Filtering Instagram hashtags through crowdtagging and the HITS algorithm

Stamatios Giannoulakis and Nicolas Tsapatsoulis

***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 crowdtagging 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 (crowdtagging) for Instagram hashtags. The crowdtagging 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 crowdtagging task and then to identify the right hashtags per image.**

***Index Terms*—Instagram hashtags, image tagging, image retrieval, crowdtagging, collective intelligence, HITS algorithm, FolkRank, bipartite graphs.**

I. Introduction

SOCIAL media are online communication channels dedicated to community-based input, interaction, content-sharing 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 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 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 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 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 [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

S. Giannoulakis and N. Tsapatsoulis are with the Department of Communication and Internet Studies, Cyprus University of Technology, 30, Arch. Kyprianos str., CY-3036, Limassol, Cyprus (e-mail: s.giannoulakis@cut.ac.cy, nicolas.tsapatsoulis@cut.ac.cy).

Manuscript received ; revised

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. Related work

The validity of crowdsourced image annotation was examined and verified by several researchers. Mitry *et al.* [28] 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.* [25] 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. They used the Crowdflower platform (now *Figure-eight*) in the categorization of large amounts of videos with diverse driving scene contents. As usual the Gold Test Questions in Crowdflower were used to verify that the annotators perform well in their job. The results indicated that crowdsourcing through the Crowdflower was effective in categorizing naturalistic driving scene contents.

The initial purpose of the Hyperlink-Induced Topic Search (HITS) algorithm was to discover and rate web-pages that are relevant to a topic (see also Section III-C). In social network analysis the HITS algorithm, and specifically the hub and authority values 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 are the bipartite networks which are usually modelled through bipartite graphs. A bipartite graph is a graph whose nodes can be divided into two distinctive groups (partitions) while its edges connect nodes among partitions but not within each partition [10, 11].

Two-mode (bipartite) networks are frequently used to model recommender systems [43], since consumers and products correspond to two different type of entities and usually the consumers choose or rate products. Mao *et al.* [26] applied HITS (and the PageRank as well) to improve user profiling in a social tagging system. The purpose of user profiling is to understand and code the personal interests of users so as to provide them 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 datasets, showed superiority of the proposed method over the traditional tag-based collaborative filtering approach that is usually adopted in recommender systems.

Zhang *et al.* [47] 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 [40] 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. London and Csendes [22] applied a modified version of the HITS algorithm called Co-HITS to evaluate the professional skills of wine tasters. In order to achieve this goal, they constructed a weighted bipartite graph composed of wine tasters, modeled as hubs, and wines, modeled as authorities. The weights correspond to the scores given by the wine-tasters to wines. According to the authors, the computed hub values can be used to filter out incompetent tasters while they are highly correlated with the competence of wine tasters.

Tseng *et al.* [44] tried to distinguish fraudulent remote phone calls from normal ones by considering that the trust value of remote phone numbers is related with the hub score of the HITS algorithm. For that purpose they used telecommunication records to create directed bipartite graphs with

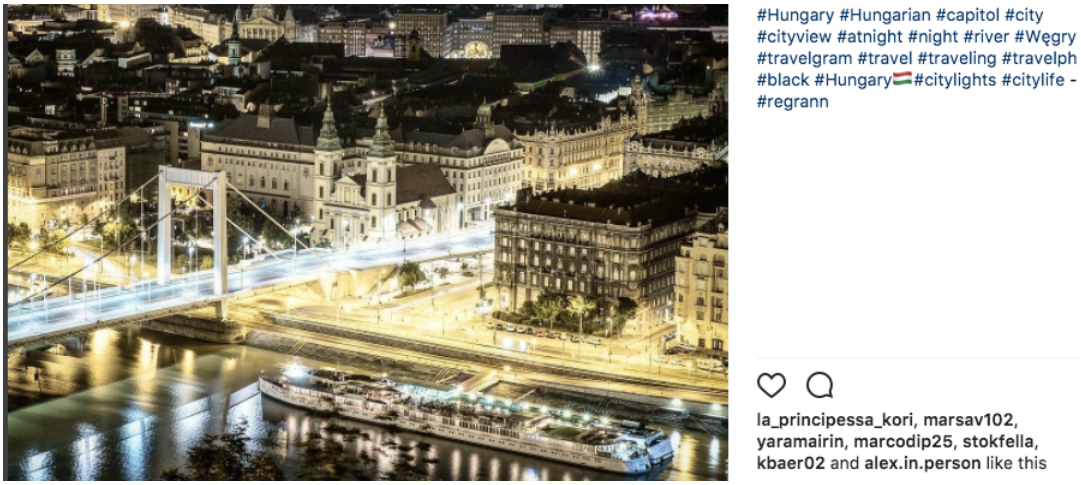

Fig. 1. An example of an instagram image: At the top right the associated hashtags attached to it.

incoming and outgoing calls between contact book entries of the users, assumed as authorities, and remote phone numbers (phone numbers not in contact books), assumed as hubs. The edge weights for each pair of user and remote phone number were computed based on duration and frequency relatedness between a user and a remote phone number. With the application of HITS the trust value for each remote phone number was computed and used to classify remote calls into fraudulent and normal.

There are also a few works in which the HITS algorithm was used in a crowdsourced environment, as we do in the current work for the specific case of image tagging. However, in the majority of cases the emphasis is put on the evaluation - enhancement of the quality of the crowdsourced data rather than to information mining. Sunahase *et al.* [36] applied the so called Pairwise HITS algorithm, a modification of the HITS algorithm which is applicable to pairwise comparisons, to three different tasks: image description, logo designing and article language translation. The aim was to estimate the quality of produced data and the ability of evaluators to assess those data through pairwise comparisons of image descriptions, logo designs and article translations created by two different creators - data producers. Schall *et al.* [34] tried to evaluate crowdsourcing participants (coordinators, supervisors and workers) used for business process. They created a two-mode social graph for each coordinator that processes a task from a customer. Supervisors, that separate the task into sub-tasks, and workers that perform the task, correspond to the two types of entities that compose the bipartite graph. The authority score is used to rank the performance of workers while the hub score is used to rank the effectiveness of supervisors to assign the right task to the right workers. Aydin *et al.* [2] tried to find the right answers to multiple-choice questions that had been aggregated from the crowd for the game "Who wants to be a millionaire?". They created a big bipartite graph composed by multiple choice answers, assumed as authorities, and users, assumed as hubs. The computed hub scores, through the HITS algorithm, of the users were used as weights in a weighted voting scheme that predicts the right answer of a multiple choice question. The authors claimed a significantly increased accuracy of right prediction on the harder questions that are posed at the end of the game while the overall accuracy of prediction reaches 95%.

