Identification of Online Users' Social Status via Mining User-Generated Data

Dissertation
for the award of the degree

Doctor of Philosophy (Ph.D.)
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral Program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Tao Zhao

from Anhui, China
Göttingen, 2019

Thesis Committee:

Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Margarete Boos
Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Members of the Examination Board:

Reviewer:
Prof. Dr. Xiaoming Fu
Institut für Informatik, Georg-August-Universität Göttingen

Second Reviewer:
Prof. Dr. Marcus Baum
Institut für Informatik, Georg-August-Universität Göttingen

Further members of the Examination Board:
Prof. Dr. Winfried Kurth
Institut für Informatik, Georg-August-Universität Göttingen

Prof. Dr. Lutz M. Kolbe
Fakultät für Wirtschaftswissenschaften, Georg-August-Universität Göttingen

Prof. Dr. Margarete Boos
Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Carsten Damm
Institut für Informatik, Georg-August-Universität Göttingen

Date of the oral examination: 05. September 2019

Acknowledgement

I would like to express my gratitude to all those who helped me during my PhD study and the writing of this thesis. Without their help, I could not finish my PhD study and this PhD thesis.

First of all, I would like to extend my sincere gratitude to my supervisor Prof. Dr. Xiaoming Fu, for his constant encouragement and guidance. I really appreciate that he gave me countless instructive advice and useful suggestions during my PhD study. His conscientious academic spirit and dedicated attitude inspire and shape me both in my research and life. Without his consistent and valuable instruction, my thesis could not reach its present form.

Second, I owe my gratitude to my second supervisor, Prof. Dr. Margarete Boos, for her kind supervision and useful suggestions in the completion of this thesis.

I am also deeply indebted to Prof. Longbing Cao, who also guided me in the last year, for his valuable suggestions and illuminating instruction on my research work.

I would like to thank all the colleagues and visitors in the Computer Network Group in the University of Goettingen, especially Dr. Hong Huang, Dr. Sameer G Kulkarni, Dr. Yali Yuan, Mr. Yachao Shao. I appreciate that they gave me much help and advice during my PhD study and the whole process of my writing.

My thanks go to Prof. Dr. Marcus Baum for being a reviewer of my thesis. I also thank Prof. Dr. Winfried Kurth, Prof. Dr. Lutz M. Kolbe, Prof. Dr. Carsten Damm for serving as the examination board for my thesis.

My sincere thanks go to China Scholarship Council (CSC). Without the financial support from CSC, my PhD study is impossible. I am also deeply grateful to my Chinese and German friends for their kind support.

Last but definitely not least, I owe a great deal to my family and my parents for their unconditional and endless love and support, which inspires me to go on. Particularly, I would like to thank my wife, Jujie Qi, who forever cares, supports and encourages me.

Abstract

With the burst of available online user-generated data, identifying online users' social status via mining user-generated data can play a significant role in many commercial applications, research and policy-making in many domains. Social status refers to the position of a person in relation to others within a society, which is an abstract concept. The actual definition of social status is specific in terms of specific measure indicator. For example, opinion leadership measures individual social status in terms of influence and expertise in an online society, while socioeconomic status characterizes personal real-life social status based on social and economic factors. Compared with traditional survey method which is time-consuming, expensive and sometimes difficult, some efforts have been made to identify specific social status of users based on specific user-generated data using classic machine learning methods. However, in fact, regarding specific social status identification based on specific user-generated data, the specific case has several specific challenges. However, classic machine learning methods in existing works fail to address these challenges, which lead to low identification accuracy. Given the importance of improving identification accuracy, this thesis studies three specific cases on identification of online and offline social status. For each work, this thesis proposes novel effective identification method to address the specific challenges for improving accuracy.

The first work aims at identifying users' online social status in terms of topic-sensitive influence and knowledge authority in social community question answering sites, namely identifying topical opinion leaders who are both influential and expert. Social community question answering (SCQA) site, an innovative community question answering platform, not only offers traditional question answering (QA) services but also integrates an online social network where users can follow each other. Identifying topical opinion leaders in SCQA has become an important research area due to the significant role of topical opinion leaders. However, most previous related work either focus on using knowledge expertise to find experts for improving the quality of answers, or aim at measuring user influence to identify influential ones. In order to identify the true topical opinion leaders, we propose a topical opinion leader identification framework called QALeaderRank which takes account of both topic-sensitive influence and topical knowledge expertise. In the proposed framework, to measure the topic-sensitive influence of each user, we design a novel influence measure algorithm that exploits both the social and QA features of SCQA, taking into account social network structure, topical similarity and knowledge authority. In addition, we propose three topic-relevant metrics to infer the topical expertise of each user. The extensive experiments along with an online user study show that the proposed QALeaderRank achieves significant improvement compared with the

state-of-the-art methods. Furthermore, we analyze the topic interest change behaviors of users over time and examine the predictability of user topic interest through experiments.

The second work focuses on predicting individual socioeconomic status from mobile phone data. Socioeconomic Status (SES) is an important social and economic aspect widely concerned. Assessing individual SES can assist related organizations in making a variety of policy decisions. Traditional approach suffers from the extremely high cost in collecting large-scale SES-related survey data. With the ubiquity of smart phones, mobile phone data has become a novel data source for predicting individual SES with low cost. However, the task of predicting individual SES on mobile phone data also proposes some new challenges, including sparse individual records, scarce explicit relationships and limited labeled samples, unconcerned in prior work restricted to regional or household-oriented SES prediction. To address these issues, we propose a semi-supervised Hypergraph-based Factor Graph Model (HyperFGM) for individual SES prediction. HyperFGM is able to efficiently capture the associations between SES and individual mobile phone records to handle the individual record sparsity. For the scarce explicit relationships, HyperFGM models implicit high-order relationships among users on the hypergraph structure. Besides, HyperFGM explores the limited labeled data and unlabeled data in a semi-supervised way. Experimental results show that HyperFGM greatly outperforms the baseline methods on individual SES prediction with using a set of anonymized real mobile phone data.

The third work is to predict social media users' socioeconomic status based on their social media content, which is useful for related organizations and companies in a range of applications, such as economic and social policy-making. Previous work leverage manually defined textual features and platform-based user level attributes from social media content and feed them into a machine learning based classifier for SES prediction. However, they ignore some important information of social media content, containing the order and the hierarchical structure of social media text as well as the relationships among user level attributes. To this end, we propose a novel coupled social media content representation model for individual SES prediction, which not only utilizes a hierarchical neural network to incorporate the order and the hierarchical structure of social media text but also employs a coupled attribute representation method to take into account intra-coupled and inter-coupled interaction relationships among user level attributes. The experimental results show that the proposed model significantly outperforms other stat-of-the-art models on a real dataset, which validate the efficiency and robustness of the proposed model.

Contents

Chapter 1

Introduction

Social status identification is a special case of profiling problem. Social status indicates the position of a person as compared to others within an online or offline society, which is actually an abstract concept. In terms of specific measure indicator, the actual definition of social status is specific. For example, in an online society, opinion leadership indicates a person's position in terms of influence and expertise, while socioeconomic status characterizes the social and economic position of a person in the offline society, i.e., real-life society. Identifying personal social status can benefit many kinds of fields. On one hand, identifying social status can help companies to promote many commercial applications and services. On the other hand, for government and academia, it can offer rich valuable population information for study and policy-making in many domains, such as health, education, politics and economics. For example, opinion leader identification can improve the information and product recommendation for companies, and enable public opinion guidance for government. Socioeconomic status prediction can not only help companies to promote personalized services to target costumers, but also assist government in assessing personal wealth and economic development in an area.

Traditional method of social status identification is survey. Although the traditional survey method can get accurate information, manually conducting a large number of personal or household interviews in an area is highly expensive and time-consuming. Especially, for some small companies and some developing countries, it is very difficult for them to identify personal social status of a population. Fortunately, the burst of available online user-generated data offers a great opportunity to efficiently identify social status with low cost. The emergence and increasing popularity of diversified applications, mobile devices and information technologies, such as online social media, smart phones and Internet, attract billions of people all over the world to participant in online activities. According to [48], there are 4.39 billion online users in 2019 and 3.26 billion people use social media on mobile devices in January 2019. These online users generate massive amounts of various data every day, such as social media data, mobile phone usage data, and other online application data. In 2018, there are 2.5 quintillion bytes of data created in a day [61]. These user-generated data contain rich personal information, such as

spatio-temporal information, published posts, and other behaviors, which can reflect to some extend individual habits, life style, and other personal traits. Therefore, identification of online users' social status via mining user-generated data has become a significant and promising research area, which has attracted some attentions from data mining fields.

Existing data mining based methods [85, 69, 10, 55] usually leverage classic machine learning methods based on specific user-generated data to identify specific social status. Although these methods are much real-time, cheap and feasible, the identification accuracy of these methods is relatively low. It is worth mentioning that the identification accuracy is really important for many practical applications. For example, improving the accuracy of opinion leader identification can enhance the information and product recommendation efficiency and increase opinion influence which can make opinion spread faster and wider. Improving the accuracy of socioeconomic status prediction can help banks to reduce loan risk and improve loan amount assessment. However, regarding specific social status identification based on specific user-generated data, the specific case has several specific challenges. Classic machine learning methods, which are general methods, fail to address these specific challenges for specific case, which lead to the low identification accuracy. Therefore, in order to improve the identification accuracy for specific application case, specific methods need to be proposed to address specific challenges.

This thesis focuses on identifying online users' social status via mining user-generated data, which considers both online and offline social status identification. Although some efforts have been made, there still exist several specific challenges that need to be addressed for some specific cases in terms of different data source and application scenario. More specifically, the thesis studies three specific cases on the identification of online users' social status, which aim at addressing corresponding challenges to enhance the identification performance respectively:

- **Identifying topical opinion leaders based on social community question answering data.** This work aims at identifying online social status of users in terms of topic-sensitive influence and topic-relevant expertise in the social community question answering sites, namely identifying topical opinion leaders.

- **Predicting individual socioeconomic status based on mobile phone data.** The purpose of this work is to predict users' socioeconomic status in the offline society via mining their mobile phone Internet data.

- **Predicting individual socioeconomic status based on social media data.** The work focuses on predicting the real-life socioeconomic status of social media users via mining their social media content.

For these three specific works, the thesis proposes novel effective methods for identifying the specific social status of users based on their specific user-generated data as accurately as possible. Section 1.1 and Section 1.2 will elaborate the detailed motivation and main contributions of these three works respectively.

1.1 Motivation

In this section, the motivation of three specific works on the user social status identification in the thesis are elaborated in details respectively.

1.1.1 Identifying Topical Opinion Leaders based on Social Community Question Answering Data

Community Question Answering (CQA) site is a popular platform for information needs [67], where users can ask or answer questions and give comments to posts (i.e., questions and answers). Compared with traditional CQA sites like Yahoo!Answers [104] and Stack Overflow [88], Social Community Question Answering (SCQA) sites, an innovative type of CQA, have become more and more popular, such as Quora [73] and Zhihu [114], which provides social network function to connect users. As two most notable SCQA sites, Quora had around 190 million users in April 2017 and Zhihu had around 220 million users by the end of 2018. In these SCQA sites, users can follow each other to receive information updates from their followees according to their interests. This built-in social network function makes SCQA become an online social media platform [97]. In addition, most users usually publish and edit posts involving various topics, resulting in different topic domains. For specific topic(s), with the Question Answering (QA) and social functions of SCQA, active users tend to publish a great number of authoritative topic-related posts, which substantially affect other users' opinions, and even guide public opinion direction. They play an important role in creating topic-related knowledge repositories, maintaining the activeness of the topic community, and even helping to controlling the development trend of public opinions on the Internet.

However, most existing researches mainly focus on the identification of general *opinion leaders*, who give influential comments and opinions, put forward guiding ideas, agitate and guide the public to understand social problems [56]. The original concept of opinion leaders ignores their specialty, which deviates from the reality in current SCQA sites. For example, Lady Gaga may be an opinion leader in the topic "music" instead of "science". Nowadays, the precision application forces us to get to know the leader in each specific field, which brings the problem - *the identification of topical opinion leader*. Compared with opinion leader who is topic-irrelevant within the field of sociology, the work refers to these active users in specific topic domains of SCQA sites as *topical opinion leaders*.

Due to the great significance of identifying topical opinion leaders, the work in the thesis mainly focuses on identifying and analyzing topical opinion leaders in SCQA sites. Despite the important role that topical opinion leaders play in SCQA, the challenge of identifying topical opinion leaders is still intractable. According to the characteristics of topical opinion leaders, a major challenge is how to identify users who have both *strong topic-sensitive influence* and *high topic-relevant knowledge expertise* in given topic(s). Most existing works either focus on the knowledge expertise to find experts for improving the quality of answers in QA sites [68,

76, 113] or mainly aim at measuring the user influence to identify influential users in social networks [13, 58, 63, 100].

In Chapter 2, a novel topical opinion leader identification method is proposed and introduced in details, which can take into account topic-sensitive influence and topic-relevant knowledge expertise in SCQA sites.

1.1.2 Predicting Individual Socioeconomic Status based on Mobile Phone Data

Socioeconomic Status (SES) is an indicator that measures an individual, a household or a region's economic and social position in relation to others, which is typically divided into three levels (high, middle, and low) [84]. The rich information carried by SES not only helps governments and research institutes study and make public policies, but also assists in meeting the needs of target clients by evaluating their purchasing power from a commercial perspective. Furthermore, SES can benefit a wide range of other fields, such as health [71, 103], education [82] and public transportation [19]. National statistical offices measure socioeconomic information typically by a large number of personal or household interviews. However, assessing SES for a whole country or region's population by this traditional method is extremely expensive and time-consuming. For example, the nationwide census for calculating SES are usually done every 5 to 10 years and is impossible for some developing countries due to the high cost. It is critical to develop a low-cost means for timely capturing and accurately assessing individual SES in a population.

Due to the worldwide ubiquity of smart phones, mobile phone data captures abundant information regarding personal social attributes, relation networks and mobility patterns in a large-scale population, which to some extent reflects SES. In view of this, mobile phone data has been used as a novel data source for efficiently inferring SES with low cost. Some efforts have been made to infer regional or household SES from mobile phone data by directly applying classic supervised machine learning methods [10, 44, 87]. Different from most existing works that concentrate on aggregated records of a region or household, this work is motivated to study the SES prediction on mobile phone data at an individual level, the first trial in the community as far as we know. Intuitively, even living in the same household, individuals probably share different SES levels. Inferring the individual SES provides the finest level of evidence and indication to improve the quality of corresponding public policies-making. Furthermore, it can enable numerous fine-grained applications at an individual level, such as precision marketing, fine service and assessment. However, the problem of individual SES prediction based on mobile phone data proposes three main challenges:

- **Sparse individual records.** Compared with aggregated records of a region or household, a large portion of individual mobile phone users actually generate sparse valid usage records every day. With the ubiquity of WiFi, individual records that telco service

providers can identify are becoming rarer. For example, 71.9% users generate less than two valid daily records in the data provided by an Internet Service Provider (ISP) in China. It is difficult to explore enough information from sparse individual records for revealing personal SES as done in the existing SES prediction work, thus causing poor prediction performance.

- **Scarce explicit relationships.** Due to the increasing popularity of mobile communication applications like WhatsApp [101] and Wechat [99], an increasing number of mobile phone users are giving up traditional voice calling and Short Message Service (SMS) [1]. Subsequently, the communication relationships built in these mobile applications are disconnected from ISP-provided mobile phone data. Therefore, explicit relationships among users extracted from mobile phone records become scarce, which makes the methods based on such relationships failed to work.

- **Limited labeled samples.** Since the cost of assessing individual SES by existing methods is extremely high, it is rather difficult to obtain enough SES-labeled samples for learning models. To the best of our knowledge, most prior works on the SES prediction only employ typical supervised learning methods to predict SES, which do not work well with limited labeled samples.

In Chapter 3, the thesis presents a semi-supervised probabilistic hypergraph based factor graph model for the individual SES prediction problem, which can address the above challenges.

1.1.3 Predicting Individual Socioeconomic Status based on Social Media Data

Predicting individual socioeconomic status (SES) from social media content recently has become an important research area. As an access to financial, social and human capital resources, inferring individual SES not only provides governments and research organizations with tools for studying and make public policies on a large scale population, but also helps promote online marketing and advertising by the analysis of user's purchasing power. It also benefits a wide range of other fields, such as education [103, 71], health [82] and public transportation [19]. With the worldwide ubiquity of online social media like Twitter, Facebook and Sina Weibo, online social media content has been used in recent research for population informatics in demographics [75, 15, 36], economics [11], social science [92, 55] and other research domains [24, 53, 54]. In consideration of the significance of SES and the ubiquity of social media applications, this work aims at predicting the SES of social media users based on their social media content. For the generalization, this work regards posted text (called social media text in the work) and platform-based user level attributes (e.g., the number of followers, the number of followees, etc) as social media content of a user since these data are ubiquitous on social media.

