Toward an Estimation of User Tagging Credibility for Social Image Retrieval

Alexandru Lucian Ginsca[1,4], Adrian Popescu[1], Bogdan Ionescu[2] , Anil Armagan[3], Ioannis Kanellos[4]
[1]CEA, LIST, Vision and Content Engineering Laboratory, 91191 Gif-sur-Yvette, France
[2]LAPI, University "Politehnica" of Bucharest, 061071, Romania
[3]Bilkent University, Ankara, Turkey
[4]TELECOM Bretagne, France
{alexandru.ginsca, adrian.popescu}@cea.fr, bionescu@alpha.imag.pub.ro,
anil.armagan@bilkent.edu.tr, ioannis.kanellos@telecom-bretagne.eu

ABSTRACT

Existing image retrieval systems exploit textual or/and visual information to return results. Retrieval is mostly focused on data themselves and disregards the data sources. In Web 2.0 platforms, the quality of annotations provided by different users can vary strongly. To account for this variability, we complement existing methods by introducing user tagging credibility in the retrieval process. Tagging credibility is automatically estimated by leveraging a large set of visual concept classifiers learned with Overfeat, a convolutional neural network (CNN) feature. A good image retrieval system should return results that are both relevant and diversified and here we tackle both challenges. Classically, we diversify results by using a k-Means algorithm and increase relevance by favoring images uploaded by users with good credibility estimates. Evaluation is performed on DIV400, a publicly available social image retrieval dataset and shows that our method is competitive with existing approaches.

1. INTRODUCTION

Existing works have identified relevance and diversity as two core properties of efficient image retrieval systems. Given that these two characteristics are antinomic, different methods have been proposed to find a good compromise between them. Classically, relevance was primarily estimated by using textual weighting schemes. However, with the improvement of low-level image descriptors, multimedia fusion schemes also gained traction. Diversity is usually improved by applying clustering algorithms which rely on textual or/and visual cues [14]. In addition, the usefulness of social cues was also explored for Web 2.0 platforms [9] but this aspect remains secondary.

Our work is focused on the estimation and exploitation of user credibility, a cue which was not previously exploited in multimedia retrieval and is complementary to those cited above. We investigate user tagging credibility in the context of social multimedia mining and address the following research questions: Q_1 - is it possible to automatically estimate user tagging credibility for Web 2.0 multimedia data? Q_2 - how should credibility be integrated in existing multimedia retrieval systems? Q_3 - what is the additional complexity of credibility estimation?

We propose an image retrieval technique that ensures a good balance of results diversity and relevance. Evaluation is performed on the DIV400 dataset, a retrieval dataset created in the context of the MediaEval 2013 Diverse Social Images Task [6]. Retrieval results are diversified using a k-Means algorithm with user-based cluster ranking. The novelty comes from relevance improvement obtained by integrating user tagging credibility. In an initial step, credibility scores are computed by probing user tag-image pairs against a large array of visual concept models learned from ImageNet [3] and by aggregating classification scores at user level. At retrieval time, images are reranked based on the credibility scores of the users who uploaded them. In addition, face and blur detection are applied to discard potentially irrelevant images. Our technique is compared to state of the art systems that were submitted to Diverse Social Images Task [6] and interesting performances are obtained for both diversification and relevance.

2. RELATED WORK

Web credibility was studied under three main aspects: informing about, analyzing and estimating credibility. The first two directions fall outside the immediate scope of our work. Automatic credibility estimation is a recent trend in Web content analysis; it is mostly applied to textual documents, such as tweets [1] or Web pages [12]. Also related is the automatic assessment of crowdsourcer credibility, which is investigated in [7]. However, none of these works is focused on multimedia content and literature regarding multimedia credibility is limited. Xu et al. [15] aim to help users filter multimedia news by targeting credible content. They propose methods to evaluate multimedia news by comparing visual descriptions and textual descriptions respectively, as well as their combination. Yamamoto and Tanaka [16] have built ImageAlert, a system that focuses on text-image credibility. While interesting, existing work on multimedia content credibility estimation is preliminary and deserves further investigation. The estimation of individual tag rel-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
MM'14, November 3–7, 2014, Orlando, Florida, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2655033.

evance is related to our work. Li et al. [10] have proposed a neighbor voting framework which exploits neighbor voting to assess tag quality. More recently, Gao et al. [5] introduce a hypergraph framework to jointly model visual and textual cues of social media images. Their approach compares favorably to other existing methods but has a high computational cost at query time. We estimate credibility independently of a given topic and thus drastically reduce processing complexity at query time. [10, 5] do not aggregate relevance at user level and focus on individual tags. Both works need a large amount of data annotated with targeted tags and their efficiency on less common tags is questionable.

