360° USER PROFILE LEARNING FROM MULTIPLE SOCIAL NETWORKS FOR WELLNESS AND URBAN MOBILITY APPLICATIONS

ALEKSANDR FARSEEV
Supervised by: Professor Tat-Seng Chua
Thesis Examiners: Associate Professor Stephane Bressan, Associate Professor Chan Mun Choon, Professor Paolo Rosso (Universitat Politècnica de València)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF SINGAPORE

2017

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Aleksandr Farseev
August 2017

Acknowledgments

This thesis would not have been possible without the support of many people.

I would first like to thank my Ph.D. advisor, Prof. Tat-Seng Chua, whose motivation, extensive knowledge, and honest opinion helped me to shape my research and, most importantly, me to be the person as I am now. During these years he patiently monitored and directed my work allowing for the expression and development of the good qualities I may possess, but most importantly, he mentored me to become a researcher and taught me to navigate in the landscape of academia in a manner that I now love and appreciate. Thank you!

Subsequently, I would like to express my gratitude to my colleagues in the Lab for Media Search (LMS) group for their friendships, for the times we had together, for the weekly meetings we had to discuss our research problems, and for all the fun we had during my Ph.D. study. Specifically, I thank Dr. Liqiang Nie who has walked with me the first steps of my academic existence and showed me the tricks. Mr. Ivan Samborskii who helped me to push the bBridge platform to production. Mr. Zaw Lin Kyaw who introduced me to the dark world of Deep Learning and Mr. Francesco Gelli for our chats about everything. I would especially like to thank Dr. Mohammad Akbari for all the fruitful discussions we had during our Ph.D. studies.

I was also lucky to have received great support from people outside the lab in NUS. Specifically, I would like to express my gratitude to Prof. Andrey Filchenkov (ITMO University), whose research group extensively collaborated with me during my Ph.D. study.

I would also like to express my deepest appreciation to my Thesis Advisory Committee: Prof. Stephane Bressan, and Prof. Chan Mun Choon, for their support and guidance during my Ph.D. journey.

Last, but not least, thanks to a beautiful flower I discovered when I was wondering around the urban jungle of Singapore sometime in the beginning of my Ph.D. — my wife Yu-Yi Chu. She supported me through my research journey and did not let me down regardless all the difficulties that I've faced.

List of Publications

Farseev, A., Nie, L., Akbari, M., & Chua, T. S. (2015 June). **Harvesting multiple sources for user profile learning: a big data study.** In Proceedings of the 5th ACM Conference on Multimedia Retrieval (pp. 235-242). ACM.

Farseev, A., Kotkov, D., Semenov, A., Veijalainen, J., & Chua, T. S. (2015 June). **Cross-social network collaborative recommendation.** In Proceedings of the 7th ACM Conference on Web Science (p. 38-39). ACM.

Farseev, A., Akbari, M., Samborskii, I., & Chua, T. S. (2016). **360° user profiling: past, future, and applications.** by Aleksandr Farseev, Mohammad Akbari, Ivan Samborskii and Tat-Seng Chua with Martin Vesely as coordinator. ACM SIGWEB Newsletter, (Summer), 4.

Farseev, A., Samborskii, I., & Chua, T. S. (2016 October). **bBridge: A Big Data Platform for Social Multimedia Analytics.** In Proceedings of the 25th ACM Conference on Multimedia (pp. 759-761). ACM.

Farseev, A., & Chua, T. S. (2017 February). **TweetFit: Fusing Multiple Social Media and Sensor Data for Wellness Profile Learning.** In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 87-93). AAAI.

Farseev, A., & Chua, T. S. (2017). **Tweet can be Fit: Integrating Data from Wearable Sensors and Multiple Social Networks for Wellness Profile Learning.** ACM Transactions on Information Systems (TOIS).

Farseev, A., Samborskii, I., Filchenkov, A., & Chua, T. S. (2017 August). **Cross-Domain Recommendation via Clustering on Multi-Layer Graphs.** In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Chowdhury, A.-K., Farseev, A., Chakraborty, P.-R., & Tjondronegoro, D. (2017 December). **Automatic Classification of Physical Exercises from Wearable Sensors using Small Dataset from Non-Laboratory Settings** In Proceedings of the First IEEE Life Science Conference. IEEE.

Buraya, K., Farseev, A., Filchenkov, A., & Chua, T. S. (2017 February). **Towards User Personality Profiling from Multiple Social Networks.** In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI.

Nie, L., Zhang, L., Wang, M., Hong, R., Farseev, A., & Chua, T. S. (2017). **Learning user attributes via mobile social multimedia analytics.** ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 36.

Contents

Abstract

The drastic change in the Web was witnessed throughout the past decade, which saw an exponential growth in social networking services. The reason of such growth is that social media users concurrently produce and consume data. In this context, millions of users, who follow different lifestyles and belong to different demographic groups, regularly contribute multi-modal data on various online social networks, such as Twitter, Facebook, Foursquare, Instagram, and Endomondo. Traditionally, social media users are encouraged to complete their profiles by explicitly providing their personal attributes such as age, gender, interest, etc. (individual user profile). Additionally, users are likely to join interest-based groups that are devoted to various topics (group user profile). Such information is essential for different applications, but unfortunately, it is often not publicly available. This gives rise of automatic user profiling, which aims at automatic inferencing of users' unobservable information based on observable information such as an individual's behavior or utterances.

This thesis focused on investigating user profiling across multiple social networks in Wellness and Urban Mobility domains. Considering that user profiling can be performed at individual and group levels, this thesis proposes two multi-source learning schemes: multi-source learning scheme for individual user profiling and multi-source community detection scheme for group user profiling. We practically applied the proposed approaches in three user profiling scenarios: demographic profiling, physical wellness profiling, and venue category recommendation.

To build a comprehensive individual user profile, we first trained an efficient ensemble to predict users' age and gender based on data from three social networks. We then proposed the Multi-Task Multi-Source Wellness Profiling ("M^2WP") framework to predict users' Body Mass Index (BMI) category and "BMI Trend" (direction of BMI change over time) based on data from multiple social networks and wearable sensors. We further presented a study on multi-source group profiling and its application in Urban Mobility Domain: Cross-Source Community-Based Collaborative Recommendation (C^3R). The proposed recommendation framework utilizes both individual and group knowledge (user communities) to solve the task of venue category recommendation. The user communities are detected based on our-proposed regularized spectral clustering approach that can perform an efficient partitioning of multi-layer user relations

graph. To show the applicability of our individual and group user profiling solutions to real-world scenarios, we integrated all the earlier proposed techniques in cloud-based Big Data Platform for Social Multimedia Analytics named bBridge. Finally, we performed a comprehensive qualitative analysis of the relationship between multiple data sources in Wellness and Urban Mobility domains.

The experimental results enable us to draw the following three key findings. First, utilization of multiple data sources improves the performance of individual and group user profiling as well as their applications. Second, it is important to take inter-category relatedness into account when dealing with multiple social networks and sensor data simultaneously. Third, in the context of group user profiling, consideration of inter-network relationship at the clustering stage is essential.

List of Figures

List of Tables

Chapter 1

Introduction

Drastic change in the Web was witnessed throughout the past decade, which saw an exponential growth in social networking services. The reason of such growth is that social media users concurrently produce and consume data. In this context, millions of users, who follow different lifestyle and belong to different demographic groups, regularly contribute multi-modal data on various online social networks, such as Twitter, Facebook, Foursquare, Instagram, and Endomondo. An example is the US region, where every third adult older than 65 years use at least one social media service, while most of the younger ones are active in multiple online social forums simultaneously [66].

There are several important properties that differentiate social media services from other information resources available on the Web. First, social media users play both “data consumer” and “data producer” roles, which result in vast amounts of user-generated content (UGC) produced in real time. Second, data from various social networks is often given in different content forms (modalities), such as text, images, video, location, sensor sequences, social connections, etc. These multi-source multi-modal data describe social media users from different perspectives, which helps to gain a more comprehensive insight into the nature of activities and interactions between users. Third, the differences in employed data modality are dictated by the distinct functionality and properties of each social networking service. For example, Linkedin conveys users’ formal career path; while Twitter microblogging service casually uncovers users’ daily activities. Lastly, it is worth mentioning that after joining a social network, users are immediately provided with a three-dimensional online identity, which consists of their profile (name, gender, location, etc.), content (multi-modal user-generated data), and a network, which connects users to each other. All the above suggest social media as a valuable and powerful information source for research and industry.

Traditionally, social media users are encouraged to complete their profiles by explicitly providing their personal attributes such as age, gender, interest, etc. (individual user profile). Additionally, users are likely to join interest-based groups that are devoted to various topics (group user profile). Such information is necessary for different applications. For example, a comprehensive **individual user profile is essential in marketing, advertisement, and wellness domains**. Specifically, automatic Body Mass Index (BMI) inference is crucial for population obesity control; weight problems, in turn, serve as a reason for more than 12 major chronic diseases [101]. In fact, such individual attributes as Age and Gender are essential in fine-tuning of wellness control and marketing strategies with respect to the particular demographic group.

Group profiling is useful for brand monitoring, facility planning, and urban mobility applications. An example of Urban Mobility application is venue category recommendation, which suggests categories of places to visit in every particular location in a collaborative manner. However, in practice, the majority of users are reluctant to provide actual personal attributes. At the same time, the group participation is often relevant to their friendship connections, rather than interests [49], which raise the problem of filtering out irrelevant user communities.

As a solution of all the above problems, automatic user profiling based on UGC is explored to gain relevant knowledge about an individual user and user groups. The scale of the user-generated data and the data multi-modality entails the necessity of automatic, scalable, multi-source user profiling techniques to be developed. These remain to be an open research challenge.

Historically, the problem of individual user profiling in wellness domain is approached based on data from various sources of information. Usually, an individual wellness user profile involves personal user attributes [49] such as BMI category, demographics [54], or chronic disease tendency [4]. Prediction of such demographic characteristics as Age, Gender, and Personality from text corpora was studied by participants of a popular evaluation forum PAN [132, 134, 136, 155]. In an attempt to infer application-specific demographic attributes, Song et al. [154, 153] predicted interests of users from multiple social networks and the tendency of users to participate in volunteer activities, respectively. Later, Buraya et al. [28] attempted to infer relationship status of users from Instagram, Foursquare, and Twitter, while Liu et al. [99] solved the problem of user occupation prediction by leveraging on temporal aspect of multi-source data. To gain more insights into users' physical condition, Chou et al. [35] conducted a qualitative study on weight-related interactions of Twitter and Facebook users, while Mejova et al. [105] and Silva et al. [152] attested to the predictive power of Foursquare-based features for the group obesity inference task and cultural differences analysis. Later, Weber et al. [174] guided machine learning models based on data from MyFitnessPal[1] to predict a success of users' personal diet plans, while Park et al. [122] studied publishing traits of social media users who openly share their individual health and fitness related information. Yuan et al. [182] studied the approaches to utilize clinical measurements to enhance the results of Alzheimer's Disease prediction, while Akbari et al. [4] proposed to learn the individual wellness profiles of microblog users concerning a taxonomy of wellness events.

[1] myfitnesspal.com

The task of group user profiling in Urban Mobility domain was tackled in a cross-disciplinary manner. Gonzalez et al. [70] studied human group mobility patterns by analyzing the one-dollar note movement across the US. Similarly, Zhao et al. [185] performed multi-modal user community detection and profiling to solve the venue recommendation problem from multiple social networks. Rhouma and Romdhane [139] proposed an approach for multi-source user community detection in partially-overlapping social network graphs, while Dong et al. [45] detected multi-source user communities on a multi-layer graph. Finally, Noulas et al. [115, 116, 117] performed a series of studies towards venue recommendation in urban environments based on data from location based social networks and cellular telecommunication companies.

Notwithstanding the fact that there were some research efforts done on group and individual user profile learning, most of these works utilize either single data source or modality, which is not practical in real world settings [49]. The above introduces a **large research gap, which is identified by low availability of automatic techniques for individual and group user profiling from multiple data sources in Wellness and Urban Mobility domains.**

1.1 Research Problems

In order to bridge the research gap, it is necessary to address the following three research challenges:

1. **Multi-Source Data Gathering and Inter-Source Alignment For Large Scale Research in Wellness and Mobility Domains:** To perform user profile learning from multiple social networks, it is necessary to align different social network accounts of the same user to each other and harvest multi-source user-generated content from his/her social media accounts. Moreover, to conduct research in wellness domain, it is essential to collect data that reflect users' actual physical condition: data from wearable sensors (i.e. Heart Rate, Oxygen Level, etc.), smart devices (i.e. Speed, Geographical Coordinates, Altitude), and external sensor devices (i.e. Wind Speed, Temperature, Humidity, etc.).

2. **Mutually-Consistent Multi-Modal Data Representation:** Besides textual data, social media services involve data of various modalities. For example, in Instagram[2], users share recently taken pictures and videos, while in Endomondo[3] users post information about their workouts, which is strongly dependent on the temporal and spatial aspects. Integration of such heterogeneous multimodal data sources requires

[2] `instagram.com`

[3] `endomondo.com`

the development of efficient and mutually consistent data representation approaches.

3. **Modeling Data Source Fusion in Wellness and Urban Mobility Domains:** To perform learning from multi-source data in Wellness and Urban Mobility application domains, it is important to develop efficient data integration techniques that will fuse multi-modal multi-source social media data and data from wearable sensors in balanced and mutually-consistent fashion.

1.1.1 Solutions Overview

To address these problems, we present a framework for automatic user profiling from multiple social networks. The framework involves both individual and group user profile estimation based on data from multiple social networks and wearable sensors. We adopt the framework for wellness and Urban Mobility domains. The framework involves three components. First, we harvest users' multi-modal social media contents from multiple social networks and wearable sensors by utilizing multi-source crawling scheme, in which Twitter plays the role of social media content "sink". Second, data representation model extracts multi-faceted features to describe each user from 360° view. Essentially, we extract psycholinguistic representations, latent topics, mobility representations, venue semantics, image concepts, sensor data frequency spectrum, and various statistics. Finally, we use the multi-source learning model to conduct individual profiling and group profiling on the multi-source data. Specifically, we model the fusion of different data sources and data from wearable sensors in a co-regularized fashion to infer individual and group user profiles.

Data Gathering and Inter-Source Alignment

To collect users' distributed social content from multiple social networks, it is necessary to identify a set of users active in multiple social networks. We consider a user to be active if he/she has recently posted tweets through the cross-linking functionality of Instagram, Endomondo, or Swarm (Foursquare) mobile apps. We utilize Twitter REST API[4] to perform the location-dependent tweets search in Singapore, London, and New York regions to crawl culture-specific data. To crawl sensor data and for Endomondo accounts alignment, we utilize Twitter REST API geo-independent search. We further start a Twitter "stream"[5] that involves all active users and downloads the multi-source user-generated content of these users via URLs to the original posts. Such a data harvesting approach offers the possibility to map Twitter user IDs to Foursquare, Endomondo, and Instagram user IDs, which allows us to overcome the multi-source user identification problem [169]. By following this, we perform

[4] dev.twitter.com/rest/public

[5] dev.twitter.com/streaming/overview

data crawling from Foursquare, Endomondo, and Instagram social networks for previously identified active Twitter users. Comparing it to the small-scale approaches (i.e. social network aggregation sites[6]) and inaccurate approaches (i.e. machine learning-based cross-network identification [169]), our-proposed crawling approach allows for balancing between a scale of the research and the requirement to have accurate ground-truth. Noticeably, this crawling strategy can also be applied to monitor other social networks as well.

Data Representation

Efficient data representation is a significant step in multi-source profile learning pipeline. Its importance is dictated by its ability to bridge the semantic gap between data of different modalities. Consistency in data representation can be achieved by describing every data source in terms the so-called user-feature matrices, where each row is a social media user, and each column — a feature (i.e. distinctive characteristic of the user). To obtain user-feature matrixes from textual data, we utilize state-of-the-art natural language processing approaches, such as Linguistic Inquiry and Word Count dictionary (LIWC [127]), latent topic learning (LDA [21]), and some heuristic behavioral traits. To represent visual data, we use GoogleNet image concept detection model [159], while to extract location and mobility features we utilized Foursquare venue category names and GPS tracks. Finally, to represent Sensor data we employed Furrier Transform (FT) [23] as well as provided workout statistics. For most of the experiments, the data representations were normalized.

Data Source Fusion

After data harvesting and representation, the automatic user profiling is performed. Particularly, we conduct both individual and group user profiling. We treat individual user profiling in wellness domain as supervised learning task. To infer users' demographic attributes, we use multi-source ensemble learning [54]; while to predict user BMI, we utilize Multi-Source Multi-Task Learning [51]. The group user profiling in Urban Mobility domain is treated as unsupervised learning problem (community detection). It is approached by performing regularized multi-layer spectral clustering, which serves a backbone of our-proposed venue category collaborative recommendation approach [53].

1.2 Research Contributions

The research contributions of this thesis are as follows::

[6] `about.me`

- We introduce a **new cross-network data gathering and alignment scheme**, where one social network (i.e. Twitter) plays the role of a "sink" for data from other social networks. The approach allows us to crawl multiple multi-modal data sources simultaneously, while automatically align accounts of the same users in different social networks. To facilitate further research in individual and group user profiling, we've released two large-scale multi-source datasets: NUS-MSS [54] and NUS-SENSE [51]. The NUS-MSS dataset is published to promote research on demographic profiling, group profiling, and cross-region Urban Mobility profiling; while NUS-SENSE dataset exceedingly supports research in Wellness and "Quantified Self" domains.
- We propose a **generic multi-source supervised learning framework** and its applications to demographic and wellness profiling domains. To infer users' demographic attributes (i.e. Age and Gender), the framework efficiently fuses multiple heterogeneous information queues in a late-fusion manner via stochastic hill climbing with the random restart. To infer users' physical health attributes (i.e. BMI category and BMI trend), the framework performs multi-task learning via collective incomplete data sources fusion, performs feature selection on-the-fly, and incorporates the new idea of inter-category smoothness regularization.
- We introduce a **generic cross-source user community detection approach**, which incorporates inter-source relationship and serves as a backbone for urban mobility recommender framework. The framework utilizes group knowledge to perform cross-source user community-based collaborative venue category recommendation. The community detection approach incorporates inter-source relationships during the process of learning individual source representations and preserves inter-source consistency at the stage of learning latent sources' representation. The inter-source relationship graph is computed automatically from the data and further utilized via novel graph-constrained regularization.
- To study the relationship between different data sources and user profile attributes, we perform the first **large-scale correlation analysis of multi-source multi-modal social media data** in Wellness and Urban Mobility domains.

1.3 Thesis Organization

The rest of this thesis is organized as follows:

Chapter 2 provides the literature survey on the user profiling in Wellness and Urban Mobility domains.

Chapter 3 outlines the data gathering and cross-source alignment approach [49] as well as provides the user profiling framework overview. It then glances through NUS-MSS [54] and

NUS-SENSE [51] datasets.

Chapter 4 describes our proposed approaches for individual user profiling in wellness domain. First, we formulate and evaluate the multi-source ensemble learning model for demographic attributes inference — “SHCR” [54]. We then define a problem of multi-source multi-task learning (MTL) based on data from multiple social networks and wearable sensors for BMI Category and “BMI Trend” inference tasks. Specifically, we represent different data source combinations as MTL “tasks” and introduce sparsity into data representations via $\ell_{2,1}$ regularization [51]. To constrain the model in learning “similar” categories in a mutually-consistent fashion, we then formulate an inter-category relationship regularization [50]. Finally, we evaluate the approach against different data source combinations and state-of-the-art baselines.

Chapter 5 proposes our multi-source group profiling approach. First, we define the task of cross-source recommendation in the urban environment. We then formulate the problem of cross-source community detection as spectral clustering on a multi-layer graph and overview the graph construction procedure. Further, we discuss the subspace regularization and inter-source regularization mechanisms [56]. Both approaches are targeted to learn a balanced latent clustering space. Furthermore, we practically evaluate the proposed scheme, where group profiling serves as a backbone for Cross-Source Venue Category Recommendation Framework [53]. The chapter is concluded by the description and architecture of the real-world group user profiling application: bBridge Multimedia Analytics Platform [55].

Chapter 6 describes a large-scale statistical analysis of the relationship between data representations within one data modality and between data from different data sources with respect to individual and group user profiling [50].

Chapter 7 concludes the thesis, highlights the limitations, and points the further research directions.

Chapter 2

Literature Review

In this chapter, we review related works on User Profiling in Wellness and Urban Mobility Domains.

2.1 Individual User Profiling and Wellness

As defined by Zukerman et al. [187], individual user profiling is the process of inferring unobservable information about a person based on observable information such as his/her behavior or utterances. Usually, this problem is treated as a supervised learning task, where one's personal attributes are inferred based his/her generated data.

2.1.1 Individual User Profiling from Conventional Data Sources

Till now, there have been multiple research efforts dedicated to individual user profile learning from conventional data sources. For example, in his early work, Fischer et. al [58] investigated the dependencies between personal attributes and morphological features, while Booklet et al. [22] attempted to identify children' age based on their voice records. Later, Argamon et al. [10] predicted Gender, Age, Native Language, Neuroticism Level of anonymous authors from language study text corpora, while Yarowsky et al. [178] utilized gender-dependent sociolinguistic differences to predict individuals' biographic attributes. Finally, Estival et al. [48] predicted five demographic and five psychometric traits by utilizing character-level, lexical, and structural features in various generic machine learning tools. Even though the above works were able to show promising individual user profiling results, they are often hindered by limited data availability, which could be explained by the limitations of the data collection procedures utilized.

2.1.2 Individual User Profiling Based on Social Media Data

With the growth of publicly-accessible online data, a large-scale individual user profiling became possible via utilizing the data from various social multimedia data sources. One of the most common individual user profiling applications in Wellness domain is demographic profiling.

Demographic Profiling

Demography as a study was known in ancient Greek, Indian, Rome and Chinese cultures [151]. In the early 20th century, it has been emerged from statistics and drastically spread among many Social Science research communities [43]. The first attempt to infer personal user attributes was made by Rao et. al [137], who utilized support vector machines (SVM) classifier ensemble to infer users' gender, age, regional origin and political views. The ensemble was trained based on a combination of sociolinguistic and n-gram features. Later, Otterbacher et al. [120] employed the statistical regression model to show that one's gender can be predicted based on stylistic, content, and metadata features in a movie reviews domain. Hu et al. [78] used browsing patterns (web page clicks) from many web portals to predict users' age and gender via Bayesian modeling. At the same time, Bi et. al [20] revealed that users' historical search queries are useful to infer users' age, gender, and political views, while Koppel et al. [89] achieved the prediction accuracy of 80% for the problem of gender detection by using simple lexical and syntactic features. Schler et al. [146] approached to solve the age detection problem. The proposed model was trained based on 71 thousand blog posts and performed with the accuracy of 80% and 75% for gender and age group (3 groups) prediction, respectively. An early attempt to microblog user demography profiling was made by Nguyen et al. [110]. The authors studied Dutch Twitter data, where the average length of a message was less than ten terms, while age was described as a continuous variable in logistic regression model.

The growing interest in individual user profiling gave rise to the series of the so-called author profiling and author identification evaluations, such as PAN [132, 133, 135, 134, 135, 136]. In PAN 2013 [132], competitors were asked to predict gender and age attributes of anonymous authors, based on English and Spanish text corpora, while in PAN 2014 [135] the same personal author characteristics were predicted based on the compiled corpus. The corpus includes the data from social media, blogs, microblogs, and hotel reviews. Later on, in PAN 2015 [134] organizers included two new languages, namely Italian and Dutch, along with a new subtask on personality recognition. In PAN 2016 [136], the effect of the cross-genre evaluation was investigated: models were trained on Twitter data and evaluated on other data sources. Finally, in PAN 2017 [133], organizers included Arabic and Portuguese languages into the corpora as well as proposed the task of language variety identification together with the gender dimension. In general, data representation techniques for demographic profiling can be divided into two types: content-based features (bag of words, words n-grams, term vectors, named entities, dictionary words, slang words, contractions, sentiment words) and style-based features (frequencies, punctuations, POS, HTML tags, readability measures, first order statistics). The best-reported results of microblog user age and gender prediction task

(English language) were achieved by PAN 2015 winners and were reported as 85% and 83% for gender and age group (4 groups) prediction, respectively [134].

Due to the immaturity of the observed research field, most of the reviewed works were focused either on single-modal or single-source demography profiling. Moreover, the proposed models are mostly built based on data from the text domain. These approaches may not be able to utilize multi-modal user-generated content efficiently, which suggests to address the problem from the multi-source perspective.

Individual User Profiling in Other Application Domains

Except demographic profiling, user profiling was broadly adopted in other related application domains. For example, it has been long known that a persona's writing, talking or social media communication style could tell a lot about one's character and emotional level. In the book "The Secret Life of Pronouns" [126], James W. Pennebaker observed that different pronouns in one's speech/writing could represent individual Sex and Age, describe user's personality and co-occur with certain emotions. Moreover, some pronounce were noticed in human speech more often when individual are not telling the truth "word I is used at far higher rates by followers than by leaders, truth-tellers than liars" [126]. It is not surprising that various data mining techniques were broadly adopted in computational social media and computational linguistic research for various applications in individual user profiling domain. One such application was proposed by Schwartz et al. [147], where authors described how Facebook status text can be used to predict users' personality by performing differential language analysis. Goel et al. [68] performed a large scale prediction devoted to age, gender, race, education, and income level based on $250,000$ users browsing history collected in one year period. The prediction framework was built on the support vector machines ensemble. Later on, Buraya et al. [28] attempted to infer relationship status of users from Instagram, Foursquare, and Twitter. The similar idea was integrated into Seneviratne et al. [148], where authors collected data from volunteers as a smartphone user and analyzed installed applications statistics data. Later, Alam et al. [6] incorporated acoustic, linguistic, and psycholinguistic features across diverse speech corpora to tackle the problem of learning the personality traits from spoken data. Meanwhile, Pennacchiotti et al. [125] described a generic framework for three-class user profiling: political affiliation, ethnicity, and the "favor". Except for the user-centric features, authors utilized social graph information, which boosted the previously-achieved classification performance. Similarly to demographic profiling, most of the approaches above are based on single-source or single-modal data, which restricts their usage in a real-world scenario. Specifically, modern social networks describe their users in a multi-view fashion, which needs to be considered at the learning stage. We, thus, review related works on multi-source individual user profiling immediately next.

2.1.3 Individual User Profiling from Multiple Sources

Multi-modal learning (also known as multi-view learning) is a paradigm in modern machine learning research, which aims to overcome the disadvantages of the two most common data integration strategies: early fusion and late fusion. In particular, it introduces a cost function to model each data modality (source) jointly with other modalities (sources), which allows for considering both inter-modality and intra-modality relationships. Below we describe existing research towards this direction.

Christoudias et al. [37] first performed multi-modal learning while taking disagreement between data views into consideration. Later on, Yuan et al. [182] presented an incomplete multi-source feature learning method where blockwise-incomplete data was split into disjoint groups (different data source combinations). Following this trend, Song et al. [153] proposed a multi-view learning framework for Volunteerism tendency prediction for users from different social networks. The missing data was induced by the proposed constrained Non-Negative Matrix Factorization (NMF) approach [121]. It worth noting that the usage of data completion techniques could introduce bias into the final data representations and, thus, may not apply to sparse real-world datasets. Meanwhile, Yuan et al. [182] proposed to treat the problem of multi-source fusion as a multi-task learning for enhancing Alzheimer's Disease prediction. Aiming to model the relationship between different inference categories, He and Lawrence [74] proposed to regularize multi-view multi-task learning by a graph structure. The research was conducted in the context of text classification. Given task pairs, the model projects them to a new latent space based on the sources' similarities. However, such an approach does not allow for generating predictive models based on data samples that do not share any common source. The latter is the major restriction in the real-world settings, which does not allow for its extensive utilization. Inspired by the limitations above, Song et al. [154] proposed a structure-constrained multi-task learning framework for user interests inference from multi-source data. To model interest relationship, authors included external knowledge in the form of "interest relationship weights", which makes the framework biased towards particular datasets and tasks. Finally, Liu et al. [99] solved the problem of user occupation prediction by leveraging on the temporal aspect of multi-source data. The above works demonstrated the great potential of individual user profiling from multi-source social media data. However, as highlighted above, most of the works are limited by utilized data completion techniques, the necessity to have an external domain knowledge or single data modality. Moreover, they do not consider specific properties of Wellness domain, which differs from conventional user profiling applications. All the above highlights the need for new multi-source individual user profiling techniques in Wellness application domain. We review several related works in next section.

2.1.4 Individual User Profiling and Physical Wellness

Physical Activity Recognition

Due to the rise of microelectronics, the huge number of sensors and mobile devices with unique characteristics came into the market during the past decade. The small size, low cost and high computational power of mobile devices have redefined the way how people interact with the devices. The trend of self-measuring was named "Quantified Self" [158]. It has widespread rapidly among industry, ordinary people and science communities.

One of the popular directions in sensor data modeling research for Wellness is devoted to human activity recognition [60, 17, 160, 181, 88, 86, 93, 92, 36]. The activities were mainly detected by incorporating wearable (e.g. accelerometer and GPS), environmental (e.g. temperature, pressure and humidity) or physiological (e.g. heart rate, skin temperature) sensors [92]. For example, Tapia et al. [160] proposed a model that can recognize physical activities (30 physical gymnasium activities) by modeling data obtained from five triaxial wireless accelerometers and wireless heart rate chest monitor. To solve the same problem, Lara et al. [93] leveraged chest sensor strap and a cellular phone. Finally, Berchtold et al. [19] approached the same task by incorporating cellular phone data only. It worth mentioning that the most of the recent activity recognition systems achieved similar accuracy level (around 92%) [92]. However, due to the differences in datasets, it is hard to compare their performance directly. Since our aim is to incorporate both multi-source social media data and sensor data for individual Wellness profiling learning, we can only rely on sensor data with comparably low sampling rate. In other words, we consider works on highly-customized sensor configurations [160, 19] as less related to our research and focus on simple sensor configurations (e.g. accelerometer, GPS, or personal wearable sensors).

The good example of a framework with simple sensor configuration is the one proposed by Maurer et al. [104]. Four types of sensors were embedded in a smart-watch like device: an accelerometer (dynamic activity sensor), a light sensor (external illumination sensor), a thermometer (external temperature sensor) and a microphone (external dynamic environment sensor). The feature set was extracted on the temporal basis. The C4.5 decision tree model was utilized, which achieved an accuracy of 92%. The smart watch concept has become more and more popular among major mobile electronics manufacturers, and hence, the emergence of big personalized user sensor data set can be expected in nearest future. This makes Maurer et al. [104] research timely and meaningful for modeling Physical Wellness. At the same time, the significance of accelerometer data can be witnessed from Berchtold et al. [19], where authors classified ambulation and phone activities. The task was completed based on fuzzy inference model, which leverages phone accelerometer data only. Later on, "Centinela" framework [93] complemented the accelerometer data with vital signs. As a result, it was able to achieve high accuracy of 97.5% in five ambulant activities recognition task. Last but not least, "Cosar" [140] system took advantage of GPS data by performing spatial-dependent activity pruning for those activities that are not likely to happen in a certain location (e.g. watching TV in Central City Park). The approach helped to boost the activity recognition performance significantly.

The one research direction, which is most relevant to us belongs to the so-called “Objective Self” trend [80, 81]. Going beyond “Quantified Self”, Jain et al. [80] raised a new research problem of simultaneous multi-source multi-modal data co-analysis in nearly real-time settings. Later, Jalali et al. [81] designed a closed-loop system that observes signals to notify an individual about appropriate sources that could facilitate the implementation of desired actions [81]. Since the research topic is quite new, in both works authors did not evaluate models that can process sensor and social media data simultaneously. However, these works raised the very important research problem and may serve a good backbone for individual user profiling. At the same time, there was research conducted towards Physical Wellness inference. These works bring research community closer to realizing the ideal of 360° user profile learning [49] in Wellness domain. We review them in next section.

Physical Wellness Attributes Inference

Inspired by the newly-discovered possibilities of social media data, medical and healthcare communities adopted social media and sensor data in different Physical Wellness applications. For example, Eggleston et al. [46] used social media to monitor food-related habits for obese and diabetes patients, while Fried et al. [62] predicted diabetes and overweight rates for 15 US cities. At the same time, Mejova et al. [105] and Silva et al. [152] attested to the predictive power of Foursquare-based features for the group obesity inference task and cultural differences analysis, while Abbar et al. [1] leveraged Twitter data in an attempt to predict obesity and diabetes statistics based on food names in tweets. Later, Weber et al. [174] guided machine learning models based on data from MyFitnessPal[1] to predict a success of users’ personal diet plans, while Park et al. [122] studied publishing traits of social media users who openly share their individual health and fitness related information. Yuan et al. [182] studied the approaches to utilize clinical measurements to enhance the results of Alzheimer’s Disease prediction, Akbari et al. [4] proposed a multi-task learning framework for individual Wellness events categorization. These research efforts were made towards Wellness lifestyle analysis, and the results show the great potential of social media data to assist in Wellness related research. However, the applicability of the above research works is limited, due to naive data analytics approaches utilized and single data modality usage. Moreover, these works a mainly descriptive in nature, which does not allow for their usage in individual Wellness user profile learning application domain. The above again suggests the necessity of developing new comprehensive individual user profile learning approaches in Wellness domain.

[1] `myfitnesspal.com`

2.2 Group User Profiling and Urban Mobility

Group user profiling aims at learning the collective behavior of users groups that are also known as user communities in social media computing. An intuitive approach for performing user profiling at a group level is to discover communities of users, which can be done by applying unsupervised learning (i.e. clustering) on user relationship graph [185].

2.2.1 User Relationship Graph Construction

As noted above, it is important to find representative user communities to leverage on a group knowledge for various applications. Such task is commonly approached by modeling users' relationship as a graph so that dense subgraphs of such graph can be treated as user communities. The graph can be constructed in a several ways: (a) based on social connections between users (i.e. follower/followee relationship) [61]; (b) based on user-generated content, when the edges of the graph are weighted as a distance between data representations of users (i.e. *cosine*, *Euclidian*, or $Heat$ distance); or (c) as a combination of the above two methods. In next sections, we review related works on user community detection, where user relationship graphs are constructed via one of the strategies above.

2.2.2 Community Detection from Single Data Source

Similar to individual user profiling, the task of group user profiling was mainly approached from a single data source/modality perspective. The conventional community detection methods and their variations can be roughly divided into five disjoint groups:

1. **Graph partitioning** is the group of community detection approaches, where G's vertices are divided into p groups of a predefined size, such that the number of edges between groups is minimized. This class of approaches requires prior knowledge about the size of clusters, which is not the case in most of the large-scale scenarios.

