Профилирование атрибутов пользователей из множества источников данных различной модальности

Aleksandr Farseev
E-mail: farseev@u.nus.edu
Slides: http://farseev.com/Presentations.html
Harvesting Multiple Sources for User Profile Learning: a Big Data Study

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua
References


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
If we want to know more?

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.

AGE, GENDER
40, MALE
User profile: Mobility + Demography

User profile

Mobility profile
- Location preference
- Movement patterns

Demographic profile
- Age
- Gender
- Personality
- Occupation
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://nusmulsitource.azurewebsites.net
NUS-MSS: Data sources

**FOURSQUARE**
Biggest LBSN

**Twitter**
Biggest English-speaking microblog

**Instagram**
Biggest photo sharing service
NUS-MSS: Data collection
NUS-MSS: Dataset Description

- 11,732,489 tweets
- 366,268 check-ins
- 263,530 images

From 7,023 users in Singapore
NUS-MSS: Dataset Description

Twitter

London

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

(a) % of users vs. Number of tweets
(b) % of users vs. Number of shared media contents
(c) % of users vs. Number of check-ins
(d) % of users vs. Number of tweets
(e) % of users vs. Number of shared media contents
(f) % of users vs. Number of check-ins
Demographic profiling
# User profile: Mobility + Demography

<table>
<thead>
<tr>
<th>Mobility profile</th>
<th>Demographic profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location preference</td>
<td>Age</td>
</tr>
<tr>
<td>Movement patterns</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Personality</td>
</tr>
<tr>
<td></td>
<td>Occupation</td>
</tr>
</tbody>
</table>
Data representation

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

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.

A text analysis software.

Dictionary Word category

![Chart showing percentage distribution of word categories](chart.png)
Words usage study for personality profiling

The smallest, most commonly used, most forgettable words serve as windows into our thoughts, emotions, and behaviors.

- Task – Word usage analysis* and correlation with personality
- Data – Various essays and questionnaires
- Approach – manual personality-related dictionaries construction
- Findings:
  - Certain word usage statistics are good indicators for human personality profiling

LIWC

- Count occurrences of each LIWC category
- Each document $D$ for user $u$ is represented as a distribution among 74 LIWC categories: $D_u = \left( \frac{LIWC_{1u}}{N}, \frac{LIWC_{2u}}{N}, \frac{LIWC_{3u}}{N}, \ldots, \frac{LIWC_{74u}}{N} \right)$
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.
Topic Modeling

- Methods for automatically organizing, understanding, searching and summarizing large electronic archives.
- Uncover hidden topical patterns in collections.
- Annotate documents according to topics.
- Using annotations to organize, summarize and search.
- Widely popular approach: Latent Dirichlet Allocation (LDA)*

Seeking Life’s Bare (Genetic) Necessities

**COLD SPRING HARBOR, NEW YORK**—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetics numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arctady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

---


---

**Topic Modeling -3**

---

**Topics**
- gene 0.04
- dna 0.02
- genetic 0.01
- life 0.02
- evolve 0.01
- organism 0.01
- brain 0.04
- neuron 0.02
- nerve 0.01
- data 0.02
- number 0.02
- computer 0.01

---

**Documents**

**Seeking Life’s Bare (Genetic) Necessities**

*COLD SPRING HARBOR, NEW YORK—* How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses, concluded that today’s organism can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough. Although the numbers don’t match precisely, those predictions are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Anderson, a genetics professor at Stanford University in Stanford, who arrived at the 800 number. But coming up with an answer may be more than just a simple matter of numbers. “The number of genes is more of a benchmark,” Anderson explains. “It may be a way of organizing any newly sequenced genome,” she says.

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland, comparing an ancient gene set with a modern one—562 genes

---

**Topic proportions and assignments**

---

• Only documents are observable (*All user’s tweets are in one document for every user*).

• Infer underlying topic structure:
  • Topics that generated the documents.
  • For each document, distribution of topics.
  • For each word, which topic generated the word.
• Each document D (one user - one document) is represented as a distribution among N LDA topics: \( D_u = \{ LDA_{1u}, LDA_{2u}, LDA_{3u}, \ldots, LDA_{Nu} \} \)
Data representation

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

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)

<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>
Data representation

- Location features
  - Location semantics

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.

We map all Foursquare check-ins to Foursquare categories from category hierarchy.

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></th>
<th>$cat_1$</th>
<th>$\cdots$</th>
<th>$cat_{rest}$</th>
<th>$\cdots$</th>
<th>$cat_{air}$</th>
<th>$\cdots$</th>
<th>$cat_{517}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\cdots$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
</tr>
<tr>
<td>$u_N$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
</tr>
</tbody>
</table>
Data representation

- Image features
  - Image concept learning

Extracted image concepts may represent user interests and be related to one's demography. For example, female users may take pictures of flowers, food, while 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
Let image $I$ be of dimension $p \times q$

- For ease in representation, need to quantize $p \times q$ potential colors into $m$ colors (for $m << p \times q$)
- For pixel $p = (x,y)$, the color of pixel is denoted by $I(p) = c_k$

Construction of Color Histogram

- Extract color value for each pixel in image
- Quantize color value into one of $m$ quantization levels

Collect frequency of color values in each quantization level, where each bin corresponds to a color in the quantized color space
Color Histogram -2

• Thus, image is represented as a color histogram $H$ of size $m$
  • where $H[i]$ gives # of pixels at intensity level $i$
• For example:

![Images showing different color histograms]

- Normalize $H$ to $NH$ by dividing each entry by size of image $p \times q$
Basic idea: use edge orientation representation

