An Integration Module for Mining Textual and Visual Social Media Data

Mohammed Ali Eltaher
Dept. of Computer Science, Faculty of Education-Ubari
University of Sebha
Sebha, Libya
E-mail: bineltaher@gmail.com

***Abstract*—In social media networks, visual data, such as images, co-exist with text or other modalities of information. In order to benefit from different data modalities, further research is obligatory. This paper introduces the first step towards an integration module, which is based on visual and textual data available in social media. We utilized tags and images of users with a novel approach of information fusion in order to enhance the social user mining. Here, two different approaches were applied to enhance social user mining: (1) content based image fusion, and (2) semantic based image fusion. Our approaches were applied to a gender classification mining application and showed the performance of our methods to discover unknown knowledge about the user.**

Keywords- social media mining; content based integration; semantic based integration.

I. INTRODUCTION

By combining features of image and textual attributes that are generated by the user, interesting properties of social user mining are revealed. These properties serve as a powerful tool for discovering unknown information about the user. However, there is a minimum amount of research reported on the combination of images and texts for social user mining.

Textual and visual information are rich sources of social user mining. It is highly desired to integrate one media with another for better accuracy. Gallagher et al. [1] use both textual and visual information to find geographical locations of Flickr images. Their model builds upon the fact that visual content and user tags of a picture can be useful to find the geo-location. Another example, in order to determine the gender of a user, a profile image and its description can be more effective than a single media, such as image itself or description only. The progress of data mining techniques makes it possible to integrate different data types in order to improve the mining tasks of social media, and thus make them more effective.

Labeling the semantic content of multimedia objects such as images with a set of keywords is known as image tagging. More details about different types of image tagging can be found in [2]. The social user mining task mainly depends on the availability and quality of tags. Furthermore, a semi-automatic tagging process, that helps to tag multimedia objects, would improve the quality of tagging and thus the overall social user mining process.

As a mining application, the gender classification problem in Flickr is one of the applications that our approach addresses. Popescu and Grefenstette [3] introduced a gender identification technique for Flickr's users based only on tags. However, the existing studies show that tags are few, impressive, ambiguous, and overly personalized [4]. In addition, recent studies reveal that users do annotate their photos with the motivation to make them better accessible to the general public [5]. In our approach, we apply tags and images to address the gender classification problem. For a user u, given his d_u (tags and images) from Flickr, we predict the gender of u based on tags, images, and a combination of both tags and images.

In this paper, we propose a novel approach for integrating Flickr's data by combining multiple types of features. We utilize tags and images of users by using two different approaches to enhance social user mining: (1) content based image fusion, and (2) semantic based image fusion. Our approaches were applied to the gender classification problem. For the classifier, we use a Naive Bayes algorithm with multinomial distributed data, where the integrated data are typically represented as feature vector, as well as Support Vector Machine (SVM). In order to evaluate the proposed algorithm, we downloaded 148,511 users profile information with 300 tags and up to 50 photos for each user from Flickr.com.

The rest of this paper is organized as follows. Section II describes the novel approach of information fusion by combined textual and visual information using image contents. Section III describes the second approach for semantic based image fusion, using a semi-automatic image tagging system. Section IV addresses the classification algorithms and the experimental result. Finally, Section V concludes the paper.

II. CONTENT BASED IMAGE FUSION

Through the content based image fusion, we combined textual and visual information by using image contents. We proposed a data integration method between the user's tags and image contents. For the image contents, we used a hue histogram and a hue in bag of words. We implemented the tags with hue histogram as a feature vector, as well as tags with hue in bag of words. Figure 1 shows the proposed module of content based image fusion.