2. **Hierarchical Clustering** is the collection of clustering approaches that discover hierarchical structures in G. The G's vertices are grouped together based on a specified similarity measure so that more similar vertices of the graph would fall into the same cluster. There are two major categories of hierarchical clustering approaches. In the case when clusters are joined iteratively, a clustering algorithm is called agglomerative [111], while when the communities are obtained by removing edges between nodes, it is said to be divisive [65].

3. **Partitional Clustering** approaches try to partition G based on a given similarity measure. The overall goal is to separate the data points into k disjoint sets, based on the distance between data points and clusters' centroids (e.g. K-Means clustering

(Alsabti et al., 1997)). The most commonly used objective functions are Minimum K-clustering and K-center. The Minimum K-clustering evaluates clusters regarding cohesion by minimizing the diameter of a cluster. K-center, in turn, minimizes the distance between each data point in a cluster and the cluster centroid by merging or removing data points as well as redefining the centroid.

4. **Spectral clustering** approaches are targeted to solve the MinCut problem [166]. For a given number k of subsets, the MinCut, essentially chooses a partition $C_1, ..., C_k$ such that it minimizes the expression: $cut(C_1, ..., C_k) = \sum_{i=1}^{k} W(C_i, \overline{C}_i)$, where $\overline{C}_i$ stands for a compliment of C_i, and $W(A, B) = \sum_{u \in A, v \in B} \rho(u, v)$, and ρ is a distance function. Such a formulation allows us to find k user communities $C_1, ..., C_k$, that are grouped according to some criteria (defined by distance function ρ). The two most commonly used formulations of the MinCut problem are RatioCut and the so-called NCut [166]. The idea behind RatioCut is based on an assumption that the resulting communities could have similar size, while NCut formulation conditions the sum of edges' weights in each community to be minimized among all communities. Though both RatioCut and NCut are NP-Hard [166], there were many approximate solutions created for the above problems. One of the existing MinCut approximations is the spectral clustering approach, which is defined as a standard trace minimization problem: $\min\limits_{U \in R^{n \times k}} tr(U^{\intercal} L_{sym} U), s.t. U^{\intercal} U = I$. According to the Rayleigh-Ritz theorem [102], the spectral clustering optimization problem can be solved as the first k eigenvectors of the normalized graph Laplacian $L_{sym} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where W is the adjacency matrix, and D is the degree matrix of the graph G, and the final community assignment can be obtained by performing clustering (i.e. K-Means clustering) in the eigenvector space.

5. **Quality optimization** approaches are based on the idea that communities can also be identified based on the measure of a quality of partition. This measure is defined by a quality function, which is established to assign a numerical score to each detected community. In such a way, communities can be further ranked based on their assigned score. The most popular quality function was by Newman and Girvan [65] and called modularity. In short, the idea is based on the assumption that a random graph does not have a community structure, so the possible existence of communities is revealed by the comparison between the actual density of edges in a community if the vertices of the graph were attached regardless of community structure. The problem of modularity maximization is proved to be $NP - complete$ [24], but was successfully approximated in a reasonable time. Specifically, in spectral graph theory, the modularity can be represented in a matrix form as follows. Let $\mathbf{d}$ denote the degrees of each vertex, where $d_i \in Z+, i = 1, ..., n$, where $Z+$ denotes the set of non-negative integers. Let $\mathbf{B}$ be the modularity matrix with $b_{ij} = a_{ij} - \frac{d_i d_j}{2m}$. The modularity then can be written in a form of a standard trace minimization problem: $\min\limits_{H \in R^{n \times k}} \frac{1}{2m} tr(H^{\intercal} B H), s.t. H^{\intercal} H = I$. According to Newman [109], the above optimization problem can be solved as the first

k eigenvectors of $B = \frac{A}{2m} - \frac{dd^{\intercal}}{4m^2}$, while the final community assignment can be obtained by performing clustering (i.e. K-Means clustering) in the eigenvector space.

2.2.3 Community Detection from Multiple Data Sources

In the case of community detection from multiple data sources, it is common to represent multi-view data in the form of hypergraphs or multi-layer graphs. A common strategy for processing such a graph is to separate it into simpler graphs. Following this strategy, Zlatic et al. projected the hypernetwork to three simple single-source networks for users, tags, and resources, respectively [186]. They then used bottom-up optimization algorithm in each of the single-modality networks. Similarly, Neubauer and Obermayer [108] split the initial multi-view hypernetwork into a user-tag network, a user-resource network, and a tag-resource network to define an extended modularity and then applied the modularity maximization clustering. Lu et al. adopted the similar projection approach followed by employing k-means clustering [100]. However, in the above works, the important information about source relationship is lost, wich does not suggest their usage in real-world settings. Meanwhile, Su et al. [157] demonstrated the usefulness of multi-source community discovery for various applications; Rhouma and Romdhane [139] proposed an approach for multi-source user community detection in partially-overlapping social network graphs; while Zhao et al. [185] identified multi-modal overlapping communities of Foursquare users. Finally, Dong et al. [45] introduced a multi-source clustering approach, where the distance between different data sources was measured on Grassmann Manifolds. The works mentioned above are supportive of utilizing multi-source community detection strategies; however, they are limited because inter-source relationships were not incorporated during the community discovery process. This could potentially lead to sub-optimal results in the real world scenario.

2.2.4 Urban Mobility Applications

There are two main approaches for evaluating community detection algorithms: explicit evaluation and implicit evaluation [185]. Explicit evaluation uses a quality measure (e.g. Modularity [65]) to explicitly compare community detection results achieved by different algorithms. However, there is no widely accepted measure that is able to quantify community detection results in the case of multi-source community detection. Moreover, many of such quality estimation measures were found to be weakly related to the actual condition of the detected communities [61]. Implicit evaluation, in turn, compares results achieved by approaches that belong to other application domains (e.g. Recommendation, Classification, etc.). Such approaches must be based on earlier obtained communities. Implicit evaluation can be used to compare both single-source and multi-source community detection approaches. In this thesis, we perform the implicit evaluation of group user profiling by comparing cross-source recommendation results in the Urban Mobility application domain. In next sections, we review the corresponding related work.

Urban Mobility Analysis

The lack of portable devices that can track human's locations over time made early Urban Mobility research difficult to be carried out at a large scale. The first large-scale research was conducted by monitoring the spread of marked 464 thousand dollar notes that were reported millions of times in different locations via the specially created web page [70]. Authors have calculated the distance between any two subsequent reports for each note followed by distance probability density estimation. As a result, they concluded that the distribution of human displacement (movement) follows the power-law distribution. After the first phone call details record (CDR) dataset had been released for open use, many research works were published in an attempt to process it for various applications. One example is the paper by Akyildizet al. [5]. The authors confirmed the power-law nature of human displacement distribution by considering human movements in long distances.

With the emergence of smartphones in 2000s, the data from location-based social networks was broadly adopted in many research works. Specifically, the authors focused on friend recommendation based on mining location patterns [145], location recommendations [123, 16], or venue category recommendation [53]. One good example of location recommendation research was performed by Ye et al. [179, 180]. First, the authors mined spatial-temporal patterns between individual points of user trajectory followed by pattern semantic meaning assignment. Later, authors extended the earlier proposed approach by considering implicit relations between venues (co-visiting and category similarity) and incorporating such statistics as check-ins distributions on daily (24 hours) and weekly (7 days) basis [180]. At the same time, Bao et al. [16] considered location sequences at different spatial granularities to measure inter-user similarities aiming to tackle the user recommendation (friend recommendation) problem. The authors incorporated venue popularities separately and proposed novel sequence matching algorithm as part of the hierarchical graph-based similarity measurement model. It is important to note that most of the proposed approaches are based on the so-called collaborative filtering (CF) paradigm. Later on, Noulas et al. [117, 118] and Scellato et al. [145] analyzed user mobility patterns based on LBSN data at a large scale. The dataset was collected from Foursquare and Gowalla LBSNs and served as a good support to examine the classical user mobility modeling approaches [156]. Noulas et al. [118] concluded that simple metrics like distance cannot describe human migration rules accurately and they need to be supported with richer data. Such data can be inferred from LBSN. Besides providing accurate user location, LBSNs also offer venues description (name, type, annotation). It gives an opportunity to make precise assumptions about the nature of venues and group them in users' interest groups (restaurants, clubs, universities, etc.). As a result, authors built a real-time supervised learning model that combines various data provided by LBSNs [118].

Later Noulas et al. [117] extended the above model to perform new venue recommendation task by adopting various recommender system algorithms from the web domain. Another complex user mobility modeling approach was proposed by Lian et al. [96]. The authors solved POI recommendation task by applying weighted matrix factorization model on LBSN data. The proposed model incorporates the spatial clustering phenomenon into the matrix factorization process. It improves POI recommendation performance as a result.

At this point, one can notice that the location-based social networks bring not just one more dimension to traditional social networks, but much more precise implicit user mobility signal. It couples precise venue locations in geographical mobility point of view and venue semantic annotations in textual and visual mobility point of view. The richness of data from location-based social networks inspired us to select LBSN as a target domain (recommendation domain) for implicit evaluation of our cross-source group profiling approach. Below, we highlight the related works on cross-source cross-domain recommendation.

Cross-Source and Cross-Domain Recommendation

Probably one of the first studies towards recommendation performance improvement based on multi-source data was conducted by Abel et al. [2]. The authors aggregated user profiles from Flickr, Twitter, and Delicious[2] to demonstrate that their cross-system user modeling strategies have a large impact on the recommendation quality in cold-start settings. At the same time, Tiroshi et al. [162] utilized network-related and domain-related features to perform user interest recommendation. Later on, Yan et al. [177] proposed a two-stage solution of cross-source video recommendation problem: first, user preferences were transferred from an auxiliary network by learning cross-network behavior correlations; next, the transferred preferences were integrated with the observed behaviors on target network in an adaptive fashion. Concurrently, Qian et al. [130] introduced a probabilistic framework that solved the problem of cross-domain recommendation by utilizing shared domain priors and modality priors for collaborative learning of a latent representation. Recently, Farseev et al. [53] performed cross-source venue category recommendation by implementing a recommender system based on their proposed Multi-Source re-Ranking approach, where the ranks of individual sources were obtained by performing nearest neighbor collaborative filtering. These works are related to our study regarding the cross-source approaches utilized. However, they ignore the relationship between data sources. Such assumption is unrealistic in the real-world settings.

[2] `del.icio.us/`

2.3 Summary

In this chapter, we reviewed related works on individual and group user profile learning in Wellness and Urban Mobility domains. First, we showed the increasingly popular research direction of individual user profiling, which is mainly focused on individual user attributes inference, such as Age, Gender, Personality, etc. We then reviewed the corresponding research attempts in Physical Wellness domain. Then, we briefly reviewed conventional community detection methods for standard networks with one dimension of interaction, as well as existing approaches for community detection from multiple sources. We finalized our literature survey by reviewing related works in the Urban Mobility application domain. From the survey, it can be seen that user profiling from a single data source/modality is a relatively popular research direction. However, till now, the research on multi-source user profiling for Wellness and Urban Mobility is sparse and requires further elaboration. Particularly, most of the works either descriptive in nature, use only a single data source, make unrealistic assumptions, or utilize naive data analytics approaches. They may not be useful to gain deeper insights from multi-source social media and wearable sensors data.

Chapter 3

Data Gathering And Framework Overview

In this chapter, we first outline our novel multi-source data gathering strategy and cross-network account disambiguation technique. We then describe the harvested multi-source datasets and glance through our proposed user profiling framework.

3.1 Data Gathering

3.1.1 Multi-Source Data

It is well-known that it is essential to incorporate multimodal data from various sources that represent users from multiple perspectives [54, 113] for building a comprehensive user profile in both Wellness and Urban Mobility domains. At the same time, to build a complete individual Wellness profile, it is necessary to utilize information about users' physical health [41, 4]. Meanwhile, to gain insights into users' Urban Mobility, it is necessary to incorporate their movement tracks while considering types of places they visit [116]. In the following, we describe the commonly-used data modalities and their potential for individual and group user profiling.

First, it was noted that textual information is one of the most valuable contributors towards user profile learning, mainly because of its high availability and its ability to describe users' daily routines comprehensively [49]. For example, users often post in microblogs about their recent activities and health updates [4]. This information is related to users' Wellness and Urban Mobility profiles [4]. Second, it was previously observed that visual data plays a significant role in age and gender prediction since it describes users' visual preferences [49]. It is reasonable to hypothesize that this data is also useful for user profile learning in Wellness and Urban Mobility domains. Third, it was reported that location data is essential for urban Mobility analysis [117], while actual location categories are helpful in wellness attributes inference [105]. This demonstrates the potential of data from location-based social networks for use in Wellness and Urban Mobility research. Finally, data from wearable sensors was discovered to be valuable for solving the activity recognition [92] and health monitoring [15] problems, which shows its potential towards Wellness profile learning.

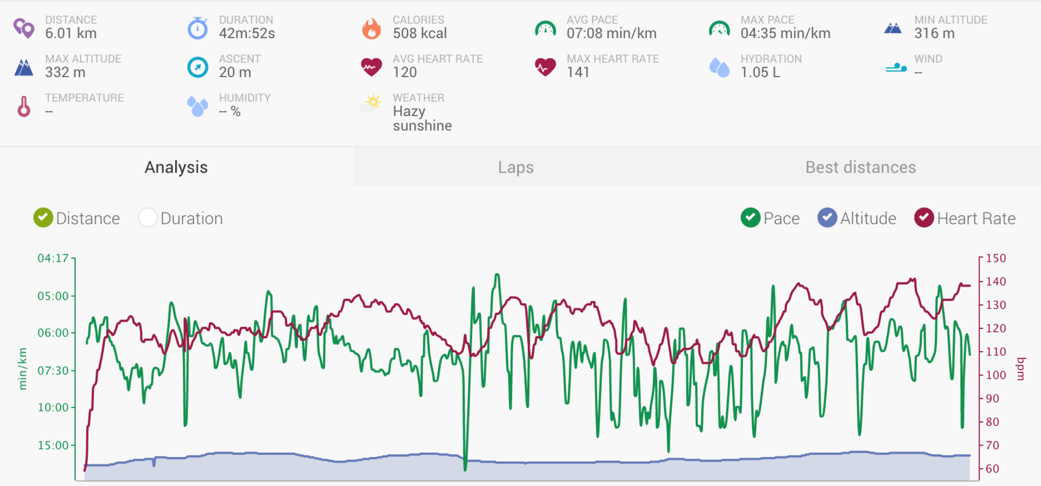
Considering the above, we selected the following social networks for our research:

- **Twitter** (the largest English-speaking microblog) micro-posts to be used as a textual data source.
- **Instagram** (the largest mobile Image-sharing service) pictures and its descriptions (comments) for use as image and textual data sources.
- **Foursquare** (the largest location-based social network) check-in records and its corresponding comments (shouts) to be used as venue semantics, Mobility, and textual data sources.
- **Endomondo** (one of the largest sports activity-tracking social networks) workouts to be used as a sensor data source and ground truth collection (BMI category and "BMI Trend" attributes).

It is worth mentioning that other than providing the so-called, exercise semantics (i.e. sports activity type, as in Fitocracy and MyFitnessPal), Endomondo also serves as a rich source of sequential data from wearable sensors and reliable wellness-related ground truth. The exercise data sequences are often publicly available and usually include a series of multi-dimensional data points (See Figure 3.1), each of which may contain such attributes as Altitude, Longitude, Latitude, Time, and Heart Rate (in cases when a heart rate sensor is set up). Additionally, the wellness-related ground truth labels can be derived from publicly accessible Endomondo user profiles' web pages, which often include such personal attributes as Country of Residence, Postal Code, Age (Birthday), Gender (Sex), Height, and Weight (See Figure 3.2). These attributes are either manually input by Endomondo App users or automatically measured by connected "smart" sensors (e.g. FitBit Aria Smart Scale[1]). The above dictates that Endomondo data goes beyond representing users from one more modality, but bridges the gap between social media-based users' representation and their actual physical activities and condition.

At the same time, it is necessary to mention the crucial importance of Foursquare data for Urban Mobility research. In addition to the information about user geographical movements and the GPS positioning, Foursquare places provide venue-specific metadata, which is also known as venue semantics. Such metadata includes semantic information about the categories of places in Foursquare, which now consists of 764 different location types. This data is essential for Urban Mobility research since it can be utilized as a proxy to model the preferences of mobile users regarding activities they like to perform in the city. Furthermore, the availability of metadata allows for examining real-time aspects of user movement by utilizing check-in time, which is available with per second granularity.

[1] `https://www.fitbit.com/sg/aria`



DISTANCE 6.01 km

DURATION 42m:52s

CALORIES 508 kcal

AVG PACE 07:08 min/km

MAX PACE 04:35 min/km

MIN ALTITUDE 316 m

MAX ALTITUDE 332 m

ASCENT 20 m

AVG HEART RATE 120

MAX HEART RATE 141

HYDRATION 1.05 L

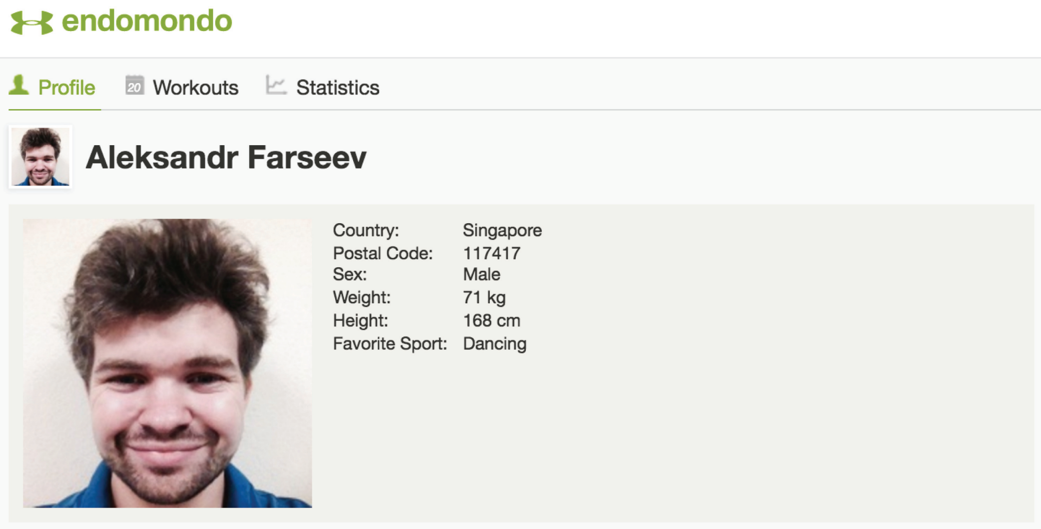
WIND --

TEMPERATURE --

HUMIDITY -- %

WEATHER Hazy sunshine

Figure 3.1: Publicly available Endomondo workout



Aleksandr Farseev

Country:	Singapore
Postal Code:	117417
Sex:	Male
Weight:	71 kg
Height:	168 cm
Favorite Sport:	Dancing

Figure 3.2: Publicly available Endomondo user profile

3.1.2 Cross-Network User Identification and Data Collection

To collect the data generated by the same individuals from multiple social networks, it is necessary to solve the task of cross-network user identification [183]. One approach to solving this issue is the utilization of social network aggregation platforms, such as About.me[2]. About.me enables people to create a public online "name card" that includes a self-described biography and links to the one's social network accounts and personal websites. This information is usually sufficient for further multi-source data collection. However, according to Alexa Internet Survey[3], About.me is ranked as an unpopular website ($7,575th$ worldwide position), while some research groups describe About.me users as atypical [97]. The tendency of About.me users to leverage on multiple social networks aiming to benefit from having an online content aggregation point could explain their abnormal online behavior [97]. Consequently, their generated data could be biased by their online goals and may lead to sub-optimal user profiling performance at both individual and group level. Furthermore, the available amounts of appropriate About.me data does not allow for conducting research in Wellness and Urban Mobility domains at a large scale.

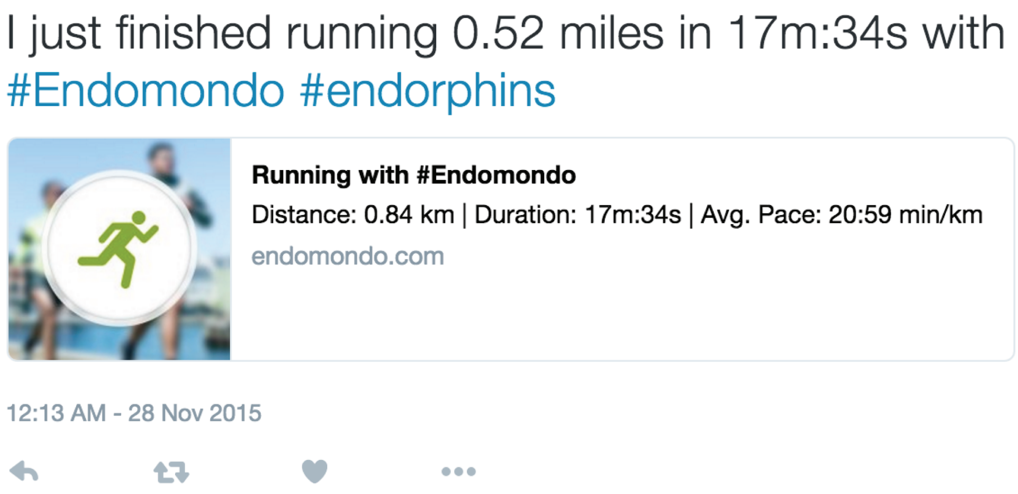
Motivated by the gaps mentioned above, in this thesis we propose the new approach for harvesting social multimedia data at a large scale. Specifically, we utilize one social network (i.e. Twitter) as a "sink" for re-post data from other social networks (Instagram, Foursquare, Endomondo), so that the linkage between social media portals can naturally be obtained with close to zero error rate. The data harvesting process can be described by the following three steps:

1. **Search of seed users:** We collect the so-called "seed" set of Twitter users, who were recently active in other social networks in particular geographical regions. For example, as we mentioned earlier, the sensor data is of crucial importance for physical Wellness profiling. We thus perform Twitter search[4] for those users who have recently posted the hashtags "#Endomondo" or "#endorphines" together with the expression "I just finished" in their Twitter timeline, which allows us to find users who are active in both Endomondo and Twitter. These users form the "seed" user set for our research in Wellness domain. Meanwhile, location semantics data from Foursquare is known to be essential for Urban Mobility research. Considering this, we utilize the Twitter geo-enabled search to perform the location-dependent tweets exploration in Singapore, London, and New York. To do so, we search for those Twitter users who have recently posted the keywords "swarmapp.com" in their tweets around Singapore, London, and New York. These users form the "seed" set for our research in Urban Mobility domain.

[2] `http://about.me/`

[3] `http://www.alexa.com/siteinfo/about.me` Retrieved on 23 January 2017

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



I just finished running 0.52 miles in 17m:34s with #Endomondo #endorphins

Running with #Endomondo

Distance: 0.84 km | Duration: 17m:34s | Avg. Pace: 20:59 min/km

endomondo.com

12:13 AM - 28 Nov 2015

Figure 3.3: Twitter repost of an Endomondo workout

2. **User-generated content collection:** We then start a Twitter "stream"[5] that involve all "seed" users and download their multi-source user-generated content (via URLs to the original posts, provided in tweets). Noticeably, in this step we solve the user identification problem, since we fulfill the cross-network account mapping whenever users perform a cross-network activity (i.e. posting an Instagram image on Twitter)[6].

3. **Ground truth collection:** During the Twitter crawling process, we conducted daily monitoring of Endomondo users' accounts and recorded all the BMI updates during the whole data collection period[7].

3.2 Multi-Source Datasets

By following the above data collection procedure, we harvested two sets of multi-source data — NUS-MSS [54] and NUS-SENSE [51]. NUS-MSS dataset was harvested in three geographical regions with the obligation to all its users to be registered in Foursquare. The applied requirement allows for NUS-MSS usage for Urban Mobility research. NUS-SENSE dataset was collected without geographical restrictions, but with the requirement to all its users to be active in Endomondo (be registered in Endomondo and perform Endomondo

[5] `dev.twitter.com/streaming/public`

[6] To be confident in cross-network account mapping correctness, we remove all retweets and ensure that all cross-network messages were automatically tweeted via cross-linking functionality of the corresponding social media App (i.e. Instagram)

[7] Users' Weight and Height updates were used for computing their BMI, which was utilized to estimate "BMI Trend" ground truth attribute.

Table 3.1: Number of data records in NUS-MSS dataset

City	Num. of users	Num. of tweets	Num. of ch-ins	Num. of 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

workouts). The last restriction equips every NUS-SENSE user with large-scale sensor data, which allows for the dataset utilization in Wellness application domain.

3.2.1 NUS-MSS dataset

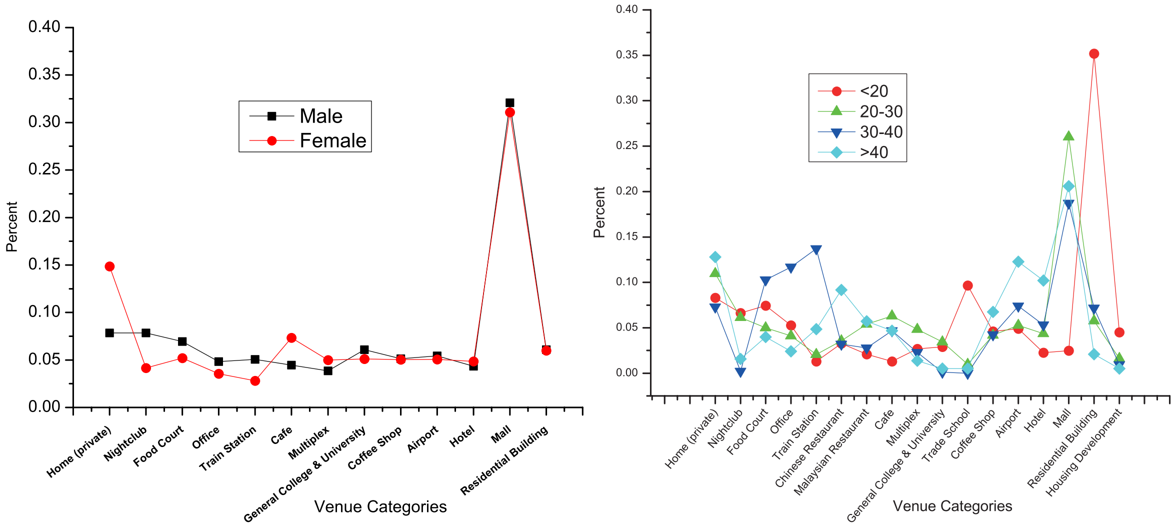
Overview

NUS-MSS [54] data crawling process started on 10 July 2014 in Singapore, London and New York regions, and ended on 20 Dec 2014. Considering the sensitivity of users' private information, the dataset was released in a form of features instead of the original data. The number of data records for each city is summarized in Table 3.1.

Ground Truth

The age, gender ground truth was crawled from the publicly available Facebook pages. The Facebook-Foursquare-Twitter account ID mapping was obtained via Foursquare REST API[8]. 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 also released other ground truth records, such as Education, Employment, Relationship, and Income Level. The venue categories (venue semantics) ground truth was collected during check-in crawling process (see Section 3.1.2), since such information is available on check-ins' web pages.

[8] `developer.foursquare.com`



(a)

(b)

Figure 3.4: Illustration of the distributions of venue categories and image concepts over genders and ages NUS-MSS dataset. (a) illustrates the distribution of Foursquare venue categories for different genders; while (b) illustrates the distribution of Foursquare venue categories for different age groups

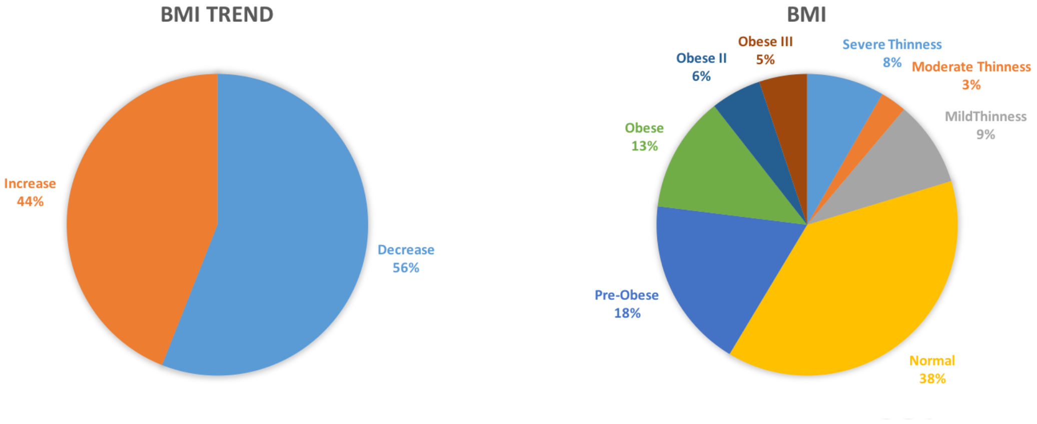
Table 3.2: Number of data records in NUS-SENSE dataset

Source	Twitter	Endomondo	Foursquare	Instagram
Num. of Posts Labels	1676310	140926	19743	48137
Num. of Users Labels	5375	4205	609	2062
Num. of BMI Categories Labels	1052	870	116	372
Num. of BMI Trands Labels	147	136	18	51

3.2.2 NUS-SENSE dataset

Overview

NUS-SENSE [51] data have been harvested in the period of 1 May 2015 to 28 Aug 2015 by following the process in Section 3.1.2. To preserve users' privacy, the dataset is released in the form of data representations (features) and anonymized multi-source user timelines, instead of the original user posts. Table 3.2 lists the dataset statistics.



(a)

(b)

Figure 3.5: Illustration of the distributions of users among different (a) BMI categories and (b) "BMI Trends" in NUS-SENSE dataset

Ground Truth

During the Twitter crawling process, we monitored the Endomondo users' accounts daily and recorded all the BMI updates during the whole data collection period. The average between a user's last Weight and Height updates was used to compute his/her BMI. The difference between a user's first and last Weight and Height updates was used to estimate his/her "BMI Trend". In the dataset, users are well distributed in all BMI categories. From Figure 3.5, it is apparent that the highest percentage of users (38%) belong to "Normal" BMI category and the lowerest percentage of users (3%) belong to "Moderate Thinness" BMI group. It suggests that there are sufficient data samples to train a supervised model for BMI category classification task, but the evaluation must be conducted on each BMI category separately to avoid the imbalanced dataset evaluation problem. It is also noticed that the distribution of users among "BMI Trend" is slightly shifted to the "Decrease" (56%) category, which can be explained by Endomondo users' general intention to "lose weight".

3.3 Overview of User Profiling Framework

After the explanation of our data and the collection procedure, we are now ready to introduce the high-level structure of our user profiling framework in Wellness and Urban Mobility domains. We outline the framework structure on Figure 3.6.

From the Figure, it can be seen that the user profiling pipeline consists of four distinct steps:

1. **Multi-Source data collection:** To harvest multiple data sources from the same

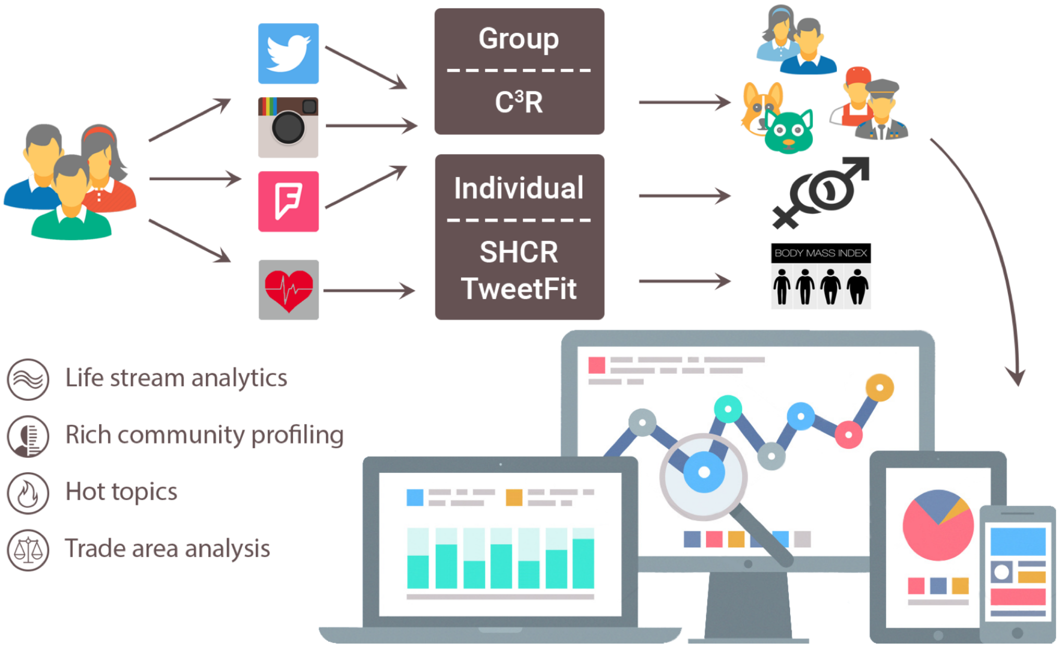


Figure 3.6: User Profiling framework block structure

set of users, we utilize the crawling strategy as outlined in Section 3.1.2. Since the framework performs seamless user profiling in Wellness and Urban Mobility domains, it is necessary to deploy at least four independent Twitter stream multi-source crawlers — one for each geographical region (Singapore, New York, London) to harvest Urban Mobility data, and one for sensor data crawling.

2. **Individual and group user profiling:** This step includes the employment of individual and group user profiling techniques on previously harvested multi-source social media data. Individual user profiling in Wellness Domain can be split into two sub-tasks:

   (a) *Individual demographic profiling* from multi-source multi-modal data via utilizing our developed cross-source ensemble learning technique "SHCR". Briefly, SHCR unites multiple machine learning approaches into a late-fused ensemble and iteratively learns the weight of each data modality by applying Stochastic Hill Climbing with Random Restart (SHCR) search. The data sources utilized are Twitter, Instagram, and Foursquare.

   (b) *Individual physical Wellness profiling* from multi-source multi-modal data and the data from wearable sensors. We treat individual user profiling in physical Wellness domain as supervised learning task. To do so, we automatically infer user BMI and "BMI Trend" attributes by applying Regularized Multi-Source Multi-Task

Learning on data from Twitter, Instagram, Foursquare, and Endomondo.

Group user profiling in Urban Mobility domain is treated as unsupervised learning problem (user community detection problem). It is approached by performing regularized multi-layer spectral clustering, which is contained by automatically-constructed social-network relationship graph.

