Harvesting Multiple Sources for User Profile Learning: a Big Data Study

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari and Tat-Seng Chua
farseev@u.nus.edu; nieliqiang@gmail.com; {akbari, chuats}@nus.edu.sg
School of Computing, National University of Singapore

ABSTRACT

User profile learning, such as mobility and demographic profile learning, is of great importance to various applications. Meanwhile, the rapid growth of multiple social platforms makes it possible to perform a comprehensive user profile learning from different views. However, the research efforts on user profile learning from multiple data sources are still relatively sparse, and there is no large-scale dataset released towards user profile learning. In our study, we contribute such benchmark and perform an initial study on user mobility and demographic profile learning. First, we constructed and released a large-scale multi-source multi-modal dataset from three geographical areas. We then applied our proposed ensemble model on this dataset to learn user profile. Based on our experimental results, we observed that multiple data sources mutually complement each other and their appropriate fusion boosts the user profiling performance.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Multiple Source Integration; User Profile Learning; Mobility Profile; Demographic Profile

1. INTRODUCTION

Regarded initially as a hub for teenagers and college students, social media in recent years has increased its impact on the way people live and communicate all around the world. Moreover, some popular social networks like Facebook and Twitter are constantly growing by involving more and more people of different genders, ages and lifestyles [6]. The rapid growth and interconnections between online social services are somehow related to the popularity of smartphones, which enable users to communicate while on the move. Furthermore, it is well known that more than half of online adults use more than one social network in their daily life[1]. These social services are often connected to each other by deploying the so-called cross-linking functionality [4]. With multi-source and mobility-related data, understanding user behaviours via user profile learning is promising and possible [13][15][17][25][26].

In this work, we define user profile as comprising of user mobility profile and user demographic profile, as shown in Figure 1. *Mobility* is a contemporary paradigm, which involves various types of people movement. It can be a physical movement in space, the movement from one social group to another or even a virtual movement between web pages while surfing the Internet [22]. The user mobility profile describes users' behavioural habits and interests. It is useful in many application scenarios, such as transport routes planning [11][20] or the control of spread of diseases [19]. On the other hand, the user *demographic profile*[2] typically includes age, gender and social class. These demographic attributes were also studied in a popular evaluations named PAN [18][21]. In general, the demographic profile is often used in marketing to describe a demographic grouping or a market segment. By considering users' demographic aspects, we are able to gain deep insights into users' behaviour and interests understanding. For example, in advertisement domain, age and gender significantly affect the list of potential products that can be routed to a given consumer: cars may be advertised mostly to adult males; while toys to kids and their parents. At the same time, it can also be used to support recommendation [13] and personalized question answering [12]. The combination of mobility and demographic profile opens the possibility of urban facility planning according to demographic properties of city areas: government organizations may consider building more kindergartens and schools in areas with high concentration of young families; while more food courts near professionals' work places.

User profile learning by jointly considering multiple sources is, however, non-trivial, due to following reasons:

- *Multiple data sources:* Users frequently enroll in multiple on-line social forums, which captures the

ICMR '15 Jun 23 - 26 2015, Shanghai, China

ACM ACM 978-1-4503-3274-3/15/06 ...$15.00.
DOI: http://dx.doi.org/10.1145/2671188.2749381

[1] www.pewinternet.org
[2] en.wikipedia.org/wiki/Demographic_profile

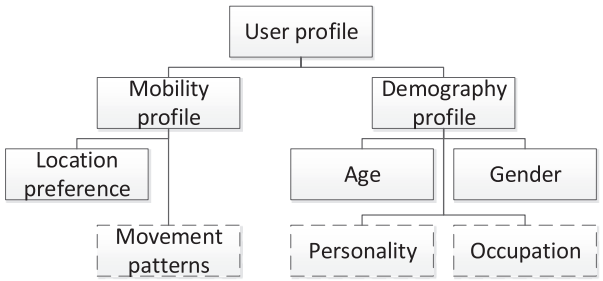

Figure 1: User profile and its compositions.

same users from different views. For example, Linkedin conveys users' formal career path; while Twitter casually uncovers users' daily activities. Effective integration of multi-view data from different sources is a tough challenge.

- *Multi-modal data:* Beyond texts, other modalities often appear in users' messages. Users may take and share pictures in photo-sharing services like Instagram, upload videos to Youtube and perform Foursquare/Swarm (Foursquare) check-ins. Users' messages on social media platform such as Twitter also include more multimedia. User profile learning on these heterogeneous information requires new theories that can seamlessly integrate such heterogeneous data in a complementary way.
- *Benchmark Dataset:* As far as we know, there is no a publicly accepted benchmark dataset towards user profile learning with multi-source multi-modal data. Due to the sensitivity of privacy, only a limited amount of data can be collected for internet active users. Even worse, after the collection of necessary data, it is a big challenge to align various social network accounts to the same user. Another problem is the lack of ground truth, which greatly hinders the development of learning algorithms.

