ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

Instagram Hashtag Filtering Application System using HITS Algorithm

[1]*K. Venkateswarlu* [2]*G. Raghuvaran*

[1] *Assistant Professor, Dept. of Master of Computer Applications, Narayana Engineering College, Gudur, Nellore Dist, AP, India*

[2] *PG Scholar, Dept. of Master of Computer Applications, Narayana Engineering College, Gudur, Nellore Dist, AP, India*

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 paper, we present a novel methodology, based on the principles of collective intelligence that helps in locating those hashtags. In particular, we show that the application of a modified version of the well-known hyper link induced topic search (HITS) algorithm, in a crowd-tagging context, provides an effective and consistent way for finding pairs of Instagram images and hashtags, which 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.

Keywords:Bipartite Graphs, Collective Intelligence, Crowd-Tagging, Folk-Rank, Hyperlink-Induced Topic Search (HITS) Algorithm, Image Retrieval, Image Tagging, Instagram Hashtags.

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. 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.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. 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 trainthe 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.

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

Hashtags are usually, words or non-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 fromsocial 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.

The main problem of the content-based image retrieval is the so-called semantic gap: 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. 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. Social media, and especially the Instagram, provide a rich source of image–tag pairs [8], [12].

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 rankingalgorithm than we could use to filter Instagram hashtags 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. 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

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

of the HITS algorithm in a real crowd-tagging environment facilitated by the Figure-eight, formerly known as Crowdflower, crowdsourcing platform.

2. RELATED WORK

The validity of crowdsourced image annotation was examined and verified by several researchers. Mitry et al. 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. investigated the effectiveness of large-scale crowdsourcing on labelling endoscopicimages and concludedthat non-trainedworkers perform comparablyto medical experts. Cabrall et al. [3] used the crowd to annotate driving scene features such as the presence of other road users and bicycles and pedestrians for drive scene categorization.

The initial purpose of the HITS algorithm was to discover and rate webpages that are relevant to a topic. In social network analysis, the HITS algorithm, and specifically the hub and authority value it computes, is used for estimating the centrality of nodes, especially in networks composed of two types of nodes known as two-mode networks. A typical example of such networks is the bipartite networks that are usually modelled through bipartite graphs. A bipartite graph is a graph whose nodes can be divided into two distinctive groups (partitions) whileits edges connect nodes among partitions but not within each partition [10], [11].

The purpose of user profiling is to understand and code the personal interests of users so as to provide advanced and personalized services. They modelled the social tagging system as a user-tag network and applied PageRank and HITS to refine the weights of tags. 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 data sets, showed superiority of the proposed method over the traditional tag-based collaborative filtering approach that is usually adopted in recommender systems.

We have seen in the previousparagraphsthat the HITS algorithm has been successfully applied in real-world problemsthat can be modelled through bipartite graphs. At the same time, crowdsourced image annotation is gaining popularity through the wide use of dedicated

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

crowdsourcing platforms. However, the problem of crowdsourced image tagging has never been modelled as a two-mode network probably because it involves three different types of entities: annotators, images, and tags. We overcome the three-entity problem by applying the HITS algorithm in two consecutive steps and on two different bipartite graphs. We first estimate the reliability of annotators (contributors in the language of Figure-eight) by utilizing the hub value of the full bipartite graph consisting of the annotators and the tags they selected and used across all images. Then, the annotator hub values are used as tie-weights on bipartite graphs constructed per Instagram image. The authority values of the tags, computed through the HITS algorithm, give us a ranking in terms of relevance between the hashtags and the image they accompany and are used to filter out the relevant from the irrelevant hashtags.

3. PROPOSED WORK

A) System Model

The system model shown in Figure-1 comprises the following entities:Instagram Server, User.