Previous related work have looked into predicting individual socioeconomic information based on social media content, such as inferring occupation category [69], SES [55] and income [70] of social media users. In these works, they devote to manually design several kinds of user level attributes and textual features, such as n-grams, from social media text, and then feed all the features into a machine learning based classifier for prediction. However, the prediction performance of these models heavily depends on the extracted features, which need effective feature engineering. Furthermore, existing methods ignore the following important information for the social media content representation.

- **Order of social media text.** Previous approaches on socioeconomic information prediction represent social media text with sparse lexical features, such as n-grams, or word embedding based features, such as neural clusters [69]. These predefined textual features cannot capture the order of social media text, which is an important information for representing long text sequence. For the microblogging that our work focuses on, the orders among words and microblogs are ignored.

- **Structure of social media text.** Previous related work directly extract user level textual features from aggregated social media text of each user. However, in fact, the social media text of each user has a hierarchical structure. For the microblogging that our work focuses on, words form microblogs, microblogs form social media text of a user. Therefore, the user level textual features ignore the hierarchical structure, which lead to information loss.

- **Relations among user level attributes.** In the real world, attributes are more or less interacted and coupled via explicit or implicit relationships [96]. For example, business and social applications always see quantitative attributes coupled with each other [18]. However, the previous work extract the user level attributes without considering relations among them, which leads to limited performance.

Chapter 4 introduces a coupled social media content representation learning model for improving the performance of individual SES prediction, which jointly considers coupled relationships among the social media text and user level attributes.

1.2 Dissertation Contributions

This section describes the main contributions of three works on the users' social status identification in the thesis.

1.2.1 Identifying Topical Opinion Leaders based on Social Community Question Answering Data

To address the challenges mentioned in Section 1.1.1, this thesis proposes a topical opinion leader identification algorithm called QALeaderRank for SCQA sites, which alleviates these shortcomings by simultaneously incorporating the *topic-sensitive influence* and the *topic-relevant knowledge expertise*. To be more specific, in order to measure the true topic-sensitive influence of users, the work proposes a novel influence measure algorithm called QARank which exploits both the *social* and *QA* features of SCQA. Two key challenges are addressed to build QARank: i) inferring the topic interest and the knowledge authority of each user from its published posts; ii) confirming the existence of *homophily* in SCQA sites, which implies that a user follows another user owing to their similar topic interests. Based on this, QARank not only takes account of the social network structure and the topical similarity between users like traditional influence measure methods (e.g., TwitterRank [100]), but also considers the topical knowledge authority. Besides, to measure the topical knowledge expertise of each user, the work proposes three topic-relevant metrics that account for knowledge capacity, satisfaction and contribution. Moreover, regarding the popularity of multi-topic, the proposed QALeaderRank can be utilized to identify multi-topic opinion leaders.

In this work, employing a dataset crawled from Zhihu as the basis of this study, a comprehensive analysis on the QA and social features of SCQA is first given. In order to validate the efficiency of the proposed model, we conduct an extensive evaluation for the proposed QALeaderRank with this dataset across the most popular ten topics in Zhihu. The experimental results, along with an online user study, show that QALeaderRank achieves significant improvement compared with the related state-of-the-art methods.

In addition, we further analyze and predict the topic interest change behaviors of users, especially topical opinion leaders, which is of great importance for many applications, such as answerer and topic recommendation. To this end, we try to answer two key questions: 1) how the user topic interest changes; 2) whether the user topic interest is predictable. Based on several analysis and experiments, we detect the change patterns of user topic interests and examine the predictability of user topic interest.

The main contributions of this work can be summarized as follows:

- We analyze the social and QA features of SCQA and confirm the existence of *homophily* in the context of SCQA.
- To the best of our knowledge, we are the first to propose an efficient algorithm called QALeaderRank to tackle the issue of topical opinion leader identification in SCQA.

- To design QALeaderRank, we propose a novel topic-sensitive influence measure algorithm for SCQA, based on the QA and social features. Additionally, we define three topic-relevant metrics to measure topical expertise.
- With extensive experiments and an online user study, we demonstrate our proposed algorithm greatly outperforms the baseline methods.
- We analyze the topic interest change behaviors of users over time and examine the predictability of user topic interest through further experiments.

1.2.2 Predicting Individual Socioeconomic Status based on Mobile Phone Data

To simultaneously address the above challenges mentioned in Section 1.1.2 for enabling individual SES prediction based on mobile phone data, this work proposes a novel semi-supervised probabilistic model called Hypergraph-based Factor Graph Model (HyperFGM). First, to reduce the performance loss caused by the individual record sparsity, leveraging the idea of factor graph model, HyperFGM utilizes customized factor functions to efficiently capture the correlations between SES and numerous attributes of users extracted from individual mobile phone records, which significantly exploits the power of sparse records compared with the prior methods on SES prediction. Second, to address the explicit relationship scarcity problem, HyperFGM leverages the advantage of hypergraph on high-order relationship modeling to model implicit high-order relationships among users based on the hypergraph structure, which avoids the performance loss caused by ignoring the implicit high-order relationships. Third, for handling the limited labeled samples, HyperFGM explores both labeled and unlabeled data on a hypergraph network in a semi-supervised way, thereby achieving better performance than supervised learning methods in prior SES prediction work.

Furthermore, compared with the proposed hypergraph-based factor graph model, traditional hypergraph-based models [33, 80, 115], focusing on the relationships among objects, need to convert the numerous attributes of objects into various relationships among objects, causing conversion loss. Traditional factor graph models [91, 95, 105] only consider objects' attributes and explicit pair-wise relationships between objects in a simple graph, which ignore implicit and high-order relationships among objects. However, in fact, there are many high-order relationships among objects [115] while implicit relationships exist among objects. Therefore, in order to solve the disadvantages of these two traditional methods, HyperFGM, combining hypergraph-based model and factor graph model into one model, predicts individual SES by not only directly considering the SES-related attributes of users but also modeling the implicit high-order mobility pattern-based relationships among users in the hypergraph structure.

We demonstrate the feasibility and power of HyperFGM on individual SES prediction using a set of anonymized real mobile phone data collected from a major ISP in China. Experimental results indicate that HyperFGM outperforms previous work on SES prediction by 5-22% w.r.t.

the F1-score and provides a considerable improvement (2-9%) compared with the state-of-the-art hypergraph-based methods and factor graph methods. It is worth to note that the proposed HyperFGM is a general semi-supervised classification method, which can be applied not only to the SES prediction problem but also to other similar tasks.

The major contributions in this work are summarized as follows.

- We first identify the issue of predicting individual SES from mobile phone data. To the best of our knowledge, no previous work has extensively studied this issue.
- We propose a semi-supervised probabilistic hypergraph model, HyperFGM, to solve the individual SES prediction problem, which jointly considers user attributes and implicit high-order relationships among users based on the hypergraph structure.
- We apply the proposed model on a collection of anonymized real mobile phone data. Experimental results show that HyperFGM outperforms the state-of-the-art baseline models.

1.2.3 Predicting Individual Socioeconomic Status based on Social Media Data

Motivated by the great success of deep learning in many fields, such as computer vision [52] and natural language processing [6], recent works utilize neural networks to learn text representation without any feature engineering and mostly achieve significantly higher performance compare with traditional machine learning methods. Inspired by this, to address the mentioned challenges in Section 1.1.3, this work proposes a coupled social media content representation learning model for individual SES prediction, utilizing neural network to represent social media content, which is the first trial in this community as far as we know. First, in order to be able to consider the order of words and microblogs in social media text, this work proposes to employ Bidirectional Long Short-Term Memory (BiLSTM) network, a variation of Recurrent Neural Network (RNN), to represent social media text due to its representational power and effectiveness at capturing long-term dependencies of a sequence. Second, since social media text have a hierarchical structure, the work likewise constructs a social media text representation by first building representations of microblogs with the corresponding words and then aggregating those into a social media text representation. Third, to consider the dependency of platform-based user level attributes, this work devises a coupled attribute representation to represent user level attributes, using intra-coupled interaction (i.e., the correlations between attributes and their own powers) and inter-coupled interaction (i.e., the correlations between attributes and the powers of others) [96]. Finally, we learn a joint coupled social media content representation with aggregating social media text representation and platform-based user level attribute representation.

We focus this work on the microblogging platform of Sina Weibo [81], a Chinese microblogging website, and build a new data set of Sina Weibo users with a SES label for each of them. To demonstrate the feasibility and efficiency of the proposed model on individual SES prediction, the proposed model is applied to the data set. Experimental results demonstrate that the proposed model significantly outperforms the baseline models in previous related work.

To sum, the main contributions of this work are as follows:

- We propose a novel coupled social media content representation framework for the individual SES prediction, which utilizes neural network and coupled representation method to integrate social media text and platform-based user level attributes. To our best knowledge, this is the first try in this community.
- We present a social media text representation method, which utilizes hierarchical recurrent neural network to take into account the order of words and microblogs as well as the hierarchical structure of social media text.
- We employ a coupled attribute representation method to analyze the intra-coupled and inter-coupled interaction among user level attributes, which can successfully capture the intrinsic couplings for SES prediction.
- We build a data set of Sina Weibo users with a SES label for each of them and demonstrate the power of our proposed model using this data set. Substantial experiments demonstrate that our model significantly outperforms the state-of-the-art models.

1.3 Dissertation Structure

This dissertation contains part of the content of the following published and submitted papers.

- Tao Zhao, Hong Huang, and Xiaoming Fu. Identifying Topical Opinion Leaders in Social Community Question Answering. In International Conference on Database Systems for Advanced Applications, pp. 372-387. Springer, Cham, 2018. DOI: 10.1007/978-3-319-91452-7_25
- Tao Zhao, Yachao Shao, Hong Huang, Baosheng Wang and Xiaoming Fu. "Identification and Analysis of Topical Opinion Leaders in Social Community Question Answering." Information Retrieval Journal. 2019. (Under review)
- Tao Zhao, Hong Huang, Xiaoming Yao, Jar-der Luo, and Xiaoming Fu. Predicting Individual Socioeconomic Status from Mobile Phone Data: A Semi-supervised Hypergraph-based Factor Graph Approach. International Journal of Data Science and Analytics. 2019. DOI: 10.1007/s41060-019-00195-z

The contents of this dissertation are organized as follows:

- Chapter 1 provides an overview of this thesis: introducing the motivation of this study, stating main contributions of this dissertation regarding the targeted problems, and presenting the structure of this thesis.

- Chapter 2 presents a novel topical opinion leader identification framework for social community question answering sites, which takes account of both the topic-sensitive influence and the topical knowledge expertise. To be more specific, Section 2.1 introduces the motivation and contributions of this work. In Section 2.2, we briefly review the related work. Section 2.3 describes data collection and initial analysis on Zhihu dataset. Section 2.4 details the proposed algorithm called QALeaderRank. Section 2.5 evaluates the performance of QALeaderRank with extensive experiments and an online user study. Section 2.6 gives an analysis on the topic interest change behaviors of users. Finally Section 2.7 concludes this work in this chapter.

- Chapter 3 proposes a novel semi-supervised probabilistic model called Hypergraph-based Factor Graph Model (HyperFGM) for enabling individual socioeconomic status prediction based on mobile phone data. More specifically, Section 3.1 first gives the description about the motivation and contributions. Section 3.2 discusses the related work on socioeconomic information analysis and prediction. Section 3.3 shows the data collection. The detailed description of the proposed HyperFGM model is presented in Section 3.4, which is composed of user attribute extraction, mobility pattern-based hypergraph construction and model description for individual SES prediction. Section 3.5 evaluates the prediction performance of HyperFGM with extensive experiments. Finally, Section 3.6 summarizes this chapter.

- Chapter 4 studies predicting individual socioeconomic status from social media content. To this end, the chapter proposes an efficient coupled social media content representation model for individual SES prediction, which not only utilizes a hierarchical neural network to incorporate the order and the hierarchical structure of social media text but also employs a coupled attribute representation method to take into account intra-coupled and inter-coupled interaction relationships among platform-based user level attributes. The motivation and contributions of the work in this chapter are firstly introduced in Section 4.1. Then, in Section 4.2, the state-of-the-art related work are reviewed, including socioeconomic-related information prediction based on social media data and representation learning of social media content. Section 4.3 describes the proposed model in details. In Section 4.4, the data collection and preprocessing are introduced. The efficiency and robustness of our proposed model are demonstrated with experimental evaluation in Section 4.5. Finally, Section 4.6 concludes the Chapter 4.

- Chapter 5 concludes the work in this dissertation and gives an outlook of the future research work with regard to the proposed methods of this dissertation.

Chapter 2

Identifying Topical Opinion Leaders based on Social Community Question Answering Data

Social community question answering (SCQA), an innovative and popular community question answering site, not only provides traditional question answering (QA) services but also allows users to follow each other. Regarding the important role of topical opinion leaders in SCQA, this chapter focuses on studying the problem of topical opinion leader identification based on SCQA data. Nevertheless, most existing works either aim at using knowledge expertise to find experts for improving the quality of answers, or measure user influence to identify influential ones. Identifying topical opinion leaders in SCQA sites has not been well investigated.

The chapter will introduce a novel topical opinion leader identification framework, taking account of both the topic-sensitive influence and the topical knowledge expertise. In the proposed framework, to measure the topic-sensitive influence of each user, we design a novel influence measure algorithm that exploits both the social and QA features of SCQA, considering social network structure, topical similarity between users and knowledge authority. To infer the topical expertise of each user, we define three topic-relevant metrics. We demonstrate that the proposed model significantly outperforms the state-of-the-art methods with extensive experiments and an online user study. Furthermore, we analyze the topic interest change behaviors of users over time and examine the predictability of user topic interest through further experiments.

Contents

2.1 Introduction

As an innovative type of community question answering (CQA) site, social community question answering provides social network function to connect users besides offering traditional question answering services. In these SCQA sites, users can follow each other to receive information updates from their followees according to their interests. This built-in social network function makes SCQA become an online social media platform [97]. Besides, most users usually publish and edit posts involving various topics, resulting in different topic domains. For specific topic(s), with the question answering (QA) and social functions of SCQA, active users tend to publish a great number of authoritative topic-related posts, which substantially affect other users' opinions, and even guide public opinion direction. In the light of the original concept of *opinion leader*, opinion leaders give influential comments and opinions, put forward guiding ideas, agitate and guide the public to understand social problems [56], who is topic-irrelevant within the field of sociology. We refer to these active users in specific topic domains of SCQA sites as *topical opinion leaders*. As topical opinion leaders, they play an important role in creating topic-related knowledge repositories, maintaining the activeness of the topic community, and even helping to controlling the development trend of public opinions on the Internet. Therefore, it is of great significance to identify and analyze topical opinion leaders in SCQA sites.

In this chapter, we mainly study identifying topical opinion leaders in SCQA sites. Most existing works either focus on the knowledge expertise to find experts for improving the quality of answers in QA sites [68, 76, 113] (see Zone I+IV in Figure 2.1) or mainly aim at measuring the user influence to identify influential users in social networks [13, 58, 63, 100] (see Zone I+II in Figure 2.1). According to the characteristics of topical opinion leaders, a major challenge in this work is how to identify users who have both *strong topic-sensitive influence* and *high topic-relevant knowledge expertise* in given topic(s), as shown in Figure 2.1.

To solve this problem, we propose a topical opinion leader identification algorithm called QALeaderRank for SCQA sites, which alleviates these shortcomings by simultaneously incorporating the *topic-sensitive influence* and the *topic-relevant knowledge expertise*. In order to measure the true topic-sensitive influence of users, we propose a novel influence measure algorithm called QARank which exploits both the *social* and *QA* features of SCQA. Two key challenges are addressed to build QARank: i) inferring the topic interest and the knowledge authority of each user from its published posts; ii) confirming the existence of *homophily* in SCQA sites, which implies that a user follows another user owing to their similar topic interests. Based on this, QARank not only takes account of the social network structure and the topical similarity between users like traditional influence measure methods (e.g., TwitterRank [100]), but also considers the topical knowledge authority. Besides, to measure the topical knowledge expertise of each user, we propose three topic-relevant metrics that account for knowledge capacity, satisfaction and contribution. Moreover, regarding the popularity of multi-topic, the proposed QALeaderRank can be utilized to identify multi-topic opinion leaders.

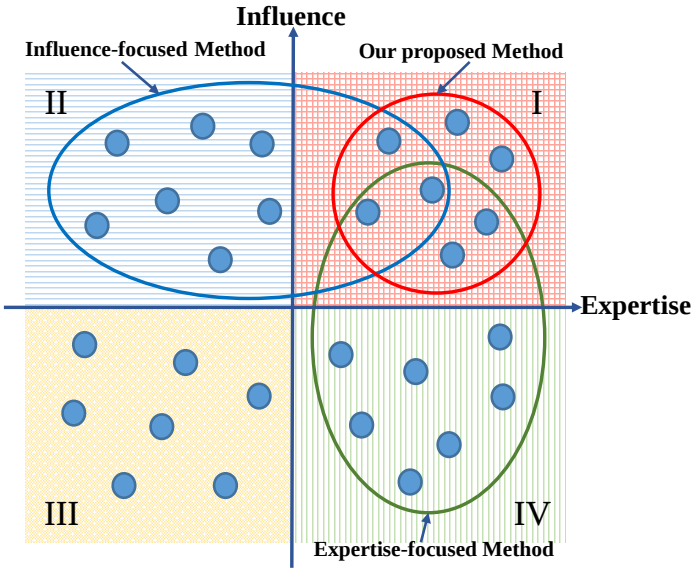


Figure 2.1: User identification in terms of influence & expertise.