3. **Application of individual and group user profiling:** At this step, we use individual and group user profiling results (predicted values of individual user attributes and user communities assignment) for particular applications. Previously inferred demographic attributes together with predicted Physical Wellness attributes are combined into online-to-offline individual Wellness Profile. Extracted user communities (group user profile), in turn, serves as a backbone of the venue category recommendation system in Urban Mobility domain.

4. **Integration into social multimedia analytics platform:** The last step is the integration of the results obtained at the previous stages into one social multimedia analytics platform. The step includes deployment of data gathering approaches together with user profile models into a cloud environment, their automatization and presentation to analytics consumers.

3.4 Summary

In this chapter, we first presented inter-network account mapping and data harvesting procedures, as well as provided an overview of our user profiling framework. Second, we introduced the technique for cross-source user identification, which is based on cross-posting activity monitoring. We then described the multi-source data harvesting procedure for Wellness and Urban Mobility application domains. Lastly, we provided an overview of our user profiling framework and described each step of its execution pipeline. The major contribution of this chapter is the cross-network data gathering and alignment scheme that allows for large-scale data gathering from an arbitrary number of social networks that are enabled with inter-network reposting feature.

Chapter 4

Integrating Sensors And Multiple Social Media Data For Individual Wellness Profile Learning

In this chapter, we propose two multi-source user profiling components that form our individual user profiling solution in Wellness domain. The first model infers users' demographic attributes via ensemble learning, while the second approach applies regularized multi-task learning to infer BMI category and "BMI Trend" attributes from social media and sensor data. Extensive experiments demonstrate the effectiveness of the proposed solutions.

4.1 Multi-Source Ensemble Learning for Demographic Profiling

In this section, we explore the so-called demographic user profiling, which is the essential part of Wellness profile. The individual *demographic profile* typically includes age and gender attributes. These demographic attributes were also studied in popular evaluations named PAN [134, 136]. In general, the demographic profile is often used in marketing to describe a demographic grouping or a market segment. By considering users' demographic aspects, marketing specialists able to gain deeper insights into users' behavior and interests understanding. Meanwhile, 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. Most importantly, demographic profiling is essential for Wellness domain. For example, to evaluate the degree of weight-related health problems based on one's Body Mass Index (BMI), it is important to consider age and gender of the person. Precisely, same BMI category may serve as a sign of multiple chronic diseases for young adults, while considered to be normal for the older ones [40].

Demographic 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 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 check-ins. User profile learning based on such heterogeneous information requires new theories that can seamlessly integrate different data sources in a complementary way.

Inspired by the challenges above, in this section we seek to address the following research question: **(RQ1) Is it possible to boost individual user profiling performance by incorporating multi-modal data from multiple social networks?**

To answer the proposed research question, we perform comprehensive demographic profiling of users from multiple social networks. In particular, we infer users' demographic attributes by modeling demographic profile learning as the ensemble of classifiers trained based on multi-source dataset [54]. We verify that multi-source multi-modal data fusion is able to improve the demographic profile performance significantly, as compared to mono source utilization.

4.1.1 Data Representation

In order to perform multi-source demographic profiling, multi-modal data from different sources must be efficiently represented in a mutually-consistent fashion. Below, we describe our adopted data representation (feature extraction) approach for Location, Text, and Image data representation.

Text Features

Prior to the textual feature extraction, the following pre-processing operations were performed:

- *Re-tweet filter:* We did not incorporate re-tweet data in our experiments, because it does not bring any information about users demographics.
- *Slang transformation:* We converted all slang words to synonyms based on the external dictionary[1] to reduce the number of possible terms in tweets.
- *User mentions, place mentions and hashtags filter:* We removed hashtags, user mentions and place mentions from the tweets texts.

[1] www.urbandictionary.com

After the cleaning process described above, we extracted and normalized 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) [21]. The tweets of all users can be represented as a mixture of different topics. We empirically set $\alpha = 0.5$, $\beta = 0.1$ and utilized KNIME[2] to train the topic model with $T = 50$ topics for $1,000$ iterations. We ultimately built the topic-based user feature vectors.
- *Linguistic Features.* We extracted LIWC linguistic features [127], which were founded to be a powerful mechanism for age and gender prediction purposes [155]. 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*[3], *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 [155]. 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.

Location Features

We obtained 764 venue categories from Foursquare[4]. 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 764 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.

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)

[2] www.knime.org

[3] sentiwordnet.isti.cnr.it

[4] developer.foursquare.com/categorytree

Table 4.1: Features summary

Modality	Features	Dimension
Text	Latent topics [21]	50
	LIWC Features [127]	74
	Heuristic Features	14
Image	ImageNet concepts [44]	1000
Location	Venue semantics [54]	764

features [44]. 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 [84] to 50 condensed features.

We summarize all the above features and their characteristics in Table 4.1.

4.1.2 Demographic Profiling

Problem Formulation

As it was mentioned before, the task of multi-source demographic profiling is a challenging problem. First, data sources are incomplete, since some users may use certain social networks but not others. The issue can be approached by using data completion techniques [121] or by amending incomplete data samples. However, data completion techniques may lead to the noise propagation while the insufficient amount of data (caused by data amendment) may result in the so-called “curse of dimensionality” issue. Second, different data modalities (and the corresponding data representations) are heterogeneous with respect to the nature of the data (e.g. text and images) and data representation dimension (e.g. 74 text LIWC features and 1000 ImageNet image concepts). Such problem is often ignored by combining data in the so-called “early-fusion” manner, where features are concentrated in one long feature vector to be fed into the classifier [171]. Such an approach often leads to overfitting [73]. Moreover, the natural difference of data sources will not be properly considered without taking into account the difference in social network usage patterns [154].

In the previously-reported demographic profiling results [132, 9, 135, 134, 136, 133], one can notice the higher predictive power of ensemble models, as compared to the single-classifier baselines. The phenomena can be explained by the ability of ensembles to efficiently combine diverse text feature types and perform joint learning from them. One of the most widely-utilized models was the Random Forest ensemble, which is essentially the combination of bagging data manipulation strategy [25] and the idea of randomized decision tree classifier [26]. Formally, given a training set of pairs, where jth data sample is the vector $\mathbf{x}_j$ ($\mathbf{x}_j \in \mathcal{R}^K$) of K features and its corresponding scalar y_j, bagging repeatedly selects a random sample with replacement of the training set and fits the decision trees (built based on a random subset of $M << K$ features) to these samples. Predictions for unseen samples can be further made by averaging the predictions from all the randomly-built regression trees by taking the majority vote [26]. The aforementioned training procedure leads to better ensemble performance because it is able to decrease the variance of the model without increasing the bias. In other words, while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Bootstrap sampling and random feature selection are the ways Random Forest ensemble de-correlates the decision trees in ensemble [77].

In earlier works, it was noticed [155] that the Random Forest often outperforms other machine learning approaches, like SVM or Naive Bayes for Age and Gender classification tasks. Inspired by the success of Random Forest classifier in solving related problems, we first trained it with 145 random trees (determined by 10-fold cross-validation on training test) based on early-fused data (concatenated feature vectors from all data modalities). However, most of the baselines outperformed such a model in terms of classification accuracy. The reason is simply that by training on early-fused data sources, Random Forest omits the significant amount of training data (users who contributed not to all social networks) and does not take into account the natural difference between data sources. To overcome these problems, we employ the late fusion (ensemble learning) approach for joint incorporation of multi-source incomplete heterogeneous data.

Demographic Profiling Ensemble

The structure of the late-fusion demographic profiling framework is presented in Figure 4.1. It can be seen that our framework consists of multi-label Random Forest classifiers trained on five feature types described earlier. The 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 utilized the "SMOTE" data oversampling technique [32].

The five Random Forest classifiers were integrated into the ensemble learning model; with

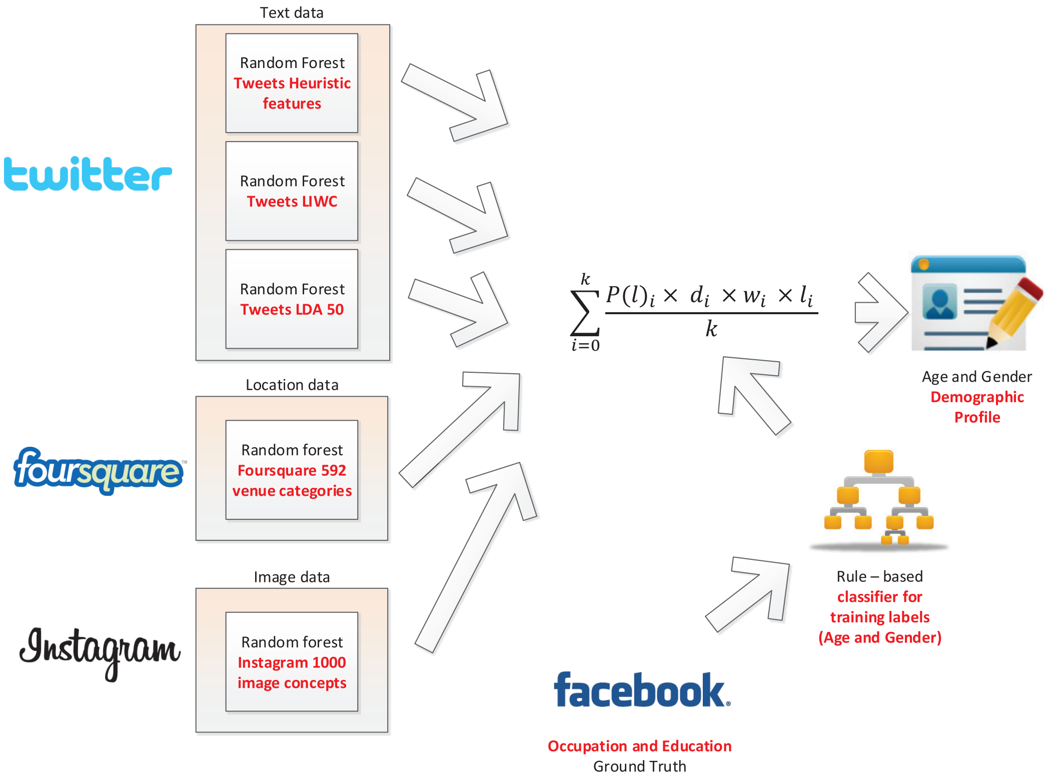

Text data
Random Forest Tweets Heuristic features
Random Forest Tweets LIWC
Random Forest Tweets LDA 50
Location data
Random forest Foursquare 592 venue categories
Image data
Random forest Instagram 1000 image concepts
$\sum_{i=0}^{k} \frac{P(l)_i \times d_i \times w_i \times l_i}{k}$
Age and Gender Demographic Profile
Rule – based classifier for training labels (Age and Gender)
Occupation and Education Ground Truth

Figure 4.1: Demographic profiling ensemble

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 [141], 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 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 one of the Twitter feature 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 scores 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 convergence, by varying the l_{i_k} parameters on each iteration aiming to maximize the overall accuracy.

4.1.3 Evaluation

As we've mentioned 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 [155, 132]. Due to the uneven users' age distribution, we've 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 overrated. 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 Tables 4.3 and 4.2 .

From the Table, it can be seen that various data sources perform differently in profiling social media users. 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.

Finally, the best results for both gender and age group prediction are obtained by the late fusion of all features using our proposed Random Forests ensemble. It outperforms all single source models, different combinations of classifiers, state-of-the-arts techniques, and model trained on early fused sources. These results in **positive reaffirmation of RQ1**[5] by highlighting the usefulness of multi-source data integration for individual user profile learning.

Statistical Analysis

To evaluate the performance of our model from the statistical point of view, we utilize weighted Cohen's Kappa (κ_w) measure (see description in Appendix A). In our case, κ_w quantifies the level of agreement between two raters: 1) multi-source demographic profiling model and 2) ground truth labels obtained from users' Facebook pages (see Section 3.2.1 for details on ground truth collection). The utilization of the weighted Kappa is preferable (as compared to ordinary Cohen's Kappa) due to its ability of dealing with ordered prediction classes (i.e. Age Groups) by making use of a predefined table of class disagreement penalization weights: the higher the disagreement the higher the weight (more details are given in Appendix A). In such a setting, the misclassification into the neighbor groups (i.e. <20 and 20-30) would be penalized less as compared to misclassification into more different categories (i.e. <20 and >40). Generally, the table of weights is a symmetric matrix **W** with zeros in the main diagonal (i.e. where there is an agreement between the two judges) and positive values off the main diagonal. For the case of Gender classification (binary classification), the weight matrix **W** was assigned as a unit matrix (all elements are 1s). For the case of Age Group classification (multi-class classification), values in **W** were assigned according to their squared distance from perfect agreement [27]. We also utilized p-value test (see Section 6.2.2) to evaluate the significance level of the computed κ_w statistic.

Statistical analysis results are presented in Table 4.3. It can be seen that the value of κ_w calculated for Age Group Classification results belongs to the "Fair" level of agreement with Ground Truth labels [91]. The imperfect agreement can be explained by the highly-imbalanced distribution of age group labels in our dataset as well as the limited applicability of the Landis and Koch's labeling scheme [91]. However, the obtained $\kappa_w = 0.297$ demonstrates the statistically-significant (with $p < 0.01$) agreement between our proposed age group classification model and the ground truth, which suggests that the problem of Age Group classification was solved satisfactorily.

The inspiring observations can be made based on Gender classification results. From Table 4.3, it can be seen that our proposed demographic profiling model is able to archive $\kappa_w = 0.745$, which is ranked as "substantial" agreement with the ground truth [91] and only 0.055 lower than "almost perfect" agreement benchmark [91]. By jointly considering classification performance results as well as statistical analysis results, we can conclude that

[5] Is it possible to boost individual user profiling performance by incorporating multi-modal data from multiple social networks?

Table 4.2: Performance of demographic profiling in terms of Macro Accuracy. Data source combinations. RF - Random Forest; NB - Naive Bayes; SVM - Support Vector Machine

Method	Gender	Age
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**

Table 4.3: Performance of demographic profiling in terms of Macro Accuracy. State-of-the-arts techniques. RF - Random Forest; NB - Naive Bayes; SVM - Support Vector Machine

Method	Gender	Age
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
Statistical Analysis		
Weighted Cohen's Kappa	$\kappa_w = 0.745, p < 0.01$	$\kappa_w = 0.297, p < 0.01$

Gender prediction model fits not just research but also industrial requirements [55] and thus can be utilized in the real-world scenario.

Error Analysis

From the previous sections, it can be seen that our proposed demographic profiling framework can achieve better prediction performance as compared to different data source combinations and other user profiling baselines. However, the age group prediction performance is not sufficiently high and thus requires additional investigation to be conducted. In Table 4.4, we present the evaluation results of demographic profiling framework trained on all data sources with respect to each age group. We use Precision, Recall, and the F_1 as evaluation metrics and color different age group performance levels as follows: green — good performance; blue — reasonable performance; brown — lower performance; red — low performance.

It can be seen that 20 – 30 age group can be predicted more accurately as compared to other age groups. The model achieved high F-measure score due to its ability to balance between high Precision (0.886) and comparably high Recall (0.634) scores. The balanced and accurate predictions were achieved due to the high percentage of positive samples in the dataset (see Table 4.4), as well as data representatives. Indeed, our computed statistics on the dataset as well as the recent research suggest that young adults (18-29 years old) are the most intensive Instagram (54%) [67], Foursquare (40%) [64], and Twitter (44%) [67] users. Accurate detection of this category is of crucial importance for public wellness domain since it is useful to have early identification of Wellness problems of young adults with the provision of timely assistance to them. We would also like to highlight the high Precision score obtained, which minimizes the number of false positives in the data, and thus allows for better patient targeting within the ethically-sensitive Wellness domain.

At the same time, the prediction of the two less popular in social media age groups (<20, 30-40) reveals lower performance in terms of F-measure, but higher Recall (see Table 4.4). The high Recall is desirable in Wellness application domain. For example, for the case of online gaming addiction analytics [71] as well as social media dependence monitoring [119], it is essential to influence as many young social media users as possible by delivering the corresponding motivational online content, such as invitations to offline activities and wellness lifestyle programs.

Lastly, let's highlight the unsatisfactory prediction performance of the >40 age group. The possible reason is the low number of training instances of this class, which is also consistent with the population of modern social networks [67, 64]. However, considering the importance of this age category for wellness application, the additional ground truth records can be obtained by applying age-group targeted data gathering strategies (e.g. search for age mentions in Twitter) as well as by utilizing social networks with older user population (e.g. Linkedin). The latter is expected to achieve significant classification results improvement.

Table 4.4: Performance of the demographic profiling framework trained on all data sources with respect to each age group category. Evaluation metrics: Precision, Recall, F_1

BMI Category	% in Dataset	Precision	Recall	F-Measure
$<$ 20	77%	0.330	0.725	0.453
20-30	18%	0.886	0.634	0.739
30-40	3%	0.400	0.667	0.500
$>$ 40	2%	0	0	-

4.2 Multi-Source Multi-Task Learning for Physical Wellness Profiling

The previous section presents an initial study of individual user profile learning in Wellness domain via integration of multiple data sources in ensemble model. We conducted first-order and higher-order learning for demographic profiling in Singapore region. User demographic profile is represented by age and gender attributes. From our experimental outputs, we can drew the conclusions that: different data sources mutually complement each other and their appropriate fusion boosts the user profiling performance. The latter positively answers our research question and encourages further research in this direction. Eventhough we already showed the way to predict such personal wellness attributes as age and gender, it is still necessary to study physical Wellness profiling, since it is related to users' physical health. These may not be done without incorporation of information about actual physical status, which can be approached by bringing in the data from wearable sensors.

This section focuses on the problem of individual wellness user profiling based on data from multiple social networks and wearable sensors. Here, an individual wellness profile involves personal user attributes [49] such as demographics (age, gender) [54], Body Mass Index (BMI) category[6], personality [28], or chronic disease tendency [4]. In this part of our work, we focus on two important physical wellness attributes - BMI category and “BMI Trend' (the direction of BMI fluctuation over time - Increase/Decrease). Both attributes are closely related and correlated to one's overall health. For example, Field et al. [57] discovered that people whose BMI is higher than 35.0 are approximately 20 times more likely to develop diabetes. Other benefits of such attributes include: a) BMI category can be further used in public health domain to monitor wellness tendencies of social media users at the global level; b) “BMI Trend” information can be utilized by users to rectify their lifestyle (i.e. via

[6] The BMI measure is defined as the body mass divided by the square of the body height. Based on BMI, an individual can be categorized into one out of 8 BMI categories, namely “Severe Thinness”, “Moderate Thinness”, “Mild Thinness”, “Normal”, “Pre Obese”, “Obese”, “Obese II”, and “Obese III” [175].

interactive mobile application or a “Smart Watch”), and by doctors to gain a complete picture of patient’s health.

There are two challenges in addressing individual physical wellness profiling: 1) **Data representation.** Besides the textual data, social media services involve data of various modalities. For example, in Instagram, users share recently taken pictures and videos, while in Endomondo[7] users post information about their workouts, which is strongly dependent on the temporal and spatial aspects. Integration of such heterogeneous multimodal data sources requires development of efficient and mutually consistent data representation approaches. 2) **Data modeling:** Efficient data integration for individual wellness profile learning is a tough challenge since the data from independent media sources is different in nature. Furthermore, multi-source data is often incomplete, which means that some users may not be active on all social networks. Lastly, personal wellness attribute categories (classes) are often inter-related (i.e. the adjacent BMI categories could unite users of similar lifestyle and social media content [105]), which must be taken into account at the learning stage. Development of a learning framework that can handle all these issues is a crucial challenge.

Inspired by previous studies and the challenges above, in this work we seek to address three research questions. First, to support the assumptions behind this study, it is important to understand: **(RQ1) Is it possible to improve the performance of BMI category and “BMI Trend” inference by fusing multiple social media and sensor data?** Second, for further wellness profiling improvement, it is essential to gain insight into: **(RQ2) What is the contribution of sensor data towards BMI category and “BMI Trend” inference?** Finally, to perform efficient data fusion it is necessary to understand: **(RQ3) Is it possible to improve the performance of BMI category and “BMI Trend” inference by incorporating inter-category relatedness into the learning process?**

To answer the above research questions, we present a new computational framework for Multi-Task Multi-Source Wellness Profiling (“M^2WP”). We introduce the techniques to represent data from a new sensor data source (the Endomondo workouts) and other social media sources: Twitter, Foursquare, and Instagram, from which we predict the users’ BMI category and “BMI Trend”. To do so, we treat individual wellness profiling as a regularized multi-task learning (MTL) problem, where different data source combinations for each inference category are represented as MTL “tasks”. Concurrently, we consider inter-category relationship by regularizing the MTL model by learning “similar” categories in a mutually-consistent fashion. Last, but not least we perform error and feature importance analysis aiming to understand which BMI categories can be inferred more accurately and which particular feature types facilitate better inference of each BMI category.

[7] endomondo.com

4.2.1 Data Representation

As mentioned earlier, efficient data representation is a significant step in multi-source profile learning pipeline. Its importance is dictated by its ability to bridge the semantic gap between different data modalities by producing mutually-consistent data representations. The extracted features are described below.

Text Features

In this piece of work, we extracted textual data from the following data sources: *Twitter tweets*, *Instagram image captions*, *Instagram image comments*, and *Foursquare check-in comments.* All textual data was aggregated at the individual user level and filtered/corrected by using Microsoft Office Spell Checker[8]. More specifically, we extracted the following features:

- *Latent Topic Features.* We merged all the textual data of each user into a document. All documents from multiple users were projected into a latent topic space using Latent Dirichlet Allocation (LDA) [21]. We trained the topic model with $T = 50$ topics (revealed the lowest perplexity score on interval from $T = 1, 5, ..., 200$), with $\alpha = 0.5$, $\beta = 0.1$ (determined empirically). Feature vector was constructed as a distribution of each user among detected latent topics: $\mathbf{u}_{i,LDAText} = (tl_{i,1}, tl_{i,2}, ..., tl_{i,50})$.
- *Writing style features.* As in [54], we extracted such writing style features as: *number of mistakes per post*, *number of slang words per post*, *average post sentiment*, etc. In total, 14 features were extracted: $\mathbf{u}_{i,HeurText} = (th_{i,1}, th_{i,2}, ..., th_{i,14})$.
- *Lexicon-based features.* We used crowd-sourced lexicon of terms associated with controversial subjects from the US press [106] and the lexicon of terms' healthiness category [105]. Additionally, we extracted food type and average calorie content from each post by incorporating the Twitter Food Lexicon [152]. Lastly, all the food-related features were grouped in one user-dependent vector of size 32: $\mathbf{u}_{i,FoodText} = (tf_{i,1}, tf_{i,2}, ..., tf_{i,32})$.

Venue Semantics Features

Similar to [54], we represented location data as a distribution of users' check-ins among 764 Foursquare venue categories. Such users' representation describes venue semantics preferences of Foursquare users and can be related to users' food and activity preferences [105]. To overcome the data sparsity problem, we further reduced the data dimensionality

[8] products.office.com

by extracting Top 86 principal components [84] (preserves 85% of variance) and representing the venue semantics data as: $\mathbf{u}_{i,LocCatPCA} = (lcPCA_{i,1}, lcPCA_{i,2}, ..., lcPCA_{i,86})$.

Mobility and Temporal Features

As stated earlier, user mobility and temporal aspects are important in user profile learning [54, 116]. The mobility features were computed based on users' *areas of interest (AOIs)* [131], which is, essentially, the geographical regions of high user's check-ins density (regardless the check-in venue semantics). AOIs were obtained by performing density-based clustering [143] over the check-ins of each user (with $\epsilon = 0.05$, the minimum number of points $MinPts = 3$) and consider the convex hull of each cluster as a new AOI. The user's AOIs represent their geographical mobility patterns [116, 131] and is correlated to their wellness lifestyle. We extracted the following 12 mobility and temporal features ($\mathbf{u}_{i,MobTemp} = (mt_{i,1}, mt_{i,2}, ..., mt_{i,12})$):

- *Average number of posts during each of the* 8 *daytime durations*, where each time duration is 3 hours long (i.e. $15 - 18$). The feature is an indicator of user's temporal online activity and related to their wellness [31].
- *Number of areas of interest (AOI).* The feature reflects the mobility side of user's physical activity. For example, users with more AOIs may maintain more intensive lifestyle. AOIs also represent user's frequent areas of activity (i.e. home, office, university/school) [131] and may indicate how far users are willing to travel on a daily basis. It, in turn, reflects users' wellness attributes, since it is related to their lifestyle.
- *Median size of user's AOIs*[9], which indicates users' mobility inside each AOI. The feature is an indicator of user's everyday traveling habits inside their main activity areas. It can, thus, indicate the approximate distance that the users are willing to travel on foot.
- *The normalized number of AOI outliers.* The feature shows how often users visit places that are not located inside their AOI, which may reflect the users' "conventionalism" level. Mainly, it shows how often users deviate from their regular mobility patterns, which is related to users' physical activity level.
- *Median distance between AOIs.* This feature reflects users' movement at intra-city/inter-city/international level. Specifically, it shows how often and how far users travel between their activity zones. It could be daily trips to job place, vocation journeys or business trips. Such a feature is related to users' lifestyle and, thus, wellness.

Even though the venue semantics and mobility features were extracted from the same data source, they are different in nature. Specifically, the venue semantics features represent users'

[9] Where AOI size is defined as the median distance between a center of mass and all points inside AOI

geo-independent location preferences [172] (i.e. Restaurant, Cinema, Church). At the same time, users' mobility features describe users' movement in space and related to their activity level in a particular geographical region [116].

Visual Features

In order to represent data from visual modality, we computed distribution of each Instagram photo among the 1000 ImageNet visual concepts [44]. We then averaged the values of each 1000 visual concept to obtain a user-dependent feature vector of size 1000. The image concept detection was performed by leveraging on GPU implementation [82] of a state-of-the-art deep learning framework GoogleNet [159]. Similar to venue semantics features, we extracted Top 150 principal components [84] (preserves 85% of variance) from image concepts data: $\mathbf{u}_{i,ImgConPCA} = (vcPCA_{i,1}, vcPCA_{i,2}, ..., vcPCA_{i,150})$.

Sensor Features

One of the most informative data sources for wellness profile learning and, at the same time, the most challenging regarding consistent representation, is the data from sensor devices. The main problem is the necessity to consider the spatial-temporal aspect of sensor data, since the static characteristics of each data sample can be biased toward individual's health and, thus, may not be suitable for further integration with other sources for joint profile learning. To represent sensor data consistently with other data modalities, we incorporated the following feature types:

- *Workout statistics.* We computed the following averaged features using all sensor data samples for each user: *Distance (Ascend/Descend)*, *Speed*, *Duration*, *Hydration*. Such features represent users' activity levels and is thus related to one's wellness attributes. In total, 10 features were extracted: $\mathbf{u}_{i,SensStat} = (ss_{i,1}, ss_{i,2}, ..., ss_{i,10})$.
- *External sensors statistics.* Apart from wearable sensor features, we leveraged data from external weather sensors (where it is available) such as *Wind Speed* and *Weather Type*. It is known that weather may affect one's exercise productivity. It can thus serve as a useful complimentary signal for wellness attribute inference. To make the feature set appropriate for model training, we represented each user as a distribution among 44 weather types from our dataset, computed through all user's workouts: $\mathbf{u}_{i,SensExt} = (se_{i,1}, se_{i,2}, ..., se_{i,44})$.
- *Workout semantics.* We also represented sensor data as a distribution of users' workouts among 96 Endomondo workout categories: $\mathbf{u}_{i,SensWCat} = (sc_{i,1}, sc_{i,2}, ..., sc_{i,96})$. The representation describes users' exercise preferences and also related to their wellness.
- *Frequency domain features.* It is important, to incorporate the temporal aspect of

Table 4.5: Features summary

Modality	Features	Dimension	Final dimension
Text	Latent topics [21]	50	50
	Lexicon-based [51]	32	32
	Writing style [54]	14	14
Image	ImageNet concepts [44]	1000	150 (PCA)
Location	Venue semantics [54]	764	86 (PCA)
	Mobility and Temporal	12	12
Sensor	Workout semantics	96	96
	Frequency domain	495	54 (PCA)
	Workout statistics	10	10

sensor data, since the signal fluctuations during an exercise are often related to one's wellness attributes. However, the user's workouts are not directly comparable due to the differences in the exercise types and duration, or the inconsistency in signal sampling rates. To incorporate the temporal aspect of sensor data without binding to actual signal duration, it is useful to project time series data into the frequency domain. To perform such a transformation, we extracted features for each workout, in three steps as follows: (i) We applied spline interpolation to fill in missing sensor signal measurements (in cases of sampling rate < 1Hz or when it is inconsistent).(ii) We transformed all data points from jth workout with r sensor signal measurements ($\mathbf{w}_{j,u_i} = (d_1, d_2, ..., d_r)$) into relative differences between subsequent sensor signal measurements: $\bar{\mathbf{w}}_{j,u_i} = (d_2 - d_1, d_3 - d_2, ..., d_r - d_{r-1})$. (iii) We applied Fast Fourier Transform [23] followed by low band-pass-filter ($0 - 0.5$ Hz) to construct the energy distribution among the 99 *frequency bins* for each of five sensor signal types, namely, *Altitude, Cadence, Speed, Heart Rate (HR), and Oxygen Consumption (Oxygen)*. We then merged these 5 vectors together and took an average among all workouts to obtain a frequency domain feature vector of size 495 for each user. Such a data representation is no longer inconsistent among different users, since it models the "changes" of sensor signal, rather than the absolute values of data points. To overcome the data sparsity problem, we reduced the frequency-domain data representation dimensionality by extracting its Top 54 principal components [84] (preserves 85% of variance) and represented the data as: $\mathbf{u}_{i,SensFreqPCA} = (sfPCA_{i,1}, sfPCA_{i,2}, ..., sfPCA_{i,54})$.

We summarize all the above features and their characteristics in Table 4.5.

4.2.2 Physical Wellness Attributes Inference

To answer research questions **(RQ1)** and **(RQ2)**, it is necessary to develop and evaluate a machine learning solution that will fuse data from multiple social media sources and wearable sensors for wellness attributes inference. Meanwhile, to answer research question **(RQ3)**, it is essential to consider inter-category relatedness at the learning stage. In the following sections, we will address these issues one by one.

Problem Formulation

The problem of multi-source personal user profile learning is not an easy task since multiple data sources are often sparse and block-wise incomplete [153]. This means that some groups of users may contribute content only to certain social networks (Figure 4.2). Traditionally, the problem can be solved by simply skipping incomplete data samples or by using data completion techniques such as NMF [121, 153]. However, these two approaches may lead to the "curse of dimensionality" problem in the first case, and to noise propagation in the second. To overcome these problems, it is necessary to develop an efficient source fusion technique [54]. Second, the categories of the wellness attributes are often inter-related. For example, the adjacent BMI categories (i.e. "Obese II" and "Obese III") are more related to each other as compared to those that represent completely different BMI groups (i.e. "Severe Thinness" and "Obese III"). Such relations must be properly modeled.

In this work, we treat the problem of multi-source individual wellness attributes inference as a multi-task learning (MTL) [30] problem. MTL is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias by learning tasks in parallel while using a shared representation [30]. In other words, MTL aims to improve the performance of multiple classification tasks by learning them jointly. Such a property of MTL makes it naturally applicable to the problem of Multi-Source learning and serves the reason why we've selected MTL idea as a backbone of our multi-source learning approach. One significant issue in multi-task learning is how to define and employ a commonality among different tasks [184]. Intuitively, different data source combinations may share common knowledge for predicting wellness attributes, and the "similar" wellness attribute categories must be correspondingly represented inside the inference model. By following this philosophy, we define a multi-task learning task as a unique combination of different sources for a given category (Figure 4.2), while constraining the overall model by considering the relatedness among adjacent wellness categories.

Notation: In the rest of this chapter, we use uppercase boldface letters (i.e. $\mathbf{M}$) to denote matrices, lowercase boldface letters (i.e. $\mathbf{v}$) to denote vectors, lowercase letters (i.e. s) to denote scalars, and uppercase letters (i.e. N) to denote constants. For matrix $\mathbf{M} = (m_j^i)$,

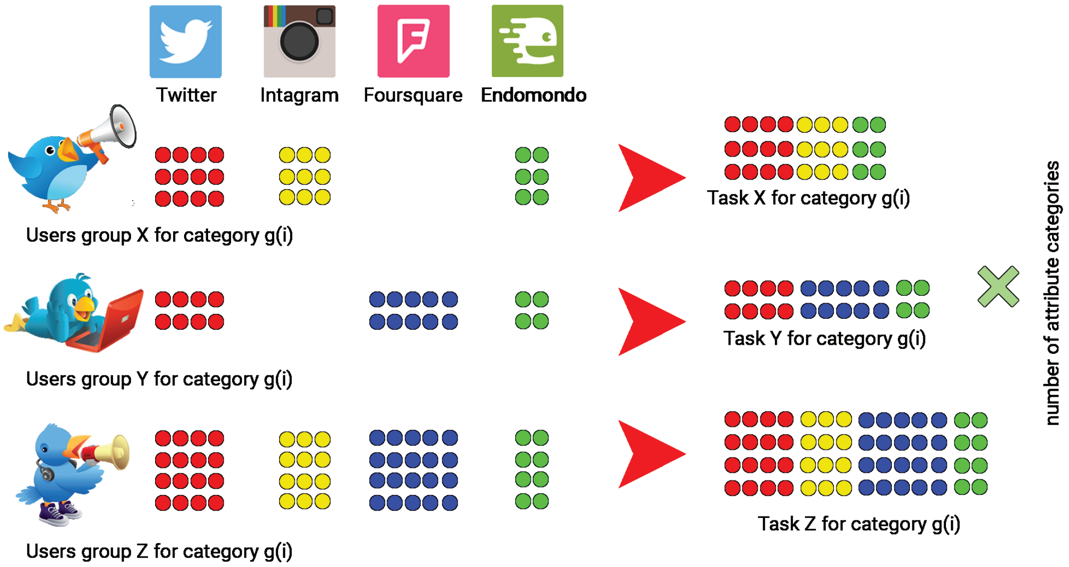

Twitter
Intagram
Foursquare
Endomondo
Users group X for category g(i)
Users group Y for category g(i)
Users group Z for category g(i)
Task X for category g(i)
Task Y for category g(i)
Task Z for category g(i)
number of attribute categories

Figure 4.2: Incorporating block-wise incomplete data into multi-task learning model.

$\|\mathbf{M}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|m_j^i\right|^2}$ is the ℓ_2 (Frobenius) norm, while the $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^{n}\left\|\mathbf{m}^i\right\|$ is the $\ell_{2,1}$ norm [98] ($\mathbf{m}^i$ is the ith row of the matrix $\mathbf{M}$). To simplify the reading process, we summarize all defined notations in Table 4.6.

Modeling Multi-Source Fusion

First, we propose a sparse model that mitigates the problem of joint learning from sensor and social media data aiming to infer BMI category and "BMI Trend" attributes.

