FILTERING INSTAGRAM HASHTAGS THROUGH CROWD TAGGING AND THE HITS ALGORITHM

Pindi Manasa , Sri.G.Ramesh Kumar, Sri.V.Bhaskara Murthy,
MCA Student, Assistant Professor, Associate Professor
Dept Of MCA
B.V.Raju College, Bhimavaram

ABSTRACT—Instagram is a rich source for mining descriptive tags for images and multimedia in general. The tags-image pairs can be used to train automatic image annotation (AIA) systems in accordance with the learning by example paradigm. In previous studies we had concluded that, on average, 20% of the Instagram hashtags are related to the actual visual content of the image they accompany, i.e., they are descriptive hashtags, while there are many irrelevant hashtags, i.e., stop-hashtags, that are used across totally different images just for gathering clicks and for searchability enhancement. In this work, we present a novel methodology, based on the principles of collective intelligence, that helps locating those hashtags. In particular, we show that the application of a modified version of the well known HITS algorithm, in a crowd tagging context, provides an effective and consistent way for finding pairs of Instagram images and hashtags, that lead to representative and noise-free training sets for content based image retrieval. As a proof of concept we used the crowdsourcing platform Figure-eight to allow collective intelligence to be gathered in the form of tag selection (crowd tagging) for Instagram hashtags. The crowd tagging data of Figure-eight are used to form bipartite graphs in which the first type of nodes corresponds to the annotators and the second type to the hashtags they selected. The HITS algorithm is first used to rank the annotators in terms of their effectiveness in the crowd tagging task and then to identify the right hashtags per image.

Index Terms—Instagram hashtags, image tagging, image retrieval, crowd tagging, collective intelligence, HITS algorithm, Folk Rank, bipartite graphs.

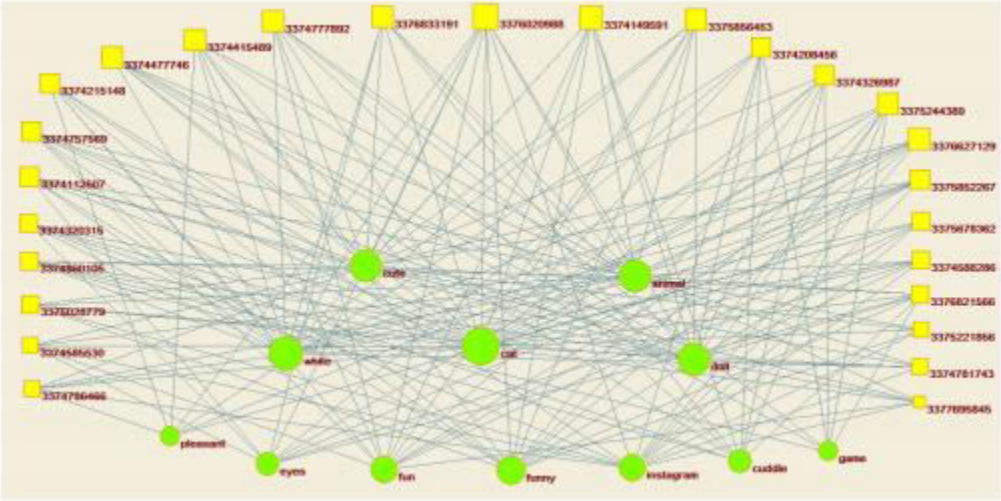
I. INTRODUCTION

SOCIAL media are online communication channels dedicated to community-based input, interaction, contentsharing and collaboration. These media give the users the opportunity to share their content such as, text, video and images [31]. Users usually accompany the content they post with text such as comments or hashtags. That alternative text(comment, hashtags etc.) provide valuable information about the users posts and other information. Preece et al. [32] to construct a Sentinel platform that can enhance social media data in order to understand different situations they based also in Youtube video comments. Sagduyu et al. [33] present a novel system that can present large-scale synthetic data from social media. In their system they use textual content (hashtags and hyperlinks in tweets) to produce topics and train n-gram revised model. The users in several of those media,

