Ivan Ivanov, Peter Vajda, Jong-Seok Lee, and Touradj Ebrahimi

In Tags We Trust

Trust modeling in social tagging of multimedia content

Tagging in online social networks is very popular these days, as it facilitates search and retrieval of multimedia content. However, noisy and spam annotations often make it difficult to perform an efficient search. Users may make mistakes in tagging and irrelevant tags and content may be maliciously added for advertisement or self-promotion. This article surveys recent advances in techniques for combatting such noise and spam in social tagging. We classify the state-of-the-art approaches into a few categories and study representative examples in each. We also qualitatively compare and contrast them and outline open issues for future research.

BACKGROUND AND SIGNIFICANCE

Social networks and multimedia content sharing Web sites have become increasingly popular in recent years. Their service typically focuses on building online communities of people who share interests and activities, or are interested in exploring the interests and activities of others. At the same time, they have become a popular way to share and disseminate information. For example, users upload their personal photos and share them through online communities, letting other people comment or rate them. This trend has resulted in a continuously growing volume of publicly available multimedia content on content sharing Web sites like Flickr [33], Picasa [34], and YouTube [35] as well as social networks like Facebook [36], which have created new challenges for access, search, and retrieval of the shared content. For instance, Flickr has hosted more than 6 billion photos since August 2011 [1], and Facebook has approximately 100 billion photos stored on its servers [48]. Every minute, 48 h of video are uploaded to YouTube [49], and 20 million videos are uploaded to Facebook every month [2].

Tagging is one of the popular methods to manage a large volume of multimedia content. It is a process by which users assign short textual annotations to the content (in the form of keywords) to describe content and to provide additional information to other users who are interested in that content. Tags, when combined with search technologies, are essential in resolving user queries targeting shared content. The success of social networks such as Flickr, YouTube, Delicious [37], and Facebook proves that users are willing to provide tags through manual annotations. Different users who annotate the same multimedia content can provide different annotations, which enrich information about that content.

The entities (or objects) that make up the model of a social tagging system [3] are shown in Figure 1. The model consists of users who interact with the system, content (resources or documents) that might be any piece of information

Signal and Information Processing for Social Learning and Networking

©iSTOCKPHOTO.COM/ANDREY PROKHOROV

Digital Object Identifier 10.1109/MSP.2011.942345
Date of publication: 17 February 2012

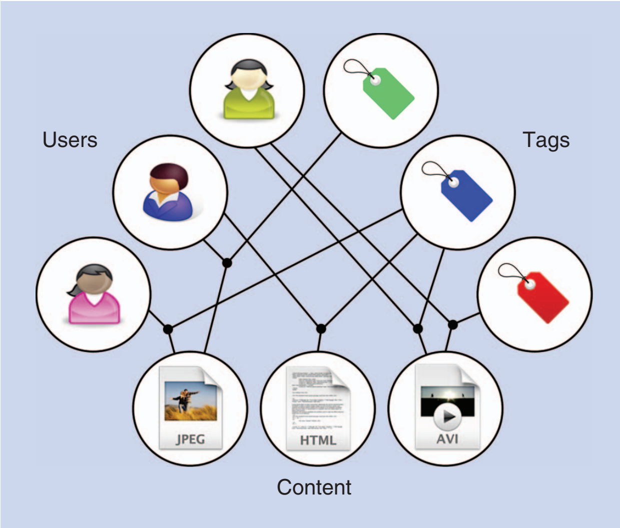
1053-5888/12/$31.00©2012IEEE

(e.g., photos, videos, textual documents, or Web pages), and tags that are descriptions attached to content by users. The action of associating a tag to a content by a user is usually referred to as tag assignment [4]. Depending on the system under consideration, a user can assign one or several tags to each content.

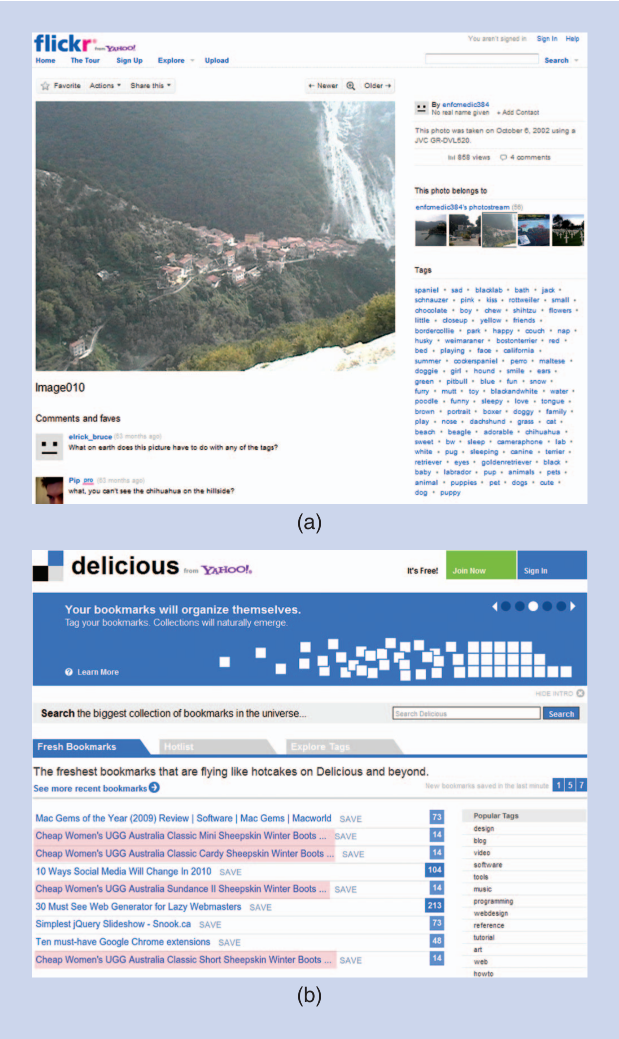
ONE IMPORTANT CHALLENGE IN TAGGING IS TO IDENTIFY THE MOST APPROPRIATE TAGS FOR GIVEN CONTENT, AND AT THE SAME TIME, TO ELIMINATE NOISY OR SPAM TAGS.

One important challenge in tagging is to identify the most appropriate tags for given content, and at the same time, to eliminate noisy or spam tags. The shared content is sometimes assigned with inappropriate tags for several reasons. First of all, users are human beings and may commit mistakes. Moreover, it is possible to provide wrong tags on purpose for advertisement, self-promotion, or to increase the rank of a particular tag in automatic search engines. Consequently, assigning free-form keywords (tags) to multimedia content has a risk that wrong or irrelevant tags eventually prevent users from the benefits of annotated content. Kennedy et al. [5] analyzed the Flickr Web site and revealed that the tags provided by users are often imprecise and only around 50% of tags are truly related to an image. Beside the tag-content association, spam objects can take other forms, i.e., possibly manifesting as a spam content or a spam user (spammer). Figure 2 shows examples of imprecise or spam tags and content on two popular social tagging systems.

