Categorization of Tags in Social Media

[1]K. Nandini, [2] P. Ashok Kumar
[1]M.Tech Student, [2]Assistant Professor
Dept. of Computer Science and Engineering,
Sri Vasavi Institute of Engineering and Technology, Nandamuru, AP, India

***Abstract* :** In this paper we propose, now a days tagging is very popular in online social networks, as it facilitates search and retrieval of multimedia data. However, noisy and spam annotations often make it difficult to perform an well-organized search. Users may make mistakes in tagging and irrelevant tags and content may be maliciously added for advertisement or self-promotion. This article studies recent advances in techniques for fight against such noise and spam in social tagging. Qualitatively compare and dividing line them and outline open issues for future research.

***Keywords* :** Tagging, Classification, Social Network, Annotation, spam.

I. INTRODUCTION

Social networks and multimedia content sharing Web sites have become more popular in now a days. Their service typically focuses on building online communities of people who share interests and activities. They have to become a popular way to share and broadcast the current information. This trend has resulted in a continuously growing volume of public accounts available in multimedia content on content sharing Web sites like Flicker, Picasa, and YouTube as well as social networks like Face book. Those have creating new challenges for access, search, and retrieval of the shared content. For resent survey, Flicker has hosted more than 6 billion photos since August 2011, and Face book has approximately 100 billion photos stored on its servers. Every minute, 48 h of video are uploaded to YouTube, and 21 million videos are uploaded to Face book per month.

Tagging is one of the most popular methods to manage a large volume of multimedia content. Tags, when combined with search technologies, are essential in deciding user queries targeting shared data. The success of social networks such as Flicker, Delicious, and Face book providing that users are willing to provide tags through manual adding notes. Different users who note the same multimedia content can provide different adding notes, which improve in quality of the information about that data.

II. EXISTING SYSTEM

One important challenge in tagging is to capture the most appropriate tags for given data, and to eliminate noisy. Sometimes the shared data assigned with un sufficient data on several reasons. Clients are human beings and may commit mistakes. It is possible to provide wrong tags on purpose for announcement, media manipulation, or to increase the rank of a particular tag in automatic search engines.

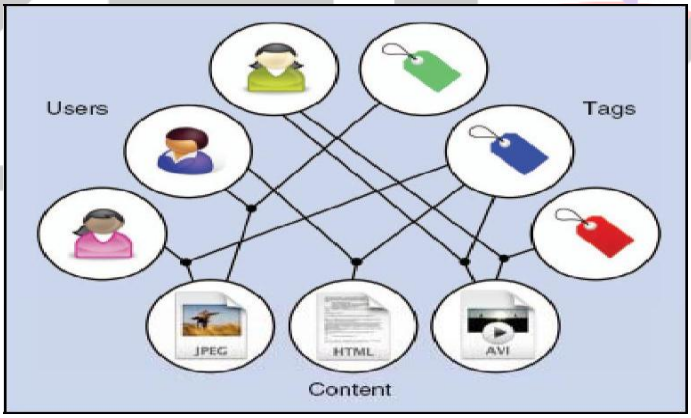

Fig. 1: General Model of a Social Tagging System is represented as a Tripartite Graph Structure

Therefore, we assigning free form tags to multimedia content has a risk that wrong or irrelevant tags. the Flicker Web site and brought out the tags providing by users are often imprecise and only around 50% of tags are truly related to an images. In the forms the tag content association to spam objects. The spam tags and content are in two popular social tagging systems.

III. PROPOSED SYSTEM

Trust provides a natural security policy specifies that users or content with low trust values should be investigated or eliminated. Trust can call the future behavior of users to keep off unwanted influences of untruth users. Trust based schemes can be used to motivate users to provide close moment and positive contrition to social network systems and punish resisters in it. The users in a social network can be used to represent the health of that network. In this article, we categorize trust modeling approaches into two classes according to the target of trust, i.e., user and content trust modeling. In this trust is tested based on image relevant analysis.

. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool.

Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

A. TRUST MODELING

When the information is exchanged on the Internet individuals are everywhere, to take advantage of the information exchange structure for their own benefit. While deviling and spamming others. Before social tagging became popular, spam content was observes in various domains:

1. Through e-mail.
2. Through Web search
3. Through Web search.
4. Through Peer to peer (P2P) networks.

We proposing, those are also influence by poisonous peers, and the various solutions based on trust and report, which carried on with collecting related data on stare behavior, scoring and position peers, and responding based on the scores. Like Amazon and Epinions, are facing more problems on this of unfair ratings by artificially amplifying or puncturing reputations.

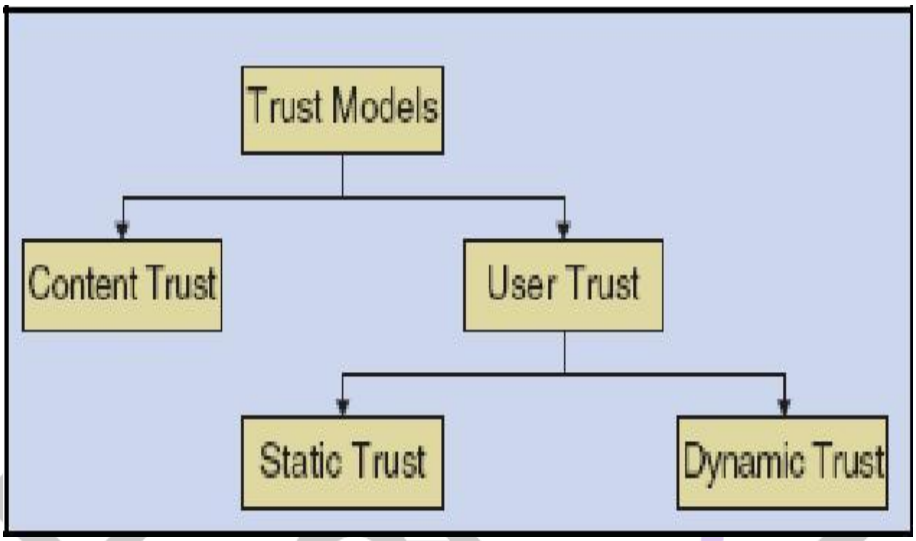

Fig. 2: Categorization of Trust Models Surveyed in this Article

B. CONTENT TRUST MODELING

Information trust modeling is used to classify data as spam or logically. In this case, the target is content and the trust score is given to each content based on its content and related tags. These models reduce the prominence of content likely spam, usually in query-based retrieval results. It was try to provide better sequence of the results to reduce the action of the spams to users. In the each incorrect content it found in a system could be simply removed by an DBA (Database administrator). The administrator step further and remove all content contributing by the user who posts the incorrect content, on the assumption that this user is a polluter. Trust Rank relies an important ethical theory scrutiny called approximate isolation. It starts from a set of seeds selected as highly capable, realistic, and popular Web pages in the mesh 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.

Although the aforementioned content trust modeling methods have shown to be effective in combating spam, the "subjectivity" in classifying spam and non spam content remains as a fundamental issues.

C. USER TRUST MODELING

The trust is given to each user based on the information extracts from a user's account by a person to person communication with other participants within the social network and the relationship between content and tags inside the social network system. Given a user trust score, the user might be drooped as a logical user or spammer. In central trust systems the user's trust models are maintained by one central authority for each user maintains own trust manager based on the previous interactions with other users. We consider information for each Investigating features user's profile, location, bookmarking activity, and context of tags. By making use of these features and SVM they were able to distinguish legitimate users from malicious ones. The aforementioned studies consider users' reliability as static at a specific moment. 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.

