FILTERING INSTAGRAM HASHTAGS THROUGH CROWD TAGGING AND THE HITS ALGORITHM

CHAMALA CHINNA SIVAKUMAR REDDY[1], KAMBHAM SALIVAHANA REDDY[2]

[1]MTech Student, Department of **CSE**, **Global College of Engineering and Technology**, kadapa, AP.

[2]Assistant professor, Department of **CSE**, **Global College of Engineering and Technology**, kadapa, AP.

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 hash tags are related to the actual visual content of the image they accompany, i.e., they are descriptive hash tags, while there are many irrelevant hash tags, i.e., stop-hash tags, that are used across totally different images just for gathering clicks and for search ability enhancement. In this paper, we present a novel methodology, based on the principles of collective intelligence that helps in locating those hash tags. In particular, we show that the application of a modified version of the well-known hyperlink induced topic search (HITS) algorithm, in a crowd tagging context, provides an effective and consistent way for finding pairs of Instagram images and hash tags, which lead to representative and noise-free training sets for content-based image retrieval. As a proof of concept, we used the crowd sourcing platform *Figure-eight* to allow collective intelligence to be gathered in the form of tag selection (crowd tagging) for Instagram hash tags. 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 hash tags 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 hash tags per image.

Keywords: HITS, Hash tags, CNN, ML.

DOI Number: 10.14704/nq.2022.20.12.NQ77052 **NeuroQuantology 2022; 20(12): 690-695**

1. 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. This alternative text (comment, hashtags, etc.) provides valuable information about the user posts and other information. Preece et al. [32] 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 the n-gram 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 nontextually annotated images led to extended research for content-based image retrieval techniques

The main problem of the content-based image retrieval is the so-called semantic gap [30], [35], [37], [42]: 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 [4]. Among the AIA methods, those based on the learning by example paradigm are probably the most common one [21]. A small set of manually annotated training images are used to train models, which 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 semiautomatically, 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 in user's clicks and findability, and the beginning of a new communication or discussion

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. Hyperlink-induced topic search (HITS) is a ranking algorithm than we could use to filter Instagram hash tags and locate the most relevant. The purpose of the HITS algorithm, developed by Jon Kleinberg, is to rate WebPages. The basic idea is that a webpage can provide information about a topic and also relevant links for a topic. Thus, webpages belong to 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 webpage both a hub and an authoritative value [27]. We have started experimenting with the HITS algorithm for mining informative Instagram hashtags in one of our previous works [14] and we extend this paper here by considering the application of the HITS algorithm in a real crowd tagging environment facilitated by the Figure-eight, formerly known as Crowd flower, crowd sourcing 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, Folk Rank is used as a baseline to evaluate the performance of the proposed method.

2. LITERATURE SURVEY

- ❖ Mitry et al. [1] compared the accuracy of crowdsourced image classification with that of experts. They used 100 retinal fundus photography images selected by two experts. Each annotator was asked to classify 84 retinal images while the ability of annotators to correctly classify those images was first evaluated on 16 practice - training images. The study concluded that the performance of naive individuals to retinal image classifications was comparable to that of experts. Giuffrida et al. [15] measured the inconsistency among experienced and non-experienced users in that task of leaf counts in images of Arabidopsis Thaliana. According to their results everyday people can provide accurate leaf counts.
- ❖ Maier-Hein et al. [2] investigated the effectiveness of large-scale crowdsourcing on labelling endoscopic images and concluded that

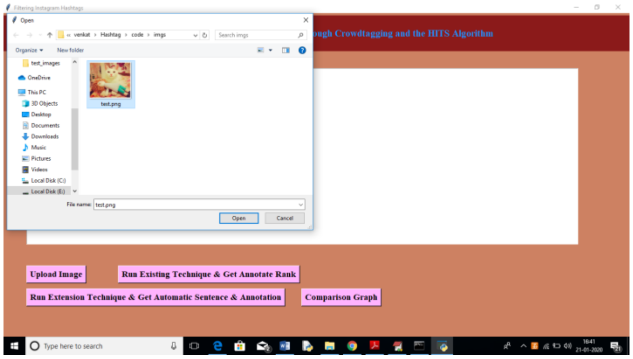
non-trained workers perform comparably to medical experts.

