ser utilizado em nosso método. Foram incluídos em nossos experimentos descritores disponíveis na LIRE [26], que é uma biblioteca de código aberto que fornece o estado-da-arte em termos de descritores CBIR. Foram avaliados os seguintes descritores: ACC [17], CEDD [8], EHD [7] e FCTH [9]. Avaliamos também o descritor BIC [38], que alcançou resultados competitivos em trabalhos anteriormente apresentados na literatura ([33, 21]). Os resultados desses experimentos são apresentados nas Tabelas 4.1 e 4.2.

Foi realizado um estudo para verificar a influência do fundo da imagem em nossa solução de busca visual, considerando a possibilidade de remover o ruído do fundo. Para evitar um grande impacto no tempo de processamento das consultas, foi adotada uma estratégia simples mas efetiva para capturar o objeto de interesse das imagens. A estratégia consiste em definir uma partição fixa para as imagens da coleção. Os experimentos real-

izados demonstraram que uma partição de 30% do tamanho original da imagem tende a capturar seu objeto de interesse, que normalmente está posicionado no centro da imagem.

Figura 4.1: Exemplo de partição fixa de imagens das coleções.

A Figura 4.1 apresenta alguns exemplos da partição fixa aplicada às imagens da base. Os resultados apresentados nas Tabelas 4.3 e 4.4 indicam que a estratégia de particionamento alcançou melhores resultados em todos os conjuntos de consulta quando comparado à utilização da imagem inteira.

Como pode ser verificado, o CEDD foi o descritor que obteve melhores resultados nas duas coleções. Esse descritor extrai informações de cor e textura e as incorpora em um vetor de características limitado a 54 bytes por imagem, o que o torna um descritor adequado para grandes coleções de imagens. Como conclusão, nosso baseline de CBIR usa o descritor CEDD e indexa somente a partição central das imagens.

Tabela 4.3: Desempenho de descritores na coleção *DafitiPosthaus* com partição. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,590*	0,560*	0,235*	0,472*	0,448*	0,200*	0,242*	0,226*	0,132*
BIC	0,428	0,393	0,137	0,294	0,274	0,087	0,140	0,128	0,079
FCTH	0,430	0,399	0,146	0,384	0,332	0,120	0,140	0,131	0,079
EHD	0,340	0,314	0,078	0,170	0,171	0,039	0,028	0,026	0,007
ACC	0,358	0,308	0,096	0,238	0,222	0,059	0,072	0,078	0,033

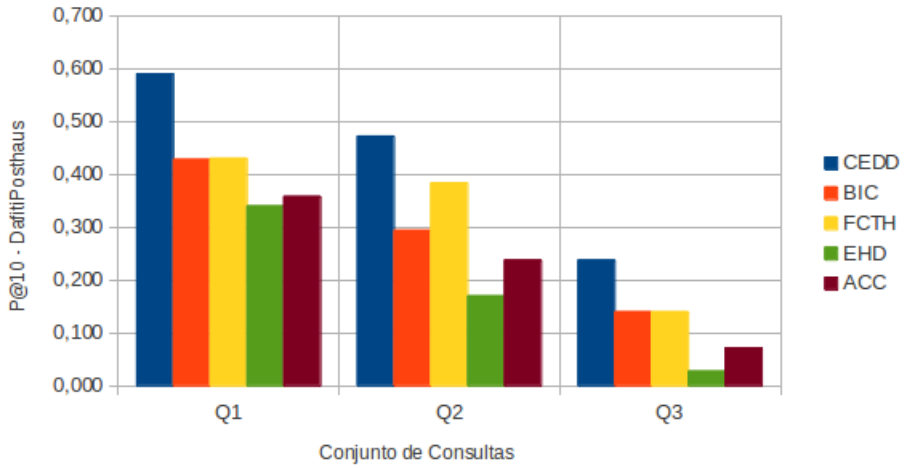

Figura 4.2: Desempenho de descritores na coleção DafitiPosthaus com partição - P@10.

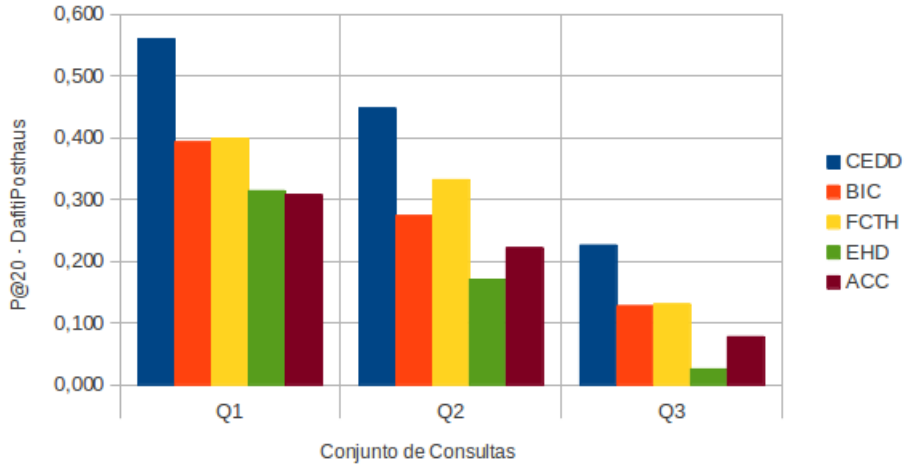

Figura 4.3: Desempenho de descritores na coleção DafitiPosthaus com partição - P@20.

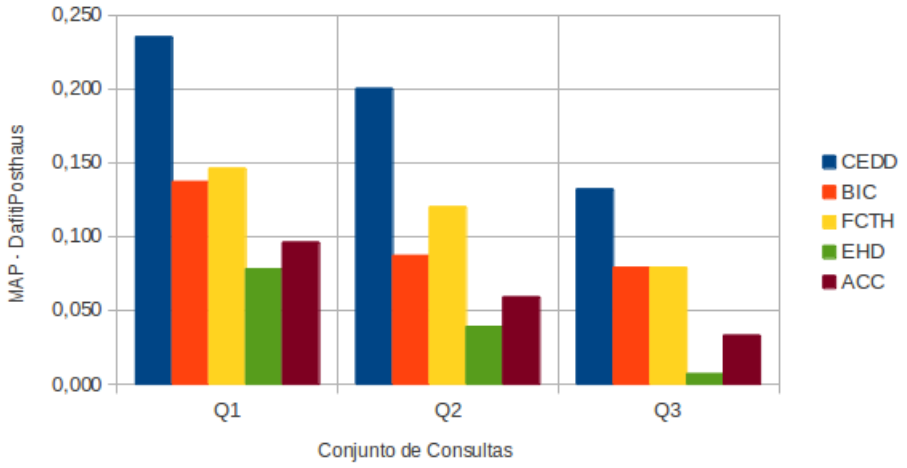

Figura 4.4: Desempenho de descritores na coleção DafitiPosthaus com partição - MAP.

Tabela 4.4: Desempenho de descritores na coleção *Amazon* com partição. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,522*	0,507*	0,232*	0,310*	0,306*	0,196*	0,242*	0,231*	0,174*
BIC	0,412	0,368	0,141	0,150	0,146	0,063	0,194	0,167	0,112
FCTH	0,434	0,398	0,157	0,196	0,186	0,105	0,198	0,169	0,137
EHD	0,220	0,212	0,046	0,068	0,059	0,018	0,044	0,046	0,017
ACC	0,410	0,361	0,124	0,192	0,156	0,059	0,126	0,111	0,072

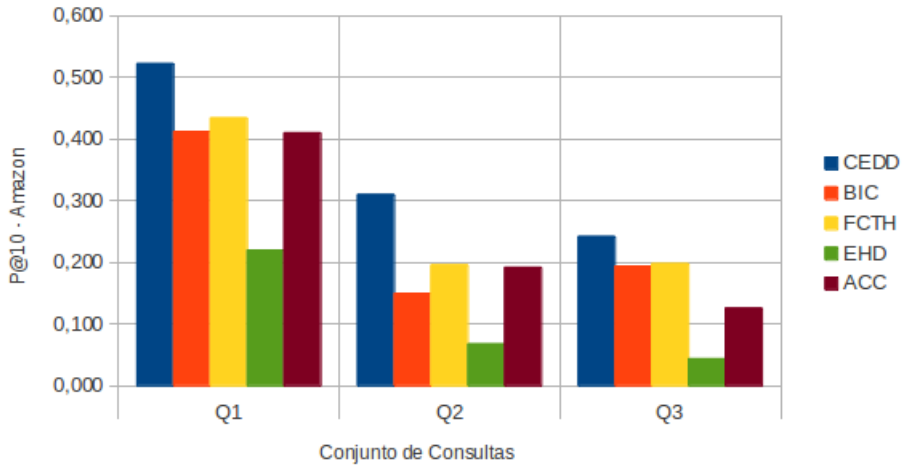

Figura 4.5: Desempenho de descritores na coleção Amazon com partição - P@10.

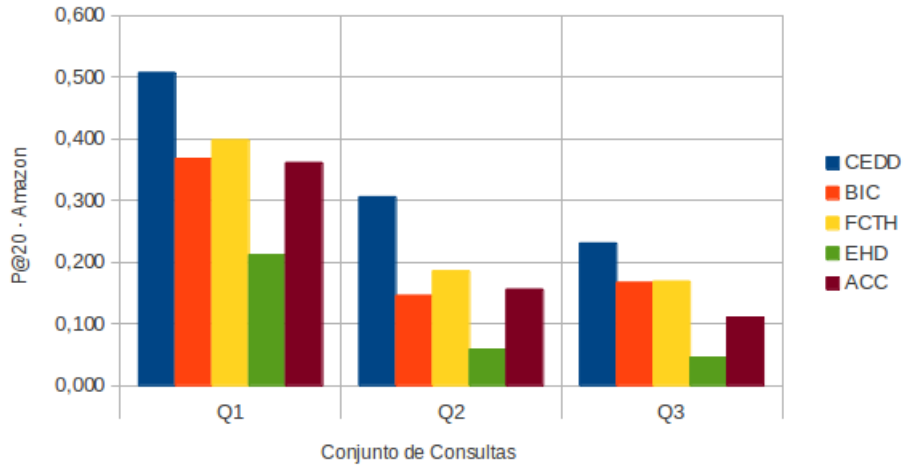

Figura 4.6: Desempenho de descritores na coleção Amazon com partição - P@20.

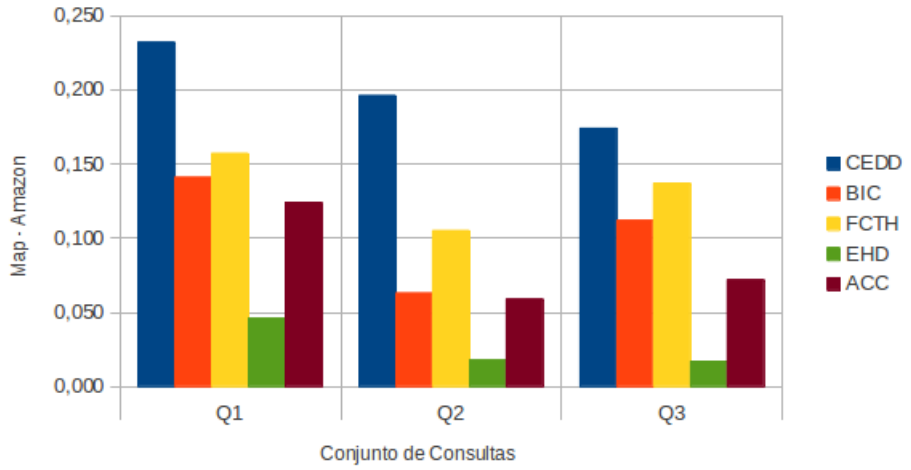

Figura 4.7: Desempenho de descritores na coleção Amazon com partição - MAP.

4.4 Resultados

Nesta seção, são apresentados e discutidos os resultados dos experimentos realizados durante a definição dos métodos *TCat-BR* e *TCatW-BR*.

4.4.1 Análise do *TCat-BR*

A primeira etapa do método *TCat-BR* consiste na geração do ranking visual. Em nossos experimentos, o ranking visual foi obtido por meio do descritor CEDD. Uma vez obtido

o ranking visual, a próxima etapa consiste em computar o *Cat-BR*. Este ranking é o resultado da reordenação do ranking visual a partir da informação de categoria estimada para a consulta. Esta categoria é estimada a partir da análise do topo-k do ranking visual. Assim, a categoria mais freqüente desse topo é definida como sendo a categoria da consulta.

Foram testadas algumas variações do parâmetro k conforme apresentado na Tabela 4.5. A intenção foi escolher um valor que resultasse em uma maior acurácia no acerto da categoria estimada para as consultas. A visualização gráfica dos resultados pode ser vista nas Figuras 4.8 e 4.9.