In this chapter, we conduct an extensive evaluation for our proposed QALeaderRank with a set of real dataset crawled from Zhihu. The experimental results, along with an online user study, show that QALeaderRank achieves significant improvement compared with the related state-of-the-art methods. In addition, we further analyze and predict the topic interest change behaviors of users, especially topical opinion leaders, which is of great importance for many applications, such as answerer and topic recommendation. To this end, we try to answer two key questions: 1) how the user topic interest changes; 2) whether the user topic interest can be predictable. Based on several analysis and experiments, we detect the change patterns of user topic interests and examine the predictability of user topic interest.

The main contributions of this work can be summarized as follows:

- We analyze the social and QA features of SCQA and confirm the existence of *homophily* in the context of SCQA.
- To the best of our knowledge, we are the first to propose an efficient algorithm called QALeaderRank to tackle the issue of topical opinion leader identification in SCQA.
- To design QALeaderRank, we propose a novel topic-sensitive influence measure algorithm for SCQA, based on the QA and social features. Additionally, we define three topic-relevant metrics to measure topical expertise.
- Through extensive experiments and an online user study, we demonstrate our proposed algorithm greatly outperforms the baseline methods.
- We analyze the topic interest change behaviors of users over time and examine the predictability of user topic interest through further experiments.

The rest of the chapter is organized as follows: In Section 2.2 we review the related work. Section 2.3 describes data collection and initial analysis on Zhihu dataset. Section 2.4 details the proposed algorithms. Section 2.5 evaluates the performance of QALeaderRank with extensive experiments and an online user study. Section 2.6 analyzes the topic interest change behaviors of users, and finally we conclude this chapter in Section 2.7.

2.2 Related Work

Due to the great importance of opinion leader, in the field of sociology, a great number of sociologists have studied to understand the concept and characteristics of opinion leaders [21, 66, 17, 77]. In this section, we mainly focus on previous related work on online communities and social media and give a summary of them, which can be divided into two main kinds of methods: expertise-focused method and influence-focused method.

2.2.1 Expertise-focused Method

Most previous works on CQA sites mainly aim at studying expert identification for the purpose of improving the quality of answers. For example, Bouguessa et al. [14] proposed a probabilistic approach based on a mixture model. The method identified which experts would answer open questions based on the number of best answers published by users in a large-scale community question answering site Yahoo!Answers. Riahi et al. [76] focused on finding experts for a newly posted question through investigating and comparing the suitability and performance of statistical topic models in the Stackoverflow website. Zhou et al. [113] developed a novel graph-regularized matrix completion algorithm for inferring the user model, thus improving the performance of expert finding in CQA systems.

With the increasing popularity of the SCQA sites, the issue of identifying important users in SCQA sites has started to draw research interests. Song et al. [85] proposed a leading user detection model for Quora, which takes into account the authority, activity and influence of each user. However, the user influence in this model is measured by its node in-degree in the social network, namely the number of followers, which cannot accurately capture the notion of influence in social networks [38, 51]. In addition, all the factors in this model are topic-irrelevant.

2.2.2 Influence-focused Method

There are also a great number of works that study the issue of opinion leader or influential user identification in social media, which mainly focus on the influence of users. For the Bulletin Board System (BBS), Zhai et al. [110] proposed interest-field based algorithms taking into account the network structure and user's interest to identify opinion leaders. For the blogosphere, Song et al. [86] proposed a novel opinion leader identification algorithm considering the importance and novelty of published blogs. Li et al. [58] proposed a framework to identify

Figure 2.2: A screen capture of user home page in Zhihu.

opinion leaders based on the information retrieved from blog contents, authors, readers and their relationships. In the microblogging sites, especially Twitter, there are amounts of works on identifying influential users [5, 20, 35, 57, 100]. One representative work is TwitterRank algorithm [100], an extension of PageRank algorithm [38]. TwitterRank is proposed to identify topic-sensitive influential users in Twitter considering both the topical similarity between users and the link structure among users. In general, most approaches mainly focus on measuring the user influence, which fail to identify topical opinion leaders in SCQA as SCQA users disseminate information by both the following relationship and the QA function.

To sum up, identifying topic-sensitive opinion leaders in SCQA has not been well investigated. To tackle this problem, we propose a topical opinion leader identification algorithm considering the topical knowledge expertise and the topical influence in the social network.

2.3 Dataset Collection and Analysis

In this section we first describe the dataset collection and then present some initial analysis of the QA and social features in SCQA sites.

2.3.1 Dataset Collection

Zhihu, as a Chinese SCQA site, has become more and more popular. The work in this chapter takes Zhihu as a case study. We collected the Zhihu dataset through web-based parallel crawls. More specifically, we started user crawls using a set of 10 popular Zhihu users. The crawls follow a Breadth-First Search (BFS) pattern through the following links of each user. Finally, we totally crawled 1.41M+ individual users from Zhihu. As shown in Figure 2.2, each user data contains the user ID, the user's followers and followees, the answers and questions posted by the user. As shown in Figure 2.3, for each question, we crawled its topics (i.e., the topic tags of each question added by its author). For each answer, we crawled its received vote

Figure 2.3: A screen capture of question and answer in Zhihu.

Table 2.1: Data summary.

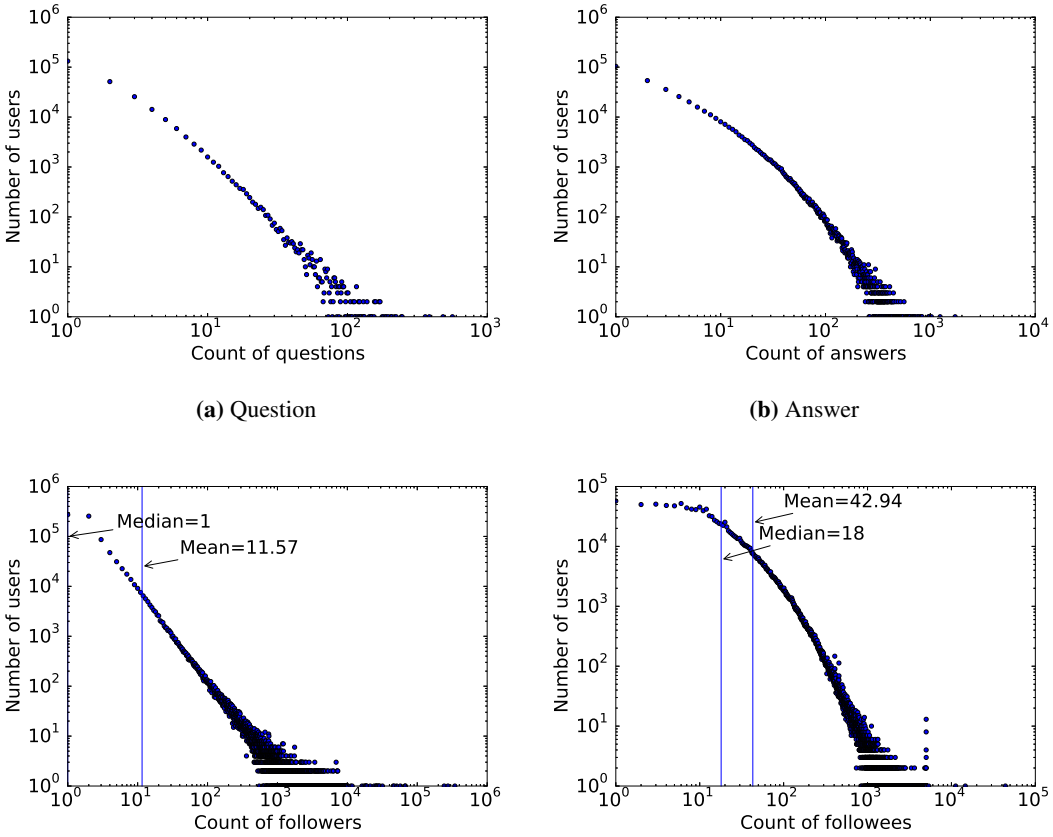
Total number of users	1,411,669
Total number of questions	701,982
Total number of answers	4,047,183
Total number of topics	160,664
Average number of followers per user	11.57
Average number of followees per user	42.94
Average number of votes per user	39.08
Average number of votes per answer	13.63

count and its corresponding question's topics. As illustrated in Table 2.1, these users posted 701K+ unique questions and 4.04M+ unique answers in total.

According to the top-down tree-like topic structure provided by Zhihu, we crawled all the unique topics in Zhihu. In the topic structure, there is only one root topic which has 6 child topics but no parent topic. Except the lowest level topics (i.e., leaf topics), the other topics have at least one parent topic and one child topic. For instance, the topic "Fitness" has two parent topics "Sport" and "Health" while it has 31 child topics, such as "Muscle", "Bodybuilding" and so on. As shown in Table 2.1, we totally obtained 160K+ unique topics in Zhihu.

2.3.2 Initial Analysis

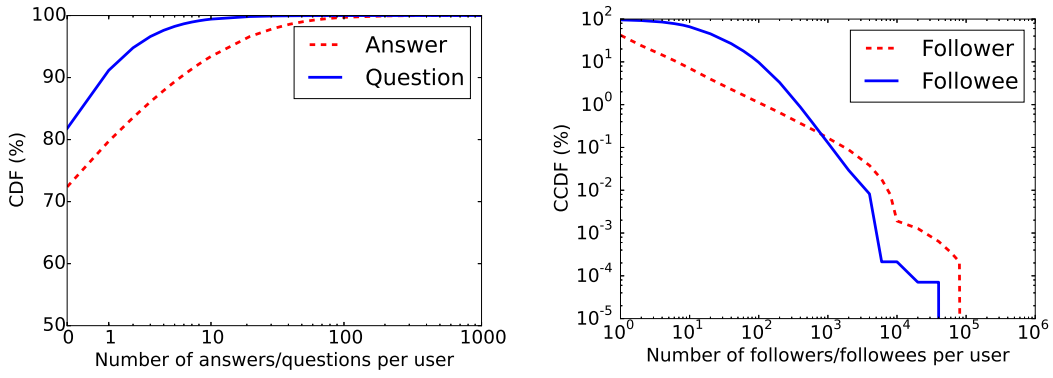
To explore the QA and social features of SCQA sites, we first present some initial analysis based on our crawled data, including the distributions of questions, answers, followers and followees. With this analysis, we find that the QA and social features of Zhihu are similar to those of Quora studied in [97].



(a) Question **(b)** Answer

(c) Follower **(d)** Followee

Figure 2.4: Power law distribution of QA and following in Zhihu.



(a) Question and answer **(b)** Follower and followee

Figure 2.5: Distribution of QA and following in Zhihu.

Questions and Answers. One main function of SCQA is to allow users ask and answer questions. In order to explore the QA features of SCQA, Figure 2.4a and Figure 2.4b show that the distributions of the number of questions and answers posted by each user, which follow power-law distribution. This means that a small portion of users posted a great number of questions or answers while most users posted a few ones. As shown in Figure 2.5a, we can observe that 81% of the users did not ask any question and 72% of the users did not give any answer, which conforms to 80/20 rule.

Followers and Followees. SCQA constructs a directed social network where users can follow each other. To explore the social feature, we analyze the number of users' followers and followees in Zhihu. Figure 2.4c and Figure 2.4d plot the distributions of the number of followers and followees per user, which also follow power-law distribution. The exponential fitting parameter α for the follower count distribution is 1.84 with standard error 0.001, which is close to that of Twitter (α=2.28) [97]. The average numbers of followers and followees per user are around 12 and 43. As illustrated in Figure 2.5b, about 38% of users have no follower and more than 99% of users have followees. This observation implies that Zhihu is a relatively dense social network like Twitter.

2.4 Topical Opinion Leader Identification Framework

This work mainly aims at identifying topical opinion leaders, who have both strong topic-sensitive influence and high topic-relevant knowledge expertise in SCQA sites. To measure the true topic-sensitive influence, we propose QARank algorithm in Section 2.4.1. To measure the topical expertise, we present three topic-relevant expertise metrics in Section 2.4.2. Based on these two factors, we build a topical opinion leader identification algorithm called QALeaderRank. With the consideration of combining both the topic-sensitive influence and the topic-relevant knowledge expertise equally, users' ranking scores in topic T ($|T| \geq 1$), denoted as LR_T, can be calculated by:

$$LR_T = Inf_T \times ES_T \tag{2.1}$$

where Inf_T denotes the topic-sensitive influence in topic T and ES_T means the topic-related expertise. Thus, for a topic T, the users who have high ranking scores are identified as topical opinion leaders.

2.4.1 Topic-sensitive Influence Measure

We first conduct topic preprocessing to represent the topic interest of each user, and then confirm the existence of *homophily* in our dataset. Based on this topic preprocessing and the finding, a novel approach to measure users' topic-sensitive influence is proposed in this section. Table 2.2 lists the descriptions of notations.

Table 2.2: Notation descriptions.

Notation	Description
n	the total number of users
s	the total number of unique topics
A, Q	$n \times s$ matrix, where $A_{i,t}$/$Q_{i,t}$ contains the number of topic t in user u_i's answers/questions
V	$n \times s$ matrix, where $V_{i,t}$ contains the number of votes received by user u_i in topic t
AM, QM	$n \times 7$ matrix, where $AM_{i,t}$/$QM_{i,t}$ contains the number of major topic t in user u_i's answers/questions
CM	$n \times 7$ matrix, where $CM_{i,t}$ contains the number of major topic t in user u_i's posts (questions and answers), i.e., $CM_{i,t} = AM_{i,t} + QM_{i,t}$

Topic Preprocessing. The purpose of topic preprocessing is to identify each user's topic interest. In Zhihu, each post of a user is always related to many unique topics so that a user has much more unique topics in the published posts. Hence, directly leveraging these unique topics to represent the topic interest of a user is very intricate because of their amount and diversity. To this end, utilizing the tree-like topic structure of Zhihu, we aggregate these topics into seven major topics, which cover all the topic fields in Zhihu. It is worth noting that, besides 6 child topics of the root topic, we select another representative topic "Science & Technology" that had not been edited into the topic structure due to some mistakes from Zhihu topic organization. Using this topic aggregation method, each post's topics of each user are transformed to the corresponding major topics according to the topic relationship in the topic structure.

To identify each user's topic interest, we first compute the topic interest of each user's questions and answers over the major topics respectively. We can row normalize AM, QM into AM', QM' such that $||AM'_{i,\cdot}||_1 = 1$ for each row $AM'_{i,\cdot}$ and $||QM'_{i,\cdot}||_1 = 1$ for each row $QM'_{i,\cdot}$. Each row of these two matrices denotes the probability distribution of a user's interest in question/answer. Using a distance metric for probability distribution [28], the topic interest difference TD between questions and answers of user u_i can be calculated as:

$$\begin{aligned} TD_{QA}(i) &= TD(AM'_{i,\cdot}, QM'_{i,\cdot}) \\ &= \sqrt{D_{KL}(AM'_{i,\cdot}||M) + D_{KL}(QM'_{i,\cdot}||M)} \end{aligned} \tag{2.2}$$

where $M = \frac{1}{2}(AM'_{i,\cdot} + QM'_{i,\cdot})$. D_{KL} is the *Kullback-Leibler Divergence* which defines the divergence from distribution H to I as: $D_{KL}(H||I) = \sum_i H(i) \log \frac{H(i)}{I(i)}$.

Figure 2.6 demonstrates the Cumulative Distribution Function (CDF) of topic interest difference between questions and answers of each user. The analysis is applied on a set of 181K+ users who posted at least one question and one answer. We can observe that the topic interests of their questions and answers for most users are similar. Hence, in this work, the major topic probability distribution of posts published by each user is utilized to present each

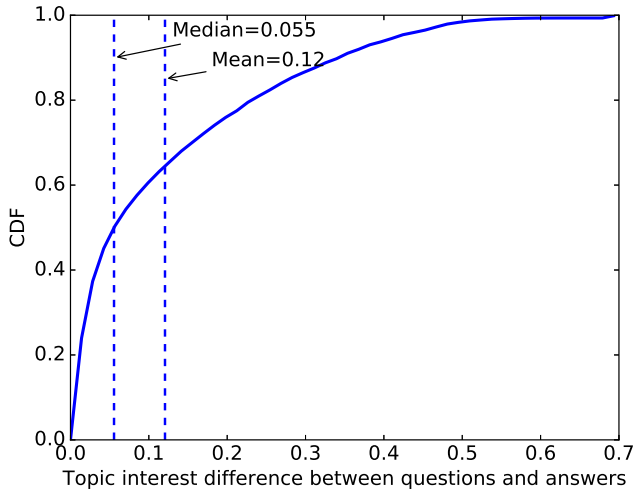


Figure 2.6: Topic interest difference between Q&A.

