Agregação Não Supervisionada de Rankings para Redução de Cold-Start em Recuperação Multimodal de Imagens

Wanderson Bezerra da Silva[1], **Rodrigo Tripodi Calumby**[1]

[1]Universidade Estadual de Feira de Santana – Feira de Santana – BA – Brazil

wbsilva@ecomp.uefs.br, rtcalumby@uefs.br

Abstract. *In content-based image retrieval systems, the objective of relevance feedback techniques is to enable the user to express her need without specific knowledge of the low-level image features. For a proper behaviour of this technique, the result set prior to the first interaction of the user must present relevant results. Aiming at attenuating the cold-start problem and improving the initial set, this work experimentally evaluated several rank aggregation methods to combine results obtained with different image ranking features. The results showed promising effectiveness when compared to the baselines considering different modalities of features.*

Resumo. *Em sistemas de recuperação de imagens por conteúdo, a técnica de realimentação de relevância visa permitir ao usuário exprimir a sua necessidade de busca sem conhecer propriedades de baixo nível das imagens. Para o bom funcionamento desta técnica, o conjunto de resultados anterior à primeira interação do usuário deve apresentar resultados relevantes. Para atenuar o problema de cold-start e aprimorar o conjunto inicial de resultados, neste trabalho realizou-se uma avaliação experimental de métodos de agregação de rankings para combinar resultados obtidos a partir de diferentes critérios de ranqueamento de imagens. Resultados considerando diferentes modalidades de características indicam eficácia promissora em relação aos baselines.*

1. Introdução

Dadas as necessidades dos usuários e a grande quantidade de informação disponível, é imprescindível o uso de técnicas eficazes para exploração destas coleções. Em se tratando de imagens, a abordagem mais comum para busca baseia-se na utilização de informações textuais (metadados, palavras-chaves, páginas web, etc.) e no uso de consultas tradicionais em bancos de dados para recuperá-las. Um outro paradigma utiliza a descrição das propriedades visuais (cor, forma, textura, etc.) das imagens para indexá-las e buscá-las. Nos sistemas de recuperação de imagens por conteúdo, a atividade consiste em, dada uma imagem de consulta, calcular a sua similaridade em relação às outras armazenadas. Várias técnicas são utilizadas para capturar e representar as informações visuais das imagens [Torres and Falcão 2006]. Neste processo, a noção de similaridade entre as imagens pode variar de acordo com o usuário que realiza a busca. Com frequência os descritores de conteúdo per si não são capazes de representar apropriadamente o conteúdo conceitual de uma imagem. Este problema é conhecido como gap-semântico e acentua-se em situações de consultas complexas. De modo a atenuar este problema, descritores são combinados para adaptar a busca às necessidades do usuário, o que não é uma tarefa

fácil [Atrey et al. 2010]. Para isso, uma técnica que tem sido empregada com sucesso é a realimentação de relevância, em que o usuário interage com o sistema, indicando a relevância dos itens no resultado e o sistema retorna outros itens possivelmente mais relevantes. Assim, a máquina de busca pode utilizar técnicas de aprendizado de máquina para determinar padrões e criar modelos de representação das necessidades do usuário.

Tradicionalmente, mecanismos de aprendizado de máquina têm sido empregados para aprimoramento de buscas utilizando realimentação de relevância. A proposta da técnica de realimentação de relevância é possibilitar ao usuário exprimir a sua necessidade sem ter que conhecer propriedades de baixo nível da imagem. Esse processo é realizado iterativamente e interativamente [Calumby et al. 2014]. A cada iteração, o algoritmo de aprendizado busca capturar quais as propriedades melhor definem as imagens informadas como relevantes pelo usuário.

Neste contexto, dada a complexidade da consulta ou a pouca informação disponível para a configuração de um sistema de busca, o conjunto inicial de resultados, anterior à primeira interação do usuário, pode não apresentar informações relevantes suficientes, caracterizando o problema do *cold-start*. Isso pode acarretar em um *feedback* pobre (com pouca informação) e consequentemente dificultar a etapa de aprendizado. Por ser uma abordagem iterativa, a influência do primeiro resultado propaga-se para as demais iterações, dado que um resultado inicial ruim limita a troca de informação entre o usuário e o sistema [Calumby et al. 2017] e consequentemente os modelos de aprendizado da intenção do usuário. Visando atenuar este problema, é imprescindível que o resultado da primeira interação seja o melhor possível.