Suppose that there is a set of N exclusively labeled data samples, $S \geq 2$ data sources (modalities), and $G \geq 1$ categories of the inference attribute. We divide the dataset into T tasks, where *each task t is represented by the unique combination of available data sources for each attribute category.* The number of features of task t is denoted as D_t; the number of data samples of task t is denoted as N_t, and the number of different existing combinations of sources is denoted as $\hat{T}$ (so that $T = \hat{T} \times G$). Formally, each of the T tasks can be defined as a set of pairs (jth data sample $\mathbf{x}_j^t$ and its corresponding label y_j^t):

$$t = \{(\mathbf{x}_j^t, y_j^t) \mid j = 1...N_t,\ \mathbf{x}_j^t \in\ R^{D_t},\ y_j^t \in\ \{-1;1\}\}^{10}.$$

The prediction for jth data sample of task t is given by the linear model:

$$f_t(\mathbf{x}_j^t; \mathbf{w}^t) = \mathbf{x}_j^{t\intercal}\mathbf{w}^t,$$

Table 4.6: Adopted notations

Notation	Description
N	Number of exclusively labeled data samples
$S(\geq 2)$	Number of data sources (data modalities)
$G(\geq 1)$	Number of inference attribute categories (for BMI category, $G = 8$; for "BMI Trend", $G = 1$)
g	Inference attribute category (class). For example, "Obese" or "Normal" in case of BMI category attribute.
T	Number of multi-task learning Tasks
t	A multi-task learning Task
D_t	Dimensionality (feature vector dimension) of the task t
D_{max}	Maximum possible dimensionality of a task
N_t	Number of data samples of the task t
$\hat{T}$	Number of different existing combinations of sources
$f_t(\mathbf{x}_j^t; \mathbf{w}^t)$	Linear prediction model for the jth data sample of task t
$\mathbf{w}^t \in \mathbb{R}^{D_t}$	Model parameter vector of task t
$\mathbf{W}$	All model parameters, denoted as linear mapping block matrix
$\Gamma(\mathbf{W})$	Objective function
$\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y})$	Loss function
$\Upsilon(\mathbf{W})$	Sparsity regularizer
$\Omega(\mathbf{W})$	Inter-category smoothness regularizer
$\rho(s, f)$	Index function that denotes all the model parameters of the fth feature from the sth source
$\xi(t, g)$	Index function that picks up the model parameter ($\mathbf{w}_{g+1}^t$), which corresponds to the attribute category $g + 1$ (adjacent to g)

where $\mathbf{w}^t \in \mathbb{R}^{D_t}$ is the model parameter vector of task t. In case of multi-attribute inference, the prediction for jth data sample of task t is given by:

$$f_{C_{D_t}}(\mathbf{x}_j^{t \in C_{D_t}}) = \arg\max_{t \in C_{D_t}} f_t(\mathbf{x}_j^t; \mathbf{w}^t),$$

where C_{D_t} is the set of D_t-dimensional[11] tasks. All model parameters are denoted as the linear mapping block matrix $\mathbf{W}$:

$$\mathbf{W} = (\mathbf{w}^{1\intercal}, \mathbf{w}^{2\intercal}, ..., \mathbf{w}^{T\intercal}) \in \mathbb{R}^{D_{max} \times T}.$$

The optimal $\mathbf{W}$ can be found by solving the following optimization problem:

$$\arg\min_{\mathbf{W}} \Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \lambda \Upsilon(\mathbf{W}), \tag{4.1}$$

where $\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ is the loss function, $\Upsilon(\mathbf{W})$ is the sparsity regularizer that selects the discriminant features to prevent high data dimensionality [98, 4], and λ controls the group sparsity.

The loss function $\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y})$ term can be replaced by a convex smooth loss function. In this work, we adopt the logistic loss:

$$\Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \log(1 + e^{-y_i^t f_t(\mathbf{x}_i^t; \mathbf{w}^t)}). \tag{4.2}$$

To incorporate feature selection into the objective [98], we define $\Upsilon(\mathbf{W})$ as:

$$\Upsilon(\mathbf{W}) = \sum_{s=1}^{S} \sum_{f=1}^{F_s} \left\| \mathbf{w}_{\rho(s,f)} \right\|, \tag{4.3}$$

where F_s is the sth source feature vector dimension, and $\rho(s, f)$ is the index function that denotes all the model parameters of the fth feature from the sth source. The Υ term is the $\ell_{2,1}$ norm [153], which leads to a sparse solution (controlled by $\lambda \geq 0$) via constraining all models that involve source s to share a common set of features.

[11] Each inference attribute category g corresponds to $k_g \leq 1$ tasks of dimension D_t.

Modeling Inter-Category Smoothness

The above optimization problem treats all prediction attribute categories equally, while in real world scenario (i.e. detection of the user's BMI category), the attribute categories are often inter-related. For example, the adjacent BMI categories "Obese I" and "Obese II" could be similar to each other, since users from these BMI groups may follow similar lifestyles. To incorporate such relatedness, we additionally regularize the model by "Inter-Category Smoothness" term, which constrains model parameters of the adjacent attribute categories to be close to each other:

$$\Omega(\mathbf{W}) = \sum_{t=1}^{\hat{T}} \sum_{g \in C_{D_t}} \kappa_{g,\xi(t,g)} \left\| \mathbf{w}_g^t - \mathbf{w}_{\xi(t,g)}^t \right\|^2 , \tag{4.4}$$

where $\xi(t, g)$ is the index function that picks up the model parameter $(\mathbf{w}_{g+1}^t)$, which corresponds to the attribute category $g + 1$ (adjacent to g) from the combination of data sources C_{D_t}. The coefficient $\kappa_{g,\xi(t,g)}$ is the pre-defined weight of the adjacent categories' relationship. In our study, we did not incorporate weighting into category inter-relatedness chain, due to the lack of prior knowledge:

$$\kappa_{g,\xi(t,g)} = \begin{cases} 1 & g = 1...G-1; \\ 0 & g = G. \end{cases} .$$

However, in the cases when such knowledge is available [154, 4], the weighting can be naturally introduced.

The final optimization framework that integrates data from multiple social networks and wearable sensors, incorporates feature selection, and consider inter-category relationship, is defined as follows:

$$\Gamma(\mathbf{W}) = \arg\min_{\mathbf{W}} \Psi(\mathbf{X}, \mathbf{W}, \mathbf{Y}) + \lambda \Upsilon(\mathbf{W}) + \mu \Omega(\mathbf{W}), \tag{4.5}$$

where $\lambda \geq 0$ and $\mu \geq 0$ are the model parameters that could be obtained, for example, by performing a grid search.

Optimization

The objective function $\Gamma(\mathbf{W})$ is not smooth, since it consists of smooth terms (Ψ, Ω) and non-smooth term (Υ). This means that the conventional optimization approaches, such as

Gradient Decent, are not directly applicable in our case. Inspired by the fast convergence rate of the Nesterov's [98], we reformulate the non-smooth problem from Eq. (4.5) as:

$$
\begin{aligned}
f(\mathbf{W}) = \arg\min_{\mathbf{W}\in\mathbf{Z}} \; & \Psi(\mathbf{X},\mathbf{W},\mathbf{Y}) + \mu\Omega(\mathbf{W}) \\
& s.t. \; \mathbf{Z} = \left\{ \mathbf{W} \mid \|\mathbf{W}\|_{2,1} \le z \right\},
\end{aligned}
\tag{4.6}
$$

where $z \geq 0$ is the radius of the $\ell_{2,1}$-ball, and there is a one-to-one correspondence between λ and z (proof is given in Liu et al. [98] Sec. 4.2).

In Nesterov's method, the solution on each step ($\mathbf{W}_{i+1}$) is computed as a "gradient" of a search point $\mathbf{S}_i$:

$$
\mathbf{W}_{i+1} = \arg\min_{\mathbf{W}} M_{\gamma_i,S_i}(\mathbf{W}),
$$

$$
M_{\gamma_i,S_i}(\mathbf{W}) = f(\mathbf{S}_i) + \langle \nabla f(\mathbf{S}_i), \mathbf{W} - \mathbf{S}_i \rangle + \frac{\gamma_i}{2}||\mathbf{W} - \mathbf{S}_i||^2,
$$

where $\mathbf{S}_i$ is computed from the past solutions:

$$
\mathbf{S}_i = \mathbf{W}_i - \alpha_i(\mathbf{W}_i - \mathbf{W}_{i-1}).
$$

where α_i is the combination coefficient, and γ_i is the appropriate step size for $\mathbf{S}_i$ (can be determined by line search according to Armijo-Goldstein rule). The detailed description of the Nesterov's optimization approach can be found in previous works [98, 182, 4].

Computational Time Complexity Analysis

Since $\Psi(\mathbf{X},\mathbf{W},\mathbf{Y})$, it costs $\mathcal{O}(D_{max}N))$ floating point operations (flops) for evaluating the function value and gradient of the objective function in Equation 4.6 at each iteration [98], where D_{max} and N are maximum possible dimensionality of the task and number of data samples, respectively. At the same time, it was shown [98] that the Euclidean projection, which is necessary for Nesterov's optimization, can be computed in a time of $\mathcal{O}(NT)$, where T is the total number of multi-task learning "tasks". Considering that Nesterov's approach requires $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ iterations for achieving an accuracy of ϵ [107], the overall time complexity is $\mathcal{O}(\frac{1}{\sqrt{\epsilon}}(D_{max}N + NT))$.

On data balancing

To tackle the data imbalance problem at the training stage, we randomly and uniformly selected equal number of negative and positive samples for each binary classification task. In other words, to train task t (where tth feature dimension is D_t) we used all its corresponding positive data samples and the same number of uniformly sampled negative data samples, which were taken from other D_t-dimensional tasks.

4.2.3 Evaluation

In this section, we present the personal wellness profiling evaluation results to answer our-proposed research questions. First, we compare performance of our proposed framework for cases when it was trained based on different data sources and its combinations, which helps to answer **RQ2**. Second, we perform evaluation against baselines, to respond **RQ1** and **RQ3**.

Evaluation Metrics

We explicitly evaluate the performance of our proposed M^2WP framework[12] by solving the problem of individual wellness profiling. Specifically, we present the inference results of two personal wellness attributes: BMI category (eight attribute classes) and "BMI Trend" (binary classification).

The distribution of users among different BMI Categories and "BMI Trends" is presented in Figure 3.5. It can be seen that users are quite well distributed in all BMI categories. The highest percent of users (38%) belong to "Normal" BMI category, and the smallest percent of users (3%) belong to "Moderate Thinness" BMI group. The above suggests that there is enough data samples to train a supervised model for BMI category classification task, but the evaluation must be conducted on each BMI category separately to avoid the imbalanced datasets evaluation problem [54]. To avoid the prevalence of popular BMI categories in evaluation, we use "Macro-Recall" (R_{Mac}), "Macro-Precision" (P_{Mac}), and "Macro-F_1" ($F_{1,Mac}$) metrics, which are the averaged "Precision", "Recall", and "F_1" measures across all categories.

Evaluation Against Individual Data Sources and Source Combinations

To study the corresponding performance of individual data sources and its combinations, we evaluated M^2WP trained on different data source permutations. The Mobility and Venue

[12] $\alpha = 0.1$, $\mu = 0.1$, $\kappa_{g,\xi(t,g)} = \begin{cases} 1 & g = 1...G-1; \\ 0 & g = G. \end{cases}$.

Table 4.7: Evaluation of the M^2WP framework trained on independent data sources and its combinations. Evaluation metrics: Macro Precision, Macro Recall, Macro F_1

Data Source Combination	BMI category prediction	
	R_{Mac}/P_{Mac}	$F_{1,Mac}$
Visual	0.049/0.188	0.077
Venue Semantics & Mobility	0.194/0.107	0.137
Sensors	0.153/0.158	0.155
Textual	0.229/0.146	**0.178**
Visual + Sensors	0.174/0.201	0.186
Visual + Text	0.126/0.245	0.166
Visual + Venue Semantics & Mobility	0.161/0.154	0.157
Text + Venue Semantics & Mobility	0.160/0.204	0.179
Sensors + Venue Semantics & Mobility	0.163/0.233	**0.191**
Sensors + Text	0.148/0.270	**0.191**
Visual + Text + Venue Semantics & Mobility	0.126/0.233	0.163
Sensors + Text + Visual	0.137/0.207	0.164
Sensors + Text + Venue Semantics & Mobility	0.182/0.236	0.205
Sensors + Venue Semantics & Mobility + Visual	0.180/0.283	**0.221**
All Data Sources	0.214/0.292	**0.246**

Semantics data representations were treated as one data source, namely, “Venue Semantics & Mobility”, since both of them were extracted from Foursquare check-ins data. It also should be noted that we did not evaluate the “BMI Trend” prediction performance on independent sources since there are only two users in the test set, who have contributed data in all modalities and have been labeled with the “BMI Trend” attribute.

First, we examine the contribution of different data sources towards user profile learning and source integration ability. An interesting observation comes from the data source combination results (see Table 4.7), where the combinations “Sensors + Text” and “Sensors + Venue Semantics & Mobility” return the best performance and seem to be the most powerful among other bi-source combinations. More impressive results can be gained from triplet combinations, where the combination “Visual + Sensors + Venue Semantics & Mobility” perform the best. Based on these results, we can answer the research question **RQ1**[13] by concluding that the Sensor data is of crucial importance for individual wellness profile learning since it is the only data source included in all best-performing data source combinations. This observation can also be interpreted by the ability of sensor data to represent users' actual physical condition, which is directly related to users' BMI category and "BMI Trend".

Let's now describe the single-source evaluation results (Table 4.7). It is interesting to note that in the case of learning from independent data sources for the BMI category inference task, our framework trained on Text modality performs the best, while those trained on Sensors and Venue Semantics & Mobility data ranks 2nd and 3rd place, respectively. First of all, the superiority of Text data over other modalities can be explained by its quantitative dominance (see Table 3.1). At the same time, the Sensor data holds the 2nd position, which again highlights its importance. Finally, being trained on visual data, M^2WP performs the worst among all other data sources, which is consistent with current data source combination results. However, it does not conform well with demographic profiling results (Section 4.1.2), where Visual data was essential for Age and Gender inference. One possible explanation of the Visual data poor performance is the high level of noise in users' Instagram photos. At the same time, the differences with the demographic profiling can be interpreted by the generality of ImageNet image concepts [44] that could be useful for the general task of demographic attributes inference [54], but less effective to the more narrow problem of individual wellness attributes inference. The last observation can also be explained by the results obtained in Buraya et al. [28], where visual data was observed to play an auxiliary role in the task of relationship status prediction. In conclusion, we would like to highlight and suggest further usage of Text and Sensor data sources as the strongest contributors towards individual wellness profile learning.

Evaluation Against Baselines

To compare M^2WP framework against state-of-the-art user profiling approaches and answer the research questions (2) and (3), we evaluate it against the following baselines[14]:

- **RandomForest** — strong baseline for the user profile learning [54]. We combined features by applying an early fusion strategy and filled in the missing values by NMF. The number of trees was selected based on 10-fold cross validation and equals to 105 and 25 for the "BMI category" and "BMI Trend" inference, respectively.
- **aMTFL$_2$** [98] — the $\ell_{2,1}$ norm regularized multi-task learning with the least squares

[13] What is the contribution of sensor data towards BMI category and "BMI Trend" inference?

[14] We note that we did not include baselines that require prior knowledge for model guidance [4, 154] due to the unavailability of such knowledge. It also worth noting that the inter-category relationship regularization (Ω) is not applicable to "BMI Trend" classification since it is a binary personal wellness attribute.

Table 4.8: Evaluation of the M^2WP framework trained on all data sources against baselines. Evaluation metrics: Macro Precision, Macro Recall, Macro F_1

Method	BMI category		"BMI Trend"	
	R_{Mac}/P_{Mac}	$F_{1,Mac}$	R_{Mac}/P_{Mac}	$F_{1,Mac}$
MSESHC [54]	0.141/0.145	0.142	0.634/0.655	0.644
Random Forest	0.135/0.226	0.169	0.333/0.863	0.480
iMSF [182]	0.171/0.174	0.172	0.649/0.649	0.649
$aMTFL_2$ [98]	0.162/0.215	0.184	0.700/0.722	0.710
$TweetFit$	0.222/0.202	0.211	0.705/0.732	**0.718**
M²WP	0.221/0.229	**0.225**	Ω is not applicable	
Statistical Analysis				
Weighted Cohen's Kappa	$\kappa_w = 0.234, p < 0.01$		$\kappa_w = 0.368, p < 0.05$	

lost and $\alpha = 0.5$.

- **iMSF** [182] — the sparse $\ell_{2,1}$ norm regularized multi-source multi-task learning, with $\alpha = 0.4$;
- **MSESHC** — weighted ensemble proposed in [54]. The modality weights **s** were learned by Stochastic Hill Climbing (SHC): $\mathbf{s} : \{0.75, 0.2, 0.25, 0.45, 0.2, 0.3, 0.45, 0.2\}$; for venue categories, image concepts, behavioral text, LDA 50 text, sport categories, sensors freq. bins, workout statistics, and mobility features, respectively.
- **TweetFit** [51] — multi-source multi-task learning framework "TweetFit" (equivalent to **M²WP** framework, trained without inter-category relatedness regularization (Equation 4.1)).
- **M²WP** — our proposed framework trained on all data sources (Eq. 4.5), where $\alpha = 0.1$, $\mu = 0.1$, $\kappa_{g,\xi(t,g)} = \begin{cases} 1 & g = 1...G-1; \\ 0 & g = G. \end{cases}$.

The evaluation results are presented in Table 4.8. One can notice that the results achieved by M^2WP are higher as compared to all the baselines (Table 4.8), which enables us to give a **positive answer to research question RQ1**[15]. Specifically, we conclude that it is possible to improve the individual wellness profiling performance by integrating multiple social media sources and sensor data. Moreover, we note that the incorporation of inter-

[15] Is it possible to improve the performance of BMI category and "BMI Trend" inference by fusing multiple social media and sensor data?

category smoothness ($\boldsymbol{\Omega}$) into model further improves the performance, as compared to the model without inter-category smoothness regularization ("*TweetFit*"), which gives **positive response to the research question RQ3**[16]. The possible reason is that M^2WP considers inter-category relatedness for adjacent BMI categories, which constrains the model parameters of the corresponding adjacent BMI categories to be "close" to each other. Consequently, similar users are treated similarly, which allows for increasing the inference performance. At last, M^2WP outperforms other state-of-the-art approaches from the multi-task learning family as well as non-linear baselines, such as Random Forest and MSESHC [54]. This result demonstrates the framework's ability to integrate data from wearable sensors and social media for wellness profile learning.

Statistical Analysis

To evaluate the performance of $\mathbf{M^2WP}$ from the statistical point of view, we utilize κ_w to quantify the level of agreement between two raters: 1) $\mathbf{M^2WP}$ framework and 2) BMI ground truth labels obtained from users' Endomondo page monitoring (see Section 3.1 for details on ground truth collection). For the case of "Fitness Trend" classification (binary classification), the weight matrix $\mathbf{W}$ was assigned as a unit matrix (all elements are 1s). For the task of BMI category classification (multi-class classification), values in $\mathbf{W}$ were assigned according to their squared distance from perfect agreement [27]. We also utilized p-value test (see Section 6.2.2) to evaluate the significance level of the computed κ_w statistic.

Statistical analysis results are presented in Table 4.8. From the Table, it can be seen that the value of κ_w calculated for BMI classification belongs to the "Fair" level of agreement with ground truth labels (see Table 7.2) [91]. The fair agreement can be explained by the skewness of the BMI distribution in our dataset. The obtained $\kappa_w = 0.234$, however, reveals that the agreement between ground truth and prediction results is statistically significant (with $p < 0.01$) and encourages further research in this direction.

According to Landi's agreement categorization schema, in the case of "BMI Trend" classification, the obtained $\kappa_w = 0.368$ lays in the border of between "Fair" and "Moderate" agreement categories [91] with the significance level of $p < 0.05$. Such results demonstrate the high potential of $\mathbf{M^2WP}$ regarding wellness-related applications and suggest its utilization in future works.

[16] Is it possible to improve the performance of BMI category and "BMI Trend" inference by incorporating inter-category relatedness into the learning process?

Error Analysis

From the previous section, it can be seen that M^2WP framework can achieve better prediction performance as compared to different data source combinations and other user profiling baselines. However, its current performance may not be sufficient for its usage in real-world scenario. To understand the reasons behind the low BMI category prediction performance, we present in Table 4.9 the evaluation results of M^2WP framework trained on all data sources with respect to each BMI category. We use Precision, Recall, and the F_1 as evaluation metrics and color different BMI performance levels as follows: green — good performance; blue — reasonable performance; brown — lower performance; red — low performance.

From the Table, it can be seen that two BMI categories (Normal, Pre-Obese) can be more accurately predicted as compared to the rest of categories. The higher performance of these two categories prediction can be attributed by the large number of samples in these categories, which increases classifiers' generalization ability. Accurate detection of Pre-Obese category is of crucial importance in public wellness applications since it helps to find social media users with "obesity risk". This allows us to offer more information support and assistance to such users, before they may be shifted to other obesity groups.

Another interesting observation comes from the BMI categories with comparably "lower performance" (brown color) that all represent intermediate levels of either obesity or thinness. It suggests that the neighboring wellness categories inside one wellness type of overweight/underweight are not always easily separatable in the social media space. This can be explained by the similar online behavior of users with similar wellness type. Such lower prediction performance also inspires us to develop new and better social media-separatable obesity categorization schemes in our future works.

Last but not least, we would like to highlight the comparably good performance of Severe Thinness category and poor performance of Obese III category. Even though both categories are extreme cases, the Severe Thinness category is often common among certain categories of well-trained amateurs and sportsmen [103]. This implies that many users that belong to Severe Thinness category can be easier distinguished from others by the way they perform their workouts. On the contrary, Obese III tends to have very few samples (as users may not want to reveal their severe obesity status to public) and users in such category behave similar to other obese cases (i.e. Obese I, Obese II). They are, thus, more challenging to classify.

Feature Importance Analysis

In Section 4.2.3, we've discussed the contribution of each data source and source combination towards BMI category inference. However, to boost the prediction performance, it is often important to know which particular feature groups are useful for predicting different BMI

Table 4.9: Performance of the M^2WP framework trained on all data sources with respect to each BMI category. Evaluation metrics: Precision, Recall, F_1

BMI Category	% in Dataset	Precision	Recall	F-Measure
Severe Thinness	8%	0.333	0.25	0.286
Moderate Thinness	3%	0.167	0.2	0.182
Mild Thinness	9%	0.222	0.138	0.17
Normal	38%	0.333	0.543	0.413
Pre-Obese	18%	0.371	0.277	0.317
Obese I	13%	0.125	0.176	0.146
Obese II	6%	0.182	0.111	0.138
Obese III	5%	0.095	0.082	0.088

categories. To gain such understanding, in this section, we perform feature importance analysis by eliminating features in two steps:

- **Correlation filtering:** For each feature, we computed its linear correlation with every other feature. We kept the feature with the largest number of correlated features and filtered out all the correlated features. This procedure was repeated until there are no more features with correlation higher than the threshold of $\sigma = 0.2$.
- **Backward feature elimination:** Each feature elimination iteration was performed as follows: first, we trained binary M^2WP[17] for each BMI category on k input features; second, we removed one input feature at a time and trained M^2WP for each BMI category on $k-1$ input features k times; and third, we removed the input feature whose removal has produced the smallest increase in the error rate. The total number of feature elimination iterations was: $D*(D+1)*G$, where $D = 230$ and $G = 8$ are respectively the data dimensionality after correlation filtering and the number of BMI categories.

Table 4.10 describes the usage frequency of different feature types for predicting the BMI categories after the feature elimination procedure. Such frequency helps to gain an insight into the feature importance for predicting each BMI category. In other words, the listed feature types are the minimal subset of features that perform the prediction of every BMI category with the lowest error rate. The most frequently used feature type is colored in red, while the feature types that are used by all the BMI inference models are colored in blue.

[17] where $\alpha = 0.1$, $\mu = 0$, $\kappa_{g,\xi(t,g)} = 0; g = 1...G$;

First, it is noted that *all the proposed feature types were utilized by the model for individual wellness profiling.* This is consistent with our source combination evaluation results and highlights the necessity of using multi-source multi-modal data for the task of individual wellness profiling.

Second, from Table 4.10, it can be seen that frequency domain features (Furrier Transform-based features extracted from sequential workout data) and workout statistics features are of crucial importance in predicting all BMI categories, while the workout semantics features (distribution over exercise types) are weak indicators of one's wellness profile. The first observation conforms well with our initial assumptions and, once again, highlights the *importance of sequential sensor data for wellness profiling.* On the contrary, latter observation suggests that the types of users' workouts are indeed not very useful for BMI prediction, which also limits the usage of exercise tracking platforms that just provide the workout description without the actual sensor measurement data sequences (i.e. MyFitnessPal, Fitocracy) [174, 85, 122]. The above can be explained by the strong relation of sensor measurements (intensity, speed, heart rate) to users' actual health status, which may not be directly related to an exercise type.

Third, we would like to highlight the *importance of temporal, and venue semantics features*, which were widely used by all the prediction models. Such observation is consistent with our source combination analysis results (see Section 4.2.3) as well as correlation analysis results (see Section 6.3). The importance of temporal and venue semantics data was also mentioned in previous studies [116, 54, 105] and can be explained by the major difference between people from different BMI groups in terms of location preferences and time schedules.

Finally, we would like to bring to attention the infrequent usage of text-based features, which could be due to the low representativeness of textual data itself as compared to multi-modal data [28]. Another possible explanation is the high noise level in Twitter [142].

4.3 Summary

In this chapter, we presented two experimental studies on joint individual wellness profile learning from sensor and social media data. To build a comprehensive individual user profile, we first trained an efficient ensemble to predict users' age and gender based on data from three social networks. We then proposed the M^2WP framework to predict users' BMI category and "BMI Trend" based on data from multiple social networks and wearable sensors. The performance of the framework was enhanced by collective data source fusion strategy, the introduction of sparsity into data representations, and incorporation of the novel idea of a multi-source multi-task learning via efficient inter-category smoothness regularization. To demonstrate the advance of our proposed individual wellness profiling solutions over other state-of-the-art baselines as well as gain deeper insight into wellness profiling performance,

Table 4.10: Usage frequency of different feature types for predicting BMI categories

Feature type	S. Th.	M. Th.	Md. Th.	Nrm.	P-Ob.	Ob. I	Ob. II	Ob. III
Latent topics	2	5	0	0	1	2	4	1
Lexicon	4	3	3	1	0	2	1	0
Writing style	2	1	1	1	0	2	1	3
Image con.	25	5	4	1	3	7	9	4
Venue sem.	19	5	1	2	11	5	11	2
Mob. & Tmp.	3	4	3	1	2	4	4	2
Work. sem.	1	0	1	0	1	2	2	2
Freq. domain	15	8	19	10	18	17	25	13
Work. stat.	1	2	2	1	1	2	2	3

we conducted extensive quantitative evaluation, classification error analysis, and feature importance analysis. Our experimental results demonstrate the crucial importance of joint learning from multiple social media and sensor data for individual wellness profiling, which encourages further research in this direction. The major contribution of this chapter is the generic multi-source supervised learning framework that is able to learn from multiple incomplete data sources while performing feature selection and inter-category regularization on-the-fly.

Chapter 5

Integrating Multi-Source Cross-Region Data For Group User Profile Learning In Urban Mobility Domain

In this chapter, we introduce our multi-source group user profiling approach and its Urban Mobility application — venue category recommendation. We then overview the real-world social multimedia analytics platform bBridge, which integrates all earlier proposed user profiling techniques to bring real-time analytics to business and individual users.

5.1 Multi-Source Community Detection for Cross-Network Recommendation

Group profiling aims at learning the collective behavior of users groups that are also known as user communities in social media computing. An intuitive approach to perform user profiling at a group level is to discover communities of users and then apply individual profile learning approach (see Chapter 4) to capture information and behaviors of the community members [185]. In other words, community detection often modeled as a two-stage framework including community discovery problem and aggregation of members' personal profiles. Group user profiling is necessary for various applications including facility planning, marketing [131], and personalized recommendation [185, 52, 53].

As we've mentioned in Section 2.2.3, there are two main approaches for evaluating group profiling algorithms: explicit evaluation and implicit evaluation. Since there is no widely accepted measure to quantify community detection results in the case of multi-source community discovery, in this thesis we evaluate our proposed multi-source community detection approach via its Urban Mobility application — Cross-Source Venue Category Recommendation.

In an attempt to make sense of the overwhelming amount of multi-source data, various recommender systems were designed to extract relevant information from individual users. However, most of them utilize data from a single-source or of a single modality, while disregarding the potential of cross-source multi-modal recommendation. Another aspect which lacks research, and is worth exploring, is the role of personal and group knowledge integration in cross-source multi-modal recommendation. It has been established in an earlier study that effective recommendation can be achieved based on the combination of group and individual knowledge [29]. While individual knowledge is useful for the incorporation of personal experience [163], group knowledge helps to improve the recommendation diversity [79].

Aiming to bridge these two research gaps, in this section, we focus on the problem of **cross-source recommendation** based on **multi-view user community discovery**. Specifically, this chapter is devoted to the emerging topic of **venue category recommendation/prediction** [7, 52, 53, 165], where we recommend a ranked list of Foursquare venue categories to users with accounts in three social networks, namely Twitter, Instagram, and Foursquare [54]. By performing filtering based on venue categories of places around user's current location, our recommender system does not depend on any particular geographical area. This means that it can be used for many Urban Mobility scenarios, such as tourist route planning [112], facility arrangement [161], or interactive mobile assistance [53]. For example, based on one's venue category preferences, we could further recommend a particular venue (i.e. Chinese Restaurant, Movie Theatre, etc.) near his/her current location, rather than a similar place located far away, even if the venue located farther away may slightly better fit his/her preferences. Last but not least, the venue category recommendation (but not venue recommendation) also helps to overcome evaluation difficulty, which often arises due to location datasets' sparsity.

Despite its potential, we recognize the challenges of multi-source group user profiling in Urban Mobility domain. One challenge is data integration. Multiple data sources often describe distinct sides of users' activities. For example, Twitter may casually reveal users' daily life updates, while Instagram may uncover their visual preferences. At the same time, social media content comes in many forms (modalities), such as text, images or videos, which also exhibit different facets of users' online experiences. Such multi-source multi-modal data must be integrated into one community detection framework in a mutually-consistent fashion, which is an open research problem. Another challenge pertains to group representation learning. Group representation can be naturally expressed in the form of user communities, which must be extracted from heterogeneous multi-source data. However, the detection of such user communities from multi-source multi-modal data is not an easy task due to the necessity of proper inter-source and intra-source relationship modeling.

Inspired by previous studies and the challenges above, we seek to address three research questions. First, to support the assumptions behind this study, it is important to answer: **(RQ1) Does incorporation of group user profiling into recommendation boost its performance?** Second, even though the topic has been discussed with respect to some problems, it is still unclear if: **(RQ2) Inter-source relationship information enable us to find better user communities.** Finally, for further recommendation improvement, it is important to understand: **(RQ3) What the contribution is of each data source (modality) towards group user profiling and venue category recommendation.**

To answer these research questions, we introduce a cross-domain recommender framework **$\mathbf{C^3R}$** that automatically extracts and utilizes group user profile and user history to perform **C**ross-Source User **C**ommunity-Based **C**ollaborative venue category **R**ecommendation. Individual knowledge is obtained from user's experience, which is modeled as the distribution among venue categories that a user has visited in past. To incorporate group knowledge, we detected cross-source user communities in a latent space, where the relationship between users is modeled as a multi-layer graph. The community detection approach incorporates inter-source relationships during the process of learning individual source representations and preserves inter-source consistency at the stage of learning latent sources representation. The inter-source relationship graph is computed automatically from the data and further utilized via new graph-constrained regularization. The experimental results show that our framework can achieve better recommendation performance in three geographical regions, as compared to state-of-the-art baselines and different data source combinations.

5.1.1 Data Representation

In addition to the features described in Section 4.1.1, we extracted a 48-dimensional temporal feature and 4 mobility features. It is known that online activity of social-media users is tightly knit to temporal and mobility aspects [116], which makes it reasonable to incorporate such data into community detection process. Intuitively, users with similar mobility patterns (i.e. often co-located) and similar temporal patterns (i.e. often perform activities at similar time intervals) may have similar interests and can form an interests-based community. Below, we give a brief description of the mobility and temporal data features that we used to form mobility and temporal layers of user relationship graph.

The mobility features were computed based on users' *areas of interest (AOIs)* [131], which are geographical regions of user's high geo-location density (regardless the geo-location semantic meaning, which could be a geo-located tweet, geo-located Instagram image, or Foursquare check-in). AOIs were obtained by performing density-based clustering[1] [143] over the geo locations of each user and considering the convex hull of each cluster as a new AOI. ε was computed by analyzing the average distance between neighbors in MinPts-distance graph ($MinPts = 3$ was selected empirically) [47]. User's AOIs represent his/her geographical mobility patterns [116, 131] that can be related to user's lifestyle and interests. We extracted the following **Mobility And Temporal** features:

- **Average number of posts during each of the** 8 **daytime durations**, where each

[1] Most of NUS-MSS's check-ins belong to three geographical regions (Singapore, New York, London), which makes it possible to compute DBScan clusters for most of the NUS-MSS users

time duration is 3 hours long. The temporal features were computed for each data source with respect to weekday/weekend factor. In total, there were $8 \times 3 \times 2 = 48$ temporal data dimensions computed. The temporal features indicate users' temporal online activity and related to their urban mobility [116].

- **Number of areas of interest (AOI)**. This feature reflects the mobility side of user's physical activity. For example, users with the higher number of AOIs may have a physically intense lifestyle. AOIs also represent users' frequent areas of activity (i.e. home, office, university/school) [131] and may indicate how far users are willing to travel on a daily basis.
- **Median size of user's AOIs**[2], which indicates users' mobility inside each AOI. The feature is an indicator of users' traveling habits inside their main activity areas.
- **Normalized number of AOI outliers**. The feature indicates how often users visit places that are not located inside their AOI, which may show how often users deviate from their regular mobility patterns.
- **Median distance between AOIs**. This feature reflects user's mobility at intra-city/inter-city/international level. Specifically, it shows how often and how far users travel between their activity zones and can be useful to infer travel-related user communities.

5.1.2 User Community Detection

Notations: To simplify the reading process, we summarize all defined notations in Table 5.1.