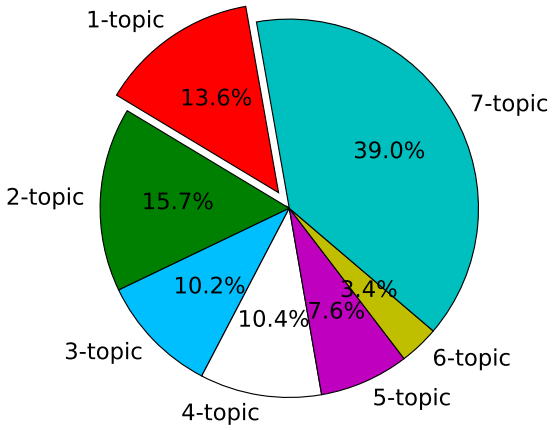


Figure 2.7: Distribution of question topic type.

user's topic interest. Namely, after the row normalization, $CM'_{i,t}$ indicates the probability that user u_i is interested in topic t. Note that the topics are transformed to the corresponding major topics only in the user topic interest calculation process.

Besides, to examine the topic diversity in SCQA sites, Figure 2.7 illustrates the distribution of question topic type in Zhihu, where k-topic means a type of questions that is relevant to k major topic(s). We can observe that multi-topic questions account for 86.4%, implying that multi-topic questions are pervasive in Zhihu. Inspired by this, our proposed algorithm is required to support identifying multi-topic opinion leaders.

Homophily. To assist in measuring the true topical influence of each user, we need to examine whether *homophily* exists in the social network of our dataset, which has been observed in many social networks [62, 100]. The phenomenon shows that users follow each other on account of similar topic interest, which means that the influence on each follower would depend

on the topic interest. The question can assist in verifying whether *homophily* exists in Zhihu: *Do users with "following" relationships have more similar topic interest than those without?*

The question can be formalized as a two-sample t-test: The null hypothesis is $H_0 : \mu_{follow} = \mu_{unfollow}$, and the alternative hypothesis is $H_1 : \mu_{follow} < \mu_{unfollow}$, where μ_{follow} is the mean topic interest difference between two users with "following" relationship, and $\mu_{unfollow}$ indicates the mean topic interest difference of those without. We design *homophily* testing and evaluation experiments based on a set of active Zhihu users who published at least 10 posts in total, denoted as U ($|U| = 124,445$). We conduct the two-sample t-test on the user congregation because around 92% of the users in our dataset have less than 30 followees. Sample 0 contains the topic interest difference of all the user pairs with "following" relationships while Sample 1 contains the topic interest difference between each user and some randomly chosen users whom he/she does not follow. Note that the number of each user's chosen non-followees is identical to the number of each user's followees. The topic interest difference between two users is calculated as $TD_u(i,j) = TD(CM'_{i,\cdot}, CM'_{j,\cdot})$. The t-test result shows that H_0 is rejected at significant level $\alpha = 0.01$ with a p-value of less than 1×10^{-17}. The t-test result depends on the extent of the dataset normality. Skewness and kurtosis of these two samples are 1.19, 2.14 and 1.21, 2.09, which are considered acceptable in order to prove normal distribution [34]. Hence, we confirm that the existence of *homophily* in Zhihu.

QARank Algorithm. Based on the above process, we propose a novel topic-sensitive influence measure algorithm called QARank, which incorporates three factors:

- *Network structure:* A user's influence is propagated to other users through following links between them in SCQA. Hence, QARank considers the link structure, similar to the authority measure of a web page.

- *Topic interest:* Based on *homophily*, a user's topical influence on his follower is stronger when their interests in this topic are more similar and vice versa. A user has different influence in different topic in the same social network.

- *Knowledge authority:* Generally a user's opinion is always accepted by his followers when his answers obtain many votes. Hence, the knowledge authority of a user plays an important role in his influence. Specifically, the more votes a user received, the more authoritative his followers think he is.

The proposed QARank, as an extension of TwitterRank, is modeled as a random surfer model. Let G be a directed graph where each node indicates a user and each directed edge denotes a "following" relationship between two users. A random surfer on the graph G visits each user with certain probability through following the corresponding edge. QARank differentiates itself from TwitterRank in that the topical knowledge authority is considered into the transition probability from one user to another meanwhile QARank can measure the multi-topic influence

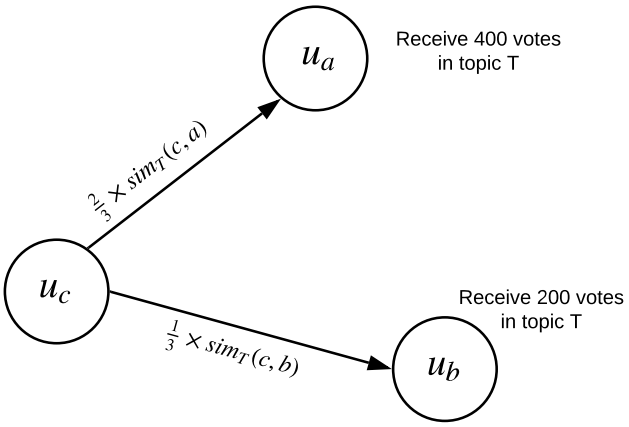


Figure 2.8: Example of transition probability calculation in QARank.

by leveraging Euclidean distance to measure the topic interest difference. Hence, each element of the transition matrix P_T for the topic set T ($|T| \geq 1$) is calculated as:

$$P_T(i,j) = \frac{|V_{j,T}|}{\sum_{k:u_i \ follows \ u_k} |V_{k,T}|} \times sim_T(i,j) \tag{2.3}$$

where

$$sim_T(i,j) = 1 - \sqrt{\sum_{t \in T} (CM'_{i,t} - CM'_{j,t})^2} \tag{2.4}$$

where $P_T(i,j)$ is the transition probability from follower u_i to followee u_j in the random surfer model. $|V_{j,T}| = \sum_{t \in T} V_{j,t}$ is the number of votes received by user u_j in topic T, and $\sum_{k:u_i \ follows \ u_k} |V_{k,T}|$ is the total number of votes received by all u_i's followees in topic set T. In the model, the number of topic-related votes received is regarded as the topical knowledge authority of a user. Figure 2.8 shows an example about three users. u_c follows u_a and u_b, who received 400 and 200 votes in topic T respectively. In this case, u_a's influence on u_c is two times of that of u_b, when the topic interest similarity among the three users is not considered. Of cause, u_c's influence on u_a and u_b are also related to the topic interest similarity between them.

In addition, in case of dangling nodes that do not have any out-degree and cyclic loops in the network, we apply random jump [38] by adding a teleportation vector E_T:

$$E_T = A''_{.,T} \tag{2.5}$$

where $A_{.,T} = \sum_{t \in T} A_{.,t}$, and $A''_{.,T}$ is the column-normalized version of $A_{.,T}$ so that $||A''_{.,T}||_1 = 1$.

Given the transition probability matrix and the teleportation vector, the topical influence scores of users in topic set T, known as Inf_T, can be calculated iteratively as:

$$Inf_T = \lambda P_T \times Inf_T + (1-\lambda) E_T \tag{2.6}$$

2.4.2 Topic-relevant Expertise Measure

Measuring topic-relevant expertise is of great significance in identifying topical opinion leaders in SCQA sites. In order to infer the topic-relevant expertise, we propose three topic-relevant metrics incorporating the knowledge capacity, satisfaction and contribution separately.

Knowledge capacity. In SCQA sites, answering many questions in specific topics means that one has rich topic-related knowledge while asking lots of topic-related questions usually indicates one lacks knowledge about these topics. Therefore, the *z-score* is adopted to measure a user's knowledge capacity in specific topics [112]. The knowledge capacity of user u_i in topic set T is calculated as:

$$KC_T(i) = \frac{|A_{i,T}| - |Q_{i,T}|}{\sqrt{|A_{i,T}| + |Q_{i,T}|}} \tag{2.7}$$

where $|A_{i,T}| = \sum_{t \in T} A_{i,t}$ is the number of answers published by user u_j in topic set T and $|Q_{i,T}| = \sum_{t \in T} Q_{i,t}$ sums up the number of questions asked by user u_i in topic set T. If answers are more than questions, KC is positive, otherwise it is negative.

Knowledge satisfaction. Another important function in SCQA is voting answers if a user agrees on them. The number of received votes indicates the satisfaction degree that an answer obtains. Hence, we use the average number of votes for the u_i's T-related answers as the knowledge satisfaction of user u_i in topic set T, which is defined as:

$$KS_T(i) = \frac{|V_{i,T}|}{|A_{i,T}|} \tag{2.8}$$

Knowledge contribution. Topical opinion leaders should be active and make a great number of contributions to SCQA sites. In our work, we choose the number of topic-related answers to measure the knowledge contribution of user u_i, which is calculated by:

$$IC_T(i) = |A_{i,T}| \tag{2.9}$$

Before combining the above three factors, Min-Max normalization is adopted to rescale the range of factors to [0, 1]. Therefore, KC_T, KS_T, and IC_T are transformed into the Min-Max normalized forms $\widetilde{KC_T}$, $\widetilde{KS_T}$, and $\widetilde{IC_T}$. Given this, the expertise score ES_T of user u_i in topic set T ($|T| \geq 1$) is calculated by:

$$ES_T(i) = \mathcal{F}(\beta\widetilde{KC_T}(i), \gamma\widetilde{KS_T}(i), (1-\beta-\gamma)\widetilde{IC_T}(i)) \tag{2.10}$$

where $\mathcal{F}(x, y, z)$ means the expertise measure method, β and γ are two parameters tunning the weight. To compare with [85] equally in the evaluation section, in our work, $\mathcal{F}(x, y, z)$ is a weighted sum of three metrics. It is worth noting that the expertise measure can be replaced with other efficient methods [68, 76].

In conclusion, the ranking process details of QALeaderRank algorithm are illustrated in Algorithm 1. Based on this, we can calculate the ranking score of each user to identify top-ranked users as topical opinion leaders.

Algorithm 1: QALeaderRank algorithm

Input : U, G, V, A, Q, CM', T, maximal iteration number max, amount of convergence required ϵ
Output : ranking score list LR

foreach *user* $u_i \in U$ **do**
  Extract u_i's followee list L from G;
  foreach *user* $u_j \in L$ **do**
    Compute the transition probability $P_T(i,j)$ from u_i to u_j using Equation 2.3;
  end
  Compute u_i's knowledge capacity, satisfaction and contribution using Equation 2.7, 2.8, and 2.9;
  Compute u_i's knowledge expertise $ES_T(i)$ using Equation 2.10;
end
Compute the teleportation vector E_T using Equation 2.5;
do
  Update $Inf_T(n) = \lambda P_T \times Inf_T(n-1) + (1-\lambda)E_T$;
  Update the convergence distance $dist(n, n-1)$;
  Update iteration number $num+ = 1$;
while $num \geq max$ & $dist(n, n-1) \leq \epsilon$;
return $LR = Inf_T \times ES_T$

2.5 Empirical Evaluation

In this section, we present an empirical evaluation of the proposed QALeaderRank over 10 popular topics in Zhihu along with an online user study.

To test the performance of **QALeaderRank (QALR)**, we compare it with two baseline algorithms in our experiments.

- **TwitterRank (TR):** It measures users' topic-sensitive influence with the consideration of the topical similarity and the link structure [100]. However, TwitterRank does not take any knowledge expertise into account.
- **InExRank (IR):** Song et.al. [85] proposed a topic-irrelevant method considering authority, activity and influence. In order to compare with this work, we extend it by incorporating topical expertise and following information (i.e., follower count) denoted as InExRank.

Two similarity metrics for comparing rankings are leveraged as follows:

Table 2.3: Ranking similarity among top 20 users identified by three algorithms.

	OSim		KSim	
	Mean	**Median**	**Mean**	**Median**
QALR	0.24	0.19	0.42	0.42
IR	0.15	0.14	0.39	0.41
TR	0.96	0.95	0.96	0.96

- $\mathbf{OSim(r_1, r_2)}$: It measures the overlap degree of two top k rankings r_1 and r_2 [42], which is defined as:

$$OSim(r_1, r_2) = \frac{|r_1 \cap r_2|}{k}$$

- $\mathbf{KSim(r_1, r_2)}$: It considers the degree to which the relative ordering of two rankings r_1 and r_2 is in agreement [26]. Let $R = r_1 \cup r_2$, and $\theta_1 = R - r_1$. We extend r_1 by appending θ_1 to the tail of r_1 to yield r_1'. r_2' is analogously extended. Thus, the $KSim$ similarity can be calculated by:

$$KSim(r_1, r_2) = \frac{|\{(u,v)|r_1', r_2' \; agree \; on \; order \; of \; (u,v)\}|}{|R| \times (|R| - 1)}$$

  where $(u, v) \in R \times R \; (u \neq v)$ means u ranks in front of v.

2.5.1 Performance Evaluation

We compare the performance of QALR and two baseline algorithms on our Zhihu dataset over 10 popular topics from some different perspectives. These topics are "Movie "(T0), "Psychology" (T1), "Travel" (T2), "Food" (T3), "Fitness" (T4), "Internet" (T5), "Fashion" (T6), "Pioneer" (T7), "Design" (T8), "Finance" (T9). For the simplicity, we assume three expertise metrics are equally essential to the expertise measure, i.e., $\beta = \frac{1}{3}$, $\gamma = \frac{1}{3}$. Teleportation parameter λ in QALR and LR are set as 0.85. As a result, we get three user rankings identified by three methods.

Performance on Topic Correlation. We look at the ranking correlation between topic pairs for the three algorithms to compare their topic sensitivity. From Table 2.3, we can observe that TR identifies much more similar leaders (with high mean/median value) than IR and QALR, while QALR and IR can yield diversified top-ranked users in each topic. This is because TR considers the number of published posts during computing transition probability, which makes one user who published many topic-irrelevant posts get high ranking score in the random surfer. Besides, the ranking similarity of IR is a little less than that of QALR, which is because QALR considers more topical influence rather than mainly focusing on the topical expertise.

Performance on User Identification. Before comparing the performance, we first divide users into 4 types according to their influence and expertise. An illustration is given in Figure 2.1. The 4 types of users are as follows:

Table 2.4: Statistic comparison of top 20 users identified by three algorithms.

	Number of followers		Number of votes		Number of answers	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
QALR	46922.59	6494.0	12245.54	4481.5	48.41	16.0
IR	43453.73	549.0	8389.55	1185.0	169.68	109.0
TR	56171.41	9261.5	4766.57	235.0	29.35	7.0

- ***Type I***: *Influential users with expertise* (Zone I in Figure 2.1) have strong influence and high expertise in specific topic(s). They always have a great number of followers, publish many posts and receive a large number of votes.
- ***Type II***: *Influential users without expertise* (Zone II in Figure 2.1) have strong influence due to their popularity in other fields but publish very few posts and get few votes in specific topic(s).
- ***Type III***: *Non-influential users without expertise* (Zone III in Figure 2.1) seldom submit posts and do not influence others in given topic(s).
- ***Type IV***: *Non-influential users with expertise* (Zone IV in Figure 2.1) are not influential and have few followers. However, they like publishing posts.

The purpose of our work is to identify ***type-I*** users from all users as accurately as possible. This section studies the detailed information of opinion leaders identified by three algorithms to compare their identification accuracy.

The results of QALR are conformant to our expectation. The top-ranked topical opinion leaders identified by QALR mostly published lots of topic-related posts and received a great number of votes. They have many followers including some important followers, who are also top-ranked users. It is evident that they belong to ***type-I***. Table 2.4 also shows that the top 20 users of QALR get much more votes than those of two baselines over 10 topics. Furthermore, we take some top 5 users of QALR for the detailed explanation. As shown in Table 2.5, "wangxing" is identified as an opinion leader in topic "Pioneer". We find that he posted mainly about pioneer and has 61,268 followers including an important user "zhou-kui". Actually most of pioneer-related top 5 opinion leaders are successful company founders in real life. For example, "wangxing" founded some popular websites such as meituan.com, fanfou.com and renren.com. "zhou-kui" is a partner of Sequoia Capital China. "dreamcog" founded a company named youxiamotors. In addition, "xiepanda", "liuniandate"and "WxzxzW" are identified as top 5 leaders in many topics because they are so-called *cewebrity*, who acquired fame by publishing a great number of posts about various topics. For instance, "xiepanda" posted mostly about movie, psychology, food, Internet and finance. He also often posted about fitness, fashion, pioneer and design. Besides, his answers always got 400+ votes in related topics.