Uma abordagem proposta para este cenário, conhecida como agregação de *rankings* [Lin 2010], é a combinação de resultados obtidos a partir de diferentes critérios de *ranking*, por exemplo, diferentes características visuais das imagens ou medidas de *ranking* baseadas no texto associado à elas. Esta fusão pode ser realizada de diferentes formas, incluindo algoritmos de agregação de *rankings* e algoritmos de rerranqueamento [Mei et al. 2014]. Utilizar métodos de fusão de *rankings*, de modo geral, permite resultados superiores à utilização dos critérios de modo isolado ou com técnicas simples de combinação de escores de relevância.

Considerando a importância de um bom conjunto inicial de resultados nos sistemas de recuperação interativa, este trabalho avalia experimentalmente a redução do *cold-start* por meio da exploração de métodos de agregação de *rankings* considerando diferentes critérios de ranqueamento de imagens.

2. Trabalhos Relacionados

As técnicas de agregação de *rankings* podem ser divididas em duas categorias principais: baseadas em escores ou baseadas em ordem. No primeiro grupo, a função de agregação utiliza as informações de pontuação dos objetos de cada lista. Na segunda, apenas a ordem relativa entre os itens é considerada [Vargas Muñoz et al. 2015]. Dentre as técnicas baseadas em escore, pode-se destacar a família Comb* [Shaw and Fox 1994] (e.g., CombMIN, CombMAX, CombSUM, CombMED, CombMNZ, e CombANZ). Em relação aos métodos baseados em posição, pode-se citar o Median Rank Aggregation (MRA) [Fagin et al. 2003], Reciprocal Rank Fusion (RRF) [Cormack et al. 2009] e Borda [Young 1974].

Para melhorar o ranking inicial de um sistema de recuperação interativa de imagens, [Calumby et al. 2014] apresenta uma adaptação do método de agregação baseado em escore proposto por [Ferreira et al. 2011]. Neste método, a similaridade entre dois objetos é definida como o valor médio de todas as medidas de similaridade disponíveis considerando múltiplas características visuais e textuais. Para consultas com mais de um objeto, os itens de coleção são classificados com base no valor mínimo de distância para cada objeto presente na consulta.

3. Metodologia Experimental

Nesse trabalho foi utilizada a ImageCLEF Photographic Retrieval Task collection [Arni et al. 2009]. Esta base de dados é formada por 20.000 imagens (fotos tiradas a partir de locais ao redor do mundo), em que cada imagem está associada a metadados textuais, como título, data e descrição dos conteúdos semânticos e visuais da imagem. Esta base de dados conta com 39 consultas compostas por um fragmento de texto e três imagens de exemplo.

Os experimentos basearam-se em sete descritores visuais globais [Penatti et al. 2012]: baseados em cor (GCH, BIC, ACC e JAC) e textura (CCOM, LAS, e QCCH). Para a modalidade textual, utilizou-se seis medidas de similaridade entre o texto da consulta e a descrição das imagens, sendo elas: Cosine [Baeza-Yates and Ribeiro-Neto 2008], BM25 [Baeza-Yates and Ribeiro-Neto 2008], Dice [Lewis et al. 2006], Jaccard [Lewis et al. 2006], tf-idf-sum [dos Santos et al. 2009] e Bag-of-words (intersecção de termos normalizada). Foram aplicadas as técnicas de remoção de *stop words* e *stemming*. Apenas os metadados em inglês foram considerados.

Nos experimentos foram utilizados métodos de agregação baseados em escore e métodos baseadas em ordem. Dentre os baseados em escore, aplicou-se os métodos da família Comb*, sendo eles: CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, e CombMED. Dentre os baseados em ordem, considerou-se: MRA, RRF e Borda. Para cada um dos descritores, foi gerado um *ranking* utilizando as 20.000 imagens da base. Estes *rankings* foram usados como entrada para os métodos de agregação. Foram definidos três cenários de busca. O primeiro cenário considerou apenas informações visuais para gerar os *rankings* de entrada para os métodos de fusão, enquanto no segundo cenário considerou-se apenas as informações textuais. O terceiro experimento representou um cenário multimodal, ou seja, utilizou-se tanto informações visuais quanto textuais.

