LAPI @ 2017 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversification Perspective

Bogdan Boteanu, Mihai Gabriel Constantin, Bogdan Ionescu
LAPI, University "Politehnica" of Bucharest, Romania
{bboteanu,mgconstantin,bionescu}@alpha.imag.pub.ro

ABSTRACT

In this paper we present the results achieved during the 2017 MediaEval Retrieving Diverse Social Images Task, using an approach based on pseudo-relevance feedback (RF), in which human feedback is replaced by an automatic selection of images. The proposed approach is designed to have in priority the diversification of the results, in contrast to most of the existing techniques that address only the relevance. Diversification is achieved by exploiting a hierarchical clustering (HC) scheme followed by a diversification strategy. Methods are tested on the benchmarking data and results are analyzed. Insights for future work conclude the paper.

1 INTRODUCTION

An efficient information retrieval system should be able to provide search results which are in the same time *relevant* for the query and cover different aspects of it, i.e., *diverse*. The 2017 Retrieving Diverse Social Images Task [7] addresses this issue in the context of a general ad-hoc image retrieval system, which provides the user with diverse representations of the queries. The system should be able to tackle complex and general-purpose multi-concept queries. Given a ranked list of photos retrieved from Flickr[1], participating systems are expected to refine the results by providing up to 50 images that are in the same time relevant and provide a diversified summary of the query. The process is based on the social metadata associated with the images and/or on the visual characteristics. A complete overview of the task is presented in [7].

Despite the current advances of machine intelligence techniques used in the area of information retrieval and multimedia, in search for achieving high performance and adapting to user needs, more and more research is turning now towards the concept of "*human in the loop*" [5]. The idea is to bring the human expertise in the processing chain, thus combining the accuracy of human judgements with the computational power of machines.

Due to good performance achieved in [3], this year we decided to follow the same work, which is an adapted version of the work in [2] that exploits the concept of RF. RF techniques attempt to introduce the user in the loop by harvesting feedback about the relevance of the search results. This information is used as ground truth for re-computing a better representation of the data needed. Relevance feedback proved efficient in improving the precision of the results [6], but its potential was not fully exploited to diversification. The main contribution of our approach is in proposing a pseudo-relevance feedback technique which substitutes the user needed in traditional RF and in proposing several diversity-adapted relevance feedback schemes.

[1] http://flickr.com/.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

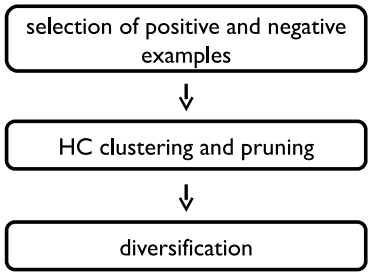

Figure 1: General scheme of the proposed approach

2 APPROACH

In traditional RF techniques, recording actual user feedback is inefficient in terms of time and human resources. The proposed approach, denoted in the following *HC-RF*, attempts to replace user input with machine generated ground truth. It exploits the concept of pseudo-relevance feedback. The concept is based on the assumption that top k ranked documents are relevant and the feedback is learned as in traditional RF under this assumption [1]. A general diagram of the approach is depicted in Figure 1.

Similarly to [3] we didn't opt for the use of the pre-processing step, i.e., the use of filters for the non-relevant images. The motivation is based on the specificity of the dataset proposed for this year [7], i.e. the use of multi-topic queries in the development and evaluation sets. An image containing people or depicting a location or a place which is geographically far away from the query, can be considered relevant as long as it is a common photo representation of the query topics (all at once). Also, we noticed in an extensive study [4] that the blur filter does not improve significantly the overall performance, thus we decided to removed it to reduce complexity. The algorithm is as follows.

First, we employ a pseudo-relevance feedback scheme based on an automatic selection of the images. We consider that the first returned results are relevant (i.e., positive examples). For instance, on *devset* [7], in average, 26 out of 50 returned images are relevant which supports our assumption. In contrast, the very last of the results are more likely non-relevant and considered accordingly (i.e., negative examples). The positive and negative examples are fed to an HC[2] scheme which yields a dendrogram of classes. For a certain cutting point (i.e., number of classes), a class is declared non-relevant if contains only negative examples or the number of negative examples is higher than the positive ones. The final

[2] http://www.mathworks.com/help/stats/hierarchical-clustering.html

Table 1: Best RF results for each modality or combination of modalities on devset (best results are depicted in bold).

metric/ run	HC-RF visual	HC-RF text	HC-RF vis-text	HC-RF CNN	HC-RF cred.	Flickr init. res.
P@20	0.575	0.575	**0.6136**	0.575	0.575	*0.5864*
CR@20	0.3969	0.3969	**0.4234**	0.3969	0.3969	*0.3646*
*F*1@20	0.4473	0.4473	**0.4773**	0.4473	0.4473	*0.4277*

step is the actual diversification scheme, which is a round robin approach. We select from each of the relevant classes one image which has the highest rank according to the initial ranking of the system. Then, we remove the selected images from the clusters and proceed by selecting the remaining ones in the same manner. The process is repeated until a maximum number of images is reached. The resulting images represent the output of the proposed system.

3 EXPERIMENTAL RESULTS

This section presents the experimental results achieved on *devset* which consists of 110 multi-topic queries and 32,487 images and *testset*, respectively, which consists of 84 multi-topic queries and 24,986 images. We optimized the parameters of the proposed approach on *devset* to obtain best precision and diversity. The final benchmarking is conducted however on *testset*.

