User Profiling of Flickr: Integrating Multiple Types of Features for Gender Classification

Mohammed Eltaher and Jeongkyu Lee
Dept. of Computer Science and Engineering, University of Bridgeport, Bridgeport, CT, USA
Email: meltaher@my.bridgeport.edu; jelee@bridgeport.edu

***Abstract*—With the pervasive use of social media sites, an extraordinary amount of data has been generated in different data types such as text and image. Combining image features and text information annotated by users reveals interesting properties of social user mining, and serves as a powerful way of discovering unknown information about the users. However, there has been few research work reported about combination of image and text data for social user mining. The progress of data mining techniques makes it possible to integrate different data types for effective mining of social media data. In this study, we propose a novel idea to classify the gender of user by integrating multiple types of features. We utilize not only text information, i.e., tag or description, but also images posted by a user with semantic based data fusion technique. Unlike the previous approaches that used a content based approach to merge multiple types of features, our approach is based on image semantic through a semi-automatic image tagging system. For the classifier, we employ Naive Bayes and SVM algorithms, where the integrated data are typically represented as feature vector. We perform the experiments with the data set, and the results show over 80% in terms of accuracy for gender classification, which outperforms the content based one.**

***Index Terms*—user profiling, semantic based integration, gender classification**

I. Introduction

The explosive growth of social media network on the Internet has led to a massive volume of information on the web. Millions of users in different social media sources connect to each other, express themselves, and share interests through the web [1]. With a growing number of users in social media, being able to discover hidden information about users becomes important in many applications. For example, mining user's demographics information, such as gender, has its potential to extract actionable patterns that can be used for business marketing. If a marketing department can discover users' demographics, such as gender, such information should be useful in targeted online marketing.

Recently, social networks for multimedia sharing such as Flickr have become more popular by allowing people to easily upload, share and annotate multimedia objects with keywords. Labelling the multimedia objects, i.e., images, with a set of keywords is known as image tagging. Most of the social user mining tasks depend on the availability and quality of the tagging system. However, the existing studies show that tags are impressive, ambiguous and overly personalize [2]. A semi-automatic tagging process that helps to tag a multimedia objects would improve the quality of tagging. To overcome the above problems with tags, we use a semi-automatic image tagging system akiwi[1] to suggest keywords for images.

Although some researchers have already studied different problems in social media mining, approaches that apply text and image data for social multimedia mining are limited. In this paper, we propose a novel approach to classify the gender of user in Flickr by combining multiple types of features. We utilize tags and images of users by using semantic based information fusion technique. Unlike the previous approaches that use a content based approach to merge multiple types of features, our approach are based on image semantic through a semi-automatic image tagging system. 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 SVM. In order to evaluate the proposed algorithm, we download 148,511 user profile information with up to 50 photos for each user from Flickr.com. Fig. 1 shows an overview of the proposed user profiling task.

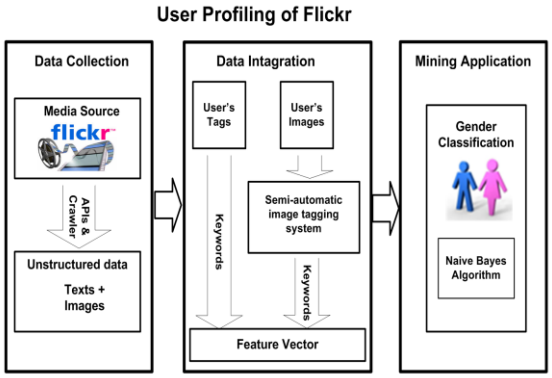

Figure 1. Overview of user profiling in Flickr

II. Related Works

We describe relevant related work in two areas, semantic based approach and content based approach. A

Manuscript received December 20, 2014; revised April 29, 2015.

[1] http://www.akiwi.eu/

doi: 10.12720/jait.6.2.84-87

semi-automatic tagging process that helps to tag a multimedia objects would improve the quality of tagging and the overall social user mining process. In general, the goal of an automated multimedia object tagging task is to assign set of semantic keywords to image or video. In [3, 4], the authors proposed a distance metric learning techniques to automated photo tagging tasks based on Flickr's images. [3] presented a Probabilistic distance metric learning techniques (PDML). First, they discover probabilistic side information from the data using a graphical model approach, and then present an effective probabilistic RCA algorithm to find an optimal metric from the probabilistic side information. On the other hand, [4] proposed an unified distance metric learning (UDML) method, which learns metrics from implicit side information hidden in massive social images on the web.

Based on a training set of tagged images, many models have been proposed to associate visual features with semantic concepts keywords. [5] proposed an annotation method which establishes the correlations between semantic concepts and low-level features. Using a local multi-label classification indictor function, their technique captures the keyword contextual correlations and exploits the discrimination between visual similar concepts. In addition, the authors in [6] introduced a probabilistic formulation for semantic image annotation. To address the limitations of unsupervised labelling, they presented a Supervised Multiclass Labelling (SML) by explicitly making the elements of the semantic vocabulary the classes of a multiclass labelling problem. Moreover, [7] proposed a semi-automatic image annotation model based on a sparse coding representation of the images. In order to remove the semantically irrelevant images, their method uses a label transfer mechanism to automatically recommend promising tags to each image by assigning each image a category label first. Based on the results, the recommended keywords can effectively reflect the image content.

