Social User Mining

Mohammed Eltaher, Advisor: Prof. Jeongkyu Lee

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

Multimedia Information Group(MIG)

Abstract

In recent years, the pervasive use of social media has generated huge amounts of data that starts to gain a lot of attentions. Each social media source utilizes different data types such as textual and visual. For example, Twitter is for a short text message, Flickr is for images and videos, and Facebook allows all of these data types. With the use of data mining techniques, the social media data opens a lot of opportunities for researchers. To address these challenges and to discover unknown information about users, we first introducing data assemble module to handle both textual and visual information from different media sources. After that, we Introducing data integration module to integrate textual and visual data. In addition, we proposed two different applications for social user mining.

Introduction

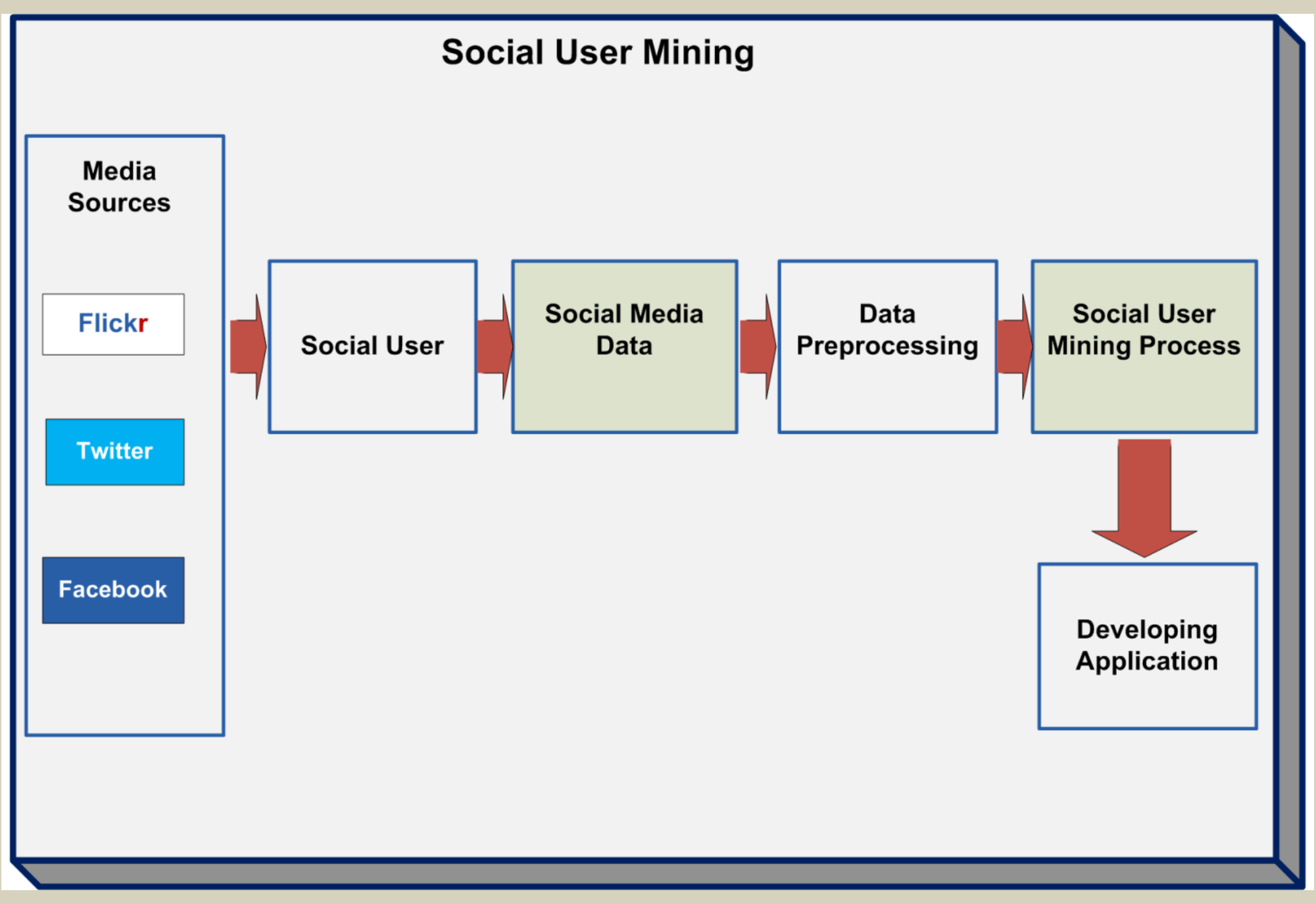


Figure 1:General diagram of social user mining

Social user mining is to discover unknown (or hidden) information about users from their publicly-available data on social media. We can define social user mining as follows:

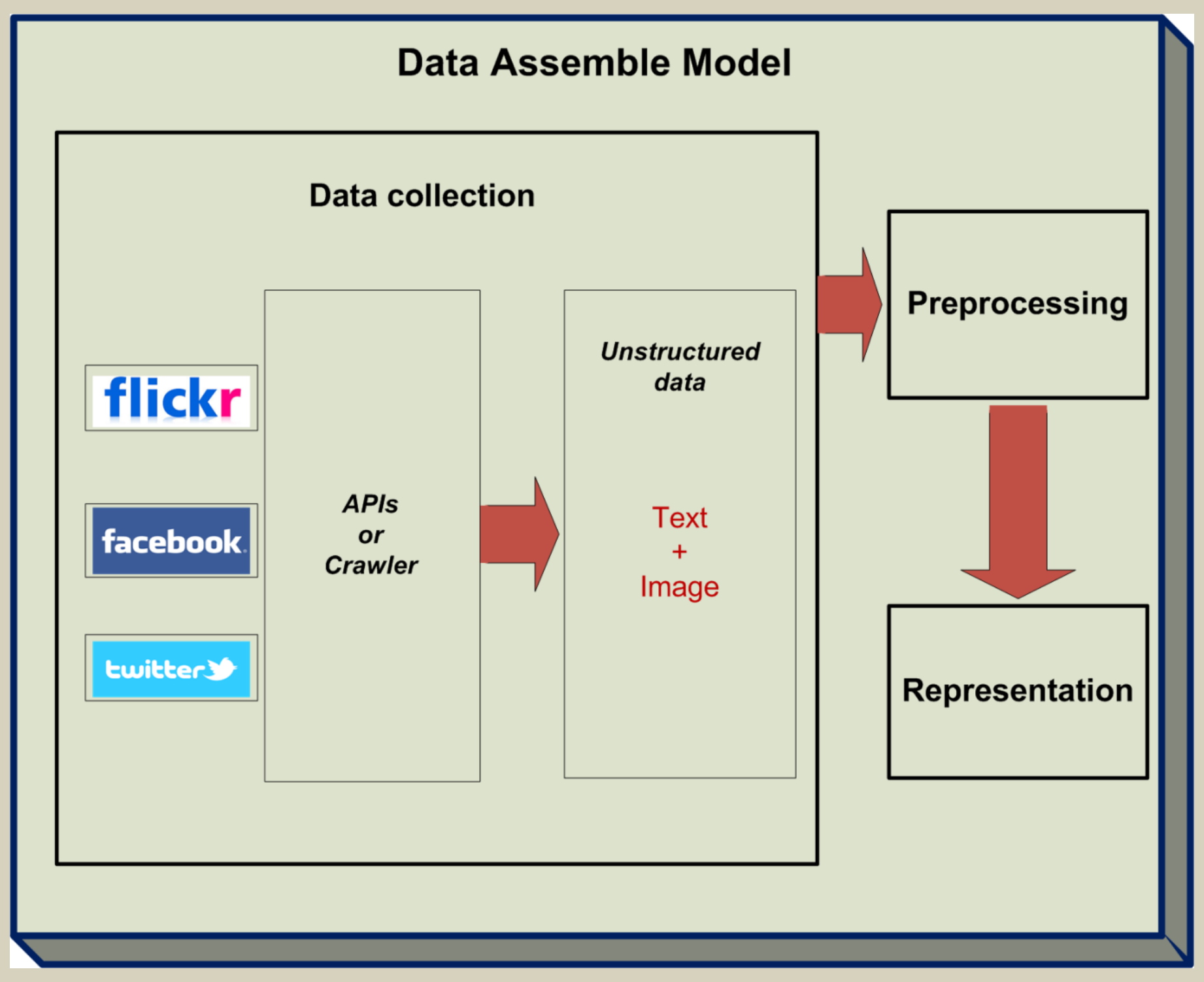
Definition1 *a social media data* is a multimedia document created by the user of a social media network. A social media data consists of multimedia data contents including text, audio, image and video. The social media data are a basic item of social user mining: for instance, a tweet in Twitter, and a photo with meta-data in Flickr.

Definition2 *Social user* creates social media data with *n* number of user attributes describing user profile information. For example, a Flickr user has profile information including user-id, location, name, gender and marital status.

Definition3 Given a set of social media data that are relevant to a user, social user mining is a process to discover one or more missing attributes of a social user.

Data Module

We introduce data assemble module to handle both textual and visual information from different media sources. It is a big challenge that making a good structured data from various media sources to discover unknown and meaningful information about users. Moreover, some missing information is also another challenge due to the privacy. In many social media sources, it is very hard to access the user information due to the privacy setting. Yet another challenging aspect of social media data is a lack of ground truth. For any social user mining task, we are required to build a ground truth data set to validate the result.



There are three main tasks associated with this model: (1) data collection, (2) data preprocessing, and (3) data representation. Depending on the data type, each step has a different way to process it. Figure 2 shows the data assemble module.

Figure 2: Social user data assemble module.

Data Integration Module

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 discovering unknown information about users. We proposed a data integration module to combine both textual and visual information based on a new approach compare to the state of the art.

