Retrieving Social Images using Relevance Filtering and Diverse Selection

Taruna Agrawal[1], Rahul Gupta[2], Shrikanth Narayanan[2]
[1]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, USA
[2]Signal Analysis and Interpretation Lab (SAIL), University of Southern California, Los Angeles, USA
tagrawal@usc.edu, guptarah@usc.edu, shri@sipi.usc.edu

ABSTRACT

Retrieving relevant and diverse images from a large set of images is problem of interest in social media. Given a set of images pertaining to a location or a concept, a subset of diverse image can summarize the attributes of the corresponding location/concept. In this work, we present a two step image retrieval model involving relevance filtering followed by diverse selection. Based on the visual features, textual descriptions and Flickr rank, relevance filtering initially determines a subset of images which have correspondence to a topic of interest. Subsequently, diverse selection determines a smaller subset of images to provide a diverse perspective of the concept. We obtain an F1 score of .509 on a test set containing 139 concepts, when computed over the top 20 images output by our system. We analyze the outcomes of our system and investigate the utility of image metadata (reviews, Flickr content) when combined with visual descriptors.

1. INTRODUCTION

"Deluge of information" is a term prevalent in present day social media [1–4], often attributed to advances in technology and social connectivity. Compact representation of relevant information is a major challenge posed by the growth of social media. Retrieving diverse social images task at MediaEval challenge 2015 [5] addresses this problem in the domain of images on social media such as Flickr. The goal is to design a query based social image retrieval engine, focusing on obtaining relevant images while covering diverse aspects of the query, for instance, various sub-topics of the query. Potential information sources include image attributes as well as image metadata such as image description, view count and image rank on social media.

Various previous works [6, 7] have focused on knowledge based image selection for relevant image selection and/or clustering based methods for diversification. The relevance selection is usually based on image attributes such as presence of people [8], image quality [7] and similarity to a standard source of images like Wikipedia [9]. In this work, we adopt a combination of supervised and unsupervised schemes for relevance filtering followed by clustering for diverse selection. After filtering out irrelevant images, we use clustering for diverse selection of images. Through our methods, we show the promise of using supervised learning methods in addition to existing knowledge based methods in such retrieval tasks. In the next section, we describe our methodology in detail followed by the results.

2. METHODOLOGY DESCRIPTION

Our system for retrieving diverse social images consists of two steps: (i) Relevance filtering, and (ii) Diverse selection. Relevance filtering helps us to filter out images that have no or little relation with the concept of interest and diverse selection provides a subset of images which are different from each other. We provide a detailed description of the two systems below.

2.1 Relevance filtering

We perform relevance filtering to filter out images unrelated to a concept. The 2015 MediaEval challenge data provides a set of visual and textual descriptors over 153 concepts for model development and 139 concepts for evaluation. Given the visual descriptors, textual information and Flickr metadata, we train several supervised and knowledge based filtering schemes. We describe these models below.

2.1.1 Supervised methods

K-nearest neighbor classifier on visual descriptors: The 2015 MediaEval challenge data set provides a set of general purpose visual descriptors such as color, texture and feature information along with a binary label indicating if an image is relevant/irrelevant to the concept under consideration [5]. We train a K-nearest neighbor (KNN) classifier on these visual descriptors using these labels. The features are z-normalized before training and K is tuned on the development set using a 3-fold cross-validation.

Maximum entropy model on textual descriptors: The textual descriptors are extracted from sources such as photo title, description as provided by the author and photo tags on Flickr. We extract features from these sources using the following steps:

1. *Feature standardization*: This step is performed to train a universal model for all the concepts instead of concept specific models. We replace any word related to a concept by a keyword. For instance, if the query is "The great wall of china", words such as "great wall", "wall of china" and "great wall china" occurring anywhere in textual descriptions are replaced by a single keyword "Place_of_interest". The list of words to be replaced is created based on the query title and contains various combinations of words in the query.

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

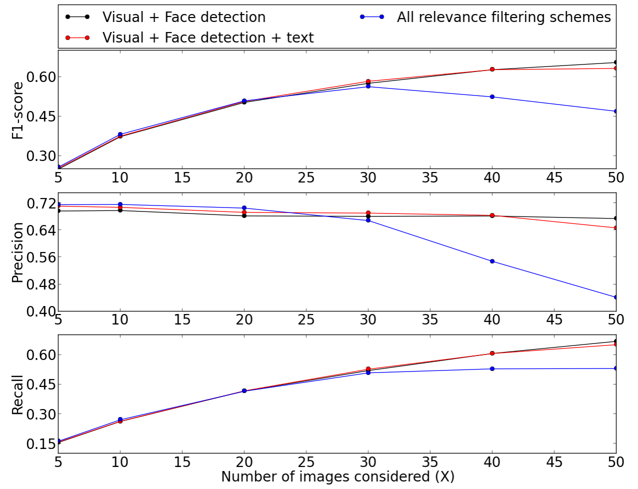

Figure 1: Results for the proposed system at different number of retrieved images (X).

2. Feature selection: Given the set of standardized features, we retain the words within the top 10% of word frequencies. This step is performed to reduce the feature dimensionality while training the model.
3. Model training: Given the set of selected features, we train a maximum entropy model to predict the binary labels (relevant/irrelevant).

2.1.2 *Unsupervised methods*

Removal of images with people in focus: Relevant images do not have a person as the subject of focus. We incorporated this fact by using the facedetect software [10] to filter out images containing people as the main subjects.

Relevance filtering based on Flickr rank: As a final relevance filtering scheme, we remove images above a certain threshold (>200) on Flickr rank. The motivation behind this scheme is that images low in rank are more likely to be not associated with the concept in question.

