Privacy Policy Generation for User Uploaded Images on Content Sharing Sites using A3P Algorithm

Sumaiya Banu V I[1], Dr. A. Rajesh[2], Feroz Khan A B[3]

[1]Sumaiya Banu V I, M. E Student/Computer Science and Engineering/C. Abdul Hakeem College of Engineering & Technology, Melvisharam, Vellore, India

[1]sumaiyabanuit@gmail.com

[2] Dr. A. Rajesh, Professor/Computer Science and Engineering/C. Abdul Hakeem College of Engineering & Technology, Melvisharam, Vellore, India

[2]a-rajesh@cahcet.edu.in

[3] Feroz Khan A B, Assistant Professor/Master of Computer Applications/C. Abdul Hakeem College of Engineering & Technology, Melvisharam, Vellore, India

[3]abferozkhan@gmail.com

ABSTRACT

With the increasing volume of images users share through social sites, maintaining privacy has become a major problem. There will be a need of tools to help users control access to their shared content is apparent. On focusing of this need, we propose an Adaptive Privacy Policy Prediction (A3P) system to help users compose privacy settings for their images. We examine the role of social context, image content, and metadata as possible indicators of users' privacy preferences. We propose a two-level framework which according to the user's available history on the site, determines the best available privacy policy for the user's images being uploaded. Our solution relies on an image classification framework for image categories which may be associated with similar policies, automated image annotation and on a policy prediction algorithm to automatically generate a policy for each newly uploaded image, also according to users' social features. We provide the results of our extensive evaluation over 5,000 policies, which demonstrate the effectiveness of our system, with prediction accuracies over 90 percent.

Index Terms—Online information services, web-based services

1. INTRODUCTION

Now a days people are sharing their images on content sharing sites with people within their social circles (e. g., Google+, Flickr, Facebook, Picasa etc.,), and also with people outside the users social circles, for purposes of social discovery to help them identify new peers and learn about peers interests and social surroundings. However, semantically rich images may reveal content-sensitive information. Consider a photo of a student's 2012 graduation ceremony, for example. It could be shared within a Google+ circle or Flickr group, but may unnecessarily expose the students' family members to other friends. Sharing images within online content sharing sites lead to unwanted disclosure and privacy violations. Also, the persistent nature of online media makes it possible for other users to collect rich aggregated information about the owner of the published content and the subjects in the published content. The aggregated information can result in unexpected exposure of one's social environment and lead to abuse or misuse of one's personal information.

The content sharing websites allow users to enter their privacy preferences. But the users struggle to set up and maintain such privacy settings. One of the main reasons provided is that given the amount of shared information this process can be tedious and error-prone.

In this paper, we propose an Adaptive Privacy Policy Prediction (A3P) system which aims to provide users a hassle free privacy settings experience by automatically generating privacy policies.

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

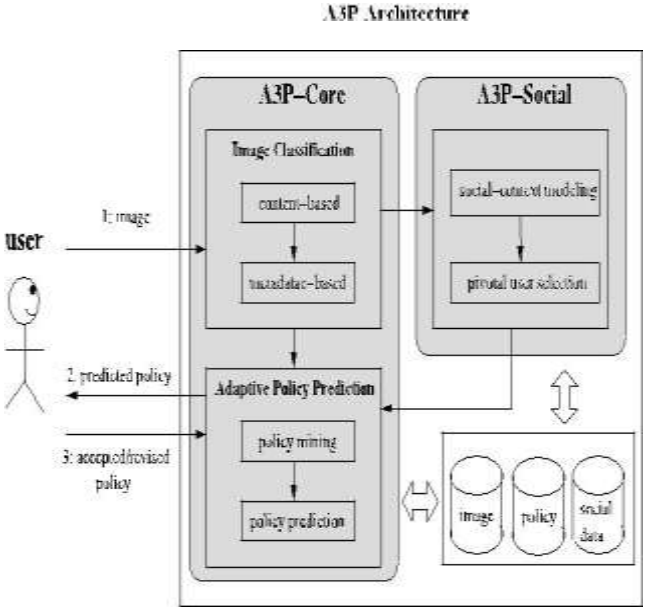

Fig - 1: System overview.