4. Resultados e Discussões

Considerando que cada um dos *rankings* gerado pelos descritores contém todas as imagens da base de dados, o método CombSUM e suas derivações (CombMED, CombANZ e CombMNZ) apresentaram o mesmo resultado, visto que as derivações diferenciam-se especificamente pelo modo como levam em consideração a quantidade de *rankings* onde cada imagem está presente. Assim, apresentamos aqui apenas os resultados do método CombSUM. Os *rankings* obtidos com os métodos de agregação foram comparados com aqueles obtidos com o método proposto por [Calumby et al. 2014] (aqui denominado MinAvg). Como critério de avaliação, utilizou-se curvas de *Precisão* (P@N). A Figura 1 apresenta os resultados obtidos utilizando apenas os descritores visuais (vis). A Figura 2 apresenta o comparativo considerando os descritores textuais (txt).

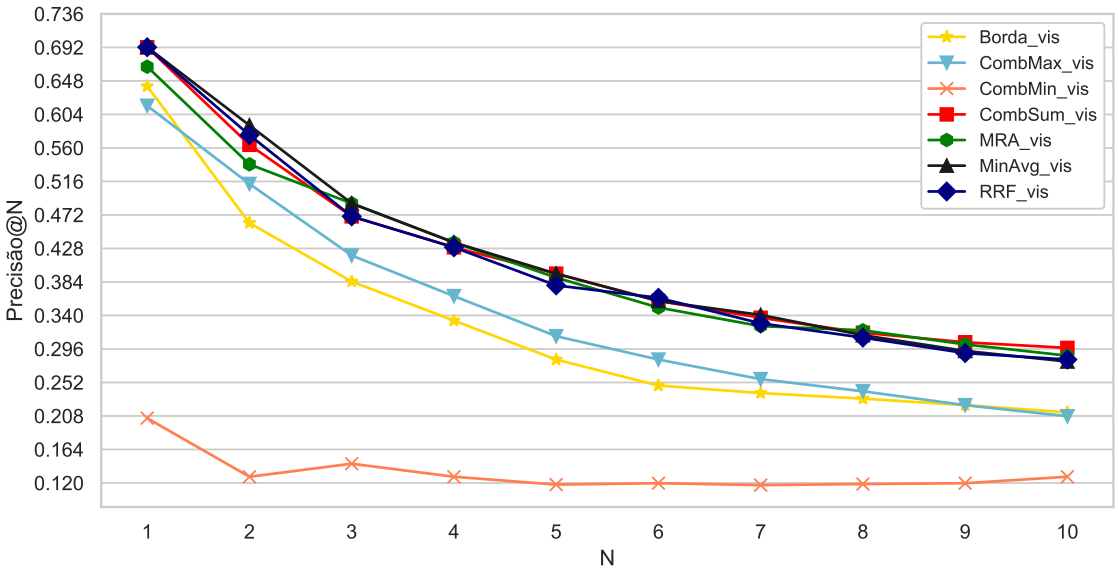

Figura 1. Comparativos dos resultados utilizando apenas informações visuais.