Distance Graph Construction

The first step in finding representative user groups is the modeling of users' relationships in the form of a graph so that dense subgraphs of such graph can be treated as user communities. The graph can be constructed based on: (a) social connections between users (i.e. follower/followee relationship) that are often hidden behind the privacy settings; or (b) user generated content, where similarity between users is estimated as a distance between data representations of users and each data source modeled as a layer in a multi-layer graph. In our work, we adopt the graph construction based on UGC to avoid privacy concerns and limitations in mining explicit user social relations. There are certain benefits of such graph construction approach [49]. First, it does not suffer from the lack of information about users' relationship, since the relations between users are simply modeled as distances between

[2] Where AOI size is defined as the median distance between the center of mass and all points inside AOI.

Table 5.1: Notations summary

Symb.	Description
vec_u	Distribution of the user u among items (venue categories) in u's past
C_u	Community of the user u
γ	Parameter that controls personal aspect of rec-n
θ	Parameter that controls group aspect of rec-n
N	Number of users
M	Number of data sources (graph layers)
L_i	Laplacian matrix of the i-th layer
U_i	Eigendecomposition matrix of the i-th layer
$\hat{L_i}$	Inter-layer relationship regularized Laplacian of the i-th layer
$\hat{U_i}$	Inter-layer relationship regularized eigendecomposition matrix of the i-th layer
$\hat{L_{mod}}$	Sub-space regularized Laplacian matrix
W_R	Adjacency matrix of inter-layer similarity graph
k	Parameter that controls the number of clusters
α	Parameter that controls sub-space regularization
β_i	Parameter that controls inter-layer regularization for the layer i

users based on UGC similarity. Second, it naturally solves the problem of cross-region recommendation, where the condition for related users to be explicitly connected in social networks is relaxed. For every graph node pair i and j, the corresponding distance was computed as *Heat* kernel [18]:

$$W_{i,j} = e^{-\frac{||x_i - x_j||^2}{\sigma}},$$

where $||x_i - x_j||$ is the Euclidean norm, and σ is parameter that could be found by i.e. grid search.

Problem Formulation

One of the commonly used formulations of the community detection problem is its representation in a form of NCut formulation, which conditions the sum of graph edges' weights in each community to be minimized among all communities [150]. This simply means that all communities are formed by users that are most "similar" to each other. Such a definition is naturally applicable to our task of group representation learning based on users' interests and from multiple data sources. We thus adopt the NCut formulation in our study. The NCut definition is given below:

$$NCut(C_1, ..., C_k) = \sum_{i=1}^{k} \frac{W(C_i, \overline{C}_i)}{vol(C_i)} = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C}_i)}{vol(C_i)},$$

where $vol(C_i)$ is the sum of weights of all edges attached to vertices in C_i.

However, the NCut problem is proven to be $\mathcal{NP}$-hard [170]. Fortunately, there exists a state-of-the-art approximation that is defined as a standard trace minimization (also known as *spectral clustering*) [166]:

$$\min_{U \in R^{n \times k}} \operatorname{tr}(U^\intercal L_{sym} U), \ s.t. \ U^\intercal U = I. \tag{5.1}$$

By the Rayleigh-Ritz theorem, the solution of the problem in Equation (5.1) is given by the first k eigenvectors of the normalized graph Laplacian $L_{sym} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where W is adjacency matrix, and D is degree matrix [102]. The mapping of each user to a cluster can be further obtained by i.e. K-Means clustering over the eigenvector space [166].

Spectral Slustering on Multi-Layer Graph

The well-known disadvantages of early fusion and late fusion data aggregation strategies [4, 154] encourage us to perform joint clustering from all social data sources (graph layers) simultaneously. The final data representation (latent representation) must be consistent with all graph layers so that clustering will not be biased towards individual data sources/modalities. One way to keep the latent representation consistent to all layers of the multi-layer graph is to apply regularization during the clustering process, which will keep graph layers "close" to the target representation. A well-adopted approach for measuring "closeness" between target latent space U and the space of each layer U_i is based on Grassman Manifolds [76], where the projected distance between two spaces S_1 and S_2 on Grassman Manifold is equal to [45]:

$$\begin{aligned} d^2_{Proj}(S_1, S_2) &= \sum_{i=1}^{k} \sin^2 \theta_i \\ &= k - \sum_{i=1}^{k} \cos^2 \theta_i \\ &= k - tr(S_1 {S_1}^\intercal S_2 {S_2}^\intercal) \\ &= \frac{1}{2}(tr(S_1 {S_1}^\intercal) + tr(S_2 {S_2}^\intercal) - 2tr(S_1 {S_1}^\intercal S_2 {S_2}^\intercal)) \\ &= \frac{1}{2}||S_1 {S_1}^\intercal - S_2 {S_2}^\intercal||^2_F, \end{aligned} \tag{5.2}$$

where $||A||_F$ is the Frobenius norm [69] of the matrix A.

Since mappings $S_1 {S_1}^\intercal$ and $S_2 S_2^\intercal$ preserve distinctness [45], we consider the projection distance as a distance measure between subspaces. The distance between target subspace S and all the other individual subspaces $\{S_i\}_{i=1}^M$ can thus be defined as the sum of squared projection distances between S and all individual subspaces $\{S_i\}_{i=1}^M$ [45]:

$$\begin{aligned} d^2_{Proj}(S, \{S_i\}_{i=1}^M) &= \sum_{i=1}^{M} d^2_{Proj}(S, S_i) \\ &= kM - \sum_{i=1}^{M} \mathrm{tr}(SS^\intercal S_i {S_i}^\intercal). \end{aligned} \tag{5.3}$$

By utilizing the Equation (5.3), the Equation (5.1) can be extended to introduce the subspace-regularized objective:

$$\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) + \alpha(kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i{}^\intercal)), \quad s.t.\ U^\intercal\ U = I, \tag{5.4}$$

where α controls sub-space regularization. We can rearrange Equation (5.3) to present it in a form of standard trace minimization problem:

$$\begin{aligned}
&\min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) + \alpha(kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i{}^\intercal)) \\
&= \min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) - \alpha \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal U_i U_i{}^\intercal) \\
&= \min_{U \in \mathbb{R}^{n \times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal L_i U) - \alpha \sum_{i=1}^{M} \operatorname{tr}(U^\intercal U_i U_i{}^\intercal U) \\
&= \min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr} \sum_{i=1}^{M} (U^\intercal L_i U - \alpha U^\intercal U_i U_i{}^\intercal U) \\
&= \min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr}(U^\intercal \sum_{i=1}^{M} (L_i U - \alpha U_i U_i{}^\intercal U)) \\
&= \min_{U \in \mathbb{R}^{n \times k}} \operatorname{tr}(U^\intercal \sum_{i=1}^{M} (L_i - \alpha U_i U_i{}^\intercal) U),
\end{aligned} \tag{5.5}$$

which, by the Rayleigh-Ritz theorem [102], can be solved as the first k eigenvectors of the modified Laplacian $L_{mod} = \sum_{i=1}^{M} (L_i - \alpha U_i U_i{}^\intercal)$. The eigen decomposition of the matrix L_{mod} is not in the scope of this thesis, but can be efficiently computed by well-known algorithms [94].

Incorporating Inter-Layer Relationship

Even though the subspace-regularized spectral clustering approach performed well on synthetic dataset [45], real world problems often require a consideration of inter-source relationship (similarity) during the learning process [154]. Such a requirement is caused by the difference between data sources and the way they describe the users [53]. Specifically, for particular applications some data sources may be more informative than others. For example, Foursquare check-ins could be more useful for venue category recommendation than textual posts, while text data is of crucial importance for demographic profiling (see Section 4.1.2).

Inspired by previous works [53, 154, 172] and based on our observations, we take a step further in multi-source clustering by introducing a new model regularization schema that guides the clustering process based on predefined inter-layer relationship graph. Given the multi-layer user relationship graph G, which is constructed as described in previous sections, let's assume that there exists a complete undirected inter-layer similarity graph R with the adjacency matrix W_R, where each value of the matrix $w_{i,j}$ represents the similarity between i-th and j-th layer of G. While W_R can be automatically computed from the data (see Section 5.1.2 for details), we first describe our inter-layer regularized spectral clustering approach.

Essentially, our goal is to regularize the conventional spectral clustering in such a way that the representation of each layer $\hat{U_i}$ would be computed with respect to the inter-layer similarities $w_{i,j}$, which are taken from adjacency matrix of the inter-layer relationship graph R. Let's also note that the approach in Equation (5.5) relies on pre-computed layers' Laplacians $\{L_i\}_{i=1}^{M}$ and their spectral spaces $\{U_i\}_{i=1}^{M}$. In our work, we apply the inter-layer regularization on the process of computing the layers' spectral spaces, so that the final multi-layer spectral space is computed as in Equation (5.5). By using previously defined distance on Grassman manifold, we define the new objective function for the i-th layer as follows:

$$\begin{aligned}\min_{\hat{U_i}\in\mathbb{R}^{n\times k}} \operatorname{tr}(\hat{U_i}^{\intercal} L_i \hat{U_i}) + \beta_i(kM - \sum_{j=1, j\neq i}^{M} w_{i,j} \operatorname{tr}(\hat{U_i}\hat{U_i}^{\intercal} U_j U_j^{\intercal})),\\ s.t.\ \hat{U_i}^{\intercal}\hat{U_i} = I,\end{aligned} \tag{5.6}$$

where $\hat{U_i}^{\intercal}$ is the new spectral space of the i-th layer, β_i — parameter that controls inter-layer regularization for the layer i, $\{U_j\}_{j=1}^{M}$ — spectral spaces of all layers after standard spectral clustering, $w_{i,j}$ — similarity between layer i and layer j. After similar rearranging procedure as in Equation (5.5), the problem in Equation (5.6) can be presented as a standard trace minimization:

$$\begin{aligned}&\min_{\hat{U_i}\in\mathbb{R}^{n\times k}} \operatorname{tr}(\hat{U_i}^{\intercal} L_i \hat{U_i}) + \beta_i(kM - \sum_{j=1, j\neq i}^{M} w_{i,j} \operatorname{tr}(\hat{U_i}\hat{U_i}^{\intercal} U_j U_j^{\intercal}))\\ &= \min_{\hat{U_i}\in\mathbb{R}^{n\times k}} \operatorname{tr}(\hat{U_i}^{\intercal}(L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j} U_j U_j^{\intercal})\hat{U_i}),\end{aligned}$$

thus, by the Rayleigh-Ritz theorem, it can be solved as the first k eigenvectors of the regularized Laplacian $\hat{L_i}$:

$$\hat{L_i} := L_i - \beta_i \sum_{j=1, j\neq i}^{M} w_{i,j} U_j U_j^{\intercal}.$$

We now presented all necessary components to define our multi-source user community detection approach with the following objective function:

$$\begin{aligned}\min_{U\in\mathbb{R}^{n\times k}} \sum_{i=1}^{M} \operatorname{tr}(U^\intercal \hat{L}_i U) + \alpha(kM - \sum_{i=1}^{M} \operatorname{tr}(UU^\intercal \hat{U}_i \hat{U}_i{}^\intercal)) \\ = \min_{U\in\mathbb{R}^{n\times k}} \operatorname{tr}(U^\intercal \sum_{i=1}^{M} (\hat{L}_i - \alpha \hat{U}_i \hat{U}_i{}^\intercal) U), \\ s.t.\ U^\intercal\ U = I.\end{aligned} \tag{5.7}$$

To make the clustering procedure clear, we present the pseudocode as shown in Algorithm 1.

From the pseudocode, it can be seen that optimization of the Equation (5.7) and further clustering consists of four main steps: 1) Perform conventional spectral clustering on each layer to obtain L_i and U_i; 2) By incorporating inter-layer relationship graph R, perform inter-layer relationship regularized spectral clustering on each layer to obtain $\hat{L_i}$ and $\hat{U_i}$; 3) Execute subspace-regularized spectral clustering on each layer to obtain $\hat{L_{mod}}$ and U; 4) Normalize U to obtain U_{norm} and execute the X-Means Clustering over it [124][3].

The value of the subspace regularization parameter α and the inter-layer regularization parameters β_i can be found by grid search. In next section, we outline the construction of inter-layer similarity graph G.

Computing Inter-Layer Relationship

Intuitively, the inter-layer relationship graph R must represent the similarity between layers in terms of clustering results, summarized for different values of k. We thus define the similarity $sim(w, q)$ between layers q and w as a normalized difference between $N \times N$ clustering co-occurrence matrices $M_{q,k}$, $M_{w,k}$, in which each value $m_{i,j}$ is equal to 1 if user i is assigned to the same cluster as user j in both layers w and q, and 0 otherwise. The clustering co-occurrence matrices are obtained by performing single-layer spectral clustering on each layer i (U_i space) and for different values of k ($k = 2..K$, where $K = \sqrt{N}$ [72]). We then take an average among relation values obtained for different k:

$$sim(w, q) = \left(\sum_{k=2}^{K} 1 - \frac{||M_{w,k} - M_{q,k}||}{\sqrt{N(N-1)}} \right)\ /\ K - 1. \tag{5.8}$$

[3] X-Means clustering was selected due to its similarity to K-Means [72] clustering and the ability of automatic inference of the number of clusters to detect

Algorithm 1 C^3R clustering

```
procedure CLUSTER($\{W_i\}_{i=1}^{M}$, $\{\beta_i\}_{i=1}^{M}$, $W_R$, $k$, $\alpha$)
▷ $\{W_i\}_{i=1}^{M}$ — $N \times N$ weighted adjacency matrices of individual graph layers $\{G_i\}_{i=1}^{M}$, $W_R$ — adjacency matrix of inter-layer similarity graph, $k$ — target number of clusters, $\{\beta_i\}_{i=1}^{M}$ — regularization parameters for each layer, $\alpha$ — parameter for subspace regularization

    for each layer $i$
        Compute $L_i$ and $U_i$ for $G_i$ [166]
▷ $L_i$ — the normalized Laplacian matrix of the layer $i$, $U_i$ — subspace representation of the layer $i$, $G_i$ — $ith$ layer graph
        Compute $\hat{L_i} := L_i - \beta_i \sum_{j=1, j \neq i}^{M} w_{i,j} U_j U_j^{\intercal}$
▷ $\hat{L_i}$ — the regularized Laplacian matrix of $ith$ layer
        Compute $\hat{U_i} \in \mathbb{R}^{N \times k}$
▷ $\hat{U_i}$ — the regularized subspace representation of the layer $i$ that is the matrix of first $k$ eigenvectors $\hat{u_1}, ..., \hat{u_k}$ of $\hat{L_i}$[94]
    end for
    Compute $\hat{L_{mod}} = \sum_{i=1}^{M} (\hat{L_i} - \alpha \hat{U_i} \hat{U_i}^{\intercal})$
▷ $\hat{L_{mod}}$ — the modified Laplacian matrix [45]
    Compute $U \in \mathbb{R}^{N \times k}$. Normalize rows of $U$ to get $U_{norm}$.
▷ $U$ — the matrix of first k eigenvectors [94] of $\hat{L_{mod}}$
    $y_j \leftarrow j$-th row of $U_{norm}$.                ▷ $y_j \in \mathbb{R}^k (j = 1, ..., N)$
    Cluster $y_j$ in $\mathbb{R}^k$ into $C_1, ..., C_k$ using X-Means [124].

    return $\{C\}_{i=1}^{k}$
▷ $C_1, ..., C_k$ cluster assignment
end procedure
```

The above formulation is the modified and normalized version of the Partition Difference measurement [95]. Being averaged over different values of k, it is able to serve as a reliable indicator of the similarity between different social networks in terms of clustering results. We explicitly would like to mention that our-proposed inter-layer relationship graph construction approach is purely automated and does not require any expert knowledge. This suggests its further usage for other graph-constraint unsupervised learning approaches.

5.1.3 Computational Time Complexity Analysis

To analyze the complexity of C^3R clustering, we need to estimate the complexity of each step of Algorithm 1. If N is the number of users, M the number of data sources (graph layers), and k is the number of first eigenvectors to compute, C^3R time complexity can be estimated as $\mathcal{O}(N^2(M^2k + MN + k^2))$. Below, we discuss the complexity in more details.

First, each Laplacian (L_i) and eigenvector matrix (U_i) computational complexity is $\mathcal{O}(N^3)$, which sums up to $\mathcal{O}(MN^3)$ for computing them for all graph layers. The computation of each $\hat{L_i}$ costs $\mathcal{O}(MN^2k)$, which gives the computational complexity of $\mathcal{O}(M^2N^2k)$. The total cost of $\hat{U_i}$ computation is $\mathcal{O}(MN^3)$. $\hat{L_{mod}}$ can be computed in $\mathcal{O}(MN^2k)$ time. The computation of matrix U takes $\mathcal{O}(N^3)$ time, while the complexity of X-Means clustering in space U_{norm} is $\mathcal{O}(N^2k^2)$ [124]. The total C^3R time complexity, thus, is $\mathcal{O}(M^2N^2k) + \mathcal{O}(MN^3) + \mathcal{O}(N^2k^2) = \mathcal{O}(N^2(M^2k + MN + k^2))$, where $M, k \ll N$.

5.1.4 Evaluation

Cross-Source Recommendation as Application of Group User Profiling

Strategic decision making is known to be influenced by external factors like personal experiences and public opinions [8]. The public opinion can be expressed by explicit and implicit user communities that are formed based on social relations and their data similarity, respectively. This phenomenon can be leveraged to enhance recommendation performance. To do so, we perform venue category recommendation based on both personal and group knowledge, which naturally models the impact of society on an individual's behavior during the selection of a new place to go. Formally, our proposed C^3R recommendation approach is defined as follows:

$$rec(u) = sort\left(\gamma \cdot vec_u + \theta \frac{\sum_{v \in C_u} vec_v}{|C_u|}\right) \tag{5.9}$$

where vec_u is the distribution of the user u among items (venue categories) in u's past, and $\frac{\sum_{v \in C_u} vec_v}{|C_u|}$ is the normalized (by the number of members in user community C_u) distribution of all community members among venue categories, γ controls the personal aspect of recommendation, while θ regulates the group experience impact of the user community C_u. Besides the benefits described earlier, incorporation of user communities reduces the search space during the recommendation process and provides better candidates to compare for further collaborative filtering [144]. While personal information like users' previous activity is often available, in most of the cases user communities are not explicitly indicated, which can be solved by our proposed multi-source community detection approach C^3R.

Evaluation Measures

To evaluate our framework against competing systems, we chose the following two widely accepted measures:

Normalized Discounted Cumulative Gain ($NDCG$) measure, which is defined as:

$$NDCG@p = \frac{DCG@p}{IDCG@p}, DCG@p = \sum_{i=1}^{p} \frac{2^{rel_i}}{\log_2(i+1)}, rel_i = \frac{Cat_i}{N_{Cat}},$$

where $IDCG$ is the maximum possible (**I**deal) DCG for a given set of queries, rel_i is the graded relevance of the result at position i, Cat_i is number of times user checked-in at venue of category i, and $N_{Cat} = 764$ is the total number of Foursquare venue categories in the dataset.

Average Precision (AP), which is defined as:

$$AP@p = \frac{1}{\sum_{i=1}^{p} r_i} \sum_{i=1}^{p} r_i \left(\frac{\sum_{j=1}^{i} r_j}{i} \right), r_i = \begin{cases} 1, & \text{item } i \text{ is relevant} \\ 0, & \text{otherwise.} \end{cases}$$

Baselines Description

In this section, we briefly describe competing recommendation approaches, and different C^3R modifications.

Community Detection Baselines:
In four $\mathbf{C^3R}$ modifications below, recommendation was performed according to Equation 5.9, and community detection was performed over all 1952 features.

$\mathbf{C^3R}$ — our proposed C^3R recommendation approach (Equation 5.9), where user communities are detected as minimization of Equation (5.7) via the Algorithm 1 from multi-layer user relationship distance graph graph ($\sigma_{tw} = 2.01, \sigma_{fsq} = 1.62, \sigma_{inst} = 0.28, \sigma_{tmp} = 1.00, \sigma_{mob} = 0.66$). The estimated parameters are: $k = 10$, $\alpha = 0.922$, $\beta_i = 1$ $(i = 1..M)$, $\gamma = 1$, $\theta = 0.248$, which are inferred by SHCR search [54]. The adjacency matrix W_R of the automatically constructed inter-layer similarity graph is given below:

$$W_R = \left(\begin{array}{c|ccccc} & \textbf{txt} & \textbf{loc} & \textbf{vis} & \textbf{tmp} & \textbf{mob} \\ \hline \textbf{txt} & 1 & 0.632 & 0.621 & 0.643 & 0.561 \\ \textbf{loc} & 0.632 & 1 & 0.614 & 0.631 & 0.570 \\ \textbf{vis} & 0.621 & 0.614 & 1 & 0.621 & 0.551 \\ \textbf{tmp} & 0.643 & 0.631 & 0.621 & 1 & 0.560 \\ \textbf{mob} & 0.561 & 0.570 & 0.551 & 0.560 & 1 \end{array}\right)$$

$\mathbf{C^3R_{-\hat{L}}}$ — C^3R recommendation without inter-layer regularization ($\beta_i = 0, i = 1..M$), where $k = 10$, $\alpha = 0.521$, $\gamma = 1$, $\theta = 0.1$ are obtained by SHCR [54].

$\mathbf{C^3R_{-\hat{L}-L_{mod}}}$ — C^3R recommendation without inter-layer regularization ($\beta_i = 0, i = 1..M$) and sub-space regularization ($\alpha = 0$), where $k = 22$, $\gamma = 1$, $\theta = 0.147$ are obtained by SHCR [54].

$\mathbf{C^3R_{-Comm}}$ — C^3R recommendation without user community detection (all users are considered to be in the same community).

$\mathbf{C^3R}$ (DBScan) — C^3R recommendation, where user communities are detected by Density-Based clustering (DBScan) [143]. The DBScan's $\varepsilon = 0.9$ and $MinPts = 6$ were obtained by grid search.

$\mathbf{C^3R}$ (X-Means) — C^3R recommendation, where user communities are detected by X-Means Clustering [124] ($k = 4$ was auto-detected by X-Means).

$\mathbf{C^3R}$ (Hierarchical) — C^3R recommendation, where user communities are detected by Hierarchical Clustering. Parameter $k = 10$ and "single linkage" inter-cluster distance measure were obtained by grid search.

Recommender System Baselines:

Popular (POP) — performs recommendation only based on user's past experience (distribution on user's check-ins among 764 Location (Venue Category) features in past). To note, it is the special case of C^3R recommendation (Equation 5.9), where $\theta = 0$.

Popular$_{All}$ (POP$_{All}$) — performs recommendation based on aggregated experience of all users, which produces user-independent recommendation output (764 Venue Category features were utilized).

Multi-Source Re-Ranking (MSRR) [53] — linearly combines recommendation results from all data sources and determines source weights via Stochastic Hill Climbing With Random Restart (SHCR) (See Farseev et al. [54] for details): $w_{tw} = 0.42,\ w_{fsq} = 0.95,\ w_{isnt} = 0.36,\ w_{temp} = 0.65,\ w_{mob} = 0.02$.

Nearest Neighbor Collaborative Filtering (CF) [3] — produces recommendation based on top $k = 20$ (determined by SHCR) most similar users from the recommendation source (Foursquare, in our case) using a similarity measure (in our case, *Cosine* similarity) over 764 Location (Venue Category) features.

Early Fusion (EF) [176] — fuses multi-source data into a single feature vector (of dimension 1952) and performs recommendation via CF.

Implicit Feedback-Enhanced Singular Value Decomposition (SVD++) [90] — makes use of the "implicit feedback" information in factorization model (trained on 764 Location (Venue Category) features), where $\lambda = 0.67,\ \gamma = 0.06,\ k = 277,\ iter = 55$ are obtained by SHCR [54].

Factorization Machines (FM) [138] — brings together the advantages of different factorization-based models via efficient regularization. In our study, we trained FM based on 764 Location (Venue Category) features, utilized the *MCMC* optimization technique [138] and the regression loss, where $k = 30,\ iter = 140$ are obtained by SHCR [54].

Comparing with Community Detection Baselines

To answer **RQ2**[4], we compare different C^3R modifications, which include different variants of our proposed community detection approach as well as other state-of-the-art clustering approaches. The evaluation results are presented in Figure 5.1.

First, we note that the non-regularized version of the framework ($C^3R_{-\hat{L}-L_{mod}}$) performs the worst against its relatives and other clustering algorithms. This is due to the inability of such non-regularized clustering to build a consistent latent representation, which does not consider inter-network relationship and leads to poor recommendation performance. At the same time, the inter-source regularized C^3R framework beats all the baselines in all three cities, which could not be achieved by its non-regularized versions ($C^3R_{-\hat{L}-L_{mod}}$, $C^3R_{-\hat{L}}$, C^3R_{-Comm}). The above observation suggests the importance of subspace regularization in combination with inter-source regularization for balanced user-community detection and, as a result, better recommendation performance. It also allows us to give a **positive response to RQ2**.

[4] Does inter-source relationship information enable us to find better user communities?

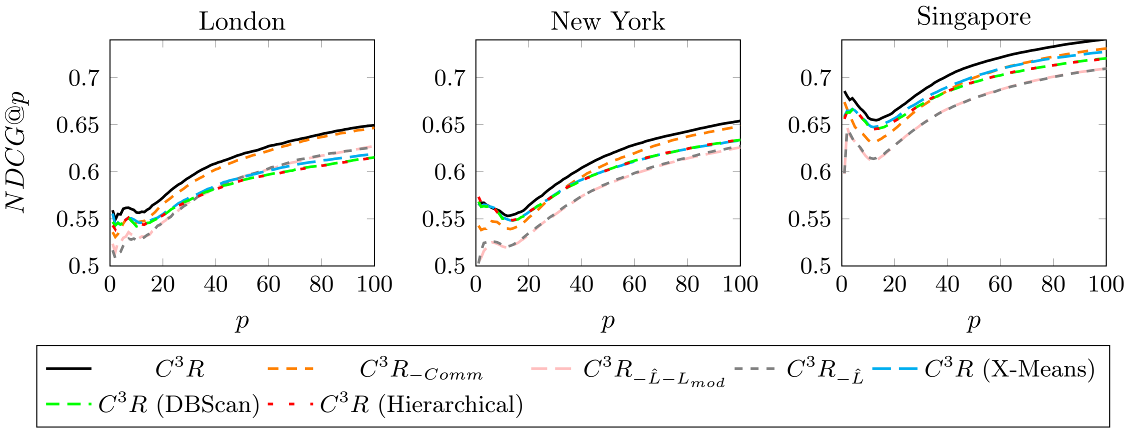


Figure 5.1: Evaluation of C^3R framework based on different clustering algorithms in terms of $NDCG@p$

Let us also outline the good performance of other clustering approaches, where the recommendation based on X-Means clustering (C^3R (X-Means)) performs slightly better than other clustering baselines. The above is consistent with the previous study [52] and can be explained by algorithm's ability to automatically infer the parameter k (number of clusters), which allows for finding more distinctive (regarding Bayesian Information Criterion) user communities and, eventually, provide more relevant recommendation.

Comparing Data Source Combinations

To gain insight into the role of different data sources (modalities) in venue category recommendation and answer **RQ3**[5], we evaluate our framework trained on all data sources (modalities) against different data source (modality) combinations. Due to space constraints, in Figure 5.2 we do not present all feature combinations (31 different combinations), but only show the combinations that include Temporal and Mobility features. For example, the combination "Text" includes Textual, Temporal, and Mobility features, while the combination "Text + Location (Venue Category)" includes Text, Location (Venue Category), Temporal, and Mobility features.

It is not surprising that the recommendation purely based on venue category distribution performs the best among all single-source baselines in all three cities [53]. The reason is that the data from the target source (Foursquare) contains explicit knowledge about the distribution of recommendation items in the training set, and thus can forecast the distribution of items in the test set accurately. Bi-source combination results also provide for

[5] What the contribution is of each data source (modality) towards group user profiling and venue category recommendation?

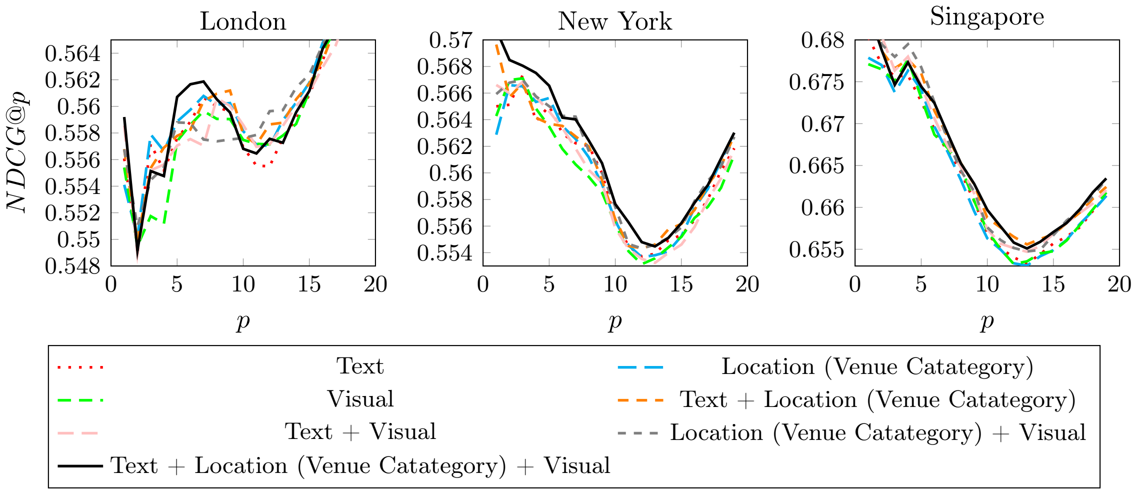


Figure 5.2: Recommendation performance of C^3R framework based on different data source combinations in terms of $NDCG@p$

interesting observations. For example, the Location (Venue Category) + Text combination achieves the best recommendation performance in London and New York regions, but not in Singapore (where text from Twitter gives way to Instagram images). The possible reason is the differences in daily social media usage patterns in different geographical regions: London and New York users mainly post interest-related messages on Twitter, while Singapore users (where Twitter is not widely used) upload pictures of their interests on Instagram (i.e. pictures of food [105]). The above is also supported by the previous study [54], where image data plays a crucial role for the task of multi-source demographic profiling in Singapore. Lastly, we also notice that the combination of all data sources performs the best for all three geographical regions, where the maximum recommendation performance is achieved in Singapore. This could be possibly because of the fact that Singapore has the largest amount of available multi-source data, in comparison with New York and London [54]. Summarizing the above, we **answer the RQ3 by highlighting the importance of venue category, temporal and location-based data** as a major contributor towards recommendation performance. At the same time, depending on geographical region and users' posting behavior, **visual data and textual data may impact differently**, while the **combination of all data sources allows for achieving the best recommendation performance** in most of the cases.

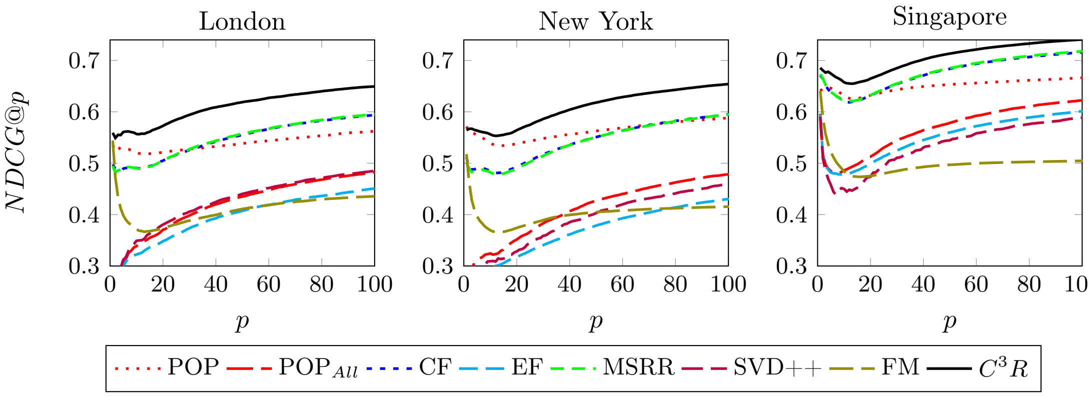


Figure 5.3: Evaluation of C^3R framework against other recommendation approaches in terms of $NDCG@p$

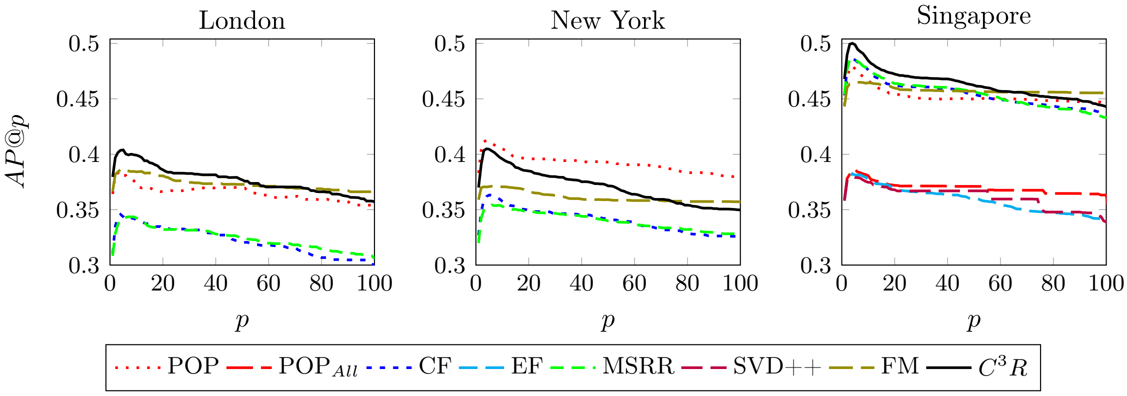


Figure 5.4: Evaluation of C^3R framework against other recommendation approaches in terms of $AP@p$

Comparing with Recommender Systems Baselines

To answer **RQ1**[6], we evaluate C^3R framework against other recommender systems. The evaluation results are presented in Figure 5.3 and Figure 5.4.

First, an interesting observation is the poor performance of the EF approach, regarding both AP and $NDCG$ evaluation metrics, which, once again [53], suggests the necessity of using proper data fusion approaches for multi-source learning. On the other hand, the recommendation approach based on user's past items distribution (POP) produces comparably good recommendation output, especially for the small values of p. It means that Foursquare users tend to re-visit venue categories that they have already visited in past, which highlights the importance of individual knowledge for venue category recommendation

[6] Does incorporation of group user profiling into recommendation boost its performance?

and allows POP for achieving high recommendation performance. At the same time, the POP$_{All}$ baseline (non-personalized popularity-based recommendation) outperforms FM and EF approaches concerning $NDCG$ metric, which suggests the usefulness of group knowledge for recommendation purposes. Finally, it can be seen that the combination of individual and group knowledge in C^3R significantly outperforms other baselines in terms of both AP and $NDCG$ evaluation metrics. This confidently confirms the necessity of group knowledge for cross-source recommendation and **positively answers the RQ1**.