To reduce or eliminate spams, various antispam methods have been proposed in the state-of-the-art research. Heymann et al. [6] classified antispam strategies into three categories: prevention, detection, and demotion. Prevention-based approaches aim at making it difficult for spam content to contribute to social tagging systems by restricting certain access types through interfaces [such as CAPTCHA [7] (which stands for "completely automated public Turing test to tell computers and humans apart") or reCAPTCHA [8]] or through usage limits (such as tagging quota, e.g., Flickr introduced a limit of 75 tags per photo [9]). Detection approaches identify likely spams either manually or automatically by making use of, for example, machine learning (such as text classification) or statistical analysis (such as link analysis), and then deleting the spam content or visibly marking it as hidden to users. Finally, demotion-based approaches reduce the prominence of content likely to be spam. For instance, rank-based methods produce ordering of a system's content, tags or users based on their trust scores. The prevention-based approaches can be considered as a type of precaution to prevent spammers. However, they cannot



[FIG1] General model of a social tagging system is represented as a tripartite graph structure that includes three kinds of nodes (objects): users, content, and tags. An edge linking a user, a tag, and a content represents a tag assignment.



By enfomedic384
No real name given

This photo was taken on October 6, 2002 using a JVC GR-DVL520.

858 views 4 comments

This photo belongs to

enfomedic384's photostream (56)

Tags

spaniel • sad • blacklab • bath • jack • schnauzer • pink • kiss • rottweiler • small • chocolate • boy • chew • shihtzu • flowers • little • closeup • yellow • friends • bordercollie • park • happy • couch • nap • husky • weimaraner • bostonterrier • red • bed • playing • face • california • summer • cockerspaniel • perro • maltese • doggie • girl • hound • smile • ears • green • pitbull • blue • fun • snow • furry • mutt • toy • blackandwhite • water • poodle • funny • sleepy • love • tongue • brown • portrait • boxer • doggy • family • play • nose • dachshund • grass • cat • beach • beagle • adorable • chihuahua • sweet • bw • sleep • cameraphone • lab • white • pug • sleeping • canine • terrier • retriever • eyes • goldenretriever • black • baby • labrador • pup • animals • pets • animal • puppies • pet • dogs • cute • dog • puppy

Image010

Comments and faves

elrick_bruce (63 months ago)
What on earth does this picture have to do with any of the tags?

Pip pro (63 months ago)
what, you can't see the chihuahua on the hillside?

(a)

Your bookmarks will organize themselves.
Tag your bookmarks. Collections will naturally emerge.

Search the biggest collection of bookmarks in the universe...

The freshest bookmarks that are flying like hotcakes on Delicious and beyond.

New bookmarks saved in the last minute 1 5 7

Mac Gems of the Year (2009) Review | Software | Mac Gems | Macworld	SAVE	73
Cheap Women's UGG Australia Classic Mini Sheepskin Winter Boots ...	SAVE	14
Cheap Women's UGG Australia Classic Cardy Sheepskin Winter Boots ...	SAVE	14
10 Ways Social Media Will Change In 2010	SAVE	104
Cheap Women's UGG Australia Sundance II Sheepskin Winter Boots ...	SAVE	14
30 Must See Web Generator for Lazy Webmasters	SAVE	213
Simplest jQuery Slideshow - Snook.ca	SAVE	73
Ten must-have Google Chrome extensions	SAVE	48
Cheap Women's UGG Australia Classic Short Sheepskin Winter Boots ...	SAVE	14

Popular Tags
design
blog
video
software
tools
music
programming
webdesign
reference
tutorial
art
web
howto

(b)

[FIG2] Examples of imprecise or spam tags and content on popular social tagging systems: (a) wrong tags in Flickr—only a few tags in the list are related to the image, while the rest is irrelevant (e.g., yellow, love, doggy) and (b) spam bookmarks in Delicious—all bookmarks are seeded by the same account and tagged by the same users.

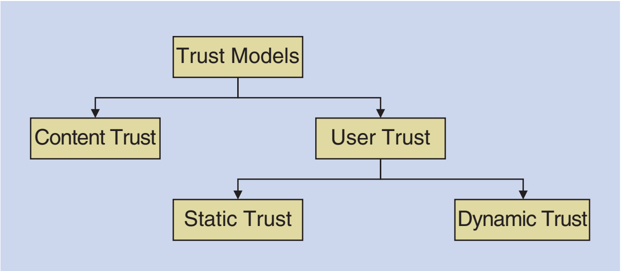
TRUST-BASED SCHEMES CAN BE USED TO MOTIVATE USERS TO POSITIVELY CONTRIBUTE TO SOCIAL NETWORK SYSTEMS AND/OR PENALIZE ADVERSARIES WHO TRY TO DISRUPT THE SYSTEM.

completely secure a social tagging system. Some studies, e.g., [10], showed that CAPTCHA systems can be defeated by computers with around 90% accuracy, using, for example, optical character recognition or shape context matching. Even if prevention methods were perfect, there would be still possibility that the social systems get polluted with spam (malicious) or irrelevant tags. Therefore, detection and demotion via trust modeling are required to keep a system free of noise and spam.

Trust provides a natural security policy stipulating that users or content with low trust values should be investigated or eliminated. Trust can predict the future behavior of users to avoid undesirable influences of untrustworthy users. Trust-based schemes can be used to motivate users to positively contribute to social network systems and/or penalize adversaries who try to disrupt the system. The distribution of the trust values of the users or content in a social network can be used to represent the health of that network.

TRUST MODELING

When information is exchanged on the Internet, malicious individuals are everywhere, trying to take advantage of the information exchange structure for their own benefit, while bothering and spamming others. Before social tagging became popular, spam content was observed in various domains: first in e-mail (e.g., [11]), and then in Web search (e.g., [12]). Peer-to-peer (P2P) networks have been also influenced by malicious peers, and thus various solutions based on trust and reputation have been proposed, which dealt with collecting information on peer behavior, scoring and ranking peers, and responding based on the scores [13]. Today, even blogs are spammed [14]. Ratings in online reputation systems, such as eBay [38], Amazon [39], and Epinions [40], are very similar to tagging systems and they may face the problem of unfair ratings by artificially inflating or deflating reputations [15]. Several filtering techniques for excluding unfair ratings are proposed in the literature (e.g., [16] and [17]). Unfortunately, the countermeasures developed for e-mail and Web spam do not directly apply to social networks [6].