However, for the results of TR, some ***type-II*** users like the users colored in red in Table 2.5 are identified. For example, "xiepanda", "liuniandate", "chuan-zhu", and "mazk" are identified

Table 2.5: List of top 5 users respectively identified by three algorithms over 10 topics.

Topic #	Topic	QALeaderRank	InExRank	TwitterRank
0	Movie	xiepanda, liuniandate, vikinglau, WxzxzW, chen-yao-39-75	leslycheung, wu-liang-si, yang-vv-20, ku-nu-ya-lu, tangyu	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
1	Psychology	xiepanda, liuniandate, WxzxzW, zhang-xiao-wei-23, yezhuang	liu-yue-61-89, xiepanda, compiler, li-fei-yang-75, WxzxzW	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
2	Travel	WxzxzW, chico-62, xu-wen-39, li-zhi-qiang-peter, qiu-shi-19-94	qi-lu-you, zzllss, duan-xiao-hui-93, ding-ding-1-50-48, WxzxzW	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
3	Food	xiepanda, anshi, wei-jiali, ji-li-ji-li, liunian-date	xiepanda, rou-si-23, dandelionpxj06, wang-xiao-jie-67-75, WxzxzW	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
4	Fitness	WxzxzW, chico-62, xiepanda, summer.li, guo-fu-lin	fitwu, lin-tu-ren-61, admeoseer, xiepanda, WxzxzW	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
5	Internet	xiepanda, WxzxzW, liuniandate, big_caaat, 8king	pirlo, zang-qi-long, am-plex, xiaoxiao, HuDP	xiepanda, liuniandate, WxzxzW, mazk, chuan-zhu
6	Fashion	WxzxzW, 8king, sick-berry, liuniandate, xiepanda	jiong-se-fu-78, hkook-kim, WxzxzW, xiepanda, halopeeka-boo	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
7	Pioneer	wangxing, zhou-kui, xiepanda, liuniandate, dreamcog	cheng-xiao-92, ding-kai-59-87, deng-li-zheng-44, he-de-wen, zhidemofang	xiepanda, liuniandate, WxzxzW, mazk, chuan-zhu
8	Design	WxzxzW, 8king, xiepanda, soulchef, xiaoxiao	WxzxzW, xiaoxiao, xiepanda, indablues, baiyanliao	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk
9	Finance	xiepanda, liuniandate, WxzxzW, ji-li-ji-li, big_caaat	li-xiao-ma-89, marx-abraham, xiepanda, tony-lee-17, zang-qi-long	xiepanda, liuniandate, WxzxzW, chuan-zhu, mazk

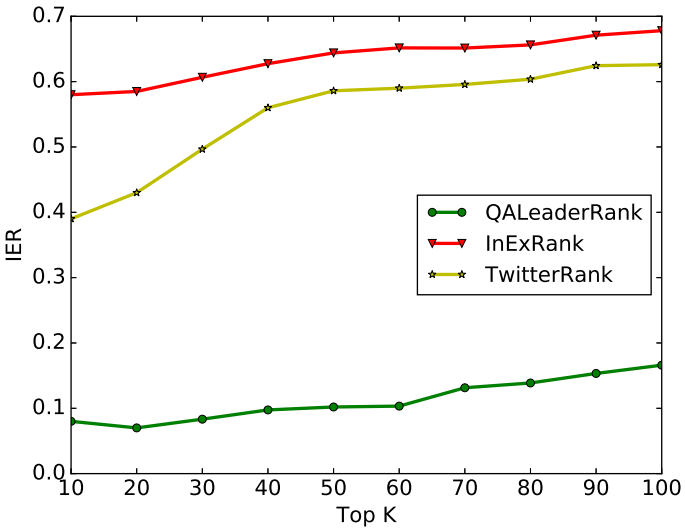


Figure 2.9: IER comparison of top users.

by TR as 4 out of top 5 users in topic "Travel". However, in fact, "xiepanda" did not post any content about "Travel", "liuniandate", "chuan-zhu", and "mazk" only posted one or two answers which received few votes. This is because the influence-focused TR ignores the topical knowledge expertise. Thus, as shown in Table 2.4, although the mean/median follower count of top 20 users identified by TR is higher than that identified by QALR, TR is much less than QALR in terms of vote/answer count.

For IR, a big problem is that IR, an expertise-focused method, yields a number of ***type-IV*** users like the users colored in blue in Table 2.5. For instance, in topic T3, "rou-si-23" only has 20 followers but published 192 related answers with 15 of maximal vote count and 0.58 of average vote count. "HuDP" posted 615 Internet-related answers that got 9 of maximal vote count and 0.34 of average vote count and only has 33 followers. One can image that these ***type-IV*** users may be paid posters, spammers or normal active but non-influential users, but cannot be indeed topical opinion leaders. This results from the accumulation of four factors in IR algorithm where one large factor (i.e., the number of answers) can greatly increase the final ranking score. Table 2.4 shows that the top 20 users identified by IR got much less votes than those identified by QALR although the users of IR posted much more answers. Meanwhile, the top-ranked users of IR have much less followers than those of QALR. This is because IR measures influence using the number of followers while QALR measures the topical influence based on the link structure of the social network.

Performance on Identification Error Rate. As mentioned above, some users who are not topical opinion leaders are wrongly identified by algorithms, such as *type-II* and *type-IV* users. In order to measure the fraction of apparently wrongly identified users, a new metric,

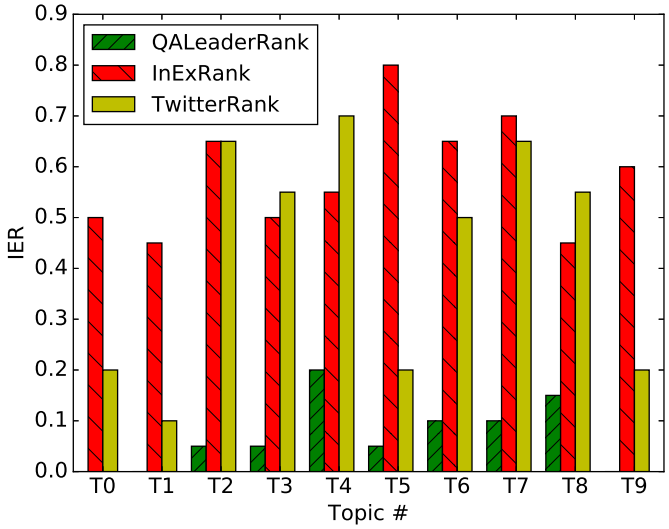


Figure 2.10: IER comparison over topics.

Identification Error Rate (IER), is proposed in this chapter. For the top k users in topic t, *IER* can be calculated as:

$$IER(k,t) = \frac{|\{l_i | v_i^t \leq nv \ or \ f_i \leq nf, i \in [0,k)\}|}{k} \tag{2.11}$$

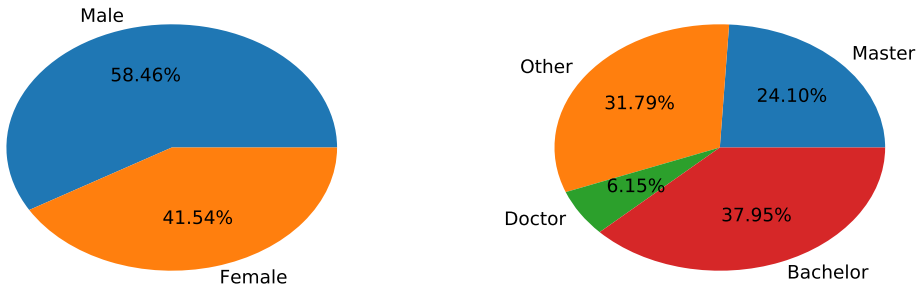
where l_i is the i-th identified leaders. v_i^t denotes average vote count of l_i in topic t and f_i is l_i's follower count. nv and nf indicates average number of votes of all answers and average number of followers of all users respectively. In our work, we assume that l_i is a wrongly identified top-ranked user if his v_i^t or f_i is less than the mean value of all users. Thus as shown in Table 2.1, we set $nv = 13.63$ and $nf = 11.57$.

Figure 2.9 illustrates the average *IER* of identified top k users over 10 topics for the three algorithms. We can observe that *IER* of QALR is always below 20% while IR and TR yield very high *IER*. As an example, Figure 2.10 illustrates *IER* comparison of the top 20 users in each topic. Note that the rankings of QALR also lead to the lowest *IER* in each topic. In particular, the rankings identified by QALR are of extremely high quality ($IER = 0$) in topics T0, T1, and T9. These observations further demonstrate that our proposed QALR greatly outperforms the two baselines in SCQA.

Performance on Multi-topic Identification. Our proposed QALeaderRank can also identify multi-topic opinion leaders. We show results for 2-topic opinion leaders identification in Table 2.6. For example, "8king" and "big_caaat" is respectively identified as a fashion-design-related opinion leader and a Internet-finance opinion leader. "big_caaat" posted frequently about Internet and Finance, who has 7939 followers including an important user "xiepanda". He published 83 Internet-related answers with 227 of average vote count and 48 finance-related answers with 163 of average vote count. "8king" posted frequently high-quality answers about fashion and design. He is also followed by a number of important users, including "WxzxzW"

Table 2.6: Top 5 multi-topic users identified by QALeaderRank.

Topic	Top 5 users
(Movie, Psychology)	xiepanda, WxzxzW, vikinglau, liuniandate, zhang-xiao-wei-23
(Fashion, Design)	WxzxzW, sickberry, 8king, xiepanda, liuniandate
(Internet, Finance)	WxzxzW, xiepanda, liuniandate, Jasonhau, big_caaat



(a) Gender **(b)** Education

Figure 2.11: Distribution of Participants.

and "sickberry". It is worth noting that "liuniandate" is ranked as the No.4 opinion leader across two topics "Movie" and "Psychology". However, the user is respectively ranked as No.2 in these two topics. This is because QAleaderRank considers the general topical influence of the social network based on topical interest and knowledge expertise instead of the individual influence for each topic.

2.5.2 User Study

To further evaluate the proposed approach, we conducted a user study to compare the proposed QALeaderRank with two baseline algorithms over 10 topics. By respectively selecting the top 20 users for each topic from the three algorithms, we obtained about 50 users in each topic due to some overlapping among the top 20 users of the three methods. Then we designed an online questionnaire that asked each participant to choose one topic that he/she focused on most frequently and rate each user's topical opinion influence using 5-point Likert scales. The questionnaire listed each user's name and three representative topic-related answers as tips. The top 3 topic-related answers are selected as the three representative answers in terms of the number of received votes. Before answering this questionnaire, each participant is required to understand 5 degrees of topical opinion influence as shown in Table 2.7. We spread the questionnaire to some professional online Zhihu discussion groups, the majority of whose members are active Zhihu users. Totally, we received about 200 valid questionnaire responses

Table 2.7: 5-point Likert scales in the questionnaire.

Degree	Description
1 point	*very weak, which means you do not know the user or never view any topic-related post published by the user*
2 point	*weak, meaning that you browsed some unimpressive topic-related posts published by the user*
3 point	*medium, meaning that you browsed and agreed on some of topic-related posts the user published*
4 point	*strong, which indicates that you browsed and agreed on most of posts and opinions of the user in this topic, and voted or commented on his/her answers*
5 point	*very strong, meaning that you agreed on the the user's opinions and posts absolutely, often checked the user's update, and have invited or would like to invite the user to answer your topic-related questions*

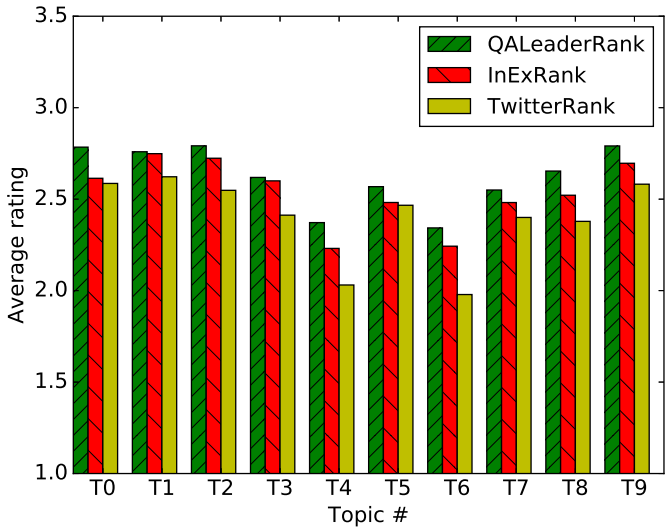


Figure 2.12: Average rating comparison.

(about 20 responses for each topic). Before comparing the results, we show some distributions of personal information from participants. As shown in 2.11a and 2.11b, we can find that the distributions of participants are similar to the distributions of users in Zhihu which are reported in [2]. Therefore, these participants are not only active users in Zhihu, but also representative in Zhihu, because they are reasonably distributed in all user classes from gender and education perspectives. In the future, we plan to invite some experts in Zhihu to rate the identified topical opinion leaders from these three methods, which would further make the evaluation results more convincing.

Average rating comparison. Using the ratings collected from those questionnaire responses, we calculate and give the comparison of average ratings of top 20 users identified by three algorithms as shown in Figure 2.12. We find that the overall ratings of QALR over 10

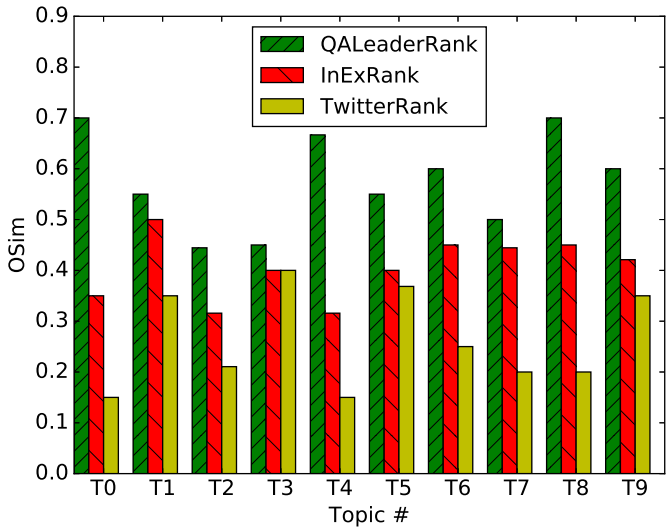


Figure 2.13: Comparison of similarity (OSim) between real rankings and identified rankings.

topics are higher than those of two baseline methods. Note that average ratings of QALR seem not high, resulting from that the user study is in a cold-start situation with limited information from each participant.

Ranking similarity comparison. Figure 2.13 and Figure 2.14 report on the ranking similarity between real rankings and the top 20 rankings generated by the three algorithms over 10 topics. The real rankings for each topic is produced by ordering the average rating of each user in each topic. From Figure 2.13, we can observe that the rankings of QALR are much closer to the real rankings than those of two baselines over 10 topics in terms of the overlap similarity $OSim$. Especially, our algorithm yields much more prominent rankings in topics T0, T4, T8 and T9. Furthermore, from Figure 2.14, for nearly all the topics, the ordering accuracy of QALR is higher than those of two baseline algorithms. As a result, a majority of participants preferred the rankings of QALR.

2.5.3 Discussion

In information retrieval, learning to rank (L2R) has been received a lot of attention from research community. For further improvement, our work proposes an preliminary framework containing two steps: top k user retrieval and L2R re-ranking. The first phase is generating and aggregating top k users using several heuristic models, which are QALeaderRank, InExRank, and TwitterRank in our work. In the second phase, a more accurate but computationally expensive L2R model is used to re-rank these users. We adopt RankNet [16], which uses gradient descent and employs cross entropy as loss function to train a neural network model. In order to enhance ranking accuracy, apart from the four factors considered in QALeaderRank, we also take into account the number of followers, the maximum number of received votes, and the number of questions in a given topic.

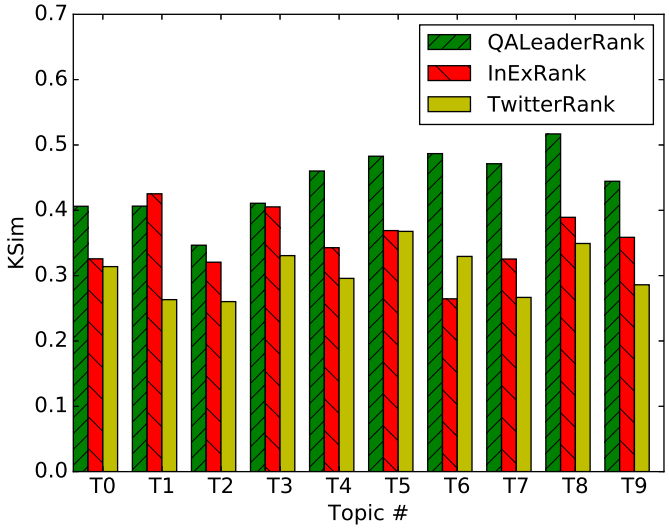


Figure 2.14: Comparison of similarity (KSim) between real rankings and identified rankings.