Tabela 4.5: Variação do topo-k em termos de acurácia de categorização. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$k = 5$	0,96	0,86	0,56	0,88	0,78	0,52
$k = 10$	0,94	0,88	0,60	0,80	0,78	0,52
$k = 15$	0,96	0,94	0,60	0,74	0,80	0,52
$k = 20$	0,96	0,96	0,62	0,74	0,78	0,54
$k = 25$	0,96	0,96	0,62*	0,76	0,78	0,48*
$k = 30$	0,96	0,96	0,62	0,76	0,74	0,48
$k = 35$	0,94	0,96	0,60	0,74	0,72	0,46
$k = 40$	0,96	0,96	0,62	0,74	0,72	0,46
$k = 45$	0,96	0,96	0,66	0,74	0,72	0,44
$k = 50$	0,96	0,96	0,68	0,76	0,72	0,42

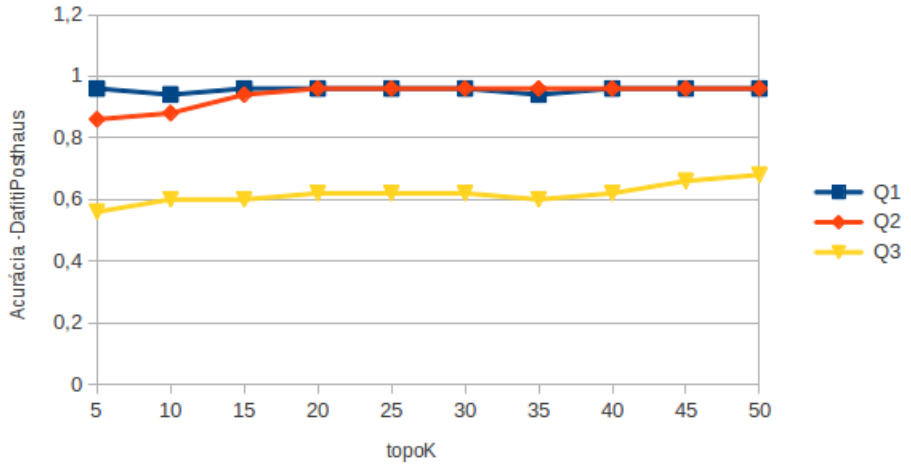

Figura 4.8: Variação do topo-k em termos de acurácia de categorização - *DafitiPosthaus*.

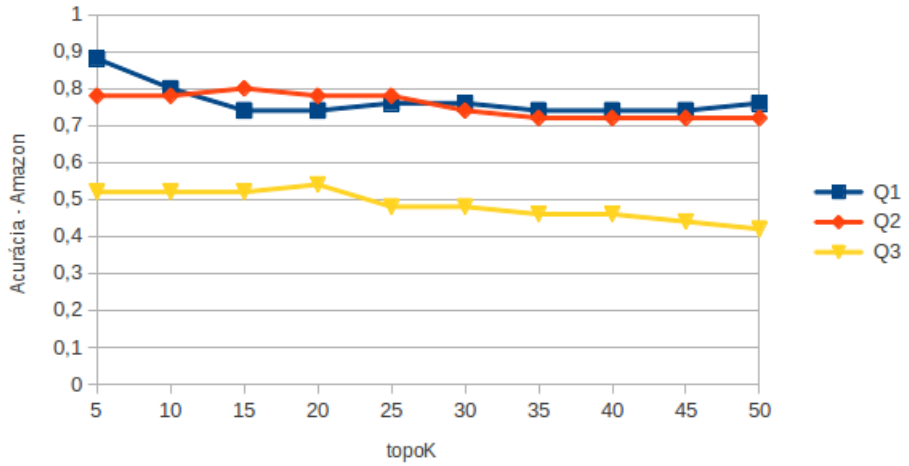

Figura 4.9: Variação do topo-k em termos de acurácia de categorização - *Amazon*.

Para imagens consideradas difíceis ($Q3$), a acurácia do nosso método de classificação (baseado no k-NN) alcança resultados que variam entre 48% e 62%. Neste ponto dos experimentos, foi possível constatar que a evidência de categoria quando utilizada no re-ranking contribuiu de forma significativa para o aumento da precisão dos resultados da consulta visual. No cenário das consultas difíceis, o re-ranking baseado na categoria estimada da consulta, resultou em um aumento, em termos de MAP, de $31,81\%$ na coleção *DafitiPosthaus* e $38,50\%$ na coleção *Amazon*. A partir disso, fizemos outros experimentos a partir dos quais concluímos que uma solução ótima para o problema de estimativa de categoria neste mesmo cenário, poderia alcançar ganhos entre $56,06\%$ e $68,96\%$. Esses resultados mostram que podemos providenciar para o futuro um estudo para encontrar melhores métodos de classificação para determinar a categoria da consulta.

Após a geração do ranking *Cat-BR*, o próximo passo é extrair evidências textuais para gerar uma consulta textual. Extraímos evidências das imagens presentes no topo-k do *Cat-BR*, considerando $k = 25$. Foram testadas três variações da função de extração de termos da descrição dos produtos (Algoritmo 2, linha 5). Essas variações consistem em extrair somente o primeiro/último termo da descrição (*TCat-BR1*), extrair os três primeiros/últimos termos da descrição (*TCat-BR3*) e extrair todos os termos da descrição (*TCat-BRALL*). Pretendemos com isso definir a melhor estratégia para obter termos que especifiquem o tipo de produto que a imagem está representando. Em nossos experimentos, foi verificado que obtemos melhor desempenho quando os substantivos que descrevem os produtos, como "vestido", "short" ou "calçado", são incluídos em nossa consulta textual.

No caso do *TCatBR1* e *TCatBR3*, são considerados os primeiros termos na coleção com descrição textual em português (*DafitiPosthaus*) e os últimos termos na coleção com descrição textual em inglês (*Amazon*). Isto é devido a forma como os produtos são descritos em cada linguagem. Em português, os substantivos que determinam o que é um produto normalmente estão presentes no início da descrição, por exemplo: " ***vestido*** *preto*", "***sandália*** *dourada*" e "***camisa*** *pólo azul*". Em inglês, estes substantivos são encontrados no final da descrição, por exemplo: "*black* ***dress***", "*golden* ***sandal***" e "*blue polo* ***shirt*** ".

As Tabelas 4.6 e 4.7 apresentam os resultados do baseline, *Cat-BR* e as variações do *TCat-BR* em termos de P@10, P@20 e MAP. Como pode ser visto, as variações do *TCat-BR* superaram nosso baseline nos três cenários de consulta e nas duas coleções. As diferenças das variações do *TCatBR* são todas estatisticamente significativas quando comparadas ao *CEDD*, o ranking visual original. Foi aplicado o teste estatístico Wilcoxon considerando somente valores maiores que 95% de confiança. Os ganhos foram expressivos em todos os casos e métricas. Uma visão gráfica dos resultados pode ser observada nas Figuras 4.10, 4.11, 4.12, 4.13, 4.14, 4.15.

Tabela 4.6: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus*. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,590	0,560	0,235	0,472	0,448	0,200	0,242	0,226	0,132
CAT-BR ($k = 25$)	0,668	0,649	0,274	0,588	0,552	0,243	0,326	0,291	0,174
TCAT-BR1	0,688	0,673	0,288	0,630*	0,589*	0,259*	0,300	0,269	0,171
TCAT-BR3	0,700*	0,688*	0,297*	0,628	0,572	0,252	0,320	0,291*	0,183*
TCAT-BRALL	0,692	0,665	0,286	0,614	0,564	0,247	0,330*	0,281	0,181

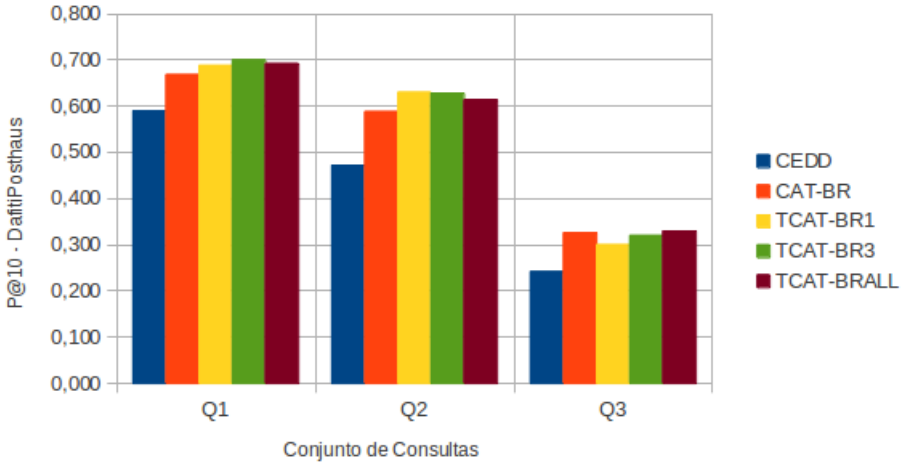

Figura 4.10: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - P@10.

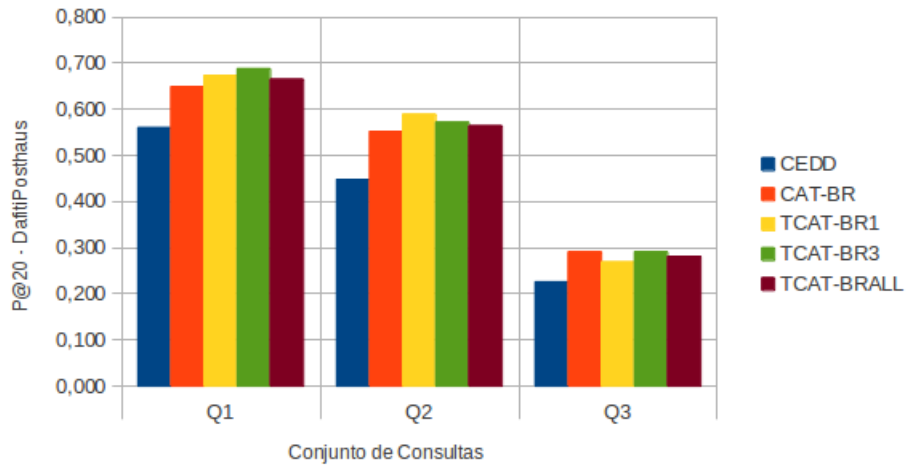

Figura 4.11: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - P@20.

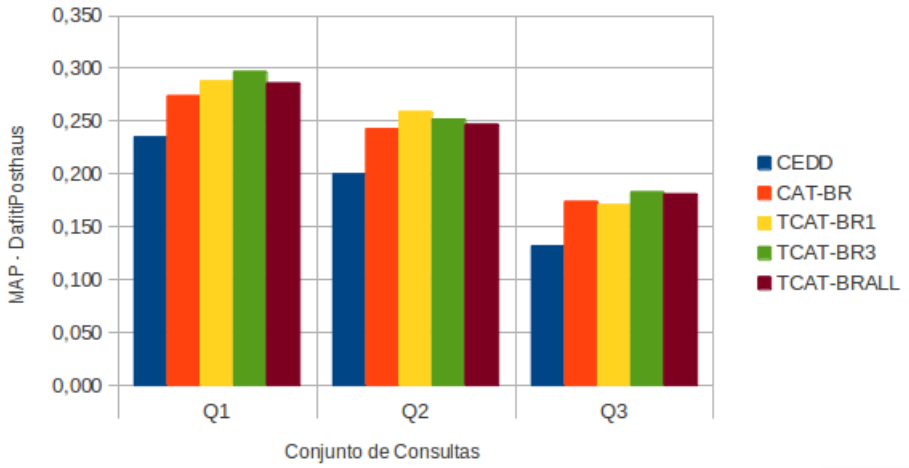

Figura 4.12: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - MAP.

Tabela 4.7: Comparação entre baseline e *TCat-BR* - *Amazon*. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,522	0,507	0,232	0,310	0,306	0,196	0,242	0,231	0,174
CAT-BR $k = 25$	0,548	0,544	0,247	0,388	0,362	0,238*	0,350*	0,311*	0,241
TCAT-BR1	0,580	0,587*	0,265*	0,402*	0,364	0,238*	0,324	0,281	0,222
TCAT-BR3	0,584*	0,568	0,263	0,398	0,367*	0,237	0,334	0,298	0,242*
TCAT-BRALL	0,582	0,566	0,260	0,396	0,357	0,228	0,320	0,289	0,237

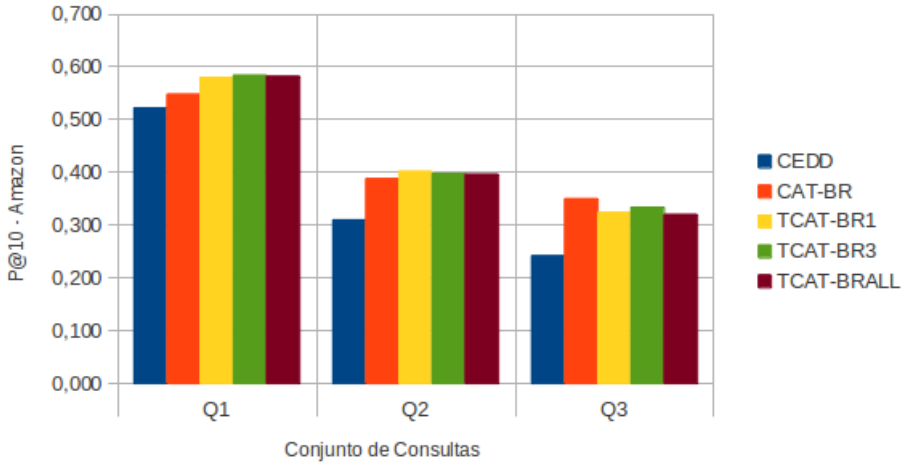

Figura 4.13: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - P@10.

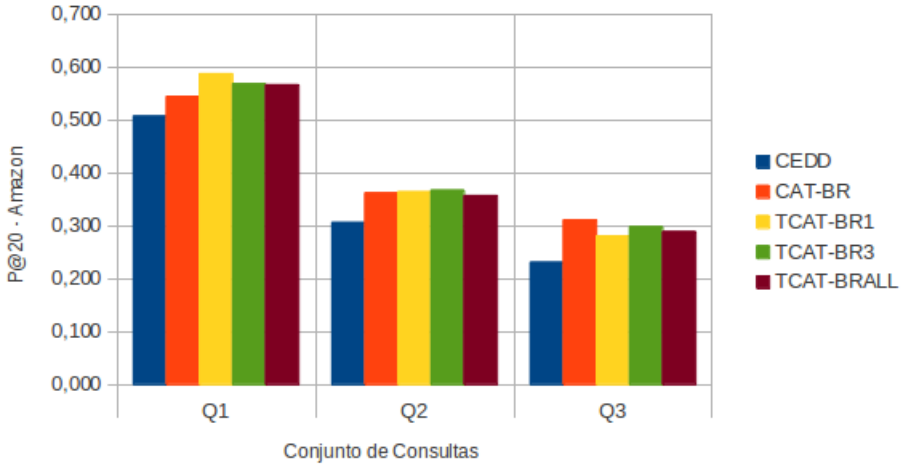

Figura 4.14: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - P@20.