Due to its large size and availability, Flickr is an interesting playground for different multimedia mining tasks. Techniques that rely on social media data in order to improve image retrieval are also related to our objectives. Van Leuken et al. [14] focus on the usage of visual features to diversify image search results. Their contribution is twofold: proposition of lightweight clustering techniques and dynamic weighting of visual features. However, they do not study the use of social cues. Kennedy and Naaman [9] exploit temporal and spatial metadata, as well as user counts, for location clustering. This approach is interesting but social cues are processed at an image level and there is no aggregation at the user level, such as the one proposed here.

The usefulness of user tagging credibility estimation for image retrieval was not evaluated in previous works. To perform evaluation, we need a dataset with identified contributors. The DIV400 dataset [6] fits our needs. It sets up a domain retrieval scenario, i.e. tourism, and the primary focus of the evaluation exercise is to improve retrieval diversity. The dataset includes images of 396 tourist points of interest (POIs) and is further described in Section 6. The most efficient approaches tried on DIV400 are described in [8] and [2]. Jain et al. [8] use re-ranking with proximity search to improve precision and a Greedy Min-Max diversifier based on temporal user information. Corney et al. [2] report their best results with a Greedy optimization of a VLAD representation of SURFs.

3. VISUAL CONTENT PROCESSING

Our diversification approach is mainly based on visual content mining. To keep abreast with the latest advances in the field, we use Overfeat [13], a powerful CNN-based feature, to model credibility and to process the DIV400 dataset. In addition, we remove faces and blurred images, which are potentially irrelevant for a part of the topics.

ImageNet Concept Learning

ImageNet [3] is a manually labeled dataset which includes over 14 million images of nearly $22,000$ concepts. This dataset is well suited here insofar as we want to model a diversified range of tag-image pairs. The basic brick in modeling user tagging credibility is a verification of tag-image content relation. For instance, if an image is tagged with *dog*, it is then compared to *dog* models from ImageNet. We select the $17,462$ ImageNet concepts which are represented by at least 100 images and build binary classifiers to model them. Image content is represented using the default configuration of Overfeat [13], followed by a L2-normalization of the features.

Let $\{I_i, y_i\}_{i=1..N}$ be the training set associated with the c-th concept of ImageNet. I_i is a training image and $y_i \in \{-1, 1\}$ its corresponding binary label. To learn models quickly, for a concept with N positive examples, we select the first N negative examples from a unique negative set, built from concepts which are not modeled. We capture the visual appearance of the c-th concept, using a linear model $\mathbf{W}^c \in \mathbb{R}^{S_f+1}$. S_f is the image feature size. The last dimension, $(\mathbf{W}^c)_{S_f+1}$, corresponds to the model bias. $\mathbf{W}^c$ is learned by minimizing the L2-regularized logistic loss:

$$\mathbf{W}^c = \arg\min_{\bar{\mathbf{W}}^c} \frac{1}{2}\|\bar{\mathbf{W}}^{\mathbf{c}}\|_2^2 + C\sum_{i=1}^{N} \log(1 + e^{-y_i \bar{\mathbf{W}}^{cT}\mathbf{f}_i}). \quad (1)$$

Here, $\mathbf{f_i} = f(I_i) \in \mathbf{R}^{S_f+1}$ is an image feature capturing I_i content information. It should be noted that $(\mathbf{f}_i)_{S_f+1} = 1\,\forall i$ since it correlates with the model bias. $C \geq 0$ is a penalty parameter. A Quasi-Newton descent [11] is used to solve (1)[1]. Classification scores are normalized between 0 and 1 using a logistic function.

