**UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA**
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
**PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO**
*COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA*

**XXV SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS**
**SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2021**

# MÉTODO DE AGREGAÇÃO PARA METABUSCA ORIENTADA POR DIVERSIDADE

**José Solenir L. Figuerêdo[1]; Ana Lúcia L. M. Maia[2]; Rodrigo Tripodi Calumby[3]**
1. Bolsista PIBIC/FAPESB, Graduando em Engenharia da Computação, UEFS, e-mail: solenir.figueredo@gmail.com
2. Orientador, Departamento de Ciências Exatas, UEFS, e-mail: anamarreiros@gmail.com
3. Coorientador, Departamento de Ciências Exatas, UEFS, e-mail: rtcalumby@uefs.br

**PALAVRAS-CHAVE:** metabusca; diversidade; agregação.

## INTRODUÇÃO

O rápido desenvolvimento de tecnologias de aquisição de dados, atrelado ao desenvolvimento e popularização das mídias sociais, levaram a uma explosão de dados, devido ao compartilhamento massivo proporcionado por esses ambientes (Lin & Zhang, 2016). A exploração dessas bases exige sistemas cada vez mais sofisticados, sendo os sistemas de Recuperação de Informação (RI) amplamente utilizados para isso. Especificamente, em relação a imagens, a abordagem mais comum que os sistemas de RI tem utilizado, é a busca baseada em informações textuais (e.g., metadados, palavras-chaves) e no uso de consultas tradicionais em bancos de dados (Chen *et al., 2010*). Outra abordagem utiliza-se da descrição das propriedades visuais (e.g., cor, forma, textura) das imagens para indexá-las e buscá-las (Korytkowski, 2017) No entanto, apesar dos avanços, a maioria desses sistemas dependem da presença de rótulos ou até mesmo da intervenção manual do usuário para obter resultados efetivos.

Em geral, sistemas diferentes tendem a apresentar respostas diferentes para a mesma necessidade de informação de um dado usuário. Neste sentido, os resultados alcançados por cada sistema costumam ser complementares. Em tais situações, métodos de fusão podem ser utilizados para explorar a complementaridade e gerar resultados de maior qualidade. A fusão refere-se a como combinar um conjunto de *rankings* diferentes, do mesmo conjunto de candidatos (*rankings* bases), a fim de que a combinação final seja melhor (mais eficaz) do que qualquer *ranking* individual (Dwork *et al.*, 2001). Uma das aplicações que tem feito uso desse processo de fusão são os sistemas de metabusca. A metabusca é a combinação de resultados obtidos de vários bancos de dados ou diferentes sistemas de pesquisa. Entretanto, em geral, os métodos de fusão propostos até o momento consideram que os resultados a serem fundidos foram construídos considerando apenas a relevância dos objetos, o que nem sempre é verdade. Estes métodos não consideram explicitamente critérios de diversidade como parte das informações de entrada.

Deste modo, neste trabalho desenvolvemos um método de agregação que considera explicitamente a informação de diversidade no processo de fusão de resultados. Para isso, foi utilizado uma estrutura denominada de grafo de preferência. Esse grafo é responsável por armazenar as relações de preferência entre um par de imagem *(x, y)* presente no *ranking*.

## MÉTODO PROPOSTO E CONFIGURAÇÃO EXPERIMENTAL

O método proposto neste estudo funciona em três etapas. A primeira etapa corresponde à conversão dos rankings em grafos de preferência, seguida da atribuição de pesos das arestas. Um grafo de preferência pode ser compreendido como uma estrutura de dados onde há uma borda ou aresta direcionada entre cada par de nós. Essa borda busca caracterizar a relação de preferência entre os itens conectados. Por exemplo, se o ranking contém uma imagem (*Img1*) acima de uma outra imagem (*Img2*), então o grafo conterá uma aresta de preferência (direcionada) do nó *Img2* para o nó *Img1*. A Figura 2 ilustra o processo de conversão.

