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

Mohammed Eltaher, Advisor: Prof. Jeongkyu Lee
Department of Computer Science and Engineering, University of Bridgeport, CT.

Multimedia Information Group(MIG)

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. 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.

Overview

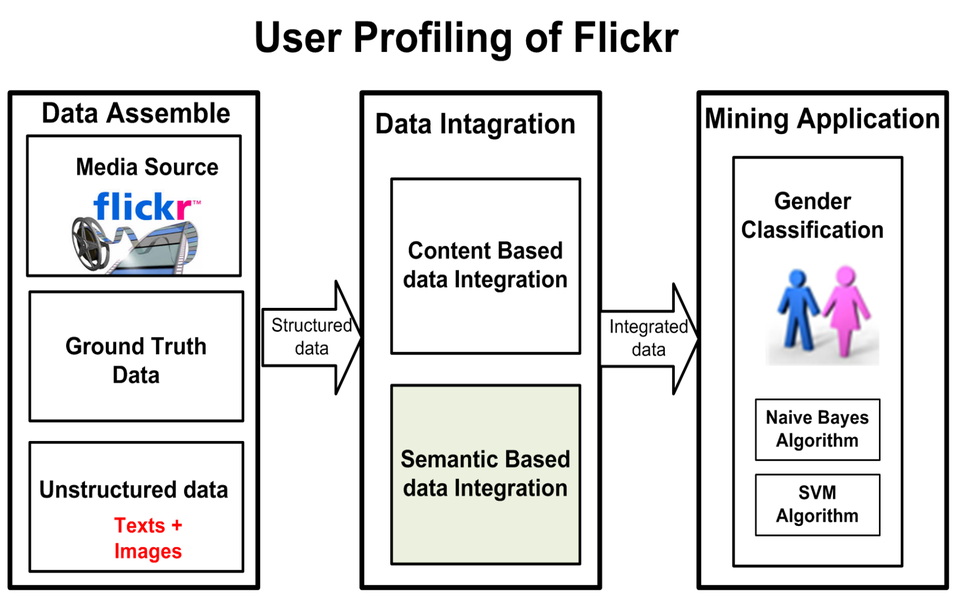


Figure 1:An overview of social user mining

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

Data Assemble

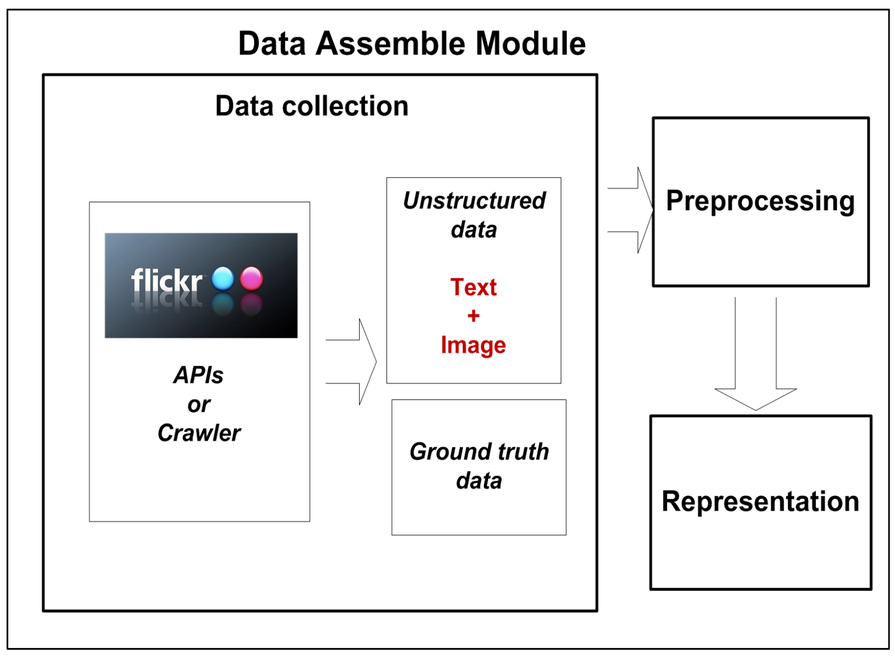


Figure 2: Data Assemble module

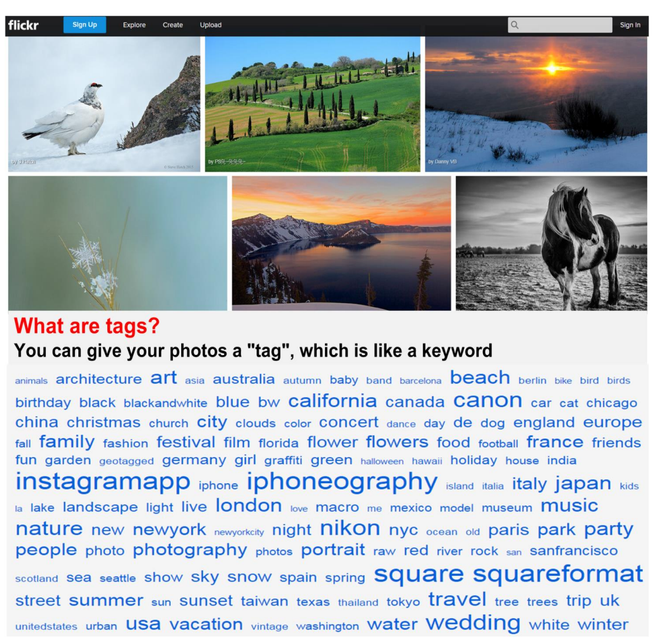


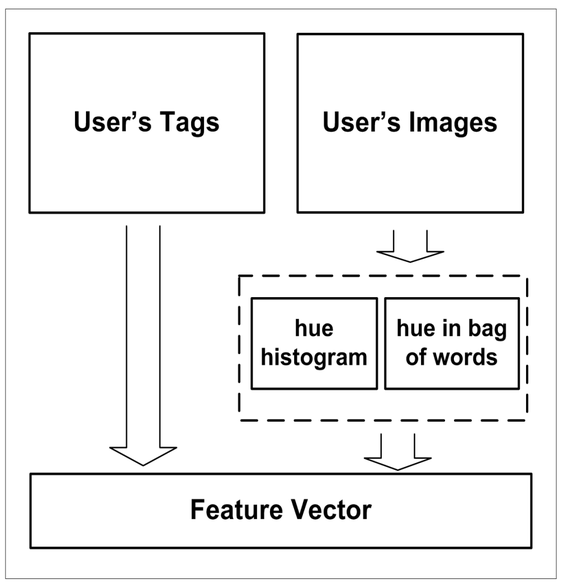
Table 1: Data details

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

Figure3. Example for textual and visual data

Data Integration

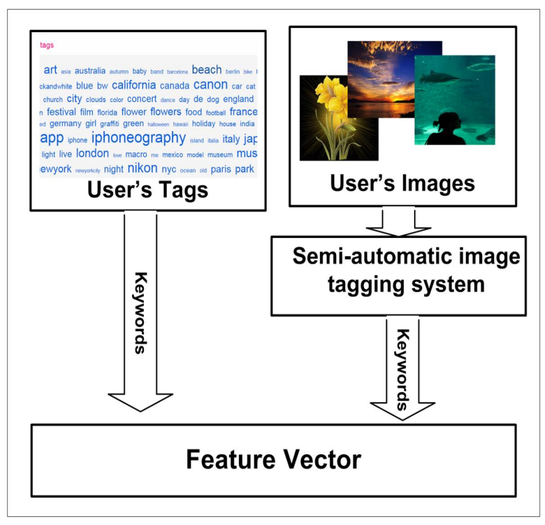
1. Content based data fusion



- Data integration between user's tags and image contents.
- For the image contents, we use hue histogram and hue in bag of words.
- Implemented all as a feature vector

Figure 4. Content based data integration

2. Semantic based data fusion



- Data integration between user's tags and keywords.
- Semi-automatic image tagging system akiwi used to suggest keywords for images.
- Implemented all as a feature vector

Figure 4. Semantic based data integration

Experiment and Results

1. Content based classification

- Multinomial Naive Bayes classifier

Table 2. Content based classification result

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

1. Semantic based classification

- Multinomial Naive Bayes classifier
- C-Support Vector Classification SVC

Table 3. Semantic based classification result

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
Key+ Tags	NB	0.80	0.80	0.80	0.79
	SVM	0.78	0.61	0.78	0.68

Conclusion

We have presented a novel idea for gender classification of Flickr's user by integrating multiple types of features, content and semantic based information fusion technique. We utilize tags and images of users. We perform the experiments with the data set, and the results show that our new semantic based approach outperforms the content based approach.