The A3P system handles user uploaded images based on the following criterion:

The impact of social environment and personal characteristics. Social context of users, such as their profile information and relationships with others may provide useful information regarding users' privacy preferences. For example, users interested in photography may like to share their photos with other photographers. Users who have several family members among their social contacts may share with them pictures related to family events.

The role of image's content and metadata. In general, similar images often incur similar privacy preferences, especially when people appear in the images. For example, one may upload several photos of his kids and specify that only his family members are allowed to see these photos. He may upload some other photos of landscapes which he took as a hobby and for these photos, he may set privacy preference allowing anyone to view and comment the photos. Analyzing the visual content may not be sufficient to capture users' privacy preferences. Tags and other metadata are indicative of the social context of the image, including where it was taken and why , and also provide a synthetic description of images, complementing the information obtained from visual content analysis.

2. LITERATURE REVIEW

Jonathan Anderson proposed a paradigm called **Privacy Suites [4]** which allows users to easily choose "suites" of privacy settings. A privacy suite can be created by an expert using privacy programming. Privacy Suites could also be created directly through existing configuration UIs or exporting them to the abstract format. The privacy suite is distributed through existing distribution channels to the members of the social sites. The disadvantage of a rich programming language is less understandability for end users.

Fabeah Adu-Oppong developed privacy settings based on the concept of **social circles [5]**. It provides a web based solution to protect personal information. The technique named Social Circles Finder, automatically generates the friend's list. It is a technique that analyses the social circle of a person and identifies the intensity of relationship and therefore social circles provide a meaningful categorization of friends for setting privacy policies. The application will identify the social circles of the subject but not show them to the subject. The subject will then be asked questions about their willingness to share a piece of their personal information. Based on the answers the application finds the visual graph of users.

Kambiz Ghazinour designed a recommender system known as **YourPrivacyProtector [5]** that understands the social net behavior of their privacy settings and recommending reasonable privacy options. It uses user's personal profile, User's interests and User's privacy settings on photo albums as parameters and with the help of these parameters the system constructs the personal profile of the user. It automatically learned for a given profile of users and assign the privacy options. It allows users to see their current privacy settings on their social network profile, namely Facebook, and monitors and detects the possible privacy risks.

Alessandra Mazzia introduced **PViz Comprehension Tool [1]**, an interface and system that corresponds more directly with how users model groups and privacy policies applied to their networks. PViz allows the user to understand the visibility of her profile according to automatically-constructed, natural sub-groupings of friends, and at different levels of granularity. Because the user must be able to identify and distinguish automatically-constructed groups, we also address the important sub-problem of producing effective group labels.

Peter F. Klemperer developed **a tag based access control of data [8]** shared in the social media sites. A system that creates access-control policies from photo management tags. Every photo is incorporated with an access grid for mapping the photo with the participant's friends. The participants can select a suitable preference and access the information. Photo tags can be categorized as organizational or communicative based on the user needs. There are several important limitations to our study design. First, our results are limited by the participants we recruited and the photos they provided. A second set of limitations concerns our use of machine generated access-control rules. The algorithm has no access to the context and meaning of tags and no insight into the policy the participant intended when tagging for access control. As a result, some rules appeared strange or arbitrary to the participants, potentially driving them toward explicit policy-based tags like "private" and "public.

Ching-man Au Yeung propose a access control system based on a **decentralised authentication protocol [11]**, descriptive tags and linked data of social networks in the Semantic Web. It allows users to create expressive policies for their photos stored in one or more photo sharing sites, and users can specify access control rules based on open linked data provided by other parties.

Sergej Zerr propose a technique **Privacy-Aware Image Classification and Search [10]** to automatically detect private images, and to enable privacy-oriented image search. It combines textual meta data images with variety of visual features to provide security policies. In this the selected image features (edges, faces, color histograms) which can help discriminate between natural and man-made objects/scenes (the EDCV feature) that can indicate the presence or absence

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

of particular objects (SIFT). It uses various classification models trained on a large scale dataset with privacy assignments obtained through a social annotation game.

Anna Cinzia Squicciarini developed an **Adaptive Privacy Policy Prediction (A3P) [3]** system, a free privacy settings system by automatically generating personalized policies. The A3P system handles user uploaded images based on the person's personal characteristics and images content and metadata. The A3P system consists of two components: A3P Core and A3P Social. When a user uploads an image, the image will be first sent to the A3P-core. The A3P-core classifies the image and determines whether there is a need to invoke the A3P-social. The disadvantage is inaccurate privacy policy generation in case of the absence of metadata information about the images.

