A brief review of User Profiling Techniques

Web Data Mining
MSc DWS AUTH, 2020-2021

Georgios Michoulis 82
Styliani Kyrama 76
Vasileios Moschopoulos 59
Dimitrios Tourgaidis 66

SCHOOL OF INFORMATICS
Data & Web Science MSc Program

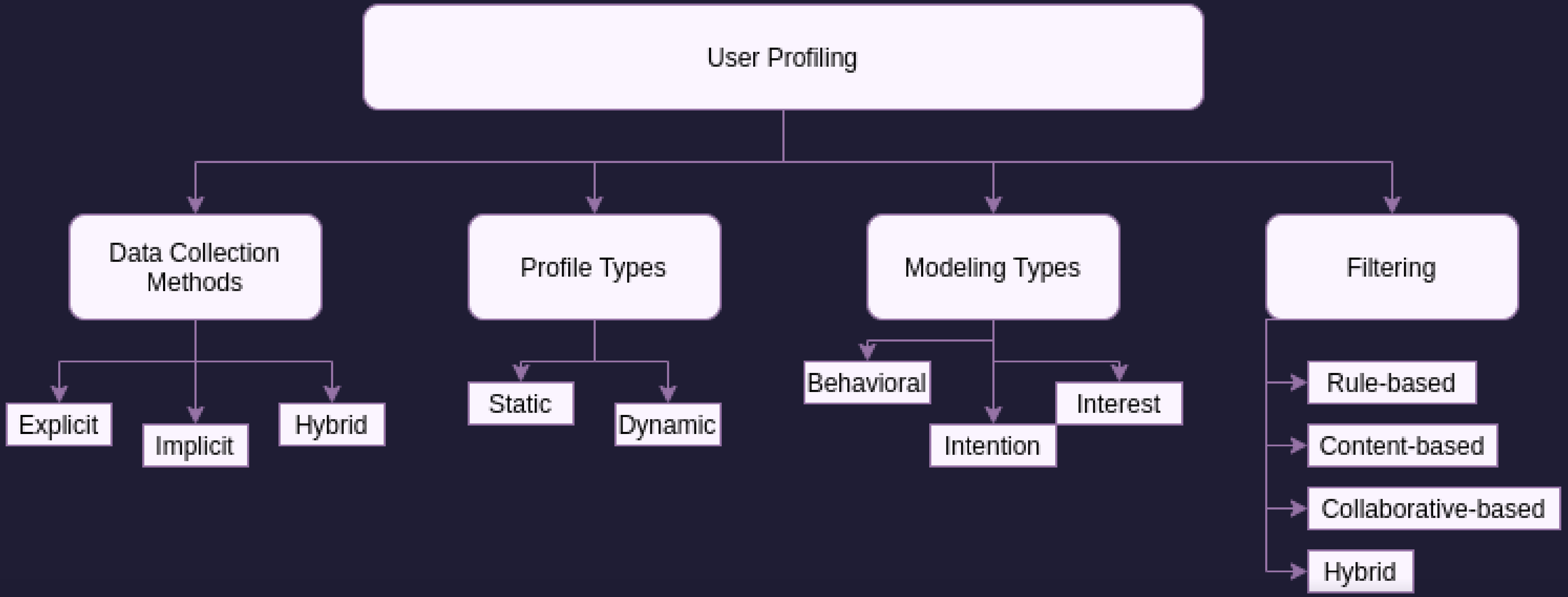
Outline

- Introduction – User Profiling
- Background
- Techniques for User Profiling
- Conclusions

User Profiling

- Modeling of users
- Obtaining useful information (demographic or social characteristics, etc.), in order to understand the user
- Challenges:
  - users' habits, interests, behavior **change over time**
  - **Cold-start** problem *(especially in recommender systems)*
  - **Multiple sources** from which we can gather information
  - **GDPR**
- Domains:
  - E-commerce
  - Healthcare
  - Cybersecurity
  - Banking
  - Social Networks

Background


User Profiling
Data Collection Methods
Profile Types
Modeling Types
Filtering
Explicit
Implicit
Hybrid
Static
Dynamic
Behavioral
Intention
Interest
Rule-based
Content-based
Collaborative-based
Hybrid