IV. EVALUATION AND PERFORMANCE

Data sets use for development and evaluation of trust modeling techniques have a wide range of diversity in terms of satisfied, numbers of possessions, tags and users, and kind of spam. Some researchers dealing with bookmarks uses in a public data set released by BibSonomy as a part of the ECML PKDD Discovery Challenge 2008 on Spam Detection in Social Bookmarking Systems. That is around 2,400 legitimate users and more than 29,000 spammers. Data provides in this data set can serve for the evaluation of both user, and content trust models. That data set since a majority of the bookmarks, 94% out of around 14 million, is spam. To model trust in other types of category systems, where spam is through videos, tweets, or user profiles.

Data is usually crowed from the corresponding social network, like Face Book, Twitter, respectively.

Trust modeling can be formulated as either a categorization problem or a ranking problem, depending on the way of handling. In the categorization problem, the results of an algorithm can be summarized by a uncertainty matrix from ground-truth data and predict labels, which contain the figure of accurate positives, accurate negative, fraud positive, and fraud negative.

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

where

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

V. CONTENT BASED IMAGE RETRIEVAL

In early days because of very large image collections the manual annotation approach was more difficult. In order to overcome these difficulties Content Based Image Retrieval (CBIR) was introduced. Content-based image retrieval (CBIR) is the application of computer vision to the image retrieval problem. In this approach instead of being manually annotated by textual keywords, images would be indexed using their own visual contents .The visual contents may be color, texture and shape. This approach is said to be a general framework of image retrieval .There are three fundamental bases for Content Based Image Retrieval which are visual excellence removal, multidimensional indexing and retrieval system design. The color aspect can be achieved by the techniques like averaging and histograms. The texture aspect can be achieved by using transforms or vector quantization .The shape aspect can be achieved by using gradient operators. Some of the major areas of application are Art collections, Medical analysis, Crime deterrence, armed thinker property, Architectural and engineering design and Geographical information and Remote sensing systems.

A. RETRIEVAL BASED ON COLOUR

Several methods for retrieving images on the basis of color similarity are being used. Each image added to the database is analyzed and a color histogram is computed which shows the proportion of pixels of each color within the image. Then this color histogram for each image is stored in the database. During the search time, the user can either specify the desired proportion of each color (75% olive green and 25% red, for example), or submit a reference image from which a color histogram is designed. The identical process then retrieves those images whose color histograms match those of the query most closely.

B. RETRIEVAL BASED ON STRUCTURE

The ability to match on texture similarity can often be useful in distinguishing between areas of images with parallel color. A multiplicity of techniques has been used for measuring texture similarity in which the best established rely on comparing values of what are known as second order statistics calculated from query and store images. Fundamentally, these compute the relative brightness of selected pairs of pixels as of each picture. From these it is possible to calculate measures of image texture such as the degree of distinction, stiffness, directionality and promptness, or periodicity, directionality and randomness.

Alternative methods of texture analysis for retrieval include the use of Gabor filters and fractals. A recent extension of the technique is the touch lexicon, which retrieves textured regions in images on the basis of similarity to automatically-derived code words representing important classes of texture within the collection.

C. RGB COLOUR MODEL

This model is represented in Cartesian coordinate system in which, three primary colours like red, green, blue occupy three corners of a cube, three secondary colours occupy three other corners of the cube. Black occupies the origin and white is at the corner farthest from black. Most of the CRT monitors and color raster graphics make use of the RGB color model. The colors in this replica are called “stabilizer primaries”, because desired colours can be produced by adding them together.

D. HSV COLOUR MODEL

HSV color model stands for Hue Saturation Value color model. This model describes colours in terms of their shades and brightness (Luminance).This model offers a more intuitive representation of relationship between colours. Basically a color model is the specification of organize system and a subspace within that, where each color is represented in single point. HUE represents the dominant colour in a colour object. It is expressed from 0º to 360º. It takes reference as red (starts at 0º), yellow (starts at 60º), green (starts at 120º), cyan (starts at 180º), blue (starts at 240º) and magenta (starts at 300º).Eventually all hues can be mixed from three basic hues known as primaries.

