Social Media Computing

Lecture 6: Case Study – Multi-Source Profile Learning

Lecturer: Aleksandr Farseev
E-mail: farseev@u.nus.edu
Slides: http://farseev.com/ainlfruct.html
References


• А. Фарсеев, Н. Жуков, И. Государев, и Ю. Заричняк. Разработка Кросплатформенной Рекомендательной Системы на Основе Извлечения Данных из Социальных Сетей Компьютерные Инструменты в Образовании. June 2014.
What is user profile?
What is human mobility?

- Mobility - contemporary paradigm, which explores various types of people movement.
What is human mobility?

• Mobility - contemporary paradigm, which explores various types of people movement.

• The movement of people
• The quality or state of being mobile
• (Physiology) the ability to move physically
• (Sociology) movement within or between social classes and occupations
• (Chess) the ability of a chess piece to move around the board
Why human mobility?

• **Urban planning:** understand the city and optimize services
• **Mobile applications and recommendations:** study the user and offer services
Mobility can describe people.
**Marketing**
Trade are analysis
Demography and interest - based marketing

**Wellness**
Health group prediction
Lifestyle recommendation

**Advertisement**
Demography and interest - based personalized advertisement

**Assistance**
Activity recommendation,
Venue recommendation,
Etc.

- Tent to stay at home, visit local pubs and shopping mall daily.
- Medium overweight, potential hypertonia and diabetes.
- Advertise new Beer brand and new car models.
- Morning excursive with medium intensity.
User profile: Mobility + Demography

User profile

Mobility profile
- Location preference
- Movement patterns

Demographic profile
- Age
- Gender
- Personality
- Occupation
Multiple sources describe user from multiple views

More than 50% of online-active adults use more than one social network in their daily life*

*According Paw Research Internet Project's Social Media Update 2013 (www.pewinternet.org/fact-sheets/social-networking-fact-sheet/)
Multiple sources describe user from multiple views
Research Problems

➢ Multi-source user profiling:
  • Geographical user mobility profiling
  • User demographic profiling
  • Data incompleteness
  • Multi–source multi–modal data integration
Multi-source dataset: NUS-MSS*

*http://lms.comp.nus.edu.sg/research/NUS-MULTISOURCE.htm
NUS-MSS: Data sources

- **Foursquare**: Biggest LBSN
- **Instagram**: Biggest photo sharing service
- **Twitter**: Biggest English-speaking microblog
NUS-MSS: Data collection
NUS-MSS: Dataset Description

11,732,489 TWEETS
366,268 CHECK-INS
263,530 IMAGES

FROM 7,023 USERS
NUS-MSS: Dataset Description

2,973,162 TWEETS
127,276 CHECK-INS
65,088 IMAGES

FROM 5,503 USERS
NUS-MSS: Dataset Description

5,263,630 TWEETS
304,493 CHECK-INS
230,752 IMAGES

FROM 7,957 USERS
NUS-MSS: Dataset Statistics in Singapore
Demographic profiling
User profile: Mobility + Demography

User profile

Mobility profile
- Location preference
- Movement patterns

Demographic profile
- Age
- Gender
- Personality
- Occupation
Data representation

• Linguistic features
  – LIWC
  – User Topics
• Heuristic features
  – Writing behavior

A text analysis software.

An efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples. Can be highly related to one's demography.
Data representation

- **Linguistic features**
  - LIWC
  - User Topics
- **Behavioral features**
  - Writing behavior

Users of similar gender and age may talk about similar topics e.g. female users – about shopping, male – about cars; youth – about school while elderly – about health.

LDA word distribution over 50 topics for collected Twitter timeline.
Data representation

- **Linguistic features**
  - LIWC
  - User Topics
- **Heuristic features**
  - Writing behavior

