UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto n°17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIENTÍFICA E TECNOLÓGICA - 2019

Fusão de *rankings* não supervisionada para metabusca de imagens diversificadas

José Solenir Lima Figuerêdo[1] e Rodrigo Tripodi Calumby[2]
1. Bolsista PIBIC/CNPq, Graduando em Engenharia da Computação, UEFS, e-mail: jslfigueredo@ecomp.uefs.br
2. Orientador, Departamento de Ciências Exatas, UEFS, e-mail: rtcalumby@uefs.br

PALAVRAS-CHAVE: fusão de rankings; diversidade; metabusca.

INTRODUÇÃO

Grandes quantidades de dados exige o uso de técnicas eficazes na sua exploração. A fim de maximizar a qualidade dos resultados da pesquisa, sistemas sofisticados são desenvolvidos para explorar o máximo de informações possíveis para determinar a relevância dos objetos nas bases de dados (Calumby *et al*., 2016). Isso melhora os algoritmos de ranqueamento e, consequentemente, às expectativas dos usuários. Contudo, dada a sua complexidade, diferentes sistemas tendem a dar respostas diferentes à mesma necessidade de informação. Assim, os resultados alcançados por cada sistema tendem a ser complementares. Uma solução proposta para esse cenário, conhecida como metabusca, é a combinação de resultados obtidos de vários bancos de dados ou de diferentes sistemas de busca. Uma abordagem popular usada para essa tarefa é aplicar algoritmos de agregação de *rankings* (Farah & Vanderpooten, 2007).

Em um cenário mais específico, os usuários podem não ser capazes de expressar adequadamente suas necessidades de informação, levando à formulação de consultas mal especificadas ou ambíguas (Santos *et al*., 2015). Além disso, dado o modelo de construção de resultados visando a maximização da relevância, os sistemas eventualmente produzem listas de resultados contendo objetos que são consideravelmente semelhantes (redundantes) ou não incluem necessariamente as diferentes possibilidades de respostas que os usuários podem encontrar na coleção (baixa cobertura). Uma abordagem usada para minimizar esses problemas é a introdução da diversidade no conjunto de resultados. Assim, este estudo descreve e investiga o uso de métodos de agregação de *rankings* para metabusca em recuperação de imagens. Embora os sistemas de metabusca normalmente sejam orientados pela relevância do resultado final, neste trabalho investigamos a diversificação de resultados via agregação, no qual a relevância e a diversificação são consideradas conjuntamente.

CONFIGURAÇÃO EXPERIMENTAL

Para a avaliação experimental, utilizou-se a coleção fornecida pela *Information Fusion for Social Image Retrieval & Diversification Task* (Ramírez-de-la-Rosa *et al*., 2018). Esta coleção inclui conjuntos de resultados de vários sistemas para várias consultas. Além disso, estes resultados são considerados relevantes e diversificados com diferentes níveis de qualidade. A coleção inclui os resultados de diferentes sistemas de

recuperação de imagens propostos e avaliados entre 2013 e 2016 nas tarefas *MediaEval Retrieving Diverse Social Images*[1]. Os dados estão organizados em conjuntos de desenvolvimento (*devset1* e *devset2*), validação (*validset*) e teste (*testset*). Uma vez que o ground-truth não estava publicamente disponível, o conjunto de teste não foi considerado na avaliação experimental. Muitos métodos de agregação de *rankings* foram considerados. Como métodos baseados em pontuação, nós consideramos: CombMAX, CombMIN, CombSUM, CombANZ, CombMNZ, CombMED and Multiplication Scores (MScores). Por sua vez, baseado em posição, consideramos o Borda Count, Median Rank Aggregation (MRA), and Reciprocal Rank Fusion (RRF) (Muñoz et al. 2015)

Precision e ClusterRecall (Zhai *et al*., 2003) são utilizadas para avaliação da eficácia. Para análise de eficácia, essas medidas foram computadas até a 50ª posição do *ranking*. Como baseline, além de usar aquele fornecido por (Ramírez-de-la-Rosa *et al*., 2018), também consideramos o melhor sistema de cada dataset. A seleção desses sistemas levou em consideração a Precision (PR@20) e ClusterRecall (CR@20). Esse ponto de corte (@20) simula o conteúdo de uma única página de um típico mecanismo de pesquisa de imagens da Web e reflete o comportamento do usuário, ou seja, inspecionando a primeira página de resultados.

RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta a eficácia dos métodos de fusão, incluindo os baselines. Os valores mais altos estão em negrito. Esses valores são usados na comparação com os baselines. Considerando os métodos de fusão, o RRF obteve melhor desempenho. Para o Devset2, além de melhorar a diversidade, não afeta negativamente a relevância dos resultados, o que é um comportamento desejável para sistemas de recuperação de imagens. A Tabela 2 apresenta uma comparação entre os métodos de fusão (tomando os de valores mais altos) e os baselines. Além disso, indica os ganhos relativos dos métodos de fusão sobre os baselines (os ganhos positivos estão em negrito). Na Tabela 2, o ICPR representa o baseline fornecido com a coleção. Por sua vez, Melhor_P e Melhor_CR representam o melhor sistema de entrada com Precision e ClusterRecall como critérios de seleção, respectivamente.

Considerando o Devset1, houve ganhos relativos para a maioria das profundidades do *ranking*. Além disso, os ganhos sobre o ICPR ocorreram em todas as profundidades do *ranking*. Na maioria das profundidades consideradas houve ganhos positivos em relação ao baseline, principalmente em relação a diversidade para o Melhor_P, com ganhos acima de 20% para todas as profundidades observadas. Já com relação ao Devset2, na maioria das profundidades consideradas houve ganhos positivos em relação ao baseline. Isso indica que, ao realizar uma busca usando uma abordagem de metabusca, o usuário obteve resultados mais relevantes e diversos. Os ganhos no validset não foram expressivos, exceto sobre o baseline do ICPR. Para os outros baselines, os ganhos obtidos foram unidirecionais, isto é, para o Melhor_P, não há ganho na Precision, mas no ClusterRecall. Para o Melhor_CR, não houve ganho considerando o ClusterRecall, mas na Precision. Isso ocorreu, possivelmente, devido à uma particularidade do validset que, ao contrário dos devsets (que têm apenas consultas single-topic), também contém consultas multi-topic. Por isso, exige investigações adicionais deste desafio e o desenvolvimento de métodos de fusão adequados.

[1] http://www.multimediaeval.org/

Tabela 1. Resultados para métodos de agregação de *rankings* e baselines. O valor máximo obtido para métodos de agregação de *rankings* está destacado em negrito.