Let us also highlight the weak performance of the factorization-based models. For example, even after proper negative sampling [75], FM is not able to perform relevant recommendation in the head of the recommendation list, which explains its poor performance in terms of $NDCG$ measure. At the same time, SVD++ fails to outperform others in terms of $Precision$, which could be explained by the limited applicability of the "implicit feedback" concept to the task of venue category recommendation. Specifically, the fact that some users did not visit a venue of a particular category does not necessarily mean that they do not appreciate such venue type. Alternatively, such behavior could be a result of the geographical [116] or social [105] constraints. Another interesting observation is the advance of popularity-based baseline (POP) over other recommendation approaches in New York region (regarding AP). The possible explanation is that Foursquare users often re-visit venues of the same category, which means that their corresponding training and test sets significantly overlap in many cases (regarding item distributions). The last is not typical for recommendation systems and may lead to the sub-optimal performance.

On User Community Profiling

The key idea of our cross-source recommendation in Urban Mobility domain is the incorporation of group knowledge into recommendation via detecting relevant user communities from multiple social multimedia data sources. The evaluation against the baselines indirectly shows the ability of C^3R framework to detect important user communities. To support our answer to **RQ2** and demonstrate the community detection performance explicitly, in Table 5.2, we have listed the profiles of the 3 largest user communities detected in Singapore. User profiles are constructed as a "bag-of-words" over textual, visual, and location data modalities. From the table, it can be seen that most popular data representations ("words") in all data modalities are consistent with each other and represent distinct user communities. For example, the user community "Com1" is represented by words: "device", "launcher", "android"; visual concepts: "mouse", "digital clock", "hard disk"; and venue categories: "electronics store", "startup", "technology building". We thus named this community as "Gadgets". The distinctness of detected user communities allows C^3R to perform collaborative filtering based on those social media users, who are semantically "close" to the user of the recommendation system, which helps to boost the recommendation performance. The inter-modal consistency of the community profiles shows that C^3R can perform balanced clustering on a multi-layer

Table 5.2: User communities profiles

Community	**Bag of Words for different modalities**		
	Text	**Visual**	**Location**
(Com1) 'Gadgets" **832 users**	device, launcher, android	mouse, digital clock, hard disc	electronics store, tech startup, technology building
(Com2) 'Arts" **538 users**	painting, landscape, reflection	obelisk, paintbrush, pencil box	arts & crafts store, arts & entertainment, museum
(Com3) "Food" **446 users**	dining, coffee, cooking	pineapple, microwave, frying pan	italian restaurant, pizzeria, macanese restaurant

graph, which is achieved via inter-layer regularization. The overall results suggest the applicability of the proposed techniques for a task of multi-source user community detection and encourage further research along these directions.

As we mentioned earlier, one of the crucial steps in our frameworks' pipeline is the integration of the results obtained from the individual and group user profiling stages into social multimedia analytics platform. The step includes deployment of data gathering approaches together with user profiling models into a cloud environment, as well as their automatization and presentation to consumers. To accomplish this task, in next section we introduce our developed cloud-based social multimedia analytics platform called "bBridge".

5.2 Real-World Application: bBridge Analytics Platform

The increasing amounts of published UGC provides opportunities as well as challenges for its consumers — users and data analysts. At the *user level*, the biggest problem is *information filtering*, due to the growing amount of irrelevant content, such as advertisement and spam. At the data analytics level, *data-preprocessing, fusion, and modeling* pose the greatest challenges to address. A comprehensive solution of these problems provides users with meaningful social-multimedia insights at both the individual and group levels [49]. For example, the group-level analytics can be utilized in Facilities Planning, Brand Monitoring, Marketing, and Geo-Enabled Advertisement domains, while the personal analytics, can be used for Personal Content Filtering and Interest-Based Recommendation.

Previously, we proposed the approaches to venue category recommendation via user community detection (see Section 5.1), and multi-source individual user profile learning (see

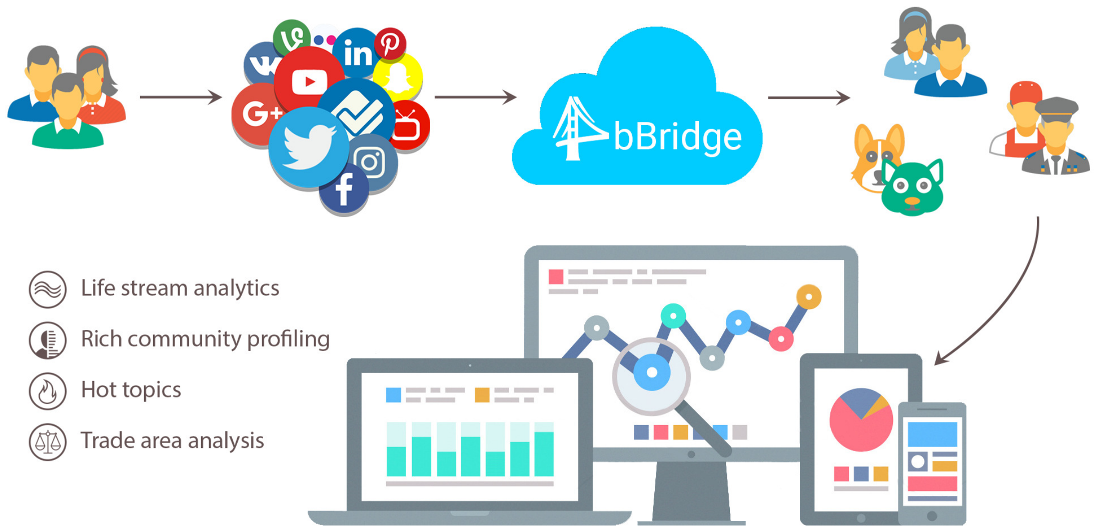


Figure 5.5: Block diagram of the bBridge platform

Chapter 4). The adopted group user profiling technique can be directly applied to live multi-source data. To perform user community profiling (automatic assignment of names and descriptions to user communities), we exploited our earlier developed multi-source individual user profiling [54] and the multi-modal topic modeling [167] techniques. Eventually, all the above approaches were integrated into the Big Social Multimedia Analytics Platform bBridge, which delivers the live stream analytics [83] to its users.

The overall data processing pipeline in bBridge is as follows: (a) the distributed cloud crawlers harvest recently posted UGC from multiple social networks [49]; (b) the community detection framework extracts user communities in a latent cross-social network space [45, 52]; (c) multi-source user profile learning [54] and multi-modal topic modeling [167] techniques construct comprehensive profiles of each detected user community and prepare it for further visualization; and (d) the stream analytics engine aggregates communities' multimedia UGC streams [11], which enables various T-SQL queries to be performed into live multimedia data [13, 114]. To the best of our knowledge, the bBridge is one of the first online solutions that offers Big Social Multimedia Analytics at both personal and group levels.

5.2.1 Platform Description

We now briefly describe the bBridge platform. The block diagram of the bBridge platform is shown in Figure 5.5.

In the platform, the user communities and its' corresponding profiles are re-computed every 12 hours, while the stream analytics and multi-source crawling are executed continuously. At any given time frame, the following steps are taken (the continuously executing operations are marked with the "*"):

I. Multi-Source Distributed Crawling*: The crawling process consists of two steps: (a) a set of active users who have recently posted tweets through the cross-linking functionality [54] of multimedia social networks is collected; and (b) Twitter Stream APIs [164] are used to perform tweets streaming with respect to specified geographical regions, so that cross-social network account mapping can be performed when users publish tweets that contain posts from other social networks (i.e. publish Instagram images on Twitter). By following the described user identification approach, we continuously harvest the data from Twitter, Instagram, Foursquare, and Facebook. At the current stage[7], we continuously monitor about 170 thousand Twitter users from Singapore, which were active (post cross-linking tweets) near Singapore region during the recent one month time frame.

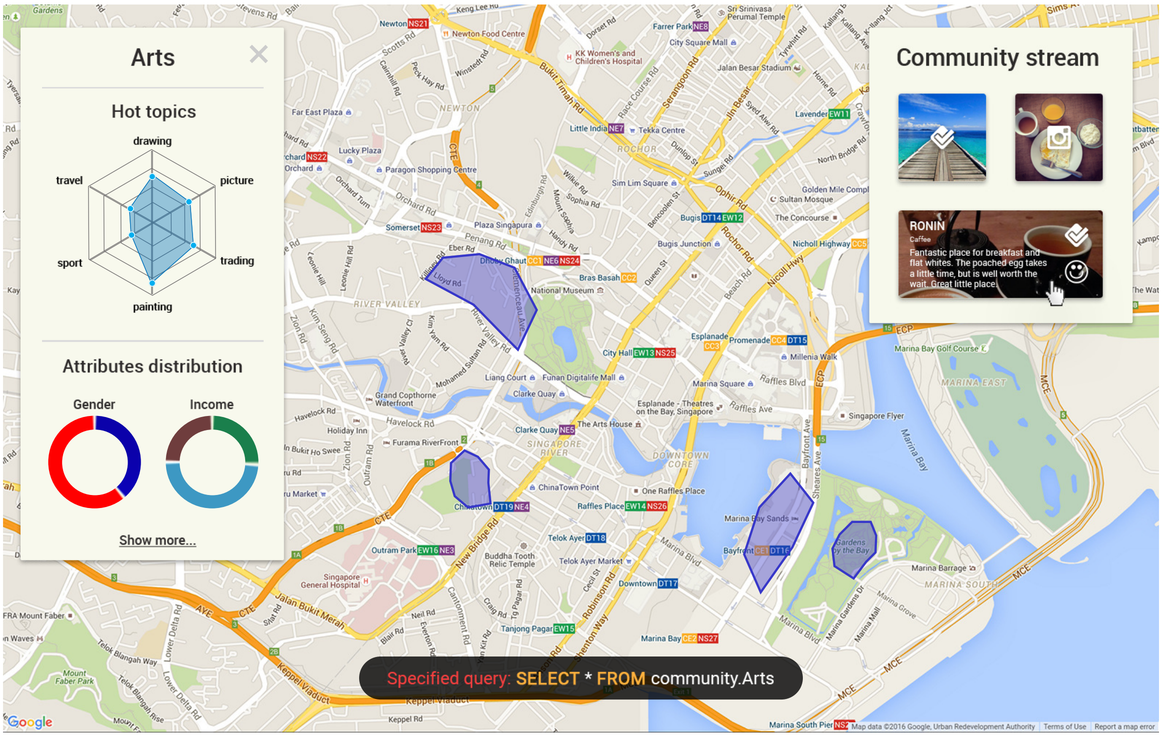
II. Multi-Modal Feature Extraction: We represented data as follows: visual data (Instagram and Twitter images) is represented in a form of image concept distribution [44]; location data (Foursquare check-ins) is represented in a form of distribution among venue categories and mobility patterns [49, 54]; textual data (Twitter tweets, Instagram image captions, and Foursquare check-in comments) is represented in a form of latent topics [167], LIWC distribution [127] and writing style features [54]. The detailed description of the adopted feature extraction processes is given in Section 4.1.1.

III. Multi-Source Community Detection: The user communities are detected via regularized spectral clustering on a multi-layer graph, as described in Section 5.1. Each layer in the graph is constructed based on UGC similarity between users, which is estimated as a $Heat$ distance between data representations in each social network (see Section 5.1.2).

IV. User Community Profiling: User Community Profiling is performed in three steps: (a) individual user profile learning [49] is applied to each community member, which results in a set of personal users' attributes, such as Age, Gender, Occupation Industry, Education Level, Income Level, Relationship Status, etc.; (b) the extracted attributes are summarized among all community members for further visualization in a form of "Donut" diagrams (see Figure 5.6).

V. Live Stream Analytics*: Except for sentiment analysis for every post [54], the live stream analytics is performed in a form of T-SQL queries [13] into a real-time social multimedia stream. bBridge adopts the Windows Azure Stream Analytics cloud service [11], which allows for using various query constructions such as "Like/Not Like patterns", "CASE statements", "Embedded procedures", etc. Below, we give an example of the query that will output distribution among extracted latent topics and content of all the multimedia posts published from "Downtown" district during the last 30 seconds by the "Arts" community

[7] As on August 1st 2017



RONIN
Caffee
Fantastic place for breakfast and flat whites. The poached egg takes a little time, but is well worth the wait. Great little place.

Specified query: SELECT * FROM community.Arts

Figure 5.6: The bBridge online GUI snapshot

with the "Positive" sentiment score:

```
SELECT Topics, Content FROM MultimediaStream
TIMESTAMP BY CreatedAt GROUP BY SLIDINGWINDOW(s, 30)
WHERE Sentiment='Positive' AND Community='Arts'
AND Region='Downtown'
```

From the example, it can be seen that the aforementioned querying approach can be used to query multimedia data at both global and local levels.

To take advantage of parallel processing, we have implemented bBridge platform as a multi-role cloud service [12]. Each crawler runs as a separate Windows Azure worker role, which enables us to harvest multiple social multimedia venues simultaneously. The community detection and profiling modules are implemented as multi-instance Windows Azure worker roles, so that it can be performed for different time frames at the same time. The stream analytics module utilizes Azure Stream Analytics engine [11] that can perform complex real-time queries [13] into the live social multimedia stream. All worker roles and stream engine are synchronized via the Windows Azure Service Bus.

Using bBridge's web GUI [55], the users can select one of the recently detected communities in a specified geographical region. Based on the selected user community, the following features are provided: (a) observe geo-activity of the community members; (b) discover hot multi-modal topics that were discussed by the community members; (c) view the distribution of the community members among automatically inferred user attributes (such as age, gender, occupation industry, education level, relationship status, and income level); and (d) perform queries of different complexity into the live social multimedia stream, which is generated by the community members.

A snapshot of the GUI is shown in Figure 5.6. The GUI shows the map of a Singapore region and the opened visualization console for the "Arts" community. The map shows the polygons that describe recent community members' geo-activity; the left pane displays the hot topics, discussed by the community members and the distribution of the community members among the detected personal user's attributes; while the right pane lists the recent cross-network posts of the community members, which are filtered according to the specified query. bBridge users can select different user communities, view visualization, specify real-time queries into the community stream via the query UI control.

5.3 Summary

In this chapter, we presented a study on multi-source group profiling and its application in Urban Mobility Domain — Cross-Source Venue Category Recommendation. The proposed recommendation framework C^3R utilizes both individual and group knowledge to solve a task of venue category recommendation. Individual knowledge is modeled as the distribution among venue categories that user has visited in past, while group knowledge is modeled as the distribution of Foursquare venue categories among user community members. The user communities are detected based on novel regularized spectral clustering approach that is able to perform an efficient partitioning of multi-layer user relations graph. By performing comprehensive evaluation against the state-of-the-art baselines and different data source combinations, we demonstrated that our proposed user-community-empowered venue category recommendation framework can achieve superior recommendation performance.

To show the applicability of our individual and group user profiling solutions to the real-world scenario, we integrated all the earlier proposed techniques in cloud-based Big Data Platform for Social Multimedia Analytics, called bBridge [55]. The platform automatically detects and profiles meaningful user communities in a specified geographical region, followed by rich analytics on communities' multimedia streams. The system executes a community detection approach that considers the ability of social networks to complement each other during the process of latent representation learning, while the community profiling is implemented based on our proposed multi-source individual user profiling techniques. The multi-source data crawler deployed as a distributed cloud jobs. Overall, the bBridge platform integrates all the above techniques to serve both business and personal objectives. The major contribution of this chapter is generic multi-layer (multi-source) graph clustering approach that considers inter-source relationship during the clustering process.

Chapter 6

User Profiling Analytics

In previous chapters, we proposed approaches for individual and group user profiling from multiple social networks and wearable sensors. The performance of the methods was evaluated by comparing with different state-of-the-art baselines from the corresponding application domains. However, till now it is still not clear what the exact relationship is between different data modalities and data sources towards individual and group user profile learning. To gain insight into such an association, in this chapter we conduct statistical analysis of multi-source user profiles.

6.1 Introduction

The growing amount of multi-modal user-generated content all over the Web inspired new directions in different research domains. One such direction is user profile learning from multiple social networks. The task is conventionally approached by utilizing various supervised and unsupervised machine learning techniques aiming at individual and group attributes inference. Most of the works towards multi-source user profile learning reported better performance by utilizing multiple data sources (modalities), as compared to single-source (single-modal) baselines [28, 51, 54, 153, 154]. However, the majority of them did not provide a comprehensive explanation of the relation between different information queues and the performance improvement.

At the same time, many groups from social science research community conducted a qualitative analysis of user-generated content. For example, there have been multiple research attempts done on the understanding of the relation between social media and various aspects of human wellness. Specifically, Culotta et al. [42] performed an analysis on Twitter data aiming to predict US citizens' health-related statistics such as obesity or teen pregnancy, while Chou et al. [35] conducted a qualitative study on weight-related interactions of Twitter and Facebook users. Mejova et al. [105] performed group obesity analysis based on data from Foursquare, while Sharma et al. [149] leveraged Instagram data in an attempt to measure food calories from image-related hashtags. Later, Jurgens et al. [85] drew first-order statistics and performed preliminary analysis at a group level of Fitocracy[1] users' exercising activity, while Park et al. [122] studied publishing traits of social media users who openly share their individual health and fitness related information. As noted, there were research efforts made towards wellness lifestyle analysis and the results showed a great potential of social media data to assist wellness related research. However, most of the works mentioned above use only a single data source and thus do not review the inter-source and cross-modal data relationship in Wellness domain. They may not be useful to gain deeper insights from multi-source social media data and wearable sensors perspective.

Meanwhile, there have been multiple research attempts made towards a qualitative evaluation of social media data relation in Urban Mobility domain. With the emergence of smartphones, the data from location-based social networks was broadly adopted in many research works. Good examples are the works performed by Ye et al. [179, 180]. First, the authors mined spatial-temporal patterns between individual points of user trajectory followed by pattern semantic meaning assignment [179]. Later, they extended the earlier proposed approach by considering implicit relations between venues (co-visiting and category similarity) and incorporating statistics such as check-ins distributions on daily (24 hours) and weekly (7 days) basis [180]. Later on, Noulas et al. [117, 118] and Scellato et al. [145] analyzed user mobility patterns based on LBSN data at a large scale. The dataset was collected from Foursquare and Gowalla LBSNs and served as a good support to examine the classical user mobility modeling approaches [156]. As a result, Noulas et al. [118] concluded that simple metrics like distance cannot describe human migration rules accurately and they need to be supported with richer data. Such data can be inferred from LBSN. Finally, Cho et al. [34] utilized cell phone location data together with data from two LBSNs to show that short-range travel is spatially and temporally periodic, while social network ties influence more long-distance travel. Many of these studies provide a good overview of various aspects of human mobility. Some of them suggest the utilization of rich multi-modal data for deeper user mobility analysis. However, till now there is not much work devoted to cross-source data analysis in the Urban Mobility domain.

It can be seen that the area lacks research towards cross-source and cross-modal analysis in Wellness and Urban Mobility domains, which introduces the large research gap. The possible reason behind such gap is the absence of multi-source multi-modal datasets for a large-scale social science research. Specifically, the necessity of inter-network account disambiguation makes it tough for researchers to obtain unbiased large-scale multi-source multi-modal data. Another possible problem is the lack of consistent data representation between different data sources: different data is often available in many diverse forms, such as video, images, text, sensor measurements, locations. These data modalities must be represented in a mutually-consistent fashion before any sort of relationship analysis can be

[1] www.fitocracy.com

conducted.

Aiming to bridge this research gap, in this chapter, we seek to address two research questions. First, to understand the role of multi-source data for individual user profiling it is important to know: **RQ1: What is the relation between different data modalities, data sources, and individual user attributes?** Second, to gain an insight into the importance of multi-source data for group user profiling it is necessary to understand: **RQ2 What is the relation between different data modalities, data sources, and detected user communities?**

To answer these two research questions, we performed a large-scale analysis of multi-source multi-modal data towards Wellness and Urban Mobility Applications. Specifically, to uncover the relationship between different data representations and wellness attributes, we've conducted an extensive correlation analysis of different data representations at both inter-source and intra-source levels. To gain further insights from the group profiling perspective, we've utilized conventional spectral clustering on each data source to detect user communities from each source independently. We then conducted inter-source community assignment correlation analysis, which helps to uncover the relationship between the obtained user communities. Our experimental results conform well with the assumptions behind this thesis. Specifically, they show that inter-source data represent users from various diverse perspectives while being significantly correlated at both inter-source and intra-source levels. The latter encourages further qualitative and quantitative research on multi-source individual and group user profiling.

6.2 Analysis Approach

In this chapter, we utilize Pearson correlation coefficient to examine the strength and direction of the linear relationship between all pairs of variables in our multi-modal multi-source data. We then employ significance testing in order to highlight and further discuss only significant data relationships. The variables to be analyzed are: multi-source data representations, ground truth labels, detected user communities.

6.2.1 Correlation Analysis: Pearson correlation

Correlation coefficient between two random variables X and Y for a population is defined as follows:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, \tag{6.1}$$

where $Cov(X, Y)$ is population covariance, and $Var(X)$ is population variance.

Since the data for the whole population is not available due to the privacy restrictions of social networks, the population correlation coefficient $\rho(X, Y)$ can not be computed. We thus utilize sample correlation coefficient r, which is an estimate of the unknown population correlation coefficient for a representative sample of size n:

$$r = \frac{\hat{Cov}(x, y)}{\sqrt{\hat{Var}(x)\hat{Var}(y)}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{6.2}$$

where x_i and y_i are the ith members of the population sample, and $\bar{x}$ and $\bar{y}$ are the population sample means: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

6.2.2 Correlation Significance Testing: P-Value Test

To evaluate the significance of each computed linear correlation between all pairs of variables in our population sample, in all further sections, we perform hypothesis testing by following the so-called "p-value approach". The hypothesis test reveals whether the value of the population correlation coefficient ρ is significantly different from 0. The latter can be decided based on the computed sample correlation coefficient r and the sample size n. Precisely, given the *null* hypothesis[2] (H_0) saying the actual correlation between X and Y within the general population is $\rho = 0$; the alternative hypothesis[3] (H_A) claiming that the actual correlation between X and Y within the general population is $\rho \neq 0$; and the actual size of the sample $n \geq 6$ (on which an observed value of r is based); the quantity

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} \tag{6.3}$$

can be used to calculate the p-value, which is the probability that, when the null hypothesis is true, the statistical summary would be the same as or more extreme than the actual observed results [173]. The p-value is determined by referring to a t-distribution with $df = n - 2$ when conducting two-tailed test. To make a decision regarding the statistical significance of the correlation between X and Y, the obtained value of p can be compared with the pre-defined significance level α. If $p \leq \alpha$, we reject H_0 in favor of H_A; if $p > \alpha$, we fail to reject H_0.

[2] **There is no sufficient evidence** at the α level that there is a linear relationship between the X and Y

[3] **There is sufficient evidence** at the α level that there is a linear relationship between the X and Y

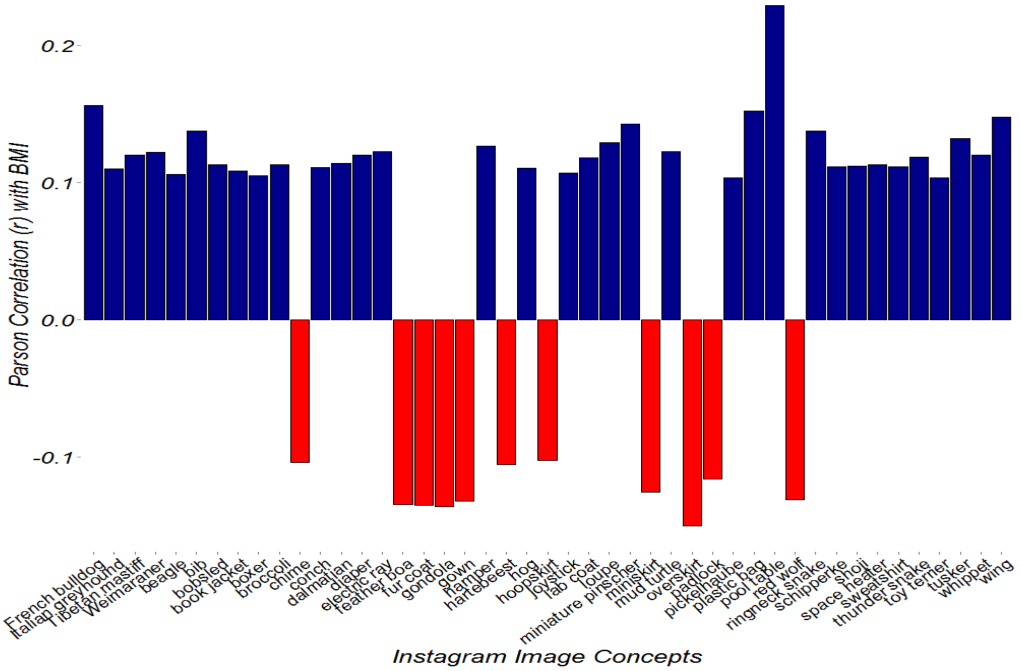


Figure 6.1: The image features are significantly ($\alpha = 0.05$) correlated to users' BMI data representations. The negative and positive correlations are colored in red and blue, respectively

6.3 Correlation Analysis Towards Individual User Profiling

To answer research question (1)[4], we first analyzed our data from wellness domain by utilizing Pearson Correlation Coefficient (r) to uncover the relationship between different data representations, wearable sensors data, and individual user attributes. The correlations are computed based on users from "NUS-SENSE" [51] dataset who have contributed to at least one of four social networks: Foursquare, Instagram, Endomondo, or Twitter. We discovered and characterized the relationship between individual wellness attributes (BMI, "BMI Trend") and different data modalities, namely text, images, locations, and sensor measurements. To do so, we highlighted significant correlations between different data representations and BMI/"BMI Trend". We utilized the significance test with the $\alpha = 0.05$. The data sample sizes n are specified in Table 3.1. It is noted that we consider negative correlation to "BMI Trend" as being favorable since it means losing weight.

[4] *What is the relation between different data modalities, data sources, and individual user attributes?*

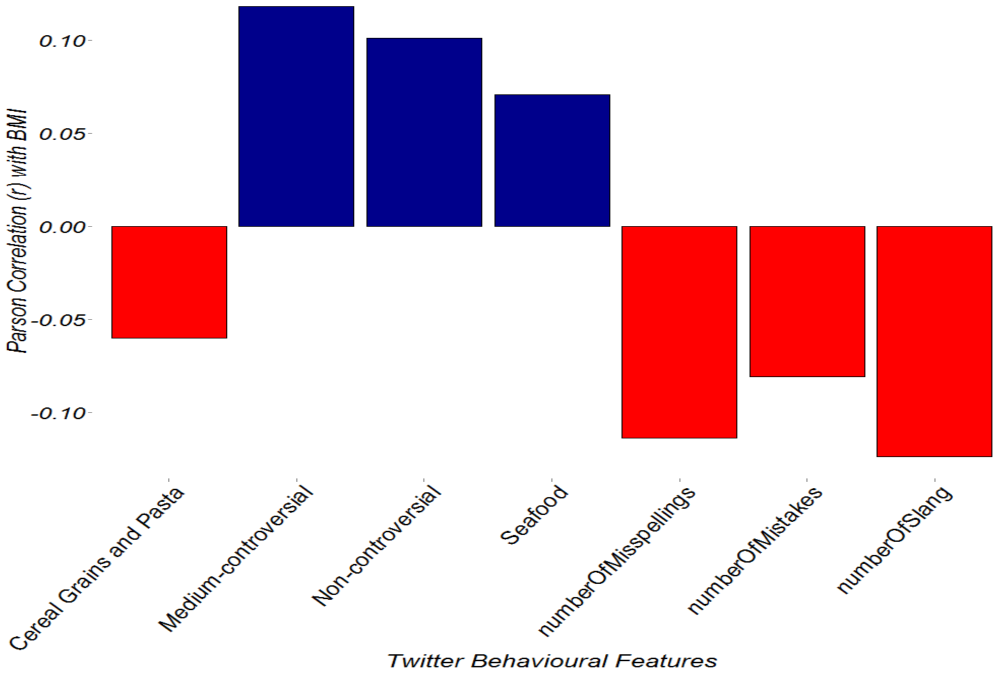


Figure 6.2: The heuristic textual features are significantly ($\alpha = 0.05$) correlated to users' BMI data representations. The negative and positive correlations are colored in red and blue, respectively

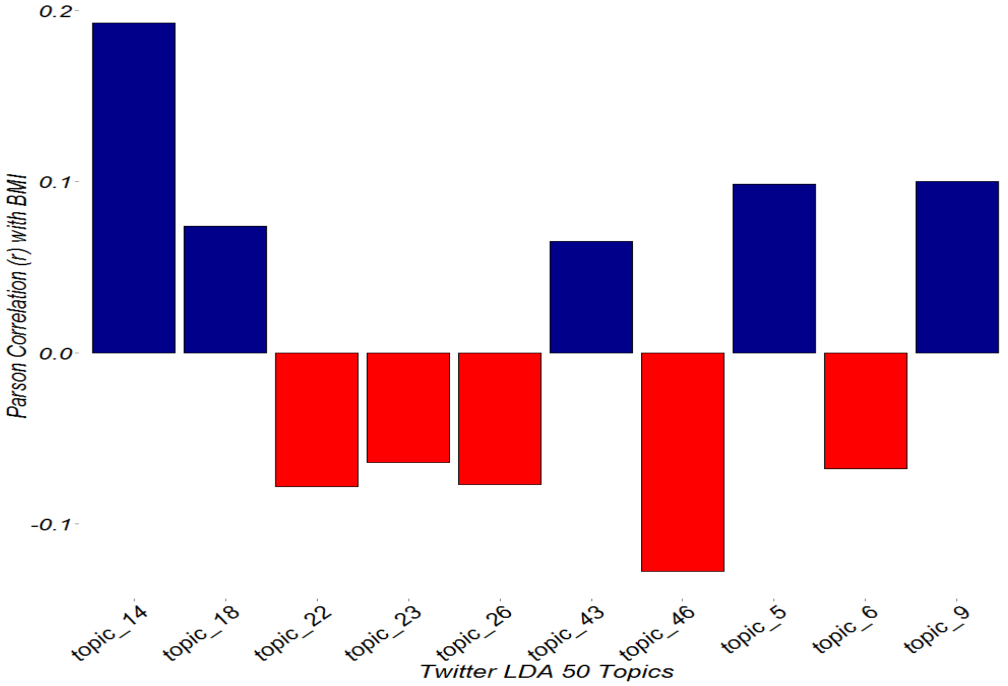


Figure 6.3: The LDA textual features are significantly ($\alpha = 0.05$) correlated to users' BMI data representations. The negative and positive correlations are colored in red and blue, respectively

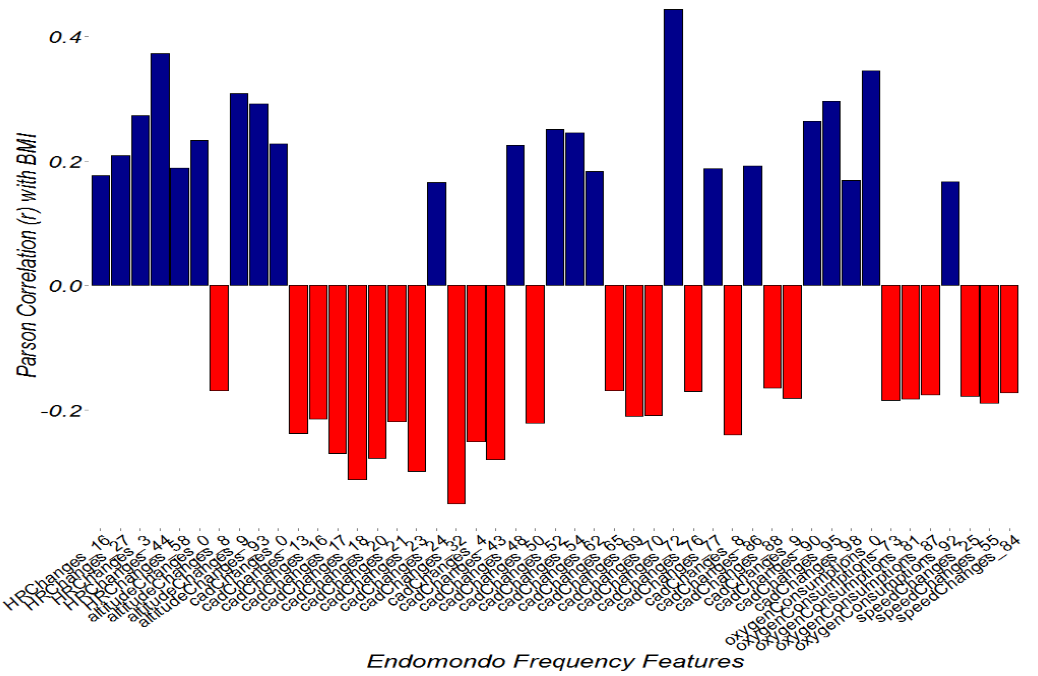


Figure 6.4: The frequency spectrum sensor features are significantly ($\alpha = 0.05$) correlated to users' BMI data representations. The negative and positive correlations are colored in red and blue, respectively

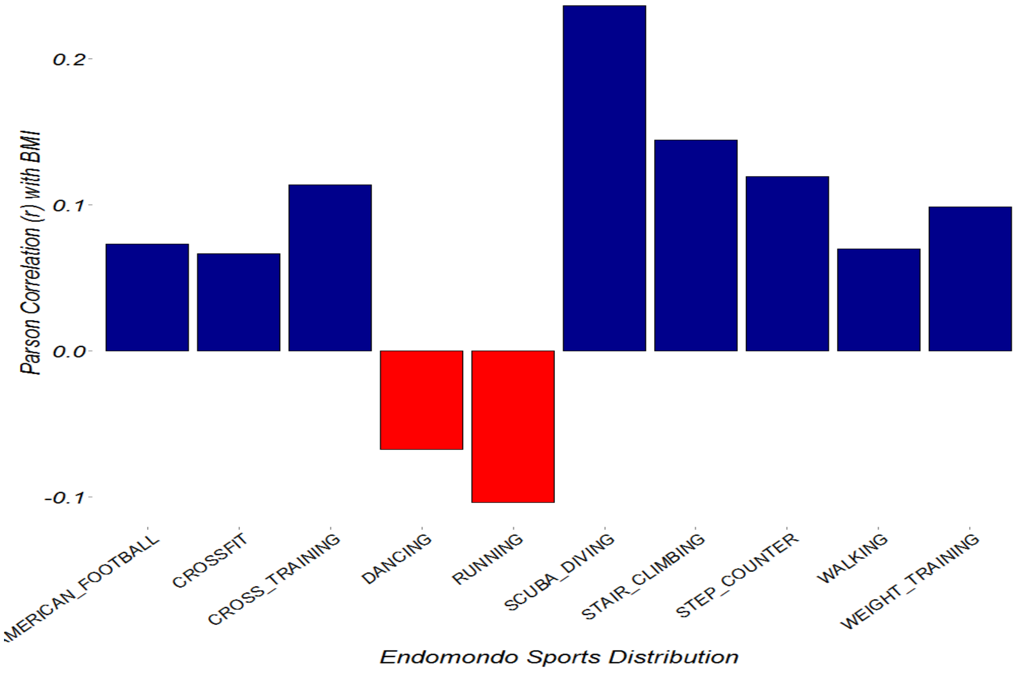


Figure 6.5: The sport type distribution sensor features are significantly ($\alpha = 0.05$) correlated to users' BMI data representations. The negative and positive correlations are colored in red and blue, respectively