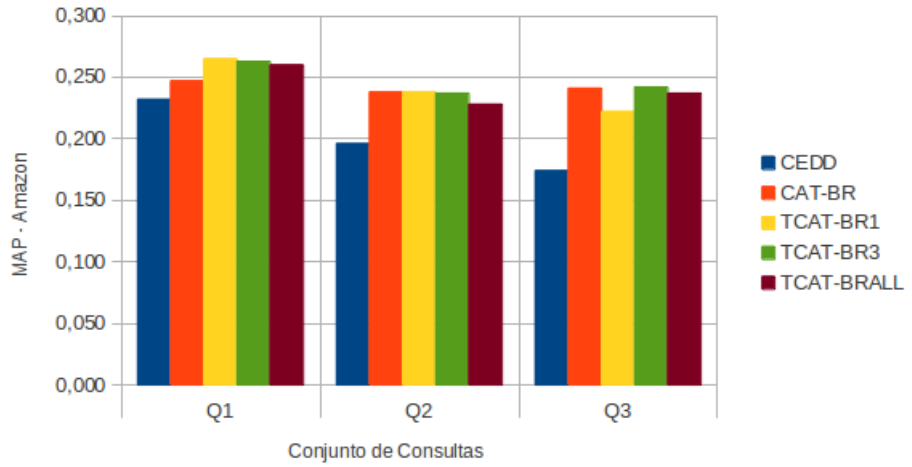

Figura 4.15: Comparação entre baseline e *TCat-BR* - *DafitiPosthaus* - MAP.

O resultado gerado pelo *Cat-BR* não é apresentado ao usuário final por ser um passo intermediário do nosso método. Entretanto, foi observado que já nessa etapa o *Cat-BR* obteve ganhos nos três conjuntos de consultas das duas coleções. A partir desse resultado é possível concluir que a informação de categoria é uma evidência importante que sozinha ajuda a melhorar de forma considerável a relevância dos resultados da busca visual de produtos. Para o nosso método, o bom resultado do *Cat-BR* garante uma qualidade maior dos termos extraídos para a montagem da consulta textual.

As variações do método *TCat-BR* não resultaram em diferenças significativas. Mas podemos observar que as estratégias *TCat-BR1* e *TCat-BR3* alcançaram resultados ligeira-

mente superiores ao *TCat-BRALL*. Como nosso objetivo com o ranking textual não é agregá-lo ao resultado final mas usá-lo para refinar os resultados em termos de subcategoria, acreditamos que os três primeiros/últimos termos são suficientes para alcançar esse propósito. Mas ainda assim, houve consultas cujo resultado da busca textual foi prejudicado devido à presença de termos não relevantes, como informações referentes a código, tamanho e marca do produto. Como conclusão, podemos dizer que um método de extração de termos mais sofisticado do que pegar somente os primeiros ou os últimos termos pode ser estudado em um trabalho futuro.

Ao analisar o desempenho geral do nosso método, percebemos que a baixa qualidade dos resultados fornecidos pelo ranking visual inicial, particularmente no cenário $Q3$, afeta consideravelmente o resultado final gerado pelo método. A função do nosso método é a de reordenar os resultados do ranking visual. Assim, se o ranking visual for muito ruim, o nosso método fica limitado e não consegue aumentar de forma significativa a precisão do resultado final. Nesse caso, seria necessário realizar um estudo para definir uma estratégia de segmentação mais eficaz e também testar o desempenho do método com outros descritores capazes de obter melhores resultados no cenário das consultas difíceis.

4.4.2 Análise do *TCatW-BR*

Neste método, a obtenção do ranking visual seguiu a mesma configuração utilizada no *TCat-BR*, ou seja, utilizamos o descritor CEDD para obter imagens visualmente semelhantes à imagem de consulta. Com o ranking visual gerado, seguimos para a próxima etapa do método, que consiste na reordenação do ranking visual a partir de pesos gerados para as categorias. Assim, o primeiro experimento dessa etapa foi a definição do topo-m do ranking visual a ser considerado para a geração dos pesos.

A Tabela 4.8 mostra, em termos de P@10, como a escolha da quantidade de imagens consideradas para a geração dos pesos afeta o ranking obtido pelo *CatW-BR*. Uma visão gráfica mais detalhada dos resultados pode ser observada nas Figuras 4.16 e 4.17. Com base nos resultados, escolhemos para a geração dos pesos para as categorias o topo-m, com $m = 25$.

Tabela 4.8: Variação do topo-*m* em termos de P@10 do *CatW-BR*. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$m = 5$	0,668*	0,550	0,310*	0,610*	0,394*	0,320
$m = 10$	0,662	0,564	0,296	0,580	0,380	0,332
$m = 15$	0,662	0,568	0,306	0,570	0,390	0,340
$m = 20$	0,662	0,566	0,308	0,562	0,386	0,340
$m = 25$	0,662	0,572	0,304	0,562	0,382	0,340
$m = 30$	0,656	0,570	0,302	0,566	0,380	0,342
$m = 35$	0,654	0,576	0,296	0,564	0,376	0,346
$m = 40$	0,656	0,576	0,294	0,560	0,376	0,346
$m = 45$	0,656	0,578	0,292	0,564	0,374	0,346
$m = 50$	0,646	0,582*	0,242	0,556	0,368	0,348*

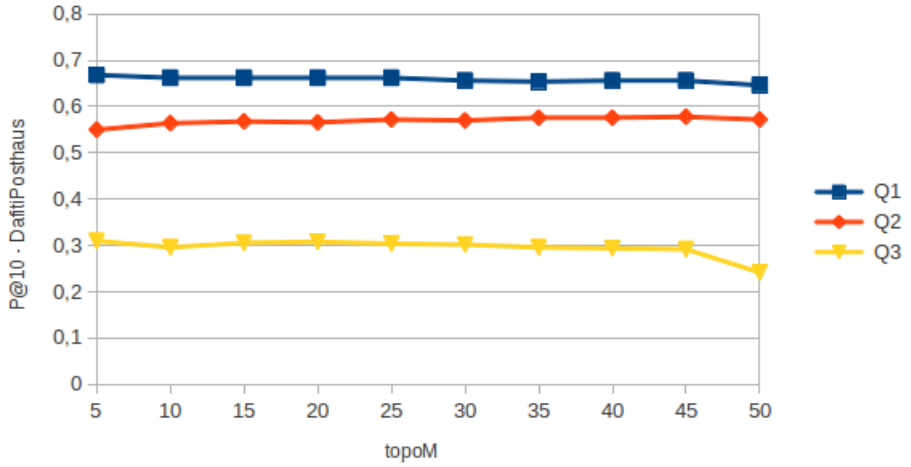

Figura 4.16: Variação do topo-*m* em termos de P@10 do *CatW-BR* - *DafitiPosthaus*.

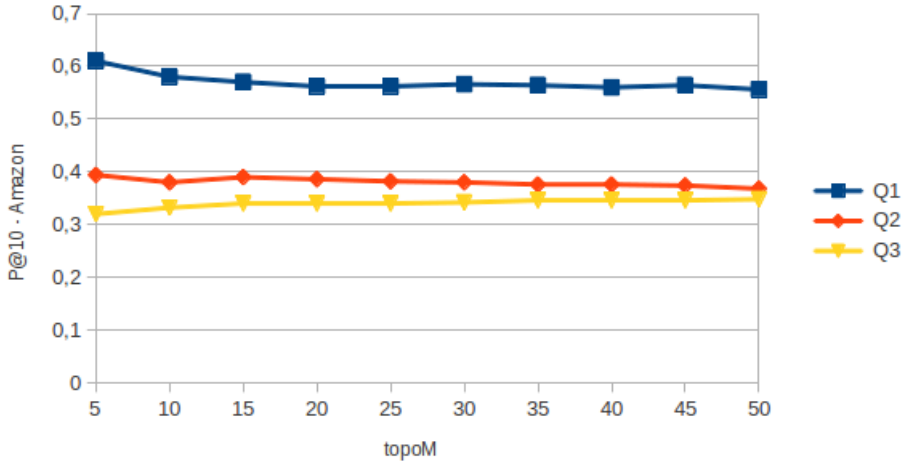

Figura 4.17: Variação do topo-*m* em termos de P@10 do *CatW-BR - Amazon*.

Após a obtenção do *CatW-BR*, o próximo passo consistiu em obter um ranking baseado em evidências textuais a partir de termos associados às imagens do topo-*n* desse ranking. Assim, os experimentos realizados nesta etapa foram feitos com o objetivo de: (i) determinar a quantidade de termos a serem extraídos e (ii) definir o topo do *CatW-BR* a ser considerado para a extração dos termos.

O método *TCatW-BR* obtém o ranking final a partir da combinação dos resultados do ranking visual com o ranking textual. Por esse motivo, resolvemos avaliar a precisão do ranking textual gerado a partir das evidências textuais extraídas. As Tabelas 4.9, 4.10 e 4.11 apresentam os resultados de P@10 obtidos a partir da variação do topo-*n* considerando, respectivamente, a extração do primeiro/último termo da descrição (*TextualRank1*), a extração dos três primeiros/últimos termos da descrição (*TextualRank3*) e a extração de todos os termos da descrição (*TextualRankALL*). Uma visão gráfica dos resultados obtidos pode ser verificada nas Figuras 4.18, 4.19, 4.20, 4.21, 4.22 e 4.23.

Tabela 4.9: Valores de P@10 do *TextualRank1* para a variação do topo-*n*. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$n = 5$	0,086*	0,086	0,036*	0,176	0,074	0,046
$n = 10$	0,084	0,102*	0,028	0,208	0,082	0,068
$n = 15$	0,082	0,100	0,020	0,184	0,080	0,080*
$n = 20$	0,082	0,088	0,014	0,194	0,076	0,072
$n = 25$	0,086*	0,082	0,014	0,192	0,078	0,066
$n = 30$	0,086*	0,078	0,008	0,216	0,084	0,058
$n = 35$	0,084	0,080	0,010	0,210	0,090	0,052
$n = 40$	0,086*	0,086	0,014	0,220	0,084	0,050
$n = 45$	0,078	0,086	0,016	0,224	0,092*	0,040
$n = 50$	0,076	0,088	0,016	0,236*	0,086	0,042

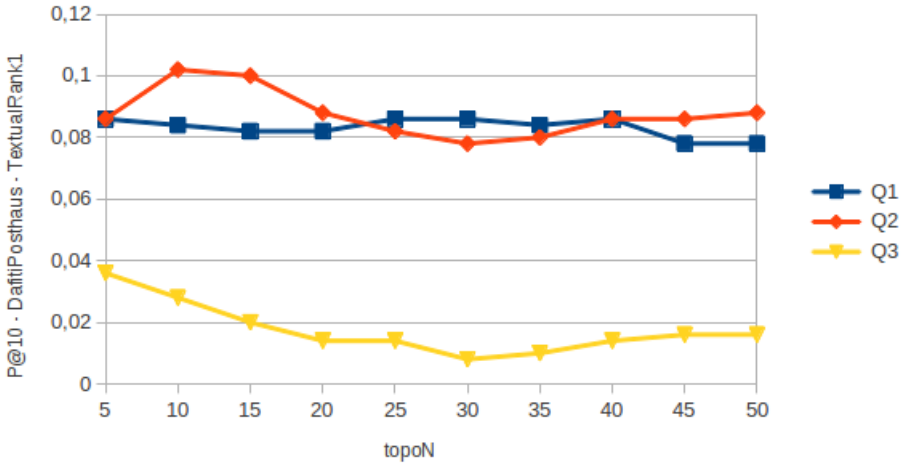

Figura 4.18: Variação do topo-*n* em termos de P@10 - *TextualRank1* - *DafitiPosthaus*.

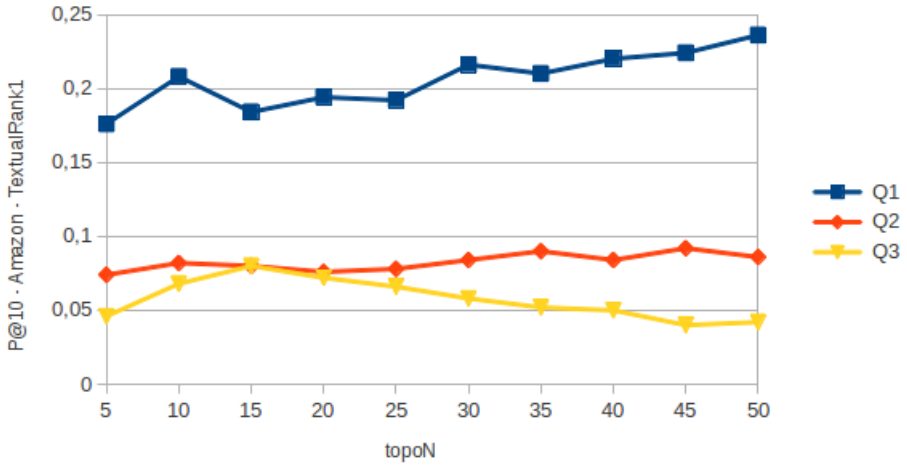

Figura 4.19: Variação do topo-*n* em termos de P@10 - *TextualRank1* - *Amazon*.