Techniques

Vector-space modeling for news recommendation

Dataset

- Keywords from tweets and re-tweets
- Nouns and hashtags
- To each word is assigned a weight (0≤weight≤1)
- This weight expresses the frequency of each word in user's tweets
- Hashtags are more important

User Profiling

- A bag-of-words for each user
- Each user is represented by a vector of weights

Keyword	Apple	Samsung	Google	Twitter	...
User 1	0.01744	0.01161	0.02037	0.01381	...
User 2	0	0.00013	0.00019	0.00034	...
User 3	0.00108	0.00111	0	0.00212	...
User 4	0.00777	0.01576	0.01176	0	...
...	...	...	...	...	...

Tweets recommendation system

- Suggests tweets to the users that are more likely to be re-tweeted by them

Dataset

- Users' tweets
- Users' friends tweets
  - Tweets of influential friends
  - Tweets of less influential friends
    - Re-tweeted
    - Not re-tweeted
  - Tweets of no influential friends

User Profiling

- Users' tweets, influential friends' tweets and less influential friends' re-tweeted tweets
- Vector-space model
- Cosine similarity between Users' profile and suggested tweets.
- 6 user profile models, with final, the most accurate

Rule-based news recommendation

- A system that recommends viral news from Facebook and Twitter based on their category

Dataset

- Users' browsing history
  - Visited URLs
  - Search engine queries
- Metrics
  - Click frequent count (CF)
  - Specific search query count (SSQ)

User Profiling

- Proxy agent that build implicit user profile (IUP)
- CF and SSQ are classified into "Low", "Medium" and "High" and given as inputs to the system
- Harcoded if-then rules, like "if Cf is low and SSQ is low for X, then the user is not interest in X"
- For each category, the if-then rules are implemented by domain experts
- Systems output: Not-interested, interested and Highly-interested

Building semantic user profile for Polish web news portal

- Gender prediction
- Secondary: similar news article retrieval
- Dataset
  - Custom corpus of 500,000 Onet articles that had pageviews within a 14-day period
  - External corpus of Polish Wikipedia and National Polish Language Corpora NKJP
  - 103,519 anonymous users split into two nearly equal gender classes, associated with their browsed articles within the 14-day period

- User profiling
  - 6 different models
  - User profiles built using the average of their browsed article vectors
  - Examination of optimal model type and size of article representation vector

Model name	Train data	Preprocessing	Embeddings	Transformed data
LDA_article	Onet articles	stopwords, lemmatization	LDA	aricle text
LDA_title	Onet articles	stopwords, lemmatization	LDA	article title
wv_article	Onet articles	stopwords, lemmatization	Word2Vec	article text
wv_title	Onet articles	stopwords, lemmatization	Word2Vec	article title
wv_article_forms	Onet articles	stopwords	Word2Vec	article text
wv_wiki_nkjp_forms	External corpus	stopwords	Word2Vec	article text