Table 2.8: Comparison of average $KSim$ similarity for 3 models.

Method	Random	QALeaderRank	RankNet
Average $KSim$	0.377	0.443	0.493

As shown in Table 2.8, we compare the ranking results of RankNet to those of QALeaderRank and Random model leveraging the KSim similarity. From Table 2.8, we find that RankNet enhances the ranking accuracy of QALeaderRank with only considering three more factors. Hence, we believe that the ranking accuracy would be improved further if much more features are taken into account in L2R model. This work will be conducted in the future.

2.6 Analysis of User Topic Interest Change

After identifying topical opinion leaders, further understanding and predicting their topic interest change behaviors is of great significance for many applications, such as answerer recommendation for askers, question invitation for topical opinion leaders, topic recommendation for users. Hence, in this section, we try to analyze and predict the topic interest change behaviors of a great number of active users including topical opinion leaders so that we can understand topical opinion leaders as well as general active users in SCQA sites. Based on the analysis and experiments, we detect the change patterns of user topic interest and examine the predictability of user topic interest change.

2.6.1 Detecting Change Patterns of User Topic Interest

As mentioned in Section 2.4.1, multi-topic posts are ubiquitous in SCQA sites. Besides, with the continuous emergence of new topics and events, some users could be attracted by new topics and events and focus on new topic domains. Therefore, we can image that there

maybe exist various kinds of users in SCQA sites: some kind of users always focus on several relatively fixed topics which means their topic interests are stable over time while some kind of users prefer more new emerging topics which means their topic interests are more or less unstable over time. Therefore, in this multi-topic era, we try to explore and answer the question: how does the user topic interest change in SCQA sites?

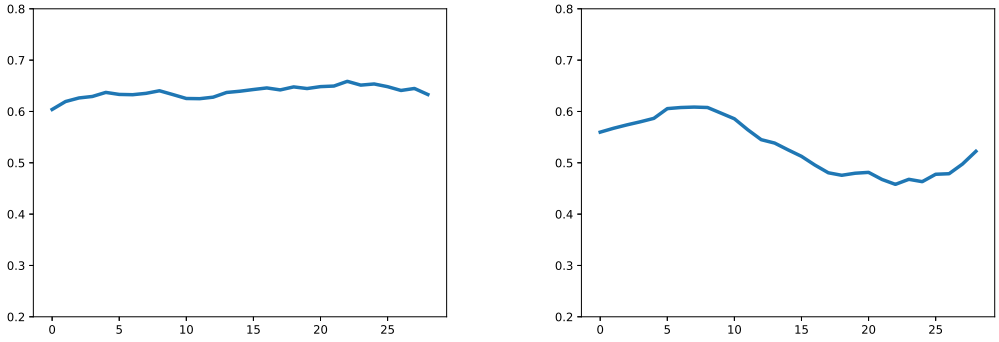
In order to find the user topic interest change patterns, we first extract active users who published more than l answers as representative samples, and then obtain a sequence of topic interests over time for each user u_i, i.e., $S_i = \{s_1, s_2, ..., s_l\}$ where s_k denotes the 7-dimension topic interest of the k-th answer in the sequence S_i and the sequence is arranged by their published time in an increasing order. Using the calculation method of user topic interest in Section 2.4.1, each answer's topic interest is denoted as the probability distribution over 7 major topics. To represent the topic interest change, we calculate the topic interest difference between s_{k-1} and s_k as:

$$\begin{aligned} c_{k-1} &= TD(s_{k-1}, s_k) \\ &= \sqrt{D_{KL}(s_{k-1}||m_{k-1}) + D_{KL}(s_k||m_{k-1})} \end{aligned}$$

where $m_{t-1} = \frac{1}{2}(s_{t-1} + s_t)$. Following the topic interest difference method, each user u_i has a sequence of topic interest change, i.e., $C_i = \{c_1, c_2, ..., c_{l-1}\}$. Here c_k has a value range between 0 and $\sqrt{2}$, where lower value denotes these two sequential topic interests are more similar. In our work, we select the active users who published more than 30 answers as samples, i.e., $l = 30$. The number of these active users are 28278.

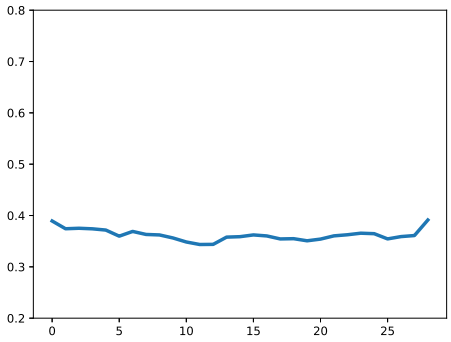
Drawing on these topic interest change sequences, we can cluster these users into several clusters to detect the change patterns of user topic interest. For this purpose, we leverage k-means clustering algorithm and set the number of clusters as 4 according to the clustering results. Figure 2.15 illustrates these four clusters' centers, which respectively represent four kinds of the user topic interest change patterns. As a result, in terms of the topic interest change patterns, SCQA users are divided into four clusters:

- Cluster 1: This type of users always change their topic interests over time, which may be because these users have a rich knowledge about various topics or they are interested in many kinds of topics.
- Cluster 2: This type of users' topic interests tend to be relatively stable from an unstable state, which may be because at the beginning, these users have not found their favorite topics yet, after finding interesting topics, they tend to focus on them during some period.
- Cluster 3: This type of users merely greatly change their topic interests over time, which may be because these users have found their favorite topics and keep focusing on these fields of topics.

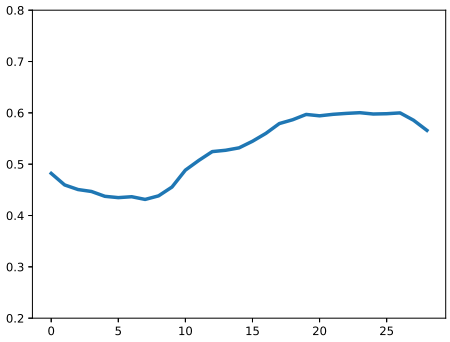


(a) Cluster 1

(b) Cluster 2



(c) Cluster 3



(d) Cluster 4

Figure 2.15: Four clusters of users in terms of topic interest change.

- Cluster 4: This type of users' topic interests tend to be relatively unstable from a stable state, which may be because at the beginning, these users have their favorite topics, and over time they want to develop their interests and involve the other topics.

As illustrated in Table 2.9, among the four clusters, Cluster 4 accounts for around 12% of the total amounts, which means that only a small body of users in SCQA sites always focus on several fixed topics. We can image that, with more and more new topics emerging, a majority of users not only concentrate on their current fixed topics but also change to involve in other interesting new topics. In a word, in current SCQA sites, a large body of users always change their topic interests.

We also specially explore the change patterns of opinion leaders' topic interest. To this end, we separately extract top-200 opinion leaders in each of ten most popular topics and obtain 1030 unique opinion leaders who published more than 30 answers. We also use the same clustering method to cluster opinion leaders' topic interest change patterns. Finally, we get very similar four patterns like Figure 2.15 and also obtain very similar distributions of clusters as shown in Table 2.9. To sum up, these topic interest change patterns exist in general active users as well as opinion leaders, which implies that every topic change pattern users have their

Table 2.9: Clusters of users.

Cluster ID	Active users	Opinion leaders
Cluster 1	31.6%	34.2%
Cluster 2	28.1%	28.4%
Cluster 3	12.3%	12.4%
Cluster 4	28.0%	25.0%

own opinion leaders. It is worth noting that Cluster 1 of opinion leaders accounts for higher proportion than that of general active users. This may be because, in order to enhance their influence and expertise, opinion leaders need to focus on and obtain richer knowledge about various topics and follow real-time new topics compared with general active users.

2.6.2 Predicting User Topic Interest Change

In this section, we first explore whether the user topic interest change are predictable and then try to predict the next topic interest. This prediction work can further assist in predicting and controlling topical opinion leader's topic interest change, which would promote many fine-grained recommendation applications.

Prediction of topic interest change. The task aims at initially examining the predictability of user topic interest change. Therefore, for simplicity, we predict the next topic interest change simply based on the previous topic interest changes without considering any other features. In order to intuitively show the change, we regard this problem as a binary classification task. More specifically, as mentioned in Section 2.6.1, for each user u_i, it has a sequence of topic interest changes $C_i = \{c_1, c_2, ..., c_{l-2}, c_{l-1}\}$. Regarding the task, we first set a topic interest change threshold T_c to label the change, i.e., if $c_{l-1} < T_c$, then $b = 0$ means no strong change, otherwise $b = 1$ means strong change. Hence, this task is to predict the topic interest change label b based on the previous $l - 2$ topic interest changes.

We choose several machine learning and deep neural network methods to predict the topic interest change, including Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM) and Long Short-Term Memory (LSTM) Network. In this experiment, 90% samples for each class are randomly selected as the training data and the rest for testing. All the prediction tasks are repeated 10 times and the average prediction performance is reported. The prediction performance is evaluated in terms of accuracy, precision, recall and F1-score.

As illustrated in Table 2.10, LSTM outperforms the other methods in terms of accuracy, precision, recall and F1-score as LSTM considers the temporal order of the previous topic interest changes. Obviously, all these methods have much higher prediction performance than the random guess method (i.e., 50% prediction performance). As a result, we can preliminarily examine that the user topic interest change can be predictable. In the future work, we plan to consider more related features to further enhance the prediction performance.

Table 2.10: Prediction of user's topic interest change.

Model	Accuracy	Precision	Recall	F1-score
LR	0.64	0.62	0.59	0.58
NB	0.62	0.60	0.59	0.59
SVM	0.64	0.62	0.59	0.58
LSTM	0.66	0.65	0.61	0.60

Prediction of next topic interest. The aim of the second task is to predict the next topic interest distribution of a user based on its previous topic interests. To be more specific, as defined in Section 2.6.1, for each user u_i, it has a sequence of topic interest distributions $S_i = \{s_1, s_2, ..., s_{l-1}\}$ where each element s_k in the sequence denotes the 7-dimension major topic interest distribution. The purpose of the task is to predict the next topic interest distribution s_l based on the previous sequence of topic interest distributions with the length of $l-1$.

This task is like predicting the next word based on previous several words. Each topic interest distribution can be regarded as embedding like word embedding. Inspired by this similarity, LSTM Network [43] is used to predict the next topic interest because of its excellent performance on sequence modeling, such as text modeling. The LSTM network introduces a memory cell that can preserve cell state over long periods of time so that it can address the problem of long-term dependencies and consider the temporal order of a sequence.

In our experiment, the LSTM network is implemented by Keras which is a deep learning library based on TensorFlow. Stochastic gradient descent using Adam optimiser is applied to update trainable parameters. The batch size is set as 128. We set the dimension of the LSTM hidden state as 32. The neural network model is trained for 100 epochs with an early stopping to report the results. Our data set is divided into the training and testing sets with a ratio of 9:1.

We define two evaluation metrics to measure the topic interest prediction performance. The first metric is Mean Topic interest Difference (MTD), which measures the mean topic interest difference between predicted topic interest and actual topic interest. The other one is Mean Pearson's Correlation coefficient (MPC), which measures the mean Pearson's product-moment correlation coefficient between predicted topic interest and actual topic interest. These two metrics are defined as:

$$MTD = \frac{1}{N}\sum_{i=1}^{N} TD(s_{il}, \tilde{s_{il}})$$

$$MPC = \frac{1}{N}\sum_{i=1}^{N} PC(s_{il}, \tilde{s_{il}})$$

where s_{il} and $\tilde{s_{il}}$ are the actual and predicted topic interest distribution of the i-th users in the test data set with N users, respectively. TD and PC denotes the topic interest difference and Pearson's correlation coefficient between two topic interest distributions.

Through the experiment, we get 0.4308 and 0.5808 in terms of MTD and MPC. As mentioned in Section 2.6.1, the topic interest difference degree that denotes weak change is around 0.4 so that we can say that MTD is low, which indicates that the actual topic distributions is similar to our predictions. Besides, the value of MPC also indicates that the model can predict the relatively similar topic interest for each user. In the future work, to enhance the prediction performance, we plan to consider more features, such as the number of votes, the number of comments, and employ attention mechanism to select informative factors for the sequence.

2.7 Chapter Summary

This chapter focuses on identifying topical opinion leaders in SCQA and proposes an efficient method called QALeaderRank, considering both the topic-sensitive influence and the topic-relevant expertise. In the proposed QALeaderRank, to measure the true topical influence, by exploring the QA and social features, we propose a novel topic-sensitive influence measure algorithm named QARank for SCQA, incorporating the network structure, the topic interest similarity between users and the topical knowledge authority. In addition, we employ three topic-relevant expertise metrics for inferring the topical expertise. The experimental results over ten popular topics, along with the feedback from an online user study, show that QALeaderRank greatly outperforms the compared state-of-the-art methods. Finally, we further analyze and predict the topic interest change behaviors of active users including topical opinion leaders. Based on the observations, we detect the change patterns of user topic interest and examine the predictability of user topic interest.

Chapter 3

Predicting Individual Socioeconomic Status based on Mobile Phone Data

Nowadays with the ubiquity of mobile phones, predicting Socioeconomic Status (SES) based on mobile phone data has become a hot research topic. In this chapter, compared with previous work on region or household's SES, we aim at addressing a new problem of predicting individual SES based on mobile phone data. Nevertheless, the task has three main challenges, i.e., sparse individual records, scarce explicit relationships and limited labeled samples. To this end, this work in the chapter proposes a semi-supervised hypergraph-based factor graph model for individual SES prediction. First, it is able to efficiently capture the associations between SES and individual mobile phone data to reduce the loss caused by the individual record sparsity. Second, to mitigate the scarcity of explicit relationships, the model can capture implicit high-order mobility pattern relationships among users by the hypergraph structure. Third, the model can explore the limited labeled data and unlabeled data in a semi-supervised way. Experimental results indicate that the proposed model outperforms previous work on SES prediction by 5-22% in terms of F1-score and provides a considerable improvement (2-9%) compared with the state-of-the-art hypergraph-based methods and factor graph methods.

Contents

3.1 Introduction

Socioeconomic Status (SES) characterizes an individual, a household or a region's economic and social position in relation to others, which is typically divided into three levels (high, middle, and low) [84]. Assessing SES not only helps governments and research institutes study and make public policies, but also assists in meeting the needs of target clients by evaluating their purchasing power from a commercial perspective. Furthermore, SES can benefit a wide range of other fields, such as health [71, 103], education [82] and public transportation [19]. National statistical offices measure socioeconomic information typically by a large number of personal or household interviews. However, assessing individual SES for a whole country or region's population by this traditional method is extremely expensive and time-consuming (e.g., usually once every 5 to 10 years). It is critical to develop a low-cost means for timely capturing and assessing individual SES in a population.

Due to the worldwide ubiquity of smart phones and mobile services, mobile phone users could generate various usage records at any time and any place. Therefore, mobile phone data captures abundant information regarding personal social attributes, relation networks and mobility patterns in a large-scale population, which to some extent reflects SES. Hence, mobile phone data has been used as a novel data source for efficiently inferring SES with low cost. Most existing work infer regional or household SES based on mobile phone data by directly applying classic supervised machine learning methods [10, 44, 87]. Compared with prior work, this work studies the SES prediction on mobile phone data at an individual level. Intuitively, even living in the same household and area, individuals probably have different SES levels. Inferring the individual SES provides the finest level of evidence and indication to improve the quality of corresponding public policies-making. Furthermore, it can enable numerous fine-grained applications at an individual level, such as precision marketing, fine service and assessment. However, individual SES prediction on mobile phone data proposes three following main challenges:

- **Sparse individual records.** Compared with aggregated records of a region or household, a large portion of individual mobile phone users actually generate sparse valid usage records every day. With the ubiquity of WiFi, individual records that telco service providers can identify are becoming rarer. For example, 71.9% users generate less than two valid daily records in the data provided by an ISP in China. It is difficult to explore enough information from sparse individual records for revealing personal SES as done in the existing SES prediction work, thus causing poor prediction performance.

- **Scarce explicit relationships.** Due to the increasing popularity of mobile communication applications like WhatsApp and Wechat, an increasing number of mobile phone users are giving up traditional voice calling and SMS services [1]. Subsequently, the communication relationships built in these mobile applications are disconnected from ISP-provided mobile phone data. Therefore, explicit relationships among users extracted

from mobile phone records become scarce, which makes the methods based on such relationships failed to work.

- **Limited labeled samples.** Since the cost of assessing individual SES by existing methods is extremely high, it is rather difficult to obtain enough SES-labeled samples for learning models. To the best of our knowledge, prior work on the SES prediction only employ supervised learning methods to predict SES, which does not work well on data with limited labeled samples.