[FIG3] Categorization of trust models surveyed in this article. Based on the target of trust, one can distinguish between a user and a content trust modeling. Further, user trust models can be divided into static and dynamic models.

In a social tagging system, spam or noise can be injected at three different levels: spam content, spam tag-content association, and spammer [18]. Trust modeling can be performed at each level separately (e.g., [18]) or different levels can be considered jointly to produce trust models, for example, to assess a user's reliability, one can consider not only the user profile, but also the content that the user uploaded to a social system (e.g., [19]). In this article, we categorize trust modeling approaches into two classes according to the target of trust, i.e., user and content trust modeling (shown in Figure 3). Table 1 summarizes representative recent approaches for trust modeling in social tagging. Presented approaches are sorted based on their complexity from simple to advanced, separately for both content and user trust models.

CONTENT TRUST MODELING

Content trust modeling is used to classify content (e.g., Web pages, images, and videos) as spam or legitimate. In this case, the target of trust is a content (resource), and thus a trust score is given to each content based on its content and/or associated tags. Content trust models reduce the prominence of content likely to be spam, usually in query-based retrieval results. They try to provide better ordering of the results to reduce the exposure of the spam to users. Koutrika et al. [20] proposed that each incorrect content found in a system could be simply removed by an administrator. The administrator can go a step further and remove all content contributed by the user who posted the incorrect content, on the assumption that this user is a spammer (polluter).

Approaches for content trust modeling utilize features extracted from content information, users' profiles and/or associated tags to detect specific spam content. Gyongyi et al. [21] proposed an algorithm called TrustRank to semiautomatically separate reputable from spam Web pages. TrustRank relies on an important empirical observation called approximate isolation of the good set: good pages seldom point to bad ones. It starts from a set of seeds selected as highly qualified, credible, and popular Web pages in the Web graph, and then iteratively propagate trust scores to all nodes in the graph by splitting the trust score of a node among its neighbors according to a weighting scheme. TrustRank effectively removes most of the spam from the top-scored Web pages, however it is unable to effectively separate low-scored good sites from bad ones, due to the lack of distinguishing features. In search engines, TrustRank can be used either solely to filter search results, or in combination with PageRank and other metrics to rank content in search results.

Heymann et al. [6] and Koutrika et al. [20] were the first to explicitly discuss methods of tackling spamming activities in social tagging systems. They studied the impact of spamming through a framework for modeling social tagging systems and user tagging behavior. They proposed a method for ranking content matching a tag based on taggers' reliability. Their coincidence-based model for query-by-tag search estimates the level of agreement among different users in the system for a given tag. A content is ranked high if it is tagged correctly by many reliable users. A user is more reliable if his/her tags more often coincide with other users' tags [6]. It was shown that spam in tag search results using the coincidence-based model is ranked lower than in results generated by, e.g., a traditional occurance-based model, where content is ranked based on the number of posts that associate the content to the query tag.

[TAGS, WHEN COMBINED WITH SEARCH TECHNOLOGIES, ARE ESSENTIAL IN RESOLVING USER QUERIES TARGETING SHARED CONTENT.]

Wu et al. [22] proposed a computer vision-based technique that discriminates spam images from legitimate ones. By assuming that images containing text are likely to be spam (e.g., banners), they identified a number of useful low-level image features detecting embedded text and computer-generated graphics. Then, pattern classification using support vector machines (SVMs) was performed to classify spam and nonspam images. Although they reported a high detection rate with a low false positive rate, this approach has limitations in that the discriminant capability of the used features may be limited and, moreover, the assumption that images containing text or computer-generated images are likely to be spam may not be true in some cases.

[TABLE 1] SUMMARY OF REPRESENTATIVE RECENT TECHNIQUES FOR COMBATTING NOISE AND SPAM IN SOCIAL TAGGING SYSTEMS.

REFERENCE	TRUST MODEL	MEDIA	METHOD	DATA SET
GYONGYI ET AL. [21]	CONTENT	WEB PAGES	AN ITERATIVE APPROACH, CALLED TRUSTRANK, TO PROPAGATE TRUST SCORES TO ALL NODES IN THE GRAPH BY SPLITTING THE TRUST SCORE OF A NODE AMONG ITS NEIGHBORS ACCORDING TO A WEIGHTING SCHEME	ALTAVISTA, REAL
KOUTRIKA ET AL. [20]	CONTENT	BOOKMARKS	A COINCIDENCE-BASED MODEL FOR QUERY-BY-TAG SEARCH WHICH ESTIMATES THE LEVEL OF AGREEMENT AMONG DIFFERENT USERS IN THE SYSTEM FOR A GIVEN TAG	DELICIOUS, REAL AND SIMULATED
WU ET AL. [22]	CONTENT	IMAGES	A COMPUTER VISION TECHNIQUE BASED ON LOW-LEVEL IMAGE FEATURES TO DETECT EMBEDDED TEXT AND COMPUTER-GENERATED GRAPHICS	SPAMARCHIVE AND LING-SPAM, REAL
LIU ET AL. [4]	CONTENT AND USER	BOOKMARKS	AN ITERATIVE APPROACH TO IDENTIFY SPAM CONTENT BY ITS INFORMATION VALUE EXTRACTED FROM THE COLLABORATIVE KNOWLEDGE	DELICIOUS, REAL
BOGERS AND VAN DEN BOSCH [23]	CONTENT AND USER	BOOKMARKS	KL-DIVERGENCE TO MEASURE THE SIMILARITY BETWEEN LANGUAGE MODELS AND NEW POSTS	BIBSONOMY AND CITEULIKE, REAL
IVANOV ET AL. [24]	USER	IMAGES	AN APPROACH BASED ON THE FEEDBACK FROM OTHER USERS WHO AGREE OR DISAGREE WITH A TAG ASSOCIATED WITH AN IMAGE	PANORAMIO, REAL
XU ET AL. [25]	USER	BOOKMARKS	AN ITERATIVE APPROACH TO COMPUTE THE GOODNESS OF EACH TAG WITH RESPECT TO A CONTENT AND THE AUTHORITY SCORES OF THE USERS	MYWEB 2.0, REAL
KRESTEL AND CHEN [26]	USER	BOOKMARKS	A TRUSTRANK-BASED APPROACH USING FEATURES WHICH MODEL TAG CO-OCCURANCE, CONTENT CO-OCCURANCE AND CO-OCCURANCE OF TAG-CONTENT	BIBSONOMY, REAL
BENEVENUTO ET AL. [27]	USER	VIDEOS	A SUPERVISED LEARNING APPROACH APPLIED ON FEATURES THAT REFLECT USERS BEHAVIOR THROUGH VIDEO RESPONSES	YOUTUBE, REAL
LEE ET AL. [28]	USER	TWEETS	A MACHINE LEARNING APPROACH APPLIED ON SOCIAL HONEYPOTS INCLUDING USERS' PROFILE AND TWEETS' FEATURES	TWITTER, REAL AND SIMULATED
KRAUSE ET AL. [19]	USER	BOOKMARKS	A MACHINE LEARNING APPROACH APPLIED ON A USER'S PROFILE, BOOKMARKING ACTIVITY AND CONTEXT OF TAGS FEATURES	BIBSONOMY, REAL
MARKINES ET AL. [18]	USER	BOOKMARKS	A MACHINE LEARNING APPROACH APPLIED ON TAG-, CONTENT- AND USER-BASED FEATURES	BIBSONOMY, REAL
NOLL ET AL. [29]	USER	BOOKMARKS	AN ITERATIVE GRAPH-BASED ALGORITHM, CALLED SPEAR, TO COMPUTE THE EXPERTISE SCORE OF A USER AND THE QUALITY SCORE OF A CONTENT CONSIDERING THE TIME OF TAGGING	DELICIOUS, REAL AND SIMULATED
CAVERLEE ET AL. [30]	USER	USER PROFILES	AN APPROACH TO COMPUTE A DYNAMIC TRUST SCORE, CALLED SOCIALTRUST, DEPENDING ON THE QUALITY OF THE RELATIONSHIP AND PERSONALIZED FEEDBACK RATINGS RECEIVED FROM NEIGHBORS IN A SOCIAL GRAPH	MYSPACE, REAL