To facilitate user profile learning via multi-source multi-modal data, we build a representative data collection from three prevailing social resources, namely, Twitter, Foursquare and Instagram. The ground truth is obtained by crawling publicly available information from users' Facebook pages. Based upon the constructed dataset, we conduct the preliminary study of profile learning. In particular, user mobility is modelled as a location-topic distribution and reflects user's location category preferences. We also perform first-order data analysis by calculating mobility-related statistics that describe users' activities for different social sources. In addition, we infer some attributes of the demographic profile, such as the age and gender. We regard the demographic profile learning as an ensemble learning task by fusing multiple and heterogeneous information cues. We claim and verify that multi-source multi-modal data fusion is able to boost the results of the demographic profile learning, as compared to the use of a mono source.

Several prior research efforts have already been dedicated to user profile learning. Some prior works were devoted to urban user mobility analysis [8][13][14]; while others focused mostly on location-based user community detection and profiling [27] or market trade area analysis [17]. The user demographic profiling task gained popularity following the works in [18][21], where participants were asked to build a model to predict user's age and gender based on users' textual posts. Users' occupation prediction from a large Twitter text corpora was also explored in [1]. However, most of these works are either based on a single-source or a single-modality dataset. Few of them explored and leveraged multi-source multi-modal data.

In this work, we constructed and released [7] a multi-source dataset, provided users demographics and mobility ground truth, and performed preliminary studies in these two directions.

2. MULTI-SOURCE BIG DATA

This section details the dataset construction. In order to cover the most popular modalities (visual, textual and location data), we incorporate following social media sources:

- Foursquare as a location data source;
- Twitter as a textual data source;
- Instagram as a visual data source;
- Facebook as a demographics-related ground truth source.

The data was obtained from three geographical regions with high Foursquare user activity, namely: Singapore London and New York. It can be used for research on *user profile learning* and other contemporary problems such as *venue recommendation* [27], *user identification* across multiple social networks [23], and *cross-region user community detection* [28]. Considering the sensitivity of users' private information, we release the extracted features instead of the original data.

We first collected a set of active users, who have recently posted tweets through the cross-linking functionality of Instagram or Swarm mobile apps. We utilized Twitter REST API[3] to perform the location-dependent tweets search in Singapore, London and New York regions. This crawling method offers the possibility to map Twitter user IDs to Foursquare and Instagram user IDs, which allows us to overcome the multi-source user identification problem [23]. Based upon the active user list, we crawled user generated contents for those users, who posted their activities on Twitter. Tweet streams can be monitored with the key words specified as "swarmapp.com" and "instagram.com" to receive tweets posted from Swarm and Instagram mobile applications, respectively. For example, each sampled check-in message contains a short link to the original check-in page, where the check-in details are available. By following this, we performed check-ins and media crawling for previously identified active users. Noticeably, this crawling strategy can be applied to crawl other social networks as well.

2.1 Statistics of the Dataset

The dataset crawling process started on 10 July 2014 in Singapore, London and New York regions, and ended on 20 Dec 2014. The number of data records for each city is summarized in Table 1. In our current work, we perform data analysis and user profile learning for users from Singapore region only. It means that all our following statements refer only to users in Singapore, while cross-region profile learning could be addressed in future work.

Figure 2(a) presents the percentage of users with respect to the number of posts. It can be seen that the users'