Tabela 4.10: Valores de P@10 do *TextualRank3* para a variação do topo-*n*. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$n = 5$	0,376	0,342	0,152	0,310	0,192*	0,100
$n = 10$	0,464	0,348*	0,180*	0,338*	0,184	0,144
$n = 15$	0,494*	0,334	0,178	0,290	0,174	0,148*
$n = 20$	0,490	0,318	0,178	0,286	0,176	0,120
$n = 25$	0,462	0,302	0,160	0,284	0,168	0,116
$n = 30$	0,436	0,312	0,144	0,280	0,184	0,124
$n = 35$	0,418	0,296	0,148	0,270	0,180	0,112
$n = 40$	0,404	0,302	0,134	0,280	0,158	0,102
$n = 45$	0,388	0,288	0,124	0,236	0,160	0,088
$n = 50$	0,386	0,276	0,134	0,224	0,148	0,094

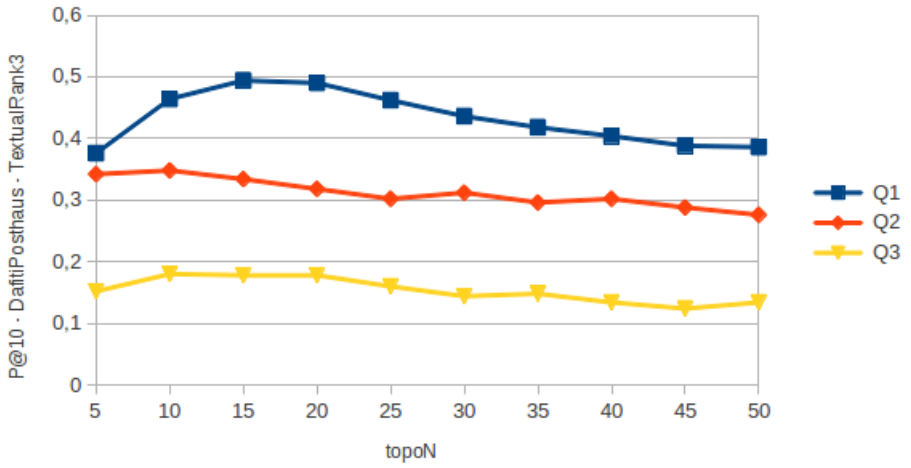

Figura 4.20: Variação do topo-*n* em termos de P@10 - *TextualRank3* - *DafitiPosthaus*.

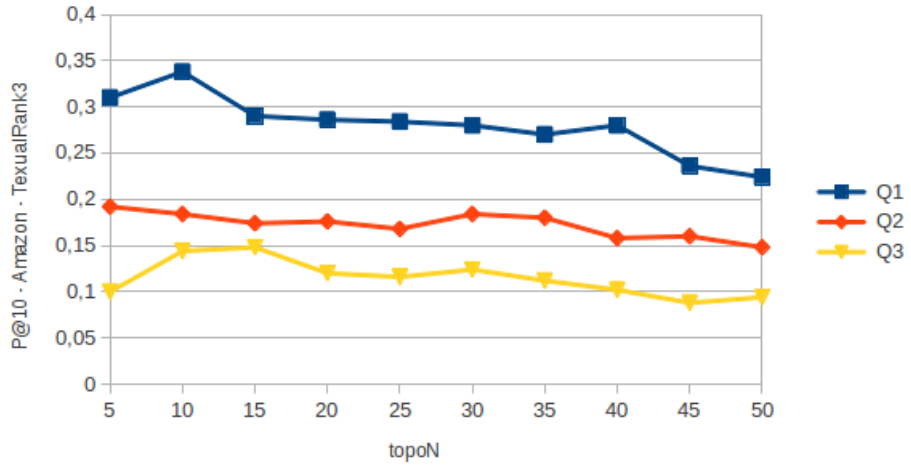

Figura 4.21: Variação do topo-*n* em termos de P@10 - *TextualRank3* - *Amazon*.

Tabela 4.11: Valores de P@10 do *TextualRankALL* para a variação do topo-*n*. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$n = 5$	0,418	0,338*	0,180	0,386	0,228	0,172
$n = 10$	0,500*	0,338*	*0,206	0,446	0,284	0,240
$n = 15$	0,486	0,320	0,202	0,476	0,300*	0,260*
$n = 20$	0,474	0,332	0,182	0,480*	0,296	0,224
$n = 25$	0,444	0,334	0,172	0,460	0,300*	0,220
$n = 30$	0,442	0,326	0,174	0,428	0,292	0,226
$n = 35$	0,428	0,314	0,172	0,414	0,270	0,202
$n = 40$	0,396	0,328	0,172	0,420	0,244	0,202
$n = 45$	0,382	0,284	0,160	0,400	0,234	0,204
$n = 50$	0,370	0,292	0,136	0,372	0,234	0,188

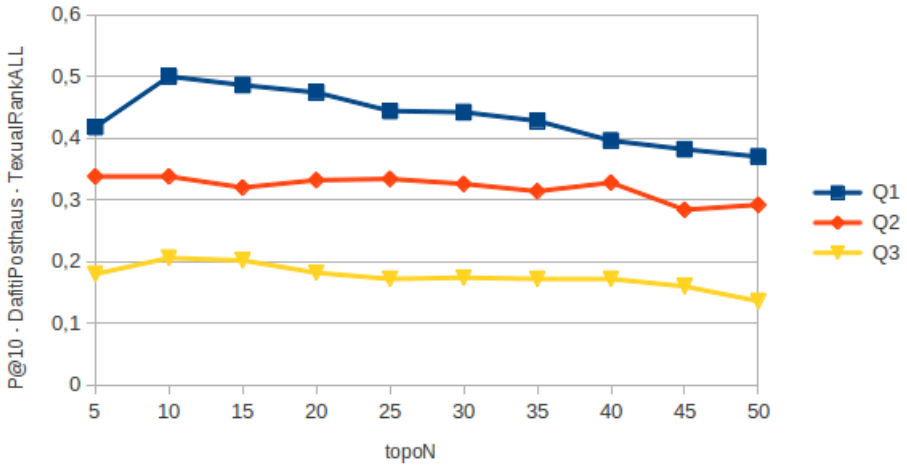

Figura 4.22: Variação do topo-*n* em termos de P@10 - *TextualRankALL* - *DafitiPosthaus*.

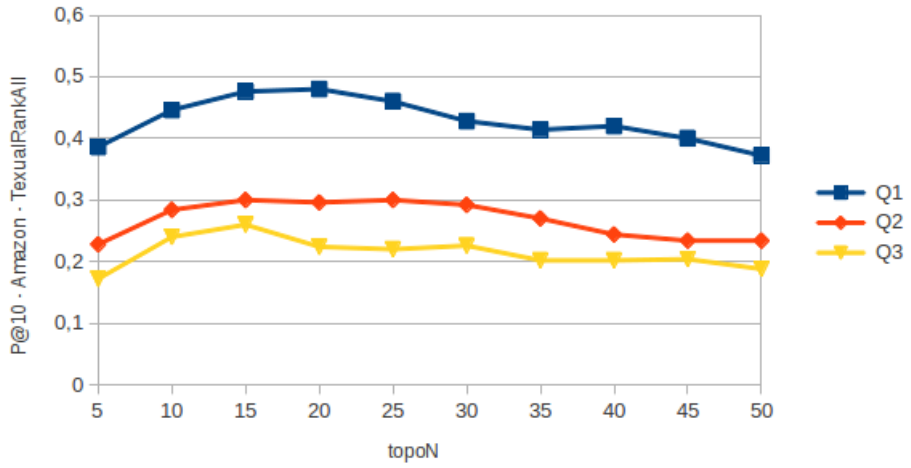

Figura 4.23: Variação do topo-*n* em termos de P@10 - *TextualRankALL* - *Amazon*.

Como pode ser observado na Tabela 4.9, fica evidente que a utilização apenas do primeiro/último termo para realizar uma busca textual alcança precisões muito baixas. Isso é claro, pois um único termo é insuficiente para descrever as características de um produto desejado.

Comparando os ganhos, os resultados do *TextualRank3* e *TextualRankALL* obtiveram ganhos significativos quando comparados aos resultados obtidos pelo *TextualRank1*. Assim, concluímos que a extração dos três primeiros/últimos termos e a extração de todos os termos são capazes de obter informações mais detalhadas sobre o produto buscado, contribuindo com o aumento da relevância dos resultados. É importante que esse ranking retorne bons resultados, pois as imagens retornadas serão agregadas ao ranking final. Com relação ao topo-*n* do *CatW-BR* a ser utilizado para a extração dos termos, consideramos o $n = 25$.

A última etapa do método consiste em realizar a combinação linear entre o ranking visual (*CatW-BR*) e o ranking textual (*TextualRank*). Os experimentos foram realizados com o intuito de verificar o peso a ser atribuído para cada evidência, de forma a identificar a combinação ideal para aumentar a relevância dos resultados.

As Tabelas 4.12, 4.13 e 4.14 apresentam os valores de P@10 associados ao resultado final da execução do método *TCatW-BR*, considerando o ranking gerado a partir da combinação entre o *CatW-BR*, representado por *v*, com respectivamente, o *TextualRank1*, o *TextualRank3* e o *TextualRankALL*, representados por *t*. Os resultados obtidos são apresentados graficamente nas Figuras 4.24, 4.25, 4.26, 4.27,4.28 e 4.29.

Tabela 4.12: Valores de P@10 do *TCatW-BR1* para combinação linear. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$v = 0,5$; $t = 0,5$	0,280	0,306	0,098	0,354	0,258	0,194
$v = 0,6$; $t = 0,4$	0,376	0,486	0,214	0,410	0,346	0,278
$v = 0,7$; $t = 0,3$	0,558	0,574*	0,274	0,542	0,380*	0,330
$v = 0,8$; $t = 0,2$	0,666*	0,574*	0,294	0,556	0,378	0,336
$v = 0,9$; $t = 0,1$	0,664	0,574*	0,304*	0,560*	0,376	0,338*

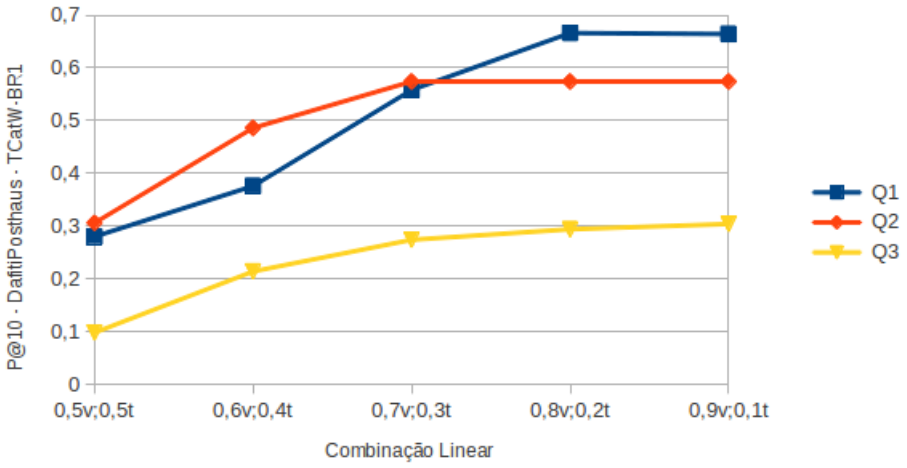

Figura 4.24: Valores de P@10 do *TCatW-BR1* para combinação linear - *DafitiPosthaus*.

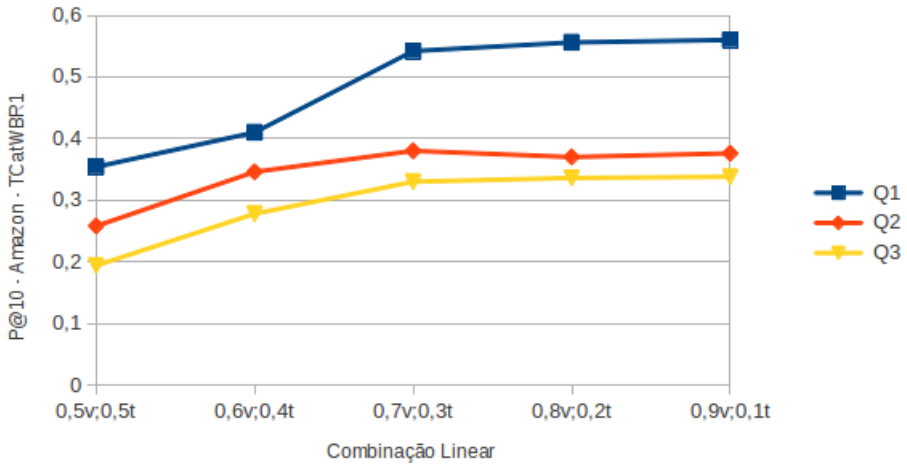

Figura 4.25: Valores de P@10 do *TCatW-BR1* para combinação linear - *Amazon*.