3. PROPOSED FRAMEWORK

3.1 A3P Framework

Users can express their privacy preferences about their content disclosure preferences with their socially connected users via privacy policies. We define privacy policies according to Definition 1.

Definition 1

A privacy policy P of user u consists of the following components:

- Subject (S): A set of users socially connected to u.
- Data (D): A set of data items shared by u.
- Action (A): A set of actions granted by u to S on D.
- Condition (C): A Boolean expression which must be satisfied in order to perform the granted actions.

In the definition, users in S can be represented by their identities, roles (e.g., family, friend, coworkers), or organizations (e.g., non-profit organization, profit organization). D will be the set of images in the user's profile. Each image has a unique ID along with some associated metadata like tags "vacation", "birthday". Images can be further grouped into albums. As for A, we consider four common types of actions: {view, comment, tag, download}. Last, the condition component C specifies when the granted action is effective. C is a Boolean expression on the grantees' attributes like time, location, and age. For better understanding, an example policy is given below.

Example - 1 Alice would like to allow her friends and coworkers to comment and tag images in the album named "vacation album" and the image named "summer.jpg" before year 2012. Her privacy preferences can be expressed by the following policy:

P :[{friend, coworker}, {vacation_album, summer.jpg}, {comment, tag}, (date < 2012)].

3.2 System Overview

The A3P system consists of two main components: **A3P-core and A3P-social**. When a user uploads an image, the image will be first sent to the A3P-core. The A3P-core classifies the image and determines whether there is a need to invoke the A3P-social. In most cases, the A3P-core predicts policies for the users directly based on their historical behavior. If one of the following two cases is verified true, A3P-core will invoke A3P-social: (i) The user does not have enough data for the type of the uploaded image to conduct policy prediction; (ii) The A3P-core detects the recent major changes among the user's community about their privacy practices along with user's increase of social networking activities (addition of new friends, new posts on one's profile etc). The A3P-social groups users into social communities with similar social context and privacy preferences, and continuously monitors the social groups. When the A3P-social is invoked, it automatically identifies the social group for the user and sends back the information about the group to the A3P-core for policy prediction. At the end, the predicted policy will be displayed to the user. If the user is fully satisfied by the predicted policy, he or she can just accept it. Otherwise, the user can choose to revise the policy. The actual policy will be stored in the policy repository of the system for the policy prediction of future uploads.

3.3 Automated Annotation

The system carries two major tasks.

- Automatic image annotation
- Annotation based image retrieval

Automatic image annotation phase makes use of a manually annotated training set taken to generate an annotated image database. Annotation based image retrieval phase gets a user query, and then find similar terms for the query with the help of WordNet. Also discover the similarity between the query and images in annotated image database. Then find the similarity between matching images. To annotate the images in database, features such as Color and texture feature are extracted by using Color Histogram and SIFT Descriptors methods.

3.3.1 Color Histogram Feature

Color histogram is simplest and most frequently used to represent color. The color histogram serves as an effective representation of the content. Color is one of the most important features of images. Color features are defined subject to a particular color space or model. A number of color spaces have been used such as RGB, LUV, and HSV. Once the color space is specified, color feature can be extracted from images or regions. An important color features namely color histogram is extracted. Color histograms are frequently used to compare images. In this gray level variations are used to compute the histogram of any image. For this purpose the color image is first converted in to gray level image. Then the histogram values are computed for gray level variations. According to histogram values, images are extracted from the database.

3.3.2 Texture Feature Extraction

Texture feature are extracted by using SIFT (Scale-invariant feature transform) descriptor. Scale-invariant feature transform (or SIFT) is an algorithm in computer vision to detect and describe texture features in images.

3.3.3 SIFT Descriptors

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

SIFT based analysis involves detecting salient locations in an image and extracting descriptors that are distinctive yet invariant to changes in viewpoint, illumination, etc. The standard SIFT interest point detector and the standard SIFT histogram-of-gradients descriptor can be used. These 128 dimension descriptors can be thought of roughly as summarizing the edge information in an image patch centered at an interest point. We term the 128 dimension descriptors the local SIFT descriptors for an image. We also compute a single global SIFT descriptor. This global descriptor is a frequency count of the quantized local descriptors. We use the clustering algorithm to cluster a large collection of SIFT descriptors and label each local descriptor with the id of the closest cluster center.