- ❖ Cabrall et al. [3] in their survey for drive scene categorization they used the crowd to annotate driving scene features such as presence of other road users and bicycles, pedestrians etc.
- ❖ Zhang et al. [4] tried to extract people's opinions on features (characteristics) of electronic products such as mobile phones, tablets etc. In order to rank the importance of those characteristics they constructed a two-mode network where features were modelled as authorities and feature relevance indicators as hubs. With the aid of the HITS algorithm they were able to identify highly-relevant features and good feature indicators by thresholding the corresponding authority and hub values respectively.
- ❖ Nguyen and Jung [5] used a variation of the HITS algorithm, called GeoHITS, to rank locations with respect to specific tags such as those related with food types. Both tags and locations were collected from geo-tagged resources on social network services. The authors used a subset of tags that shared across several locations to act as hubs while the locations were considered as the authorities.
- ❖ Cui et al. [6] proposed a healthcare fraud detection approach which is based on the trustworthiness of doctors to distinguish fraud cases from normal records. They created a doctor-patient two-mode network which was represented as a weighted bipartite graph. The prescription behavior in patients' healthcare records was used to compute the edge weights. According to the authors the hub scores of the HITS algorithm provide a good estimation of the trustworthiness of doctors.

PROPOSED SYSTEM:

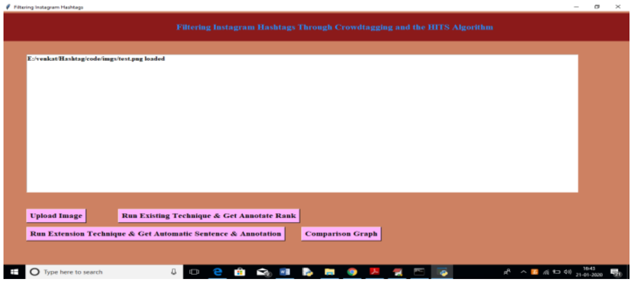
In the proposed system, the system presents a novel methodology, based on the principles of collective intelligence that helps locating those hash tags. In particular, we show that the application of a modified version of the well known HITS algorithm, in a crowd tagging context, provides an effective and consistent way for finding pairs of Instagram images and hash tags, that lead to representative and noise-free training sets for content based image retrieval. As a proof of concept we used the crowd sourcing platform Figure-eight to allow collective intelligence to be gathered in the form of tag selection (crowd tagging) for Instagram hash tags. 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 hash tags 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 hash tags per image.

3. METHODOLOGY



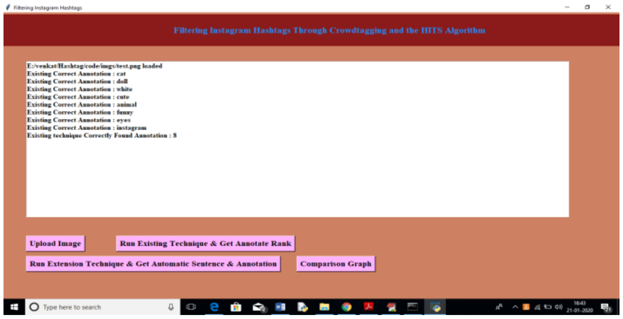
File name: test.png

In above screen I am uploading one image and by seeing that image anybody can say that cat or kitten sitting on a bed with some stuff and our extension will describe same sentence or extract same data from image but existing technique just will check whether given hash tag and annotator tags are similar or relevant or not relevant. After uploading image will get below screen


Filtering Instagram Hashtags
Filtering Instagram Hashtags Through Crowdtagging and the HITS Algorithm
Upload Image
Run Existing Technique & Get Annotate Rank
Run Extension Technique & Get Automatic Sentence & Annotation
Comparison Graph
Type here to search
16:43
21-01-2020

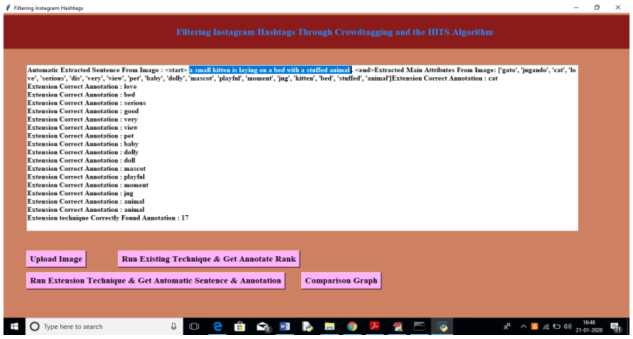
```
E:/venkat/Hashtag/code/imgs/test.png loaded
```

Now click on 'Run Existing Technique & Get Annotation Rank' button to get below screen


Filtering Instagram Hashtags
Filtering Instagram Hashtags Through Crowdtagging and the HITS Algorithm
Upload Image
Run Existing Technique & Get Annotate Rank
Run Extension Technique & Get Automatic Sentence & Annotation
Comparison Graph
Type here to search