[3] `dev.twitter.com/rest/public`

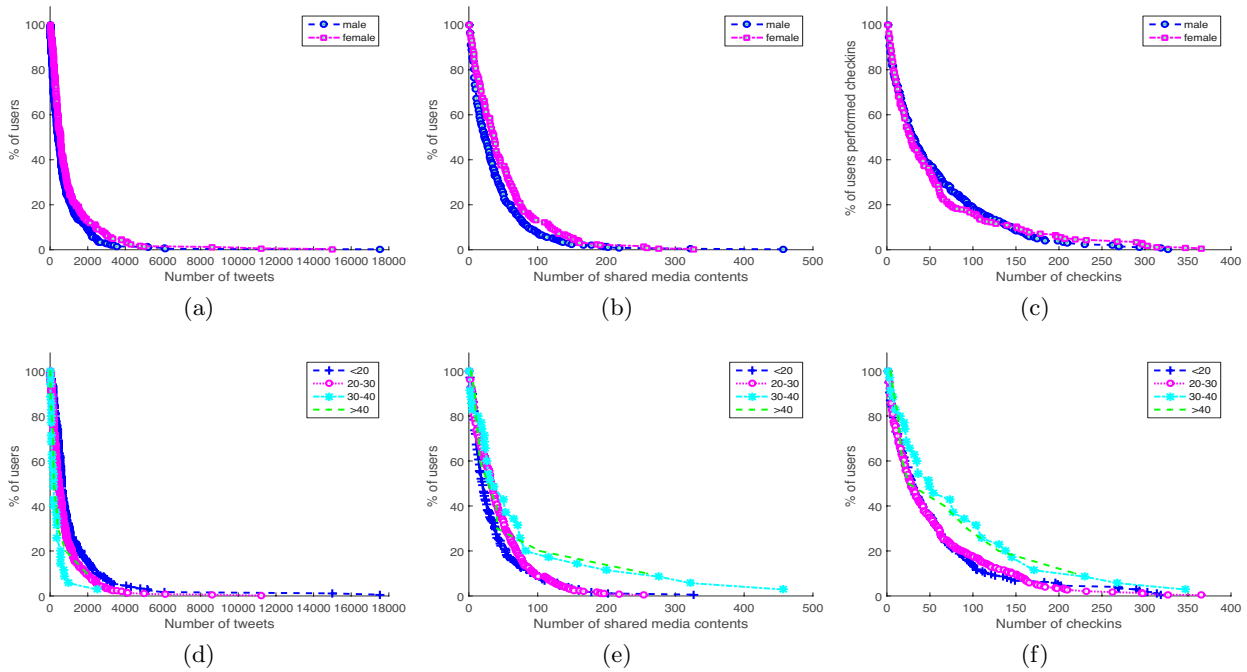

Figure 2: The statistics of Singapore dataset: (a) illustrates the percentage of users with respect to the number of posted tweets based on gender; (b) and (c) respectively show users posted multimedia on Instagram and checkins in Foursquare. (d) illustrates the percentage of users with respect to the number of posted tweets based on their age group; (e) and (f) respectively show users posted multimedia on Instagram and checkins in Foursquare.

Table 1: Number of data records in our constructed dataset from Singapore, London and New York.

City	#users	#tweets	#ch-ins	#images
Singapore	7,023	11,732,489	366,268	263,530
London	5,503	2,973,162	127,276	65,088
NY	7,957	5,263,630	304,493	230,752

microblog activity is not very correlated to their gender. Figure 2(b) shows the percentage of users with respect to the amount of multimedia contents which they shared through Twitter. It shows that female users in general share more media content than male users. Figure 2(c) shows the analysis results of user check-in behaviors. It can be seen that males tent to perform more check-ins than females.

We also analyzed users' posting behavior with respect to users' age. Figure 2(d), 2(e) and 2(f) respectively show user behavior distributions in terms of writing microblog messages, sharing media contents, and posting check-ins. It is interesting but surprising to notice that, users from 30 - 40 age group are less active in microblog than younger users (< 20 y.o; 20 - 30 y.o); while in image and location sharing services, older users in average (30 - 40 y.o; >40 y.o) more active.

In order to visualize the semantic component of data, we plot the distributions of users among popular Foursquare venue categories and Instagram image concepts for users of different age and gender. Figure 3(a) shows the most popular venue categories in our dataset with respect to users' gender. As it can be seen, there is a meaningful difference in venue preference between male and female users. In particular, men are more interested to visit night clubs and food courts, while female users often perform check-ins from home or cafe. Figure 3(c), in contrast, visualizes user mobility with respect to age. It shows that adult users (older than 20 y.o) often perform check - ins in malls, while children and teenagers (<20 y.o.) often stay at home.

Besides, we analyzed users' media sharing behaviours on Instagram. Figures 3(b) and 3(d) demonstrate the distribution of the extracted image concepts for users' gender and age groups, respectively. From Figure 3(b), we can see that images posted by woman often include hairs and dresses, which are represented by "Wig" and "Pajama" concepts, while men mostly shared images with cloth-related ("Ski mask", "Bulletproof vest") and daily life ("Comic book", "Chocolate sauce") concepts. Figure 3(d) explicitly reflects that distinct age groups are interested in different concepts. For example, food-related concepts such as "Plate", "Pot-pie", and "Burrito" are frequently shared by senior age group, while "Comic book" and "Toy shop" are often shared by users of between 30-40 years old[4].

2.2 Ground Truth

The ground truth was crawled from the publicly available Facebook pages. The Facebook-Foursquare-Twitter account ID mapping was obtained via Foursquare REST API[5]. The significant part of users in Singapore (3,849 records) have mentioned their gender, while only a few have provided their actual age (305 records). In order to complement age-related ground truth, we estimated the missing age from their education path. To encourage other broad research, we have also released [7] ground truth records about user's Education, Employment, Demography, Location and Family.

[4]Image concepts were extracted from ImageNet [5].
[5]`developer.foursquare.com`