 ISBN: 978-1-61208-681-1

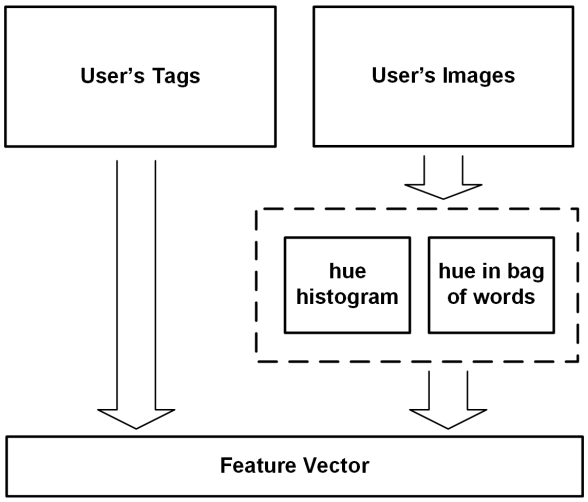

Figure 1. Content based image fusion

A. Integrated Data Units

Flickr allows their users to annotate their photos with textual labels called "tags". In many social media sites, tags are accurate descriptors of content, and could be used in many mining applications [6]. This section captured the content of users' Flickr photos through the user-generated tags. The first element of our content based data integration module is the user's tags. For each user u, we utilized up to 300 tags T_u, whereby each tag is represented as a word denoted as:

$T_u = \langle t_1, t_2, \dots, t_n \rangle$, where n is the number of tags for each user.

For the second element of the proposed content based data integration module, we use the user's images as visual data. Specifically, we represent the images through two features known as hue histogram and hue in bag of words. Hue histogram is based on the hue value of all the pixels in the user's images. Hue, Saturation, Value (HSV) and Hue, Saturation, Lightness/Luminance (HSL) represent other color models that are used in multimedia mining. In HSV, the brightness of a pure color is equal to the brightness of white. In HSL the lightness of a pure color is equal to the lightness of a medium gray. Each hue value in the HSL or the HSV color space represents an individual color. For the hue in bag of words feature, we selected the top two colors by assigning "1" to the feature value for these top colors and "0" to the others. The hue histogram and hue in bag of words can be denoted as follows:

$HS_u = \langle hs_1, hs_2, \dots, hs_n \rangle$,

$HBW_u = \langle hbw_1, hbw_2, \dots, hbw_n \rangle$, where n represents the number of colors for each user u.

B. Integration Scheme

We continued to implement the users' tags with the hue histogram and the hue in bag of words as a feature vector. Figure 2 shows the scheme of the content based image fusion. For the tag features, each user u has a feature vector F_t that corresponds to all the users' tags. This feature vector can be defined as:

$F_t = \langle T_u \rangle$, where T_u is the users' tags.

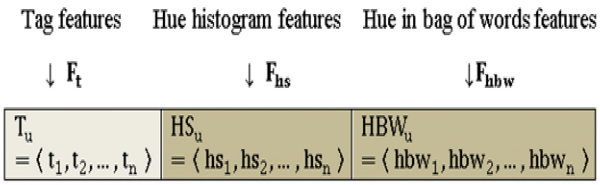

Figure 2. Feature vector of content based data integration

For the hue histogram, each user had up to 50 images. We calculated the hue histogram for each user based on their images, and determined the average based on 50 images for each color. For the hue in bag of words, we selected the top two colors for each user based on the images. The feature vectors of the hue histogram and the hue in bags of words can be defined as:

$F_{hs} = \langle HS_u \rangle$, where HS_u is hue histogram per user.

$F_{hbw} = \langle HBW_u \rangle$, where HBW_u is a hue in bag of words per user.

III. SEMANTIC BASED IMAGE FUSION

In this section, we address the problem of integrating textual and visual data semantically to perform the social user mining tasks. We determined that the integration of the two data types will be more beneficial than using an individual data type. We proposed a data integration module that combined both textual and visual information. First, we applied a semi-automatic image tagging system called *Akiwi* to suggest keywords for images. *Akiwi* uses an enormous collection of 15 million images tagged with keywords. Basically, *Akiwi* retrieves images that are visually very similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image. Then, we integrated these keywords for individual users' tag. Figure 3 illustrates the data integration module for the sematic based image fusion.