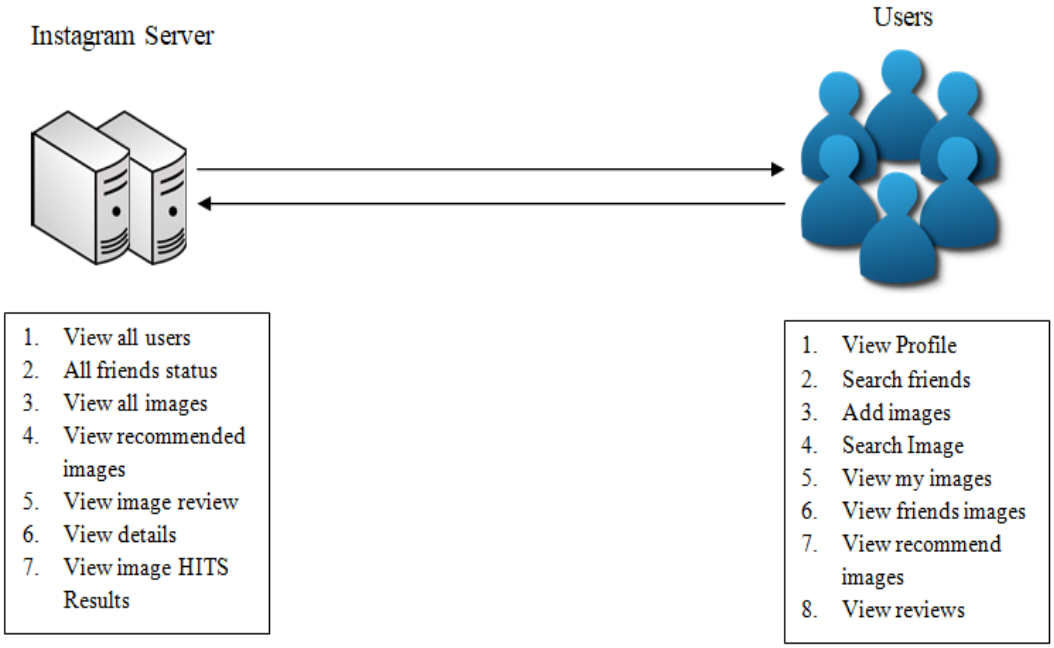

Fig. 1. System Model

This system model consists and implements the following modules:

Instagram Server:

- In this module, Instagram server login to system.

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

- Here, he can perform the various operations like, view all users and authorize them, view all images with reviews, view dislikes, view the images with HITs algorithms.
- He can view the Instagram image tags.

Users:

- In this module, user can register and login to the system.
- Here, he can post images, search images, view all images he posted and recommend images to friends, and add reviews to images.
- He can send the friend request and responses.

Hyperlink Induced Topic Search

- The HITS algorithm was initially introduced and in order to analyze a collection of webpages, relevant to a topic, and locate the most "authoritative" ones in that topic. It performs link analysis on those webpage's in order to rank them in terms of two measures: hub value and authoritativeness. The authority score estimates the importance of the content of the page, while the hub score estimates the quality of its links to other pages.
- The HITS algorithm is commonly used for the analysis of two-mode networks represented as bipartite graphs. In that case, both authority and hub values are used as measures of centrality2; however, their interpretation differs significantly. A vertex with high authority score is considered as an expert, while a vertex with high hub value is assumed as a good recommender.

Crowd Sourcing

- The validity of crowdsourced image annotation was examined and verified by several researchers. Mitry*et al.* 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.

B) Design Goals

Input Design:

- In this design we maintain the user details and data set.
- We design the following pages to collect the data.

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

- They are:

Registration:

This page collects the data from users.

Login:

This page collects username and password from user, validate the data and store.

Output Design:

- In the output design, we design the output pages to represent the results of our proposed method.
- For that we design different page as follows:

Search Result:

This page shows the search results of the system.

And other pages carry the details of users, user search history, location details and so on.