Dataset Processing

Overfeat features are also extracted for DIV400 images. PCA is applied to these features to obtain a more compact representation of images and thus accelerate retrieval. Preliminary tests have already shown that results obtained with the first 256 PCA dimensions are equivalent to those obtained with the default Overfeat configuration (4096 dimensions). Inspired by [8], face and blur detection is applied to the dataset. Face detection is implemented with the standard OpenCV algorithm[2]. Another set of tests has shown that direct removal of images containing faces does not improve results. Consequently, given a set of POI images and the associated user set t_u, face removal is performed based on p_u, the proportion of users from the set t_u which upload face images. Face images are retained for p_u values lower than a threshold ($th(p_u)$) and discarded otherwise. In order for p_u to be meaningful, we impose face removal only on the POIs with at least $th(t_u)$ contributors. p_u exploits social consensus about usefulness of face images and is optimized on the devset of DIV400. Blur detection is performed using thresholded gradient. Similar to face retrieval, a threshold $th(b)$ for blur removal is learned on the devset of DIV400.

4. USER CREDIBILITY ESTIMATION

User tagging credibility is estimated by exploiting ImageNet visual models (Section 3). For each user, we download at most 300 images whose textual annotations match at least one ImageNet concept. Flickr annotations are selected either from tags or from the image title and are all referred as tags hereafter. We perform multiword detection in order to match multiwords from ImageNet. Tags are tested against corresponding ImageNet concepts to obtain individual relevance scores. User tagging credibility estimation ($cred(U)$) is obtained by averaging scores from individual tag-image pairs.

Visual models are built on top of ImageNet concepts, which are often ambiguous, and tested for Flickr annotations. For instance, if an unknown image annotated with *dog* is tested, which of the three senses of *dog* from Figure 1 should be used? An inspection of Flickr results shows that most images annotated with *dog* depict *animals* but there are some of them which depict *dog* as *food* and *dog*

[1] We rely on the liblinear implementation from [4].

[2] http://opencv.org/

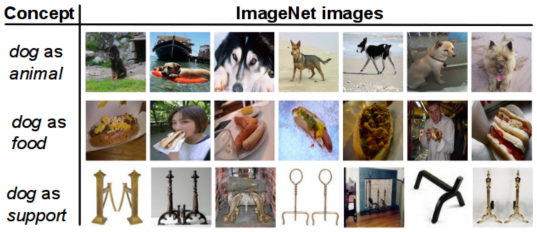

Figure 1: Different senses of *dog* in ImageNet.

as *support*. Our credibility estimator should be able to automatically select the right sense of *dog* for the content of the tested image. A simple way to process ambiguity is to compare the tag-image pair to all models available for the tag and retain only the maximum classification score. Preliminary tests showed that this procedure has good behavior and it is thus used in the experiments. Beyond ambiguity, another problem is the coverage of ImageNet, with some important senses of words not being included. For instance, *berlin* is represented as *car* but not as *city*. These problems represent limitations of our method and their tackling would probably improve credibility estimations.

5. IMAGE RETRIEVAL METHOD

We propose a retrieval method which diversifies images using k-Means and improves relevance with credibility estimations. Let $\mathbf{L_F} = \{(I_1, U_1), (I_2, U_2), (I_3, U_1), ..., (I_N, U_M)\}$ be the ranked list of Flickr images which should be reranked. Here (I_i, U_j) denote image-user pairs. Our retrieval method can be broken down into three steps: initial filtering, cluster ranking and image ranking.

Image Filtering In this step, we remove from $\mathbf{L_F}$ all pairs (I_i, U_j) for which I_i qualifies for face or blur removal.

Cluster Ranking After image filtering, we perform k-Means clustering to diversify the topic representation. Let $\mathbf{C_F} = \{C_1, C_2, ..., C_k\}$ be the clustered version of $\mathbf{L_F}$. Inspired by [9], we rank clusters based on *#Users*, the number of distinct users which contribute to each cluster. Ranking based on *#Users* gives priority to clusters which show social consensus. When ties appear with *#Users*, they are broken by using the top Flickr rank among the images of the user with the highest credibility score $cred(U)$ from each cluster. As a result, we obtain $\mathbf{C_F^R} = \{C_3, C_k, C_2, ..., C_1\}$, a list of clusters ranked using social cues. For comparison, we also rank clusters based on their raw image count (*#Images*).

Image Sorting We exploit credibility estimation to sort images within clusters. Let $C_c = \{(I_1, U_1), (I_3, U_5), (I_8, U_1)\}$ be a cluster with its images ranked by Flickr. Assuming that $cred(U_5) > cred(U_1)$, the sorted representation of the cluster will be $C_c^R = \{(I_3, U_5), (I_1, U_1), (I_8, U_1)\}$. In C_c^R, priority is given to images uploaded by users with higher credibility score.