In addition to the semantic based approach, there has been huge expansion of user generated content in social networks such as Twitter, YouTube, and Flickr. Mining demographics information such as gender, ethnicity, age, and marital status is an interesting topic for the researcher. As an example, Peersman et al. [8] applied a text categorization approach for the prediction of age and gender on a corpus of chat texts. Their study investigates the automatic prediction of age and gender using short chat messages from Netlog[2]. Moreover, Burger et al. [9] investigated the development of high performance classifiers for identifying the gender of Twitter users using content of the tweet text as well as three fields from the Twitter user profile: full name, screen name, and description. [10] addressed the task of predicting the gender of the YouTube users based on comments and profile.

III. GENDER CLASSIFCATION

We introduce a gender classification problem in Flickr as one of the applications in social user mining, which our proposed module should be able to work with. [11] introduces a gender identification technique for Flickr's users based only on tags. Different from their approach, we apply tags and images. The problem of social user mining can be defined as follows:

Problem Definition: *For a user u , given his d_u (multimedia objects) from Flickr, we predict the gender of u based on his multimedia objects*

For the multimedia objects, we extract two types of features representing it as following:

Textual feature: The text information used to describe the image by users, such as tags, titles, descriptions and comments. In this study, we used tags as textual feature because it reflect what users consider important in their images, and also reveal the users' interest. This feature is denoted as:

$$T = < t_1, t_2, ... t_n > \quad (1)$$

Visual feature: Visual information in social media is mainly represented by images or videos. To represent the visual feature, we label the semantic content of user's images with a set of keywords using a semi-automatic image tagging system. This feature is denoted as:

$$K = < k_1, k_2, ..., k_m > \quad (2)$$

We assume that male and female tagging vocabularies are different to some levels, and this difference can be used to identify their gender. To test our assumption, we build a dictionary which has female and male tagging vocabularies. We compute the importance of a tag in a gender vocabulary by counting the number of different users of that gender who used the respective tag and find the probability of a gender given the tags. Table I shows a sample of gender dictionary tested with sample Flickr data set.

TABLE I. SAMPLE OF TAGS GENDER DICTIONARY

Tag	*Male Frequency*	*P(male | tag)*	*Female Frequency*	*P(female | tag)*
panorama	6921	0.785	1896	0.215
cupcakes	776	0.309	1738	0.609
lake	9887	0.628	5869	0.372
fisherman	2125	0.67	1045	0.33
piazza	1085	0.679	514	0.321
dessert	1815	0.442	2290	0.558
police	4350	0.728	1623	0.272

Male frequency: number of male users that have used the tag at least once
Female frequency: number of feamle users that have used the tag at least once
P(male | tag): probability of male given the tag
P(female | tag): probability of female given the tag

IV. DATA INTEGEATION

Combining the features from image attributes and textual generated by user reveals interesting properties of social user mining and serves as a powerful way of

[2] http://www.netlog.com

discovering unknown information about users. However, there has been few research work reported about combination of images and texts for social user mining. In this section, we study the problem of integrating textual and visual data to perform gender classification task, and show that such combination may lead to better results comparing with using individual data type.

We propose a data integration module to combine both textual and visual information. First, we use a semi-automatic image tagging system *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. After that, we integrate these keywords of each user with his or her tag. Fig. 2 shows the data integration process.

TABLE II. DATA SET DETAILS

Data type	*Quantity*
Ground truth	148,511 user with known gender
User's tags	Up to 300 tag per user
User's images	Up to 50 image per user

V. EXPERIMENT AND DISCUSSION

A. Data Set

Flickr is one of the best online photo management and user can shares their photos and images on the social media site. In order to allow developers to access the information, Flickr offers a comprehensive API that allows developers to create any application for their data. In order to evaluate the proposed algorithm, we build a ground truth data based on 215k users. Specifically, we collected the Flickr users' profile information using crawler. For the gender attribute, we are able to collect 148,511 user with known gender. Textual and visual data can be obtained through the Flickr public API, which allows us to download information with the user's authorization. We download 148,511 user's tags and images. Table II shows more details about our data set.