C) Implementation Methods and Algorithms

Method:

- Step 1: The relevance of each hashtag with respect to the visual content of the associated image is assessed by a set of N users (annotators) with the aid of a crowdsourcing platform.
- Step 2: Step 2: Given that all users assessed all image hashtags, we can rank their effectiveness by considering the HITS algorithm. For that purpose, we construct a bipartite graph.
- Step 3: The effectiveness (reliability) of annotators is approximated with the set of hub values computed with the aid of the HITS algorithm.
- Step 4: For each image, we construct a weighted bipartite graph.
- Step 5: A ranked set of tags, for each Instagram image is achieved through the set of authority values, computed with the aid of the HITS algorithm when it is applied on the weighted bipartite graphs that were created in the previous step.

Algorithm:

- ✓ The HITS algorithm was initially introduced in order to analyze a collection of webpages, relevant to a topic, and locate the most “authoritative” ones in that topic.
- ✓ It performs link analysis on those webpages in order to rank them in terms of two measures: hub value and authoritativeness.

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

- ✓ The authority score estimates the importance of the content of the page, while the hub score estimates the quality of its links to other pages.
- ✓ Thus, a webpage that has many in-links from other pages with high hub value is considered an authority, while a page with many out-links to high authority webpages is a hub.

D) Requirements

Hardware Requirement: Processor:Dual Core 1.6 GHz,RAM :2 GB,Hard Disk:500 GB**;Software Requirement:**Operating System:Window 7 or above, Programming Language:JAVA, Front End:HTML& CSS,Back End:JSP& Servlets,Database:MySQL 5.0, Server:Apache Tomcat.

4. EXPERIMENTAL RESULTS & DISCUSION

In this section, we provide the performance assessment of the proposed system model.

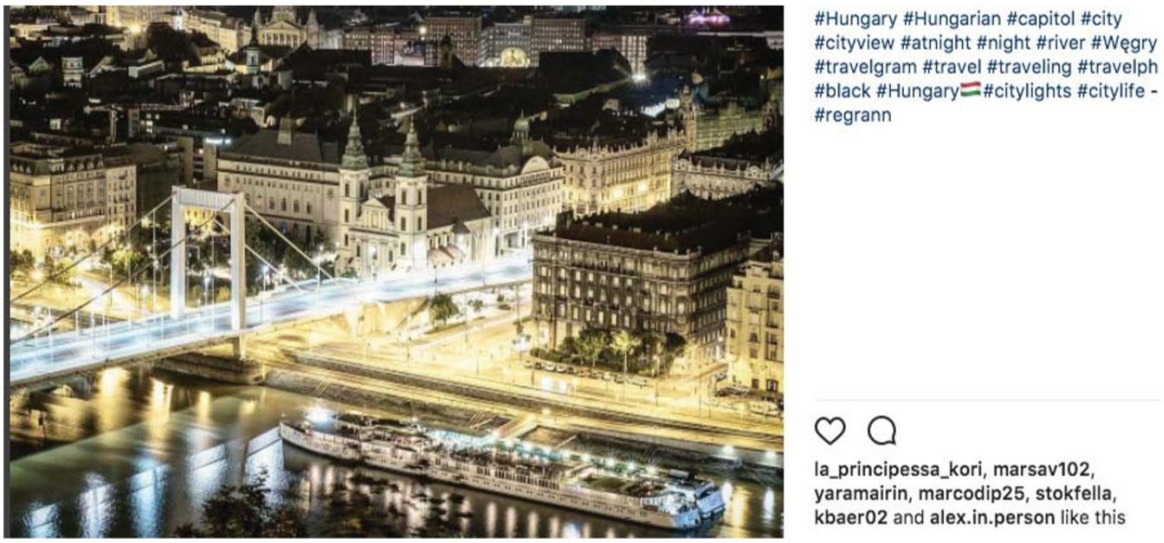

Fig. 2. Example of an Instagram image: at the top right, the associated hashtags are attached to it.

