TUW @ MediaEval 2015
Retrieving Diverse Social Images Task

Serwah Sabetghadam, João Palotti, Navid Rekabsaz, Mihai Lupu, Allan Hanbury
Favoriten Strasse 9-11/188
Vienna University of Technology
Vienna, Austria
{sabetghadam, palotti, rekabsaz, lupu, hanbury}@ifs.tuwien.ac.at

ABSTRACT

This paper describes the contributions of Vienna University of Technology (TUW) to the MediaEval 2015 Retrieving Diverse Social Images challenge. Our approach consists of 3 phases: (1) Precision-oriented-phase: in which we focus only on the relevance of the documents; (2) Recall-oriented-phase: in which we focus only on the diversity aspect; (3) Merging phase: in which we explore ways to find a balance between the relevance and diversity factors. We use two fusion methods for this last part. Our best run reached a F1@20 of 0.582.

1. INTRODUCTION

Result diversification has recently attracted much attention in the IR community. Often, the information need requested by the users cannot be found by displaying items related only to one facet of the query topic. Ideally an IR system displays pieces of information covering diverse subtopics of the query. The same idea has been used in the Recommender System area, where diversification techniques has shown to increase user satisfaction [9, 10].

This paper describes the second participation of our team at MediaEval Retrieving Diverse Social Images task [3]. We build our solution upon our previous participation [7]. Last year, we had good results in both precision and recall, but in separated runs. Therefore, we decided to explore different strategies for better fusing our individual runs.

2. METHODS

We leveraged a distinct set of methods for each run. We show the combinations used for each run in Table 1.

2.1 Relevancy

Regarding the experience of the previous year [7], we use only textual features (i.e. title, tag, description) for finding the relevant documents. We extend the usual term-frequency-based methods to more semantic-based approach [8]. We create word embeddings using the Wikipedia corpus with 400 dimension by Word2Vec method [6]. We calculate the similarity between the query and the text documents (concatanation of title, tag, and description) using the *SimGreedy* method [8].

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2.2 Diversity

To find diverse images we experiment with different clustering methods. From [7] we learned that an approach based on ensemble of clusters can perform better than using only one single clustering method. We also learned that a prefiltering step can potentially remove irrelevant images that harm the process of clustering creation. Here we briefly comment on these two aspects:

Pre-Filtering: We use hand-coded rules previously shown to perform well in this task, to exclude probably irrelevant pictures [4]. We exclude pictures based on three rules: without any views, geo-tagged 8km away from the POI, or with description length greater than 2000 characters.

Clustering solution: The basic idea is that, given a clustering algorithm A, a feature set F that describes an image and a distance measure Di, we can create a cluster set $C = (A, F, Di)$. For example, C_1 can be the result of applying K-Means (A) using the Color Histogram of the images (F), based on the cosine distance (Di): C_1 = (K-Means, ColorHistogram, Cosine).

A common strategy used by a number of teams in 2013 was to go one by one of the clusters made in C_1 and pick the "best“ image from each cluster to form the final ranked list. We noticed that small differences, for example having C_2 = (K-means, NeuralNetworkFeatures, Cosine), could have a large impact in the clusters formed, consequently strongly influencing the final ranked list. As described in [7], our solution is to use the development set to learn what are the best clustering algorithm, features sets, and distance measures. After that, we combine the results of different Cs and count the frequency that any two images end up in the same cluster. Based on this simple frequency, we rerank the initial Flickr list (Run 1) or the list generate by the algorithm in Section 2.1 (Run 3).

2.3 Fusion

Atrey et al. [1] performed a survey on fusion methods of combining multiple modalities. In their view, there are three category of methods for fusion: rule-based methods, classification-based methods and estimation-based methods. Our approach is inspired from these fusion methods for combining relevancy and diversity results. We leverage the weighted linear method from the first category, and Bayesian inference from the second category.

Weighted Linear: We use the optimization technique proposed by Deselaers et al. [2] based on weighted linear fusion. Having the relevance of the query to each document (R) and also the diversification measure for each set of documents (D), we formulate the diversification issue as

Table 1: Official runs setup. Features used in clustering are Combined on CN3x3 and CNN in all runs [3]. The relevancy is based on the Word2Vec method [6] - see Section 2.1.; diversity approach is presented in Section 2.2.; fusion mechanism in Section 2.3.

Run	Type	Relevancy	Diversity		Fusion
			Pre-Filtering	Clustering	
1	image	-	Based on [4]	✓	-
2	text	✓	-	-	-
3	text, image	✓	-	✓	-
4	text, image	✓	-	✓	Linear Fusion
5	text, image	✓	Based on [4]	✓	Bayesian Fusion

Table 2: Results for the development and test set at various cutoff points. Official metric is F1@20.

Run	2015 Development Set						2015 Test Set					
	P@10	CR@10	F1@10	P@20	CR@20	F1@20	P@10	CR@10	F1@10	P@20	CR@20	F1@20
1	0.8137	**0.2873**	**0.4189**	**0.7817**	**0.4713**	**0.5806**	0.7201	0.2908	0.4025	0.7058	0.4705	0.5487
2	**0.8203**	0.2188	0.3389	0.798	0.3485	0.4766	**0.7842**	0.2576	0.3728	**0.7687**	0.3914	0.4968
3	0.7804	0.2748	0.4014	0.7546	0.4531	0.5583	0.7633	**0.3163**	**0.4309**	0.7309	0.4963	**0.5727**
4	0.8157	0.2867	0.4184	0.7782	0.4616	0.5741	0.7345	0.3005	0.4128	0.7291	0.4767	0.5601
5	0.8137	**0.2873**	**0.4189**	0.7804	0.4706	0.5796	0.7216	0.2906	0.4026	0.7076	0.4702	0.5492

Table 3: Results based on single and multi topics - the best run according to the official metric is Run3.