Bogers and Van den Bosch [23] used language models based on features such as title, description, tags, and URL of posts, to automatically detect spam content in social bookmarking systems such as BibSonomy [41] and CiteULike [42]. Their method is based on the intuitive notion that different users (legitimate users versus spammers) tend to use different language when posting. To detect spam content, they learned a language model for each post and then measured its similarity to the incoming posts by making use of Kullback-Leiber (KL) divergence. The spam status of the content in a new post takes the status of the most similar language model. Furthermore, all posts uploaded by a user are collated together to construct a language model of the user profile. Grouping posts of one user to form large document that can be considered as the user profile makes the spam detection reliable, as shown in [23].

MOST OF USER TRUST MODELING TECHNIQUES USE MACHINE LEARNING APPROACHES APPLIED TO FEATURES SPECIFIC TO CONSIDERED SOCIAL NETWORK DOMAINS.

Liu et al. [4] proposed a simple but effective approach for detecting spam content in Delicious, by harvesting the wisdom of crowds. An information value of a content is defined as the average number of times that each tag of the content is assigned by different users. A low information value of a content indicates a divergence from crowds, which can be considered as a spam content. Furthermore, this method was extended to user trust modeling by aggregating the information values for each user.

Although the aforementioned content trust modeling methods have shown to be effective in combating spam, the "subjectivity" in classifying spam and nonspam content remains as a fundamental issue, i.e., what is spam content to one user may be interesting to another, and vice versa [18].

USER TRUST MODELING

In user trust modeling, trust is given to each user based on the information extracted from a user's account, his/her interaction with other participants within the social network, and/or the relationship between the content and tags that the user contributed to the social tagging system. Given a user trust score, the user might be flagged as a legitimate user or spammer.

User trust can be established in a centralized or distributed manner [15]. In centralized trust systems, users' trust models are maintained by one central authority, i.e., manager, while in distributed trust systems each user maintains his/her own trust manager based on the previous interactions with other users. Distributed trust models are mainly used in P2P networks [15], while social networks usually use centralized systems (e.g., [18], [27], [24], and [30]).

Most of user trust modeling techniques use machine learning approaches applied to features specific to considered social network domains. Krause et al. [19] employed a machine learning approach to identify spammers in BibSonomy. They investigated features considering information about a user's profile (e.g., number of digits in the username and the e-mail address), location (e.g., number of spam users with the same IP), bookmarking activity (e.g., number of tags per post), and context of tags (e.g., user co-occurrences with spammers related to tags, content and tag-content pairs). By making use of these features and SVM or naive Bayes classifier, they were able to distinguish legitimate users from malicious ones. It was found that the co-occurrence features describing the usage of a similar vocabulary and content usage are the most promising.

The recent approach by Lee et al. [28] uses social honeypots for uncovering spammers in Twitter [43]. Social honeypots are system resources that monitor spammers' behaviors and log their information (e.g., their profiles and content created by them). Lee et al. examined different users' profile features (e.g., longevity of their accounts, the ratio of the number of followings and the number of followers) and features extracted from tweets (e.g., the number of URLs and @usernames), and found that a majority of spammers post a series of nearly identical tweets just by changing @username or @replies, and also post tweets with URLs. Therefore, text-based features extracted from tweets and the ratio of the number of URLs in recently posted top 20 tweets among a user's tweets showed the best discrimination power in detecting spammers. On the other hand, when this approach is applied over a large collection of profiles and tweets, the results are significantly worse than what was observed over the small data set of controlled data, because of the mismatch due to the time difference between the harvested social honeypots and the large data set of Twitter profiles and tweets used for test.

Benevenuto et al. [27] proposed various features for detecting spammers and promoters in the user feedback of YouTube videos. Spammers were defined as users who post an unrelated video as response to a popular one (e.g., pornographic content posted as response to a cartoon video), aiming at increasing the likelihood of the response being viewed by a larger number of users. On the other hand, promoters are those who try to gain visibility of a specific video by posting a large number of (potentially unrelated) responses to boost the rank of the responded video (e.g., a sequence of 100 unrelated video responses to a single video, often with very short durations), making it appear in the top of the lists in the system. In this approach, they considered features that reflect users' behavior through video responses, such as video attributes (e.g., duration, number of views, ratings, number of times the video was selected as favorite), user attributes (e.g., number of friends, number of videos uploaded, number of videos watched, numbers of video responses posted and received) and social network attributes (e.g., UserRank). A nonlinear SVM classifier was then applied on these features to classify users into three classes: legitimate users, spammers and promoters. It was shown that this approach is able to correctly identify majority of promoters, while misclassifying only a small percentage of legitimate users. Distinguishing spammers was much harder, because legitimate users also post video responses to popular videos, which is a typical behavior of spammers.