<table>
<thead>
<tr>
<th>Feature name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of hash tags</td>
<td>Number of hash tags mentioned in message</td>
</tr>
<tr>
<td>Number of slang words</td>
<td>Number of slang words one use in his tweets. We calculate number of slang words / tweet and compute average slang usage</td>
</tr>
<tr>
<td>Number of URLs</td>
<td>Number of URL’s one usually use in his/her tweets</td>
</tr>
<tr>
<td>Number of user mentions</td>
<td>Number of user mentions – may represent one’s social activity</td>
</tr>
<tr>
<td>Number of repeated chars</td>
<td>Number of repeated characters in one tweets (e.g. noooooooo, wahhhhhhh)</td>
</tr>
<tr>
<td>Number of emotion words</td>
<td>Number of words that are marked with not – neutral emotion score in Sentiment WordNet</td>
</tr>
<tr>
<td>Number of emoticons</td>
<td>Number of common emoticons from Wikipedia article</td>
</tr>
<tr>
<td>Average sentiment level</td>
<td>Module of average sentiment level of tweet obtained from Sentiment WordNet</td>
</tr>
<tr>
<td>Average sentiment score</td>
<td>Average sentiment level of tweet obtained from Sentiment WordNet</td>
</tr>
<tr>
<td>Number of misspellings</td>
<td>Number of misspellings fixed by Microsoft Word spell checker</td>
</tr>
<tr>
<td>Number Of Mistakes</td>
<td>Number of words that contains mistake but cannot be fixed by Microsoft Word spell checker</td>
</tr>
<tr>
<td>Number of rejected tweets</td>
<td>Number of tweets where 70% of words either not in English or cannot be fixed by Microsoft Word spell checker</td>
</tr>
<tr>
<td>Number of terms average</td>
<td>Average number of terms per / tweet</td>
</tr>
<tr>
<td>Number of Foursquare check-ins</td>
<td>Number of Foursquare check-ins performed by user</td>
</tr>
<tr>
<td>Number of Instagram medias</td>
<td>Number of Instagram medias posted by user</td>
</tr>
<tr>
<td>Number of Foursquare tips</td>
<td>Number of Foursquare Tips that user post in a venue</td>
</tr>
<tr>
<td>Average time between check-ins min</td>
<td>Average time between two sequential check-ins - represents Foursquare user activity frequency</td>
</tr>
</tbody>
</table>

As we mention from our research – user’s writing behavioral patterns are highly correlated with e.g. age (individuals from 10 – 20 years old are making two times less grammatical errors than 20 -30 years old individuals)
Data representation

- Location features
  - Location semantics
  - Location topics

Venue semantics such as venue categories can be related to users demography. E.g. individuals who tend to visit night clubs are usually belong to 10 – 20 or 20 – 30 years old age groups.

For case when user performed check-ins in two restaurants and airport but did not perform check-ins in other venues:

<table>
<thead>
<tr>
<th>Category₁</th>
<th>…</th>
<th>Categoryrestaurant</th>
<th>…</th>
<th>Categoryairport</th>
<th>…</th>
<th>Categoryₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>U₁</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>…</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Uₙ</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
Data representation

- **Image features**
  - **Image concept**

Extracted image concepts may represent user interests and be related to one’s demography. For example, a female user may take pictures of flowers, food, while a male – of cars or buildings.

*The concept learning Tool was provided by Lab of Media Search LMS. It was evaluated based on ILSVRC2012 competition dataset and performed with average accuracy @10 - 0.637*
Ensemble learning

\[
\sum_{i=0}^{k} \frac{P(l_i) \times d_i \times w_i \times l_i}{k}
\]

- **Text data**
  - Random Forest
    - Tweets
      - Additional features
  - Random Forest
    - Tweets LIWC
  - Random Forest
    - Tweets LDA 50

- **Location data**
  - Random forest
    - Foursquare 546 venue categories

- **Image data**
  - Random forest
    - Instagram 1000 image concepts

**Occupation and Education**
Ground Truth

- Age and Gender
  - Demographic Profile
- Rule-based classifier for training labels
Ensemble learning

\[ d_i \times w_i \times l_i \]

**AGG, GENDER CONFIDENCE SCORES**

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

- \( P(l)_i \) - model prediction confidence
- \( d_i \) - normalized data records number
- \( w_i \) - model trust weight
- \( l_i \) - model “strength” – learned by “Hill Climbing” optimization with step 0.0ε
Ensemble learning details

- According to our evaluation, the bias of estimated ages does not exceed ±2.28 years. It is thus reasonable to use the estimated age for age group prediction task.
- We have adopted SMOTE* oversampling to obtain balanced age-group labeling.
- By performing 10-fold cross validation, we determine the optimal number of constructed random trees for each classifier with iteration step equal to 5 as 45, 25, 35, 40, 105 random trees for Random Forest Classifiers learned based on location, LIWC, heuristic, LDA 50, and image concept features respectively.
- We jointly learn the $I_i$ model “strength” coefficient by performing “Hill Climbing” optimization** with step 0.05. The randomized “Hill Climbing” approach is able to obtain local optimum for non-convex problems and, thus, can produce resolvable ensemble weighting.