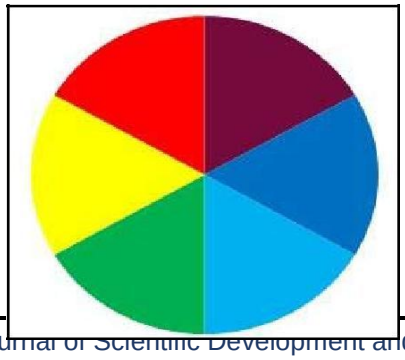

Fig. 3: Colours of Hue

Saturation represents the degree of diluteness of a pure color with white colour. It can also be thought as the intensity of the color. It is defined as the degree of purity of color. A highly saturated color is vivid, whereas a low saturated color is muted. When there is no saturation in the image, then the image is said to be a grey image. Value describes the brightness or intensity of the color. In other words value is defined as a relative lightness or darkness of color.

CONVERTING RGB TO HSV COLOR MODEL

HSV colours are said to lie within a triangle whose vertices are defined by the three primary colours in RGB space. The hue of the point P is given by the angle between the line connecting P to the center of the triangle and line connecting the RED point to the center of the triangle. The saturation of the point P is the vector between the point P and origin of the triangle. The value (intensity) of the point P is represented as height of the line perpendicular to the triangle and passing through its center. The grayscale points are situated onto the same line. Conversion formulas are given below as follows.

$$S = 1 - \frac{3}{R+G+B}[\min(R,G,B)]$$

$$\left(\sqrt{(R-G)^2 + (R-B)(G-B)}\right)$$

$$V = \frac{1}{3}(R+G+B)$$

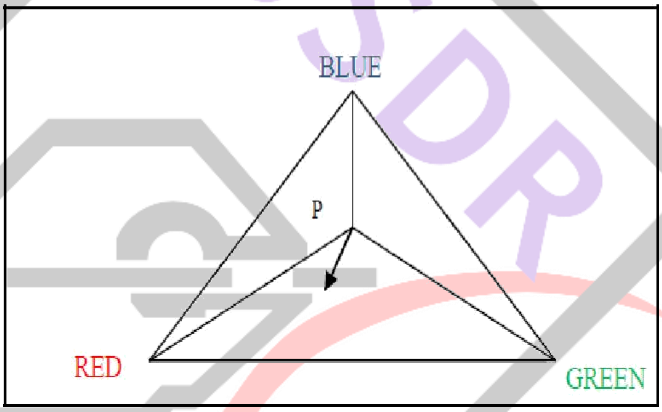

Fig. 4: RGB to HSV Conversion

Where R, G & B represents RED, GREEN & BLUE, H:Hue, S:Saturation,V:Value

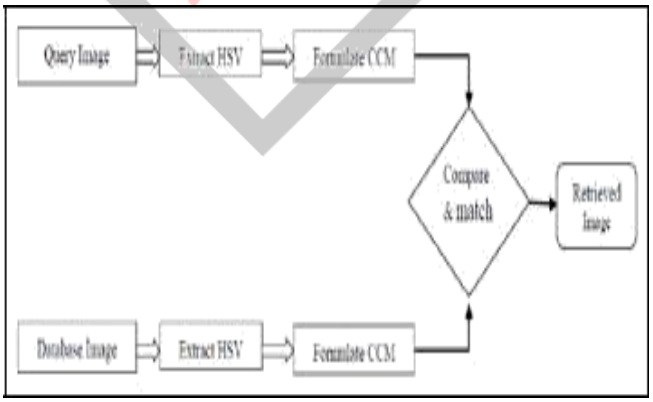

Fig. 5: CBIR Using HSV Model

E. COLOR CO-OCCURRENCE MATRIX (CCM)

A co-occurrence matrix is a matrix that is defined over an image to be the distribution of co-occurring values at a given counteract. Value of an image is originally the gray scale value of a specified pixel. In our case we take the values to be the Hue, saturation and values of a specified pixel. The co-occurrence matrix is mainly used to measure the texture of the image and hence it is used for texture analysis. Features generated using these techniques are also called as Haralick skin tone, because this concept was first introduced by Robert M Haralick. Texture measures like co-occurrence matrix and wavelet transforms have found applications in medical image analysis in particular. To carry out the retrieval process, a Query image is selected and the Hue, saturation and value of the pixels are taken and finally the color co-occurrence matrix if formed using the formula below.