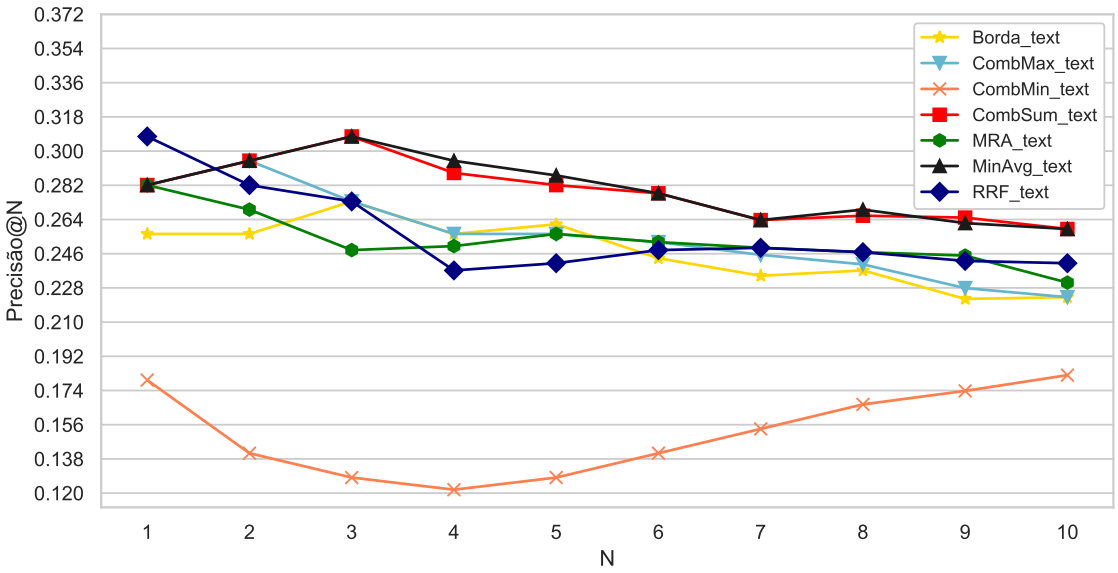

Figura 2. Comparativos dos resultados utilizando apenas informações textuais.

Os resultados demonstram que os métodos avaliados, quando utilizaram rankings baseados em apenas uma modalidade (visual ou textual), não conseguiram obter resultados superiores ao *baseline*. Entretanto, ao analisar os resultados apresentados na Figura 1, percebe-se que há uma sobreposição em termos de P@N entre o MinAvg e os métodos CombSUM, MRA, RRF. Na Figura 2, pode-se perceber que também há uma equivalência entre os métodos MinAvg e CombSUM.

A Figura 3 apresenta os resultados obtidos com cada um dos métodos de agregação no cenário multimodal (mm – informações textuais e visuais). Percebe-se que os métodos MRA e RRF apresentaram ganhos expressivos em termos de P@N em relação ao MinAvg nas primeiras posições do ranking. Além dos ganhos numéricos em termos de P@N nas

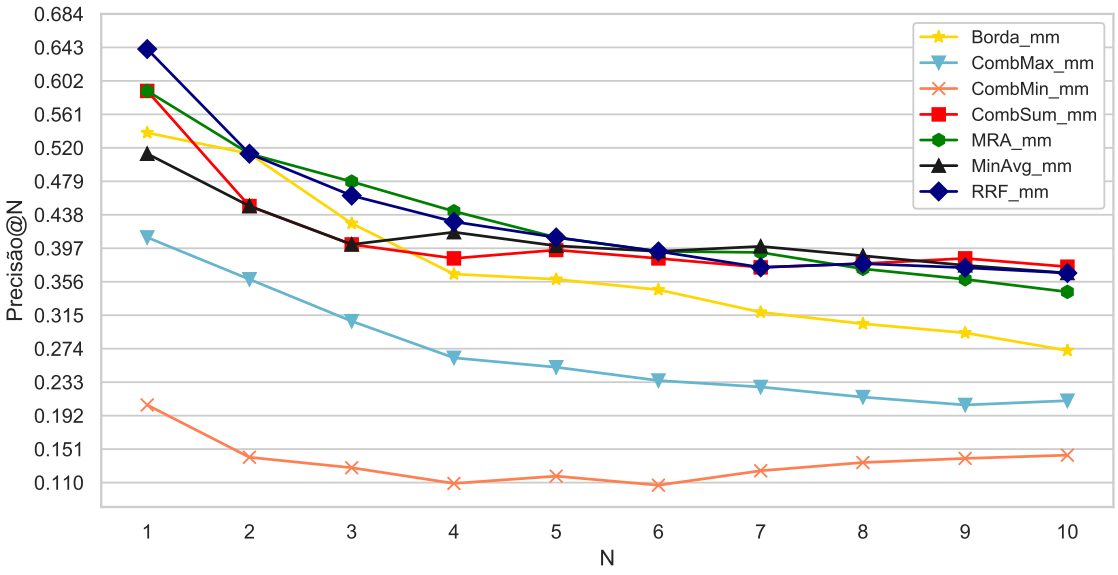

Figura 3. Comparativos dos resultados para o cenário multimodal.

primeiras posições, verificou-se com 95% de confiança por meio do teste de Wilcoxon que esses ganhos são estatisticamente significativos. Para as demais posições dos *rankings* os métodos foram considerados estatisticamente equivalentes.