2.2 Diverse selection

After obtaining the set of images based on relevance filtering, we use image clustering for diverse selection. Given a query size of $\hat{K}$ images, we perform $\hat{K}$-means clustering on the visual descriptors. We hypothesize that similar images fall into a single cluster and retain only image per cluster. We select the image closest to the cluster centroid as the cluster representative.

In order to compute the selection score for each image, we use the output of the KNN classifier, maxent model and distance of image from cluster centroid. The score is given by an unweighted sum of the ratio of relevant images amongst closest K images, the maxent output probability for image being relevant and inverse of Euclidean distance of image from cluster centroid. The last term is added based on the assumption that images closer to centroids are more representative of the cluster. In the next section we present our results and discussion.

3. RESULTS

In run 1, we only use the relevance filtering model developed on visual descriptors (K-nearest neighbors classifier) and face detection. In run 2, we append filtering using maximum entropy model on textual descriptors. Finally, run5 uses all the relevance filtering schemes (visual, face detection, text and Flickr rank based). Note that in all the three runs diverse selection is based on visual descriptors only. The evaluation metric is cluster recall (CR) and precision (P) for top X ranked images as predicted by the system. We show the CR@X and P@X along with corresponding F-score F1@X for $X = 5, 10, 20, 30, 40, 50$ in Figure 1. All these outcomes are based on cluster with $\hat{K}$ set to 50. Also, in the 2015 challenge, separate metrics were reported for concepts which share images with other concepts (multi-concept) along with single-concept images. We report the official score of CR,P and F1 @X=20 for the multi and single concept images in Table 1.

Run #	Relevance filter	Single concept F1/P/CR	Multi concept F1/P/CR	All F1/P/CR
1	Visual desc.	.492/ .664/.408	.514/ .700/.426	.504/ .682/.417
3	+ Textual desc.	.497/ .677/.410	.517/ .708/.426	.507/ .692/.418
5	+ Flickr rank	.512/ .702/.421	.507/ .708/.411	.509/ .705/.416

Table 1: Results (F1 score/Precision/Cluster recall) for the proposed system @$X = 20$.

From the results, we observe that for low values of X the combined system (visual + face detection +text + Flickr rank) marginally (although insignificantly) outperforms the system using only the visual cues. However the performance degrades significantly at higher value of X. Note that this decrease in performance is not due to additional filtering schemes not performing well. Instead, this decrease in performance is due to the fact that additional filtering leads to decrease in data points available for diverse selection. Therefore we had to reduce the number of clusters in relevance selection, sometimes to the extent that our model returned less than 50 images. However, better performance at lower X (e.g. $X = 20$ in Table 1) shows the promise of using additional modalities. In Table 1, we observe minor improvements in F1@20 after adding subsequent relevance filtering schemes. One interesting observation is that while using Flickr ranks, F1 for multiple concepts decreases, whereas for single concepts increases. This indicates that Flickr ranks are more reliable in the case of single concept images than multiple-concept images. This factor can be regarded in future system designs.

4. CONCLUSION

In this work, we present a two stage system for social image retrieval. In the first stage, we perform relevance filtering to remove irrelevant images and in the second stage we perform diverse selection using clustering in the visual descriptor space. Our relevance filtering system involves a combination of supervised and unsupervised methods. In the future, we can extend the work presented by exploring other methods (filtering, clustering) under a similar system development paradigm. We can also reformulate the problem as a diverse system development and can be inspired from several of the existing works [11, 12]. Finally, we would also like additional metadata like Flickr user credibility [5, 13] and other image properties (CNN features) to further improve our system.

5. REFERENCES

[1] Holly M Bik and Miriam C Goldstein. An introduction to social media for scientists. 2013.

[2] Sophia B Liu. Trends in distributed curatorial technology to manage data deluge in a networked world. *The European Journal for the Informatics Professional*, 11(4):18–24, 2010.

[3] C Szongott, Benjamin Henne, G von Voigt, et al. Big data privacy issues in public social media. In *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, pages 1–6. IEEE, 2012.

[4] Duc-Tien Dang-Nguyen, Luca Piras, Giorgio Giacinto, Giulia Boato, and F De Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. *MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2014.

[5] Bogdan Ionescu, Alexandru L Gınsca, Bogdan Boteanu, Adrian Popescu, Mihai Lupu, and Henning Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[6] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru L Gınsca, and Henning Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[7] Tomas Brodsky. Relevant image detection in a camera, recorder, or video streaming device, April 4 2006. US Patent App. 11/397,780.

[8] Alexandru Lucian Ginsca, Adrian Popescu, and Navid Rekabsaz. Cea listâĂŹs participation at the mediaeval 2014 retrieving diverse social images task. In *Proceedings of the MediaEval Multimedia Benchmark Workshop, CEURWS. org*, volume 1263, pages 1613–0073, 2014.

[9] Maia Zaharieva and Patrick Schwab. A unified framework for retrieving diverse social images. 2014.

[10] Robert Frischholz. The face detection homepage. https://facedetection.com/.

[11] Rahul Gupta, Kartik Audhkhasi, and Shrikanth Narayanan. A mixture of experts approach towards intelligibility classification of pathological speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1986–1990. IEEE, 2015.

[12] Rahul Gupta, Kartik Audhkhasi, and Shrikanth Narayanan. Training ensemble of diverse classifiers on feature subsets. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2927–2931. IEEE, 2014.

[13] Bogdan Ionescu, Adrian Popescu, Mihai Lupu, Alexandru Lucian Gînsca, Bogdan Boteanu, and Henning Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. *ACM Multimedia Systems-MMSys, Portland, Oregon, USA*, 2015.