Tabela 4.13: Valores de P@10 do *TCatW-BR3* para combinação linear. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$v = 0,5$; $t = 0,5$	0,678	0,594	0,304	0,500	0,362	0,308
$v = 0,6$; $t = 0,4$	0,676	0,602	0,294	0,542	0,400	0,338
$v = 0,7$; $t = 0,3$	0,680	0,608*	0,298	0,572	0,418*	0,338
$v = 0,8$; $t = 0,2$	0,688*	0,602	0,298	0,580*	0,400	0,342
$v = 0,9$; $t = 0,1$	0,678	0,592	0,310*	0,572	0,384	0,344*

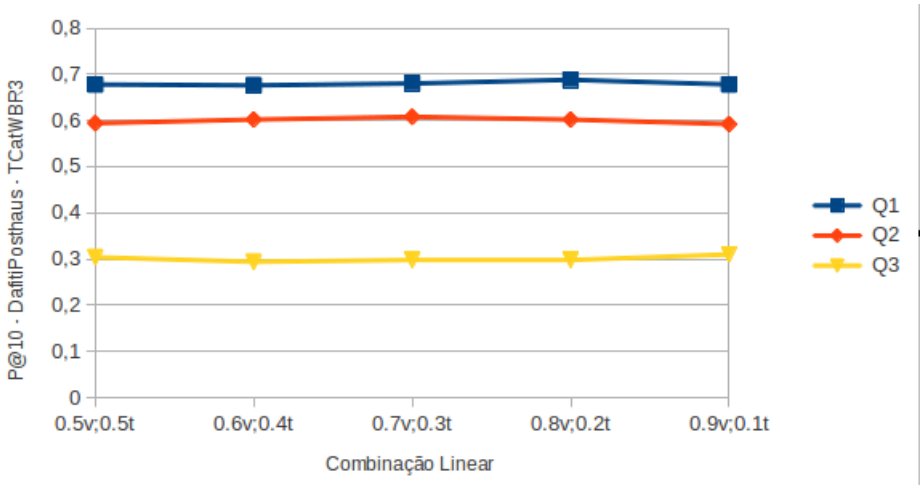

Figura 4.26: Valores de P@10 do *TCatW-BR3* para combinação linear - *DafitiPosthaus*.

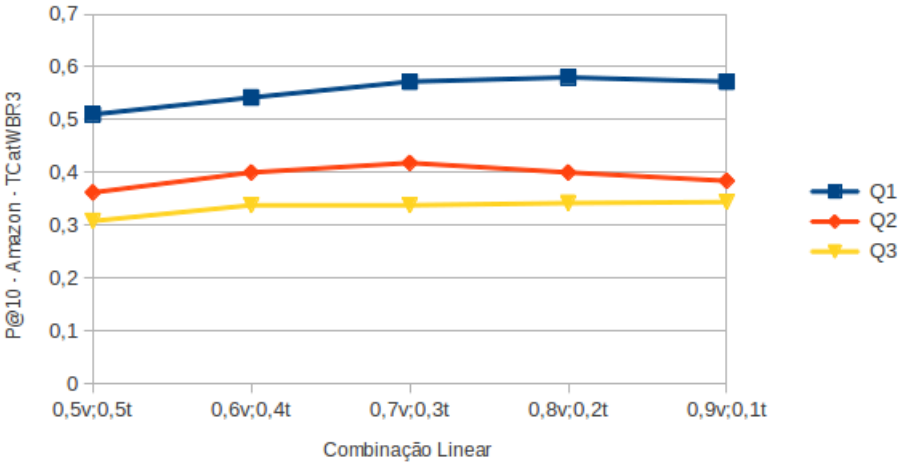

Figura 4.27: Valores de P@10 do *TCatW-BR3* para combinação linear - *Amazon*.

Tabela 4.14: Valores de P@10 do *TCatW-BRALL* para combinação linear. Os maiores valores são apresentados com *.

	DafitiPosthaus			Amazon		
	Q1	Q2	Q3	Q1	Q2	Q3
$v = 0,5$; $t = 0,5$	0,684	0,566	0,298	0,580*	0,396	0,342
$v = 0,6$; $t = 0,4$	0,694*	0,590	0,292	0,560	0,396	0,354*
$v = 0,7$; $t = 0,3$	0,678	0,600*	0,292	0,564	0,398*	0,348
$v = 0,8$; $t = 0,2$	0,686	0,600*	0,306	0,568	0,390	0,340
$v = 0,9$; $t = 0,1$	0,676	0,592	0,308*	0,572	0,382	0,348

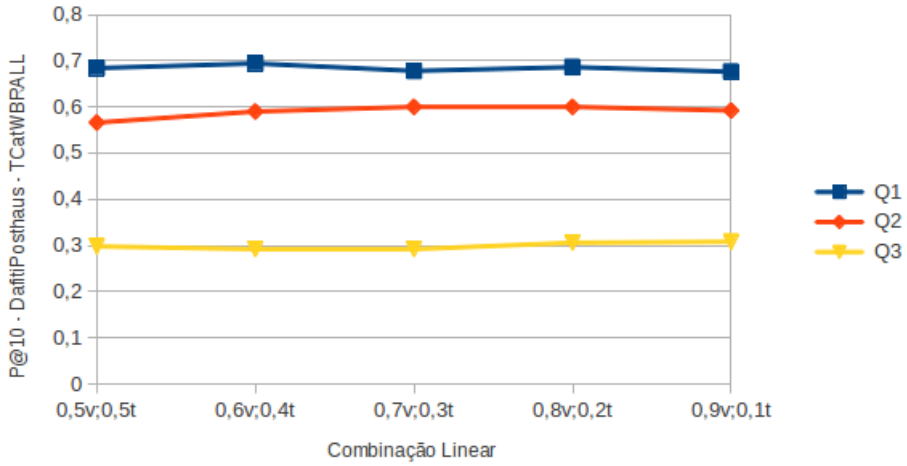

Figura 4.28: Valores de P@10 do *TCatW-BRALL* para combinação linear - *Dafiti-Posthaus*.

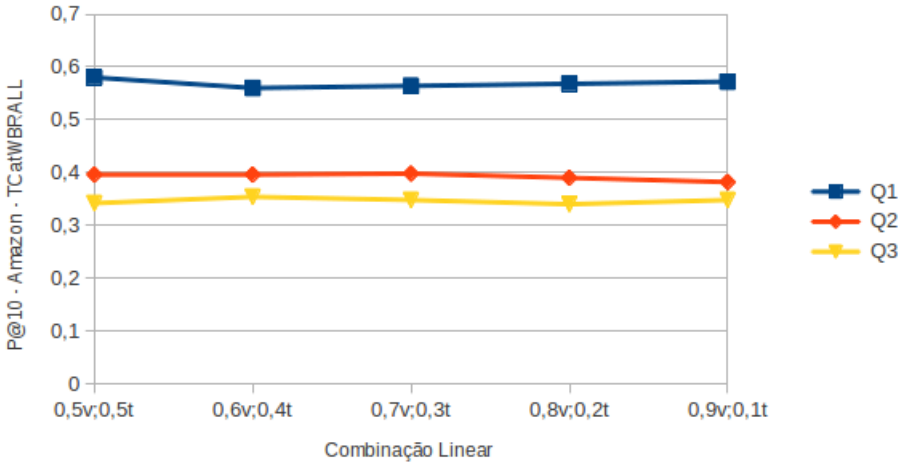

Figura 4.29: Valores de P@10 do *TCatW-BRALL* para combinação linear - *Amazon*.

Analisando os resultados, podemos observar que os melhores valores são obtidos com combinações que atribuem um peso maior para a evidência visual. Isso pode ser moti-

vado pela baixa precisão alcançada pelo ranking textual. Entretanto, mesmo com a baixa precisão, o ranking textual ajuda a aumentar a relevância de imagens com semântica mais próxima à imagem de consulta seja em termos de categoria ou subcategoria, além de complementar o ranking visual com novos resultados relevantes.

As Tabelas 4.15 e 4.16 apresentam os resultados a partir dos quais podemos comparar o desempenho do baseline, o *CatW-BR* e as variações do *TCatW-BR* em termos de P@10, P@20 e MAP. Como pode ser visto, o *CatW-BR* e as variações do *TCatW-BR* superaram o baseline sem re-ranking nos três cenários de consulta e nas duas coleções. Os resultados do *CatW-BR* e as diferenças das variações do *TCatW-BR* são todas estatisticamente significativas quando comparadas ao *CEDD*, o ranking visual original. Foi aplicado o teste estatístico Wilcoxon considerando somente valores maiores que 95% de confiança. Os ganhos foram expressivos em todos os casos e métricas. Uma visão gráfica dos resultados pode ser observada nas Figuras 4.30, 4.31, 4.32, 4.33, 4.34 e 4.35.

Tabela 4.15: Comparação entre baseline e *TCatW-BR* - *DafitiPosthaus*. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,590*	0,560*	0,235*	0,472*	0,448*	0,200*	0,242*	0,226*	0,132*
CATW-BR ($m = 25$)	0,662	0,638	0,269	0,572	0,532	0,236	0,304	0,268	0,165
TCATW-BR1 ($v = 0,8$; $t = 0,2$)	0,666	0,619	0,266	0,574	0,535	0,244	0,294	0,268	0,159
TCATW-BR3 ($v = 0,8$; $t = 0,2$)	0,688*	0,664*	0,350*	0,602*	0,554*	0,278*	0,298	0,272*	0,195
TCATW-BRALL ($v = 0,8$; $t = 0,2$)	0,686	0,662	0,344	0,600	0,545	0,275*	0,306*	0,271	0,196*

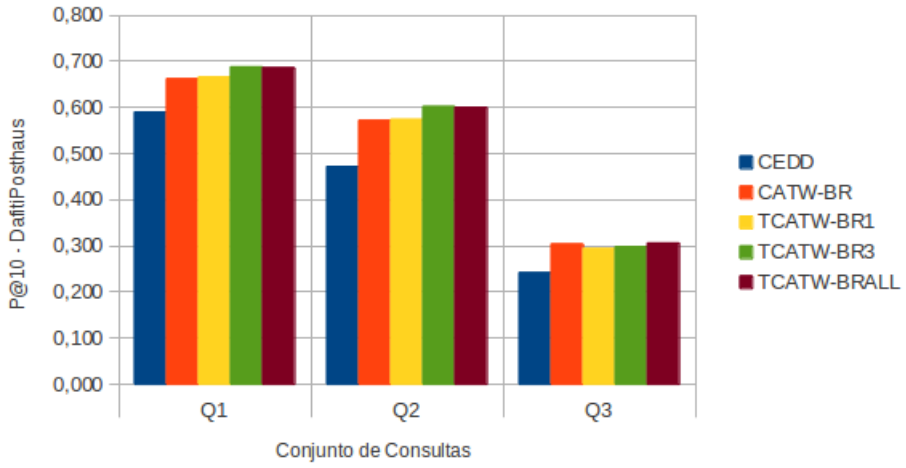

Figura 4.30: Comparação entre baseline e *TCatW-BR* - *DafitiPosthaus* - P@10.

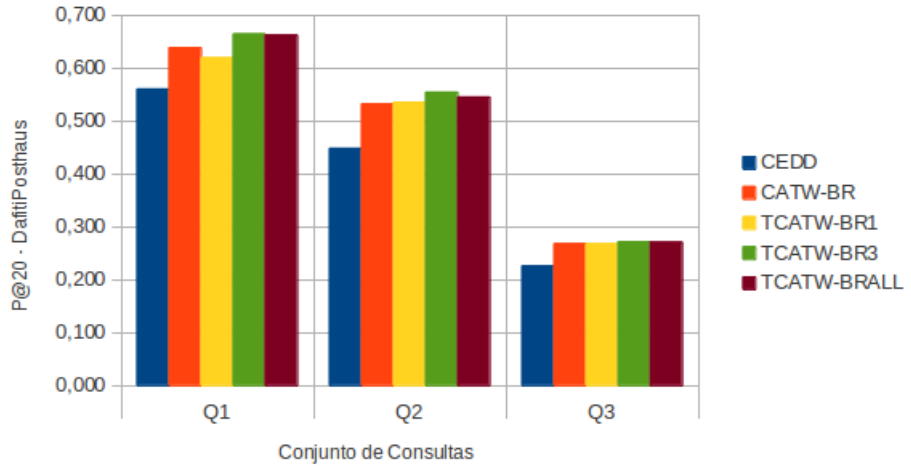

Figura 4.31: Comparação entre baseline e *TCatW-BR* - *DafitiPosthaus* - P@20.

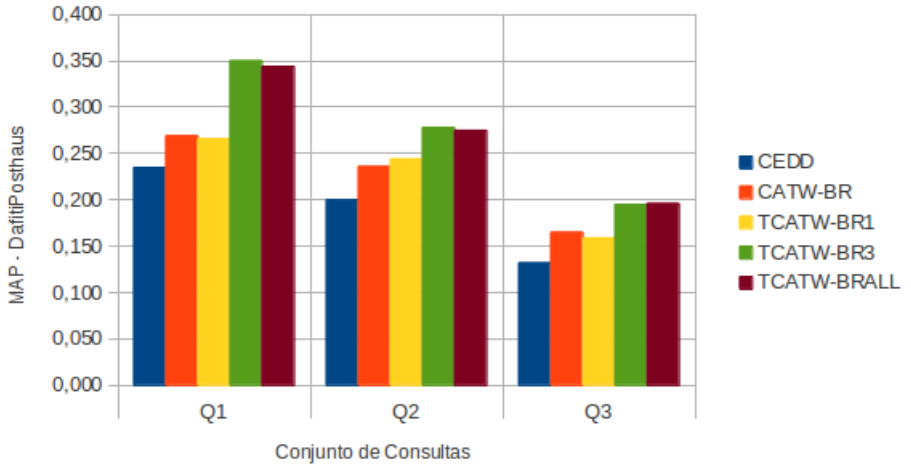

Figura 4.32: Comparação entre baseline e *TCatW-BR* - *DafitiPosthaus* - MAP.