The structure of tuples $\{$user, item, tags$\}$ in tagging systems has been termed folksonomy, being composed of folk, i.e., the users of the tagging system, and a taxonomy, i.e., a hierarchy is built from an "is-a" relationship. Traditional ranking algorithms such as the PageRank and HITS were proposed for ranking folksonomies [16]. However, the fact that folksonomies are composed from three different types of entities, and, therefore, can be only modelled as tripartite graphs, makes the direct application of those algorithms for ranking folksonomies problematic. As a result several modifications of the original PageRank and HITS algorithms were proposed. The FolkRank [17] is one of the algorithms that are based on the PageRank algorithm while a modification, called *differential FolkRank*, appropriate for ranking folksonomies that are modeled as uni-directed tripartite graphs was also proposed by the same authors [18]. We further discuss this algorithm in Section III-D.

We have seen in the previous paragraphs that the HITS algorithm has been successfully applied in real-world problems that can be modeled through bipartite graphs. At the same time crowdsourced image annotation is gaining popularity through the wide use of dedicated crowdsourcing platforms. However, the problem of crowdsourced image tagging has never been modeled as a two-mode network probably because it involves three different types of entities: annotators, images and tags. We overcome the three entities 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-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 is used to filter out the relevant from the irrelevant hashtags.

There are different approaches in tag filtering including Xia *et al.* [46] they propose a bi-layer clustering framework to locate relevant tags to social images images. In the first layer they try to locate relevant tags and images. In the second layer the image groups are divided into smaller using Affinity Propagation. Then they calculate the frequency of tags and relevance to keep only the relevant tags. Wang [45] *et al.* inspired by topic model and deep learning they propose a novel method called regularized latent Dirichlet allocation to filters tags. In the deep learning model they use four layers combining tags and image features. Argyrou [1] *et al.* in their research they used Latent Dirichlet Allocation (LDA) model to retrieve the relevant Instagram hashtags that are related to the content of the image and can be used for Automatic Image Annotation. Based on hashtags from a sample of 1000 Instagram the researchers trained an LDA model.

III. Materials and Methods

In this section we present the problem, and describe the methodology we follow to solve it along with the main concepts formulated within this methodology, and we explain the data we used in our experiments along with the data collection procedure.

A. Problem formulation

Let us assume an Instagram image I_j and the set $\mathcal{T}^j$=$\{t_1^j, t_2^j, \ldots, t_k^j \ldots, t_{K_j}^j\}$ of K_j hashtags that accompany it (see Figure 1 for an example). We denote by r_k^j the relevance of hashtag t_k^j with the visual content of image I_j. We assume that the relevance scores $R[t_k^j]$, $k = 1, 2, \ldots, K_j$, $j = 1, 2, \ldots, M$ are computed with the aid of a crowd of N annotators (crowdtaggers) as explained in Section III-E.

The aim of this study is to create a ranked set of tags for each one of the Instagram images I_j in terms of their relevance with its visual content, such as:

$$\mathcal{T}_r^j = \{t_{r,1}^j, t_{r,2}^j, \ldots, t_{r,k}^j \ldots, t_{r,k+1}^j \ldots, t_{r,K_j}^j\} \quad (1)$$

where $R[t_{r,k}^j] > R[t_{r,k+1}^j]$

B. Methodology

We assume that a set $\mathcal{I} = \{I_1, I_2, \ldots, I_M\}$ of M Instagram images along with their associated hashtags $\mathcal{T} = \{\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^j, \ldots, \mathcal{T}^M\}$ crawled according to the procedure described in Section III-E. The methodology we follow to solve the problem mentioned in the previous section consists of the following steps. For the convenience of the readers who are interested to re-run the process detailed Python code is given in Appendix.

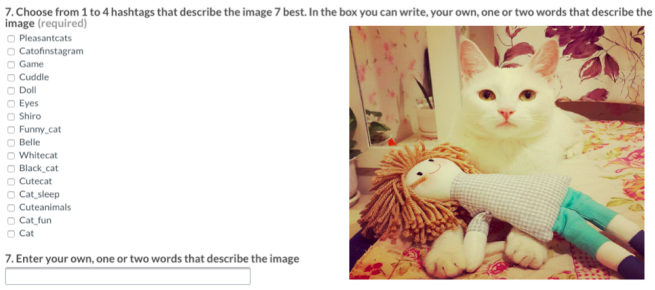

Fig. 2. An example of hashtag selection process that took place via *Figure-eight*

- *Step 1*: The relevance $R[t_k^j]$, $k = 1, 2, \ldots, K_j$ of each hashtag with respect to the visual content of the associated image I_j is assessed by a set $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N users (annotators) with the aid of a crowdsourcing platform as it can be seen in Figure 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:

$$\begin{aligned} \mathcal{B} &= \{\mathcal{V}, \mathcal{E}\} \\ \mathcal{V} &= \mathcal{V}_U \bigcup \mathcal{V}_T \\ \mathcal{V}_U &\bigcap \mathcal{V}_T = \emptyset \end{aligned} \quad (2)$$

  where $\mathcal{V}_U$ and $\mathcal{V}_T$ are the sets of vertices corresponding to the annotators and hashtags, respectively, while $\mathcal{E} = \{e_{ik}^j\}$ is the set of edges denoting that the i-th user selected (considered as visually relevant) the tag t_k^j of image I_j.
- *Step 3*: The effectiveness (reliability) of annotators is approximated with the set of hub values $\mathcal{H} = \{h[v_1], h[v_2], \ldots, h[v_i], \ldots, h[v_N]\}$, where $h[v_i]$ is the hub value of vertex $v_i \in \mathcal{V}_U$, computed with the aid of the HITS algorithm (see also Section III-C).
- *Step 4*: For each image I_j we construct a weighted bipartite graph as follows:

$$\begin{aligned} \mathcal{B}^j &= \{\mathcal{V}^j, \mathcal{E}^j\} \\ \mathcal{V}^j &= \mathcal{V}_U \bigcup \mathcal{V}_T^j \\ \mathcal{V}_U &\bigcap \mathcal{V}_T^j = \emptyset \\ \mathcal{E}^j &= \{(v_i, v_k, h[v_i]) | v_i \in \mathcal{V}_U, v_k \in \mathcal{V}_T^j, h[v_i] \in \mathcal{H}\} \end{aligned} \quad (3)$$

  where $\mathcal{V}_U$ is the set of vertices corresponding to the annotators, $\mathcal{V}_T^j$ is the set of vertices corresponding to the hashtags of the j-th image and $\mathcal{E}^j$ is the set of weighted edges denoting that the i-th-user selected (considered as visually relevant) the tag t_k^j of image I_j.
  In Figure 3 it is shown, for better visualization, the k-core[1] (k=6) of the bipartite graph corresponding to image 7 (the one shown in Figure 2). The radius of each tag is analogous to the weighted degree of the corresponding vertex. The whole bipartite graph for image 7 consists of

[1]https://networkx.github.io/documentation/stable/reference/algorithms/core.html

607 vertices: 499 annotators (users), the 16 hashtags of image 7 and another 92 tags suggested by the annotators.