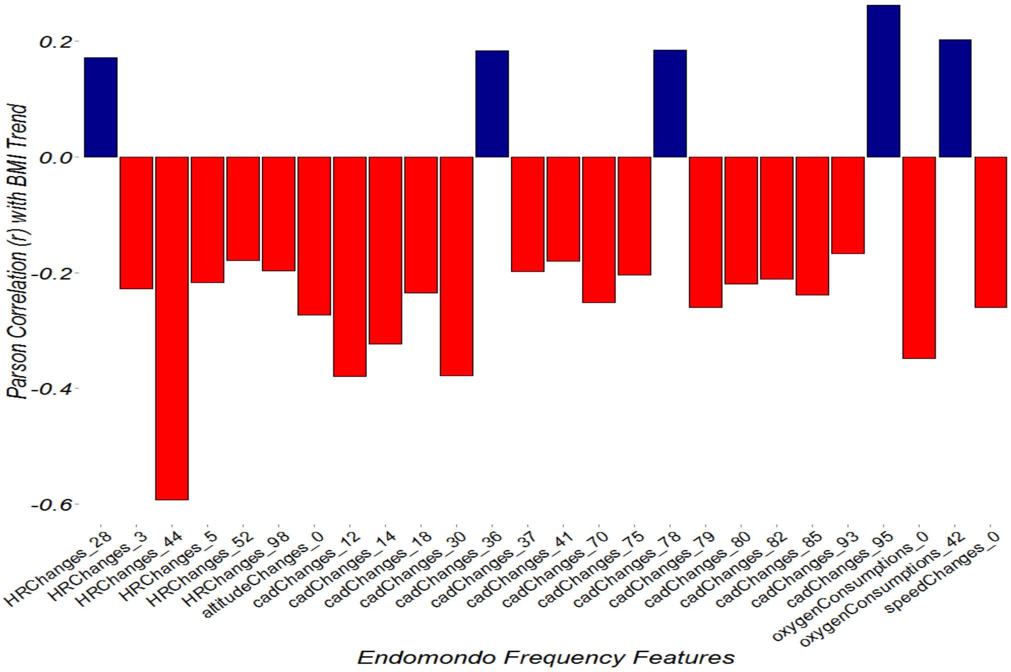


Figure 6.6: The frequency spectrum sensor features are significantly ($\alpha = 0.05$) correlated to users' "BMI Trend" data representations. The negative and positive correlations are colored in red and blue, respectively

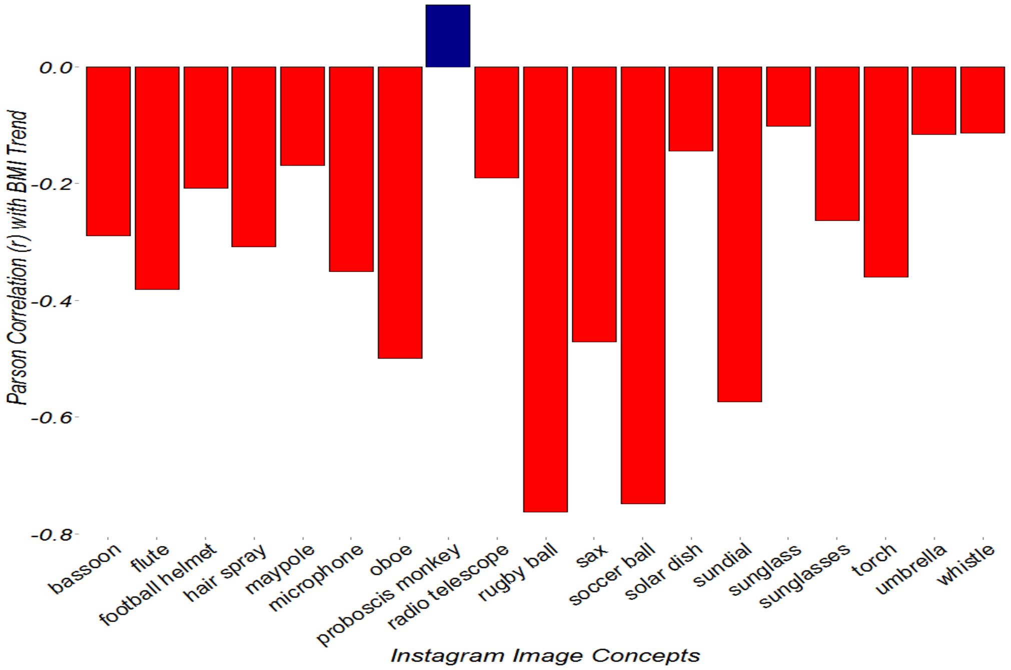


Figure 6.7: The image concept features are significantly ($\alpha = 0.05$) correlated to users' "BMI Trend" data representations. The negative and positive correlations are colored in red and blue, respectively

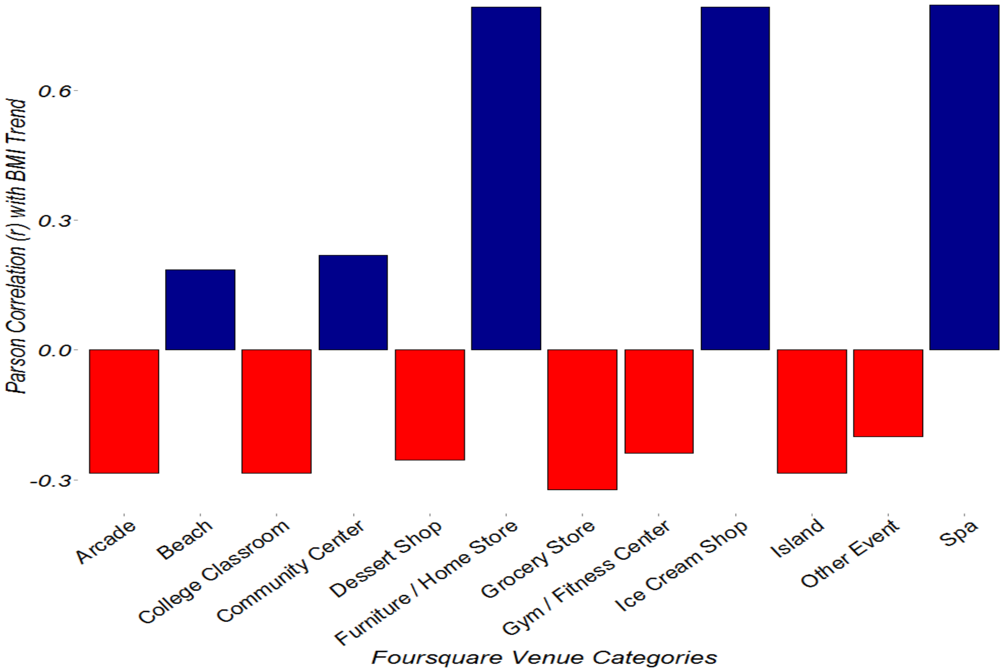


Figure 6.8: The venue category features are significantly ($\alpha = 0.05$) correlated to users' "BMI Trend" data representations. The negative and positive correlations are colored in red and blue, respectively

6.3.1 Data from Wearable Sensors

To study sensor data, we performed correlation analysis between wellness attributes and two types of sensor data representations: distribution among Fourier spectrum bins and workout categories (exercise semantics).

First, it can bee seen that users' BMI is negatively correlated to active sport types ("Dancing", "Running"), while positively correlated to moderately active sport types ("Scuba Diving", "Walking", "Cross-Training") (see Figure 6.5). The possible reason is that people with higher BMI may select less intensive workout types due to their mobility restrictions and fitness level, while fitter ones often prefer more intensive workouts [85].

At the same time, there is a significant number of frequency bins from each sensor type that are correlated to users' BMI (Figure 6.4): 5 "HR" bins, 4 "Altitude" bins, 30 "Cadence" bins, 5 "Oxygen" bins, and 3 "Speed" bins. The prevalence of "Cadence" bins features can be explained by the relatedness of cadence to exercises' tempo, which, in turn, is related to the actual fitness level of users: fitter users may perform sports activity with different intensity as compared to less fit ones. Lastly, the significant correlation to user' "BMI Trend" (Figure 6.6) could be explained by the connection of exercise intensity and energy spending, which is related to weight loss.

To summarize, *both exercise semantics and frequency bins features are tightly knit to BMI and "BMI Trend"*, which can be seen from the correlation plots. Intuitively, it can be explained by the ability of sensor data to reveal knowledge about users' actual physical status, which is not available from other data sources. This again supports the incorporation of sensor data into the learning process (as pointed out in Section 4.2.1).

6.3.2 Data From Social Media

The Instagram image concepts "Diaper" and "Sweatshirt" are positively correlated to users' BMI, while "Over Skirt" and "Gown" image concepts reveal negative correlation (Figure 6.1). The potential reason might be the difference of dressing preferences between users of different BMI: users with overweight problems may take pictures of themselves wearing baggy garments, while people with lower BMI may prefer slinky clothes instead. It is also interesting to note that "Rugby Ball" and "Soccer Ball" show the highest negative correlation level to "BMI Trend". The explanation could be that sports activities (represented by sport-related image concepts) help users to lose weight. Generally, it can be observed that *many Instagram image concepts are significantly correlated to personal wellness attributes.* The possible reason is that these image concepts represent wellness-related aspects of human life, such as diet, sports activities, and dressing preferences. They are useful indicators of users' wellness.

As mentioned in a previous study [105], Foursquare data is correlated to the country-level obesity traits. The above conforms well with our observations (Figure 6.8), where the venue categories "Ice Cream Shop" and "Fitness Center" are significantly correlated to the tendency of gaining/losing weight, respectively. The relation of *multiple venue categories to users' dietary habits* supports the idea of utilizing such data for individual user profile learning in wellness domain (see Section 4.2.1).

Another interesting dependency can be observed between lexicon-based text features and the BMI categories — the *significant correlation between automatically extracted (from user-generated text) food groups* ("Cereal Grains and Pasta", "Seafood"). Such observation is also consistent with previous studies [105, 149, 4].

Let's summarize the above observations. First, the global data analysis results are consistent with previous works [54, 153] regarding the fact that multiple data sources describe users from different perspectives. Second, many data representations are significantly correlated to BMI and "BMI Trend" individual user attributes. At the same time, some data representations are not linearly correlated to them. These highlights the necessity of feature selection at the data modeling stage (as in Equation 4.3). Third, it can be seen that BMI index negatively/positively correlated to numerous data representations. Essentially, this means that the data representations expose proportional (or inversely proportional) trends to the wellness attributes. This suggests that the adjacent BMI categories may represent users with similar social media behavior, which must be considered at the learning stage (see

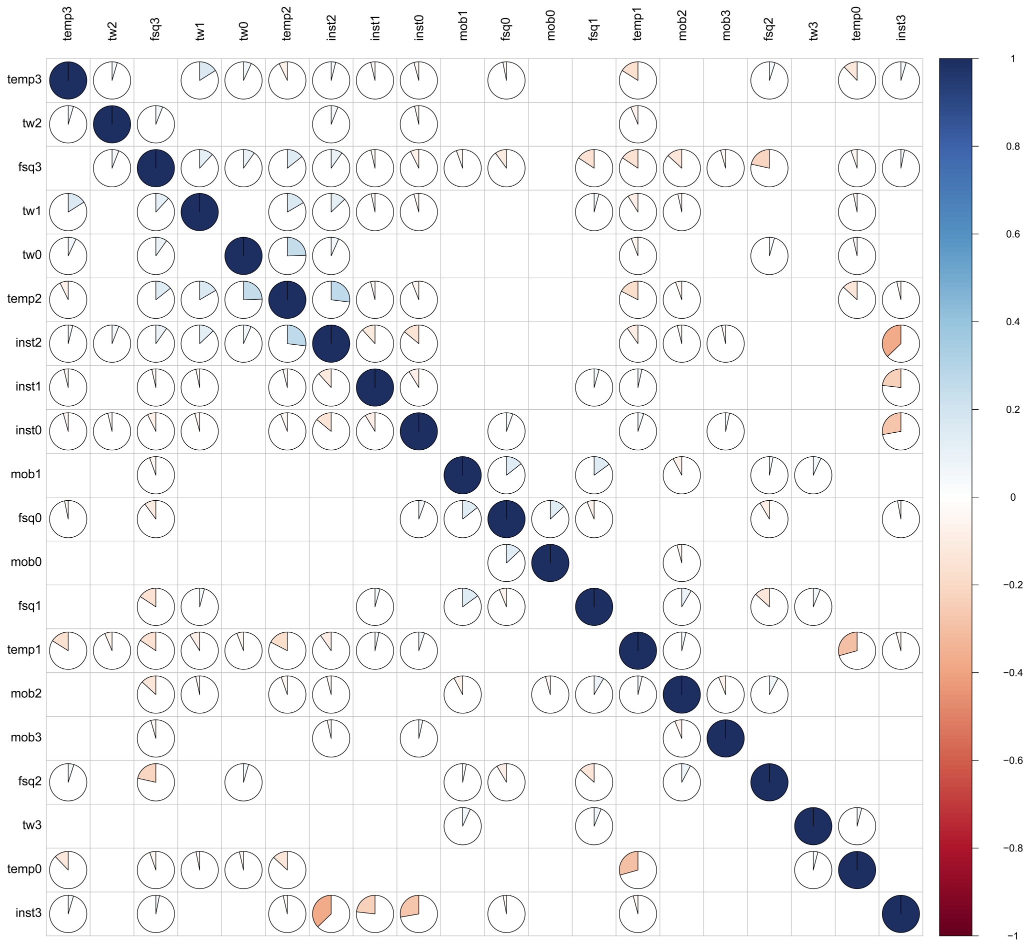


Figure 6.9: Inter-community correlation plot

Section 4.2.2). Fourth, we would like to note that our correlation analysis results are also consistent with the results of feature importance analysis in Section 4.2.3. Specifically, the observed significant correlation between food groups and wellness attributes is well aligned with the utilization of lexicon-based features for predicting the 6 BMI categories. At the same time, the significant correlation of Instagram image concepts and venue semantics is consistent with its extensive usage by all the predictive models. Finally, the significant correlation of multiple frequency bins has resulted in the dominance of the frequency domain features in the final models as compared to other feature types. Based on the above investigations, **we respond to the research question (1) by concluding that it is reasonable to incorporate all the harvested data sources and data from wearable sensors into individual user profiling**.

Table 6.1: Community assignment matrix, where communities are detected from each data source separately and $k = 4$

User	Foursquare usr. com.				Instagram usr. com.			
	4sq0	4sq1	4sq2	4sq3	inst0	inst1	inst2	inst3
user 1	0	0.27	0	0	0.23	0	0	0
...	...	...	...	...	...	...	...	...
user k	0.14	0	0	0	0	0	0.33	0

6.4 Correlation Analysis Towards Group User Profiling

To answer research question (2)[5], we utilized Pearson Correlation Coefficient (r), which depicts the relationship between user communities exposed from each data source independently. In such a setting, each data source is represented by a graph structure, where all the graphs share the vertex set. To detect user communities, we utilize the standard spectral clustering [102]. We empirically set $k = 4$ (number of clusters to detect), since it allowed us to achieve the best recommendation performance in the case when spectral clustering is executed on each social network separately (see Chapter 5.1.4).

The correlation analysis is performed based on data from the NUS-MSS dataset [54]. Specifically, the correlation matrix is built based on users in Singapore who have contributed to all three social networks: Foursquare, Instagram, and Twitter. The Pearson's r is computed based on community assignment matrix, where the matrix is computed as follows: for every data source, user is assigned to one community and the value in matrix is equal to $Heat$ kernel (see Section 5.1.2) distance between the community centroid and user representation in that network. To explain the idea, we give a toy example in Table 6.1, where the number of detected communities for each source (k) is equal to 4; *user 1* is assigned to community *4sq1* in Foursquare and *inst0* in Instagram, while *user k* is assigned to community *4sq0* in Foursquare and *inst2* in Instagram.

We utilize the significance test with the $\alpha = 0.05$ and data sample size of 4490 users. The significantly correlated data representations are shown in Figure 6.9. For easier interpretation, we have named and profiled some of the detected user communities and listed them in Table 6.2. To achieve better visibility, we reorder the correlation matrix according to the angular order of the eigenvectors [63] and highlight only the significantly correlated pairs ($\alpha = 0.05$).

From the inter-community correlation matrix, it can be seen that community *inst1* from

[5] *What is the relation between different data modalities, data sources, and detected user communities?*

Table 6.2: User communities detected from each graph layer independently for $k = 4$

ID	Name	Profile (Terms)
Twitter		
1	Work/Active	#misspellings, #words/tweet, work
Foursquare		
1	Shop/Dine	Mall, Chinese Restaurant, Fast Food, Japanese Restaurant, Food Court
3	Business/Travel	Airport, Café, Hotel, Bakery, Bus Station
Instagram		
1	Beach/Leisure	Lotion, Sunscreen, Sea slug, Swan
3	Home/Family	Gong, Backpack, Dog, Pot, Library (Book)
Temporal		
2	Misc.	Weekends 12 – 15
3	Misc.	Weekends 6 – 12, Weekdays 18 – 24
Mobility		
3	Explorers	#AOI outliers

Instagram (related to “Beach/Leisure” activities) is positively correlated to community *4sq1* from Foursquare (“Shopping/Dining”), but negatively correlated to temporal community *temp2* (“Weekends 12 – 15”) and Foursquare’s community *4sq3* (“Business/Travel”). The first relation could be explained by the similarity between communities members lifestyle, where leisure activities are often followed by shopping/dining activities and vice versa. The negative relation, in turn, could be hypothesized by the lack of business trips and business-related activities for the users from the “Beach/Leisure” community.

At the same time, Twitter community *tw1* (“Work/Online Active”) is positively correlated to temporal communities *temp2* and *temp3* (“Weekends 6 – 15, 18 – 24”), which could be explained by the working schedule of work-oriented users: those who are more occupied by daily routines on weekdays, but are able to spend more time online during weekends.

At last, Foursquare’s community *4sq3* (“Business/Travel”) is positively correlated to the Instagram’s community *inst3* (“Home/Family”), while negatively correlated to mobility community *mob3* (“Explorers”). Both examples could be explained by demographics and lifestyle of “Business/Travel” users. Specifically, it could be hypothesized by the tendency of business-oriented users to spend more time with family in the first case (positive correlation with Home/Family community), and by everyday habits of visiting certain venues for dining or outing in the second (negative correlation to the “Explorers” community).

It is worth noting that most of the communities extracted from the same layers are negatively correlated to each other, which shows the good clustering ability of the spectral clustering and the separability of graphs constructed based on each data source. **The large number of the significant inter-community correlations shows that the data sources of**

different modalities are able to complement each other, which responds to the research question (2). The latter observation also supports the idea of joint community detection from multi-source multi-modal data (see Section 5.1.2).

6.4.1 On Inter-Source Relationship

In past works [80, 174, 49], the research community highlighted the importance of studying the relationship between different social media sources and sensors at the inter-source (cross-modal) level. To gain more insights into such a relationship, in this section, we additionally perform correlation analysis at the inter-source level. To do so, we extracted LDA topics from each data source and modality [168], where the number of topics K is dictated by the lowest perplexity values and equals to 50, 9, 7, 6, for Twitter, Instagram, Endomondo, and Foursquare data sources, respectively. For all data types, we treat all user data as a document (one document corresponds to one user) and each data unit (a word in a Twitter message, Instagram image concept, Endomondo workout type, Foursquare venue category) as a "word". After obtaining the latent topics, we computed Pearson's r between all the extracted topics, user demographic attributes, temporal and mobility features. The correlations are computed based on users from "NUS-SENSE" [51] dataset who have contributed to all four social networks: Foursquare, Instagram, Endomondo, and Twitter. The results are presented in Figures 6.10, 6.11, 6.12, 6.13, where inter-source and intra-source high order relations are highlighted in a correlation matrix. For better readability, we name some of the extracted topics and list them in Table 6.3. To achieve better visibility, we reorder the correlation matrix according to the angular order of the eigenvectors [63] and highlight only significantly correlated pairs in it. We utilize the significance test with the $\alpha = 0.05$ and data sample sizes n are indicated in Table 3.1. Several interesting significant correlations are discussed below.

Inter-Source correlation: Except the abundance of intra-source correlations, due to obvious reasons, the strong correlation can also be observed at the inter-source level. For example, topic 7 from Endomondo (intensive sports categories) is positively correlated to topic 8 from Instagram (represented by beach/leisure image concepts), but negatively correlated to topic 28 from Twitter (one of the sport-related textual topics). Such a relation could be hypothesized by the difference of information that users tend to uncover in different social venues [153]: users who prefer more intensive workouts may not post their sports activities on Twitter due to the nature of their sports life, but may like to take more pictures of leisurely occasions.

Temporal dependencies: We observed that most of the important $3h$ time durations $(0-3, 6-9, 9-12, 18-21, 21-24)$ are significantly correlated to latent topics from all data sources: topic 28 from Twitter and topic 3 from Endomondo are negatively correlated to time duration $18-21$, but positively correlated to time duration $9-12$ [116].

Mobility dependencies: From the correlation matrix, it can be seen that many of user

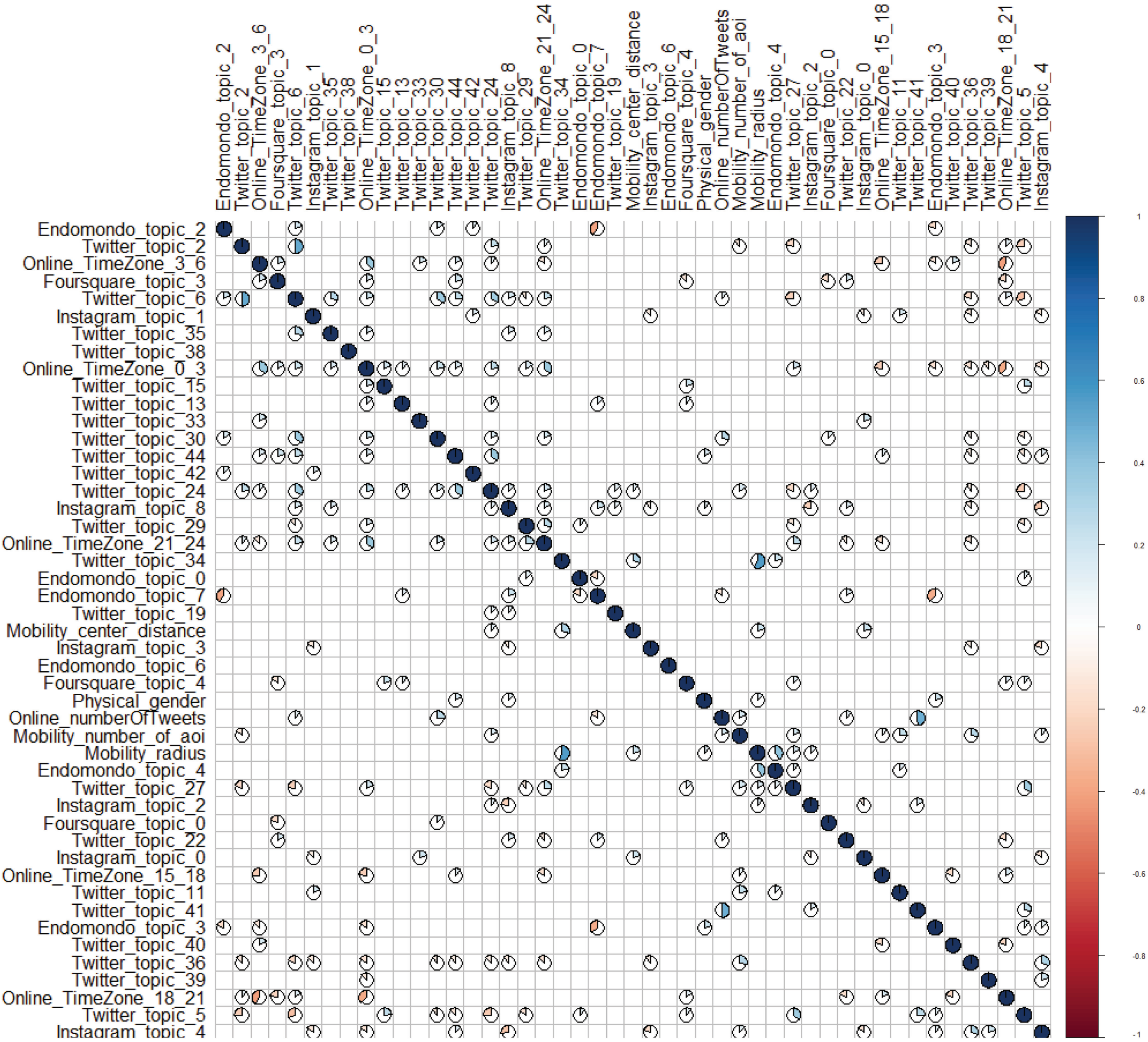


Figure 6.10: Significantly ($\alpha = 0.05$) inter-source correlated data representations (matrix 1 out of 4). The negative and positive correlations are colored in red and blue, respectively

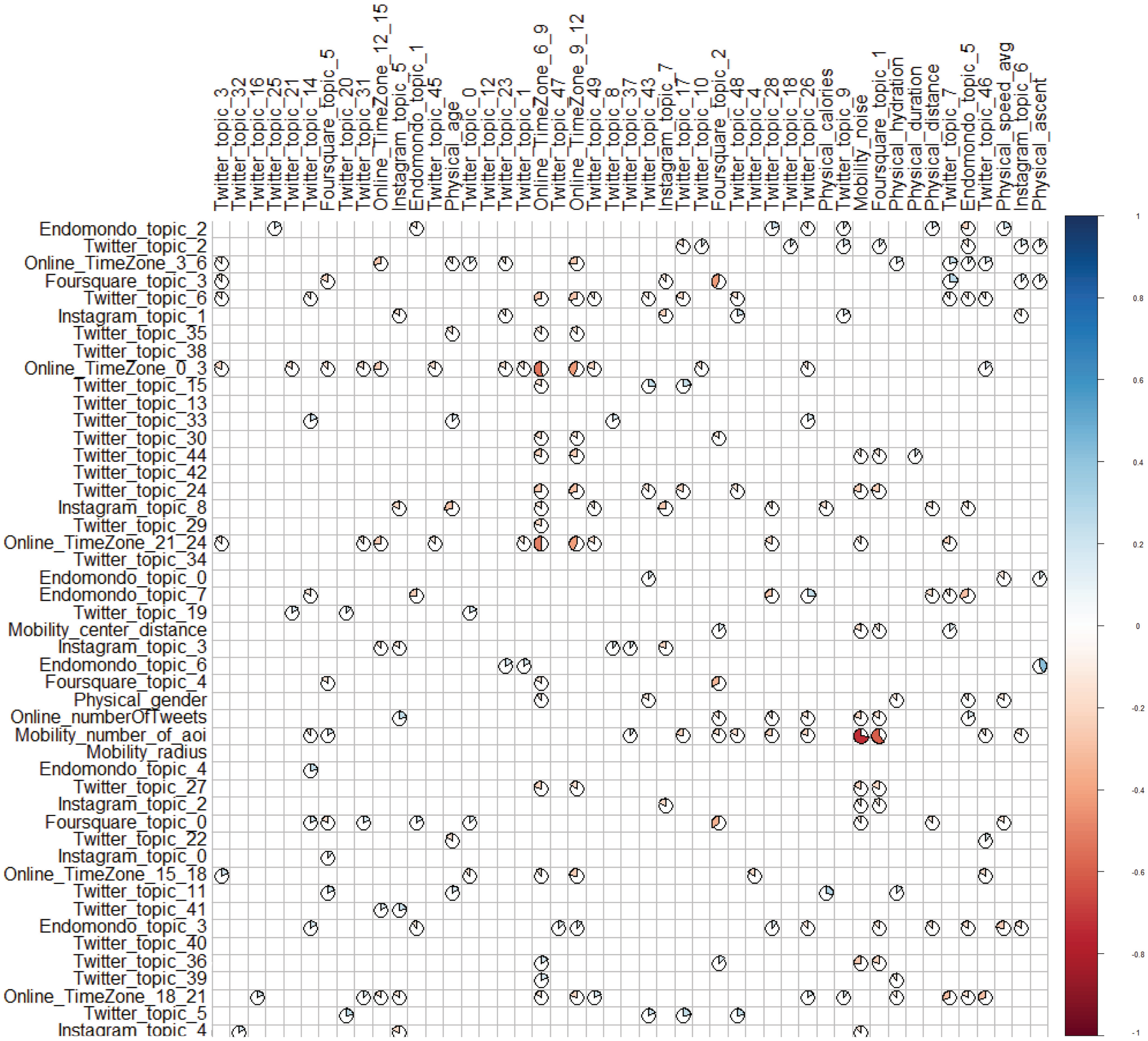


Figure 6.11: Significantly ($\alpha = 0.05$) inter-source correlated data representations (matrix 2 out of 4). The negative and positive correlations are colored in red and blue, respectively

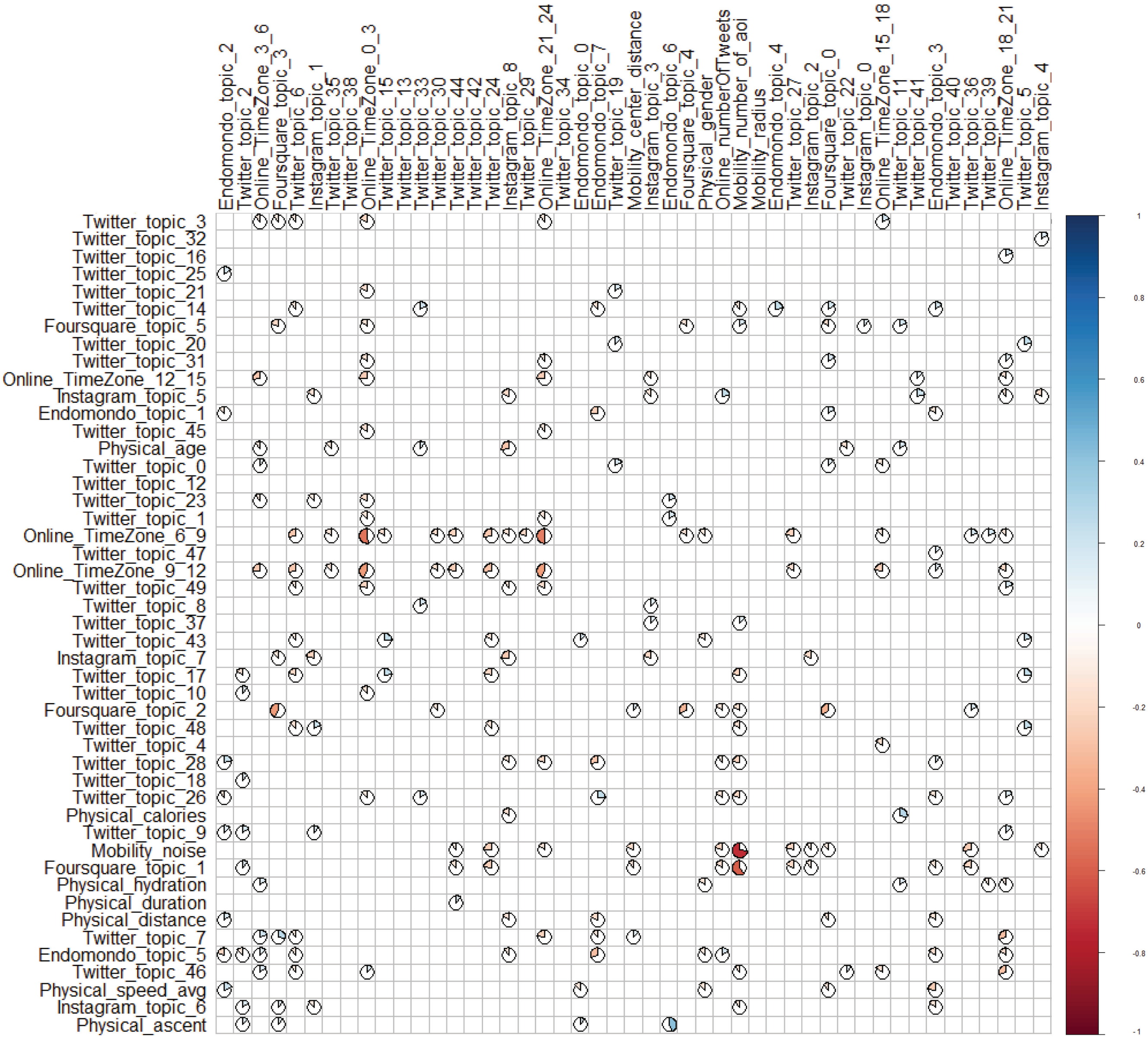


Figure 6.12: Significantly ($\alpha = 0.05$) inter-source correlated data representations (matrix 3 out of 4). The negative and positive correlations are colored in red and blue, respectively

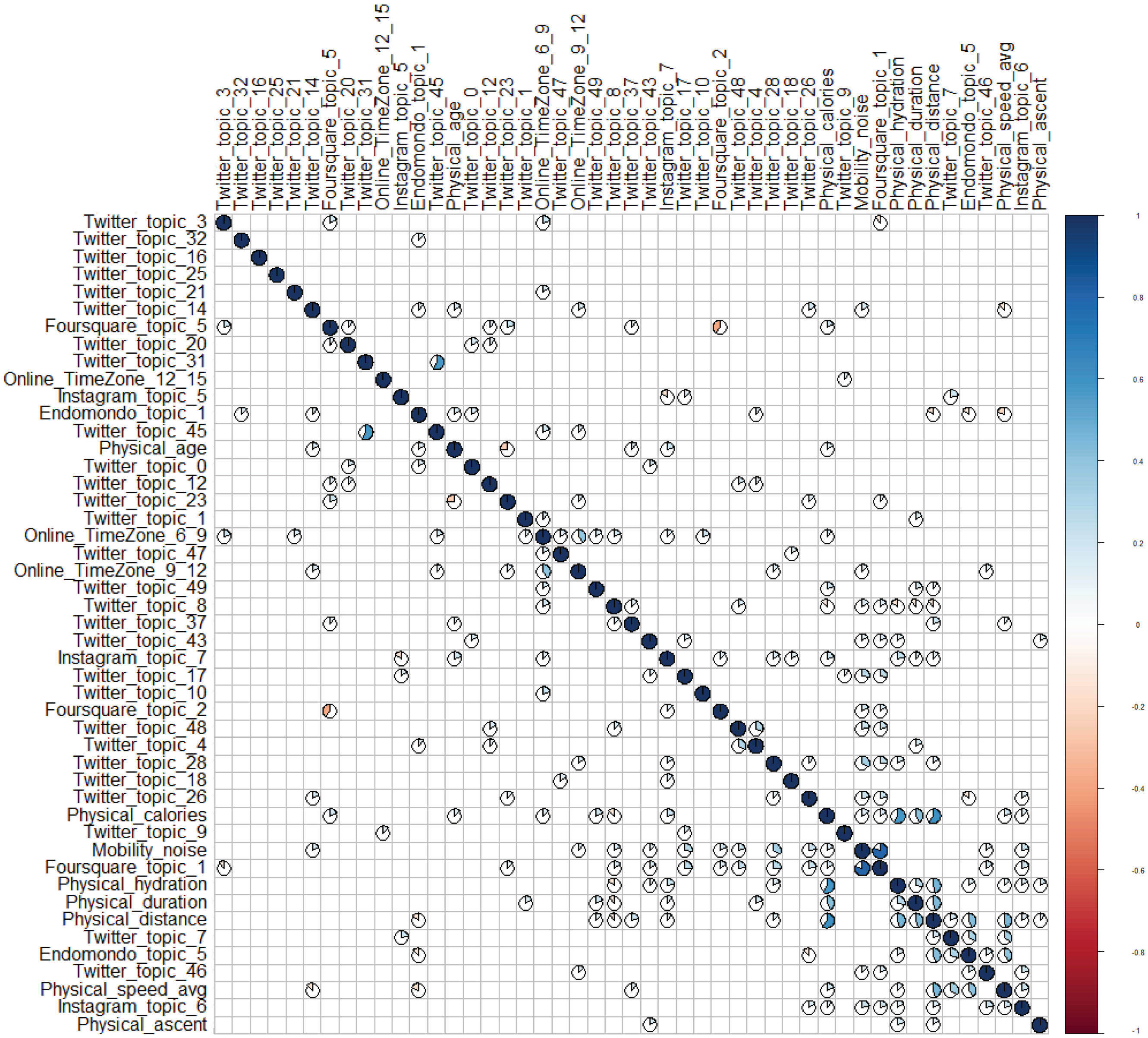


Figure 6.13: Significantly ($\alpha = 0.05$) inter-source correlated data representations (matrix 4 out of 4). The negative and positive correlations are colored in red and blue, respectively

Table 6.3: Terms distribution among LDA topics extracted from different data sources

Topic ID	Topic Name	Topic Terms (Words)
	Text latent topics	
14	Sport	mile, do, walk, fit, train, bit, step, travel, 00s, weight
28	Sport	finish, kiss, cycl, walk, 1h, bike, mountain, 2h, 3h
34	Happy	follow, pleas, list, sunshin, happiest, love, girl
	Foursquare venue categories latent topics	
0	Sport	Gym, Home, Office, Train Station, Bakery, Athletics
1	Shop	Restaurant, Shop, Store, College, Place, School, Bar
	Instagram image concepts latent topics	
6	Biking	mountain, bike, unicycle, tricycle, moped, helmet
8	Leisure	bathing, sunglasses, tie, wig, sunscreen, bow, maillot
	Endomondo workout categories latent topics	
3	H.Int.	walking, skating, gymnastics, cricket, baseball
4	M.Int.	step counter, walking, tr. walking, am. football
7	H.Int.	running, boxing, snow shoeing

mobility representations are significantly correlated to other data modalities. For example, the median size of users' AOIs (the size of the area, where users are active: "Home", "Work", etc. [131]) is significantly positively correlated to topic 34 from Twitter and to the Endomondo's topic 4 (related to walking). The possible reason is the difference of users' lifestyle: users, who, on average, travel by foot more (i.e. during the lunch break inside Work AOI) may have larger AOI size, as compared to those who use public/private transport.