 ISBN: 978-1-61208-681-1

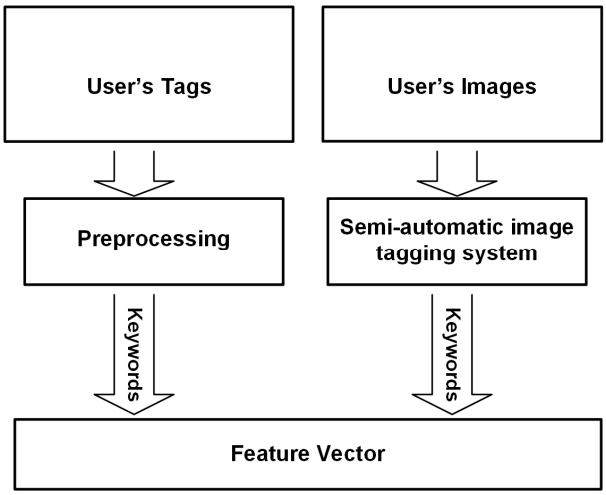

Figure 3. Semantic based image fusion

A. Integrated Data Units

Social media networks provide their users with the ability to describe any photos' contents by manually annotating the photos. Similar to the above represented content based integration module, our first element of the semantic based data integration module is users' tags. Some users' tags are unreliable due to excess noise in tags provided by users. These tags prove to be irrelevant or incorrectly spelled.

For example, only about 50% of the tags provided by Flickrs' users are in fact related to the images [7]. Due to tagging inaccuracies, we use a semi-automatic image tagging system to suggest keywords for our images. These keywords are considered as the second element of our proposed semantic based data integration module. We applied *Akiwi* to suggest keywords for the images. *Akiwi* uses an enormous gathering of 15 million images tagged with keywords. Essentially, *Akiwi* retrieves images that are visually precise similar to the query image. Based on the keywords of these images, *Akiwi* tries to predict the keywords for the unknown image.

B. Integration Scheme

For the semantic based data integration module, we implemented the users' tags with the keywords retrieved from *Akiwi* as a feature vector. The main difference in this module focuses on tags and keywords generated by users, as opposed to keywords generated by *Akiwi*. Figure 4 shows the scheme of the semantic based image fusion. These feature vectors of tags and keywords for each user can be defined as:

- F_t=<T_u>, where T_u is users' tags.
- F_k=<K_u>, where K_u is users' keywords.

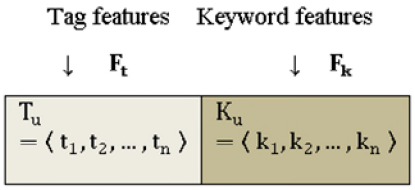

Figure 4. Feature vector of semantic based data integration

IV. CLASSIFICATION ALGORITHMS

Generally, there are various mining techniques that can be used in evolving social user mining. To apply our approaches for the gender classification problem, we selected two popular classifiers: the Naive Bayes and the SVM.

The Naive Bayes classifier is one of the most efficient and effective inductive learning algorithms for machine learning and data mining [8]. Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. In this experiment, we adopted a multinomial Naive Bayes model. This model implements the Naive Bayes algorithm for multinomial distributed data, where the data is typically represented as vector. Given the gender classification problem having G classes $\{g_1, g_2\}$ with probabilities $P(g_1)$ and $P(g_2)$, we assign the class label G to a Flickr user u based on the feature vector $D_u = (d_1, d_2, \ldots, d_N)$, where d_N represent the user data, such as tags and images:

$$g = arg \max_g P(g|D_u) \quad (1)$$

The above equation is to assign a class with the maximum probability given the feature vector of user data D_u. This probability can be formulated by using Bayes theorem as follows:

$$P(g|D_u) = \frac{P(g) \times \prod_{i=1}^{N} P(d_i|g)}{P(D_u)} \quad (2)$$

Here, the objective is to predict the most probable class to the user u giving the feature vector D_u that contains N features to the most possible class.

The SVM is a popular machine learning method for classification and other learning tasks [9]. In our experiment, we adopted the C-Support Vector Classification (SVC), which is implemented based on libsvm [10]. The main idea of applying SVM on classification is to find the maximum-margin hyperplane to separate classes in the feature vector space. Given a set of Flickr data D_u, that is relevant to a user u and class labels for training $\{(d_u, g_u)|u = 1, \ldots, N\}$, where d_u represent the feature vectors of user data and g_u

is the target class label, the SVM will map these feature vectors into a high dimensional space.

A. Content Based Classification

As the amount of social media content grows, researchers should identify robust ways to discover unknown knowledge about users, based on these contents. In this section, we explain how the contents of tags and images are used for gender classification.