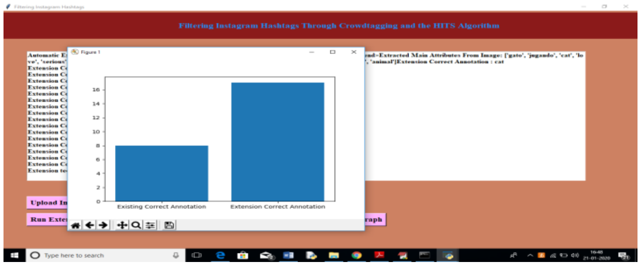
```
E:/venkat/Hashtag/code/imgs/test.png loaded
Existing Correct Annotation : cat
Existing Correct Annotation : doll
Existing Correct Annotation : white
Existing Correct Annotation : cute
Existing Correct Annotation : animal
Existing Correct Annotation : funny
Existing Correct Annotation : eyes
Existing Correct Annotation : instagram
Existing technique Correctly Found Annotation : 8
```

In above screen we can see from loaded images above annotations are correct as image contains cat, doll, cute etc. Existing technique able to extract 8 correct annotation from all annotated text. Now click on ‘Run Extension Technique & get Automatic Sentence & Annotation’ button to describe image in sentence and to check extracted words are matching with annotators words or not.


Filtering Instagram Hashtags
Filtering Instagram Hashtags Through Crowdtagging and the HITS Algorithm
Upload Image
Run Existing Technique & Get Annotate Rank
Run Extension Technique & Get Automatic Sentence & Annotation
Comparison Graph
Type here to search

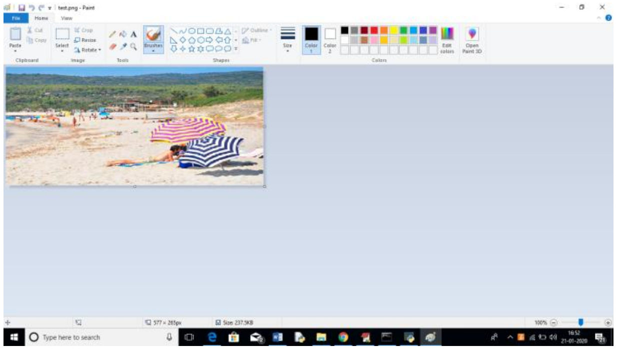
```
Automatic Extracted Sentence From Image : <start> a small kitten is laying on a bed with a stuffed animal . <end>Extracted Main Attributes From Image: ['gato', 'jugando', 'cat', 'lo
ve', 'serious', 'dis', 'very', 'view', 'pet', 'baby', 'dolly', 'mascot', 'playful', 'moment', 'jug', 'kitten', 'bed', 'stuffed', 'animal']Extension Correct Annotation : cat
Extension Correct Annotation : love
Extension Correct Annotation : bed
Extension Correct Annotation : serious
Extension Correct Annotation : good
Extension Correct Annotation : very
Extension Correct Annotation : view
Extension Correct Annotation : pet
Extension Correct Annotation : baby
Extension Correct Annotation : dolly
Extension Correct Annotation : doll
Extension Correct Annotation : mascot
Extension Correct Annotation : playful
Extension Correct Annotation : moment
Extension Correct Annotation : jug
Extension Correct Annotation : animal
Extension Correct Annotation : animal
Extension technique Correctly Found Annotation : 17
```

In above screen in selected text you can see our extension technique describing image in a sentence and then extracting words from image and compare with annotator’s tags to get relevant details. Extension technique able to extract 17 related annotations.Now click on ‘Comparison Graph’ button to get below graph

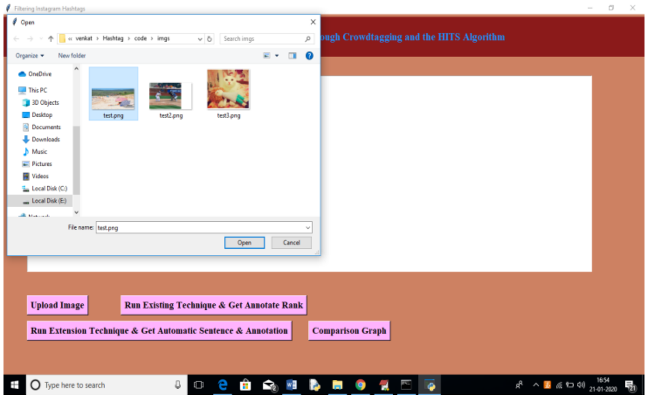


In above graph x-axis represents technique name and y-axis represents count of extracted matching annotations and we can see extension technique able to extract more related words compare to existing technique.

