Per Query Subtopic Discovery for Diverse Image Retrieval

José Solenir L. Figuerêdo[1], **Rodrigo Tripodi Calumby**[1]

[1] Universidade Estadual de Feira de Santana
Av. Transnordestina, s/n, Novo Horizonte, Feira de Santana – BA, Brasil – 44036-900

solenir.figueredo@gmail.com, rtcalumby@uefs.br

Abstract. *Given the complex search tasks imposed to multimedia retrieval systems, the similarity-based results often represent redundant item sets. Several real-world search tasks demand broad coverage of multiple implicit subtopics of a given query. Many works have proposed the use of clustering-based result diversification for addressing such problem. However, the definition of the number of clusters (subtopics) to be discovered is a long-lasting challenge. In order to attenuate such problems, this work proposes a novel diverse image retrieval approach as an unsupervised query-adaptive subtopic discovery based on intrinsic clustering quality optimization. Our experimental analysis have shown significant improvements, both in terms of relevance and diversity.*

1. Introduction and Background

Multimedia retrieval systems face multiple obstacles related to user query definition. In general, these queries may present subjectivity, ambiguity, or are under-specified. Furthermore, the results generated based solely on similarity may introduce duplicates or non-representative items. In order to address these challenges, some works have proposed the introduction of result diversification methods into search engines [Santos et al. 2015]. Clustering is one of the most used techniques to promote diversity. Diversification is achieved by grouping similar images from an original result set. With this approach, several parameters must be selected, specially the number of clusters to be generated. The research community on clustering optimization methods have proposed several alternatives for the selection of the best number of clusters [Muhlenbach and Lallich 2009, Salvador and Chan 2004]. Different approaches, both offline and online have already been proposed. However, they often do not generalize effectively.

Aiming at optimizing the clustering parameters some works have relied on separate training sets for the selection of a general top performing configuration [Tollari 2016]. However, considering the high heterogeneity of the context (users, queries, subtopics, datasets, features, etc.), such approaches usually suffer from under or over-fitting achieving sub-optimal effectiveness on validation sets. Thereby, in order to reduce such problems, this work proposes a novel diverse image retrieval approach based on a query-adaptive unsupervised clustering optimization. The results show that our method, with proper configurations, allows the display of more relevant and more diverse images for the users. Thus, there is an information gain with respect to the baseline. To the best of our knowledge this is the first work to propose a query-adaptive clustering optimization for adjusting the number of clusters in the context of diverse image retrieval.

2. Proposed Method

Our method follows a common diverse image retrieval workflow, which consists of the following steps: i) relevance-based dataset ranking according to the query; ii) ranked list filtering based on multiple criteria (non-relevant images removal); iii) selection of the top of the ranking for clustering (most relevant images); iv) implicit subtopic discovery (clustering); and v) representative selection (round-robin) and diverse result presentation. Different from some works, rather of using a fixed clustering configuration for any given query, in step iv, an unsupervised optimization is conducted. The proposed method is illustrated in Figure 1. In order to optimize the number of clusters during the diversification process, for each query, this work proposes an unsupervised method based on intrinsic quality measures for clustering. Hence, instead of using a single number of clusters for all queries, a query-specific configuration is selected.

For reducing the computational cost, our method only considers a reasonable interval for the possible number of clusters ($K_{min} \leq K \leq K_{max}$). The input ranked list is submitted to clustering using each value in the interval of the possible number of clusters. Each configuration is evaluated and the best performing one is selected for the clustering-based diversity promotion. Given it is an adaptive optimization framework, the proposed method may be instantiated with different features, clustering algorithms, clustering quality measures, and the interval for the number of clusters.

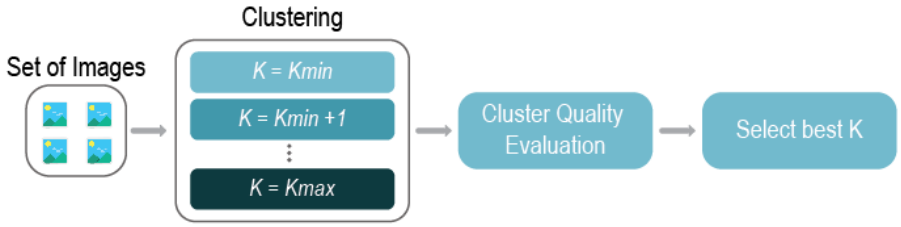

Figure 1. Proposed method workflow.