- *Step 5*: A ranked set of tags, $\mathcal{T}_r^j$=$\{t_{r,1}^j, t_{r,2}^j, \ldots, t_{r,k}^j, t_{r,k+1}^j \ldots, t_{r,K_j}^j\}$, for each Instagram image I_j is achieved through the set of authority values $\mathcal{A}^j = \{a^j[v_1], a^j[v_2], \ldots, a^j[v_k], a^j[v_{k+1}], \ldots, a^j[v_{K_j}]\}$, where $a^j[v_k]$ is the authority value of vertex $v_k \in \mathcal{V}_T^j$, computed with the aid of the HITS algorithm when it is applied on the weighted bipartite graphs that were created in the previous step.

Table I shows the authority values for the hashtags associated with image 7 along with the hub values of the 16 most reliable annotators (for this specific image) after the application of the proposed methodology.

TABLE I
AUTHORITY AND HUB VALUES FOR THE BIPARTITE NETWORK OF IMAGE #7 (SEE ALSO FIG. 3) - ONLY THE 16 MOST RELIABLE ANNOTATORS ARE SHOWN

Hashtag	**Authority**	**Annotator ID**	**Hub** ($x10^{-2}$)
cat	0.2027	3376020988	0.5582
doll	0.1314	3374149591	0.5163
white	0.1264	3374415489	0.4872
cute	0.1171	3374112507	0.4806
animal	0.0635	3374477746	0.4680
funny	0.0621	3376833191	0.4563
eyes	0.0471	3375771052	0.4556
instagram	0.0434	3375856453	0.4513
fun	0.0389	3374757569	0.4489
game	0.0279	3374777892	0.4256
pleasant	0.0267	3374647452	0.4037
cuddle	0.0256	3374505202	0.4029
belle	0.0092	3376453894	0.3996
shiro	0.0077	3374248101	0.3981
sleep	0.0060	3375852267	0.3976
black	0.0040	3374781743	0.3964

C. The HITS Algorithm in bipartite and weighted bipartite graphs

The HITS (Hyperlink-Induced Topic Search) algorithm was initially introduced by Kleinberg [19, 20] in order to analyze a collection of web-pages, relevant to a topic, and locate the most "authoritative" ones in that topic. It performs link analysis on those web pages 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. Thus, a web-page that has many inlinks from other pages with high hub value is considered an authority while a page with many outlinks to high authority web-pages is a hub [29, 41]. In simple words, the main principle of the HITS algorithm is that an informed hub points to many effective authorities and an effective authority is pointed out by many informed hubs. Thus, authorities and hubs have a mutual reinforcement relationship [9].

As already discussed in the Introduction, 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 centrality[2], 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. The authority $a[v]$ and hub value $h[v]$ of a vertex v in a bipartite graph are (iteratively) computed with the aid of the following equations:

$$a[v] = \sum_{v_i \in \mathcal{N}_{v,U}} h[v_i] \qquad \mathcal{N}_{v,U} = \{v_i | v_i \in \mathcal{V}_U,\ (v_i, v) \in \mathcal{E}\} \tag{4}$$

$$h[v] = \sum_{v_i \in \mathcal{N}_{v,T}} a[v_i] \qquad \mathcal{N}_{v,T} = \{v_i | v_i \in \mathcal{V}_T,\ (v, v_i) \in \mathcal{E}\} \tag{5}$$

where $\mathcal{N}_{v,U}$ is the set of vertices in $\mathcal{V}_U$ that point to vertex v and $\mathcal{N}_{v,T}$ is the set of vertices in $\mathcal{V}_T$ that vertex v points to (see also eq. 2).

It can be seen in eq 4 and 5 that a vertex's authority value is the sum of the hub score of all vertices pointing to it while its hub value is the sum of authority sores of all vertices that it points to. The final hub-authority values of a vertex are determined after infinite repetitions of the algorithm but in practice typical convergence tests, based on the number of iterations or the change of hub - authority scores between consecutive iterations, are applied. Given that directly and iteratively applying the above equations leads to diverging values, it is necessary to normalize hub and authority values after every iteration so as to sum to 1, i.e., $\sum_v h[v] = 1$, $\sum_v a[v] = 1$. By definition the initial values of $a[p]$ and $h[p]$ are set to 1.

For weighted undirected bipartite graphs $\mathcal{B}^j$, such as those corresponding to a user-tag bipartite network for a specific image I_j (see eq. 3), the equations of the HITS algorithm are modified as follows:

$$a^j[v] = \sum_{v_i \in \mathcal{N}_{v,U}^j} h[v_i] \cdot h^j[v_i] \qquad \mathcal{N}_{v,U}^j = \{v_i | v_i \in \mathcal{V}_U,\ (v_i, v, h[v_i]) \in \mathcal{E}^j\} \tag{6}$$

$$h^j[v] = \sum_{v_i \in \mathcal{N}_{v,T}^j} h[v_i] \cdot a^j[v_i] \qquad \mathcal{N}_{v,T}^j = \{v_i | v_i \in \mathcal{V}_T^j,\ (v, v_i, h[v_i]) \in \mathcal{E}^j\} \tag{7}$$

where $\mathcal{N}_{v,U}^j$ is the set of vertices in $\mathcal{V}_U$ that point to vertex v and $\mathcal{N}_{v,T}^j$ is the set of vertices in $\mathcal{V}_T^j$ that vertex v points to (see also eq. 3).

D. Folksonomies and the FolkRank algorithm

While our approach is a modification of the HITS algorithm to handle {user, images, hashtags} folksonomies, the FolkRank [17] is a known modification of the PageRank

[2]https://en.wikipedia.org/wiki/Centrality

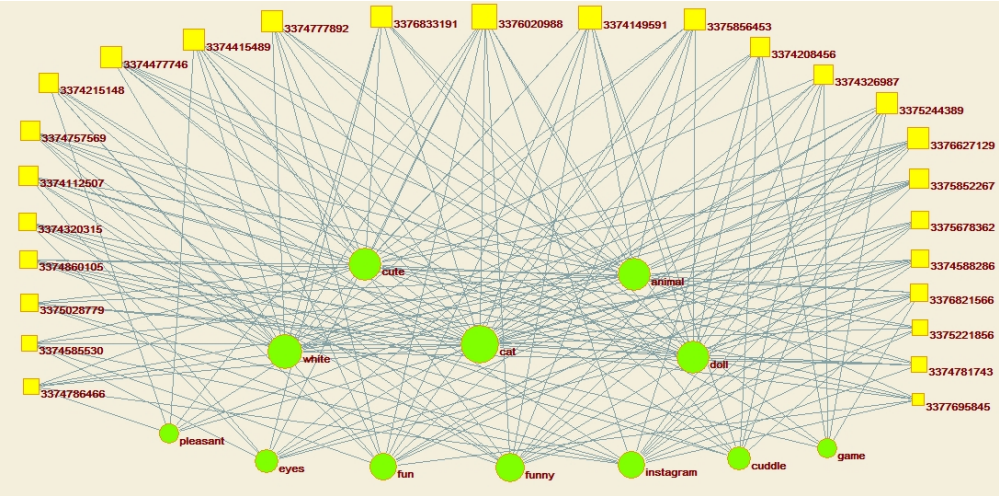

Fig. 3. A subgraph of user-tag bipartite network for image #7. Circles show the tags while the boxes show the annotators that selected those tags.