Devset1												
Método	P@5	P@10	P@20	P@30	P@40	P@50	CR@5	CR@10	CR@20	CR@30	CR@40	CR@50
CombMAX	0.7602	0.7512	0.7512	0.7349	0.7242	0.7031	0.2531	0.4176	0.6069	0.7315	0.8158	0.8639
CombMIN	0.6544	0.6626	0.6744	0.6790	0.6736	0.6582	0.2049	0.3580	0.5455	0.6822	0.7647	0.8201
CombSUM	0.8287	0.8243	0.8034	0.7867	0.7626	0.7315	0.2694	0.4410	0.6255	**0.7437**	**0.8261**	0.8738
CombANZ	0.7573	0.7561	0.7469	0.7347	0.7205	0.6971	0.2500	0.4020	0.5905	0.7234	0.8073	0.8552
CombMED	0.8287	0.8243	0.8034	0.7867	0.7626	0.7315	0.2694	**0.4410**	0.6255	0.7437	0.8261	0.8738
CombMNZ	0.8456	0.8289	**0.8098**	**0.7888**	**0.7645**	**0.7344**	0.2753	0.4370	**0.6258**	0.7412	0.8238	0.8708
MScores	0.8240	0.8228	0.8044	0.7870	0.7624	0.7313	0.2672	0.4383	0.6235	0.7413	0.8235	**0.8745**
Borda Count	0.6129	0.6143	0.6213	0.6256	0.6246	0.6150	0.1884	0.3090	0.4642	0.5883	0.6780	0.7437
RRF	**0.8491**	**0.8316**	0.8092	0.7874	0.7616	0.7314	**0.2781**	0.4327	0.6235	0.7421	0.8203	0.8655
MRA	0.7614	0.7567	0.7361	0.7262	0.7092	0.6872	0.2544	0.4157	0.6168	0.7351	0.8186	0.8690
Baseline(ICPR)	0.7883	0.7558	0.7289	0.7194	0.7080	0.6877	0.2331	0.3649	0.5346	0.6558	0.7411	0.7988
Baseline(melhor_P)	0.8947	0.8936	0.8788	0.8569	0.8265	0.7891	0.2047	0.2963	0.4410	0.5476	0.6416	0.7122
Baseline(melhor_CR)	0.7409	0.7330	0.7487	0.7603	0.7145	0.5915	0.2614	0.4291	0.6314	0.7228	0.7473	0.7484
Devset2												
CombMAX	0.7467	0.7417	0.7558	0.7639	0.7333	0.7257	0.1378	0.2534	0.4209	0.5375	0.6186	0.6725
CombMIN	0.4667	0.5000	0.5025	0.5189	0.5312	0.5463	0.0878	0.1663	0.2804	0.3914	0.4790	0.5565
CombSUM	0.8667	0.8550	0.8267	0.8194	0.8075	0.7980	0.1650	0.2861	0.4430	0.5579	0.6390	0.7046
CombANZ	0.4733	0.5433	0.5808	0.5944	0.6096	0.6147	0.0917	0.1882	0.3374	0.4480	0.5427	0.6045
CombMED	0.8667	0.8550	0.8267	0.8194	0.8075	0.7980	0.1650	0.2861	0.4430	0.5579	0.6390	0.7046
CombMNZ	0.8933	0.8650	0.8417	0.8361	**0.8292**	0.8123	0.1685	0.2820	0.4525	**0.5692**	**0.6417**	0.7041
MScores	0.8733	0.8517	0.8308	0.8250	0.8075	0.8013	0.1665	0.2797	0.4433	0.5616	0.6362	0.7018
Borda Coun	0.4700	0.4567	0.4750	0.4889	0.4925	0.4993	0.0885	0.1541	0.2694	0.3665	0.4238	0.4948
RRF	**0.8967**	**0.8817**	**0.8508**	**0.8372**	**0.8292**	**0.8260**	**0.1692**	**0.2906**	**0.4586**	0.5676	0.6367	**0.7052**
MRA	0.6633	0.6733	0.6775	0.6733	0.6725	0.6797	0.1287	0.2370	0.4066	0.5185	0.6141	0.6831
Baseline(ICPR)	0.8100	0.8067	0.8058	0.8056	0.8025	0.7917	0.1316	0.2135	0.3435	0.4517	0.5364	0.5977
Baseline(melhor_P)	0.8933	0.8933	0.8550	0.8167	0.8021	0.7850	0.1461	0.2377	0.3809	0.4750	0.5531	0.6210
Baseline(melhor_CR)	0.8867	0.8600	0.8492	0.8189	0.8017	0.7933	0.1682	0.3003	0.4697	0.5594	0.6274	0.6788
Validset												
CombMAX	0.7439	0.7209	0.7065	0.7041	0.7007	0.7014	0.1658	0.2704	0.4207	0.5230	0.6084	0.6716
CombMIN	0.5597	0.5813	0.5968	0.5986	0.6031	0.6029	0.1188	0.2019	0.3282	0.4152	0.4876	0.5440
CombSUM	0.7626	0.7554	0.7320	0.7271	0.7214	0.7173	0.1757	0.2812	0.4256	0.5404	0.6246	0.6875
CombANZ	0.6230	0.6201	0.6399	0.6441	0.6550	0.6544	0.1390	0.2274	0.3755	0.4639	0.5436	0.5956
CombMED	0.7626	0.7554	0.7320	0.7271	0.7214	0.7173	0.1757	0.2812	0.4256	0.5404	0.6246	0.6875
CombMNZ	0.7683	0.7662	0.7471	0.7331	0.7243	0.7219	0.1772	0.2893	0.4344	0.5494	0.6326	0.6951
MScores	0.7612	0.7597	0.7367	0.7254	0.7212	0.7168	0.1760	0.2825	0.4298	0.5391	0.6254	0.6861
Borda Count	0.5669	0.5727	0.5770	0.5878	0.5946	0.5994	0.1208	0.2049	0.3161	0.3972	0.4643	0.5191
RRF	**0.7813**	**0.7698**	**0.7536**	**0.7405**	**0.7318**	**0.7281**	**0.1776**	**0.3008**	**0.4570**	**0.5596**	**0.6388**	**0.7025**
MRA	0.6619	0.6590	0.6680	0.6743	0.6761	0.6753	0.1546	0.2641	0.4085	0.5310	0.6119	0.6744
Baseline(ICPR)	0.7281	0.7086	0.7000	0.6952	0.6838	0.6776	0.1489	0.2402	0.3684	0.4616	0.5284	0.5851
Baseline(melhor_P)	0.8101	0.8108	0.7906	0.7736	0.7646	0.7534	0.1904	0.2908	0.4051	0.5005	0.5703	0.6246
Baseline(melhor_CR)	0.7755	0.7633	0.7309	0.7002	0.6899	0.6790	0.1935	0.3163	0.4963	0.6112	0.6933	0.7514

Tabela 2. Ganhos relativo entre o maior valor obtido pelo método de fusão e baselines