To this end, this work in the chapter proposes a novel semi-supervised probabilistic model called Hypergraph-based Factor Graph Model (HyperFGM). First, to reduce the performance loss caused by the individual record sparsity, leveraging the idea of factor graph model, HyperFGM utilizes customized factor functions to efficiently capture the correlations between SES and numerous attributes of users extracted from individual mobile phone records, which significantly exploits the power of sparse records compared with the prior methods on SES prediction. Second, to address the explicit relationship scarcity problem, HyperFGM leverages the advantage of hypergraph on high-order relationship modeling to model implicit high-order relationships among users based on the hypergraph structure, which avoids the performance loss caused by ignoring the implicit high-order relationships. Third, for handling the limited labeled samples, HyperFGM explores both labeled and unlabeled data on a hypergraph network in a semi-supervised way, thereby achieving better performance than supervised learning methods in prior SES prediction work.

Compared with the proposed hypergraph-based factor graph model, traditional hypergraph-based models [33, 80, 115], focusing on the relationships among objects, need to convert the numerous attributes of objects into various relationships among objects, causing conversion loss. Traditional factor graph models [91, 95, 105] only consider objects' attributes and explicit pair-wise relationships between objects in a simple graph, which ignore implicit and high-order relationships among objects. However, there actually exist many complex high-order relationships among objects [115]. Therefore, in order to solve the disadvantages of these two traditional methods, HyperFGM, combining hypergraph-based model and factor graph model into one model, predicts individual SES by not only directly considering the SES-related attributes of users but also modeling the implicit high-order mobility pattern-based relationships among users in the hypergraph structure.

We demonstrate the feasibility and power of HyperFGM on individual SES prediction using a set of anonymized real mobile phone data collected from a major ISP in China. Experimental results show that HyperFGM outperforms previous work on SES prediction by 5-22% w.r.t. the F1-score and provides a considerable improvement (2-9%) compared with the state-of-the-art hypergraph-based methods and factor graph methods. It is worth to note that the proposed HyperFGM is a general semi-supervised classification method, which can be applied not only to the SES prediction problem but also to other similar tasks.

The major contributions in this work are summarized as follows.

- We first identify the issue of predicting individual SES from mobile phone data. To our knowledge, no previous work has extensively studied this issue.
- We propose a semi-supervised probabilistic hypergraph model, HyperFGM, to solve the SES prediction problem, which jointly considers user attributes and high-order relationships among users based on the hypergraph structure.
- We apply our model on a collection of anonymized real mobile phone data. Experimental results show that HyperFGM outperforms the baseline models.

The rest of the chapter is organized as follows: Section 3.2 discusses related work. Section 3.3 shows the data collection. Section 3.4 describes the proposed HyperFGM model. Section 3.5 evaluates the prediction performance of HyperFGM with extensive experiments. Finally, Section 3.6 concludes the chapter.

3.2 Related Work

This section reviews the related work, containing SES prediction based on mobile phone data, factor graph based model, and hypergraph based model.

3.2.1 SES Prediction based on Mobile Phone Data

SES prediction based on mobile phone data emerges as a very recent application of Artificial Intelligence (AI) for social and economic good. One research direction is to investigate the relation between regional economic development and mobile phone usage. [83] analyzed the aggregated call detail records of mobile phone subscribers from two developing countries and extracted a set of important features that are strongly correlated with poverty indexes. [60] defined several indicators of mobile phone usage to analyze their correlations with economic status indicators. [31] presented a study on large-scale datasets of cell phone records with country-wide census data to analyze the relationship between specific socioeconomic factors and the way people use cell phones in an emerging economy in Latin America. Their main results show correlations between socioeconomic levels and social network or mobility patterns among others.

Other efforts are on applying classic supervised machine learning techniques to predict regional or household SES. [87] studied whether the information derived from the aggregated use of cell phone records can be used to identify the socioeconomic levels of a population and applied SVM and Random Forest (RF) on the aggregated cell phone records to predict regional socioeconomic levels. [44] leveraged a supervised Latent Dirichlet Allocation (LDA) to extract latent recurring patterns of co-occurring behaviors across regions and then used them to infer regional SES from large-scale spatio-temporal calling data. [10] developed a Deterministic

Finite Automation (DFA)-based method to generate a large number of features and relied on a linear regression method (elastic net) to predict the SES of each household in Rwanda on mobile phone data. However, these classic supervised learning methods cannot solve three challenges mentioned in Section 3.1, would lead to poor performance in predicting individual SES from mobile phone data.

3.2.2 Factor Graph based Model

Factor graph based models, as a specific type of graphical models, have been widely applied in many areas, such as social network modeling, disease forecasting and medical informatics. [91] formalized the social relationship learning into a semi-supervised framework and proposed a partially-labeled pairwise factor graph model by considering pairwise relations and attribute factors to infer the type of social ties. [105] proposed a sparse factor graph model, which projects sparse features into a lower-dimensional latent space and is able to capture the associations between complications and lab tests, to forecast potential diabetes complications. [95] presented a factor graph based model with customized factor functions defined based on domain knowledge, which can be used to infer characteristics of instantaneous brain activities by jointly analyzing spatial, temporal and observational relationships in electroencephalograms. However, these works do not consider implicit relationships between objects and these traditional factor graph models are unable to exploit high-order relationships among objects.

3.2.3 Hypergraph based Model

To formulate the complex relationships among objects beyond pairwise relationship, hypergraph learning has obtained some interest recently. [115] extended spectral clustering methods from undirected graphs to hypergraphs in which an edge can connect more than two vertices, and further proposed a transductive learning model on the basis of the the spectral hypergraph clustering approach. [46] proposed to employ the hypergraph structure to formulate the relevance relationship among images. [33] proposed to employ the weighted multiple hypergraphs to formulate the higher order relationships among objects. [80] modeled multi-way relations as hypergraphs and extended the Discriminative Random Walk (DRW) framework, originally proposed for transductive inference on single graphs, to the case of multiple hypergraphs. These hypergraph learning methods focus on the relationship called hyperedge and need to convert the attributes of objects into various relationships among objects, which causes some conversion loss. To our knowledge, there is no effort on directly considering the attributes of objects to well exploit the dependencies between a large number of real-valued attributes of objects and labels.

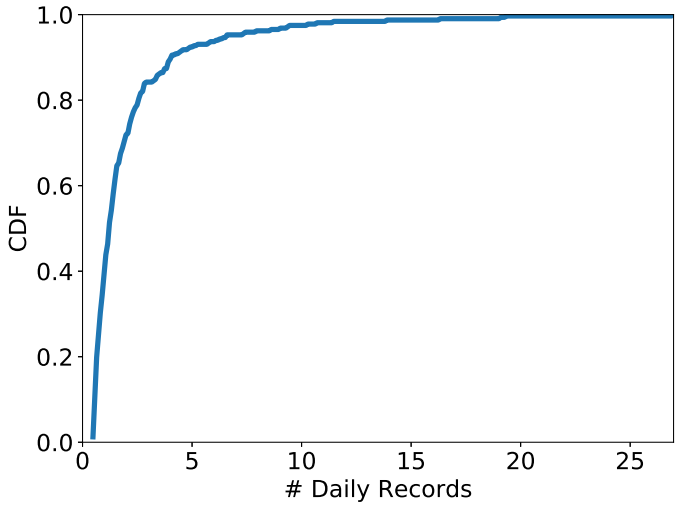


Figure 3.1: Distribution of average daily record count of each user.

Table 3.1: Description of dataset.

Location	Shanghai, China
Time duration	Oct.31, 2016-Feb.13, 2017
Number of users	317
Number of records	69621

3.3 Data Collection

Before presenting our proposed model, we first describe the dataset used in this work. The data was provided by our collaborator, a major ISP in China. We got the anonymized mobile phone's Internet records of 317 active mobile phone users (each user generated more than 50 records in a given period) who agree to provide their personal SES-related information, including occupation, education, income for this research. The dataset contain these users' Internet records from October 31, 2016 to February 13, 2017 from the city of Shanghai, one of the largest cities in China. For the user privacy and ISP's privacy agreement, the data can be only used for our research. Each user averagely generated about 219 valid records in the period. As shown in Figure 3.1, most of users generate very sparse daily records. The key data statistics are summarized in Table 3.1.

In our dataset, each user has an Internet record sequence generated from his/her mobile phone during the given period. Each record provided by the ISP contains the anonymized userID, the occurred time and the Uniform Resource Locator (URL). Figure 3.2 shows a Internet record sample, where contains encrypted user ID, time and URL. A URL indicates the address of a resource on the Internet, specifically, the HTTP and HTTPS requests issued from user to the cellular towers. As shown in the previous studies [45, 59], a user's spatio-temporal mobility pattern is correlated with his/her socioeconomic status. Besides, in order to make our

Figure 3.2: Mobile phone data sample in this work.

work applicable to traditional Call Detail Record (CDR) and other location-based data, this work mainly focuses on spatio-temporal information, i.e., latitude-longitude pair and timestamp, to exploit the power of mobile phone records for predicting individual SES. Through analyzing the content of URL, we find that the URLs generated from location-based mobile applications mostly contain Global Positioning System (GPS) location information. For example, the URL in Figure 3.2 contains a latitude-longitude pair. After extracting the location information from URLs, we can obtain a set of spatio-temporal data from the raw data, which will be used for the SES prediction in this work.

In order to obtain the SES label for each mobile phone user, a sociologist is invited to map users into three SES levels, which are high (level A), middle (level B), or low (level C) level, according to their personal SES-related information [7, 40, 78]. Finally, in our data, the resultant user distribution across classes is 70 users with Level A, 160 users with Level B and 87 users with Level C. Consequently, like most previous work [44, 55, 87] on SES level prediction, our work regards the SES prediction as a classification problem. To be more specific, the aim of this work is to predict a SES label (high, middle or low) for each individual user as accurately as possible.

3.4 The HyperFGM Model

The purpose of this work is to predict individual SES based on mobile phone user's records, which proposes three main challenges, i.e., sparse individual records, scarce explicit relationships, and limited labeled samples. To address these challenges, in this section, we propose and elaborate the details of the proposed HyperFGM for the individual SES prediction.

- For the sparse individual records, we leverage factor graph model to efficiently capture the correlations between SES and mobile phone records, which can greatly enhance the performance compared with classic machine learning methods [105]. To this end, we first extract SES-related user attributes from sparse mobile phone data through employing a

DFA-based method and a relief-based feature selection method. Then in the HyperFGM, we define attribute factor functions to represent the correlations between SES and each attributes.

- For the scarce explicit relationships, we first extract semantic mobility pattern similarity between users as implicit relationships, and then construct a hypergraph network structure among users based on the mobility pattern similarity to capture more implicit high-order relationships among users. Through defining hyperedge factor function in HyperFGM, we utilize these implicit high-order relationships among users to enhance the prediction performance.

- For the limited labeled samples, HyperFGM can explore both labeled and unlabeled data in a semi-supervised way. Specifically, the input data to our model is partially labeled so that the prediction model is learned by leveraging the labeled data and unlabeled data on the hypergraph network to infer the unknown label.

In this section, we first present the SES-related user attribute extraction from mobile phone's Internet records, and then propose a hypergraph construction method based on based on their semantic mobility patterns for exploring the implicit high-order relationships among users, as shown in Figure 3.4. Lastly, HyperFGM is conducted based on the user attributes and the hypergraph structure to infer the SES of each mobile phone user.

3.4.1 SES-related User Attribute Extraction

Since the raw latitude-longitude points contain no semantic meaning like the spot name or place attributes, we first need to preprocess the spatio-temporal data for the user attribute extraction. The first step utilizes a stay point estimating method proposed by Ye et al. [107] to obtain the stay points of each user from the raw latitude-longitude points. A stay point represents a geographic region in which the user stays for a while, which carries its semantic meaning, such as home, working place and the spot the user traveled. To obtain the stay points' semantic information about each user's real life style, the second step employs the Baidu Map API and a land price crawler to obtain each stay point's Point of Interest (POI), visited area name (district, city, country) and nearby housing price.

After the data preprocessing, the user attribute extraction transforms each user's semantic spatio-temporal data into a set of SES-related attribute metrics. To this end, we first employ a deterministic finite automation method [74] to generate a large number of potentially correlated attributes. In this method, several data transition operations are defined to transform the data input into a different data output using several legal operations like filter, group, select or computation. Consequently, the deterministic finite automaton takes a sequence of data as input and generates numerical metrics as output by a complete traversal of the automata. In our work, the structured and combinatorial method automatically generates more than 400 attributes. These user attributes are generated from different attribute spaces including record volume,

movement behavior, POI type, city level and housing price. For each attribute space, we obtain numerous real-value attributes such as mean, maximum, minimum, standard deviation, sum, radius of gyration and count/fraction of unique values over time.

To eliminate irrelevant attributes, we utilize a relief-based feature selection method, MultiSURF* [37, 94] to select SES-related user attributes according to the importance score ranking. In this work, top 20% attributes are selected as the final attribute input for the best prediction performance. As a result, for each user v_i, there is an associated attribute vector $\mathbf{x}_i$, in which each element denotes a user attribute.

3.4.2 Mobility Pattern-based Hypergraph Construction

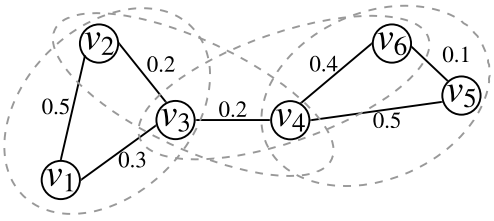
In this part, we aim at generating the implicit high-order mobility pattern relationships among users, namely, high-order relationships among users on a hypergraph structure based on users' semantic mobility patterns. As mentioned above, users with the same SES are more likely to have similar mobility patterns. For instance, persons, who typically stay in office during the daytime of a workday and visit entertainment places on the weekend, might belong to the same SES level. Inspired by this intuition, we first extract the semantic mobility pattern of each user by leveraging POI types and occurred time. A user's semantic mobility motifs can be defined as follows.

Definition 1. **Semantic Mobility Motifs.** A user v_i has a set of semantic spatio-temporal records $\{s_{i1}, s_{i2}, ..., s_{im}\}$. Each record is a tuple of $s = (t, p)$, which means that the user visited the POI type p at time t. Our work divides the time into workday/weekend and day/night so that a semantic mobility motif is defined as $smm = (w, d, p)$ if a user was at the POI p at time (w, d) where $w = 1$ if the time is in a workday otherwise 0; $d = 1$ if it is daytime otherwise 0. Hence, a user v_i's semantic mobility sequence is represented as $\mathbf{sm}_i = \{smm_{i1}, smm_{i2}, ..., smm_{im}\}$.

Given the defined semantic mobility motifs, we employ Latent Dirichlet Allocation (LDA) [9], a topic modeling method, to extract individual's semantic mobility patterns from mobility motifs. Each semantic mobility motif is regarded as a word and a user's semantic mobility sequence is treated as a document. As a result, each user's mobility pattern is represented as a topic distribution vector. Given two users' topic distribution vectors $\mathbf{m}_i, \mathbf{m}_j$, using a distance metric for probability distribution called *Jensen–Shannon Divergence* [25], the mobility pattern distance between each pair of users can be calculated as:

$$Mdistance(i, j) = \frac{1}{2} D_{KL}(\mathbf{m}_i || M) + \frac{1}{2} D_{KL}(\mathbf{m}_j || M) \tag{3.1}$$

where $M = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j)$, D_{KL} is the *Kullback-Leibler Divergence* which defines the divergence from distribution $\mathbf{p}$ to $\mathbf{q}$ as: $D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_i \mathbf{p}(i) \log \frac{\mathbf{p}(i)}{\mathbf{q}(i)}$.



	e_1	e_2	e_3	e_4	e_5	e_6
v_1	1	1	0	0	0	0
v_2	1	1	1	0	0	0
v_3	1	1	1	1	0	0
v_4	0	0	1	1	1	1
v_5	0	0	0	0	1	1
v_6	0	0	0	1	1	1

Figure 3.3: Left: A graph of six vertices, where pairwise distances between v_i and its 2 nearest neighbors are marked on the corresponding edges. Right: The H matrix of the hypergraph shown above. The entry (v_i, e_j) is set to 1 if a hyperedge e_j contains v_i, or 0 otherwise.

Based on the mobility pattern distance, we build a hypergraph structure $G = (V, E)$, where V represents a set of vertices (users), E is the hyperedge set such that for any hyperedge $e_i \in E, e_i \subseteq V$. Different from a simple graph that only contains pair-wise edges, the hypergraph is a graph where an edge called hyperedge can connect more than two vertices. Accordingly, to build the hypergraph, by using the star expansion strategy [46], we take each vertex as a centroid and generate a hyperedge for this vertex by connecting this centroid and its k-1 nearest neighbors. The strength of connectivity is determined by the mobility pattern distance between the centroid vertex and the other vertices. That is, each hyperedge connects k vertices. Following this construction method, we can choose different k (e.g., $k = 2, 3, 4, 5$) to generate different hyperedges in a hypergraph. Finally, the hypergraph can be represented by a $|V| \times |E|$ incidence matrix H:

$$h(v_i, e_j) = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

Figure 3.3 demonstrates an example to explain how to construct a hypergraph. We note that the employed methods of user attribute extraction and hypergraph construction are flexible and can be expanded/replaced by other methods.