- Obtain interest points from scale-space extrema of differences-of-Gaussians (DoG)
- Take 16x16 square window around detected interest point
- Compute edge orientation for each pixel
- Throw out weak edges (threshold gradient magnitude)
- Create histogram of surviving edge orientations

http://www.scholarpedia.org/article/Scale_Invariant_Feature_Transform
Detected Interest Points
Scale Invariant Feature Transform (SIFT) descriptor - 2

- A popular descriptor:
  - Divide the 16x16 window into a 4x4 grid of cells (we show the 2x2 case below for simplicity)
  - Compute an orientation histogram for each cell
  - 16 cells X 8 orientations = 128 dimensional descriptor
Scale Invariant Feature Transform (SIFT) descriptor -3

- Invariant to
  - Scale
  - Rotation

- Partially invariant to
  - Illumination changes
  - Camera viewpoint
  - Occlusion, clutter
Examples of SIFT matching

80 matches

34 matches
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the cerebral cortex. However, Hubel and Wiesel demonstrated that the visual cortex was a more complex system to speak, upon which the image in the eye was projected.

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004's $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. The figures are likely to further annoy the US, which has long argued that China's exports are unfairly helped by a deliberately undervalued yuan.
Overall Representation: as Bag of Visual Words

- Can images be represented as Bag-of-Visual Words?

- Idea: quantize SIFT descriptors of all training images to extract representative visual words!
Overall Representation: as Bag of Visual Words -3

Step 1: Extract interest points of all training images
Overall Representation: as Bag of Visual Words -4

Step 2: Features are clustered to quantize the space into a discrete number of visual words.
Overall Representation: as Bag of Visual Words -5

Get the final visual word Tree

Hierarchical K-means clustering
Overall Representation: as Bag of Visual Words

Step 3: Summarize (represent) each image as histogram of visual words

and use as basis for matching and retrieval!
Overall Representation: as Bag of Visual Words -7

- Another example:

![Visual Word Histogram](image)

frequency

Visual words codebook
Concept Recognition: Bag-of-Word Model

BASIC IDEA:

- Representative set of images in each category is collected
- An image is represented by a collection of “visual words”

- Object categories are modeled by the distributions of these visual words
~10,000 to 30,000
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

- **Twitter**
- **Foursquare**
- **Instagram**
- **Facebook**

- Age and Gender
  - Demographic Profile
- Rule-based classifier for training labels
- Occupation and Education
  - Ground Truth
Ensemble learning

\[ d_i \times w_i \times l_i \]

**AGE, GENDER CONFIDENCE SCORES**

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

- \( P(l)_i \) - 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 0.05
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>0.483</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>
</tbody>
</table>
| RF Multi-source (Late Fusion)               | **0.878** | **0.509**

**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)
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
Geographical user mobility: venue semantics profiling

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

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

Every venue category

| ID | Categories                                      | LDA Topics          |
|----|------------------------------------------------|
| T1 | Malay Res-t, Mall, University, Indian Res-t, Asian 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)    |

Location topics may serve as user interest clusters for distinguishing user demography attributes such as age or gender.
Geographical user mobility: venue semantics profiling
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.
Semantic user mobility:

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).

\[ \text{Rank}_f(\text{item}_i) = \frac{1}{n} \sum_{s=1}^{n} \frac{w_s}{\text{Rank}_s(\text{item}_i)} \]

- where \( \text{Rank}_s(\text{item}_i) \) 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.
Results

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 else can be done?
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
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 functioning

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 and venue recommendation performance.

• We believe that we can predict one’s social media data and the data from wearable sensors.
You, actually can join us as Intern or Research Engineer

[http://next.comp.nus.edu.sg/opportunities](http://next.comp.nus.edu.sg/opportunities)

E-mail to: farseev@u.nus.edu
As Research Intern

The Project

User profile learning, such as mobility, wellness, or 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. In our project, we construct large-scale multi-source multi-modal datasets, apply machine learning techniques on it to infer various user profile attributes, deploy large-scale data analytics platforms.

Requirements and Benefits

Requirements:

- Bachelor/Master/PhD full-time student (or graduate)
- Strong programming background (C#, Java, Python, R, MathLab, etc.)
- Strong mathematical background (Probability Theory, Linear Algebra, Convex Optimization)
- Machine learning background

Benefits:

- 1 year (extendable) Internship in NUS - leading Asia's university and of Top10 in the world
- Internship in leading Asia’s Social Media Lab LMS@NUS
- Opportunity to publish your work in Top-ranked journals and conferences
- Finance allowance provided.
- Opportunity to apply for PhD in NUS based on the internship results

http://nusmulsitource.azurewebsites.net
As Java Research Engineer

Types of UGC’s Gathered

**Type 1:**
- Contents: Tweets; Comments; cQA
- User Comments/cQA
- Social News

**Type 2:**
- Images/Videos

**Type 3:**
- Location/Check-in Apps

**Type 4:**
- Structured Data

NExT Social Observatory

Structured Contents

People, Domain, Social, Loc & Mobile

Users

http://live.nextcenter.org
You, actually can join us as Intern or Research Engineer

http://next.comp.nus.edu.sg/opportunities

E-mail to:

farseev@u.nus.edu
AINL-ISMW FRUCT OPEN DAY

Social Media Analytics: What has changed over the last 5 years.

Registration:
You Name and Job Place to: office@ainlfruct.com or +7 (921) 438-80-77

Tat-Seng CHUA
Chair Professor School of Computing, National University of Singapore

7-9, Universitetskaya nab. (Здание Двенадцати Коллегий)
Start: 11:00 AM