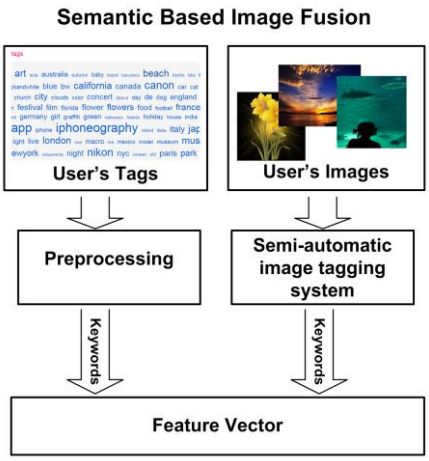

Figure 2. Semantic based data integration

B. Implementation

Experiment utilizes Scikit-Learing tools in Python [12]. We use two different classification methods, i.e., Naive Bayes and Support Vector Machine. Naive Bayes classifier is one of the most effective inductive learning algorithms for machine learning and data mining [13]. 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 adopt a multinomial Naive Bayes model. This model implements the Naive Bayes algorithm for multinomial distributed data, where the data are typically represented as vector. For the SVM, we adopt C-Support Vector Classification SVC which is implemented based on libsvm. For both classifier, we use the method fit (X,Y). This method fit the classifier according to the given training data. After that, we use the method predict(X) to perform the classification in sample of X. In our case, X represents the feature matrix of the data, while Y represents the user label.

C. Experiment Results

To compare the performance of our approach, we use the classification accuracy (*Acc*), precision (*Pre*) and recall (*Rec*) metrics as well as F1 score as defined in Equation 3,4,5, and 6.

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$\Pr e = \frac{tp}{tp + fp} \tag{4}$$

$$\mathrm{Re}\, c = \frac{tp}{tp + fn} \tag{5}$$

$$F1 = 2\left(\frac{\Pr e \times \mathrm{Re}\, c}{\Pr e + \mathrm{Re}\, c}\right) \tag{6}$$

where *tp* is true positive, *tn* is true negatives, *fp* false positives, and *fn* is false negatives.

TABLE III. EXPERIMENT RESULTS

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

We perform the experiments with sampling of the data set for different features and classifiers, and tested the performance of each classifier and feature. The results are presented in Table III. 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 term of classifier, we observe that Naive Baye is slightly better than SVM, specifically with tags. This is because the Naive Bayes

classifier can work better even if there are some missing data.

VI. Conclusion

In this study, we have presented a novel idea for gender classification of Flickr's user by integrating multiple types of features. Using semantic based information fusion technique, we utilize not only text information, i.e., tags and description, but also images of users. Different from the previous approaches that used a content based approach to merge multiple types of features, our approach is based on image semantic through a semi-automatic image tagging system. We perform the experiments with the data set, and the results show that our semantic based approach outperforms the content based approach.

References

[1] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: Inferring user profiles in online social networks," in *Proc. Third ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, 2010, pp. 251–260.

[2] 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. 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA: ACM, 2006, pp. 249–258.

[3] L. Wu, S. C. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging," in *Proc. 17th ACM International Conference on Multimedia*, New York, NY, USA: ACM, 2009, pp. 135–144.

[4] P. Wu, S. C. H. Hoi, P. Zhao, and Y. He, "Mining social images with distance metric learning for automated image tagging," in *Proc. Fourth ACM International Conference on Web Search and Data Mining*, New York, NY, USA: ACM, 2011, pp. 197–206.

[5] M. Wang, X. Zhou, and T. S. Chua, "Automatic image annotation via local multi-label classification," in *Proc. International Conference on Content-based Image and Video Retrieval*, New York, NY, USA: ACM, 2008, pp. 17–26.

[6] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, March 2007.

[7] W. Zhang, Z. Qin, and T. Wan, "Semi-automatic image annotation using sparse coding," in *Proc. International Conference on Machine Learning and Cybernetics*, vol. 2, July 2012, pp. 720–724.

[8] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proc. 3rd International Workshop on Search and Mining User-Generated Contents*, New York, NY, USA: ACM, 2011, pp. 37–44.

[9] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on twitter," in *Proc. Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1301–1309.

[10] K. Filippova, "User demographics and language in an implicit social network," in *Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 1478–1488.

[11] A. Popescu, G. Grefenstette, *et al.*, "Mining user home location and gender from flickr tags." in *ICWSM*, 2010.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] H. Zhang, "The optimality of naive bayes," *A A*, vol. 1, no. 2, pp. 3, 2004.

Mohammed Eltaher Mr. Mohammed is a full-time Ph.D. student of Computer Science and Engineering at the University of Bridgeport. He received his B.S degree in Computer Science from Sebha University, Libya in 2000 and the M.S. degree in Intelligent System from University Utara Malaysia in 2005. He worked as assistant lecturer at Department of Computer Science, Sebha University from November 2005 to January 2008. Mohammed has research interests in the areas of social media data mining, information retrieval and multimedia database.

Jeongkyu Lee Dr. Jeongkyu Lee is currently associate professor in Department of Computer Science and Engineering at University of Bridgeport. He received his Ph.D. in Computer Science from the University of Texas at Arlington in 2006. Previously, he received a BS from Sungkyunkwan University in Mathematics Education and an MS from Sogang University in Computer Science, both of South Korea. Before he pursued his doctorate, Dr. Lee worked as a database administrator for seven years with companies including Boram Bank, Hana Bank and IBM Korea. His primary research area is in the multimedia database management system and analytics. Research interests include graph-based multimedia data modeling, indexing structure, query processing, content/semantic based multimedia retrieval and summary. In addition, his interests include big data system/analytics, and social media data mining.