Tags reflect what users consider important in their images and also reveal the users' interest. We assume that male and female tagging vocabularies are different, and this difference can be used to identify their gender. To test our assumption, we built a dictionary containing female and male tagging vocabularies. In order to determine the importance of tags, we compute the gender tagging vocabulary by counting the number of different gender users who used the respective tag. Then, we calculated the probability of a gender given the utilized tags.

The color histogram is a representation of color distribution in an image. For image data, a color histogram represents the number of pixels that have colors in each of a fixed list of color ranges that span the color space for the image. The color histogram can be built for any kind of color space. The hue histogram is based on the hue value of all the pixels in an image. Each hue value in HSL or HSV color space represents a color by itself.

The hue in bag of words approach is motivated by an analogy of learning methods that applies the bag-of-words representation for text categorization [11], visual categorization with bags of keypoints [12], and bags of features [13]. For this approach, we selected the top two colors by assigning "1" to the feature value for these top colors and "0" to the other colors.

B. Semantic Based Classification

Using the semantic content of multimedia data added by the user could be beneficial to the social user mining research. However, manually annotating images requires time and effort, and it is difficult for users to provide all relevant tags for each image. Thus, a semi-automatic image tagging system emerged and has recently involved in different task. To improve the quality of tags, we applied a semi-automatic image tagging system *Akiwi* to suggest keywords for images. The goal of using semi-automatic image tagging system is to assign a few relevant keywords to the image to reflect its semantic content. This process improves the quality of tags by utilizing image content. For the gender classification problem, the semantic based approach is conducted based on the collected keywords from *Akiwi*.

Similar to the tags data in the content based classification, the assumption is that male and female keyword vocabularies are different. This difference can be used to classify their gender. To test our assumption, we built a dictionary containing female and male keyword vocabularies. In order to determine the importance of keywords, we computed the gender tagging vocabulary by counting the number of different gender users who used the respective keywords. Then, we calculated the probability of a gender given the utilized keywords.

In addition, we proposed a data integration module to combine both semantic based and content based data. Particularly, we utilized this module for the gender classification problem by combining the keywords of the user with his/her tags. We used a feature vector to merge both the keywords and the tags of the user.

V. EXPERIMENT RESULT

This experiment utilizes Scikit-Learing tools in Python [14]. Two different classification methods, i.e., Naive Bayes and SVM were used. In this experiment, we adopted the multinomial Naive Bayes model. This model implements the Naive Bayes algorithm for multinomial distributed data, where the data are typically represented as feature vector. For the SVM, we adopted SVC, which is implemented based on libsvm. For both classifiers, we use the fit (X, Y) method. This method fit the classifier according to the given training data. Next, we used predict (X) method to perform the classification in a sample of X. In our case, X represents the feature matrix of the data, while Y represents the user label.

A. Data Set

One of the greatest online photo management and sharing application is Flickr. In this site, user can shares their photos and organize them in many ways. Textual and visual data can be obtained through the Flickr public API, which allows us to download information with the user's authorization. We downloaded 148,511 user's tags and images. Table 1 shows more details about our data set. To evaluate the proposed algorithms, we build a ground truth data based on 215k users. Precisely, we collected the Flickr users' profile information using crawler. For the gender attribute, we were able to collect 148,511 users with known gender.

TABLE I. DATA SET

Data Category	Size
Tags	Up to 300 tags per user
Images	Up to 50 images per user
Ground truth	148,511 user

B. Experiment Result for Content Based Classification

For the content based experiment, we implemented a multinomial Naive Bayes classifier. We examined the performance of different features, and we observed the difference that appeared in the classification. Table 2 shows the result of different features, such as tags, hue histogram, and hue in bag of words.