e.g. Twitter, Instagram and Facebook, use hashtags to annotate the digital content they upload. Hahshtags are, usually, words or nonspaced phrases preceded by the symbol # that allow creators / content contributors to apply tagging that makes it easier for other users to locate their posts. A great portion of the digital content shared on social media platforms consists of images and short videos. Thus, effective retrieval of images from social media and the web in general, becomes harder and more challenging day by day. Contemporary search engines are basically based on text descriptions to retrieve images, however, inaccurate text descriptions and the plethora of non-textually annotated images, led to extended research for content-based image retrieval techniques [23]. The main problem of content-based image retrieval is the socalled semantic gap [30, 35, 37, 42]: Content-based retrieval is associated with low-level features while humans use highlevel concepts for their search. To overcome this problem, Automatic Image Annotation (AIA) methods were developed, that is, processes by which computing systems automatically assign metadata in the form of captions or keywords to images [4]. Among the AIA methods those based on the learning by example paradigm are probably the most common [21]. A small set of manually annotated training images are used to train models, that learn the correlation between image features and textual words (high level concepts) and then, allow automatic annotation of other (unseen) images. Obviously, good training examples, i.e., representative and accurate pairs of images and related tags are vital in this case [38]. Social media, and especially the Instagram, provide a rich source of image - tag pairs [8, 12]. Mining the right ones, automatically or semi-automatically, so as to be used as training examples is extremely important. We have to consider, however, that, in many cases, hashtags that accompany images in social media are not related with the image's content but serve several other purposes such as the expression of user's emotional state, the increase of user's clicks and findability, and the beginning of a new communication or discussion [7]. In our previous research we have shown that the percentage of the Instagram hashtags that describe the visual content of the image they are associated with, does not exceed 25% [12]. We have also noticed that many Instagram hashtags are used across images that have nothing in common, just for searchability enhancement. We named those hashtags as stophashtags [13]. Thus, filtering the Instagram hashtags in terms of the visual content of the image they accompany is required. HITS is a ranking algorithm than we could use to filter Instagram hashtags and locate the most relevant. The purpose of HITS algorithm, developed by Jon Kleinberg, is to rate Web pages. The basic idea is that web page can provide information about a topic and also relevant links for a topic. Thus, web pages belong into two groups: pages that provide good information about a topic ("authoritative") and those that give to the user good links about a topic ("hubs"). The HITS algorithm gives to each web page both a hub and an authoritative value [27]. We have started experimenting with the HITS algorithm for mining informative

Instangram hashtags in one of our previous works [14] and we extend this study here by considering the application of HITS algorithm in a real crowdtagging environment facilitated by the Figure-eight, formerly known as Crowdflower, crowdsourcing platform. In addition, we have increased the number of annotations per image to 500, we formed the bipartite graphs for all images and we calculated the performance of annotators across all those images. Moreover, FolkRank is used as baseline to evaluate the performance of the proposed method.

II. SYSTEM ARCHITECTURE



III. EXISTING SYSTEM

- In our previous research we have shown that the percentage of the Instagram hashtags that describe the visual content of the image they are associated with, does not exceed 25%.
- We have also noticed that many Instagram hashtags are used across images that have nothing in common, just for searchability enhancement. We named those hashtags as stop hashtags
- Thus, filtering the Instagram hashtags in terms of the visual content of the image they accompany is required.
- A diffusion process on the tag-item bipartite graph of the collection was then applied by using the estimated tag weights. The experiments, conducted on three different datasets, showed superiority of the proposed method over the traditional tag-based collaborative filtering approach that is usually adopted in recommender systems.

Disadvantage:

- An additional difficulty comes from the fact folksonomies are usually modelled as unidirected graphs, i.e., humans select tags for an item.

IV. PROPOSED SYSTEM

- The users in several of those media, e.g. Twitter, Instagram and Facebook, use hashtags to annotate the digital content they upload. Hashtags are, usually, words or no spaced phrases preceded by the symbol # that allow creators / content contributors to apply tagging that makes it easier for other users to locate their posts.
- A great portion of the digital content shared on social media platforms consists of images and short videos.
- Thus, effective retrieval of images from social media and the web in general, becomes harder and more challenging day by day.
- Content-based retrieval is associated with low-level features while humans use high level concepts for their search.
- To overcome this problem, Automatic Image Annotation (AIA) methods were developed, that is, processes by which computing systems automatically assign metadata in the form of captions or keywords to images

Algorithm:

HITS algorithm. Hyperlink-Induced Topic Search (**HITS**; also known as hubs and authorities) is a link analysis **algorithm** that rates Web pages, developed by Jon Kleinberg.