3.4 A3P-Core

There are two major components in A3P-core: (i) Image classification and (ii) Adaptive policy prediction. For each user, his/her images are first classified based on content and metadata. Then, privacy policies of each category of images are analyzed for the policy prediction.

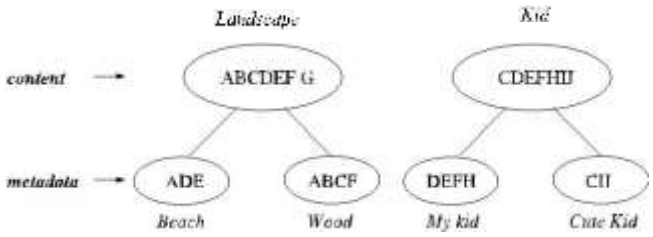

Fig – 2: Two-level Image classification.

3.4.1 Image Classification

To obtain groups of images that may be associated with similar privacy preferences, we propose a hierarchical image classification which classifies images first based on their contents and then refine each category into subcategories based on their metadata. Images that do not have metadata will be grouped only by content. Such a hierarchical classification gives a higher priority to image content and minimizes the influence of missing tags.

Fig. 2 shows an example of image classification for 10 images named as A, B, C, D, E, F , G, H, I, J, respectively. The content-based classification creates two categories: “landscape” and “kid”. Images C, D, E and F are included in both categories as they show kids playing outdoor which satisfy the two themes: “landscape” and “kid”. These two categories are further divided into subcategories based on tags associated with the images. As a result, we obtain two subcategories under each theme respectively. The image G is not shown in any subcategory as it does not have any tag; image A shows up in both subcategories because it has tags indicating both “beach” and “wood”.

3.4.1.1 Content-Based Classification

The classification algorithm (Decision Tree Classification Algorithm) compares image signatures defined based on quantified and sanitized version of Haar wavelet transformation. For each image, the wavelet transform encodes frequency and spatial information related to image color, size, invariant transform, shape, texture, symmetry, etc. Then, a small number of coefficients are selected to form the signature of the image. The content similarity among images is then determined by the distance among their image signatures.

When a user uploads an image, it is handled as an input query image. The signature of the newly uploaded image is compared with the signatures of images in the current image database. To determine the class of the uploaded image, we find its first m closest matches. The class of the uploaded image is then calculated as the class to which majority of the m images belong. If no predominant class is found, a new class is created for the image. Later on, if the predicted policy for this new image turns out correct, the image will be inserted into the corresponding image category in our image database, to help refine future policy prediction.

3.4.1.2 Metadata-Based Classification

The metadata-based classification groups images into subcategories. The process consists of three main steps.

The first step is to extract keywords from the metadata associated with an image. The metadata considered in our work are tags, captions, and comments. We identify all the nouns, verbs and adjectives in the metadata and store them as metadata vectors.

The second step is to derive a representative hypernym (denoted as h) from each metadata vector. For example, consider a metadata vector t ¼ f“cousin”,“first steps”, “baby boy”. We find that “cousin” and “baby boy” have the same hypernym “kid”, and “first steps” has a hypernym “initiative”. Corre-spondingly, we obtain the hypernym list h = f(kid, 2), (initiative, 1). In this list, we select the hypernym with the highest frequency to be the representative hypernym, e.g., “kid”. In case that there are more than one hypernyms with the same frequency, we consider the hypernym closest to the most relevant baseline class to be the representative hypernym. For example, if we have a hypernym list h = f(kid, 2), (cousin, 2), (initiative, 1), we will select “kid” to be the representative hypernym since it is closest to the baseline class “kids”.

The third step is to find a subcategory that an image belongs to. This is an incremental procedure. At the beginning, the first image forms a subcategory as itself and the representative hypernyms of the image becomes the subcategory’s representative hypernyms. Then, we compute the distance between representative hypernyms of a new incoming image and each existing subcategory.

3.5 Adaptive Policy Prediction

The policy prediction algorithm provides a predicted policy of a newly uploaded image to the user for his/her reference. The prediction process consists of three main phases: (i) policy normalization; (ii) policy mining; and (iii) policy prediction.

