LAPI @ 2014 Retrieving Diverse Social Images Task: A Relevance Feedback Diversification Perspective

Bogdan Boteanu[1], Ionuţ Mironică[1]*, Anca-Livia Radu[1,2]†, Bogdan Ionescu[1,2]
[1]LAPI, University "Politehnica" of Bucharest, Romania
[2]DISI, University of Trento, Italy
{bboteanu,imironica,bionescu}@alpha.imag.pub.ro, ancalivia.radu@unitn.it

ABSTRACT

In this paper we approach the 2014 MediaEval Retrieving Diverse Social Images Task from the perspective of relevance feedback techniques. Two methods are introduced. A first approach exploits real user feedback with a multi Support Vector Machine classification scheme and a confidence score based image selection mechanism. The second approach replaces human feedback with an automatic hierarchical clustering pseudo-relevance feedback. The proposed relevance feedback approaches are designed to have in priority the diversification of the results, in contrast to most of the existing techniques that address only the relevance. 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 2014 Retrieving Diverse Social Images Task [1] addresses this issue in the context of a tourism real-world usage scenario. Given a ranked list of location 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 location. These results will help potential tourists in selecting their visiting locations. The refinement and diversification 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 [1].

Despite the current advances of machine intelligence techniques, 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*" [2]. 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.

In this work we propose a novel perspective that exploits the concept of relevance feedback (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 [4], but its potential was not fully exploited to diversification. The main contribution of our approach is in proposing several diversity-adapted relevance feedback schemes.

2. HUMAN RELEVANCE FEEDBACK

The first proposed relevance feedback approach (*SVM-RF*) is based on real user input. We implemented the method in [3]. It involves the following steps: (1) For each target image class obtained via user feedback (users select both relevant and diverse images from the results) we train an individual Support Vector Machine (SVM) classifier. We use an optimized version that determines the SVM's parameter C (tradeoff between margin maximization and error minimization) using a two-fold optimization on the user recorder feedback. Once trained, the SVMs are fed with all the images generating a confidence score for each of the output classes; (2) diversification is then achieved by analyzing the resulting confidence score matrix (of size number of images x number of classes): for each image class, the images are analyzed by decreasing the confidence scores. Each highest confidence score image, different from the others already selected, is added to the output. The process is repeated by visiting the classes in a circular way to ensure the highest diversity among the selected images.

3. PSEUDO RELEVANCE FEEDBACK

Recording actual user relevance feedback is inefficient in terms of time and human resources. The second approach (*HC-RF*) attempts to replace user input with machine generated ground truth. It exploits the concept of pseudo-relevance feedback. We consider that most of the first returned results are relevant (i.e., positive examples). For instance, on *devset* [1], in average, 40 out of 50 returned images are relevant which support our assumption. In contrast, the very last of the results are more likely un-relevant and considered accordingly (i.e., negative examples). The positive and negative examples are feed to an Hierarchical Clustering scheme which yields a dendrogram of classes. For a certain cutting point (i.e., number of classes), a class is declared un-relevant if contains only negative examples or the number of negative examples is higher than the positive ones. The resulting images are generated using images from each of the relevant classes in their initial order.

4. EXPERIMENTAL RESULTS

This section presents the experimental results achieved on *devset* (30 locations, 8,923 images) and *testset* (123 locations, 36,452 photos), respectively. For *devset*, ground truth was provided with the data for preliminary validation of the approaches. The final benchmarking is conducted however on *testset*.

*The work was funded by the ESF POSDRU/159/1.5/S/132395 InnoRESEARCH programme.

†The work was funded by the ESF POSDRU/159/1.5/S/134398 KNOWLEDGE programme.

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

Copyright is held by the author/owner(s).

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: Best method - modality combination relevance feedback results on devset (best results are depicted in bold).

metric /method	SVM-RF expert text TF	SVM-RF user text TF	HC-RF text TF	HC-RF visual CM	HC-RF cred.	HC-RF text-cred.	HC-RF visual-text-cred.	HC-RF visual-text	HC-RF visual-cred.	*Flick initial res.*
P@20	0.8817	**0.91**	0.8117	**0.83**	0.7367	0.735	0.7033	0.805	0.6967	*0.8333*
CR@20	**0.5363**	0.3965	0.4423	0.4135	0.4347	0.4236	0.4081	**0.4434**	0.4104	*0.3455*
$F1$@20	**0.6607**	0.546	**0.568**	0.5454	0.5419	0.53	0.5137	0.5649	0.5118	*0.4768*

In our approaches, images are represented with the content descriptors that were provided with the task data, i.e., visual (e.g., color, feature descriptors), text (e.g., term frequency - inverse document frequency representations of metadata) and user annotation credibility (e.g., face proportions, upload frequency) information [1]. Performance is assessed with Precision at X images (P@X), Cluster Recall at X (CR@X) and F1-measure at X (F1@X).