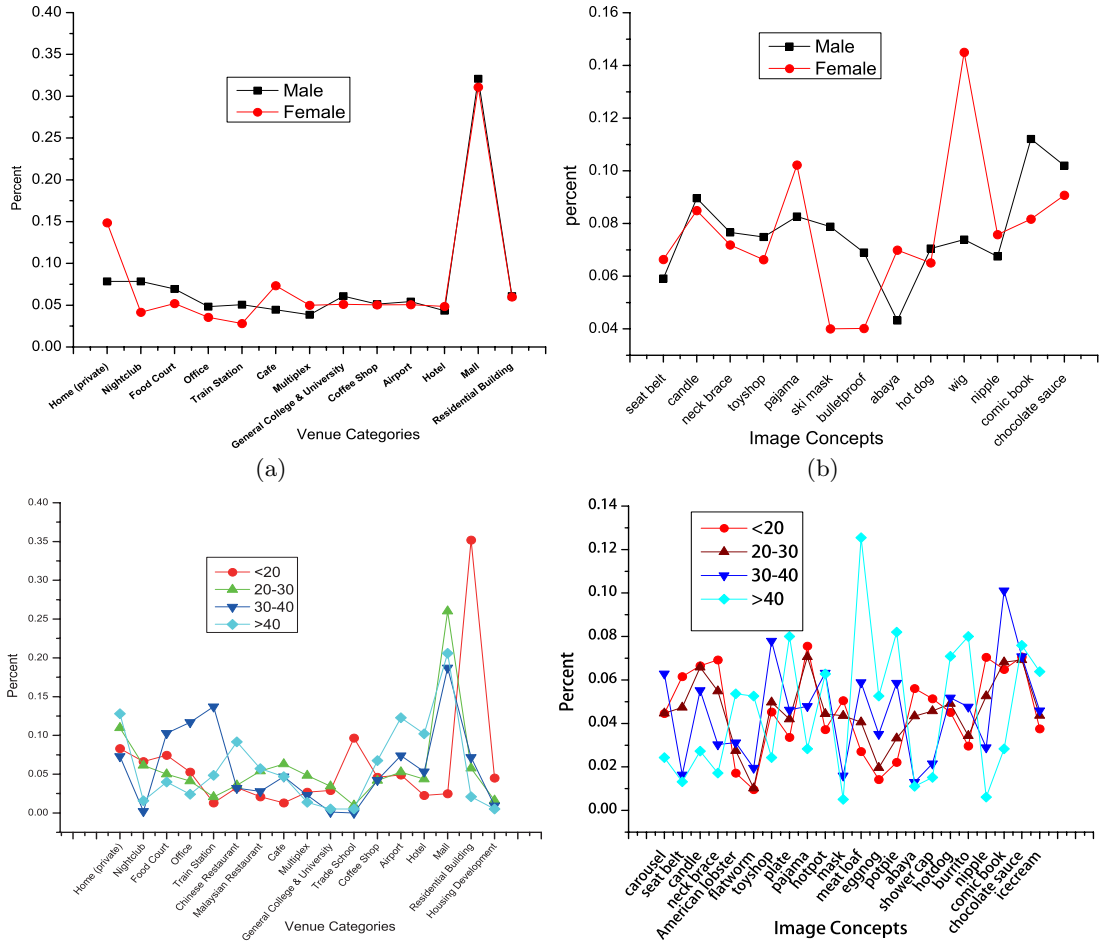

Figure 3: Illustration of the distributions of venue categories and image concepts over genders and ages for the Singapore dataset. (a) and (b) respectively illustrate the distribution of categories and image concepts for different genders; while (c) and (d) illustrate the distribution of categories and image concepts for different age groups, respectively.

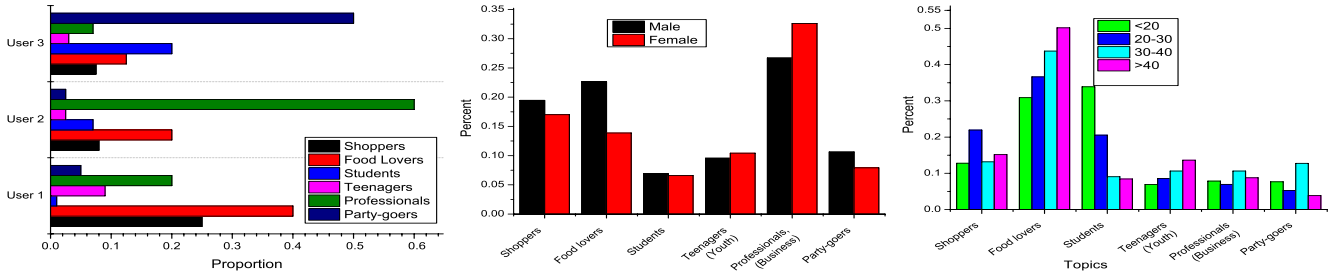

Figure 4: The personal mobility statistics for the Singapore dataset: (a) shows a category-based user mobility profiles for three users; (b) and (c) show the distribution of users among mobility categories based on gender and age respectively

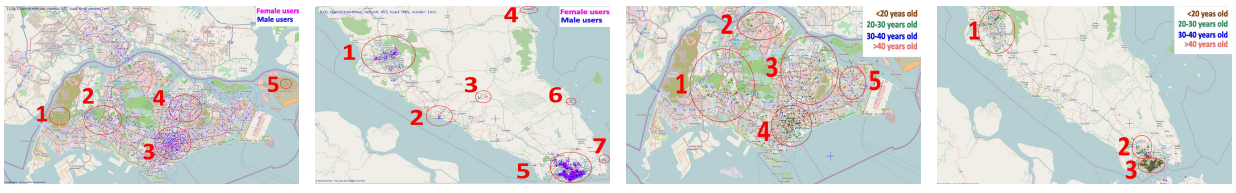

Figure 5: Visualization of user check-ins. (a) and (b) illustrate the spatial distribution of check-ins for different gender on Singapore and regional map respectively; (c) and (d) illustrate the spatial distribution of check-ins for different age groups on Singapore and regional map respectively.

3. USER PROFILE LEARNING

3.1 User Mobility Profiling

3.1.1 Individual User Mobility

One way to create category-based user mobility profile is by counting the check-ins frequency for each venue category. As can be seen from previous sections, most of the time users preferred to visit venues that belong to popular categories such as "Mall", "Home", or "Nightclub". However, the dimension of such user profile would be the same as the number of categories in Foursquare category hierarchy (i.e. 592 in Singapore), and thus cannot be easily interpreted or visualized. Moreover, it is more useful to extract individual users' interests that are represented by the tail of user-category distribution.