algorithm towards this direction. FolkRank makes use of the personalization component of the PageRank algorithm and applies single entity optimization. By doing so, Folk rank is capable of handling the inherent difficulty to adapt a single entity ranking algorithm (PageRank) to a three entity structure (folksonomy). An additional difficulty comes from the fact folksonomies are usually modelled as uni-directed graphs, i.e., humans select tags for an item. In order to handle this problem Hotho *et al.* [18] proposed a modified version of the FolkRank algoritm, called *differential FolkRank*. It is this algorithm that is used for comparison with the proposed method in the next section.

E. Data collection, crowdtagging and software tools

A set of 50 Instagram images, along with their hashtags, were automatically crawled with the aid of a Python[3] program (see [12] for more details on the crawling process). The collected Instagram images were uploaded to *Figure-eight* for crowdtagging in the form of tag selection as indicated in Figure 3 for image #7. To simplify the process all hashtag choices were presented to the annotators as checkboxes. The annotators were invited to select 1-4 hashtags and were given also the opportunity to provide their own tags. Despite these guidelines many annotators select much more than 4 tags and in several cases the extra tags they provided were already among the given choices. Therefore, 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 *wordsegment*[4] Python library. For instance, the hashtag *#picoftheday* is decomposed into the words *pic*, *of*, *the* and *day*.

Every image was annotated by 500 annotators for experimentation purposes. In practice much fewer annotations per image are enough while there is absolutely no reason that all images must be assessed by all annotators. Nevertheless, we made those choices to allow us generalize the conclusions of our study as much possible. One of the annotators turned out to be dishonest as indicated by the *_trust* value of *Figure-eight* as well as by the corresponding hub value of the HITS algorithm when it was applied on the full bipartite graph (eq. 2), and she / he was excluded from the experiments. Comparison between hub values and *trust* scores are given in Section IV-A. The full bipartite graph and the bipartite graphs per image were constructed and analyzed with help of the NetworkX[5] library of Python. We also used the NetworkX implementation of the HITS algorithm to extract the overall hub values (reliability scores for the annotators) and authority scores of the tags of each image.

F. Evaluation framework

The 50-Instagram-Image questionnaire was given to the *Figure-eight* annotators. Additionally, two image retrieval experts have acess to the same data set. The annotations of the experts, aggregated together and pre-processed in the same way as the crowdsourced data, consist our gold standard across which the effectiveness of the proposed methodology is evaluated through the measures defined below. In total 145 different tags were proposed by the experts for the 50 images. On the other hand, the 499 annotators proposed a total of 2571 different tags. However, only 135 of the tags proposed by the experts were also proposed by the annotators.

Let us denote with $\mathcal{G} = \{\mathcal{G}^1, \mathcal{G}^2, \ldots, \mathcal{G}^M\}$ the set of hashtags in the gold standard set, where $\mathcal{G}^j$ is the gold standard set for the j-th image. Let us also denote with $\mathcal{T}^j_{r,\theta} = \{t^j_{r,1}, t^j_{r,2}, \ldots, t^j_{r,k}\}$ the ordered set of tags for image I_j such that $a^j[t^j_{r,1}] \geq a^j[t^j_{r,2}] \geq \ldots \geq a^j[t^j_{r,m}] \ldots \geq a^j[t^j_{r,k}]$ and $a^j[t^j_{r,k}] > \theta$, where $a^j[t^j_{r,m}]$ is the authority value of the vertex of bipartite graph $\mathcal{B}^j$ corresponding to the tag $t^j_{r,m}$.

The recall value $R_{j,\theta}$ for image I_j at the authority threshold value θ, i.e., the portion of tags in the gold standard set that

[3]https://www.python.org/

[4]http://www.grantjenks.com/docs/wordsegment/

[5]https://networkx.github.io/

were identified by the HITS algorithm when only the annotator tags with authority score higher than θ were kept, is given by:

$$R_{j,\theta} = \frac{||\mathcal{T}^j_{r,\theta} \cap \mathcal{G}^j||}{||\mathcal{G}^j||} \tag{8}$$

where $\cap$ denotes the set intersection operation and $||\Omega||$ refers to the cardinality of set Ω.

In a similar manner we define the precision value $P_{j,\theta}$ for image I_j at the authority threshold value θ, as the portion of the tags that were identified by the HITS algorithm that are included in the gold standard set of image I_j:

$$P_{j,\theta} = \frac{||\mathcal{T}^j_{r,\theta} \cap \mathcal{G}^j||}{||\mathcal{T}^j_{r,\theta}||} \tag{9}$$

With the aid of eq. 8 and 9 we can compute the Recall, Precision and F_1-measure, at the authority threshold value θ, for the whole image dataset as follows:

$$R_\theta = \frac{1}{M}\sum_{j=1}^{M} R_{j,\theta} \tag{10}$$

$$P_\theta = \frac{1}{M}\sum_{j=1}^{M} P_{j,\theta} \tag{11}$$

$$F_{1,\theta} = \frac{2 \cdot P_\theta \cdot R_\theta}{P_\theta + R_\theta} \tag{12}$$

The effectiveness of the proposed method is also evaluated with the aid of Mean Reciprocal Rank (MRR) [5]. The MRR of an image I_j is computed as follows:

$$MRR_j = \frac{1}{||\mathcal{T}^j_r \cap \mathcal{G}^j||} \sum_{i=1, t^j_{r,i} \in \mathcal{G}^j}^{K_j} \frac{1}{r^j_i} \tag{13}$$

where $\mathcal{T}^j_r$=$\{t^j_{r,1}, t^j_{r,2}, \ldots, t^j_{r,K_j}\}$ is the ordered set of tags for image I_j, $\mathcal{G}^j$ is the corresponding gold standard set, and r^j_i is the ranking of tag $t^j_{r,i}$.

The MRR is computed as the average of MRR_j across all images.