Tabela 4.16: Comparação entre baseline e *TCatW-BR* - *Amazon*. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,522	0,507	0,232	0,310	0,306	0,196	0,242	0,231	0,174
CATW-BR ($m = 25$)	0,562	0,561	0,256	0,382	0,354	0,237	0,340	0,296	0,236
TCATW-BR1 ($v = 0,8$; $t = 0,2$)	0,556	0,557	0,284	0,378	0,358	0,242	0,336	0,297	0,251
TCATW-BR3 ($v = 0,8$; $t = 0,2$)	0,580*	0,561	0,288	0,400*	0,365*	0,250*	0,342*	0,298	0,253
TCATW-BRALL ($v = 0,8$; $t = 0,2$)	0,568	0,575*	*0,292	0,390	0,360	0,249	0,340	0,307*	*0,262

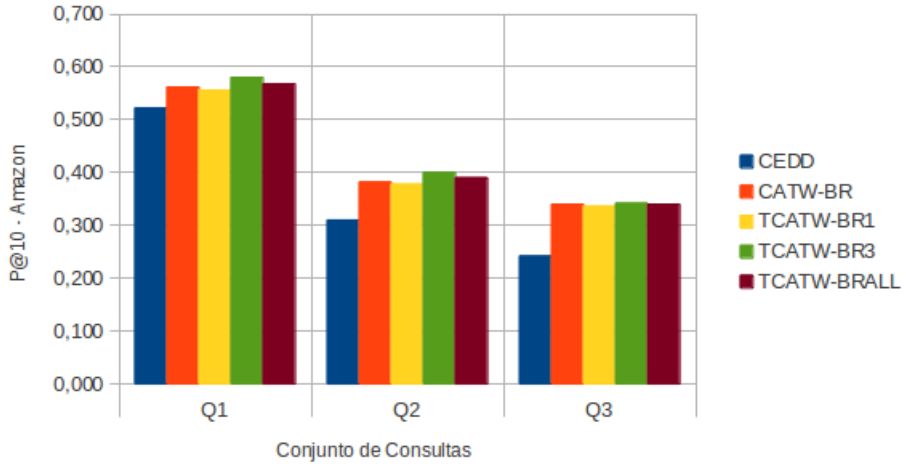

Figura 4.33: Comparação entre baseline e *TCatW-BR* - *Amazon* - P@10.

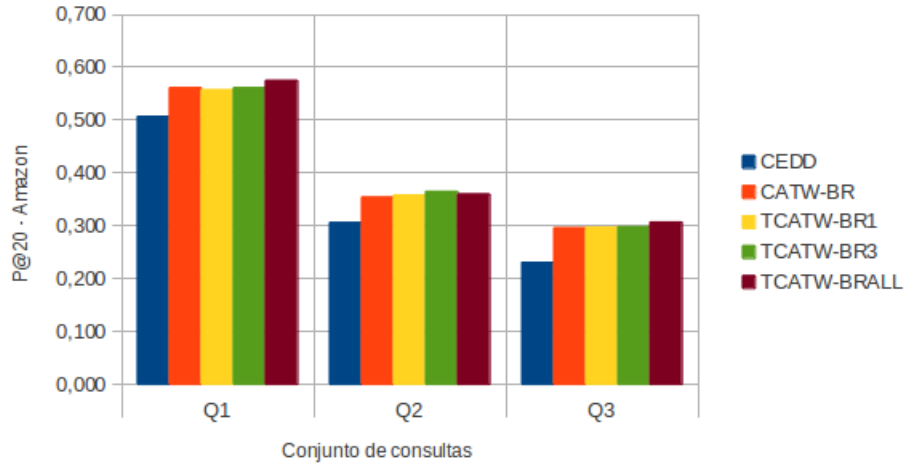

Figura 4.34: Comparação entre baseline e *TCatW-BR* - *Amazon* - P@20.

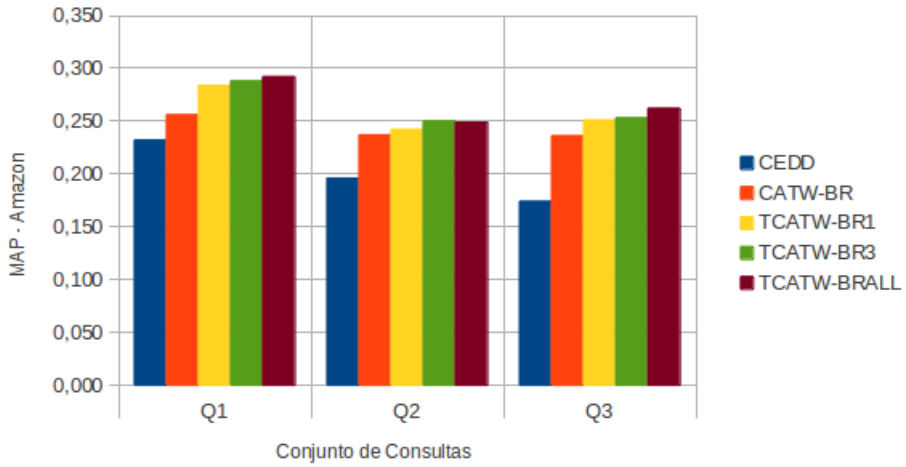

Figura 4.35: Comparação entre baseline e *TCatW-BR* - *Amazon* - MAP.

4.4.3 Análise comparativa entre *TCat-BR* e *TCatW-BR*

Prosseguindo com os experimentos, comparamos os resultados obtidos pelas estratégias *Cat-BR* e *CatW-BR*. A diferença entre as duas está na forma como as informações de categoria são utilizadas para fazer o re-ranking. Os resultados são apresentados nas Tabelas 4.17 e 4.18. Em termos de MAP, as diferenças significativas só foram apresentadas nos conjuntos $Q1$ e $Q2$ da coleção *DafitiPosthaus*. Uma análise gráfica dos resultados pode ser feita por meio das Figuras 4.36, 4.37, 4.38, 4.39, 4.40 e 4.41.

Embora, a estratégia *Cat-BR* tenha obtido ganhos parcialmente superiores, acreditamos que essa abordagem está sujeita a erros que podem prejudicar de forma considerável os resultados apresentados ao usuário. Usar apenas a categoria mais freqüente para fazer o re-ranking dos resultados, como o *Cat-BR* faz, funciona muito bem quando a estimativa está correta, mas no pior caso, ou seja, quando a estimativa da categoria está incorreta, a precisão das primeiras posições do ranking é quase nula. Acreditamos que considerar a distribuição da freqüência das categorias presentes no topo, como o *CatW-BR* faz, é uma estratégia mais robusta e garante um re-ranking mais justo.

Tabela 4.17: Comparação entre *Cat-BR* e *CatW-BR* - *DafitiPosthaus*. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CAT-BR	0,668*	0,649*	0,274*	0,588*	0,552*	0,243*	0,326*	0,291*	0,174*
CATW-BR	0,662	0,638	0,269	0,572	0,532	0,236	0,304	0,268	0,165

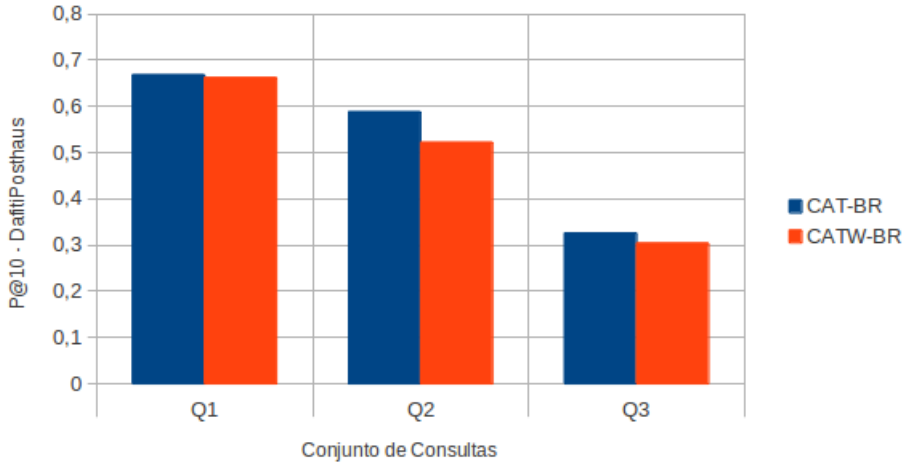

Figura 4.36: Comparação entre *Cat-BR* e *CatW-BR* - *DafitiPosthaus* - P@10.

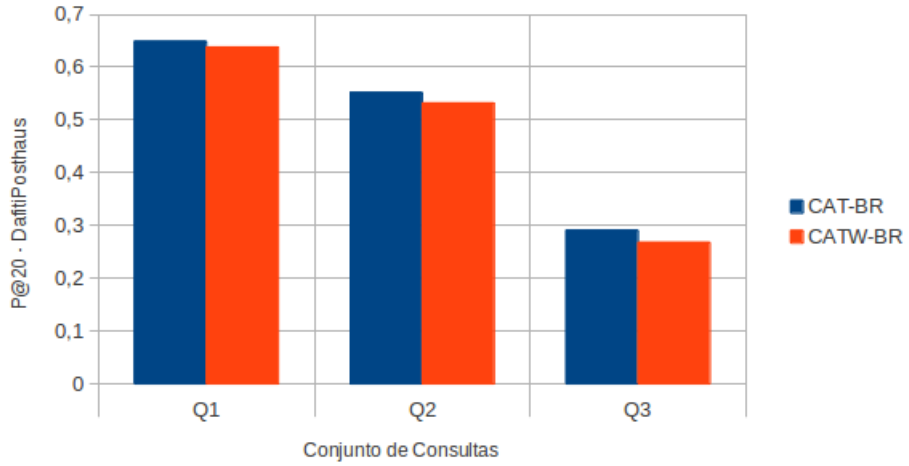

Figura 4.37: Comparação entre *Cat-BR* e *CatW-BR* - *DafitiPosthaus* - P@20.

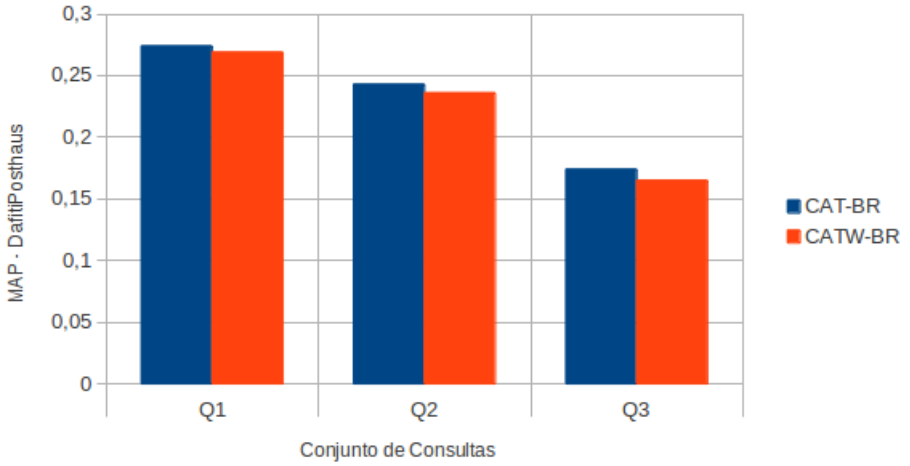

Figura 4.38: Comparação entre *Cat-BR* e *CatW-BR* - *DafitiPosthaus* - MAP.

Tabela 4.18: Comparação entre *Cat-BR* e *CatW-BR* - *Amazon*. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CAT-BR	0,548	0,544	0,247	0,388*	0,362*	0,238*	0,350*	0,311*	0,241*
CATW-BR	0,562*	0,561*	0,256*	0,382	0,354	0,237	0,340	0,296	0,236

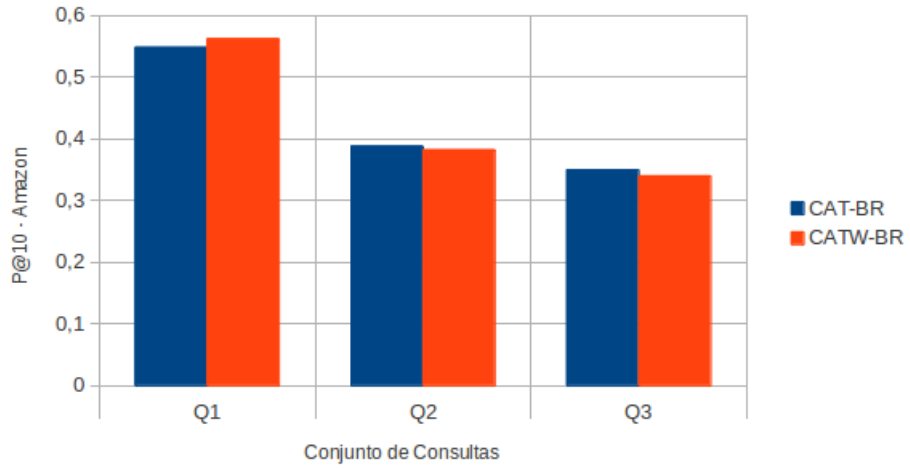

Figura 4.39: Comparação entre *Cat-BR* e *CatW-BR* - *Amazon* - P@10.