The final image ranking $\mathbf{L_F^R}$ is obtained by iterating over $\mathbf{C_F^R}$, the ranked list of clusters, and by selecting each time the first unseen image from C_c^R, the sorted images of C_c.

6. EVALUATION

Dataset Description We evaluate our retrieval method with the DIV400 dataset, which is thoroughly described in [6]. It consists of a development dataset (50 tourist POIs, 5,118 photos) and a testing dataset (346 POIs, 38,300 photos). Each POI is represented with up to 150 photos and associated metadata retrieved with Flickr's default "relevance" algorithm. Relevance and diversity annotations are available for each photo. Photos are considered relevant if they depict a common photo representation of the POI. A set of photos is considered to be diverse if it depicts complementary visual characteristics of the target POI. Clusters are manually built from relevant images of each POI. The main objective of the evaluation from [6] is diversity, which is captured with cluster recall at N (CR@N). However, since a good retrieval method should find a good compromise between relevance and diversity, we also report precision (P@N) and their combination, F1@N.

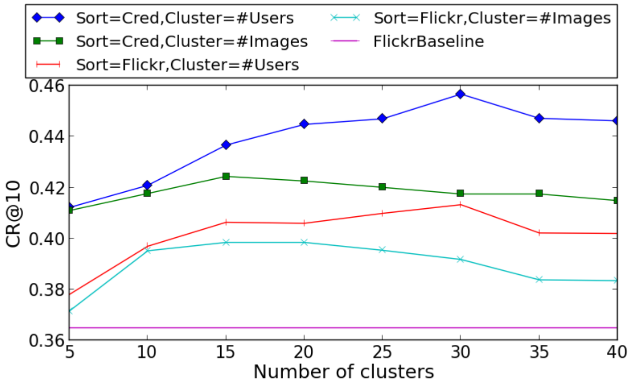

Figure 2: CR@10 performances with different clustering methods and different numbers of clusters on the testset of DIV400. *Sort* denotes the type of image sorting used within clusters. *Cred* is a sorting based on user credibility and *Flickr* is the original Flickr ordering. "Cluster" denotes the cluster ranking method. *#Users* and *#Images* represent the user and image counts of a cluster.

Clustering Analysis In Figure 2, we illustrate the impact of the number of clusters on clustering performances. Within each cluster, *Cred*, the credibility based image sorting outperforms the use of the initial Flickr sorting in all settings. Intuitively, the best overall results are obtained when *#Users* and *Cred* are combined for inter- and intra-cluster ranking. With 30 clusters, *Flickr* + *#Users* brings a 2 CR@10 points improvement of results compared to *Flickr* + *Images*. This result confirms the conclusions of [9], namely that the use of social cues for cluster ranking is beneficial. More importantly, the introduction of credibility estimation (*Cred* + *#Users*) further improves CR@10 by 4 points. We present results on the testset here because they are obtained by averaging a larger number of topics. However, similar results are obtained on the devset and *Cred* + *#Users* with 30 clusters is used for further experiments.

Global performances In table 1, we present the results obtained with the best credibility based retrieval method, described in Section 5. It combines clustering and user credibility estimates and produces a reranked list of images $\mathbf{L_F^R}$. For comparison, we also present results obtained by the two most efficient existing methods tested on DIV400 [6].

Method	metrics	@10	@20	@30
SOTON-WAIS [8]	P	**0.8158**	**0.7788**	0.7414
	CR	0.4398	0.6197	0.7216
	F1	0.5455	**0.6607**	0.7019
SocSens [2]	P	0.733	0.7487	**0.7603**
	CR	0.4291	0.6314	0.7228
	F1	0.5209	0.6595	**0.7087**
$\mathbf{L_F^R}$	P	0.7822	0.7154	0.6927
	CR	**0.4567**	**0.6582**	**0.7801**
	F1	**0.5526**	0.659	0.7073

Table 1: Comparison of retrieval results obtained with different methods on DIV400 and CR@N, P@N and F1@N metrics. SOTON-WAIS [8] and SocSens [2] are the two most efficient retrieval methods proposed at MediaEval Diverse Images 2013. $\mathbf{L_F^R}$ corresponds to a setting with *Cred+#Users* and 30 clusters (figure 2).