Another key performance metric in information retrieval is Mean Average Precision (MAP). The purpose of MAP is to calculate the average of the precision value of the top set of k results. It is defined as follows:

$$MAP_j = \frac{1}{||\mathcal{T}^j_r \cap \mathcal{G}^j||} \sum_{k=1}^{K_j} \frac{||\mathcal{T}^j_{r,k} \cap \mathcal{G}^j||}{||\mathcal{T}^j_{r,k}||} \tag{14}$$

where $\mathcal{T}^j_{r,k}$=$\{t^j_{r,1}, t^j_{r,2}, \ldots, t^j_{r,k}\}$ is the ordered set of the k first tags of image I_j.

A practical example on how the MAP and MRR scores are computed is shown in Table V for the particular case of Image #6.

IV. Results

The Precision, Recall and F_1 measure, as defined in eq. 10-12, were computed for a variety of authority threshold values θ and are presented in Table II. Moreover, we present the Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) results according to the eq. 13-14, in Table IV. The corresponding Receiver Operating Characteristic curves[6] (ROC) are shown in Figure 4. For convenient juxtaposition with the values presented in Table II, in this ROC curve it is plotted the Precision versus Recall instead of the typical case of ROC curves in which are usually plotted the True Positive Rate versus the False Positive Rate. We observe from both Table II and Figure 4 that the best results in terms of the F_1 measure is obtained for an authority score threshold value θ=0.11. However, as in most information retrieval systems we usually prefer a higher value of Recall, that is identifying more tags even if they are not that accurate, instead of Precision. Thus, an authority score threshold θ=0.09 give us also a reasonable choice.

With a MAP score equal to 0.891 (see Table IV) we can conclude that applying the HITS algorithm for the selection of the appropriate hashtags, for Instagram images, in a crowdsourcing environment is, at least promising. Since, MAP ranges [0,1] and the result is close to 1, we can conclude that the algorithm located almost all the relevant hashtags of the collection. Another indication that the proposed methodology is suitable for locating relevant hashtags is the MRR results (see also Table IV). Values for MRR range from 0 to 1, with higher values signify that the relevant hashtags are ranked higher. Thus, MRR=0.5 corresponds to the correct hashtags being in the top two returned by the HITS algorithm.

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 random information retrieval system (or, similarly, a uninformative two-class classifier) while an AUC equal to 1 represents the perfect information retrieval system. The AUC corresponding to the ROC curve of Figure 4 is equal to 0.692. As we show in the Appendix (*Step 6*) the computation was done with the aid of the *metrics*[7] Python library of *Sklearn*[8].

In our previous study [12] we concluded that on average four of the hashtags accompanying each Instagram image are related to its visual content. This conclusion was inline with the findings of Ferrara *et al.* [8] who studied users' behavior while they annotate their photos with hashtags and concluded that users use quite a few hashtags in order to annotate image

[6]https://en.wikipedia.org/wiki/Receiver_operating_characteristic
[7]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html
[8]https://scikit-learn.org/stable/

TABLE II
RECALL, PRECISION AND F_1-MEASURE SCORES FOR M=50 IMAGES AND VARIOUS THRESHOLD VALUES W.R.T. AUTHORITY SCORE (HITS), *_trust* WEIGHTING AND FOLKRANK RANKING SCORE

		Authority threshold value θ/ FolkRank ranking score threshold value										
Algorithm	(M=50, N=499)	0.25	0.21	0.17	0.15	0.13	0.11	0.09	0.07	0.05	0.03	0.01
HITS	Recall (R)	0.136	0.223	0.359	0.440	0.527	0.620	0.679	0.712	0.766	0.804	0.842
AUC = 0.692	Precision (P)	0.962	0.932	0.904	0.862	0.822	0.755	0.654	0.604	0.504	0.396	0.265
	F_1-measure (F)	0.238	0.360	0.514	0.583	0.642	**0.681**	**0.667**	0.653	0.608	0.530	0.403
FolkRank	Recall (R)	0.158	0.261	0.370	0.424	0.504	0.603	0.663	0.707	0.755	0.804	0.832
AUC = 0.689	Precision (P)	0.935	0.923	0.895	0.876	0.823	0.766	0.709	0.613	0.529	0.418	0.277
	F_1-measure (F)	0.270	0.407	0.523	0.571	0.626	**0.675**	**0.685**	0.657	0.622	0.550	0.415
_trust	Recall (R)	0.168	0.272	0.353	0.424	0.527	0.609	0.652	0.696	0.739	0.798	0.856
AUC = 0.680	Precision (P)	0.929	0.903	0.877	0.847	0.813	0.772	0.698	0.601	0.517	0.412	0.267
	F_1-measure (F)	0.286	0.418	0.504	0.565	0.640	**0.681**	**0.674**	0.645	0.609	0.543	0.407

TABLE III
RECALL, PRECISION AND F_1-MEASURE SCORES FOR M=50 IMAGES AND VARIOUS VALUES OF THE TOP RANKED HASHTAGS BASED ON THE AUTHORITY SCORE

	Number of mined hashtags kept (k)										
(M=50, N=499)	1	2	3	4	5	6	7	8	9	10	11
Recall (R)	0.234	0.467	0.603	0.685	0.750	0.772	0.808	0.815	0.837	0.842	0.848
Precision (P)	0.862	0.858	0.740	0.630	0.552	0.473	0.426	0.375	0.342	0.310	0.284
F_1-measure (F)	0.368	0.605	**0.665**	**0.656**	0.636	0.587	0.558	0.514	0.486	0.453	0.425

TABLE IV
MEAN AVERAGE PRECISION AND MEAN RECIPROCAL RANK FOR FOR M=50 IMAGES

	Mean	Min	Max
Average Precision	0.89	0.51	1.00
Reciprocal Rank	0.52	0.16	1.00

TABLE V
AVERAGE PRECISION AND MEAN RECIPROCAL RANK FOR IMAGE #6 HASHTAGS ACCORDING TO AUTHORITY SCORE RANK (ASR)

Hashtag	In Gold Standard	ASR	Precision	RR
vacation		1	0	
beach	x	2	1/2 (0.500)	1/2 (0.500)
sand	x	3	2/3 (0.667)	1/3 (0.333)
sun		4	0	0
bikini	x	5	3/5 (0.600)	1/5 (0.200)
sea	x			
sky	x			
woman	x			
hat	x			
Sum			1.767	1.033
Average			0.589	0.344

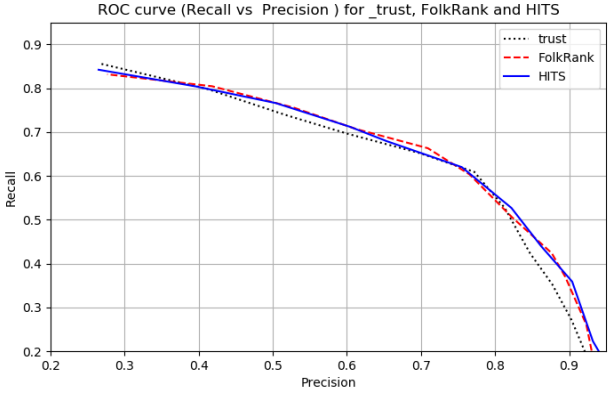

Fig. 4. Recall vs precision ROC curves for the *_trust* (AUC = 0.680), the FolkRank (AUC = 0.689) and the HITS (AUC = 0.692) weighting schemes

content. In order to verify these findings we also evaluated, again with the aid of the gold standard set, the effectiveness of hashtags' selection through the HITS algorithm by keeping the k top ranked hashtags per image based on their authority scores. The results, for a variety of k values, are shown in Table III while the corresponding ROC curve is shown in Figure 5. We see that the best F_1 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. We see also in Figure 5 that the area under the curve (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; so keeping the k top ranked hashtags based on the authority score is another option for mining tags from Instagram hashtags accompanying images.

