

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

Multimedia Information Group(MIG)

Mohammed Eltaher, Advisor: Prof. Jeongkyu Lee

Department of Computer Science and Engineering, University of Bridgeport, CT.

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

User Profiling of Flickr

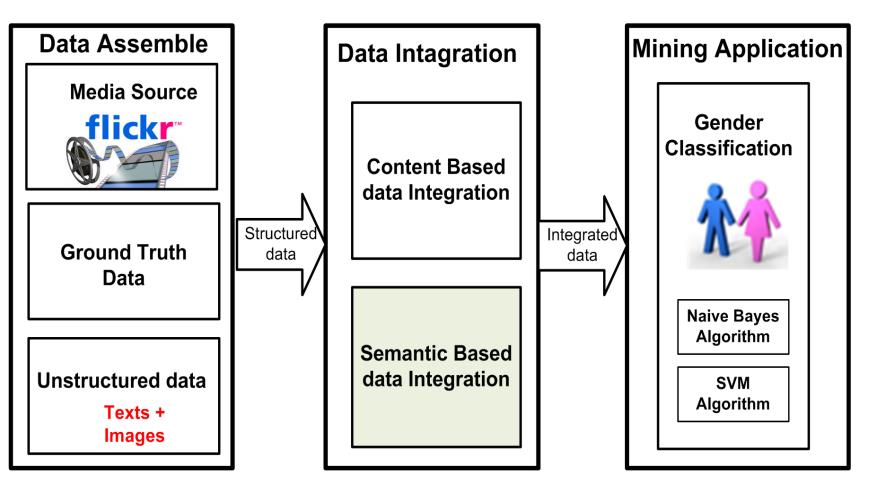


Figure 1:An overview of social user mining

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

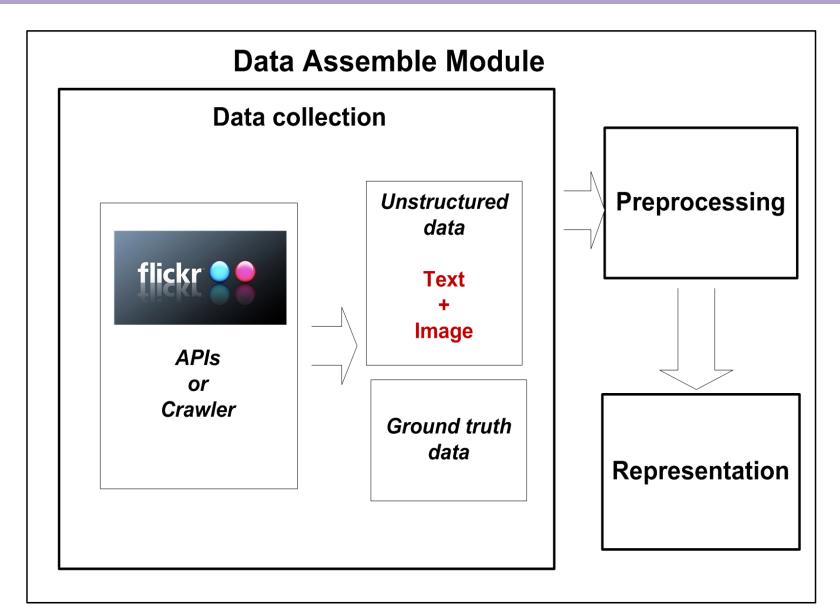
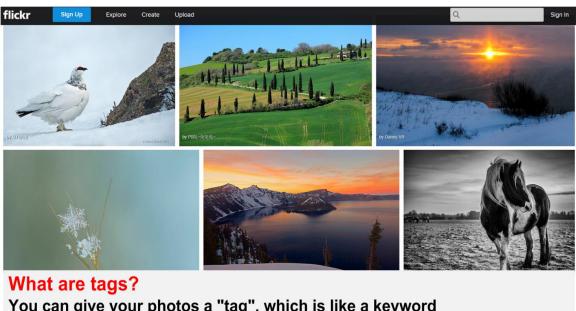


Figure 2: Data Assemble module



You can give your photos a "tag", which is like a keyword

animals architecture art asia australia autumn baby band barcelona beach berlin bike bird birds birthday black blackandwhite blue bw california canada canon car cat chicago china christmas church city clouds color concert dance day de dog england europe fall family fashion festival film florida flower flowers food football france friends fun garden geotagged germany girl graffiti green halloween hawaii holiday house india instagramapp iphone iphoneography island italia italy japan kids la lake landscape light live london love macro me mexico model museum music nature new newyork newyorkcity night nikon nyc ocean old paris park party people photo photography photos portrait raw red river rock san sanfrancisco scotland sea seattle show sky snow spain spring Square squareformat street summer sun sunset taiwan texas thailand tokyo travel tree trees trip uk unitedstates urban USA vacation vintage washington water Wedding white winter

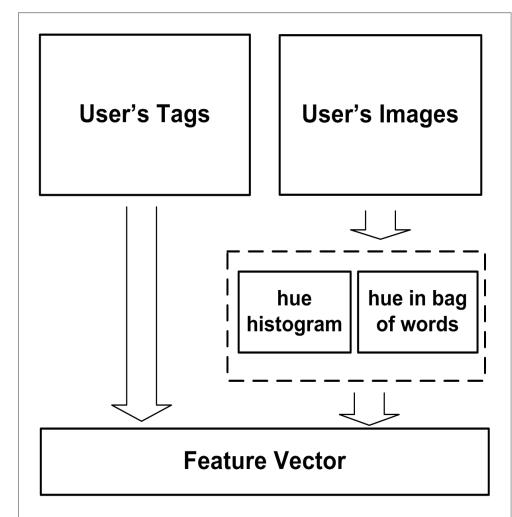
Table 1: Data details

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

Figure 3. Example for textual and visual data

Data Integration

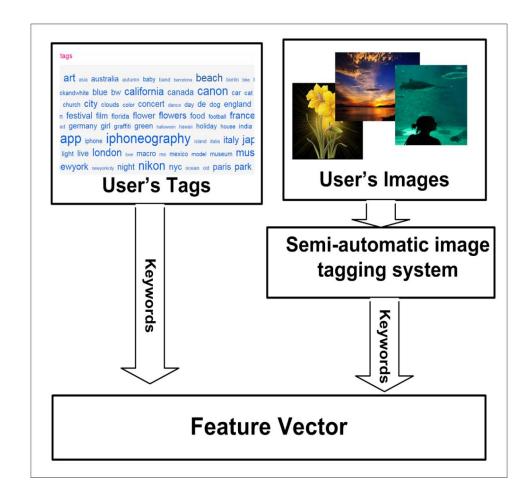
1. Content based data fusion



- 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

Table 2. Content based classification result

Multinomial Naive Bayes classifier

| 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

Table 3. Semantic based classification result

Multinomial Naive Bayes classifier

C-Support Vector Classification SVC

| 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.