To understand the impact of face and blur removal, we briefly present results obtained when we skip one of these steps. When no prefiltering is used CR@10 is 0.4437. The use of blur removal or of face removal augments the score to 0.4476 and to 0.4536 respectively. While image filtering is beneficial, the main contribution comes from the use of credibility and of user centered clustering.

A comparison of our method to [8] and [2] shows that cluster recall is improved at all cut-off points. For CR@10, the official metric associated to DIV400, the improvement is close to 2 and 3 points. Confirming other results obtained on DIV400, which show that clustering hurts precision, the P@10 obtained with $\mathbf{L_F^R}$ is lower than those obtained in [8]. However, the F1@10 score of our method is slightly better. This comparison shows that our approach is competitive. It also departs from existing retrieval methods by the central role given to social cues and particularly to credibility.

7. CONCLUSIONS

This paper proposes an exploration of the introduction of user tagging credibility estimation in image retrieval systems. Evaluation results show that credibility is a good complement to direct text and/or visual content analysis.

Preliminary answers are provided to the research questions listed in the introduction. A credibility model based on text-image pairs assessment was proposed in response to Q_1. The main limitations of our method come from: the mismatch between the background visual resource and the datasets used for retrieval, the limited amount of modeled tags available for some users, the imperfection of visual models and the exclusive use of tag-image content relations. Credibility scores were obtained by comparing user tags to ImageNet concepts with no adaptation whatsoever to the evaluation dataset, which is made of tourist POIs. The fact that credibility estimations are effective even in this difficult setting accounts for their usefulness. In the future, we will extend the background visual resource in order to narrow the gap between it and retrieval datasets. New concepts can be learned from noisy Web datasets and their availability would contribute to the reduction of the number of users for which reliable credibility scores cannot be obtained. Visual models creation was focused on scalability but, since these models are learned offline, they will be improved by using larger and adapted negative sets. Following work in the textual domain [1], credibility estimation can be cast as a machine learning problem, with the addition of other cues than tag-image pairs.

In response to Q_2, credibility estimations were integrated with a classical clustering algorithm. The performance gains obtained through the use of credibility account for its usefulness in retrieval. The use of credibility in more sophisticated retrieval schemes ([8], [2]) will be investigated. Finally, additional complexity is added to the retrieval framework (Q_3) but affects retrieval steps which are performed offline. These steps, including feature extraction, visual model learning and credibility estimations, can be repeated periodically to follow the dataset evolution. At query time, only a reranking of images which accounts for credibility is required and this procedure has negligible effects compared to clustering.

8. ACKNOWLEDGMENT

This research was supported by the MUCKE project funded within the FP7 CHIST-ERA scheme.

9. REFERENCES

[1] C. Castillo and al. Information credibility on twitter. In *Proc. of WWW 2011*, pages 675–684.

[2] D. Corney and al. Socialsensor: Finding diverse images at mediaeval 2013. In *Proc. of MediaEval Wksp. 2013.*

[3] J. Deng and al. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. of CVPR 2009.*

[4] R.-E. Fan and al. Liblinear: A library for large linear classification. *JMLR*, 2008.

[5] Y. Gao and al. Visual-textual joint relevance learning for tag-based social image search. *IEEE TIP*, 22(1), 2013.

[6] B. Ionescu and al. Div400: A social image retrieval result diversification dataset. *ACM MMSys 2014.*

[7] P. G. Ipeirotis and al. Quality management on amazon mechanical turk. In *HCOMP 2010.*

[8] N. Jain and al. Experiments in diversifying flickr result sets. In *Proc. of MediaEval Wksp. 2013.*

[9] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. of WWW 2008*, pages 297–306.

[10] X. Li and al. Learning tag relevance by neighbor voting for social image retrieval. In *ACM MIR 2008.*

[11] C.-J. Lin and al. Trust region newton method for logistic regression. *JMLR*, 2008.

[12] A. Olteanu and al. Web credibility: Features exploration and credibility prediction. *ECIR 2013.*

[13] P. Sermanet and al. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.

[14] R. H. van Leuken and al. Visual diversification of image search results. In *Proc. of WWW 2009.*

[15] L. Xu and al. Credibility-oriented ranking of multimedia news based on a material-opinion model. *Web-Age Inf. Mgmt.*, pages 290–301, 2011.

[16] Y. Yamamoto and K. Tanaka. Imagealert: credibility analysis of text-image pairs on the web. *SAC 2011.*