In our approaches, images are represented with the content descriptors that were provided with the task data, i.e., visual (e.g., convolutional neural network based descriptors), text (e.g., term frequency - inverse document frequency representations of metadata) and user annotation credibility (e.g. upload frequency) information. Detailed information about provided content descriptors is available in [7]. Performance is assessed with Precision at X images (P@X), Cluster Recall at X (CR@X) and F1-measure at X (F1@X).

3.1 Results on devset

Several tests were performed with different descriptor combinations and various cutoff points. Descriptors are combined with an early fusion approach (normalization and concatenation). We varied the number of initial images considered as positive examples (Np) from 100 to 280 with a step of 20 images, the number of last images considered as negative examples (Nn) from 0 to 20 with a step of 10, and the inconsistency coefficient threshold for which HC divides the data into well-separated clusters (Nc) from 0.5 to 1.3 with a step of 0.2. We select the combinations yielding the highest $F1@20$, which is the official metric.

While experimenting, we observed that, by increasing the number of analyzed images, precision tends to decrease as the probability of obtaining non-relevant images increases; in the same time, diversity increases as having more images is more likely to get more diverse representations. For brevity reasons, in the following we focus on presenting only the results at a cutoff of 20 images which is the official cutoff point. These results are presented in Table 1. We present also the Flickr initial retrieval results to serve as baseline for the evaluation. From the modality point of view, visual-text information (visual-all textual-all) with the parameter setup (Np-Nn-Nc)=(180-0-1.1) lead to the highest results ($F1@20$=0.4773). This performance was followed by visual (visual-all without CNN), textual (textual-all), CNN and credibility descriptors ($F1@20$=0.4473) all with (180-20-1.1) parameter setup.

Table 2: Results for the official runs on testset (best results are depicted in bold).

metric/run	*Run1*	*Run2*	*Run3*	*Run4*	*Run5*
P@20	**0.6333**	0.6214	0.6196	0.5845	0.6018
CR@20	0.5791	0.5794	0.5729	0.5216	**0.6045**
*F*1@20	0.5753	0.5733	0.5741	0.5253	**0.5777**

3.2 Official results on testset

Following the previous experiments, the final runs were determined for the best modality/parameter combinations obtained on *devset* (see Table 1). We submitted five runs, computed as following: *Run1* - automated using visual information only: HC-RF all visual; *Run2* - automated using text information only: HC-RF all text; *Run3* - automated using visual-text information: HC-RF all visual-all text; *Run4* - everything allowed: HC-RF CNN.; and *Run5* - everything allowed: HC-RF cred. Results are presented in Table 2.

What is interesting to observe is the fact that the highest precision is achieved using visual information, (*Run1* - $P@20 = 0.6333$), whereas maximum diversification is achieved using credibility information (*Run5* - $CR@20 = 0.6045$), with more than 2% over other types of descriptors. Relevance was also preserved, which leads to the conclusion that credibility information was useful in the context of overall diversification. Credibility information estimates the quality of tag-image content relationships, telling which users are most likely to share relevant images in Flickr. Best diversification is achieved in this case due to a high probability that different and relevant images belong to different users with a good credibility score. In terms of $F1$ metric score, the use of credibility information, *Run5* - $F1@20 = 0.5777$, allows for best performance, followed closely by visual descriptors, *Run1* - $F1@20 = 0.5753$. Visual-textual information achieved also good performance, *Run3* - $F1@20 = 0.5741$, followed by textual information, *Run2* - $F1@20 = 0.5733$. The CNN descriptors had the lowest performance, by more than 5% under the credibility information, *Run4* - $F1@20 = 0.5253$.

4 CONCLUSIONS

We approached the image search result diversification issue from the perspective of relevance feedback techniques, when user feedback is substituted with an automatic pseudo-relevance feedback approach. Results show that in general, the automatic techniques improve the precision and diversification, which proves the real potential of relevance feedback to the diversification. Future developments will mainly address different efficient exploitations of re-ranking approaches, e.g., relevance-score estimation techniques, to improve the relevance and consequently the overall diversification. Another perspective is to also exploit the advantages of deep neural networks and use them in the context of automatic relevance-feedback-based diversification scenarios, by classifying the selected positive and negative examples using unsupervised deep-learning-based classifiers.

REFERENCES

[1] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. 2015. Hierarchical Clustering Pseudo-Relevance Feedback for Social Image Search Result Diversi□cation. *Content-Based Multimedia Indexing (CBMI), 2015 13th International Workshop on* (September 2015), 1–6.

[2] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. 2015. LAPI @ 2015 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversi□cation Perspective. *MediaEval 2015 Workshop* (September 2015).

[3] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. 2016. LAPI @ 2016 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversi□cation Perspective. *MediaEval 2016 Workshop* (October 2016).

[4] Bogdan Boteanu, Ionuţ Mironică, and Bogdan Ionescu. 2016. Pseudo-Relevance Feedback Diversication of Social Image Retrieval Results. *Multimedia Tools and Applications* 76, 9 (2016), 11889–11916.

[5] Bruno Emond. 2007. Multimedia and Human-in-the-loop: Interaction as Content Enrichment. *ACM International Workshop on Human-Centered Multimedia* (2007), 77–84.

[6] Jing Li and Nigel M Allinson. 2013. Relevance Feedback in Content-Based Image Retrieval: A Survey. *Handbook on Neural Information Processing* 49 (2013), 433–469.

[7] Maia Zaharieva, Bogdan Ionescu, Alexandru Lucian Gînscă, Rodrygo L.T. Santos, and Henning H. Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation. *MediaEval 2017 Workshop* (September 2017).