 ISBN: 978-1-61208-681-1

TABLE II. EXPERIMENT RESULT OF CONTENT BASED CLASSIFICATION

Features	Accuracy	F1
tags	0.7362	0.7349
huehist	0.6141	0.6140
huebow	0.5866	0.5786
tags+huebow	0.7365	0.7351
tags+huehist	0.7251	0.7228
huehist+huebow	0.6151	0.6150
tags+huehist+huebow	0.7181	0.7141

To assess the performance of our model, we used the standard classification accuracy (*Acc*) and F1 score as defined in the equations 3 and 4, shown below. For evaluation purposes, all classes are grouped into four categories: 1) true positives (TP), 2) true negatives (TN), 3) false positives (FP), and 4) false negatives (FN). For instance, the true positives are the users that belong to the positive class and are in fact classified to the positive class, whereas the false positives are the users not belonging to the positive class but incorrectly classified to the positive class.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$F1 = \frac{2TP}{(2TP + FP + FN)} \quad (4)$$

C. Experiment Result of Semantic Based Classification

To compare the performance of our approach, we use the classification accuracy (*Acc*) as defined in equation 3, precision (*Pre*), and recall (*Rec*) metrics, as well as F1 score as defined in the following equations:

$$\mathrm{Pr}e = \frac{TP}{TP + FP} \quad (5)$$

$$\mathrm{Re}c = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = 2\left(\frac{\mathrm{Pr}e \times \mathrm{Re}c}{\mathrm{Pr}e + \mathrm{Re}c}\right) \quad (7)$$

where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

We performed the experiments by sampling of the data set for different features and classifiers, and then tested the performance of each classifier and each feature. The results are presented in Table 3. As seen in the table, the results show over 80% in terms of accuracy for gender classification when using keywords with both classifiers. This indicates that the proposed semantic based approach outperforms the content based one. In terms of classifier, we observed that Naive Bayes is slightly better than SVM, specifically with tags. This is because the Naive Bayes classifier can work better even if there is some missing data.

TABLE III. EXPERIMENT RESULT OF SEMANTIC BASED CLASSIFICATION

Features	Approach	Acc	Pre	Rec	F1
keywords	NB	0.82	0.81	0.82	0.81
	SVM	0.82	0.83	0.82	0.80
Tags	NB	0.78	0.82	0.78	0.78
	SVM	0.74	0.55	0.74	0.63
Keywords +Tags	NB	0.80	0.80	0.80	0.79
	SVM	0.78	0.61	0.78	0.68

VI. CONCLUSION

In this paper, we proposed a new data integration method that integrates textual and visual data. Unlike the previous approaches that used a content based approach to merge multiple types of features, our main approach is based on image semantic through a semi-automatic image tagging system. Our method was applied to gender classification mining application and showed the performance of our method to discover unknown knowledge about user. For gender classification, we performed the experiments with the data set, and the results for the semantic based approach indicate over 81% accuracy for gender classification, which outperforms the content based approach by 10%.

REFERENCES

[1] A. Gallagher, D. Joshi, J. Yu, and J. Luo, "Geo-location inference from image content and user tags," in Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops CVPR Workshops 2009, pp. 55–62.

[2] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Ht06, tagging paper, taxonomy, flickr, academic article, to read," in Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, ser. HYPERTEXT '06. New York, USA: ACM, 2006, pp. 31–40.

 ISBN: 978-1-61208-681-1

[3] A. Popescu and G. Grefenstette, “Mining user home location and gender from flickr tags.” in ICWSM, 2010.

[4] 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 Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, ser. MIR ’06. New York, USA: ACM, 2006, pp. 249–258.

[5] M. Ames and M. Naaman, “Why we tag: Motivations for annotation in mobile and online media,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI ’07. New York, USA: ACM, 2007, pp. 971–980.

[6] P. Heymann, D. Ramage, and H. Garcia-Molina, “Social tag prediction,” in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’08. New York, USA: ACM, 2008, pp. 531–538.

[7] B. Sigurbjornsson and R. van Zwol, “Flickr tag recommendation based on collective knowledge,” in Proceedings of the 17th International Conference on World Wide Web, ser. WWW ’08. New York, USA: ACM, 2008, pp. 327–336.

[8] H. Zhang, “The optimality of Naive Bayes,” A A, vol. 1, no. 2, p. 3, 2004.

[9] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[10] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[11] N. Cristianini, J. Shawe-Taylor, and H. Lodhi, “Latent semantic kernels,” J. Intell. Inf. Syst., vol. 18, no. 2-3, pp. 127–152, Mar. 2002.

[12] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in In Workshop on Statistical Learning in Computer Vision, ECCV, 2004, pp. 1–22.

[13] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, 2006, pp. 2169–2178.

[14] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python ,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

 ISBN: 978-1-61208-681-1