Color co-occurrence matrix = 9H + 3S + V....

Here H, S and V write to Hue, Saturation and Value. The Hue values are from 20 to 316, saturation values are 0 to 1 and values ranges from 0 to 1.The CCM in the same way is found for the database images and the feature distinction is found using the Euclidean distance. The images are matched for similarity using both the textural features and HSV of the pixels of the image. This process provides accurate retrieval results so that the query image is retrieved.

F. EUCLIDEAN DISTANCE

Euclidean distance measures the similarity between the two different feature vectors. The formula for Euclidean distance is shown below. Q and D are feature vectors of the Query image and database image.

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}[Qi - Di]^2} \ldots.$$

VI. RETRIEVAL BASED ON SHAPE

The ability to retrieve by shape is perhaps the most obvious requirement at the prehistoric level. Unlike texture, shape is a fairly well-defined concept and there is considerable evidence that natural objects are primarily recognized by their outline. A number of skin texture characteristic of object shape (but independent of size or orientation) are computed for every object known within each stored image. Queries are then answered by computing the same set of features for the query image, and retrieving those stored images whose features most closely match those of the query. Two main types of shape feature are regularly used global features such as aspect ratio, circularity and instant invariants and local features such as sets of consecutive state line segments. Alternative methods planned for shape matching have incorporated elastic deformation of templates, comparison of directional histograms of edges extracted from the image, skeletal representations of object shape that can be compared using graph identical techniques. Queries to shape retrieval systems are formulated either by identifying an example image to act as the query, or as a user-drawn sketch. Shape matching of three-dimensional objects is a more challenging task particularly where only a single 2-D view of the object in question is available.