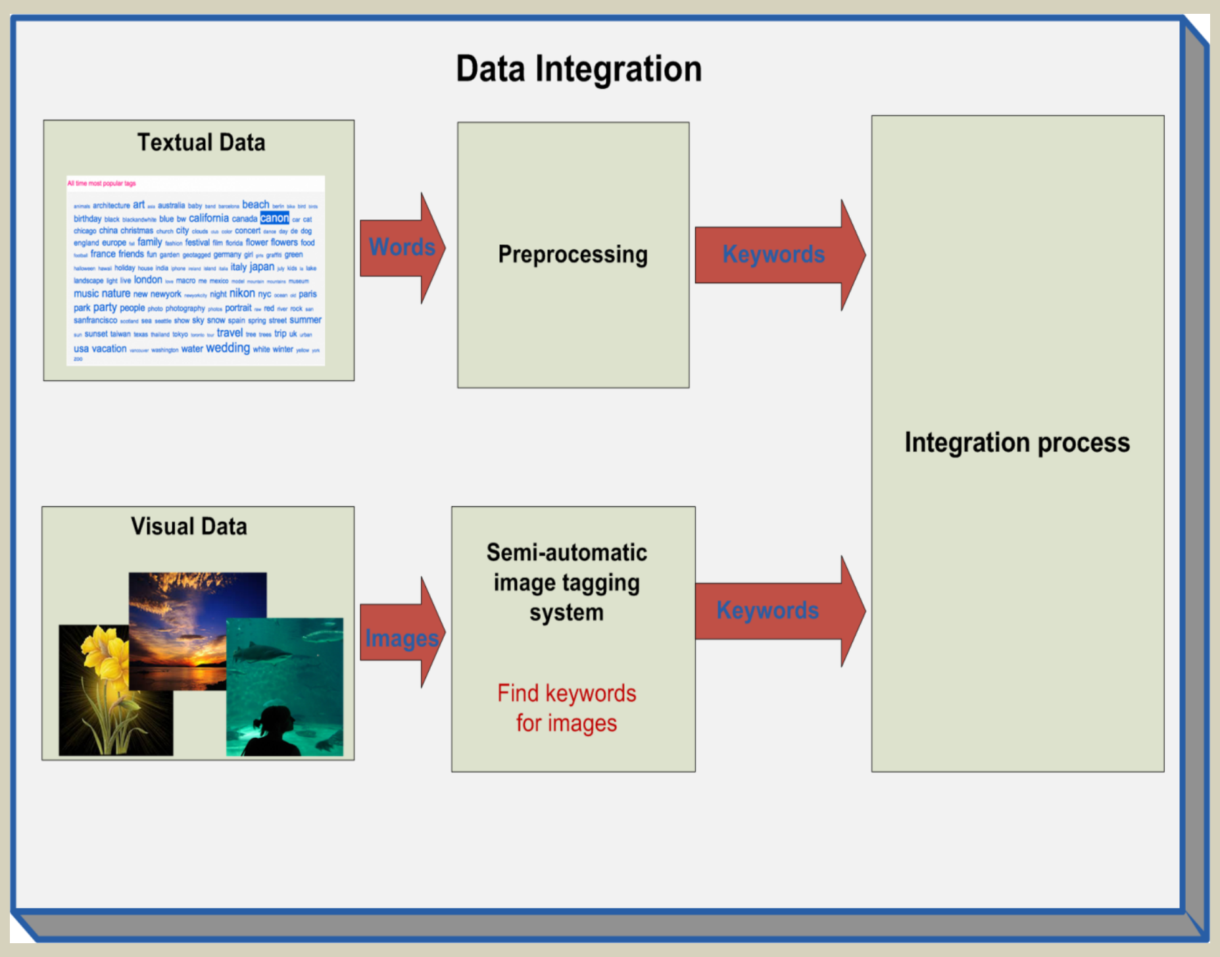


Figure 3:Data Integration module

- Semi-automatic image tagging system to suggest keywords for images.
- Combined the keywords from tags with suggest keywords that we got from images.

Applications

1. User Location

@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage

Content-based Home Location Estimation.

Figure 4: Twitter user location

Dataset

- Collected by Cheng et al. (CIKM10)
- Training set: 130K users with 4M tweets, location in city levels
- Testing set: 5,119 users, each with around 1,000 tweets, locations in coordinate
- 5,913 US cities with more than 5,000 of population in Census 2000 U.S. Gazetteer

Problem definition

For a user u, given a set of his tweet messages Tu = {t1,……,t|Tu|} ,where ti is a tweet message up to 140 characters, and a list of candidate cities, C, predict a city c that is most likely to be the home location of u.

Summary of predicting user Location

- Model location with spatial word usage
- Estimation with Gaussian Mixture Model (GMM)
- Unsupervised local word selection

2. Gender Detection

Problem definition

For a user u, given a set of his tags and images from Flickr, predict the gender of u as following:Classification based on tags, images and both.

Dataset

- Flickr data collected using Flickr's APIs
- We have 150k users with their tags and favorite images

Summary of predicting user gender

- Building data based on the module and create good quality ground truth
- combine different social media data to improve the performance of overall solutions using the proposed data integration module.

Conclusion

We introduce a novel mining approach for social user mining which has the following: data assemble module for different media source and data integration module. We have highlighted the need for Social user mining in view of the rapidly growing amounts of social data. In conclusion, social user mining is a promising field for research and still in its early years and many issues need to be resolved.