Markines et al. [18] proposed six different tag-, content- and user-based features for automatic detection of spammers in BibSonomy. First, tag- and content-based features are averaged across each user's posts, then combined with user-based features, and finally fed into a supervised learning algorithm (such as LogitBoost or AdaBoost) to discriminate spammers from legitimate users. It was shown that TagSpam feature (probability that a particular tag is used to spam, aggregated across all tags assigned to a content) is the best predictor of spammers among all other features, because spammers tend to use certain "suspect" tags more than legitimate users. The DomFp feature (likelihood that a content is spam based on its structure) also appeared important but may not be available since it relies on an infrastructure to enable access to the content, and therefore its feasibility depends on the circumstances of a particular social tagging system.

Recently, Ivanov et al. [24] explored features extracted from the knowledge accumulated in photo sharing social networks such as Panoramio [44]. They proposed an approach to model the user trust by making use of the feedback from other users who agree or disagree with a tag associated with an image. The more disagreement a user has, the more distrusted he/she is. Therefore, a user's trust score is calculated as the ratio between the number of correctly tagged images and the number of all images tagged by that user. Further, they introduced user trust modeling in the framework of a geotag propagation system, which will be described in the section "Illustrative System," and showed that by considering a simple user trust model the accuracy of the geotag propagation system could be considerably improved.

Xu et al. [25] introduced the concept of "authority" in social bookmarking systems, where they measured the goodness of each tag with respect to a content by the sum of the authority scores of the users who have assigned the tag to the content. Authority scores and goodness are iteratively updated by using hyperlink-induced topic search (HITS), which was initially used to rank Web pages based on their linkage on the Web [31]. In contrast, Krestel and Chen [26] iteratively updated scores for users only. They proposed to use a spam score propagation technique to propagate trust scores through a social graph, similar to that shown in Figure 1, where edges between nodes (in this case, users) indicate the number of common tags supplied by users, common content annotated by users and/or common tag-content pairs used by users. Starting from a manually assessed set of nodes labeled as spammers or legitimate users with the initial spam scores, a TrustRank metric is used to calculate spam scores for all users. This approach is more sophisticated than the approach in [25] in that multiple relationships, such as tag cooccurance, content cooccurance and tag-content cooccurance, can be taken into account, rather than considering only the tag-content pairs shared by users.

The aforementioned studies consider users' reliability as static at a specific moment. However, a user's trust in a social tagging system is dynamic, i.e., it changes over time. The tagging history of a user is better to consider, because a consistent good behavior of a user in the past can suddenly change by a few mistakes, which consequently ruins his/her trust in tagging.

Noll et al. [29] introduced the time of tagging as an additional dimension for assessing the trust of a user in Delicious. They proposed a graph-based algorithm, called spamming-resistant expertise analysis and ranking (SPEAR). It computes the expertise score of a user and the quality score of a content which are dependent on each other. The time of tagging is considered so that the earlier a user tags a content, the more expertise score he/she receives. These two scores are calculated iteratively in a similar way to that of the HITS algorithm. It was shown that SPEAR produces better ranking of users than the HITS method. SPEAR was able to demote different types of spammers (flooders, promoters, and trojans [29]) and remove them from the top of the ranking.

Caverlee et al. [30] proposed the use of users' behavior and feedback for dynamic trust establishment in MySpace [45]. A dynamic trust score, called SocialTrust, is derived for each user. It depends on the quality of the relationship with his/her neighbors in a social graph and personalized feedback ratings received from neighbors so that trust scores are updated as the social network evolves. The dynamics of the system is modeled by including the evolution of the user's trust score to incent long-term good behavior and to penalize users who build up a good trust rating and suddenly "defect." It was shown that SocialTrust is resilient to the increase in number of malicious users, since the highly trusted users manage to keep them under control thanks to the trust-aware feedback scheme introduced in this approach. It was also shown that SocialTrust outperforms TrustRank-based models, because SocialTrust model incorporates relationship quality and feedback ratings into the trust assessment so that bad behavior is punished.

[IT IS NOTICEABLE THAT USER TRUST MODELING IS MORE POPULAR THAN CONTENT TRUST MODELING.]

It is noticeable that user trust modeling is more popular than content trust modeling. One reason of this is that the former has a less complexity when compared to the latter, i.e., the number of models required is usually much larger in content trust modeling. The other reason is that user trust models can quickly adapt to the constantly evolving and changing environment in social systems due to the type of features used for modeling, and thus be applicable longer than content trust models, without need for creation of new models. On the other hand, user trust modeling has a disadvantage of "broad brush," i.e., it may be excessively strict if a user happens to post one bit of questionable content on otherwise legitimate content. The trustworthiness of a user is often judged based on the content that the user uploaded to a social system, and thus "subjectivity" in discriminating spammers from legitimate users remains an issue for user trust modeling as in content trust modeling.

[DATA SETS USED FOR DEVELOPMENT AND EVALUATION OF TRUST MODELING TECHNIQUES HAVE A WIDE RANGE OF DIVERSITY IN TERMS OF CONTENT, NUMBERS OF RESOURCES, TAGS AND USERS, AND TYPE OF SPAM.]

EVALUATION

DATA SET

Data sets used for development and evaluation of trust modeling techniques have a wide range of diversity in terms of content, numbers of resources, tags and users, and type of spam.

Social bookmarking is the most popularly explored domain for trust modeling, especially user trust modeling, as shown in Table 1.

Some researchers dealing with bookmarks used a public data set released by BibSonomy as a part of the ECML PKDD Discovery Challenge 2008 on Spam Detection in Social Bookmarking Systems [32]. This data set consists of around 2,400 legitimate users and more than 29,000 spammers, which were manually labeled as spammers and nonspammers, and in all their posts. Data provided in this data set can serve for the evaluation of both user (e.g., [18], [23], and [26]) and content trust models (e.g., [23]). However, the skewness is present in this data set since a majority of the bookmarks, 94% out of around 14 million, are spam. Markines et al. overcame this issue by selecting randomly only a subset of users (around 250 legitimate users and 250 spammers) to achieve a balance with respect to the number of users, since some features in their approach for the user trust modeling rely on the statistics of the data set [18].

Other researchers have collected data about bookmarks by crawling Delicious or BibSonomy during a limited period of time. For example, Krause et al. [19] collected around 20,000 users and 1.2 million bookmarks from BibSonomy, which is big enough to match real-world applications. The largest data set for trust modeling in Delicious was collected by Liu et al. [4] and had around 82,000 users, 1.1 million tags, 9.3 million bookmarks, and 17.4 million tag-bookmark associations.