To address these problems, we apply LDA [2] on venue category and their descriptions, which has been found to be a good approach for location-based user profiling [17]. In LDA terminology, each user is modelled as a LDA document, while each venue category is considered as a LDA word. We empirically found the number of LDA topics equals to 6, since it gives more human-interpretable category-topic distribution results. The venue category distribution for different LDA topics is presented in Table 2.

We label each topic by considering the corresponding top 5 most popular venue categories. As a result, we can clearly define 6 interest groups, namely, "Food Lovers", "Travellers (Business)", "Party Goers", "Family Guys", "Students" and "Teenagers (Youth)". It is worth noting that in most of the obtained mobility groups (LDA topics), users are represented by the popular categories that obviously reflect *personal* user mobility profile. Such user mobility profiling tackles the human interpretability and visualization problems as well. Each user profile can be visualized based on the mobility groups learned such as that based on 3 users described in Figure 4(a), where "User 1" often post about food, while "User 2" and "User 3" are Professional and Party-goer, respectively. Next, we compare the obtained users' profiles for different demographic categories. Specifically, we draw the distribution of users among mobility "topics" based on users' gender (Figure 4(b)) and age group (Figure 4(c))). From Figure 4(b), it can be seen that females often manifest their professional side, while male users often demonstrate student's lifestyle. Regarding the age differences, most of the "Party-goers" belong to 30-40 age group, while people > 40 tend to contribute food-related posts. Therefore, the obtained mobility patterns clearly separate users from different age groups and gender, and can thus be considered as a powerful approach for *individual* user mobility profiling.

3.1.2 Group Mobility

To make deeper insights into user group mobility, we visualize users' check - ins on Singapore map for different age groups and genders. We also provide smaller scaled map to visualize inter - city user mobility. From Figure 5(a), it can be seen that the city population is concentrated in several dense regions, which represent peoples' housing (Regions 2 and 3) and working (Region 3) areas. Generally, male and female users are evenly distributed in these areas. However, there are some regions where male (Blue markers) user check-in density is much higher than that of female (Pink markers). It can be explained by the special purposes of the land usage, which includes military objects and army polygons (Regions 1 and 5). From the inter-city perspective (Figure 5(b)), it can be seen that both female and male users often perform trips to nearby cities for shopping and leisure purposes (Regions 1, 2, 4, 5). However, there are some geographical areas that are popular only by female users (Regions 2, 3). It can be explained by attractiveness of these regions to users who travel with family. For example, Region 2 is the city of "Malacca" which is famous by it's resorts, while Region 3, the "Segamat Distinct", is the leisure area with natural waterfall and national park, which could be popular among family groups. From the user age distribution map (Figure 5(c)), we can see that people of different ages perform check-ins in different city areas, i.e. teenagers and children (Brown markers) mostly perform check-ins in housing city areas and around schools (Regions 1,2,3,5), while students (Green markers) and working professionals (Blue and Red markers) are concentrated in city center (Region 4). this clearly shows the places where these users spend most of their time; for example, young people spend most of their time studying in school or at home with their parents, while students and working professionals tend to make longer trips from home to office or university. From the small scale map (Figure 5(d)), it can be seen that young users (brown circles) are rarely travel to nearby cities due to their age (Region 3), while young adults (green circles) often make such trips (Regions 1 and 2). These users may be students or young professionals who visit their families during weekends. The temporal aspect of user movement is not considered in our work. However, it is an important factor for user mobility study, thus need to be analyzed in future. In summary, user behaviour patterns are tightly knit to users' demographic profiles, such as gender and age.

3.2 User Demographic Profiling

This section describes the feature extraction from multi-source multi-modal data, and the ensemble learning model we proposed for demographic profiling.

3.2.1 Location Features

We obtained 592 venue categories from Foursquare[6]. Take a user as an example. The user has performed three check-ins in two restaurants and airport, but did not perform any check-in in other venues; then the user's venue category feature vector will consist of 592 real values with c_r (the feature representing restaurant) equals to 2/3 and another c_a (the feature representing airport) equals to 1/3. All the other values of feature vector will be equal to 0.

3.2.2 Text Features

We extracted the following textual features:

- *LDA-based Features.* We merged all the tweets of each user into a document. All documents were projected into a latent topic space using Latent Dirichlet Allocation (LDA) [2]. All users' tweets can be represented as a mixture of different topics. We empirically set $\alpha = 0.5$, $\beta = 0.1$ and utilized KNIME[7] to train the topic model with $T = 50$ topics for $1,000$ iterations. We ultimately built the topic-based user feature vectors.