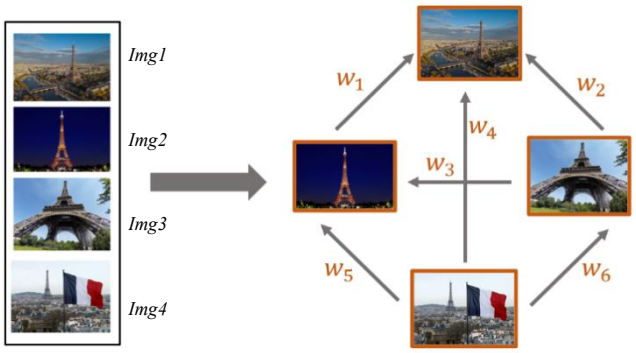


Figura 1. Conversão de um ranking de entrada em um grafo de preferência.

A atribuição de pesos utiliza a equação 1. Esta equação visa capturar a relação de diversidade existente. Com isso, busca-se favorecer os pares de imagens diversificadas, considerando que esses *rankings* bases além de terem sido gerados considerado a relevância, também usaram a diversidade como critério.

$$W = 1 - \frac{1}{1 + \frac{\mu}{\alpha + \beta}} \qquad (1)$$

Considerando um par de imagens quaisquer *(x, y)*, $\mu$ corresponderia à média da posição(decrescente) ocupada por elas no *ranking*. $\alpha$ é a quantidade de nós (imagens) que são mais preferíveis que *y*, ou seja, estão acima de *y* no *ranking*. Por sua vez, $\beta$ representa a quantidade nós entre o nó principal *(y)* e o nó secundário *(x)*. Ou seja, indica quantos pares de imagens são mais preferíveis que o par em avaliação *(x, y)*.

Na segunda etapa, tem-se a formação de um grafo agregado, considerando os grafos individuais formados anteriormente. O grafo agregado é construído tomando a combinação das relações de preferência obtidas dos *rankings*. A combinação dos pesos da aresta *(x, y)* para o novo grafo é calculada como indicado na equação 2.

$$GA_{xy} = \sum G_{xyk} \qquad (2)$$

O somatório percorre todos os grafos individuais que fornecem relações de preferência para o par *(x, y)*. $G_{xyk}$ denota o peso da borda de preferência de *x* para *y* no grafo de preferência correspondente ao *ranking* de entrada $R_k$ .

Por fim, tem-se a etapa 3. Nesta etapa é feita a indução do *ranking* final, a partir das relações de preferência armazenadas no *GA*. Para tanto, diferentes abordagens podem ser aplicadas. A partir dos experimentos utilizados, definiu-se como critério selecionar os nós principais com o maior valor acumulado.

O método desenvolvido foi avaliado experimentalmente a partir da coleção utilizada na *Information Fusion for Social Image Retrieval & Diversification Task*

(Ramírez-de-la-Rosa *et al*., 2018). Essa base inclui resultados dos diversos sistemas de busca de imagens com diferentes níveis de qualidade, propostos e avaliados entre 2013 e 2016 nas tarefas *MediaEval Retrieving Diverse Social Images.* Nesta base existem resultados ranqueados para várias consultas. Além disso, inclui resultados relevantes e diversos com diferentes níveis de qualidade. O conjunto de dados é organizado em conjuntos de desenvolvimento, validação e teste. O conjunto de teste não foi considerado na avaliação experimental, uma vez que o método proposto tinha como pré-requisito a existência de *rankings* bases, ausentes para o conjunto de teste disponibilizado.

Considerando o processo de diversificação como tema central do trabalho desenvolvido, a avaliação da eficácia do método utilizou a medida de Cluster-Recall. O Cluster-Recall calcula a percentagem de *clusters* conceituais que estiveram representados no resultado diversificado. Essa medida foi calculada com base no *ground-truth* disponível. Para a análise de eficácia, essas medidas foram computadas até a 50ª posição do ranking. Como *baselines*, além da fornecida por Ramírez-de-la-Rosa *et al.* (2018), também consideramos um conjunto de algoritmos de agregação clássicos, tanto baseados em *score* quanto baseados em ordem. Os métodos baseados em score avaliados foram CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, CombMED e Multiplication Scores (MScores). Por sua vez, os métodos baseados em ordem Borda Count, Median Rank Aggregation (MRA) e Reciprocal Rank Fusion (RRF).

## RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

A Tabela 1 apresenta a eficácia do método desenvolvido, incluindo os baselines. Os valores mais altos estão em negrito. Observa-se que o método proposto neste estudo permitiu melhorar a diversidade do resultado frente aos baselines, com a vantagem de que o método desenvolvido não depende de outras informações, como o *score* de relevância, nem sempre disponível. Esse resultado é observado para a maioria dos níveis de profundidade do *ranking*, sendo mais evidente para níveis mais profundos (CR@30 a CR@50). No *Validset*, que envolve multi-tópicos, o método proposto alcança um bom desempenho frente a maioria dos baselines, com exceção do RRF, quando se considera os níveis iniciais do *ranking*. De modo geral, considerando a diversidade, os resultados apresentados revelam que a utilização do método desenvolvido é o mais indicado em ambientes reais, pois permitiria fornecer resultados mais diversificados para o usuário, melhorando a experiência desse usuário.

**Tabela 1.** Resultados para o método proposto e baselines. Os melhores resultados são destacados em negrito.

| Dataset | Método | CR@5 | CR@10 | CR@20 | CR@30 | CR@40 | CR@50 |
|---|---|---|---|---|---|---|---|
| Devset 1 | Baseline (ICPR) | 0.2331 | 0.3649 | 0.5346 | 0.6558 | 0.7411 | 0.7988 |
| | CombMAX | 0.2536 | 0.4171 | 0.6067 | 0.7316 | 0.8158 | 0.8643 |
| | CombMIN | 0.2049 | 0.358 | 0.5455 | 0.6822 | 0.7647 | 0.8201 |
| | CombSUM | 0.2694 | **0.4407** | 0.6255 | 0.7437 | 0.8261 | 0.8738 |
| | CombANZ | 0.2500 | 0.4020 | 0.5905 | 0.7237 | 0.8073 | 0.8552 |
| | CombMED | 0.2694 | **0.4407** | 0.6255 | 0.7437 | 0.8261 | 0.8738 |
| | CombMNZ | 0.2753 | 0.4370 | 0.6258 | 0.7412 | 0.8238 | 0.8708 |
| | MScores | 0.2672 | 0.4380 | 0.6235 | 0.7413 | 0.8235 | 0.8745 |
| | Borda Count | 0.1884 | 0.3090 | 0.4642 | 0.5883 | 0.6780 | 0.7437 |
| | RRF | **0.2781** | 0.4327 | 0.6235 | 0.7421 | 0.8203 | 0.8655 |
| | MRA | 0.2544 | 0.4157 | 0.6168 | 0.7353 | 0.8186 | 0.8690 |
| | **Método Proposto** | 0.2740 | 0.4393 | **0.6266** | **0.7532** | **0.8273** | **0.8798** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Devset2 | Baseline (ICPR) | 0.1316 | 0.2135 | 0.3435 | 0.4517 | 0.5364 | 0.5977 |
| | CombMAX | 0.1378 | 0.2534 | 0.4209 | 0.5375 | 0.6186 | 0.6725 |
| | CombMIN | 0.0878 | 0.1663 | 0.2804 | 0.3914 | 0.4790 | 0.5565 |
| | CombSUM | 0.1650 | 0.2861 | 0.443 | 0.5579 | 0.6390 | 0.7046 |
| | CombANZ | 0.0917 | 0.1882 | 0.3374 | 0.4480 | 0.5427 | 0.6045 |
| | CombMED | 0.1650 | 0.2861 | 0.4430 | 0.5579 | 0.6390 | 0.7046 |
| | CombMNZ | 0.1685 | 0.282 | 0.4525 | 0.5692 | 0.6417 | 0.7041 |
| | MScores | 0.1665 | 0.2797 | 0.4433 | 0.5616 | 0.6362 | 0.7018 |
| | Borda Count | 0.0885 | 0.1541 | 0.2694 | 0.3665 | 0.4238 | 0.4948 |
| | RRF | **0.1692** | **0.2906** | 0.4586 | 0.5676 | 0.6367 | 0.7052 |
| | MRA | 0.1665 | 0.2797 | 0.4433 | 0.5616 | 0.6362 | 0.7018 |
| | **Método Proposto** | 0.1668 | 0.2878 | **0.4658** | **0.5695** | **0.6428** | **0.7138** |
| Validset | Baseline (ICPR) | 0.1489 | 0.2402 | 0.3684 | 0.4616 | 0.5284 | 0.5851 |
| | CombMAX | 0.1658 | 0.2704 | 0.4207 | 0.5230 | 0.6084 | 0.6716 |
| | CombMIN | 0.1188 | 0.2019 | 0.3282 | 0.4152 | 0.4876 | 0.5440 |
| | CombSUM | 0.1757 | 0.2812 | 0.4256 | 0.5404 | 0.6246 | 0.6875 |
| | CombANZ | 0.1390 | 0.2274 | 0.3755 | 0.4639 | 0.5436 | 0.5956 |
| | CombMED | 0.1757 | 0.2812 | 0.4256 | 0.5404 | 0.6246 | 0.6875 |
| | CombMNZ | 0.1772 | 0.2893 | 0.4344 | 0.5494 | 0.6326 | 0.6951 |
| | MScores | 0.1760 | 0.2825 | 0.4298 | 0.5391 | 0.6254 | 0.6861 |
| | Borda Count | 0.1208 | 0.2049 | 0.3161 | 0.3972 | 0.4643 | 0.5191 |
| | RRF | **0.1776** | **0.3008** | **0.4570** | **0.5596** | 0.6388 | 0.7025 |
| | MRA | 0.1546 | 0.2641 | 0.4085 | 0.5310 | 0.6119 | 0.6744 |
| | **Método Proposto** | 0.1744 | 0.2955 | 0.4482 | 0.5594 | **0.6451** | **0.7118** |