Age group classification of Twitter users

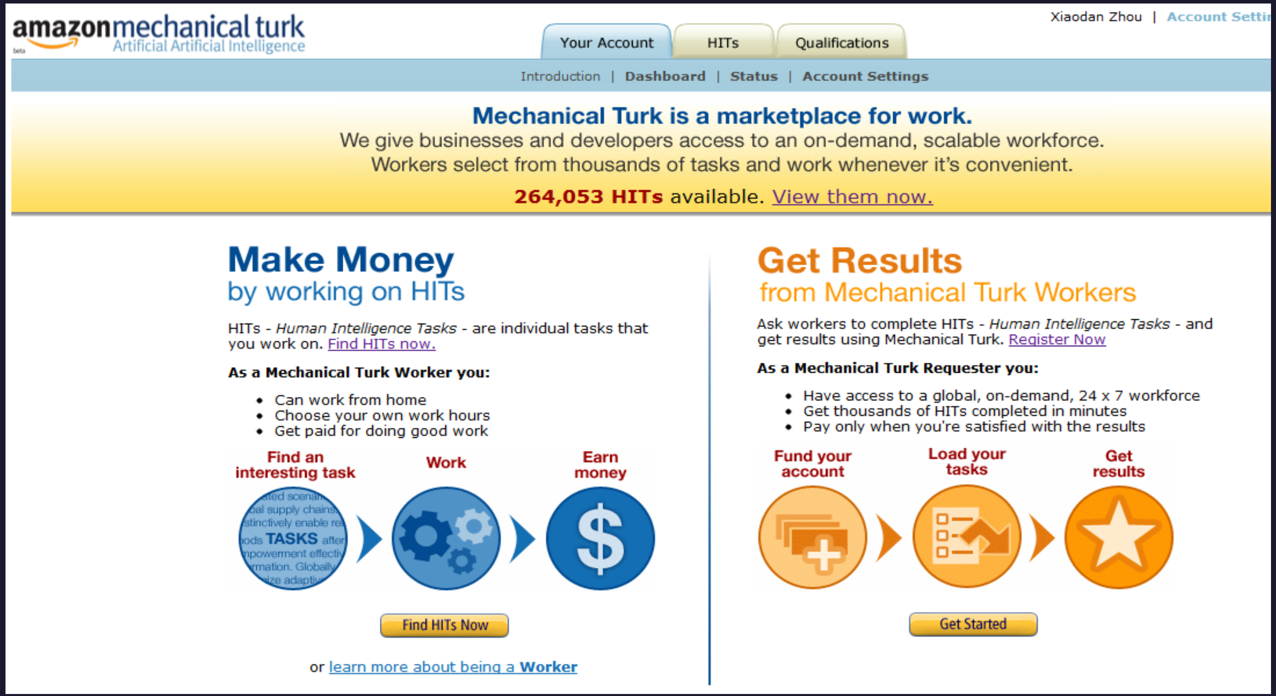
- Classification of Twitter users into adults and teenagers
- Evaluation of the importance of age group information and usefulness of predicted age group in tweet sentiment analysis
- Datasets
  1. 6,387 tweets with estimated sentiment score provided by 76 assessors for sentiment analysis
  2. 6,280 tweets over 7 different topics, collected through Twitter API, for age group prediction
     13 features: 7 tweet features, 5 tweet author profile features, 1 age group

- Sentiment analysis
  - Sentiment scoring between -5 and 5
  - Sentimeter-Br2 score, does not take user profile information into account
  - eSM, takes user profile information into account, proven superior
- User profiling - Age group prediction
  - Evaluation of 5 different models; MLP, CNN, Decision Tree, Random Forest, SVM
  - CNN was selected
    Proven useful in improving sentiment score prediction using the predicted age group information when no age group is provided

Using First Names as Features for Gender Inference in Twitter

Dataset

- Incorporate a user's name into a gender classifier
- Collected the most recent tweets and combined with profile information to create the user's textual content.
- Validated the quality of their dataset with the real Twitter statistics



User Profiling

- Three methods of SVM classifiers:
  1. Baseline: Only user-feature vector (K-top words, Frequency statistics, etc.)
  2. Integrated: Added the gender-association score.
  3. Threshold: Threshold using the gender-association score.

The highest average accuracy derived from the threshold algorithm, because most names are either strongly associated with a given gender or unknown.

Twitter-based user occupational classification

Dataset

- Identify the most likely job class for a given user based on their Twitter profile and a variety of textual features.
- 4-digit UK Standard Occupational Classification system.
- Divided the dataset into two types features:
  1. The UserLevel. General user information or aggregated statistics about the tweets.
  2. The Textual. Represent each user as a distribution over features.

- 4 extraction methods for the textual features using the Normalized Pointwise Mutual Information matrix:
  1. SVD-Embeddings. Each user's function representation is calculated by adding all the embedding dimensions in all words.
  2. SVD-Classifier. The NPMI matrix creates clusters of words which represent the "topics" using spectral clustering.
  3. W2V-Embeddings. Use skip-gram model, with negative sampling.
  4. W2C-Classifier. Clusters of related words.