A. Reliability measures for the annotators

Figure-eight, as many other crowdsourcing platforms, provides its own measure to identify dishonest annotators. In particular it uses the *_trust* variable which is computed on a subset of the data, known as Gold Test Questions, for which the creators provide the correct answers and which is considered as a type of gold standard. In our case, an additional

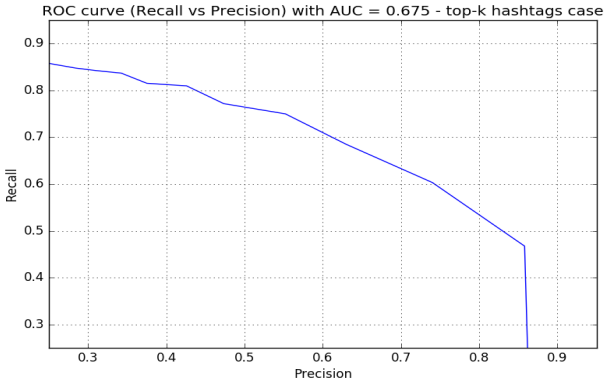

Fig. 5. Recall vs precision ROC curve with an area under the curve (AUC) equal to 0.675 - the case of top-k hashtags

set of Instagram images corresponding to 10% of the data was assessed (crowdtagged) by the creators. The performance of each one of the annotators is the recall value of the tags used by the creators that the annotator correctly identified.

As already mentioned, in the proposed method the reliability of the annotators is estimated with the aid of the hub value computed on the full graph composed from all images and all tags (see eq. 5). So the annotators reliability is based on the total number of image for hub value on contrast to the calculated for all the _trust value that is based on 10% of the data. In Table VI we present the hub values of the top 10 reliable annotators based on our method along with the corresponding *_trust* value as computed by *Figure-eight*. In the same table we show also, the corresponding ranking of the differential FolkRank algorithm. While the rankings of annotators based on the hub scores and the FolkRank algorithm are identical, as they both based on the same principle, we observe large differences between them and the *_trust* values (fifth column) of *Figure-eight*. In fact the *_trust* values, of the top 10 annotators based on the hub scores and FolkRank, are below the average *_trust* value (0.7675) and in almost all cases the corresponding ranking is in the last 100. We remind here that the total number of annotators is N=499.

TABLE VI
TOP 10 USERS ACCORDING TO THE HUB VALUE ALONG WITH THEIR CORRESPONDING RANKING BASED ON *Figure-eight*'S *_trust* VALUE

User ID	**hub value** $x10^{-2}$	**hub based ranking**	**FolkRank value** $x10^{-2}$	**FolkRank ranking**	***_trust* value**	***_trust* based ranking**
xx7892	0.3195	1	0.1444	1	0.6665	490
xx5795	0.3060	2	0.1372	2	0.7104	462
xx7746	0.3045	3	0.1363	3	0.6688	487
xx9591	0.3020	4	0.1350	4	0.6504	496
xx8610	0.2964	5	0.1320	5	0.7308	419
xx3452	0.2939	6	0.1306	6	0.6547	493
xx0988	0.2931	7	0.1302	7	0.6351	497
xx1052	0.2912	8	0.1291	8	0.7306	422
xx8286	0.2909	9	0.1290	9	0.7367	404
xx2687	0.2888	10	0.1278	10	0.7402	389

We observe also, by examining the extreme values of hub and *_trust*, that the hub scores provide a more subtle diversification than the *_trust* scores. Therefore, our choice to weight the bipartite graphs for each image (see eq. 6) with the hub scores of the full bipartite graph rather than the *_trust* values seems justified. However, in order to empirically check this assumption we repeated our experiments by using as weights in the bipartite graphs for each image the *_trust* scores of the annotators. The results are summarized in Table II and illustrated in Figure 4. We see a quite similar performance in terms of the F_1 metric although some differentiation between Recall and Precision for the same values of the authority threshold θ do exist. The area under the curve achieved when using the *_trust* scores to weight the bipartite graphs is 0.680, not very much lower than that of the hub score weighting of the bipartite graphs. We further discuss this finding in Section V.

V. DISCUSSION & 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. 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 study 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 with 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.

A final remark of the current study refers to the importance of using weighted user-tag bipartite graphs for the crowdtagged images. It appears that weighting the bipartite graphs with the hub scores of the annotators provides the best results. However, even in the case that the reliability metric of the crowdsourcing platform itself (the *_trust* variable of *Figure-eight* in our case) is used to weight the bipartite graphs the results are not significantly worse. We are a little bit reluctant to generalize this conclusion because in the current study we have used too many annotations (499) per image. Thus, one of our future tests will involve a more typical image crowdtagging scenario in which much more images will be used and much fewer (typically less than five) annotations per image will be considered. In that case only partial co-annotation of the same images by the same annotators will take place in contrary to the current study where all annotators annotated all images.

We are currently working to check in practice that the image - hashtags pairs mined from the Instagram through the approach described in this paper can be used, indeed, for a large scale Automatic Image Annotation in a content-based image retrieval scenario as proposed by Theodosiou and Tsapatsoulis [37].

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, "Projection-based link prediction in a bipartite network," *Information Sciences*, vol. 376, pp. 158–171, 2017.

[11] S. I. Gass and C. M. Harris, "Bipartite Graph," in *Encyclopedia of operations research and management science*, Boston: Springer, 2013, pp. 126.

[12] S. Giannoulakis and N. Tsapatsoulis, "Evaluating the descriptive power of instagram hashtags," *Journal of Innovation in Digital Ecosystems*, vol. 3, no. 2, pp. 114–129, 2016.

[13] S. Giannoulakis and N. Tsapatsoulis, "Defining and identifying stophashtags in instagram," in *Proc. INNS Conference on Big Data*, 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 Intl. Conf. on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/-DataCom/CyberSciTech)*, IEEE, 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.