To model trust in other types of tagging systems, where spam is introduced through videos, tweets, or user profiles, data are usually crawled from the corresponding social network, like YouTube, Twitter, or MySpace, respectively. For example, Lee et al. [28] collected around 215,000 users and 4 million tweets from Twitter. Since this raw data are missing ground truth for evaluation, they manually labeled a small portion of users distinguishing between legitimate users, spammers, and promoters and then evaluated their approach. The same approach for the evaluation of a trust model was followed by Benevenuto et al. [27], who crawled a large-scale data set from YouTube containing more than 260,000 users and 1 million videos. Caverlee et al. [30] crawled 892,000 user profiles and around 20 million relationship links from MySpace.

Not all data sets for trust modeling have large-scale data. For example, Ivanov et al. [24] argued that their approach requires a small number of images to learn models for geotag propagation with user trust modeling, and they evaluated their approach on a data set of 1,320 images of famous landmarks downloaded from Google Images [46], Flickr, and Wikipedia [47], and 44 users who annotated images.

Koutrika et al. [20] performed a variety of evaluations of their trust model on controlled (simulated) data set by populating a tagging system with different user tagging behavior models, including a good user, bad user, targeted attack model, and several other models. Using controlled data, interesting scenarios that are not covered by real-world data could be explored.

Most of the data sets are not publicly available, except data sets in [22] and [32].

PERFORMANCE METRICS

Trust modeling can be formulated as either a classification problem or a ranking problem, depending on the way of treatment.

In the classification problem, the results of an algorithm can be summarized by a confusion matrix from ground-truth data and predicted labels, which contains the number of true positives, true negatives, false positives, and false negatives. From these values, classical measures such as a receiver operating characteristic (ROC), the area under the ROC curve (AUC), precision-recall (PR) curves, and F-measure can be derived.

SpamFactor is a performance metric for a ranking problem [20]. It measures the impact of a spam object, either user or content, in a ranking list. Assuming the scenario in which the trust model returns a ranked list of N contents, c_i for $i \in [1, N]$, for a query tag t, SpamFactor is defined as

$$\text{SpamFactor}(N, t) = \frac{\sum_{i=1}^{N} \omega(c_i, t) \cdot \frac{1}{i}}{\sum_{i=1}^{N} \frac{1}{i}},$$

where

$$\omega(c_i, t) = \begin{cases} 1, & \text{if } t \text{ is a bad tag for } c_i, \\ 0, & \text{if } t \text{ is a good tag for } c_i. \end{cases}$$

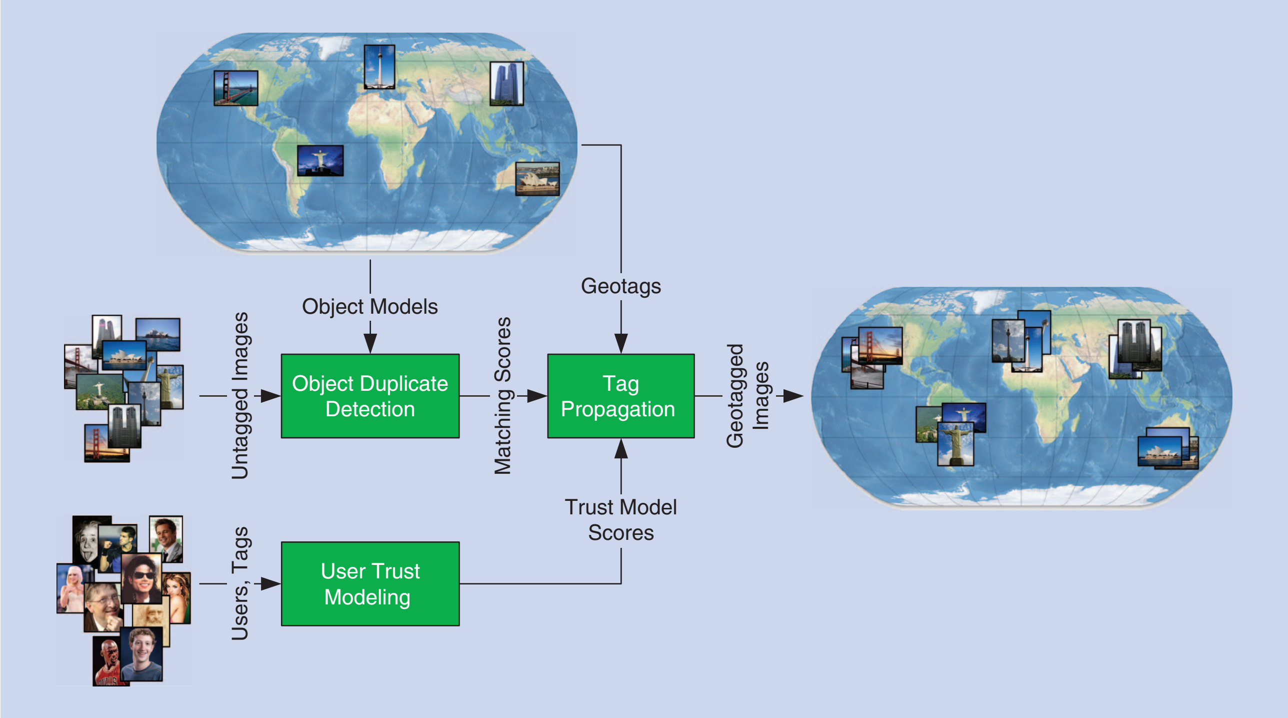
SpamFactor is normalized between zero and one. It takes into consideration both the number of spam objects and their position in the result list. The higher the position of a spam content in the ranking list, the greater contribution of the content to SpamFactor is. A higher SpamFactor represents existence of more spam objects or higher ranks of spam objects in the result list.

ILLUSTRATIVE SYSTEM

In this section, an example system using trust modeling is described to demonstrate how a social tagging system can benefit from trust modeling. Particularly, the geotagging scenario is considered, which is now a very popular application due to the fact that a large portion of Internet images in social networks are related to travel. Travel is an important type of event for which people like to share, annotate, and search pictures. For the majority of travel images on the Internet, however, proper geographical annotations are not available. In most cases, the images are annotated by users manually. To speed up this time-consuming manual tagging process, geotags can be propagated based on the similarity between image content (usually famous landmarks) and context (associated geotags).

GEOTAGS CAN BE PROPAGATED BASED ON THE SIMILARITY BETWEEN IMAGE CONTENT (USUALLY FAMOUS LANDMARKS) AND CONTEXT (ASSOCIATED GEOTAGS).