---


**An iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.
## Experimental results (Singapore)

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

**AGE GROUPS:**

- **< 20 YEARS OLD**
- **20 – 30 YEARS OLD**
- **30 – 40 YEARS OLD**
- **> 40 YEARS OLD**
Demographic mobility
User profile: Mobility + Demography

User profile

Mobility profile
- Location preference
- Movement patterns

Demographic profile
- Age
- Gender
- Personality
- Occupation
Geographical user mobility: users movement (city level)
Geographical user mobility: users movement (city level)

• Singapore population is concentrated in several regions, which represent peoples' housing (Regions 2 and 3) and working (Region 3) areas.
• There are some regions where male (Blue markers) user check-in density is much higher than female (Pink markers).
Geographical user mobility: users movement (region level)
Both female and male users often perform trips to nearby cities for shopping and leisure purposes (Regions 1, 2, 4, 5).

Regions 2 and 3 are popular among female users, since 2 is “Malacca resorts”, while 3 – National park. Both regions are famous by it’s family time spending facilities.
Geographical user mobility: users movement (city level)
Geographical user mobility: users movement (city level)

- Teenagers and children (Brown markers) mostly perform check-ins in housing city areas and around schools (Regions 1,2,3,5).
- Students (Green markers) and working professionals (Blue and Red markers) are concentrated in city center (Region 4).
Geographical user mobility: users movement (region level)
Geographical user mobility: users movement (region level)

- Young users (brown circles) are rarely travel to nearby cities due to their age (Region 3).
- 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.
Dataset Statistics: Content

(a) 

(b) 

(c) 

(d)
Geographical user mobility: venue semantics profiling

- We extract location topics based on venue categories to model user mobility semantics.

Location topics may serve as an user interest clusters for distinguishing user demography attributes such as age or gender.

LDA word distribution over 6 topics for collected Foursquare check-ins.

Every venue category

<table>
<thead>
<tr>
<th>ID</th>
<th>Categories</th>
<th>LDA Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>Malay Res-t, Mall, University, Indian Res-t, Aisan Res-t</td>
<td>Food Lovers</td>
</tr>
<tr>
<td>T2</td>
<td>Cafe, Airport, Hotel, Coffee Shop, Chinese Res-t</td>
<td>Travelers (Business)</td>
</tr>
<tr>
<td>T3</td>
<td>Nightclub, Mall, Food Court, Trade School, Res-t, Coffee Shop</td>
<td>Party Goers</td>
</tr>
<tr>
<td>T4</td>
<td>Home, Office, Build., Neighbor-d, Gov. Build., Factory</td>
<td>Family Guys (Youth)</td>
</tr>
<tr>
<td>T5</td>
<td>University (Collage), Gym, Airport, Hotel, Fitness Club</td>
<td>Students</td>
</tr>
<tr>
<td>T6</td>
<td>Train St., Apartment, Mall, High School, Bus St.</td>
<td>Teenagers (Youth)</td>
</tr>
</tbody>
</table>
Geographical user mobility: venue semantics profiling

![Bar chart showing the proportion of different user categories for three users.](chart.png)
Geographical user mobility: venue semantics profiling

- Male users more often do shopping than male, while female users often show-up in job-related venues.
- > 30 years old users often show-up in dining-related places, while < 20 – often visit education-related venues.
Venue Category Recommendation
Which category(s) of 4sq venues to go next?
Evaluation – split time on train and test periods
We use Collaborative Filtering (CF)
Multi-Source re-ranking

- Seeking to boost the recommendation performance, we developed late fusion re-ranking approach. We linearly combined the outputs from different sources, where the weight of each source is learned based on a stochastic hill climbing with random restart (SHCR)

\[
Rank_f(item_i) = \frac{1}{n} \sum_{s=1}^{n} \frac{w_s}{Rank_s(item_i)}
\]

- where \(Rank_s(item_i)\) is the rank of \(i\)th item in recommendation list for source \(s\); \(w_s\) corresponds to the weight of the source \(s\); \(n\) is a total number of sources (in our case, \(n = 4\)). The venue categories in final recommendation list are sorted in increasing order according to their rank.
To measure the recommendation performance we use F-measure@K, where $P@K$ and $R@K$ are precision and recall at $K$, respectively, and $K$ indicates the number of selected items from the top of the recommendation list.
What we are doing now?
Something much bigger…
You, actually can join us as Intern or Research Engineer

http://next.comp.nus.edu.sg/opportunities
Extended User Profiling

- Extended Demographic Profiling:
  - Occupation detection;
  - Personality detection;
  - Social status detection.