[16] M. Gupta, R. Li, Z. Yin, and J. Han. 2010. Survey on social tagging techniques. *SIGKDD Explor. Newsl.* 12, 1, November 2010, pp. 58-72.

[17] A. Hotho, R. Jäschke, C. Schmitz and G. Stumme, "FolkRank: A Ranking Algorithm for Folksonomies," in *Proc. of the 12th Workshop on Knowledge Discovery, Data Mining, and Machine Learning*, 2006, pp. 111–114

[18] A. Hotho, R. Jäschke, C. Schmitz and G. Stumme, Trend Detection in Folksonomies. In *Semantic Multimedia (SAMT 2006). Lecture Notes in Computer Science*, vol 4306. Springer, Berlin, Heidelberg, 2006, pp. 56–70

[19] J. M. Kleinberg, "Citizen crowds and experts: observer variability in image-based plant phenotyping,"Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[20] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins "The web as a graph: measurements, models, and methods," in *Proc. International Computing and Combinatorics Conference*, Springer, 1999, pp. 1–17.

[21] H. Kwaśnicka and M. Paradowski, "Machine learning methods in automatic image annotation,"in *Advances in Machine Learning II*, Springer, 2010, pp. 387–411.

[22] A. London and T. Csendes, "HITS based network algorithm for evaluating the professional skills of wine tasters," in *Proc. of the 8th IEEE International Symposium on Applied Computational Intelligence and Informatics*, IEEE, 2013, pp. 197–200.

[23] B. Luo, X. Wang, and X. Tang, "World wide web based image search engine using text and image content features,"in *Internet Imaging IV*, vol. 5018. International Society for Optics and Photonics, 2003, pp. 123–131.

[24] C. Manning, P. Raghavan and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2009.

[25] L. Maier-Hein, S. Mersmann, D. Kondermann, S. Bodenstedt, A. Sanchez, C. Stock, H. G. Kenngott, M. Eisenmann, and S. Speidel, "Can masses of non-experts train highly accurate image classifiers?" in *Proc. International conference on medical image computing and computer-assisted intervention*, Springer, 2014, pp. 438–445.

[26] J. Mao, K. Lu, G. Li, and M. Yi, "Profiling users with tag networks in diffusion-based personalized recommendation," *Journal of Information Science*, vol. 42, no. 5, pp. 711–722, 2016.

[27] J. C. Miller, G. Rae and F. Schaefer, "Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records," in *Proc. 24th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2001, pp. 444-445.
[28] D. Mitry, K. Zutis, B. Dhillon, T. Peto, S. Hayat, K.-T. Khaw, J. E. Morgan, W. Moncur, E. Trucco, and P. J. Foster, "The accuracy and reliability of crowdsource annotations of digital retinal images," *Translational Vision Science and Technology*, vol. 5, no. 5, pp. 6–6, 2016.
[29] I. Nagasinghe, "Computing principal eigenvectors of large web graphs: Algorithms and accelerations related to pagerank and hits," Ph.D. dissertation, Southern Methodist University, 2010.
[30] K. Ntalianis, N. Tsapatsoulis, A. Doulamis, and N. Matsatsinis, "Automatic annotation of image databases based on implicit crowdsourcing, visual concept modeling and evolution," *Multimedia Tools and Applications*, vol. 69, no. 2, pp. 397–421, 2014.
[31] K. Ntalianis, N. Tsapatsoulis, A. Doulamis, and N. Matsatsinis, "Social Relevance Feedback based on Multimedia Content Power,"*IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 109-117, 2018.
[32] A. Preece, I. Spasic, K. Evans, D. Rogers, W. Webberley, C. Roberts, and M. Innes, "Sentinel: A Codesigned Platform for Semantic Enrichment of Social Media Streams," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 118-131, 2018.
[33] Y. E. Sagduyu, A. Grushin and Y. Shi, "Synthetic Social Media Data Generation " *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 605-620, 2018.
[34] D. Schall, B. Satzger, and H. Psaier, "Crowdsourcing tasks to social networks in BPEL4People," *World Wide Web*, vol. 17, no. 1, pp. 1–32, 2014.
[35] C. G. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2009.
[36] T. Sunahase, Y. Baba, and H. Kashima, "Pairwise HITS: Quality estimation from pairwise comparisons in creator - evaluator crowdsourcing process," in *Proc. 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 977–984.
[37] Z. Theodosiou and N. Tsapatsoulis, "Image retrieval using keywords: the machine learning perspective," in *Semantic Multimedia Analysis and Processing*, pp. 3–30, 2014.
[38] Z. Theodosiou and N. Tsapatsoulis "Crowdsourcing annotation: Modelling keywords using low level features," in *Proc.of the 5th IEEE International Conference Internet Multimedia Systems Architecture and Application*, IEEE, 2011, pp. 1–4.
[39] Z. Theodosiou, O. Georgiou, and N. Tsapatsoulis "Evaluating annotators consistency with the aid of an innovative database schema," in *Proc. 6th International Workshop on Semantic Media Adaptation and Personalization*, IEEE, 2011, pp. 74–78.
[40] N. T. Tri and J. J. Jung, "Exploiting geotagged resources to spatial ranking by extending hits algorithm," *Computer Science and Information Systems*, vol. 12, no. 1, pp. 185–201, 2015.
[41] P. Tsaparas, "Link analysis ranking," Ph.D. dissertation, University of Toronto, 2004.
[42] N. Tsapatsoulis, "Web image indexing using wice and a learning-free language model," in *Proc. IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, 2016, pp. 131–140.
[43] N. Tsapatsoulis and O. Georgiou, "Investigating the scalability of algorithms, the role of similarity metric and the list of suggested items construction scheme in recommender systems," *International Journal on Artificial Intelligence Tools*, vol. 21, no. 04, p. 1240018, 2012.
[44] V. S. Tseng, J.-C. Ying, C.-W. Huang, Y. Kao, and K.-T. Chen, "Fraudetector: A graph-mining-based framework for fraudulent phone call detection," in *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2015, pp. 2157–2166.
[45] J. Wang, J. Zhou, H. Xu, T. Mei, X.-S. Hua and S. Li "Image tag refinement by regularized latent Dirichlet allocation," *International Journal on Artificial Intelligence Tools*, vol. 124, p. 61-70, 2014.
[46] Z. Xia, X. Feng, J. Peng and J. Fan "Content-irrelevant tag cleansing via bi-layer clustering and peer cooperation," *Journal of Signal Processing Systems*, vol. 81, no. 1, p. 29-44, 2015.
[47] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and ranking product features in opinion documents," in *Proc. 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 1462–1470.