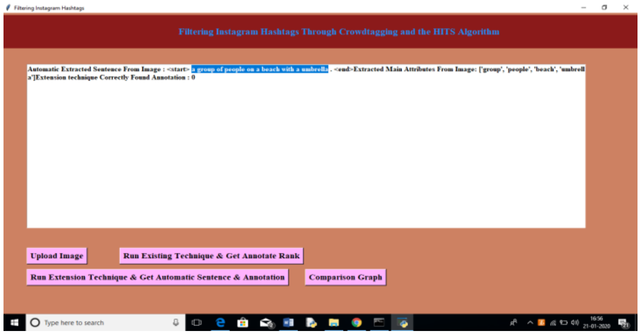
Note: existing technique can able to check with only one image as author given only one image details in paper and what other images he has used that information is not available. But extension technique can work with any image. See another image example



In above image we can see peoples are on beach with umbrellas and extension technique can extract this information but cannot compare with existing technique as author not include this image in his annotation dataset



In above screen uploading same image and then click on ‘Run Extension Technique & get Automatic Sentence & Annotation’ button to get below results


Filtering Instagram Hashtags
Filtering Instagram Hashtags Through Crowdtagging and the HITS Algorithm
Upload Image
Run Existing Technique & Get Annotate Rank
Run Extension Technique & Get Automatic Sentence & Annotation
Comparison Graph
Type here to search

```
Automatic Extracted Sentence From Image : <start> a group of people on a beach with a umbrella . <end>Extracted Main Attributes From Image: ['group', 'people', 'beach', 'umbrell
a']Extension technique Correctly Found Annotation : 0
```

In above screen in selected text you can see sentence describing image and its related attributes or hashtag also displaying

CONCLUSION

In this paper, we have presented an innovative methodology, based on the HITS algorithm and the principles of collective intelligence, for the identification of Instagram hash tags 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 25 000 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 to this graph, called hereby full bipartite graph, give us a measure of the reliability of the annotators. The aforementioned approach is based on the findings of Theodosius et al. [39], in which the reliability of annotators is better approximated if we consider all the annotations they have performed rather than the subset of gold test questions. In the 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 this paper are briefly summarized here. The first refers to the value of crowdtagging itself. In several studies before, we found that the crowd can substitute the experts in the evaluation of images with respect to relevant tags. However, even with a large number of annotators (499 in our case), it seems that a perfect agreement between annotators and experts cannot be achieved. In particular, it was found that from the 145 different tags suggested for the 50 images used in this paper by the two experts, only 135 were also identified by the 499 annotators. This leads to a maximum achievable recall value equal to 0.931. Thus, in subjective evaluation tasks, such as those referring to the identification of tags that are related to the visual content of images, no perfect agreement between the experts and the crowd should be expected.

A second finding is that crowdtagging of images can be effectively modeled through user–tag bipartite graphs, one per image. Thresholding the authority score of the HITS algorithm applied on these graphs is a robust way to identify the tags that characterize the visual content of the corresponding images. Getting the top ranked tags based on the authority score is an alternative solution, but, with a little bit lower effectiveness.

REFERANCES

[1] A. Argyrou, S. Giannoulakis, and N. Tsapatsoulis, “Topic modelling on Instagram hashtags: An alternative way to automatic image annotation?” in Proc. 13th Int. Workshop Semantic Social Media Adaptation 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 question answering,” in Proc. 28th. AAAI Conf. Artif. Intell., 2014, pp. 2946–2953.

[3] C. D. D. Cabrall et al., “Validity and reliability of naturalistic driving scene categorization judgments from crowdsourcing,” Accident Anal. Prevention, vol. 114, pp. 25–33, May 2018

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

[5] N. Craswell, “Mean reciprocal rank,” in Encyclopedia of Database Systems. London, U.K.: Springer, 2009, p. 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, 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. Design Commun. CD-ROM, 2014, Art. no. 16.

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

[9] J. M. Fletcher and T. Wennekers, “From structure to activity: Using centrality measures to predict neuronal activity,” Int. J. Neural Syst., vol. 28, no. 2, 2018, Art. no. 1750013.

[10] M. Gao, L. Chen, B. Li, Y. Li, W. Liu, and Y.-C. Xu, “Projection-based link prediction in a bipartite network,” Inf. Sci., vol. 376, pp. 158–171, Jan. 2017.

[11] S. I. Gass and C. M. Harris, “Bipartite graph,” in Encyclopedia of Operations Research and Management Science. Boston, MA, USA: Springer, 2013, p. 126.