We present the MAP and MRR results according to (13)–(14). The correspondingreceiveroperatingcharacteristic curves6 (ROC) are shown in Fig. 4. For convenient juxtaposition with the values presentedin this ROC curve,the precision versus

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

recall is plotted instead of the typical case of ROC curves in which the true positive rate versus the false positive rate are usually plotted.

Fig. 3. Example of hashtag selection process that took place via Figure-eight.

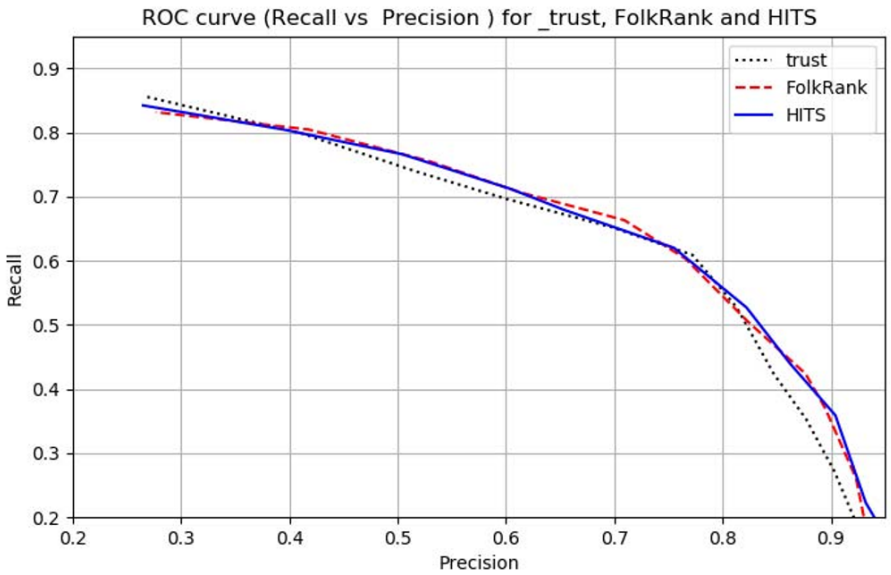

Fig. 4. Recall versus precision ROC curves for the _trust (AUC = 0.680), the Folk-Rank (AUC = 0.689), and the HITS (AUC = 0.692) weighting schemes.

We observe inFig. 4 that the best results in terms of the F1 measure are obtained for an authority score threshold value θ = 0.11. However, as in most information retrieval systems, we usually prefera highervalueofrecall that is identifyingmoretags even if it is notthat accurate,instead of precision.Thus,an authority score threshold θ =0.09 also gives us a reasonable choice.

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

Another important metric that is used to evaluate the performance of information retrieval systems is the area under the (ROC) curve (AUC or AUROC). Since both precision and recall take values in the range [0,1], AUC also ranges in [0,1]. The intuition behind this metric is that an AUC of 0.5 represents a randominformationretrieval system (or, similarly, an uninformative two-class classifier), while an AUC equal to 1 represents the perfect informationretrieval system. The AUC corresponding to the ROC curve of Fig. 4 is equal to 0.692.

The results, for a variety of k values, are given in the correspondingROC curve is shown in Fig. 5. We see that the best F1 scores are achieved by keeping either the top three or the top four ranked hashtags per image. Keeping four hashtags per image favors the recall value which, as already discussed above, is preferable for the majority of information retrieval systems.