Devset1												
Baseline	P@5	P@10	P@20	P@30	P@40	P@50	CR@5	CR@10	CR@20	CR@30	CR@40	CR@50
Gain Over ICPR	**7.71%**	**10.03%**	**11.10%**	**9.65%**	**7.98%**	**6.79%**	**19.31%**	**20.86%**	**17.06%**	**13.40%**	**11.47%**	**9.48%**
Gain Over Melhor_P	-5.10%	-6.94%	-7.85%	-7.95%	-7.50%	-6.93%	**35.86%**	**48.84%**	**41.90%**	**35.81%**	**28.76%**	**22.79%**
Gain Over Melhor_CR	**14.60%**	**13.45%**	**8.16%**	**3.75%**	**7.00%**	**24.16%**	**6.39%**	**2.77%**	-0.89%	**2.89%**	**10.54%**	**16.85%**
Devset2												
Gain Over ICPR	**10.70%**	**9.30%**	**5.58%**	**3.92%**	**3.33%**	**4.33%**	**28.57%**	**36.11%**	**33.51%**	**26.01%**	**19.63%**	**17.99%**
Gain Over Melhor _P	**0.38%**	-1.30%	-0.49%	**2.51%**	**3.38%**	**5.22%**	**15.81%**	**22.25%**	**20.40%**	**19.83%**	**16.02%**	**13.56%**
Gain Over Melhor_CR	**1.13%**	**2.52%**	**0.19%**	**2.23%**	**3.43%**	**4.12%**	**0.59%**	-3.23%	-2.36%	**1.75%**	**2.28%**	**3.89%**
Validset												
Gain Over ICPR	**7.31%**	**8.64%**	**7.66%**	**6.52%**	**7.02%**	**7.45%**	**19.27%**	**25.23%**	**24.05%**	**21.23%**	**20.89%**	**20.06%**
Gain Over Melhor_P	-3.56%	-5.06%	-4.68%	-4.28%	-4.29%	-3.36%	-6.72%	**3.44%**	**12.81%**	**11.81%**	**12.01%**	**12.47%**
Gain Over Melhor_CR	**0.75%**	**0.85%**	**3.11%**	**5.76%**	**6.07%**	**7.23%**	-8.22%	-4.90%	-7.92%	-8.44%	-7.86%	-6.51%

CONCLUSÃO

Os resultados experimentais sugerem que os métodos de fusão tendem a permitir melhores resultados de busca do que sistemas independentes. Além disso, observou-se que os maiores ganhos foram em termos de diversidade, embora também tenha havido ganhos em termos de relevância. Os achados validaram a ideia de que os sistemas de metabusca podem permitir melhorias na relevância e diversidade dos resultados. Verificou-se também que alguns métodos de fusão foram eficazes o suficiente para melhorar um dos objetivos, mantendo o desempenho satisfatório para o outro. Para alguns casos, além do método de fusão alcançar altos ganhos em termos de diversidade, também manteve resultados aceitáveis de relevância. No entanto, observa-se que, para as consultas multi-topic, os resultados obtidos a partir da fusão não foram tão eficazes, em comparação com o melhor sistema individual. Isso destaca a necessidade de novas investigações, a fim de desenvolver métodos de fusão capazes de otimizar tanto a diversificação quanto a relevância para consultas multi-topic.

REFERÊNCIAS

CALUMBY, R. T., GONÇALVES, M. A., and da SILVA Torres, R. (2016). On interactive learning-to-rank for IR: overview, recent advances, challenges, and directions. Neurocomputing, 208:3–24.

FARAH, M. and VANDERPOOTEN, D. (2007). An outranking approach for rank aggregation in information retrieval. In SIGIR'07, Amsterdam, The Netherlands, July 23-27, 2007, páginas 591–598.

MUÑOZ, J. A. V., da Silva Torres, R., and Gonçalves, M. A. (2015). A soft computing approach for learning to aggregate rankings. In CIKM'15, Melbourne, VIC, Australia, Outubro 19 - 23, 2015, páginas 83–92.

RAMÍREZ-DE-LA-ROSA, G., VILLATORO, E., Ionescu, B., Escalante, H. J., Escalera, S., Larson, M., Müller, H., and Guyon, I. (2018). Overview of the multimedia information processing for personality & social networks analysis contest. In ICPR'18, Beijing, China, Agosto 20-24, 2018, Revised Selected Papers, páginas 127–139.

SANTOS, R. L. T., MACDONALD, C., and OUNIS, I. (2015). Search result diversification. Foundations and Trends in Information Retrieval, 9(1):1–90.

ZHAI, C. X., COHEN,W.W., and LAFFERTY, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In ACM SIGIR.