5. Conclusões e Trabalhos Futuros

Neste trabalho foi apresentada uma avaliação experimental de métodos de agregação de *rankings* em cenários de buscas textuais, visuais e multimodais. Considerando a abordagem multimodal, resultados promissores foram alcançados em relação ao *baseline*. Neste cenário, ganhos expressivos no topo dos *rankings* foram observados. Estima-se que este comportamento tenha ocorrido devido à complexidade das consultas e a variabilidade de qualidade dos descritores na representação quantitativa da relevância. Nestes casos, métodos que usam critérios mais rigorosos para definir a relevância final de uma imagem permitem selecionar um conjunto reduzido, porém altamente relevante e colocá-lo no topo do *ranking*.

Os ganhos no topo do *ranking* são benéficos ao usuário, pois uma maior presença de imagens relevantes nas primeiras posições permite ao usuário fornecer *feedback* significativo sem a necessidade de inspecionar toda a lista de resultados. Vale destacar que nos cenários utilizados nos experimentos, cada um dos descritores gerou um *ranking* utilizando todas as 20.000 imagens contidas na base de dados. Portanto, uma nova etapa de experimentação pode ser realizada para avaliar a qualidade dos *rankings* gerado pelos métodos de agregação quando apenas *rankings* contendo os top-k itens são utilizados como entrada. O ajuste do tamanho dos *rankings* gerados com cada descritor poderá eliminar itens menos representativos presentes nas posições mais profundas, trazendo aos métodos de agregação a possibilidade de operar apenas com os itens mais relevantes encontrados com cada *feature*.

Agradecimentos

Este trabalho contou com o apoio do PIBIC/CNPq (processo nº 134848/2018-7).

Referências

Arni, T., Clough, P., Sanderson, M., and Grubinger, M. (2009). Overview of the imageclefphoto 2008 photographic retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 500–511. Springer Berlin Heidelberg.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Syst.*, 16(6):345–379.

Baeza-Yates, R. and Ribeiro-Neto, B. (2008). *Modern Information Retrieval: The Concepts and Technology Behind Search*. USA, 2nd edition.

Calumby, R., R. da S, T., and M. A, G. (2014). Multimodal retrieval with relevance feedback based on genetic programming. *MTAP*, (69):991–1019.

Calumby, R. T., Gonçalves, M. A., and da Silva Torres, R. (2017). Diversity-based interactive learning meets multimodality. *Neurocomputing*, 259:159 – 175.

Cormack, G. V., Clarke, C. L. A., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32Nd SIGIR*, pages 758–759. ACM.

dos Santos, K. C. L., de Almeida, H. M., Gonçalves, M. A., and da Silva Torres, R. (2009). Recuperação de imagens da web utilizando múltiplas evidências textuais e programação genética. In *Proceedings of the XXIV SBBD*, pages 91–105.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Efficient similarity search and classification via rank aggregation. In *Proceedings of the SIGMOD*, pages 301–312. ACM.

Ferreira, C., Santos, J., da S. Torres, R., Gonçalves, M., Rezende, R., and Fan, W. (2011). Relevance feedback based on genetic programming for image retrieval. *Pattern Recognition Letters*, 32(1):27 – 37.

Lewis, J., Ossowski, S., Hicks, J., Errami, M., and Garner, H. R. (2006). Text similarity: an alternative way to search MEDLINE. *Bioinformatics*, 22(18):2298–2304.

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570.

Mei, T., Rui, Y., Li, S., and Tian, Q. (2014). Multimedia search reranking: A literature survey. *ACM Comput. Surv.*, 46(3):38:1–38:38.

Penatti, O. A., Valle, E., and da S. Torres, R. (2012). Comparative study of global color and texture descriptors for web image retrieval. *Journal of Visual Communication and Image Representation*, 23(2):359 – 380.

Shaw, J. A. and Fox, E. A. (1994). Combination of multiple searches. In *TREC-2*, pages 243–252.

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

Vargas Muñoz, J. A., da Silva Torres, R., and Gonçalves, M. A. (2015). A soft computing approach for learning to aggregate rankings. In *Proceedings of the 24th CIKM*, pages 83–92. ACM.

Young, H. (1974). An axiomatization of borda's rule. *Journal of Economic Theory*, 9(1):43 – 52.