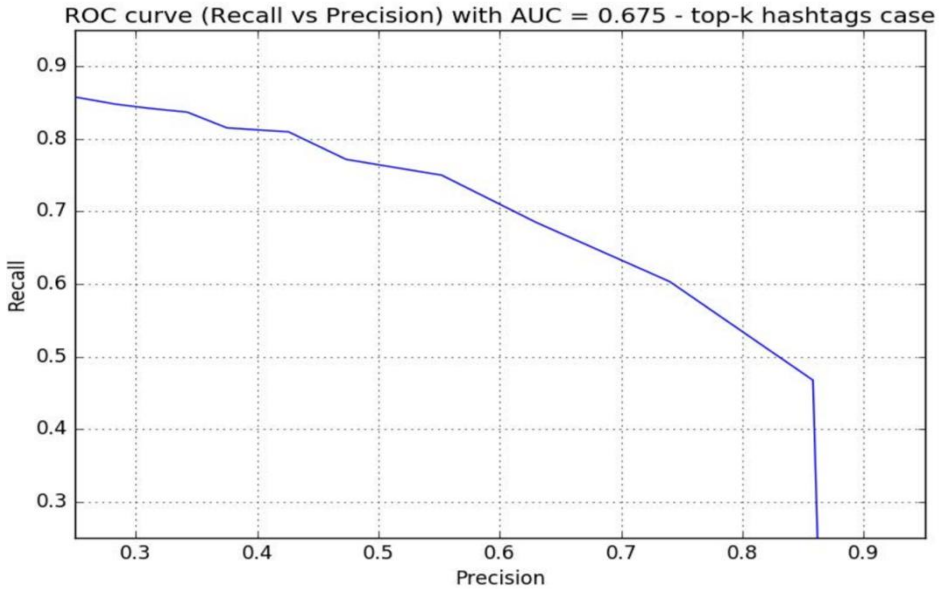

Fig. 5. Recall versus precision ROC curve with an AUC equal to 0.675—the case of top-k hashtags.

We see also in Fig. 5 that the AUC is 0.675, which is comparable with the authority score thresholding case. This means that there is no significant variation of the agreed hashtags per image; therefore, keeping the k top-ranked hashtags based on the authority score is another option for mining tags from Instagram hashtags accompanying images.

5. 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 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

ISSN: 2278-4632
Vol-10 Issue-5 No. 14 May 2020

an easy and effective way to locate pairs of Instagram images and hashtags that can be used as trainingsets 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 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 Theodosiou et al., 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 annotator's evaluation.

REFERENCES

[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. Cabrallet 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.

[12] S. Giannoulakis and N. Tsapatsoulis, "Evaluating the descriptive power of Instagram hashtags," J. Innov. Digit. Ecosyst., vol. 3, no. 2, pp. 114–129, 2016.

[13] S. Giannoulakis and N. Tsapatsoulis, "Defining and identifying stophashtags in instagram," in Proc. INNS Conf. Big Data. Cham, Switzerland: Springer, 2016, pp. 304–313.

[14] S. Giannoulakis, N. Tsapatsoulis, and K. Ntalianis, "Identifying image tags from Instagram hashtags using the HITS algorithm," in Proc. 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech), 2017, pp. 89–94.

[15] M. V. Giuffrida, F. Chen, H. Scharr, and S. A. Tsaftaris, "Citizen crowds and experts: Observer variability in image-based plant phenotyping," Plant Methods, vol. 14, no. 1, p. 12, 2018.

Author's Profile:

K. Venkateswarlu has received his MCA degree from Saraswathi Velu College of Engineering, Vellore affiliated to Anna University, Chennai in 2010 and MTech degree in Computer Science from PBR VITS, Kavali affiliated to JNTU, Ananthapur in 2014 respectively. He is dedicated to teaching field from the last 6years. He has guided 25 P.G students. At present he is working as an Assistant Professor in Narayana Engineering College, Gudur, AndhraPradesh, India.

G. Raghuvaranhas Received his B.Sc Degree in Computer Science from Krishna Chaitanya Degree College, Nellore affiliated to VikramaSimhapuri University, Nellore in 2017 and pursuing PG Degree in Master of Computer Applications (M.C.A) from Narayana ngineering College, Gudur affiliated to JNT University, Anantapur, Andhra Pradesh, India.