3. Experimental Setup

The image collection from the *Retrieving Diverse Social Images Task from MediaEval 2015* [Ionescu et al. 2015] was used for the experimental evaluation of the proposed method. The collection has 222 queries. Each query has a textual phrase which was submitted to the Flickr [1] image search engine. The relevance and diversity ground-truth were generated by human annotators and provided along with the collection. For the clustering step, the k-Medoids algorithm was applied considering it has been effectively adopted in previous works [Calumby et al. 2017]. The similarity between images were computed based on multiple visual features (ACC, CN3x3, LUM, SCH, Gabor, Tamura, CEDD, FCTH, JCD, PHOG, and CNN_AD) and text similarity measures (Cosine, BM25, Dice, and Jaccard). Figure 2 illustrates the evaluation workflow for a given query. The images retrieved from Flickr are initially reranked according to their textual relevance based on BM25 and the top-150 images are selected for diversification.

Many clustering quality measures were considered for the unsupervised optimization such as: Silhouette Coefficient, Davies-Bouldin Index, Dunn Index, Squared Error,

[1] http://www.flickr.com (As of Jan 2019).

and Xie-Bie Index [Xie and Beni 1991]. The queries are executed for all $15 \leq K \leq 25$. This interval encompasses negative and positive variations, w.r.t the number of clusters commonly used in the literature (K =20). $K = 20$ is taken as the baseline configuration for the approach with a fixed number of clusters. Precision and Cluster-Recall [Zhai et al. 2003] are used for effectiveness assessment. While precision represents the quality of the ranking in terms of relevance, the Cluster-Recall measure computes the percentage of conceptual clusters that are covered in the diversified result. For effectiveness analysis, these measures where computed up to the 50th position of the ranking. For an strict comparative to the baseline, the Wilcoxon's Signed Rank Test is performed in order to assess the statistical significance with 95% confidence.

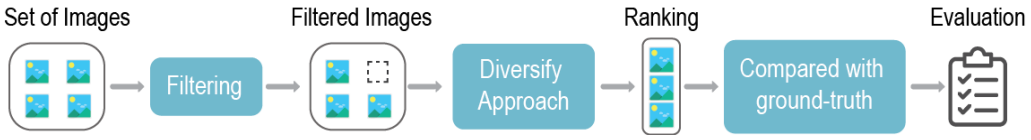

Figure 2. Experimental evaluation workflow.

4. Results

Table 1 presents a summary of the effectiveness results of the baseline and the proposed method for different clustering quality measures. Only the most significant and illustrative results covering a subset of the evaluated image similarity measures are presented. The analysis of the experiments demonstrated superior effectiveness achievements of the proposed method. Except for the CEDD, in general, our method allowed important improvements in terms of relevance at the top of the ranking ($N \leq 20$). On the other hand, the enhancements in terms of diversity, were generally achieved at the end of the ranking ($N \geq 30$). Assuming relevance and diversity as generally opposite optimization objectives, our method effectively allowed improvements for both of them. Beyond it, while one of the objectives was statistically improved the other as kept equivalent in relation to the baseline. In addition, considering the execution based on textual information with the cosine measure, our method produced statistically superior results for both objectives simultaneously. These results show that the proposed unsupervised query-adaptive optimization method was able to capture intrinsic quality information and use it as an indicator for a better discovery of implicit query subtopics.

5. Conclusions

The proposed method achieved statistical significant improvements against the baseline, both in terms of relevance and diversity. The results also highlighted that achieving the best effectiveness is a consequence of the adequate combination of features and clustering quality optimization method. It is important to notice that while the proposed method achieves equivalent or superior effectiveness, it eliminates the necessity of offline optimization or system re-training for new data. As future work, it may be promising to integrate this solution to other unsupervised query-adaptive strategies, include other parameters in the optimization process and consider additional quality measures or even a combination of them.

Table 1. Experimental Effectiveness Results. Statistical significance is reported in relation to the baseline: Superiority is highlighted in boldface; All the rest is equivalent.