[6]developer.foursquare.com/categorytree

[7]www.knime.org

Table 2: Category distribution among LDA topics

ID	Categories	LDA Topics
T1	Malay Res-t, Mall, University, Indian Res-t, Aisian Res-t	Food Lovers
T2	Cafe, Airport, Hotel, Coffee Shop, Chinese Res-t	Travelers (Business)
T3	Nightclub, Mall, Food Court, Trade School, Res-t, Coffee Shop	Party Goers
T4	Home, Office, Build., Neighbor-d, Gov. Build., Factory	Family Guys (Youth)
T5	University (Collage), Gym, Airport, Hotel, Fitness Club	Students
T6	Train St., Apartment, Mall, High School, Bus St.	Teenagers (Youth)

- *Linguistic Features.* We extracted LIWC linguistic features [16], which were founded to be a powerful mechanism for age and gender prediction purposes [21]. For each user tweet list, we extracted 71 LIWC features.
- *Heuristically-inferred Features.* We extracted some heuristically inferred features. First, we counted the *number of URLs*, *number of hash tags* and *number of user mentions*, since these features are correlated with user's social network activity level and can, thus, indicate user's age. Second, we counted the *number slang words*, *number of emotion words*[8], *number of emoticons* and computed *an average sentiment score*. These features can be good signals of user personality traits, which in turn are gender and age dependent [21]. Third, we computed the linguistic style features: "Number of repeated characters" in words, "number of misspellings", and "number of unknown to the spell checker words", which are often reflected by user's age.

3.2.3 Visual Features

We mapped each Instagram photo to the pre-defined image concept dictionary. The concept dictionary, comprising of 1000 daily life concepts (such as ball, flower, and chair), is constructed from ImageNet. We trained the concept classifiers based on the concatenation of "Bag-of-visual-words", "Local Binary Patterns"(LBP) and "64-D Color Histogram"(CH) features [5]. For the bag-of-visual-words, we used the difference of Gaussians to detect keypoints in each image and extracted their SIFT descriptors. By building a visual codebook of size 1000 based on K-means, we obtained a 1000-dimensional bag-of-visual-words histogram for each image. We then applied Principal Component Analysis (PCA) to reduce the dimension of image feature space [10] to 50 condensed features.

3.2.4 Age and Gender Prediction

Two most popular approaches for multi-source fusion are early and late fusion. In early fusion strategy, features are concentrated in one long feature vector to feed into the classifier, while late fusion incorporates the independently learned results from each modality [24]. In our work, we have evaluated both approaches. First, we performed an early fusion and trained a Random forest classifier with 145 random trees (the number of random trees was estimated empirically). However, as it can be seen from the experimental results (Table 3), most of the introduced approaches outperform the early fusion models. Inspired

[8]sentiwordnet.isti.cnr.it

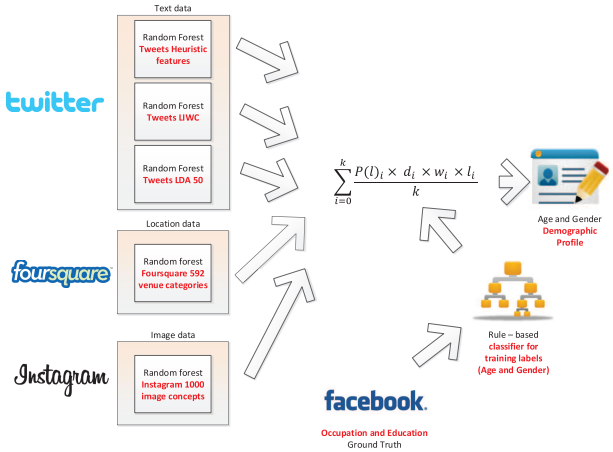

Figure 6: Demographic profiling ensemble.

by the previously reported results [1, 18] and considering the comparably high performance of single source features (See Table 3), we employed the late fusion approach to train models from different combinations of feature sets. The structure of the demographic profiling framework is presented in Figure 6. It can be seen that our framework consists of multi-label Random Forest classifiers trained on five feature types described earlier. The Random Forest classification model was selected due to its well-known ability to perform classification on high dimensional space by randomized feature selection approach. Moreover it was noticed [21] that the Random Forest classifier often outperforms other machine learning approaches, like SVM or Naive Bayes for Age and Gender prediction tasks. The optimal number of random trees for each classifier is found by 10-fold cross-validation. The number of random trees for "Location", "LIWC", "Heuristic", "LDA 50", and "Image concepts" features are set to 140, 120, 75, 70, 130 respectively. To obtain balanced label distribution in the training set, we performed the "SMOTE" data oversampling technique [3].

The five Random Forest classifiers were integrated into the ensemble learning model; with the final classification score for each label computed as:

$$Score(l) = \sum_{i=0}^{k} \frac{P(l)_i \times d_i \times w_i \times l_i}{k},$$

where $P(l)_i$ is a positive prediction confidence for label l of the i-th classification model, d_i is the number of data records that were collected from current user to train the i-th classification model, divided by average number of data records of given type collected for all users, w_i is the accuracy of the i-th classification model, and l_i is the "strength" of the i-th classification model. In our experiments, we set w_i to be equal to the macro accuracy level of the corresponding i classifier, which is estimated after performing 10-fold cross validation on each individual features set. We jointly learned the l_i coefficient for each modality by performing the "Stochastic Hill climbing with Random Restart (SHCR)" optimization[9], which is able to obtain local optimum for non-convex problems and, can, thus, produce reasonable ensemble weighting.

We present an example on how ensemble model works for

[9]en.wikipedia.org/wiki/Hill_climbing

different data sources. Suppose users u_x and u_y have 100 and 300 tweets respectively and we have already obtained the classification accuracy for some Twitter features set as 0.321. Then the score of classifier trained on these features $P(l)_{i_k}$ will be multiplied by: $w_{u_x} = 0.321$, $d_{u_x} = 100/200 = 0.5$ and $w_{u_y} = 0.321$, $d_{u_y} = 300/200 = 1.5$ for u_x and u_y respectively. These score will also be multiplied by the current iteration assignment of l_{i_k}. After all the five classifiers are weighted as described above, the output scores will be averaged and the result will represent the classification ensemble score assigned to label l on the current iteration k. The above procedure will be repeated until the SHCR coveraged, by varying the l_{i_k} parameters on each iteration aiming to maximize the overall accuracy.