3.5.1 Policy Normalization

The policy normalization is a simple decomposition process to convert a user policy into a set of atomic rules in which the data (D) component is a single-element set.

3.5.2 Policy Mining

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

We propose a hierarchical mining approach for policy mining. Our approach leverages association rule mining techniques to discover popular patterns in policies. Policy mining is carried out within the same category of the new image because images in the same category are more likely under the similar level of privacy protection. The basic idea of the hierarchical mining is to follow a natural order in which a user defines a policy. Given an image, a user usually first decides who can access the image, then thinks about what specific access rights (e.g., view only or download) should be given, and finally refine the access conditions such as setting the expiration date.

3.5.3 Policy Prediction

The policy mining phase may generate several candidate policies while the goal of our system is to return the most promising one to the user. Thus, we present an approach to choose the best candidate policy that follows the user's privacy tendency.

To model the user's privacy tendency, we define a notion of strictness level. The strictness level is a quantitative metric that describes how "strict" a policy is. In particular, a strictness level L is an integer with minimum value in zero, wherein the lower the value, the higher the strictness level. It is generated by two metrics: major level (denoted as l) and coverage rate (a), where l is determined by the combination of subject and action in a policy, and (a) is determined by the system using the condition component.

3.6 A3P-Social

The A3P-social employs a multi-criteria inference mechanism that generates representative policies by leveraging key information related to the user's social context and his general attitude toward privacy. A3P-social will be invoked by the A3P-core in two scenarios. One is when the user is a newbie of a site, and does not have enough images stored for the A3P-core to infer meaningful and customized policies. The other is when the system notices significant changes of privacy trend in the user's social circle, which may be of interest for the user to possibly adjust his/her privacy settings accordingly.

3.6.1 Modeling Social Context

We observe that users with similar background tend to have similar privacy concerns, as seen in previous research studies and also confirmed by our collected data. This observation inspires us to develop a social context modeling algorithm that can capture the common social elements of users and identify communities formed by the users with similar privacy concerns. The identified communities who have a rich set of images can then serve as the base of subsequent policy recommendation. The social context modeling algorithm consists of two major steps. The first step is to identify and formalize potentially important factors that may be informative of one's privacy settings. The second step is to group users based on the identified factors.

3.6.2 Identifying Social Group

We now introduce the policy recommendation process based on the social groups obtained from the previous step. Suppose that a user U uploaded a new image and the A3P-core invoked the A3P-social for policy recommendation. The A3P-social will find the social group which is most similar to user U and then choose the representative user in the social group along with his images to be sent to the A3P-Core policy prediction module to generate the recommended policy for user U. Given that the number of users in social network may be huge and that users may join a large number of social groups, it would be very time consuming to compare the new user's social context attributes against the frequent pattern of each social group. In order to speed up the group identification process and ensure reasonable response time, we leverage the inverted file structure to organize the social group information. The inverted file maps keywords (values of social context attribute) occurring in the frequent patterns to the social groups that contain the keywords. Specifically, we first sort the keywords (except the social connection) in the frequent patterns in an alphabetical order. Each keyword is associated with a link list which stores social group ID and pointers to the detailed information of the social group. Next, given a new user, we search his/her attribute values in the inverted file and obtain a set of candidate social groups. We also count the number of occurrence of the candidate groups during the search. We select the candidate group with the highest occurrence as the social group for the new user. In the identified social group, we further examine its subgroups by comparing the strictness levels of the sub-groups with the new user's preferred privacy strictness level if provided. We select the sub-group whose strictness level matches the new user's privacy requirements best. If the new user did not specify privacy preference, we select the sub-group with the largest members. Then, in this selected sub-group, we look for the user who is most similar to the new user. We just need to compare the new users and the group members' remaining attributes that are not included in the frequent pattern. The selected user and his/her images and policies are sent to the A3P-Core module to generate the recommended policy for the new user. Finally, we update the social group information by including the new user as a probational member. The probational member will not be chosen by A3P-Social module to until he/she uploaded sufficient images and becomes a regular member.