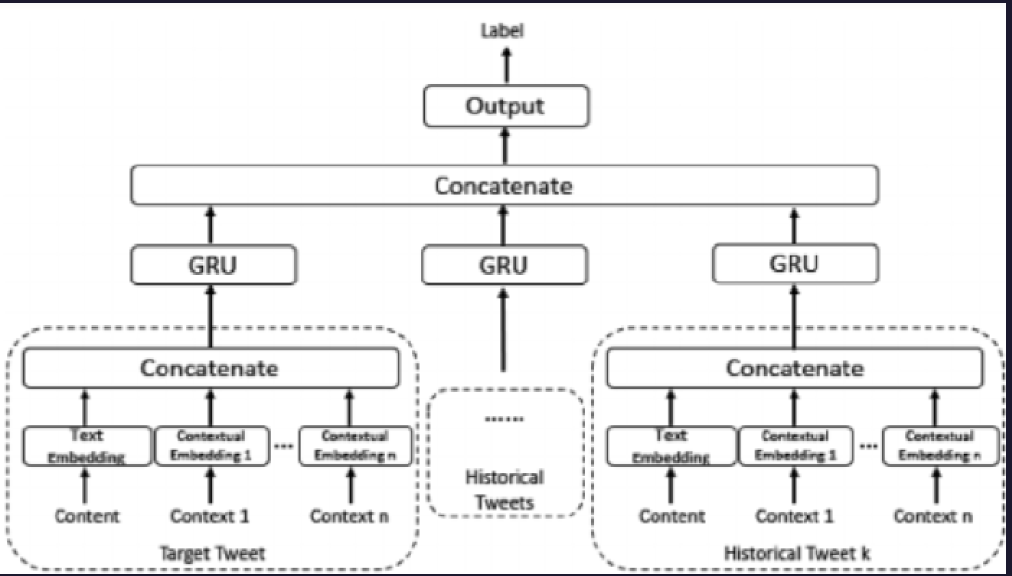
User Profiling

- Implemented the Gaussian Processes (GPs) using Expectation Propagation which offers very good empirical results for many different likelihoods.
- Perform a separate one-vs-all classification for each class and then determine the label

Activity recognition using hybrid GRNN and Tweets

Dataset

- Detect the “offine” activity.
- Location of the tweets to assign Location–activity labels using Google Map API (e.g., user is traveling, dining etc).
- Tweet filtering based on point of interest (POI).



User Profiling

- GloVe pre-trained word embeddings.
- HD-LSTM takes text input, along with contextual features of historical information , POS tag sequence, and post time, in the form of embeddings.
- Produces a flat vector representation
- Label is produced from the output layer which has as input the above concatenated flat vector.
- Created a second–similar algorithm GRU

Purpose	Dataset source	Profile filtering	Profile type	Modeling type	Profiling technique
News Recommendation	Twitter	Content-Based	Static	Interest	Vector-Space Model
Tweets Recommendation	Twitter	Content-Based	Static	Interest	Vector-Space Model
News Recommendation	DMOZ	Rule-Based	Static	Interest	Expert rules
Gender prediction	Onet	Content-Based	Static	Interest	Vector-Space Model
Age group classification / Sentiment analysis	Twitter	Content-Based	Static	Behavioral	Machine Learning
Gender prediction	Twitter	Content-Based	Static	Interest	Machine Learning
Job classification	Twitter	Content-Based	Static	Interest	Machine Learning
Activity Recognition	Twitter	Content-Based	Dynamic	Behavioral	Deep Learning

Conclusion

Overview

- Data collection and preprocessing
- User profiling techniques

Observations

- Most works use implicit data collection methods
- Not many works perform dynamic user profiling
- Most works model user interest
- Not many publicly available datasets, most create custom-made datasets using the Twitter API
- Vector-Space modelling and Machine Learning techniques are the most common
- User profiling techniques are commonly used in recommendation systems

Increasing interest due to the need for

- More personalized recommendations
- Demographic analysis

Technical Implementation Plan

- Study the effect of gender on one's development in academia.
- Collect information for several academics, both male and female, from different universities across Europe.
- Build a profile, which will reflect
  - their whole progress in academia
  - their interests
  - their involvement in projects and conferences
- Study the impact of other factors as well, either demographic, or social

Questions?