3.2.5 *Experimental results*

As we mention before, the number of male and female users in our dataset is similar while the users' age is mainly in the range of between 10 and 40 years old. The similar age distribution was also observed in related works [21][18]. Due to the uneven users' age distribution, we have divided users' on age groups as follows: "< 20", "20-30", "30-40", "> 40".

We evaluated the performance of our approach in terms of *macro accuracy*, which is the average accuracy between all classification labels. The evaluation metric was selected due to its capability of equally treating all data labels in unbalanced datasets, so that the overall performance will not be overpriced. The evaluation of gender and age prediction was based on 222 users who have reported their exact age and perform activity in all three data sources, and can thus be considered as reliable ground truth. The age group prediction model was trained based on other users whose age group was estimated from their education-related information by performing rule-based classification (548 users). According to our experiment, the bias of estimated ages does not exceed ± 2.31 years. It is thus reasonable to use the estimated age for training the age group prediction classifier. The gender prediction model was trained based on $3,544$ users who have reported their gender and are not in evaluation set. The age and gender prediction results are presented in Table 3.

From Table 3, it can be seen that different data sources perform differently in predicting different demographic attributes. For example, in gender prediction task, the LDA features play a crucial role and perform the best among other single sources. The reason is that there exists high difference in interests between male and female users. For example, female users may discuss family and children, while male users tent to tweet about cars and finance. By integrating all text-based features together with visual features, it introduces the best combination for age and gender prediction among all bi-source ensembles, while the location plays a minor role. The possible reason is that there are distinct differences in picture taking preferences and writing style among different users' age groups and genders. For example, older people may take pictures of food and scenery, while younger users could be more emotional in their tweets.

However, the best results for both gender and age group prediction are obtained by late fusion of all features using the Random Forests ensemble. It outperforms all single source models, different combinations of classifiers in ensemble models, state-of-the-arts techniques, and model trained on early fused sources.

Table 3: Performance of demographic profiling in terms of Macro Accuracy. RF - Random Forest; NB - Naive Bayes; SVM - Support Vector Machine

Method	Gender	Age
State-of-the-arts techniques		
SVM Location Cat. (Foursquare)	0.581	0.251
SVM LWIC Text(Twitter)	0.590	0.254
SVM Heuristic Text(Twitter)	0.589	0.290
SVM LDA 50 Text(Twitter)	0.595	0.260
SVM Image Concepts(Instagram)	0.581	0.254
NB Location Cat. (Foursquare)	0.575	0.185
NB LWIC Text(Twitter)	0.640	0.392
NB Heuristic Text(Twitter)	0.599	**0.394**
NB LDA 50 Text(Twitter)	**0.653**	0.343
NB Image Concepts(Instagram)	0.631	0.233
Single-Source		
RF Location Cat. (Foursquare)	0.649	0.306
RF LWIC Text(Twitter)	0.716	0.407
RF Heuristic Text(Twitter)	0.685	**0.463**
RF LDA 50 Text(Twitter)	**0.788**	0.357
RF Image Concepts(Instagram)	0.784	0.366
Multi-Source combinations		
RF LDA + LIWC(Late Fusion)	0.784	0.426
RF LDA + Heuristic(Late Fusion)	0.815	0.480
RF Heuristic + LIWC (Late Fusion)	0.730	0.421
RF All Text (Late Fusion)	0.815	0.425
RF Media + Location (Late Fusion)	0.802	0.352
RF Text + Media (Late Fusion)	**0.824**	**0.483**
RF Text + Location (Late Fusion)	0.743	0.401
All sources together		
RF Early fusion for all features	0.707	0.370
RF Multi-source (Late Fusion)	**0.878**	**0.509**

4. RELATED WORK

User profile learning via multi-sources is a comparably new topic in social media research, and there are little work devoted to it. The emergence of new cross-linking functionality in modern social Web platforms is related to our work. For instance, the work conducted by Wang et al. [25] performed high level comparison between Facebook, Twitter and Foursquare profiles for randomly sampled Foursquare users; they performed both content level and network level analysis. From the network point of view, it was noted that "67 percent of friend pairs on Foursquare are also friends on Facebook, and 43 percent of friend pairs on Foursquare follow each other on Twitter" [25]. The authors observed that active Foursquare users are usually active in Twitter as well, which is consistent with the observations derived from our dataset. In other study, Chen et al. [4] tried to discover cross-site linking regularities in large-scale online social networks dataset. By gathering almost all Foursquare user network graph and applying first order statistical analysis, the authors [4] were able to extract comprehensive knowledge about the Foursquare network structure and users cross social network behaviour. At the same time, Zhong et al. [29] attempted to discover network evolution relations across various social forums datasets (Tencent, Epinion, Facebook, Xiaonei, Twitter, Sina Weibo, Github, StackOverflow). As a result, the proposed by authors statistical approach outperformed the baselines in link prediction task. Finally, Ottoni et al. [15] discovered user generated data relations across Pinterest and Twitter social services; however, the work focused mostly on data description, but not model for profile learning.