Ivanov et al. [24] developed an efficient system for automatic geotag propagation in images by associating locations with distinctive landmarks and using object duplicate detection. The system overview is shown in Figure 4. The robust graph-based object duplicate detection approach reliably establishes the correspondence between a small set of tagged images and a large set of untagged images by searching for the same landmark depicted in different images, to propagate geotags from the former to the latter. In tag propagation, trust modeling is especially crucial because propagating a wrong/spam tag can easily damage the integrity and reliability of the whole system. A user trust modeling derived for each user is introduced in the geotagging system of [24] by making use of the feedback from other users who agree or disagree with a tag associated with an image, so that only reliable geotags are propagated. It was shown that the proposed user trust model can be generalized to photo sharing platforms, such as Panoramio or Flickr. The performance of the proposed geotag propagation system was evaluated on a set of 1,320 images depicting 66 famous landmarks that were obtained from Google Images, Flickr, and Wikipedia. Due to the lack of a suitable data set that provides user feedback from Panoramio to compute trust scores for users, the evaluation of the user trust model is based on the simulation of the social network environment. In the simulated social network, 44 users were asked to tag 66 photos from the data set, putting the name of the landmark depicted in the image. The user trust model is then created based on the correlation of these geotags with the landmark in the image. It was shown that consideration of the user trust model leads to an increased accuracy of the tag propagation (from 46% without trust modeling to 65% with trust modeling in terms of recognition rate [24]) and a decrease of tagging efforts (more than 10,000 tags from



[FIG4] Overview of the system for geotag propagation in images. The object duplicate detection is trained with a small set of images with associated geotags. The created object (landmark) models are matched against untagged images. The resulting matching scores serve as an input to the tag propagation module, which propagates the corresponding tags to the untagged images. Given a user trust model, only the tags from reliable users are propagated [24]. (Figure used with permission from [24].)

trusted users can be automatically propagated, while keeping accuracy higher than 46%).

OPEN ISSUES AND CHALLENGES

There have been a variety of data sets from different social networks and even different data sets of one social network for evaluation of trust modeling approaches, as shown in the "Evaluation" section. However, publication of such data sets is rarely found, which makes it difficult to compare results and performance of different trust modeling approaches. Therefore, it would be desirable to promote researchers to make their data sets publicly available to the research community, which can be used for comparison and benchmarking of different approaches. Furthermore, most of the data sets provide data for evaluating only one aspect of trust modeling, either user or content trust modeling, while evaluation of the other aspect requires introducing simulated objects in the real-world social tagging data sets (e.g., [20] and [29]). However, for the thorough evaluation of a trust model it is necessary that real-world data sets have ground-truth data for both users and content.

We already noted that a user's trust tends to vary over time according to the user's experience and evolvement of social networks. However, only a few approaches (e.g., [29] and [30]) deal with dynamics of trust by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling of phenomenon in real-world applications.

Most of the existing trust modeling approaches based on text information assume monolingual environments. However, many social network services are used by people from various countries, so that various languages simultaneously appear in tags and comments. In such cases, some text information may be regarded as wrong due to the language difference. Therefore, incorporating the multilingualism in trust modeling would be useful to solve this problem.

Today, it is observed that interaction across social networks becomes popular. For example, users can use their Facebook accounts to log in some other social network services. Thus, a future challenge in trust modeling is to investigate how trust models across domains can be effectively connected and shared.

As shown in the section "Trust Modeling," most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis, while audio and visual content features of multimedia content can also provide useful information about the relevance of the content and content-tag relationship (e.g., [22]). In the future, a promising research direction would be to combine multimedia content analysis with conventional tag processing and user profile analysis.

> **AS ONLINE SOCIAL NETWORKS AND CONTENT SHARING SERVICES EVOLVE RAPIDLY, WE BELIEVE THAT THE RESEARCH ON ENHANCING RELIABILITY AND TRUSTWORTHINESS OF SUCH SERVICES WILL BECOME INCREASINGLY IMPORTANT.**

CONCLUSIONS

In this article, we dealt with one of the key issues in social tagging systems: combatting noise and spam. We classified existing studies in the literature into two categories, i.e., content and user trust modeling. Representative techniques in each category were analyzed and compared. In addition, existing databases and evaluation protocols were reviewed. An example system was presented to demonstrate how trust modeling can be particularly employed in a popular application of image sharing and geotagging. Finally, open issues and future research trends were prospected. As online social networks and content sharing services evolve rapidly, we believe that the research on enhancing reliability and trustworthiness of such services will become increasingly important.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management (IM2), the Swiss National Science Foundation Grant "Multimedia Security" (200020-113709), and partially supported by the European Network of Excellence PetaMedia (FP7/2007-2011).

AUTHORS

Ivan Ivanov (ivan.ivanov@epfl.ch) is a research assistant and Ph.D. student in the Multimedia Signal Processing Group at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He received the Dipl. Ing. (M.Sc.) degree in electrical engineering from the University of Belgrade, Serbia, in 2006. In 2006 and 2007, he worked as a hardware design engineer for Texas Instruments, France, where he participated in the development of low-power very large-scale integration multimedia applications for portable devices. He also worked as a radio access network conceptual planning expert in Vip mobile, Serbia, focusing on the implementation of second and third generation radio access technologies. His research interests include multimedia content analysis and the tagging and retrieval of multimedia in social networking environments.

Peter Vajda (peter.vajda@epfl.ch) received his M.Sc. degree in computer science from the Vrije Universiteit, Amsterdam, The Netherlands, in 2006 as well as in program designer mathematics from Eötvös Loránd University, Budapest, Hungary, in 2007. He performed his diploma work on selection mechanisms in evolution computing and on using prediction algorithms in human-computer interaction. He has been a research assistant and Ph.D. student in the Multimedia Signal Processing Group at EPFL since September 2007. His research interests include mobile visual search, multimedia content analysis, and social networks.

Jong-Seok Lee (jong-seok.lee@yonsei.ac.kr) received the Ph.D. degree in electrical engineering from KAIST, Korea, in 2006. He worked as a postdoctoral researcher from 2006 to 2008 and an adjunct professor in 2007, both at KAIST. From 2008 to 2011, he was with the Multimedia Signal Processing Group at EPFL, Switzerland. Currently, he is an assistant professor at the School of Integrated Technology, Yonsei University, Korea. He was the chair of the First Spring School on Social Media Retrieval held in 2010 and an organizing committee member of its second edition in 2011. His research interests include multimedia processing, multimedia quality assessment, and multimodal human-computer interface.

Touradj Ebrahimi (touradj.ebrahimi@epfl.ch) is currently a professor at EPFL heading its Multimedia Signal Processing Group. He has been the recipient of various distinctions and awards, such as the IEEE and Swiss National ASE Award, the SNF-PROFILE grant for advanced researchers, seven ISO-certificates for key contributions to MPEG-4, JPEG 2000, JPSearch, and JPEG XR standards, and the *IEEE Transactions on Consumer Electronics* Best Paper Award. He is the head of the Swiss delegation to MPEG, JPEG, and SC29 and chair of the Advisory Group on Management in SC29. His research interests include still, moving, and three-dimensional image processing and coding, visual information security, new media, and human computer interaction.