4.1 Results on devset

Several tests were performed with different descriptor combinations and various cutoff points. Descriptors are combined with an early fusion approach. SVM-RF was run with a number of $N_{class} = 20$ image classes (which is the predicted average number of diversity classes from *devset* ground truth) and using a linear kernel (which provided the best results). User feedback was recorded from two users, one *expert* familiarized with the data and a common *user*. For HC-RF, we varied the number of initial images considered as positive examples, N_{start}, from 100 to 150 with a step of 10 images, the number of last images considered as negative examples, N_{end}, from 0 to 20 with a step of 5, and the number of image diversity classes, N_{class}, from 20 to 30 with a step of 1. We select the N_{start}-N_{end}-N_{class} combinations yielding the highest $F1$@20, which is the official metric.

By increasing the number of analyzed images, precision tends to slightly decrease as the probability of obtaining un-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. Apart for the use of the Color Moments (CM) and term-frequency (TF) descriptors, all the other modalities reflect the combination of all the task provided descriptors. SVM-RF results are presented only for the best performing descriptors (text TF). To serve as baseline for the evaluation, we present also the Flickr initial retrieval results.

If an expert user is used, human-based relevance feedback provides a significantly higher performance than other approaches, SVM-RF text TF — $F1$@20 $= 0.6607$, which is an improvement of more than 9 percentage points compared to the best pseudo-relevance feedback, HC-RF text TF — $F1$@20 $= 0.568$, and of 18 percentage points compared to Flickr's baseline, $F1$@20 $= 0.4768$. In contrast, a common user feedback allows to achieve lower/similar results compared to the pseudo-relevance feedback. However, in average, human input provides better results than the automated version (average $F1$@20 is 0.6034). From the modality point of view, text descriptors lead to the highest results for both approaches, followed closely by the combination of visual and text descriptors and then visual Color Moments and credibility information.

4.2 Official results on testset

Following the previous experiments, the final runs were determined for best modality/parameter combinations obtained on *devset* (see Table 1). We submitted five official runs, computed as following: *Run1* - automated using visual information only: HC-RF visual CM, *Run2* - automated using text information only: HC-RF text TF, *Run3* - automated using visual-text information: HC-RF visual-text, *Run4* - automated using credibility information only: HC-RF cred., and *Run5* - everything allowed: SVM-RF text TF (to simulate a real scenario, in this case the feedback was recorded from a common user). Results are presented in Table 2.

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

metric/run	*Run1*	*Run2*	*Run3*	*Run4*	*Run5*
P@20	0.7687	0.7882	0.7675	0.674	**0.876**
CR@20	0.3994	**0.4431**	0.4335	0.4149	0.3859
$F1$@20	0.5187	**0.5583**	0.5472	0.5071	0.5261

What is interesting to observe is the fact that the highest precision is achieved with a human-based approach, *Run5*, P@20 $=$ 0.876, whereas the automatic methods allow for the best diversification, *Run2*, CR@20 $= 0.4431$. In terms of modality, the use of text information allows for the best performance, *Run2*, $F1$@20 $=$ 0.5583. These results are consistent with the results on *devset*.

5. CONCLUSIONS

We approached the image search result diversification issue from the perspective of relevance feedback techniques. Two scenarios were considered: (1) user feedback is recorded from actual users, (2) user feedback is substituted with an automatic pseudo-feedback approach. Results show that in general, real user feedback allows for achieving better precision while the automatic techniques improve the diversification. Overall, the best results in terms of both precision and diversity are achieved with the automatic pseudo-relevance feedback approach which proves the real potential of relevance feedback to the diversification. Future developments will mainly address a more efficient exploitation of different modalities (visual-text-credibility), e.g., via late fusion techniques.

6. REFERENCES

[1] B. Ionescu, A. Popescu, M. Lupu, A.L. Gînscă, H. Müller, "*Retrieving Diverse Social Images at MediaEval 2014: Challenge, Dataset and Evaluation*", MediaEval 2014 Workshop, October 16-17, Barcelona, Spain, 2014.

[2] B. Emond, "*Multimedia and Human-in-the-loop: Interaction as Content Enrichment*", ACM Int. Workshop on Human-Centered Multimedia, pp 77-84, 2007.

[3] B. Boteanu, I. Mironică, B. Ionescu, "*A Relevance Feedback Perspective to Image Search Result Diversification*", IEEE ICCP, September 4-6, Cluj-Napoca, Romania, 2014.

[4] J. Li, N.M. Allinson, "*Relevance Feedback in Content-Based Image Retrieval: A Survey*", Handbook on Neural Information Processing, 49, pp 433-469, Springer 2013.