## CONSIDERAÇÕES FINAIS

Este trabalho propõe um método de agregação orientado à diversidade, aplicado ao contexto de metabusca. Nossos resultados experimentais sugerem que o método proposto permitiu obter os melhores resultados em termos de diversidade. Futuramente pretende-se incluir uma estratégia de ponderação dos *rankings* de entrada, tendo em vista que os sistemas apresentam níveis de qualidade diferentes. Também será avaliada outras formas de atribuição dos pesos para as relações de preferência.

## REFERÊNCIAS


LIN, X., ZHANG, T. (2016). Image search reranking with relevance, diversity and topic coverage. In: Proceedings of the International Conference on Internet Multimedia Computing and Service. p. 105-109.

CHEN, Y., YU, N., LUO, B., & CHEN, X. W. (2010). iLike: Integrating visual and textual features for vertical search. In: Proceedings of the 18th ACM international conference on Multimedia. 2010. p. 221-230.

KORYTKOWSKI, M. (2017). Novel visual information indexing in relational databases. Integrated Computer-Aided Engineering, v. 24, n. 2, p. 119-128.

DWORK, C., KUMAR, R., NAOR, M., & SIVAKUMAR, D. (2001). Rank aggregation methods for the web. In: Proceedings of the 10th international conference on World Wide Web. p. 613-622.

RAMÍREZ-DE-LA-ROSA, G., VILLATORO, E., IONESCU, B., ESCALANTE, H. J., ESCALERA, S., LARSON, M., MÜLLER, H.,GUYON, I. (2018). Overview of the multimedia information processing for personality & social networks analysis contest. In ICPR'18, Beijing, China. Revised Selected Papers, páginas 127–139.