VII. RESULTS

Fig. 6: Start

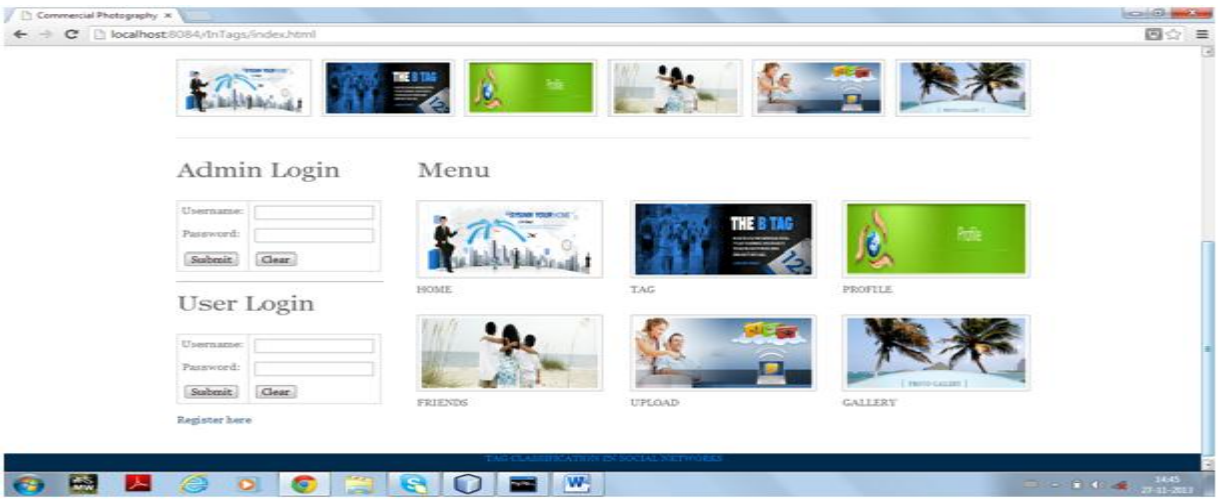

Fig. 7: Administration and user registration page

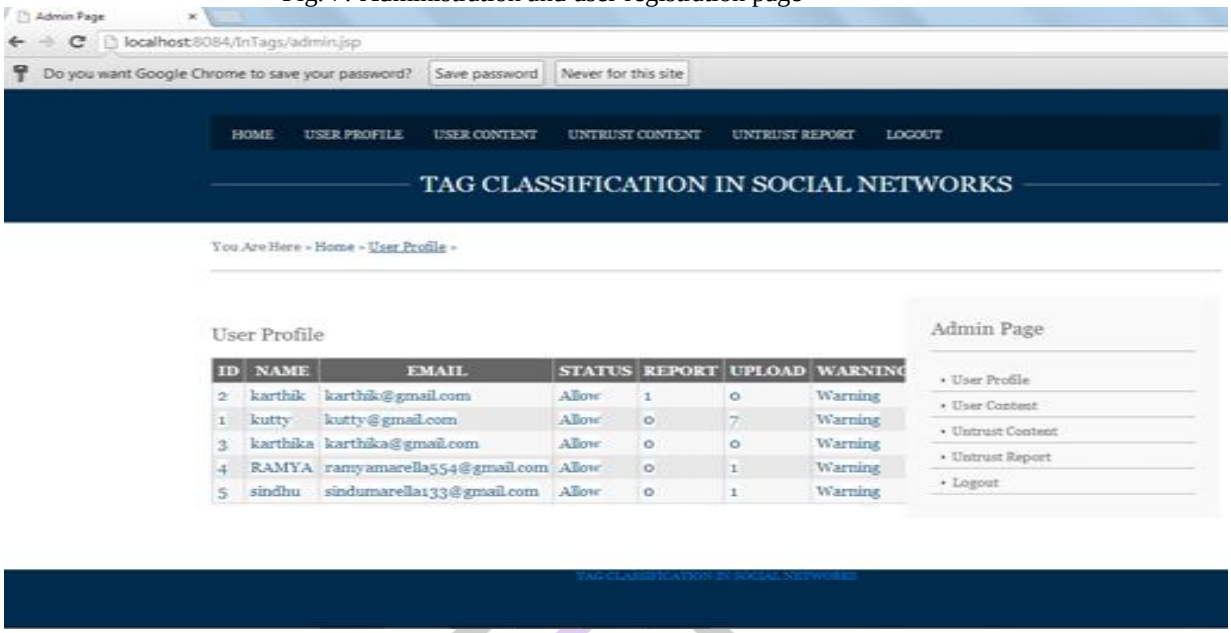

Fig. 8: User profiles

VIII. CONCLUSION

We proposes with one of the key issues in social tagging systems combating noise and spam in social networks. We relegate existing studies in the literature into two categories are analyze and compare. In addition, existing databases and evaluation protocols were are view. How trust modeling can be particularly employed in a popular application of image sharing and geo tagging. Open issues and future research trends were explored.

REFERENCES

[1] Wikimedia Foundation Inc. (2011, Dec.). Flicker. [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, M. Davis,"HT06, tagging paper, taxonomy, Flicker, academic article, to read", In Proc. ACM HT, Aug. 2006, pp. 31–40.
[4] K. Liu, B. Fang, 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, I. V. Kozintsev,"To search or to brand?: Predicting the performance of search-based automatic image classifiers", In Proc.ACM MIR, Oct. 006, pp. 249–258.
[6] P. Heymann, G. Koutrika, H. Garcia-Molina,"Fighting spam on social Websites: A survey of approach and future challenges", IEEE Internet Comput., Vol.11, No. 6, pp. 36– 45, Nov. 2007.
[7] Ritendra Datta, Dhiraj Joshi, Jia Li, James Z. Wang, "Image retrieval:Ideas, influences, and trends of the new age", ACMComputing Surveys, Vol. 40, pp. 1-60, 2008.
[8] Subrahmanyam Murala, R. P. Maheshwari, R. Balasubramanian,"Local Tetra Patterns: A New Feature Descriptor for Content-Based Image Retrieval", IEEE Trans. Image Process.