5. CONCLUSIONS AND FUTURE WORK

This work presented an initial study of user profile learning via integration of multiple data sources. We constructed a dataset by crawling information from Twitter, Foursquare and Instagram. To facilitate research by other researchers, we have released this dataset [7]. Based upon this dataset, we conducted first-order and higher-order learning for user mobility and demographic profiling in Singapore region. User mobility was modelled as a location-topic distribution, while demographic profile was represented by age and gender attributes. From our experimental outputs, we can draw the conclusions that: multi-source data of the same users mutually complements each other and their appropriate fusion boosts the user profiling performance.

We list some potential research topics that can be conducted on our released dataset:

- *Complete demographic profiling.* Researchers are encouraged to learn other demographics attributes, such as occupation, personality and social status [1].
- *Extended mobility profiling.* In current study, we focused on category-specific user mobility profiling; while it would be useful to incorporate spatio-temporal factors of users' movement [17].
- *Causality pattern extraction.* It is important to discover potential causal relationships between events from multiple data sources [9]. For example, the "flower" image concept could be temporally related with flower shop check-ins or tweets about flowers.
- *Cross-source user identification.* The alignment of user accounts across multiple social resources [23] can benefit from user profile compilation [25].
- *Cross-region user profiling and community matching.* This direction [28] may offer insight on differences and similarities between users' preferences.

6. ACKNOWLEDGMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

The Microsoft Windows Azure Cloud and Microsoft MSDN subscription were provided by "MB-Guide"[10] project and as part of Microsoft BizSpark program.

7. REFERENCES

[1] E. Amigó, J. Carrillo-de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, and D. Spina. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. In *Information Access Evaluation. Multilinguality, and Interaction*. 2014.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

[4] Y. Chen, C. Zhuang, Q. Cao, and P. Hui. Understanding cross-site linking in online social networks. In *Proceedings of the Workshop on Social Network Mining and Analysis*, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] M. Duggan and J. Brenner. *The demographics of social media users*. 2013.

[7] A. Farseev and T.-S. Chua. NUS-MSS: A Multi-Source Social Dataset from National University of Singapore. `http://lms.comp.nus.edu.sg/research/NUS-MULTISOURCE.htm`, 2015.

[8] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 2008.

[9] R. Jain and L. Jalali. Objective self. *MultiMedia, IEEE*, 2014.

[10] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[11] P. Kasemsuppakorn and H. A. Karimi. Pedestrian network data collection through location-based social networks. In *IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing*, 2009.

[12] L. Nie, M. Wang, Z.-J. Zha, G. Li, and T.-S. Chua. Multimedia answering: Enriching text qa with media information. In *Proceedings of the International ACM SIGIR Conference*, 2011.

[13] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust, International Conference on Social Computing*, 2012.

[14] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *The International Conference on Weblogs and Social Media*, 2011.

[15] R. Ottoni, D. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. 2014.

[16] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[17] Y. Qu and J. Zhang. Trade area analysis using user generated mobile location data. In *Proceedings of the International Conference on World Wide Web*, 2013.

[18] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, 2013.

[19] A. Sadilek, H. A. Kautz, and V. Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Association for the Advancement of Artificial Intelligence*, 2012.

[20] S. Shang, D. Guo, J. Liu, and K. Liu. Human mobility prediction and unobstructed route planning in public transport networks. In *IEEE International Conferebce on Mobile Data Management*, 2014.

[21] E. Stamatatos, W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M. A. Sanchez-Perez, and A. Barrón-Cedeño. Overview of the author identification task at pan 2014. *Analysis*, 2014.

[22] C. Von Der Weth, V. Hegde, and M. Hauswirth. Virtual location-based services: Merging the physical and virtual world. *IEEE International Conference on Web Services*, 2013.

[23] J. Vosecky, D. Hong, and V. Y. Shen. User identification across multiple social networks. In *Networked Digital Tech.*, 2009.

[24] M. Wang, H. Li, D. Tao, K. Lu, and X. Wu. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing*, 2012.

[25] P. Wang, W. He, and J. Zhao. A tale of three social networks: User activity comparisons among facebook, twitter, and foursquare. 2014.

[26] X. Wang, Y.-L. Zhao, L. Nie, Y. Gao, W. Nie, Z.-J. Zha, and T.-S. Chua. Semantic-based location recommendation with multimodal venue semantics. *IEEE Transactions on Multimedia*, 2015.

[27] Y.-L. Zhao, Q. Chen, S. Yan, T.-S. Chua, and D. Zhang. Detecting profilable and overlapping communities with user-generated multimedia contents in lbsns. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2013.

[28] Y.-L. Zhao, L. Nie, X. Wang, and T.-S. Chua. Personalized recommendations of locally interesting venues to tourists via cross-region community matching. *ACM Transactions on Intelligent Systems and Technology*, 2014.

[29] E. Zhong, W. Fan, Y. Zhu, and Q. Yang. Modeling the dynamics of composite social networks. In *Proceedings of the ACM International KDD Conference*, 2013.

[10] mb-guide.com