- Extended Mobility Profiling:
  - User communities detection and profiling (in terms of demographics, movement patterns, multi-source interests) – in progress
  - Cross-region mobility profiling (comparison of users’ mobility across different regions and cultures) – in progress
Sensor Data Incorporation & Wellness Research

- Wellness lifestyle recommendation via:
  - Chronic diseases tendency prediction
  - Cross-source causality relationships analysis (just like Ramesh Jain proposed*)

Future work: How the framework may look like
Other tasks based could be approached

1. Demographic profile learning
2. Multi-source data fusion
3. Individual and group mobility analysis
4. Cross-source user identification
5. Cross-region user community detection
6. Cross-source causality relationships extraction
7. Users’ privacy-related and cross-disciplinary research
User Profile Learning in Wellness Domain
People are often now aware of their wellness problems
It is not easy to follow doctor’s prescriptions
Personal and continuous assistance is necessary
Continuous patients monitoring for better prescription
Weight Problems Consequences*

- All-causes of death (mortality)
- **High blood pressure (Hypertension)**
- High LDL cholesterol, low HDL cholesterol, or high levels of triglycerides (Dyslipidemia)
- **Type 2 diabetes**
- **Coronary heart disease**
- Stroke
- Gallbladder disease
- **Osteoarthritis (a breakdown of cartilage and bone within a joint)**
- Sleep apnea and breathing problems
- **Some cancers (endometrial, breast, colon, kidney, gallbladder, and liver)**
- Low quality of life
- **Mental illness such as clinical depression, anxiety, and other mental disorders**
- Body pain and difficulty with physical functioning⁶

User Profiling: Next Step

User profile

Wellness profile
- Diabetes
- Asthma
- Obesity

Mobility profile
- Location preference
- Movement patterns

Demographic profile
- Age
- Gender
- Personality
- Occupation
Data sources describe user in multiple views
Research Problems

- Multi-source user profiling:
  - Wellness profiling
  - Predict one’s obesity level by leveraging multi-source multi-modal data (in other words – BMI prediction)
  - Data gathering, noise, sensitivity and incompleteness
  - Multi-source multi-modal data integration
Summary

• We constructed and released a large multi-source multi-modal cross-region “NUS-MSS” dataset;
• We conducted first-order and higher-order learning for user mobility and demographic profiling;
• New multi-modal features were proposed for a demographic profile learning.
• Based on our experimental results, we can conclude that multi-source data mutually complements each other and their appropriate fusion boosts the user profiling performance.
• We believe that we can predict one’s social media data and the data from wearable sensors.
Next Lesson

• Wrap Up
Short KNIME* Tutorial

1. *We select Knime since it could be used even without any programming experience.

2. Download and Install Knime together with all extensions from here: [http://knime.org](http://knime.org)

3. Go to [http://nusmultisource.azurewebsites.net](http://nusmultisource.azurewebsites.net)

Further steps are based on London dataset, but it is applicable to all the other sources.
Open Knime and Create New Workflow
Add two “CSV Reader” nodes – one for features, one for ground truth
Set up CSV readers to read features file and ground truth file, execute the workflow.
Set up CSV readers to read features file and ground truth file, execute the workflow.
Add two “Row Filter” nodes – to separate users with real age indicated and without by excluding and including missing rows, respectively.
Add two “Joiner” nodes – to join (Inner Join by RowId) ground truth and features in one table for train and one table for test set.
And i.e. “Naïve Bayes Learner” and “Naïve Bayes Predictor” nodes to train and test data flows, respectively. Set up learner to train based on i.e. “gender”.
And “Scorer” node to the output of “Predictor” and set it up to compare predicted results and ground truth. Execute the workflow.
That’s All! The evaluation metrics are computed in “Scorer” node and can be flushed to file (“CSV Writer”) or to UI. Try it different features.
Summary

• Knime is easy to use but you must understand the principles of each node you used.
• Knime is not capable to solve custom tasks easily, but very helpful to test assumptions or run baselines.
• Sometimes it is useful to implement a model from scratch. It may help to understand results better, so we encourage it.
• You have two days to implement your assignment and prepare presentations. You can use whatever software (language) you like. Just make it work on time and present to us.