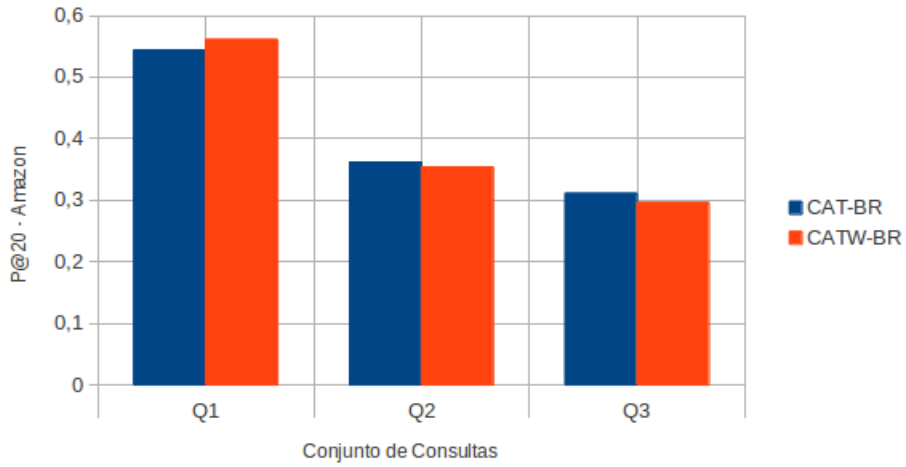

Figura 4.40: Comparação entre *Cat-BR* e *CatW-BR* - *Amazon* - P@20.

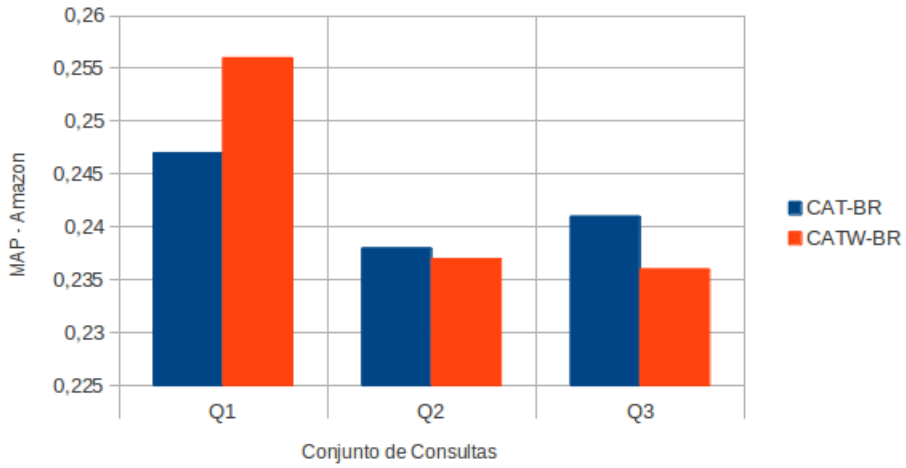

Figura 4.41: Comparação entre *Cat-BR* e *CatW-BR* - *Amazon* - MAP.

As Tabelas 4.19 e 4.20 apresentam um quadro com os resultados das variações do *TCat-BR* e do *TCatW-BR*. Uma análise gráfica desses resultados pode ser feita por meio das Figuras 4.42, 4.43, 4.44, 4.45, 4.46 e 4.47.

Em termos de precisão, o método *TCat-BR* foi superior ao método *TCatW-BR*, mas com ganhos significativos apenas nos conjuntos $Q1$ e $Q2$ da coleção *DafitiPosthaus*; e no conjunto $Q3$ da *Amazon*. A precisão um pouco inferior do *TCatW-BR* com relação ao *TCat-BR* pode ser devido à baixa precisão do ranking textual, que quando combinado ao ranking visual faz com que algumas imagens relevantes do topo caiam algumas posições. Isso não acontece com o *TCat-BR*, pois os resultados do ranking textual não é agregado

ao resultado final.

Por outro lado, o ranking textual gerado pelo *TCatW-BR* complementa o ranking visual com novas imagens relevantes que estão semanticamente associadas à imagem de consulta, aumentando assim a relevância dos resultados do ranking como um todo. Podemos observar isso nos valores de MAP obtidos pelo *TCatW-BR*, que obteve ganhos significativos com relação ao *TCat-BR* nas duas coleções para os três conjuntos de consulta.

Tabela 4.19: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *DafitiPosthaus*. Os maiores valores são apresentados com *.

	DafitiPosthaus								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,590	0,560	0,235	0,472	0,448	0,200	0,242	0,226	0,132
TCAT-BR1	0,688	0,673	0,288	0,630*	0,589*	0,259	0,300	0,269	0,171
TCAT-BR3	0,700*	0,688*	0,297	0,628	0,572	0,252	0,320	0,291*	0,183
TCAT-BRALL	0,692	0,665	0,286	0,614	0,564	0,247	0,330*	0,281	0,181
TCATW-BR1	0,666	0,619	0,266	0,574	0,535	0,244	0,294	0,268	0,159
TCATW-BR3	0,688	0,664	0,350*	0,602	0,554	0,278*	0,298	0,272	0,195
TCATW-BRALL	0,686	0,662	0,344	0,600	0,545	0,275	0,306	0,271	0,196*

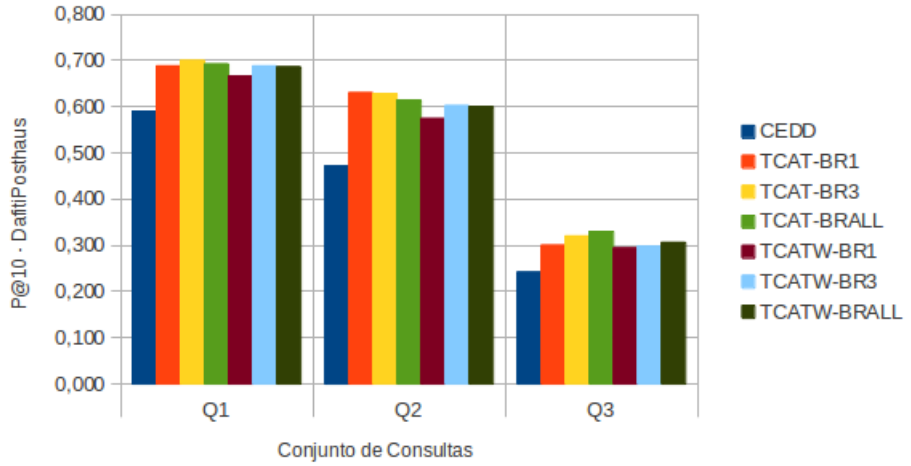

Figura 4.42: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *DafitiPosthaus* - P@10.

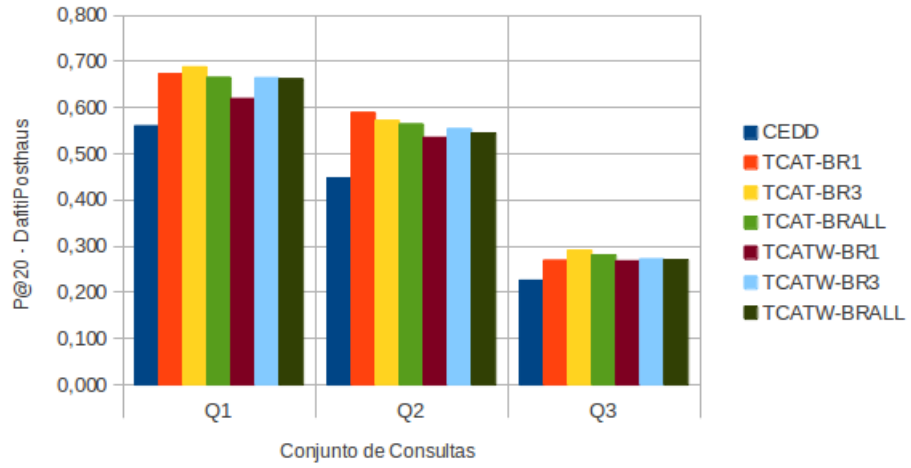

Figura 4.43: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *DafitiPosthaus* - P@20.

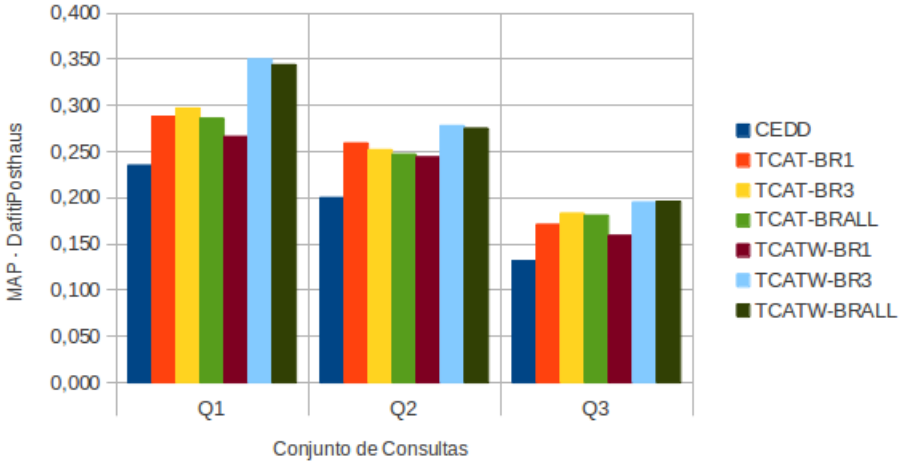

Figura 4.44: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *DafitiPosthaus* - MAP.

Tabela 4.20: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *Amazon*. Os maiores valores são apresentados com *.

	Amazon								
	Q1			Q2			Q3		
	P@10	P@20	MAP	P@10	P@20	MAP	P@10	P@20	MAP
CEDD	0,522	0,507	0,232	0,310	0,306	0,196	0,242	0,231	0,174
TCAT-BR1	0,580	0,587*	0,265	0,402*	0,364	0,238	0,324	0,281	0,222
TCAT-BR3	0,584*	0,568	0,263	0,398	0,367*	0,237	0,334	0,298	0,242
TCAT-BRALL	0,582	0,566	0,260	0,396	0,357	0,228	0,320	0,289	0,237
TCATW-BR1	0,556	0,557	0,284	0,378	0,358	0,242	0,336	0,297	0,251
TCATW-BR3	0,580	0,561	0,288	0,400	0,365	0,250*	0,342*	0,298	0,253
TCATW-BRALL	0,568	0,575	0,292*	0,390	0,360	0,249	0,340	0,307*	0,262*

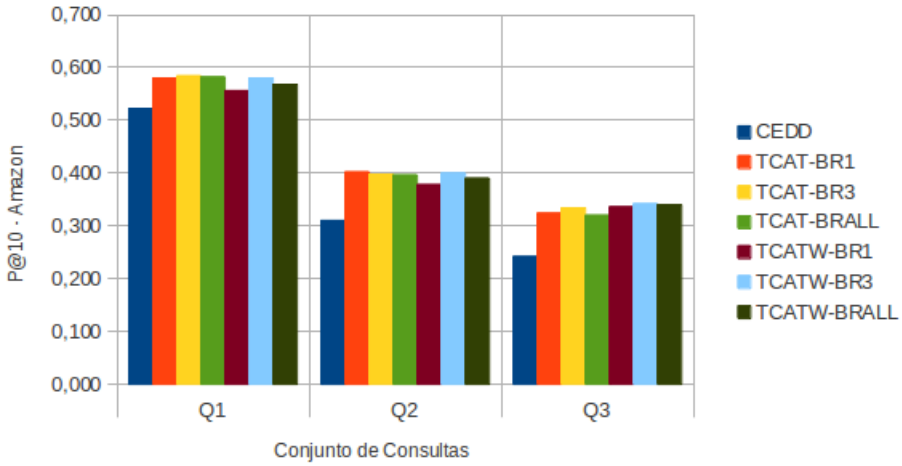

Figura 4.45: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *Amazon* - P@10.

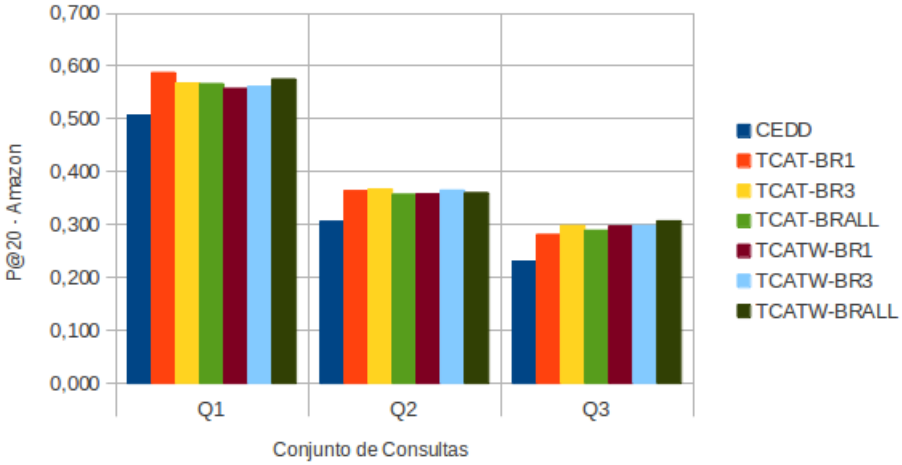

Figura 4.46: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *Amazon* - P@20.

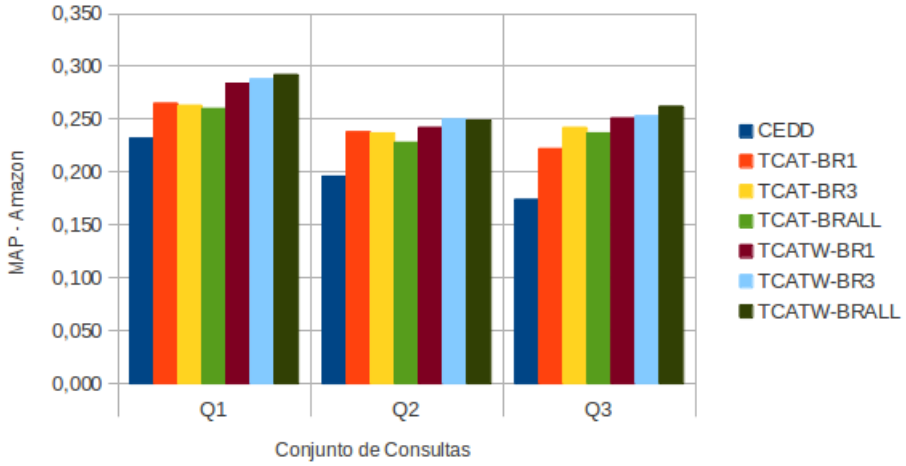

Figura 4.47: Comparação entre baseline, *TCat-BR* e *TCatW-BR* - *Amazon* - MAP.