Online-to-Offline dependencies: The encouraging observations can be made from the correlation between sensor data statistics and latent topics. Specifically, the average speed is negatively correlated to topic 0 from Foursquare and topic 14 from Twitter, but positively correlated to topic 6 from Instagram. We note that the negatively correlated topics are both related to fitness sports activities, where topic 6 from Instagram is related to biking and cycling. The possible explanation could be that people who perform stationary fitness activities are, on average, attain lower speed during their exercises as compared to those who use bikes for workouts and transportation. It also supports the idea that different data modalities are inter-correlated and mutually complement each other [53, 54, 153], which allows us to discover users' activities in much wider perspective as compared to modeling each modality independently (see Sections 4.1.2, 4.2.2, 5.1.2).

Aiming to bring attention to the problem of inter-source integration, we listed some interesting examples of inter-source data relationships above. These examples show that even completely different, at first sight, data modalities, such as Sensor Data and Data from Image Sharing Social Networks, are, indeed, significantly inter-correlated to each other and, thus, must be properly integrated for joint learning purposes. Such correlation may negatively affect linear models like in Equation 4.5 and may lead to sub-optimal performance. However, considering that we applied $\ell_{2,1}$ regularization as feature selection mechanism (see Equation 4.1), we do not expect the linear correlation between features to affect the final performance significantly.

6.5 Summary

In this chapter, we studied the relationship between different data sources, user attributes, and user communities. To uncover the relationship between different data representations and wellness attributes, we've conducted an extensive correlation analysis of different data representations at both inter-source and intra-source levels. To gain further insights from the group profiling perspective, we've utilized conventional spectral clustering on each data source to detect user communities from each source independently. We then conducted inter-source community assignment correlation analysis, which helps to uncover the relationship between the obtained user communities. Our experimental results show that inter-source data represent users from various diverse perspectives while being significantly correlated at both inter-source and intra-source levels, which encourage further research on multi-source individual and group user profiling. The major contribution of this chapter is the large-scale cross-source correlation data analysis and its discussion.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis' focus is on investigating user profiling across multiple social networks in Wellness and Urban Mobility application domains. Considering that user profiling can be performed on individual and group levels, this thesis proposed two multi-source learning schemes: multi-source learning approach for individual user profiling and multi-source community detection scheme for group user profiling. We practically applied the proposed approaches in three user profiling scenarios: demographic profiling, physical wellness profiling, and venue category recommendation. To demonstrate the applicability of the proposed methods in a real-world scenario, we integrated them into cloud-based social multimedia analytics platform, called bBridge.

In this thesis, we first performed individual user profiling. To build a comprehensive individual user profile, we trained the efficient ensemble to predict users' age and gender based on data from three social networks. We then leveraged the M^2WP framework to predict users' BMI category and "BMI Trend" based on data from multiple social networks and wearable sensors. The performance of the framework was enhanced by collective data source fusion strategy, the introduction of sparsity into data representations, and the idea of a multi-source multi-task learning via efficient inter-category smoothness regularization. To demonstrate the advance of our proposed individual wellness profiling solutions over other state-of-the-art baselines as well as gain deeper insight into wellness profiling performance, we conducted an extensive quantitative evaluation, classification error analysis, and feature importance analysis. Our experimental results demonstrate the crucial importance of joint learning from multiple social media and sensor data for individual wellness profiling, which encourages further research in this direction.

In this thesis, we further presented a study on multi-source group profiling and its application in Urban Mobility Domain — Cross-Source Collaborative Venue Category Recommendation. The proposed recommendation framework C^3R utilizes both individual and group knowledge to solve a task of venue category recommendation. Individual knowledge is modeled as the distribution among venue categories that user has visited in the past, while group knowledge is modeled as the distribution of Foursquare venue categories among user community members. The user communities are detected based on new regularized spectral clustering approach that can perform an efficient partitioning of multi-layer user relations graph. By performing comprehensive evaluation against the state-of-the-art baselines and different data source combinations, we demonstrated that our proposed user-community-empowered venue category recommendation framework can achieve superior recommendation performance.

To show the applicability of our individual and group user profiling solutions to the real-world scenario, we integrated all the earlier proposed techniques in cloud-based Big Data Platform for Social Multimedia Analytics, called bBridge [55]. The platform automatically detects and profiles meaningful user communities in a specified geographical region, followed by rich analytics on communities' multimedia streams. The system executes a community detection approach that considers the ability of social networks to complement each other during the process of latent representation learning, while the community profiling is implemented based on our proposed multi-source individual user profiling techniques. The stream analytics is performed via cloud-based stream analytics engine, while the multi-source data crawler deployed as a distributed cloud jobs.

Last but not least, we conducted a comprehensive analysis of the relationship between multiple data sources in Wellness and Urban Mobility domains. The study's results conform well with the assumptions behind this thesis. Specifically, it shows that inter-source data represents users from various diverse perspectives while being significantly correlated at both inter-source and intra-source levels. The latter suggests further qualitative and quantitative research on multi-source individual and group user profiling.

Let's once again summarize the major contributions of this thesis:

- Cross-network data gathering and alignment scheme that allows for large-scale data gathering from an arbitrary number of social networks that are enabled with inter-network reposting feature.
- Generic multi-source supervised learning framework that is able to learn from multiple incomplete data sources while performing feature selection and inter-category regularization on-the-fly.
- Generic multi-layer (multi-source) graph clustering approach that considers inter-source relationship during the clustering process.
- Large-scale cross-source correlation data analysis and its discussion.

7.2 Future Work

Except for prediction of user demographics and weight-related individual attributes, the inference of other individual and group wellness attributes can be approached in future works. For example, the so-called wellness-related hashtags could be used for users' wellness status prediction, wellness-related life events detection [4], and chronic disease tendency estimation. Precisely, the #bgnow hashtag could be utilized for diabetes-related social-media content identification and individual wellness timeline construction [4]. At the same time, such hashtags as #foodporn or #asthma could be employed in food consumption-related research [105] and asthma inference-related research [128], respectively. In a real-world scenario, such research must be conducted based on multi-source multi-modal data [54, 55, 51], which was not broadly utilized yet and thus expected to be widely adopted in future. Additionally, we would like to highlight that "NUS-SENSE" dataset naturally supports all the above research directions by providing not just personal wellness and demographics ground truth, but also the wellness-related hashtags as well as users' sports preferences and demographic attributes.

It is worth noting that one possible limitation of our proposed group user profiling scheme is its computational time complexity, which is introduced by the graph Laplacians computation. In future works, this issue could be resolved via reformulating the spectral clustering problem in a smaller space [33]. Specifically, by replacing $N \times N$ graph Laplacians with $P \times P$ matrices ($P \ll N$), a solution of mathematically-equivalent problem in significantly reduced space could be obtained. The final clustering could be then efficiently recovered from the obtained solution, while P can be chosen to be much smaller than N so that the running time could grow almost linearly to N [33]. Except for addressing the limitation above, our future work includes the development of semi-supervised techniques for user community detection guidance based on domain knowledge. Exactly, the domain knowledge about really-existing user communities (e.g. Facebook groups) or multi-source social graphs (e.g. social ties from different social networks) could potentially improve the quality of user community detection and recommendation in Urban Mobility domain. Finally, we plan to utilize additional data sources, such as a social graph, to diversify the community detection and recommendation outputs. Again, we would like to highlight that "NUS-SENSE" and "NUS-MSS" datasets support the research mentioned above.

Lastly, considering the wide applicability and the recent popularity of neural networks, in our future works, we will investigate the application of deep learning to individual and group user profiling [129]. Particularly, we plan to apply Long-Short Term Memory Networks (LSTMs) to leverage on the temporal dependencies within multi-source data for inference of such individual user attributes as Personality and Depression.

References

[1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[2] F. Abel, S. Araújo, Q. Gao, and G.-J. Houben. Analyzing cross-system user modeling on the social web. In *International Conference on Web Engineering*. Springer, 2011.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 2005.

[4] M. Akbari, X. Hu, N. Liqiang, and T.-S. Chua. From tweets to wellness: Wellness event detection from twitter streams. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, 2016.

[5] I. F. Akyildiz and W. Wang. The predictive user mobility profile framework for wireless multimedia networks. *IEEE/ACM Transactions on Networking (TON)*, 12(6):1021–1035, 2004.

[6] F. Alam and G. Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 955–959. IEEE, 2014.

[7] M.-D. Albakour, R. Deveaud, C. Macdonald, I. Ounis, et al. Diversifying contextual suggestions from location-based social networks. In *Proceedings of the 5th Information Interaction in Context Symposium*, pages 125–134. ACM, 2014.

[8] A. Ambrus, B. Greiner, P. Pathak, et al. Group versus individual decision-making: Is there a shift. *W. Paper 91*, 2009.

[9] 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 *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2014.

[10] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2009.

[11] M. Azure. Microsoft azure stream analytics service. `https://azure.microsoft.com/en-us/services/stream-analytics/`. Online; accessed 3 Aug 2016.

[12] M. Azure. Microsoft cloud services overview. `https://azure.microsoft.com/en-us/documentation/articles/cloud-services-choose-me/`. Online; accessed 3 Aug 2016.

[13] M. Azure. Stream analytics query language reference. `https://msdn.microsoft.com/en-us/library/azure/dn834998.aspx`. Online; accessed 3 Aug 2016.

[14] R. Bakeman and J. M. Gottman. *Observing interaction: An introduction to sequential analysis*. Cambridge university press, 1997.

[15] H. Banaee, M. U. Ahmed, and A. Loutfi. Data mining for wearable sensors in health monitoring systems: a review of recent rrends and challenges. *Sensors*, 2013.

[16] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th international conference on advances in geographic information systems*, pages 199–208. ACM, 2012.

[17] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. In *International Conference on Pervasive Computing*, pages 1–17. Springer, 2004.

[18] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[19] M. Berchtold, M. Budde, D. Gordon, H. R. Schmidtke, and M. Beigl. Actiserv: Activity recognition service for mobile phones. In *International Symposium on Wearable Computers (ISWC) 2010*, pages 1–8. IEEE, 2010.

[20] B. Bi, M. Shokouhi, M. Kosinski, and T. Graepel. Inferring the demographics of search users: social data meets search queries. In *Proceedings of the International Conference on World Wide Web*. ACM, 2013.

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

[22] T. Bocklet, A. Maier, and E. Noth. Age determination of children in preschool and primary school age with gmm-based supervectors and support vector machines/regression. *Text, Speech and Dialogue*, pages 253–260, 2008.

[23] R. Bracewell. The fourier transform and it's applications. *New York*, 1965.

[24] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. Maximizing modularity is hard. *arXiv preprint physics/0608255*, 2006.

[25] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[26] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] H. Brenner and U. Kliebsch. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, pages 199–202, 1996.

[28] K. Buraya, A. Farseev, A. Filchenkov, and T.-S. Chua. Towards user personality profiling from multiple social networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 2017.

[29] R. Burke. Integrating knowledge-based and collaborative-filtering recommender systems. In *Proceedings of the Workshop on AI and Electronic Commerce*, 1999.

[30] R. Caruana. Multitask learning. *Machine Learning*, 1997.

[31] H. Chahal, C. Fung, S. Kuhle, and P. Veugelers. Availability and night-time use of electronic entertainment and communication devices are associated with short sleep duration and obesity among canadian children. *Pediatric obesity*, 2013.

[32] 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.

[33] 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.

[34] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

[35] W.-y. S. Chou, A. Prestin, and S. Kunath. Obesity in social media: a mixed methods analysis. *Translational behavioral medicine*, 4(3):314–323, 2014.

[36] A. K. Chowdhury, A. Farseev, P. R. Chakraborty, D. Tjondronegoro, and V. Chandran. Automatic classification of physical exercises from wearable sensors using small dataset from non-laboratory settings. In *Proceedings of the first IEEE Life Sciences Conference*. IEEE, 2017.

[37] C. Christoudias, R. Urtasun, and T. Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.

[38] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[39] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[40] P. S. Collaboration et al. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. *The Lancet*, 373(9669):1083–1096, 2009.

[41] C. B. Corbin, G. Welk, W. R. Corbin, and K. Welk. *Concepts of Fitness and Wellness.* McGraw-Hill, 2001.

[42] A. Culotta. Estimating county health statistics with twitter. In *Proceedings of the Conference on Human Factors in Computing Systems (SIGCHI)*. ACM, 2014.

[43] H. A. De Gans and F. van Poppel. Contributions from the margins. dutch statisticians, actuaries and medical doctors and the methods of demography in the time of wilhelm lexis. In *workshop on 'Lexis in Context: German and Eastern& Northern European Contributions to Demography*, volume 1910, pages 1–24, 1860.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009*. IEEE, 2009.

[45] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov. Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. *IEEE Transactions on Signal Processing*, 2014.

[46] E. M. Eggleston and E. R. Weitzman. Innovative uses of electronic health records and social media for public health surveillance. *Current Diabetes Reports*, 2014.

[47] D. Eppstein, M. S. Paterson, and F. F. Yao. On nearest-neighbor graphs. *Disc. & Comp. Geometry*, 1997.

[48] D. Estival, T. Gaustad, S. B. Pham, W. Radford, and B. Hutchinson. Author profiling for English emails. *10th Conference of the Pacific Association for Computational Linguistics*, 2007.

[49] A. Farseev, M. Akbari, I. Samborskii, and T.-S. Chua. 360° user profiling: past, future, and applications. *ACM SIGWEB Newsletter*, (Summer), 2016.

[50] A. Farseev and T. Chua. Tweet can be fit: Integrating data from wearable sensors and multiple social networks for wellness profile learning. *ACM Transactions on Information Systems (TOIS)*, 2017.

[51] A. Farseev and T.-S. Chua. Tweetfit: Fusing multiple social media and sensor data for wellness profile learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI, 2017.

[52] A. Farseev, N. Gukov, I. Gossoudarev, and U. Zarichnyak. Cross-platform online venue and user community recommendation based upon social networks data mining. *Computer Instruments in Education*, 6:28–38, 2014.

[53] A. Farseev, D. Kotkov, A. Semenov, J. Veijalainen, and T.-S. Chua. Cross-social network collaborative recommendation. In *Proceedings of the 7th ACM International Conference on Web Science*. ACM, 2015.

[54] A. Farseev, L. Nie, M. Akbari, and T.-S. Chua. Harvesting multiple sources for user profile learning: a big data study. In *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 2015.

[55] A. Farseev, I. Samborskii, and T.-S. Chua. bbridge: A big data platform for social multimedia analytics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 759–761. ACM, 2016.

[56] A. Farseev, I. Samborskii, A. Filchenkov, and T.-S. Chua. Cross-domain recommendation via clustering on multi-layer graphs. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017.

[57] A. E. Field, E. H. Coakley, A. Must, J. L. Spadano, N. Laird, W. H. Dietz, E. Rimm, and G. A. Colditz. Impact of overweight on the risk of developing common chronic diseases during a 10-year period. *Archives of Internal Medicine*, 2001.

[58] J. L. Fischer. Social influences on the choice of a linguistic variant. 1958.

[59] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[60] F. Foerster, M. Smeja, and J. Fahrenberg. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Computers in Human Behavior*, 15(5):571–583, 1999.

[61] S. Fortunato. *Community detection in graphs*. Physics Reports, 2010.

[62] D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, and D. Bell. Analyzing the language of food on social media. In *Proceedings of the International Conference on Big Data*. IEEE, 2014.

[63] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 2002.

[64] B. Gaille. 26 great foursquare demographics. `http://brandongaille.com/26-great-foursquare-demographics/`, 2017.

[65] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[66] GlobalWebIndex. Gwi social report q3 2016. `http://insight.globalwebindex.net/social`, 2016.

[67] GlobalWebIndex. Social media fact sheet. `http://www.pewinternet.org/fact-sheet/social-media/`, 2017.

[68] S. Goel, J. M. Hofman, and M. I. Sirer. Who does what on the web: A large-scale study of browsing behavior. In *ICWSM*, 2012.

[69] G. H. Golub and C. F. Van Loan. *Matrix computations.* 2012.

[70] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[71] S. M. Grüsser, R. Thalemann, and M. D. Griffiths. Excessive computer game playing: evidence for addiction and aggression? *CyberPsychology & Behavior*, 10(2):290–292, 2006.

[72] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1979.

[73] D. M. Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[74] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 25–32, 2011.

[75] X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the ACM SIGIR Conference*, 2016.

[76] S. Helgason. The radon transform on euclidean spaces, compact two-point homogeneous spaces and grassmann manifolds. *Acta Mathematica*, 1965.

[77] T. K. Ho. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis & Applications*, 5(2):102–112, 2002.

[78] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th international conference on World Wide Web*, pages 151–160. ACM, 2007.

[79] R. Hu and P. Pu. Helping users perceive recommendation diversity. In *Proceedings of the Workshop on novelty and diversity in recommender systems, DiveRS. Chicago*, 2011.

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

[81] L. Jalali and R. Jain. Bringing deep causality to multimedia data streams. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 221–230. ACM, 2015.

[82] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the International Conference on Multimedia*. ACM, 2014.

[83] K. Jiang, D. Shao, S. Bressan, T. Kister, and K.-L. Tan. Publishing trajectories with differential privacy guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, page 12. ACM, 2013.

[84] I. Jolliffe. *Principal Component Analysis*. Wiley Library, 2002.

[85] D. Jurgens, J. McCorriston, and D. Ruths. An analysis of exercising behavior in online populations. In *Proceedings of the Int. Conference on Web and Social Media*. AAAI, 2015.

[86] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE transactions on information technology in biomedicine*, 14(5):1166–1172, 2010.

[87] Y. S. Kharin. Optimality and robustness in statistical forecasting. In *International encyclopedia of statistical science*, pages 1034–1037. Springer, 2011.

[88] E. Kim, S. Helal, and D. Cook. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 9(1):48–53, 2010.

[89] M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 2002.

[90] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.

[91] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[92] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys & Tutorials*, 15(3):1192–1209, 2013.

[93] Ó. D. Lara, A. J. Pérez, M. A. Labrador, and J. D. Posada. Centinela: A human activity recognition system based on acceleration and vital sign data. *Pervasive and mobile computing*, 8(5):717–729, 2012.

[94] R. B. Lehoucq and D. C. Sorensen. Deflation techniques for an implicitly restarted arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 1996.

[95] T. Li, M. Ogihara, and S. Ma. On combining multiple clusterings. In *Proceedings of the 13th ACM international conference on Information and knowledge management*, 2004.

[96] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 831–840. ACM, 2014.

[97] B. H. Lim, D. Lu, T. Chen, and M.-Y. Kan. # mytweet via instagram: Exploring user behaviour across multiple social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, pages 113–120. IEEE, 2015.

[98] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009.

[99] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum. Fortune teller: Predicting your career path. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.

[100] C. Lu, X. Chen, and E. Park. Exploit the tripartite network of social tagging for web clustering. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1545–1548. ACM, 2009.

[101] N. H. Lung and B. Institute. What are the health risks of overweight and obesity? `www.nhlbi.nih.gov/health/health-topics/topics/obe/risks`.

[102] H. Lutkepohl. Handbook of matrices. *Computational Statistics and Data Analysis*, 1997.

[103] M. Martinsen, S. Bratland-Sanda, A. K. Eriksson, and J. Sundgot-Borgen. Dieting to win or to be thin? a study of dieting and disordered eating among adolescent elite athletes and non-athlete controls. *British Journal of Sports Medicine*, 2010.

[104] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher. Activity recognition and monitoring using multiple sensors on different body positions. In *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*, pages 4–pp. IEEE, 2006.

[105] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber. # foodporn: Obesity patterns in culinary interactions. In *Proceedings of the 5th International Conference on Digital Health*. ACM, 2015.

[106] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*, 2014.

[107] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[108] N. Neubauer and K. Obermayer. Towards community detection in k-partite k-uniform hypergraphs. In *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, pages 1–9, 2009.

[109] M. E. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 2006.

[110] D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder. " how old do you think i am?" a study of language and age in twitter. 2013.

[111] N. P. Nguyen, T. N. Dinh, S. Tokala, and M. T. Thai. Overlapping communities in dynamic networks: their detection and mobile applications. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 85–96. ACM, 2011.

[112] A. S. Niaraki and K. Kim. Ontology based personalized route planning system using a multi-criteria decision making approach. *Expert Systems with Applications*, 2009.

[113] L. Nie, L. Zhang, M. Wang, R. Hong, A. Farseev, and T.-S. Chua. Learning user attributes via mobile social multimedia analytics. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3):36, 2017.

[114] I. Nikolaidis, K. Iniewski, B. Malhotra, H. Jeung, T. Kister, S. Bressan, and K.-L. Tan. Maritime data management and analytics. In *Building Sensor Networks: From Design to Applications*. CRC Press, 2013.

[115] A. Noulas, C. Mascolo, and E. Frias-Martinez. Exploiting foursquare and cellular data to infer user activity in urban environments. In *2013 IEEE 14th International Conference on Mobile Data Management*, volume 1, pages 167–176. IEEE, 2013.

[116] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Proceedings of the 12th International Conference on Data Mining (ICDM)*. IEEE, 2012.

[117] 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 (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153. Ieee, 2012.

[118] 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.

[119] G. S. O'Keeffe, K. Clarke-Pearson, et al. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804, 2011.

[120] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378. ACM, 2010.

[121] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 1994.

[122] K. Park, I. Weber, M. Cha, and C. Lee. Persistent sharing of fitness app status on twitter. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 2016.

[123] M.-H. Park, J.-H. Hong, and S.-B. Cho. Location-based recommendation system using bayesian user's preference model in mobile devices. In *International Conference on Ubiquitous Intelligence and Computing*, pages 1130–1139. Springer, 2007.

[124] D. Pelleg, A. W. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, 2000.

[125] M. Pennacchiotti and A.-M. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[126] J. W. Pennebaker. The secret life of pronouns. *New Scientist*, 211(2828):42–45, 2011.

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

[128] S. Pongpaichet, V. K. Singh, M. Gao, and R. Jain. Eventshop: recognizing situations in web data streams. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1359–1368. ACM, 2013.

[129] C. Pérez-Estruch, P. Roberto, and P. Rosso. Multimodal gender profile learning using neural networks. In *International Conference Recent Advances in Natural Language Processing*. RANLP, 2017.

[130] S. Qian, T. Zhang, R. Hong, and C. Xu. Cross-domain collaborative learning in social multimedia. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, 2015.

[131] Y. Qu and J. Zhang. Trade area analysis using user generated mobile location data. In *Proceedings of the 22nd International conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013.

[132] 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.

[133] F. Rangel, P. Rosso, M. Potthast, and B. Stein. Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. *Working Notes Papers of the CLEF*, 2017.

[134] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. Overview of the 3rd author profiling task at pan 2015. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, 2015.

[135] F. Rangel, P. Rosso, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, W. Daeleman, et al. Overview of the 2nd author profiling task at pan 2014. In *CEUR Workshop*. CEUR, 2014.

[136] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*, 2016.

[137] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the ACM Workshop on Search and Mining User-generated Contents*. ACM, 2010.

[138] S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3).

[139] D. Rhouma and L. B. Romdhane. An efficient algorithm for community mining with overlap in social networks. *Expert Systems with Applications*, 41(9), 2014.

[140] D. Riboni and C. Bettini. Cosar: hybrid reasoning for context-aware activity recognition. *Personal and Ubiquitous Computing*, 15(3):271–289, 2011.

[141] S. Russell, P. Norvig, and A. Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25:27, 1995.

[142] I. Samborskii, A. Filchenkov, G. Korneev, and A. Farseev. Person, organization, or personage: Towards user account type prediction in microblogs. 2016.

[143] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 1998.

[144] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, 2001.

[145] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

[146] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205, 2006.

[147] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[148] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti. Your installed apps reveal your gender and more! *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(3):55–61, 2015.

[149] S. S. Sharma and M. De Choudhury. Measuring and characterizing nutritional information of food and ingestion content in instagram. In *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015.

[150] J. Shi and J. Malik. Normalized cuts and image segmentation. *Transactions on Patt. Anal. and M. Intelligence*, 2000.

[151] S. Shrivastava. Studies in demography. 1980.

[152] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, M. Musolesi, and A. A. F. Loureiro. You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM*, 2014.

[153] X. Song, L. Nie, L. Zhang, M. Akbari, and T.-S. Chua. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2015.

[154] X. Song, L. Nie, L. Zhang, M. Liu, and T.-S. Chua. Interest inference via structure-constrained multi-source multi-task learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. ACM, 2015.

[155] E. Stamatatos, W. Daelemans, B. Verhoeven, P. Juola, A. López-López, M. Potthast, and B. Stein. Overview of the author identification task at pan 2014. In *CLEF (Working Notes)*, pages 877–897, 2014.

[156] S. A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867, 1940.

[157] W.-C. Su. Integrating and mining virtual communities across multiple online social networks: Concepts, approaches and challenges. In *4th International Conference on digital Inf. and comm-n Tech. and it's App.* IEEE, 2014.

[158] M. Swan. The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2):85–99, 2013.

[159] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[160] E. M. Tapia, S. S. Intille, W. Haskell, K. Larson, J. Wright, A. King, and R. Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart monitor. In *In: Proc. Int. Symp. on Wearable Comp.* Citeseer, 2007.

[161] J. C. Teixeira and A. P. Antunes. A hierarchical location model for public facility planning. *European Journal of Operational Research*, 2008.

[162] A. Tiroshi, S. Berkovsky, M. A. Kaafar, T. Chen, and T. Kuflik. Cross social networks interests predictions based on graph features. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013.

[163] S. Trewin. Knowledge-based recommender systems. *Encyclopedia of Library and Information Science: Volume 69-Supplement 32*, 2000.

[164] Twitter. The streaming apis. `https://dev.twitter.com/streaming/overview`. Online; accessed 3 Aug 2016.

[165] M. H. Veiga and C. Eickhoff. Privacy leakage through innocent content sharing in online social networks. *arXiv preprint arXiv:1607.02714*, 2016.

[166] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 2007.

[167] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the Workshop on Topic Models: Post-Processing and Applications*. ACM, 2015.

[168] K. Vorontsov, A. Potapenko, and A. Plavin. Additive regularization of topic models for topic selection and sparse factorization. In *Statistical Learning and Data Sciences*. Springer, 2015.

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

[170] D. Wagner and F. Wagner. *Between min cut and graph bisection*. Springer, 1993.

[171] 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.

[172] 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.

[173] R. L. Wasserstein and N. A. Lazar. The asa's statement on p-values: context, process, and purpose. *Am Stat*, 70(2):129–133, 2016.

[174] I. Weber and P. Achananuparp. Insights from machine-learned diet success prediction. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 540–551. World Scientific, 2016.

[175] W. H. O. WHO. Global database on body mass index. 2011. *Global Database on Body Mass Index*, 2011.

[176] P. Winoto and T. Tang. If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations. *New Generation Computing*, 2008.

[177] M. Yan, J. Sang, and C. Xu. Unified youtube video recommendation via cross-network collaboration. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 2015.

[178] D. Yarowsky and N. Garera. Modeling latent biographic attributes in conversational genres. In *Joint Conference of ACL and the International Joint Conference on Natural Language Processing*, 2012.

[179] M. Ye, K. Janowicz, C. Mülligann, and W.-C. Lee. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111. ACM, 2011.

[180] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM, 2011.

[181] J. Yin, Q. Yang, and J. J. Pan. Sensor-based abnormal human-activity detection. *IEEE Transactions on Knowledge and Data Engineering*, 20(8):1082–1090, 2008.

[182] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *Proceedings of the 18th International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2012.

[183] R. Zafarani and H. Liu. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 41–49. ACM, 2013.

[184] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512. ACM, 2015.

[185] 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.

[186] V. Zlatić, G. Ghoshal, and G. Caldarelli. Hypergraph topological quantities for tagged social networks. *Physical Review E*, 80(3):036118, 2009.

[187] I. Zukerman and D. W. Albrecht. Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):5–18, 2001.

Appendix A

7.3 Inter-Rater Agreement Measures

Cohen's Kappa

Cohen's Kappa coefficient [38] is a statistical measure of the degree of agreement between two independent raters (e.g. Rater A and Rater B) that takes into account the possibility that agreement could occur by chance. It is often used in machine learning as a complementary evaluation measure. The reason for its extensive use is its ability to measure robustness [87]. The robustness of the measure is achieved by taking into account the possibility of the agreement between classification results and ground truth labels occurring by chance. In the case of two independent raters (i.e. Rater A and Rater B), Cohen's Kappa is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{7.1}$$

where $p_o = \sum_i^k p_{ii}$ is the overall proportion of observed agreement among raters, and $p_e = \sum_i^k p_{i.}p_{.i}$ is the the overall proportion of chance-expected agreement (see Table 7.1 for notation details).

Table 7.1: Example of joint proportions of ratings by two raters on a scale with k categories

Rater B Rater A	1	2	...	k	Total
1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
....	...	...	...	...	...
k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	...	$p_{.k}$	1

Table 7.2: Landis and Koch Interpretation of Cohen's κ [91]

Kappa Statistics	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Weighted Cohen's Kappa

From the Cohen's Kappa description, it can be seen that κ statistic equally treats all disagreements between raters. In the case of binary classification ($k = 2$), such behavior does not affect the results. However, in the case when more than two categories are involved ($k > 2$), κ computation in its original form could be misleading due to the equal penalization of all disagreements between raters [39]. For example, in the case of age group classification, where the categories are ordered, misclassification into neighboring age groups (e.g. 20-30 and 30-40) must be penalized less than misclassification of completely different age groups (e.g. < 20 and > 40) [14]. To overcome the above shortcomings, Cohen et al. [39] further proposed the weighted version of the statistics κ_w, where the weight matrix $\mathbb{W} = (w_{i,j}) \in \mathbb{R}^{k \times k}$ is involved. $\mathbb{W}$'s main diagonal cells usually contain zeros, which represents perfect agreement between taters [39]. Off-diagonal cells contain weights indicating the seriousness of that disagreement. The weighted versions of proportions can be defined as $p_o(w) = \sum_i^k \sum_j^k w_{ij} p_{ij}$ and $p_e(w) = \sum_i^k \sum_j^k w_{ij} p_{i.} p_{.j}$ for weighted observed and chance-expected proportions, respectively [59]. The weighted κ statistics is defined as follows:

$$\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}, \tag{7.2}$$

Table 7.2 indicates the widely-adopted rater agreement labels proposed by Landis and Koch [91] for unweighted Cohen's Kappa in the Healthcare domain. The proposed labeling schema can be referred during inter-rater agreement assessment. However, assuming that the proposed labels were originally used to measure the disagreements between human raters and for the small populations (< 150 samples), their ranges are advisory but not obligatory for inter-rater agreement assessment.