CEDD												
N	Precision						Cluster-Recall (CR)					
	K=20	DB	Dunn	Silhouette	Error	XB	K=20	DB	Dunn	Silhouette	Error	XB
5	0.7946	0.8027	0.8009	0.8009	0.7937	0.8090	0.1536	0.1601	0.1558	0.1568	0.1548	**0.1624**
10	0.7586	0.7563	0.7635	0.7563	0.7595	0.7640	0.2697	0.2688	0.2670	0.2656	0.2637	0.2727
20	0.7565	0.7637	0.7628	0.7590	0.7534	0.7653	0.4224	0.4142	0.4138	0.4189	0.4260	0.4212
30	0.7599	**0.7740**	**0.7727**	**0.7682**	0.7571	**0.7760**	0.5254	0.5165	0.5181	0.5212	0.5302	0.5208
40	0.7658	0.7709	0.7709	0.7671	0.7624	**0.7725**	0.5917	0.5842	0.5870	0.5887	0.5933	0.5855
50	0.7596	0.7632	**0.7646**	0.7622	0.7605	**0.7662**	0.6474	0.6404	0.6422	0.6473	0.6511	0.6428
ACC												
N	Precision						Cluster-Recall (CR)					
	K=20	DB	Dunn	Silhouette	Error	XB	K=20	DB	Dunn	Silhouette	Error	XB
5	0.7964	0.7973	0.8072	0.7973	**0.8162**	0.8063	0.1523	0.1537	0.1568	0.1535	0.1552	0.1572
10	0.7896	0.7910	0.7914	0.7946	0.7977	0.7937	0.2621	0.2607	0.2620	0.2606	0.2646	0.2620
20	0.7752	0.7750	0.7829	0.7782	0.7829	0.7755	0.4198	0.4133	0.4176	0.4204	0.4162	0.4217
30	0.7698	0.7734	0.7745	0.7730	0.7754	0.7736	0.5202	0.5204	0.5228	0.5241	0.5249	0.5289
40	0.7609	0.7668	0.7660	0.7626	0.7662	0.7637	0.5891	0.5958	0.5903	**0.6029**	0.5935	**0.6014**
50	0.7554	0.7595	0.7559	0.7581	0.7576	0.7600	0.6462	0.6508	0.6476	**0.6611**	0.6502	0.6535
COSINE												
N	Precision						Cluster-Recall (CR)					
	K=20	DB	Dunn	Silhouette	Error	XB	K=20	DB	Dunn	Silhouette	Error	XB
5	0.7892	**0.8063**	0.8027	0.7973	0.7937	0.7964	0.1589	0.1619	0.1630	0.1609	0.1573	0.1579
10	0.7734	**0.7928**	0.7811	0.7757	0.7766	0.7770	0.2728	**0.2832**	**0.2824**	0.2739	0.2741	0.2726
20	0.7565	**0.7691**	0.7651	0.7649	0.7604	0.7606	0.4272	0.4284	0.4319	0.4308	0.4252	0.4250
30	0.7553	0.7619	0.7586	0.7598	0.7571	0.7569	0.5256	0.5243	0.5315	0.5336	0.5252	0.5278
40	0.7519	0.7586	0.7546	0.7570	0.7550	0.7530	0.5961	0.5947	0.6006	0.6039	0.5938	0.5959
50	0.7516	0.7571	0.7542	0.7546	0.7530	0.7523	0.6486	0.6498	0.6546	0.6566	0.6486	0.6521
CN3X3												
N	Precision						Cluster-Recall (CR)					
	K=20	DB	Dunn	Silhouette	Error	XB	K=20	DB	Dunn	Silhouette	Error	XB
5	0.8027	**0.8252**	0.8054	0.8162	0.8027	**0.8207**	0.1587	0.1625	0.1607	0.1639	0.1606	0.1637
10	0.7950	0.7982	0.7982	0.7905	0.8041	0.7973	0.2747	0.2750	0.2781	0.2762	0.2775	0.2697
20	0.7755	0.7759	0.7795	0.7773	0.7802	0.7827	0.4380	0.4356	0.4347	0.4391	0.4333	0.4301
30	0.7686	0.7656	0.7694	0.7671	0.7718	0.7688	0.5281	0.5318	0.5310	**0.5429**	0.5274	0.5314
40	0.7556	0.7587	0.7589	0.7581	0.7625	0.7609	0.6009	0.6047	0.5909	0.6062	0.5937	0.6059
50	0.7526	0.7520	0.7517	0.7525	0.7586	0.7555	0.6517	**0.6610**	0.6471	**0.6602**	0.6446	**0.6613**

References

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

Ionescu, B., Gînsca, A., Boteanu, B., Popescu, A., Lupu, M., and Müller, H. (2015). Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval*.

Muhlenbach, F. and Lallich, S. (2009). A new clustering algorithm based on regions of influence with self-detection of the best number of clusters. In *IEEE ICDM*.

Salvador, S. and Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *IEEE ICTAI*.

Santos, R. L. T., Macdonald, C., and Ounis, I. (2015). Search result diversification. *Found Trends Inf Ret*, 9(1):1–90.

Tollari, S. (2016). UPMC at mediaeval 2016 retrieving diverse social images task. In *MediaEval*.

Xie, X. L. and Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE TPAMI*, 13(8):841–847.

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR*.