Capítulo 5

Conclusão

Este trabalho abordou o problema da busca visual de produtos no contexto de sites de comércio eletrônico. Com os resultados obtidos, foi possível verificar que a falta de semântica associada à consulta visual torna a utilização unicamente de técnicas de CBIR insuficiente para resolver de forma eficiente o problema apresentado.

As informações textuais associadas aos produtos disponíveis nos sites de comércio eletrônico mostraram ser uma importante fonte de evidência que contribui de forma significativa para a melhoria da relevância dos resultados apresentados para uma busca puramente visual. Em particular, podemos dizer que a informação de categoria foi fundamental para o sucesso alcançado pelos métodos propostos neste trabalho. No cenário das consultas difíceis, utilizar apenas a categoria estimada da consulta já resultou em um aumento, em termos de MAP, de $31,81\%$ na coleção *DafitiPosthaus* e $38,50\%$ na coleção *Amazon*. Verificamos que uma solução ótima para o problema de estimativa de categoria, poderia alcançar, neste mesmo cenário, ganhos entre $56,06\%$ e $68,96\%$.

Considerando a aplicação da estratégia multimodal, embora a evidência textual contribua em parte para o aumento da precisão dos resultados, a evidência visual obtida pelo CBIR exerce um papel fundamental na recuperação de imagens relevantes. Isso pode ser observado na aplicação da fusão de evidências, onde os melhores resultados foram atribuídos à combinação que atribuiu um peso maior à evidência visual.

Nossa proposta de solução resultou na definição de dois métodos de re-ranking baseados em informações de categoria normalmente encontradas nesses tipos de Web sites. Os métodos propostos apresentaram melhorias significativas em termos de precisão das consultas nas duas coleções e nos três cenários adotados nos experimentos realizados. Em

termos de MAP, o método *TCatW-BR* foi o que obteve os melhores resultados. O menor ganho obtido por este método ocorreu no cenário das consultas consideradas difíceis, para as quais alcançou um aumento de $48,48\%$ na coleção *DafitiPosthaus* e $50,57\%$ na coleção *Amazon* quando comparado à busca puramente visual, em termos de MAP. Os resultados indicam que usar informação multimodal para realizar o re-ranking dos resultados é uma solução promissora para a busca visual de produtos.

Assim podemos concluir que este trabalho alcançou o objetivo para o qual foi proposto, gerando como contribuições: (i) a disponibilização de duas estratégias para descobrir de forma automática a categoria de um produto dada uma imagem de consulta, sem a necessidade de interação com o usuário ou a aplicação de técnicas de aprendizagem de máquina e (ii) a definição de um novo método para busca visual de produtos que combina a categoria estimada e a descrição do produto para reordenar o ranking visual.

5.1 Trabalhos Futuros

Apesar dos bons resultados obtidos com a aplicação dos métodos propostos, é possível realizar trabalhos futuros no sentido de alcançar uma maior eficácia em termos de busca visual de produtos.

Um ponto a ser explorado é o problema da estimativa de categoria da consulta, uma vez que nossos experimentos indicaram que a precisão pode aumentar de forma considerável quando a estimativa é feita de forma correta. Isso é particularmente verdade nos casos das consultas difíceis, ou seja, aquelas que a imagem não pertence à coleção consultada e que apresentam muito ruído em seu plano de fundo. Neste sentido, podem ser feitos experimentos nas seguintes direções: (i) identificar alternativas para determinar dinamicamente o topo do resultado a ser usado para estimar a categoria da consulta e (ii) verificar o desempenho de técnicas de aprendizagem de máquina na tarefa de determinar propriedades que ajudam na estimativa da categoria da consulta. Outro aspecto que pode ser estudado é o impacto do número de categorias sobre o desempenho do método, ou seja, verificar o que acontece ao aumentar ou diminuir a quantidade de categorias da coleção.

Com relação às técnicas de recuperação multimodal, existe muito a ser explorado. Um exemplo seria a realização de um estudo sobre a seleção de melhores evidências para

representar os conteúdos combinados. Experimentos poderiam ser feitos para verificar a utilização de mais de um descritor para a geração do ranking visual inicial. Outros pontos a serem explorados são: estudo sobre a melhor função de fusão de evidências a ser utilizada, aplicação de técnicas de aprendizagem de máquina para definir estratégias de expansão de consulta e modelos de re-ranking. Nesse sentido, é necessário verificar em quais situações a utilização de expansão de consultas se torna viável e qual o impacto da utilização das técnicas de aprendizagem no processamento das consultas. Além disso, seria interessante definir um modelo formal para os métodos propostos utilizando, por exemplo, *language models*.

Referências Bibliográficas

[1] ARAMPATZIS, A., ZAGORIS, K., AND CHATZICHRISTOFIS, S. Dynamic two-stage image retrieval fromÂ largeÂ multimodalÂ databases. In *Advances in Information Retrieval*, vol. 6611 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011, pp. 326–337.

[2] ARICA, N., AND F.T., Y. Bas: a perceptual shape descriptor based on the beam angle statistics. *Pattern Recognition Letters 24*, 9 (2003), 1627–1639.

[3] BAEZA-YATES, R., AND RIBEIRO-NETO, B. Modern information retrieval: the concepts and technology behind search, harlow, 2011.

[4] BELONGIE, S., MALIK, J., AND PUZICHA, J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 4 (2002), 509–522.

[5] CHA, S. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences 1*, 4 (2007), 300–307.

[6] CHANDRASEKHAR, V. R., CHEN, D. M., TSAI, S. S., CHEUNG, N.-M., CHEN, H., TAKACS, G., REZNIK, Y., VEDANTHAM, R., GRZESZCZUK, R., BACH, J., AND GIROD, B. The stanford mobile visual search data set. In *ACM Multimedia Systems* (2011), pp. 117–122.

[7] CHANG, S., SIKORA, T., AND PURL, A. Overview of the mpeg-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology 11*, 6 (2001), 688–695.

[8] CHATZICHRISTOFIS, S., AND BOUTALIS, Y. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. *Computer Vision Systems* (2008), 312–322.

[9] CHATZICHRISTOFIS, S. A., AND BOUTALIS, Y. S. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *International Workshop on Image Analysis for Multimedia Interactive Services* (2008), IEEE, pp. 191–196.

[10] CHEN, Y., YU, N., LUO, B., AND CHEN, X.-W. ilike: integrating visual and textual features for vertical search. In *Proceedings of the international conference on Multimedia* (2010), ACM, pp. 221–230.

[11] CUI, J., WEN, F., AND TANG, X. Real time google and live image search reranking. In *Proceedings of the 16th ACM international conference on Multimedia* (2008), pp. 729–732.

[12] DEL BIMBO, A. *Visual information retrieval*. Morgan Kaufmann, 1999.

[13] DEPEURSINGE, A., AND MÜLLER, H. Fusion techniques for combining textual and visual information retrieval. *ImageCLEF* (2010), 95–114.

[14] DEPEURSINGE, A., AND MÜLLER, H. Fusion techniques for combining textual and visual information retrieval. *ImageCLEF* (2010), 95–114.

[15] HOU, A., LIU-QING, Z., AND DONG-CHENG, S. Garment image retrieval based on multi-features. In *International Conference on Computer, Mechatronics, Control and Electronic Engineering* (2010), vol. 6, pp. 194–197.

[16] HSU, W., KENNEDY, L., AND CHANG, S. Reranking methods for visual search. *IEEE Multimedia 14*, 3 (2007), 14–22.

[17] HUANG, J., KUMAR, S., MITRA, M., ZHU, W., AND ZABIH, R. Image indexing using color correlograms. In *IEEE Computer Vision and Pattern Recognition* (1997), pp. 762–768.

[18] JAIN, V., AND VARMA, M. Learning to re-rank: query-dependent image re-ranking using click data. In *ACM WWW* (2011), pp. 277–286.

[19] KEJIA, W., HONGGANG, Z., LUNSHAO, C., YING, H., AND PING, Z. A comparative study of moment-based shape descriptors for product image retrieval. In *International Conference on Image Analysis and Signal Processing (IASP)* (2011), IEEE, pp. 355–359.

[20] KHERFI, M., ZIOU, D., AND BERNARDI, A. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Computing Surveys (CSUR) 36*, 1 (2004), 35–67.

[21] KIMURA, P., CAVALCANTI, J., SARAIVA, P., TORRES, R., AND GONÇALVES, M. Evaluating retrieval effectiveness of descriptors for searching in large image databases. *Journal of information and data management 2*, 3 (2011), 305–321.

[22] KOVALEV, V., AND VOLMER, S. Color co-occurrence descriptors for querying-by-example. In *ACM Multimedia Modeling* (1998), IEEE, pp. 32–38.

[23] LIN, X., GOKTURK, B., SUMENGEN, B., AND VU, D. Visual search engine for product images. In *Proc. SPIE 6820, Multimedia Content Access: Algorithms and Systems II* (2008), pp. 1–9.

[24] LIU, Y., MEI, T., AND HUA, X. Crowdreranking: exploring multiple search engines for visual search reranking. In *ACM SIGIR* (2009), pp. 500–507.

[25] LOWE, D. Distinctive image features from scale-invariant keypoints. *International journal of computer vision 60*, 2 (2004), 91–110.

[26] LUX, M. Content based image retrieval with lire. In *ACM Multimedia Modeling* (2011), pp. 735–738.

[27] MANJUNATH, B., OHM, J., VASUDEVAN, V., AND YAMADA, A. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology 11*, 6 (2001), 703–715.

[28] MCGILL, M., AND SALTON, G. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[29] MEHTRE, B., KANKANHALLI, M., AND LEE, W. Shape measures for content based image retrieval: a comparison. *Information Processing & Management 33*, 3 (1997), 319–337.

[30] OUYANG, Y. Clothes image searching system based on sift features. In *International Conference on E-Business and Information System Security* (2009), pp. 1–5.

[31] PASS, G., ZABIH, R., AND MILLER, J. Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia* (1997), ACM, pp. 65–73.

[32] PEDRONETTE, D. C. G., AND DA S TORRES, R. Exploiting contextual spaces for image re-ranking and rank aggregation. In *Proceedings of the first ACM International Conference on Multimedia Retrieval* (2011), ACM, p. 13.

[33] PENATTI, O. A., VALLE, E., AND TORRES, R. D. S. Comparative study of global color and texture descriptors for web image retrieval. *Journal of Visual Communication and Image Representation 23*, 2 (2012), 359–380.

[34] POPESCU, A., MOËLLIC, P., KANELLOS, I., AND LANDAIS, R. Lightweight web image reranking. In *ACM Multimedia Systems* (2009), pp. 657–660.

[35] RO, Y. M., KIM, M., KANG, H. K., MANJUNATH, B., AND KIM, J. Mpeg-7 homogeneous texture descriptor. *Journal of Electronics and Telecommunications Research Institute 23*, 2 (2001), 41–51.

[36] SANTOS, J., CAVALCANTI, J., SARAIVA, P., AND MOURA, E. Multimodal reranking of product image search results. In *European Conference on Information Retrieval* (2013), Springer, p. to appear.

[37] STEHLING, R., NASCIMENTO, M., AND FALCÃO, A. An adaptive and efficient clustering-based approach for content-based image retrieval in image databases. In *Database Engineering & Applications, 2001 International Symposium on.* (2001), IEEE, pp. 356–365.

[38] STEHLING, R. O., NASCIMENTO, M. A., AND FALCÃO, A. X. A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management* (2002), ACM, pp. 102–109.

[39] STEHLING, R. O., NASCIMENTO, M. A., AND FALCAO, A. X. Techniques for color-based image retrieval. *Multimedia Mining* (2002), 61–82.

[40] SWAIN, M., AND BALLARD, D. Color indexing. *International journal of computer vision 7*, 1 (1991), 11–32.

[41] TORRES, R., AND FALCÃO, A. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada 2*, 13 (2006), 161–185.

[42] TORRES, R., AND FALCÃO, A. Contour salience descriptors for effective image retrieval and analysis. *Image and Vision Computing 25*, 1 (2007), 3–13.

[43] TORRES, R., ZEGARRA, J., SANTOS, J., FERREIRA, C., PENATTI, O., ANDALÓ, F., AND ALMEIDA JR, J. Recuperação de imagens: Desafios e novos rumos. In *XXXV Seminário Integrado de Software e Hardware (SEMISH)* (2008), SBC, pp. 223–237.

[44] TSAY, J., LIN, C., TSENG, C., AND CHANG, K. On visual clothing search. In *International Conference on Technologies and Applications of Artificial Intelligence* (2011), pp. 206–211.

[45] TSENG, C.-H., HUNG, S.-S., TSAY, J.-J., AND TSAIH, D. An efficient garment visual search based on shape context. *WSEAS Transactions on Computers 8*, 7 (2009), 1195–1204.

[46] WILCOXON, F. Individual comparisons by ranking methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[47] XIE, X., LU, L., JIA, M., LI, H., SEIDE, F., AND MA, W.-Y. Mobile search with multimodal queries. In *Proceedings of the IEEE* (2008), pp. 589–601.

[48] YAO, T., MEI, T., AND NGO, C. Co-reranking by mutual reinforcement for image search. In *ACM CIVR* (2010), pp. 34–41.