Advantage:

- Duplicate tags for the same image were identified and removed. Another important pre-processing step was the splitting of hashtags into their constituting words with the help of the library

V. IMPLEMENTATION

Admin:

Admin login into their account. View the detail of user and all friend's status. In this project are admin working us performing just view and allow an Instagram for user. view all images and then view recommended images, view image reviews, view dislikes. Finally, admin viewer HITS result.

User:

User register their own detail and login their account. Admin are accepted a user then view Instagram post. However, upload an image sharing him/her friends and then recommended an image should be given a review. A user is adding an image Instagram. Searching a friend and then view to friend's request. Searching an image and give reviews.

VI. CONCLUSION

In the current work, we have presented an innovative methodology, based on the HITS algorithm and the principles of collective intelligence, for the identification of Instagram hashtags that describe the visual content of the images they are associated with. We have empirically shown that the application of a two-step HITS algorithm in a crowdtagging context provides an easy and effective way to locate pairs of Instagram images and hashtags that can be used as training sets for content based image

retrieval systems in the learning by example paradigm. As a proof of concept we have used 25000 evaluations (500 annotations for each one of 50 images) collected from the Figure-eight crowdsourcing platform to create a bipartite graph composed of users (annotators) and the tags they selected to describe the 50 images. The hub scores of the HITS algorithm applied on this graph, called hereby full bipartite graph, give us a measure of reliability of the annotators. The aforementioned approach is based on the findings of Theodosiou et al. [39] who claim that the reliability of annotators better approximated if we consider all the annotations they have performed rather than the subset of Gold Test Questions. In a second step a weighted bipartite graph for each image is composed in the same way as the full bipartite graph. The weights of these graphs are the hub scores computed in the previous step. By thresholding the authority scores of the per image graphs, obtained by the application of the HITS algorithm on the weighted graphs, we can rank and then effectively locate the hashtags that are relevant to their visual content as per the annotators evaluation. Some important findings of the current work are briefly summarized here. The first refers to the value of crowdtagging itself. As in several studies before we found that the crowd can substitute the experts in the evaluation of images w.r.t. relevant tags.

REFERENCES

[1] A. Argyrou, S. Giannoulakis and N. Tsapatsoulis,“Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation?” in Proc. 13th International Workshop on Semantic and Social Media Adaptation and Personalization, 2018, pp. 61-67.

[2] B. I. Aydin, Y. S. Yilmaz, Y. Li, Q. Li, J. Gao, and M. Demirbas,“Crowdsourcing for multiple-choice ques-tion answering” in Proc. 28th. AAAI Conference on Artificial Intelligence, 2014, pp. 2946–2953.

[3] C. D. Cabrall, Z. Lu, M. Kyriakidis, L. Manca, C. Dijksterhuis, R. Happee, and J. de Winter, “Validity and reliability of naturalistic driving scene categorization judgments from crowdsourcing,” Accident Analysis & Prevention, vol. 114, pp. 25–33, 2018.

[4] Q. Cheng, Q. Zhang, P. Fu, C. Tu, and S. Li, “A survey and analysis on automatic image annotation,” Pattern Recognition, vol. 79, pp. 242–259, 2018.

[5] N. Craswell, “Mean Reciprocal Rank,” in Encyclopedia of Database Systems, London : Springer, 2009, pp. 1703-1703.

[6] H. Cui, Q. Li, H. Li, and Z. Yan, “Healthcare fraud detection based on trustworthiness of doctors,” in Proc. Trustcom/BigDataSE/I SPA, IEEE, 2016, pp. 74–81.

[7] A. R. Daer, R. Hoffman, and S. Goodman, “Rhetorical functions of hashtag forms across social media applications,” in Proc. 32nd ACM Int. Conf. on the Design of Communication CD-ROM, ACM, 2014, p. 16.

[8] E. Ferrara, R. Interdonato, and A. Tagarelli, “Online popularity and topical interests through the lens of instagram,” in Proc. 25th ACM Conf. on Hypertext and Social Media, ACM, 2014, pp. 24–34.

[9] J. M. Fletcher and T. Wennekers, “From structure to activity: Using centrality measures to predict neuronal activity,”

International Journal of Neural Systems, vol. 28, no. 02, p. 1750013, 2018.
[10] M. Gao, L. Chen, B. Li, Y. Li, W. Liu, and Y.-c. Xu, “Projectionbased link prediction in a bipartite network,” Information Sciences, vol. 376, pp. 158–171, 2017.