REFERENCES

[1] Wikimedia Foundation Inc. (2011, Dec.). Flickr. [Online]. Available: http://en.wikipedia.org/wiki/Flickr

[2] Pingdom Blog. (2011, Jan.). Internet 2010 in numbers. [Online]. Available: http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers

[3] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "HT06, tagging paper, taxonomy, Flickr, academic article, to read," in *Proc. ACM HT*, Aug. 2006, pp. 31–40.

[4] K. Liu, B. Fang, and Y. Zhang, "Detecting tag spam in social tagging systems with collaborative knowledge," in *Proc. IEEE FSKD*, Aug. 2009, pp. 427–431.

[5] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label?: Predicting the performance of search-based automatic image classifiers," in *Proc. ACM MIR*, Oct. 2006, pp. 249–258.

[6] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting spam on social Web sites: A survey of approaches and future challenges," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 36–45, Nov. 2007.

[7] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Proc. Eurocrypt*, May 2003, pp. 294–311.

[8] L. von Ahn, B. Maurer, C. Mcmillen, D. Abraham, and M. Blum, "reCAPTCHA: Human-based character recognition via Web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, Aug. 2008.

[9] Yahoo, Inc. (2011, Dec.). Flickr—Tags. [Online]. Available: http://www.flickr.com/help/tags

[10] G. Mori and J. Malik, "Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA," in *Proc. IEEE CVPR*, June 2003, pp. I-134–I-141.

[11] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A Bayesian approach to filtering junk e-mail," AAAI Workshop on Learning for Text Categorization, Madison: WI, *Tech. Rep. WS-98-05*, July 1998.

[12] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages," in *Proc. ACM WebDB*, June 2004, pp. 1–6.

[13] S. Marti and H. Garcia-Molina, "Taxonomy of trust: Categorizing P2P reputation systems," *Comput. Netw.*, vol. 50, no. 4, pp. 472–484, Mar. 2006.

[14] A. Thomason, "Blog spam: A review," in *Proc. CEAS*, Aug. 2007.

[15] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Syst.*, vol. 43, no. 2, pp. 618–644, Mar. 2007.

[16] A. Whitby, A. Jøsang, and J. Indulska, "Filtering out unfair ratings in Bayesian reputation systems," in *Proc. IEEE AAMAS*, July 2004, pp. 106–117.

[17] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proc. ACM SAC*, Mar. 2009, pp. 1308–1315.

[18] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in *Proc. ACM AIRWeb*, Apr. 2009, pp. 41–48.

[19] B. Krause, C. Schmitz, A. Hotho, and G. Stum, "The anti-social tagger: Detecting spam in social bookmarking systems," in *Proc. ACM AIRWeb*, Apr. 2008, pp. 61–68.

[20] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, "Combating spam in tagging systems: An evaluation," *ACM TWEB*, vol. 2, no. 4, pp. 22:1–22:34, Oct. 2008.

[21] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating Web spam with TrustRank," in *Proc. VLDB*, Aug. 2004, pp. 576–587.

[22] C.-T. Wu, K.-T. Cheng, Q. Zhu, and Y.-L. Wu, "Using visual features for anti-spam filtering," in *Proc. IEEE ICIP*, Sept. 2005, vol. 3, pp. 509–512.

[23] T. Bogers and A. Van den Bosch, "Using language models for spam detection in social bookmarking," in *Proc. ECML PKDD*, Sept. 2008, pp. 1–12.

[24] I. Ivanov, P. Vajda, J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Geotag propagation in social networks based on user trust model," *Multimedia Tools Applicat.*, pp. 1–23, July 2010.

[25] Z. Xu, Y. Fu, J. Mao, and D. Su, "Towards the semantic Web: Collaborative tag suggestions," in *Proc. ACM WWW*, May 2006, pp. 1–8.

[26] R. Krestel and L. Chen, "Using co-occurence of tags and resources to identify spammers," in *Proc. ECML PKDD*, Sept. 2008, pp. 38–46.

[27] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves, "Detecting spammers and content promoters in online video social networks," in *Proc. ACM SIGIR*, July 2009, pp. 620–627.

[28] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *Proc. ACM SIGIR*, July 2010, pp. 435–442.

[29] M. G. Noll, C. A. Yeung, N. Gibbins, C. Meinel, and N. Shadbolt, "Telling experts from spammers: Expertise ranking in folksonomies," in *Proc. ACM SIGIR*, July 2009, pp. 612–619.

[30] J. Caverlee, L. Liu, and S. Webb, "SocialTrust: Tamper-resilient trust establishment in online communities," in *Proc. ACM JCDL*, June 2008, pp. 104–114.

[31] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *JACM*, vol. 46, no. 5, pp. 604–632, Sept. 1999.

[32] A. Hotho, D. Benz, R. Jäschke, and B. Krause, Eds. (2008, Sept.). *ECML PKDD Discovery Challenge* [Online]. Available: http://www.kde.cs.uni-kassel.de/ws/rsdc08

[33] Flickr Web site. [Online]. Available: http://www.flickr.com

[34] Picasa Web site. [Online]. Available: http://picasa.google.com

[35] YouTube Web site. [Online]. Available: http://www.youtube.com

[36] Facebook Web site. [Online]. Available: http://www.facebook.com

[37] Delicious Web site. [Online]. Available: http://www.delicious.com

[38] eBay Web site. [Online]. Available: http://www.ebay.com

[39] Amazon Web site. [Online]. Available: http://www.amazon.com

[40] Epinions Web site. [Online]. Available: http://www.epinions.com

[41] BibSonomy Web site. [Online]. Available: http://www.bibsonomy.org

[42] CiteULike Web site. [Online]. Available: http://www.citeulike.org

[43] Twitter Web site. [Online]. Available: http://www.twitter.com

[44] Panoramio Web site. [Online]. Available: http://www.panoramio.com

[45] MySpace Web site. [Online]. Available: http://www.myspace.com

[46] Google Images Web site. [Online]. Available: http://images.google.com

[47] Wikipedia Web site. [Online]. Available: http://www.wikipedia.org

[48] E. Barnett. (2011, Oct.). 3.4 billion photographs on Google+ in 100 days. [Online]. Available: http://www.telegraph.co.uk/technology/google/8838196/3.4-billion-photographs-on-Google-in-100-days.html

[49] Google Inc. (2011, Dec.) YouTube statistics. [Online]. Available: http://www.youtube.com/t/press_statistics