Run	2015 Test Set - Single-concept queries						2015 Test Set - Multi-concept queries					
	P@10	CR@10	F1@10	P@20	CR@20	F1@20	P@10	CR@10	F1@10	P@20	CR@20	F1@20
1	0.7232	0.2904	0.4045	0.6942	0.4807	0.553	0.7171	0.2912	0.4005	0.7171	0.4605	0.5445
2	**0.8188**	0.2589	0.3772	**0.808**	0.4038	0.5202	**0.7500**	0.2563	0.3684	**0.7300**	0.3793	0.4738
3	0.7928	**0.3237**	**0.4443**	0.7326	**0.5037**	**0.5802**	0.7343	**0.3091**	**0.4177**	0.7293	**0.489**	**0.5654**
4	0.7507	0.3062	0.4206	0.7355	0.4798	0.5664	0.7186	0.2948	0.4052	0.7229	0.4737	0.554
5	0.7261	0.2907	0.4053	0.6935	0.4788	0.5515	0.7171	0.2905	0.4000	0.7214	0.4617	0.5469

an optimization problem where one tries to maximize the linear combination of these two values.

$$U(S|q) = w * R(S|q) + (1 - w) * D(S) \qquad (1)$$

where U denotes the score for the selected set S regarding to the query q, and w is a parameter which controls the importance of relevance and diversity. The parameter w is tuned using the development set.

Bayesian Inference: In this method the information is combined based on the rules of the probability theory [5]. The probability of a hypothesis H of diversification is:

$$P(H|R, D) = 1/2P(D|H)^{w_d} P(R|H)^{w_r} \qquad (2)$$

where w_d and w_r are weights given to diversity and relevancy results.

3. EXPERIMENTS

We submitted 5 runs, varying on the use of relevancy results, pre-filtering, different clustering algorithms, and fusion methods. Details of the run configurations are shown in Table 1. Run 1 is based on pure diversity results using image features. Run 2 uses only text information, we apply Word2Vec [8] semantic similarity. In Run 3, the input of diversity algorithm is the Run 2 ranked results. In this run, we leverage both modalities of text and image similarity in clustering the images. In the Run 4 and Run 5 we use two fusion methods of weighted linear and Bayesian reference on Run 1 (diversity) and Run 2 (relevancy) results.

Based on our development tests, we expected the Run 1 and Run 5 to achieve better results according to the F1 measure (Table 2). However, based on the test set results, we observe that the Run 3 obtains the best value for F1@10 with 0.43 and F1@20 with 0.57. One reason could be the multi-concept queries in the test runs. It shows that the semantic text similarity result (Run 2) as input to the clustering algorithms (Run 3) improved the F1 measure by 4%. We receive the best precision (0.82) in the Run 2 which is purely based on text similarity results.

In the experiments of this year, we added two runs based on fusion of relevancy and diversity results. In the development tests we reached the optimum weighting of $0.2 \cdot R + 0.8 \cdot D$ for both methods. Although we obtained better result with Bayesian inference approach in the development tests, with the test data, weighted linear fusion has the second place in F1@20 measure. This confirms the approach that Deselaers et al. [2] used in the score combination. However, Bayesian inference is usually used on classification results, which may explain why in our case the linear combination performed better on the test data.

In Table 3 we show separate results for single and multi-concept topics. We observe the same order of results here. The Run 3 keeps the best value of F1@20 and Run 2 the highest result in P@20.

4. CONCLUSION

Our experiments show that the cluster ensemble with input of relevancy results (Run 3) provides robust results for this task. The input of this run was our relevancy results based on text semantic similarity results. This demonstrates that the combination of text similarity and diversity approach leads to higher F1@20 value. This year we added two fusion methods of weighted linear and Bayesian inference. Their results were indistinguishable on the devset, but the weighted linear fusion outperfomed the Bayesian on the testset.

5. REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 2010.

[2] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM international conference on image and video retrieval*, 2009.

[3] B. Ionescu, A. Ginsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation. 2015.

[4] N. Jain, J. Hare, S. Samangooei, J. Preston, J. Davies, D. Dupplaw, and P. H. Lewis. Experiments in diversifying flickr result sets. In *MediaEval 2013*, 2013.

[5] R. C. Luo, C.-C. Yih, and K. L. Su. Multisensor fusion and integration: approaches, applications, and future research directions. *Sensors Journal, IEEE*, 2002.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] J. Palotti, N. Rekabsaz, M. Lupu, and A. Hanbury. Tuw@ retrieving diverse social images task 2014. In *MediaEval*, 2014.

[8] N. Rekabsaz, R. Bierig, B. Ionescu, A. Hanbury, and M. Lupu. On the use of statistical semantics for metadata-based social image retrieval. In *CBMI*, 2014.

[9] E. Vee, U. Srivastava, J. Shanmugasundaram, P. Bhat, and S. A. Yahia. Efficient computation of diverse query results. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, 2008.

[10] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, 2005.