4. IMPLEMENTATION DETAILS

This study involved 88 participants (48 female and 40 males) who were recruited from a large US university community (staff, students, and the community at large). Their average age is 26.3 years old (Range: 18-39). The participants completed at least 90 percent of the ques -tionnaire consisting of two parts. The first part contains questions related to one's background information and online privacy practices and the second part is to collect user-specified policies.

In the first part of the questionnaire, the participants were asked to indicate any social networks they were a part of (98 percent indicated Facebook and 37 percent also indicated others like Myspace). In terms of usage frequency, 95 percent of the respondents accessed social network sites at least once a week, with 76 percent of reporting that they were daily users.

We also asked participants if they have had concerns about their privacy due to shared images. Over 51 percent of the participants indicated that they had privacy concerns. Users also reported that image content is an important factor when

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED

determining privacy settings for an image with 87 per-cent of people agreeing or strongly agreeing with the statement "When I set privacy settings for a certain image I usually think about the content of the image", and over 91 percent of users agreeing or strongly agreeing with the statement "The content of an image determines whether I upload the image to a social network site." Surprisingly, however, many users indicated that they never changed privacy settings for images (38 percent) or changed their set-tings only 1 or 2 times (36 percent) since joining the social network. There seems to be a clear disconnect between users privacy inclinations and their practice of setting privacy policies. The possible reason could be "Changing privacy set-tings for every image uploaded on a social site can be very time consuming", as strong agreed or agreed by 70 percent of users.

In the second part of the questionnaire, we presented each user 30 images selected by us. For each image, we asked the user to input the privacy settings by assuming these photos as his/her own images. We collected around 3,000 policies.

We invited another 41 people to use our A3P system. The goal of this experiment is to assess our system's acceptability, i.e., whether users would consider the predicted policies reasonable, and inline with their over-all preferences. We asked participants to input policies for a few images at first for training purposes. Three images from a given class are sufficient to bootstrap the algorithm. Next, participants enter privacy settings for a set of images that they would upload in their fictitious profile. Upon showing the image, privacy settings for it are suggested to the user. The participant has the option to accept the predicted policy as is, revise some components of it, or disagree with the predicted result and e re-enter preferred settings.

5. CONCLUSION

We have proposed an Adaptive Privacy Policy Prediction (A3P) system that helps users automate the privacy policy settings for their uploaded images. The A3P system provides a comprehensive framework to infer privacy preferences based on the information available for a given user.

REFERENCES

[1] Alessandra Mazzia Kristen LeFevre and Eytan Adar, The PViz Comprehension Tool for Social Network Privacy Settings, Tech. rep.University of Michigan, 2011.

[2] Anna Cinzia Squicciarini, "Privacy Policy Inference of User-Uploaded Images on Content Sharing Sites", IEEE Transactions On Knowledge And Data Engineering, vol. 27, no. 1, January 2015.

[3] J. Bonneau, J. Anderson, and L. Church, "Privacy suites: Shared privacy for social networks," in Proc. Symp. Usable Privacy Security, 2009.

[4] Kambiz Ghazinour, Stan Matwin and Marina Sokolova, "Yourprivacyprotector: A Recommender System For Privacy Settings In Social Networks", International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 2, No 4, August 2013.

[5] A. Kapadia, F. Adu-Oppong, C. K. Gardiner, and P. P. Tsang, "Social circles: Tackling privacy in social networks," in Proc. Symp. Sable Privacy Security, 2008.

[6] Karthikram, Mailaivasan, Parthiban and Ganesh, "Tag Based Image Retrieval (TBIR) Using Automatic Image Annotation", IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

[7] Peter F. Klemperer, Yuan Liang, Michelle L. Mazurek, "Tag, You Can See It! Using Tags for Access Control in Photo Sharing", Conference on Human Factors in Computing Systems, May 2012.

[8] Sangeetha. J et al, An Improved Privacy Policy Inference over the Socially Shared Images with Automated Annotation Process, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3166-3169.

[9] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, Elena Demidova , I Know What You Did Last Summer!:Privacy-Aware Image Classification and Search , Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012.

[10] C. A. Yeung, L. Kagal, N. Gibbins, and N. Shadbolt, "Providing access control to online photo albums based on tags and linked data," in Proc. Soc. Semantic Web: Where Web 2.0 Meets Web 3.0 at the AAAI Symp., 2009, pp. 9–14.

 COPYRIGHT © 2016 IJREST, ALL RIGHT RESERVED