APPENDIX A
PYTHON CODE

Here we provide the full Python code that allows anyone who wishes to re-run the experiments and test their validity. The graphs as Pajek[9] files are also publicly available at https://irci.eu/insta-hashtags/

- *Step 1*: Read the datafile produced through crowdsourcing (already converted to *json*[10] format)

```
>>> import json
>>> with open('../data/F8_data.json', 'r') as fp:
...   data = json.load(fp)
>>> users = list(data.keys())
>>> data[users[0]].keys()
```

- *Step 2*: Create a full bipartite graph composed by annotators and all available tags in order to rank the annotators.

```
>>> import networkx as nx
>>> import numpy as np
>>> exec(open('csv2imageGraphs.py').read())
>>> G = FullGraph(data,50,no_split,
      '../data/full499.net')
```

- *Step 3*: Apply the HITS algorithm and get the hub values (h).

```
>>> [h,a] = nx.hits(G)
```

- *Step 4*: Use the hub values (h) computed in the previous step to initialize the bipartite graphs for each one of the images.

```
>>> ImageGraphs(data,50,h,no_split)
>>> G7 = nx.read_pajek('../data/img7.net')
>>> [annotators,tags] = nx.bipartite.sets(G7)
>>> list(sorted(tags))[:9]
['acosta', 'amigo', 'amores', 'and', 'animal',
    'animales', 'baby', 'bau', 'beautiful']
>>> list(sorted(annotators))[:5]
['3374092858', '3374094788', '3374097114', '3374098976',
    '3374107231']
>>> G7['3374092858']
{'cat': {'weight': 0.1629}, 'doll': {'weight': 0.1629},
    'white': {'weight': 0.1629}}
>>> G7['3374098976']
{'cat': {'weight': 0.1248}}
```

- *Step 5*: For each image graph apply the HITS algorithm to rank the tags according to the computed authority value (a).

```
>>> import operator
>>> G7 = nx.DiGraph(G7)
```

[9]http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[10]https://www.json.org/

```
>>> [h7, a7] = nx.hits(G7)
>>> sorted_a7 = sorted(a7.items(),
      key=operator.itemgetter(1), reverse=True)
>>> sorted_a7[:4]
[('cat', 0.2030), ('doll', 0.1318), ('white', 0.1268),
      ('cute', 0.1171)]
```

- *Step 6*: Compute various recall and precision values for different authority score thresholds θ and plot the result.

```
>>> Thresholds = [0.25, 0.21, 0.17, 0.15, 0.13, 0.11,
      0.09, 0.07, 0.05, 0.03, 0.01]
>>> p = []; r = []
>>> for t in Thresholds:
 ...  [R, P] = computeROC('img', 'data/gold.json', 50, t)
  ...  p += [P]; r += [R]
  ...
>>> from sklearn import metrics
>>> metrics.auc(r,p)
>>> import matplotlib.pyplot as plt
>>> plt.plot(p,r)
>>> plt.axis([0.2, 0.95, 0.2, 0.95])
>>> plt.title('ROC curve (Recall vs Precision) with
      AUC = 0.692')
>>> plt.xlabel('Precision'); plt.ylabel('Recall')
>>> plt.grid(True); plt.show()
```

The proprietary Python functions that were developed and used in the experimentation (file *csv2imagGraphs.py*) are listed below:

```
import networkx as nx
import numpy as np
from wordsegment import load, segment
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import TweetTokenizer
load()

def FullGraph(data,M,no_split, file_out):
  G = nx.DiGraph()
  for j in np.arange(M):
    img = str(j+1)+'_choose'
    img1= str(j+1)+'_own'
    users = data.keys()
    for u in users:
      key_list = list(set(tknzr.tokenize(data[u][img])+
          tknzr.tokenize(data[u][img1])))
      keys = []
      for key in key_list:
        if key in no_split:
          keys +=[key]
        else:
          keyX = segment(key)
          keyX = [lemmatizer.lemmatize(w) for w in keyX if
              len(w)>2]
          keys +=keyX
      keys = sorted(list(set(keys)))
      for key in keys:
        G.add_edge(u, key)
  nx.write_pajek(G, file_out, encoding='UTF-8')
  return G

def ImageGraphs(data,M,h,no_split):
  for j in np.arange(M):
    G1 = nx.DiGraph()
    img = str(j+1)+'_choose'
    img1= str(j+1)+'_own'
    users = data.keys()
    for u in users:
      key_list = list(set(tknzr.tokenize(data[u][img])+
          tknzr.tokenize(data[u][img1])))
      key_list = [w.lower() for w in key_list]
      keys = []
      for key in key_list:
        if key in no_split:
          keys +=[key]
        else:
          keyX = segment(key)
          keyX = [lemmatizer.lemmatize(w) for w in keyX if
              len(w)>2]
          keys +=keyX
      keys = sorted(list(set(keys)))
      for key in keys:
        G1.add_edge(u, key, weight=h[u]*100)
    filename = 'img'+str(j+1)+'.net'
    nx.write_pajek(G1,filename, encoding='UTF-8')

def computeROC(filestart, goldfile, N, thresh_level):
  with open(goldfile, 'r') as fp:
    Gold = json.load(fp)
  retrieved = []; matched = []; gold = []
  tp = []; fp = []; fn = []
  for i in np.arange(N):
    filename = filestart+str(i+1)+'.net'
    gold_current = Gold[filestart+str(i+1)]
    G1 = nx.read_pajek(filename, encoding='UTF-8')
    G1 = nx.DiGraph(G1)
    [h, a] = nx.hits(G1)
    keys = [key for key in a.keys() if a[key]>thresh_level]
    tp +=[key for key in keys if key in gold_current]
    fp +=[key for key in keys if key not in gold_current]
    fn +=[key for key in gold_current if key not in keys]
    gold += gold_current
    retrieved += keys
  R = len(tp)/len(gold)
  P = len(tp)/len(retrieved)
  return R, P
```

Stamatios Giannoulakis received his BSc degree in Library Science and Information Systems in 2005 from the Technological Educational Institute of Thessaloniki, Greece. In 2007 he received his Msc in Information Science, Library Management and Organization specializing in new information technologies from Ionian University. In 2014 he received his Msc in Cultural Organisations Management from the Hellenic Open University and started his Ph.D. in the Department of Communication and Internet Studies of Cyprus University of Technology. His current research interest include image retrieval, metadata and digital libraries.

Nicolas Tsapatsoulis received his master's and Ph.D. degrees from the School of Electrical Engineering, National Technical University of Athens, Greece, in 1994 and 2000, respectively. He joined the Department of Communication and Internet Studies, Cyprus University of Technology, Athens, Greece, in 2008, where he is currently a full Professor. He has authored over 160 papers in international journals, book chapters, and conference proceedings while his citation record includes over 3500 non self-citations. His current research interests include multimedia information retrieval, web data mining, social computing, computer vision and biometrics.