3.4.3 Model Description

This work focuses on investigating the prediction of individual SES through combining traditional hypergraph model and a probabilistic factor graph model into one model. Given the above constructed hypergraph, we define the input of our problem as a partially labeled hypergraph network. The hypergraph network is denoted as $G = (V^L, V^U, E, Y^L, \mathbf{X})$, where V^L is a set of labeled users (vertices) and V^U is a set of unlabeled users with $V^L \cup V^U = V$; E is a set of hyperedges; Y^L is a set of SES labels corresponding to the users in V^L. Let an attribute matrix $\mathbf{X} = \{\mathbf{x}_i\}$ which means each user v_i is associated with an attribute vector $\mathbf{x}_i$.

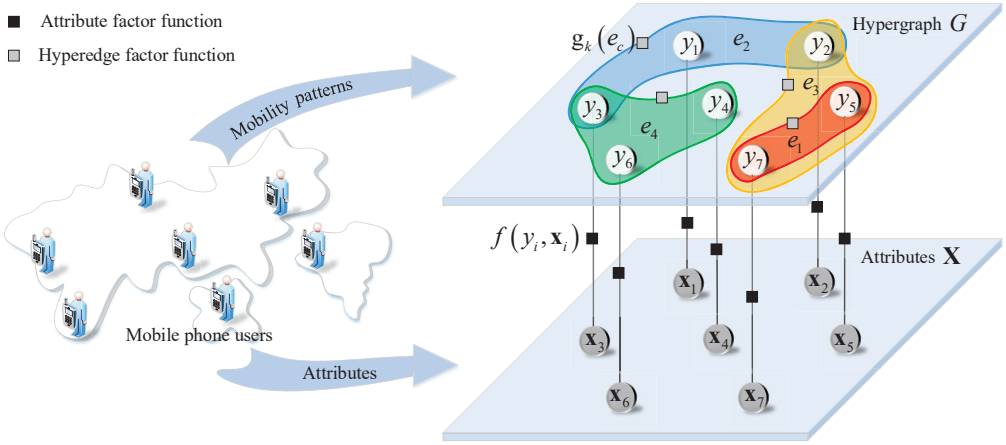


Figure 3.4: Graphical representation of the HyperFGM model.

Given the partially labeled hypergraph network, the goal of our work is to predict the labels (SES) of all SES-unknown users in the network, which is formulated as the following prediction problem.

Problem 1. **Individual Socioeconomic Status Prediction.** Given a hypergraph network $G = (V^L, V^U, E, Y^L, \mathbf{X})$, the objective is to learn a classification function:

$$f : G = (V^L, V^U, E, Y^L, \mathbf{X}) \rightarrow Y \tag{3.3}$$

As defined above, the input data is partially labeled. Therefore, to solve this problem, the HyperFGM model is learned in a semi-supervised way, i.e., exploring the labeled data as well as the unlabeled data on the hypergraph network to infer the unknown labels. Figure 3.4 shows the graphical representation of the HyperFGM model, where each user has a corresponding attribute vector $\mathbf{x}_i$ while the implicit complex relationships among users are exploited and represented on the hypergraph G. For example, y_1, y_2 and y_3 are connected by the hyperedge e_2. Furthermore, to efficiently model the power of the user attributes and the implicit high-order relationships among users, we define the following two kinds of factor functions respectively:

- **Attribute factor:** $f(y_i, \mathbf{x}_i)$ (denoted as black rectangles in Figure 3.4) represents the correlation between y_i and its attribute vector $\mathbf{x}_i$.
- **Hyperedge factor:** $g_k(e_c)$ (denoted as gray rectangles in Figure 3.4) represents the complex correlation among users, where e_c denotes the c-th hyperedge in the hypergraph and k denotes the vertex number of the hyperedge.

According to the proposed model, given a partially labeled hypergraph network $G = (V^L, V^U, E, Y^L, \mathbf{X})$, we first define the posterior probability of $P(Y|\mathbf{X}, G)$ according to Bayes' theorem as follows:

$$\begin{aligned} P(Y|\mathbf{X}, G) &= \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \\ &\propto P(\mathbf{X}|Y)P(Y|G) \\ &\propto (\prod_i P(\mathbf{x}_i|y_i))P(Y|G) \end{aligned} \tag{3.4}$$

We assume that the generative probability of user attributes given each user's label is conditionally independent, and the attributes and the network structure G are conditionally independent given labels Y. In Equation 3.4, $P(\mathbf{X}|Y)$ denotes the probability of generating the attributes $\mathbf{X}$ given their labels Y and $P(\mathbf{x}_i|y_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i; $P(Y|G)$ indicates the labels' probability in a given hypergraph network structure G.

These two kinds of factors can be instantiated in different ways. In this work, we use exponential-linear functions. Accordingly, the probability of generating attributes $\mathbf{x}_i$ given the label y_i is instantiated as:

$$P(\mathbf{x}_i|y_i) = \frac{1}{Z_\alpha} exp\{\sum_{j=1}^{m} \alpha_j f_j(y_i, x_{ij})\} \tag{3.5}$$

where $f_j(y_j, x_{ij})$ denotes the attribute factor function of an attribute x_{ij} associated with user v_i; α_j is the weight of the attribute function f_j, and Z_α is a normalization factor. $f_j(y_i, x_{ij})$ can be defined as either a binary function or a real-valued function. Without losing generality, we define it as a real-valued function, e.g., the land price of the place that user v_i visited most frequently.

For the hyperedge factor function, we define it as a binary function based on the hypergraph network. For instance, if there is a 3-node hyperedge $e_4 = \{y_3, y_4, y_6\}$ among three users in Figure 3.4, then the value of the corresponding hyperedge factor function $g_3(e_4) = 1$; otherwise 0. Hyperedges in the network can be obtained from the incidence matrix H. We accumulate all hyperedge factor functions and obtain the probability of labels given the hypergraph as follows:

$$P(Y|G) = \frac{1}{Z_\beta} exp\{\sum_{e_c \in E} \sum_k \beta_k g_k(e_c)\} \tag{3.6}$$

where $g_k(e_c)$ denotes a hyperedge factor function of a hyperedge e_c which connects k nodes (vertices), and β_k is the weight of the k-node hyperedge factor function.

According to Equations 3.4-3.6, a hypergraph-based factor graph model is constructed as follows:

$$P(Y|\mathbf{X},G)=\frac{1}{Z}exp\{\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j f_j(y_i,x_{ij})+\sum_{e_c\in E}\sum_{k}\beta_k g_k(e_c)\} \tag{3.7}$$

where $Z=Z_\alpha Z_\beta$ is a normalization factor; m denotes the length of the attribute vector $\mathbf{x}_i$; $n=|V|$ is the number of users.

The goal of learning the model is to estimate a parameter configuration $\theta=(\alpha,\beta)$, based on the input hypergraph structure and the attributes, to maximize the log-likelihood objective function $\mathcal{L}(\theta)=\log P_\theta((Y|\mathbf{X},G)$, i.e.,

$$\begin{aligned}\theta^* &= \operatorname*{argmax}_\theta \mathcal{L}(\theta)\\ &= \operatorname*{argmax}_\theta \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_j f_j(y_i,x_{ij})+\sum_{e_c\in E}\sum_{k}\beta_k g_k(e_c)-\log Z\end{aligned} \tag{3.8}$$

Solution. We use a gradient descent method (or a Newton-Raphson method) to solve the objective function. The gradient for each parameter θ is calculated as:

$$\begin{aligned}\frac{\partial\mathcal{L}(\theta)}{\partial\alpha} &= \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]-\mathbb{E}_{P_\alpha(Y)}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]\\ \frac{\partial\mathcal{L}(\theta)}{\partial\beta} &= \mathbb{E}[\sum_{e_c\in E}\sum_{k}h_k(e_c)]-\mathbb{E}_{P_\beta(Y)}[\sum_{e_c\in E}\sum_{k}g_k(e_c)]\end{aligned} \tag{3.9}$$

where $\mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]$ is the expectation of factor function $f_j(y_i,x_{ij})$ given the data distribution in the training data, and $\mathbb{E}_{P_\alpha(\mathbf{y})}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]$ is the expectation of factor function $f_j(y_i,x_{ij})$ under the distribution $P_\alpha(Y)$ (i.e., $P_\alpha(Y|\mathbf{X},G)$) given by the estimated model. For the other equation, the expectation has the similar notations.

Algorithm 2: Learning algorithm for HyperFGM

Input: attribute matrix $\mathbf{X}$, hypergraph G, learning rate η
Output: estimated parameters θ

Initialize $\theta\leftarrow 0$;
repeat
  Call LBP to calculate $\mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]$ and $\mathbb{E}_{P_\alpha(Y)}[\sum_{i=1}^{n}\sum_{j=1}^{m}f_j(y_i,x_{ij})]$;
  Call LBP to calculate $\mathbb{E}[\sum_{e_c\in E}\sum_{k}g_k(e_c)]$ and $\mathbb{E}_{P_\beta(Y)}[\sum_{e_c\in E}\sum_{k}g_k(e_c)]$;
  Compute $\frac{\partial\mathcal{L}(\theta)}{\partial\alpha}$ and $\frac{\partial\mathcal{L}(\theta)}{\partial\beta}$ according to Equation 3.9;
  Update the parameter θ with the learning rate η:

$$\alpha_{new}=\alpha_{old}-\eta\frac{\partial\mathcal{L}(\theta)}{\partial\alpha}$$
$$\beta_{new}=\beta_{old}-\eta\frac{\partial\mathcal{L}(\theta)}{\partial\beta}$$

until *Convergence;*;

As shown in Algorithm 2, to solve the intractable problem of calculating the marginal distributions (e.g., $P_\alpha(Y)$), which is caused by the arbitrariness and the possible cycles of the graphical structure in the HyperFGM, we adopt Loopy Belief Propagation (LBP) [65] to calculate the marginal probability of Y and all hyperedges E such that the gradient for each parameter can be calculated. Then, with the gradient, we update α and β with a learning rate η. With the learned parameters, we can predict the label of unknown users Y^U by finding a label configuration which maximizes the objective function, i.e., $Y^* = \operatorname{argmax} P(Y|\mathbf{X}, G)$. We need to utilize LBP to compute the marginal probability of each user $P(y_i|\mathbf{x}_i, G)$ again and then assign each user the label with the maximal marginal probability. Please notice that the proposed HyperFGM is a general framework, which can be utilized to other similar tasks with appropriate definitions of factor functions and their hypergraphs.

Finally, we present a case study to further demonstrate the proposed model. As shown in Figure 3.4, each user v_i has an attribute vector $\mathbf{x}_i$, containing SES-related attributes, and has its own mobility pattern $\mathbf{m}_i$ extracted from its mobility motifs. With LDA, each user's mobility pattern is represented as a probability distribution over some latent topics, while each topic is represented as a probability distribution over a number of mobility motifs. Then, a hypergraph is constructed based on each user's mobility pattern. For example, user v_1 has an attribute vector $\mathbf{x}_1$ and has a hyperedge e_2 to connect with v_2 and v_3, which means they have similar mobility patterns. The SES label y_1 of the user may be known or unknown according to the actual case. Next, the attribute factor and hyperedge factor are used to capture the correlations between SES and attributes and the mobility pattern relationships among users respectively. Based on Algorithm 2, the labeled and unlabeled users can be used to infer these unknown label on the hypergraph network.

3.5 Experiments

In this section, we apply the proposed HyperFGM to a real-life data for predicting individual SES levels. We first describe the experimental setup, and then report the experimental results to demonstrate the efficiency of HyperFGM compared with the baseline methods.

3.5.1 Experimental Setup

To evaluate the performance of our model, all the previous related work on SES prediction, traditional hypergraph-based methods and traditional factor graph methods are considered below for comparison.

Logistic Regression (LR): [10] relied on the elastic net model for SES prediction. We choose LR with the elastic net regularization as a baseline model for SES prediction.

SVM & Random Forest (RF): [87] utilized SVM and RF for SES prediction.

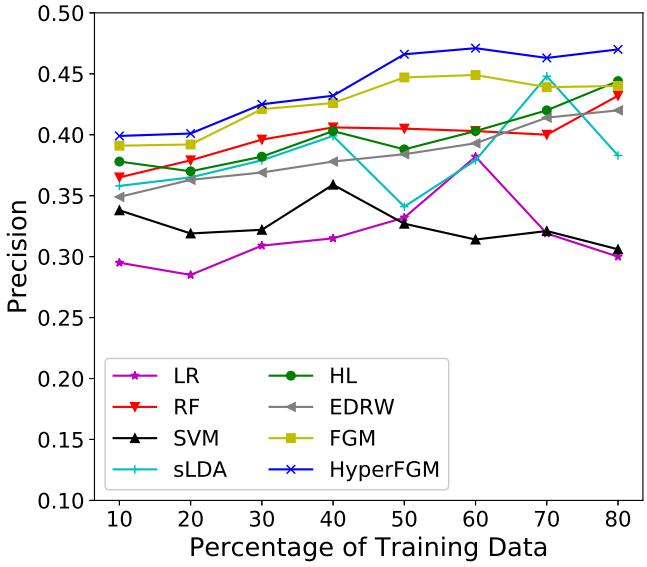


Figure 3.5: Performance (Precision) comparison of different methods with different percentages of training data.

Supervised Latent Dirichlet Allocation (sLDA): [44] employed the supervised topic model to infer SES. We use mobility pattern vectors as the input.

Hypergraph Learning (HL): A classic hypergraph learning model [33].

Extended Discriminative Random Walk (EDRW): A hypergraph-based model[80] that extends the discriminative random walk framework.

Factor Graph Model (FGM): A traditional factor graph model [95] that does not consider the implicit relationship factors.

LR, SVM and RF use the same user attributes and mobility pattern vectors as their inputs. For the hypergraph-based models HL and EDRW, two kinds of hyperedges that are respectively based on the user attributes and mobility pattern vectors are considered.

In our experiments, in order to evaluate the performance of our model with different percentages of training data (i.e., labeled data), 10% to 80% samples for each SES level are randomly selected as the labeled training data and the rest as the unlabeled testing data. More specifically, we consider several kinds of data splitting, i.e., we randomly select $k\%$ samples for each SES level as the labeled training data and the rest samples for the unlabeled testing data. In our work, we set $k = [10, 20, 30, 40, 50, 60, 70, 80]$. In order to ensure the soundness and robustness of experimental results, like the traditional evaluation method of semi-supervised method [89], this procedure with different percentages of training data repeats 10 times and we report the averaged prediction performance as final results. The prediction performance is evaluated in terms of precision, recall, and macro F1-score (F1-macro). In the presence of

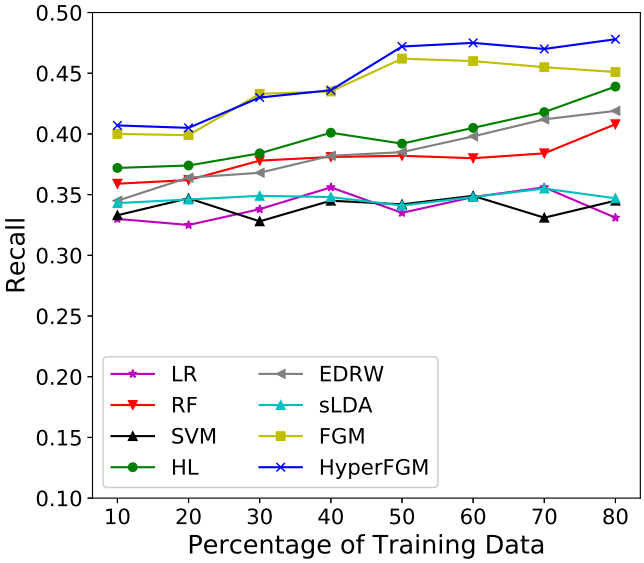


Figure 3.6: Performance (Recall) comparison of different methods with different percentages of training data.

class-imbalance, the F1-macro that balances precision and recall is deemed to be better than other measures such as accuracy [80]. For each SES level, Precision is defined as the fraction of correctly predicted positive observations over the total predicted positive observations. Recall is calculated as the number of correctly predicted positive observations divided by the number of the all observations in actual class. F1-score is the harmonic mean of Precision and Recall, which is calculated as:

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{3.10}$$

3.5.2 Prediction Performance

Figure 3.5, 3.6 and 3.7 compare the prediction performance of different methods with different training data percentage (10%-80%) in terms of precision, recall and F1-macro respectively. The proposed HyperFGM achieves the highest performance under any percentages in terms of all metrics. Specifically, on the sparse individual records, HyperFGM significantly outperforms previous models for SES prediction, i.e., LR, SVM, RF and sLDA, by 11-22%, 7-19%, 5-9% and 9-20% respectively in terms of F1-macro. There is a similar improvement in terms of precision and recall. This is because HyperFGM, taking advantage of factor graph model, can effectively capture the relations between SES and numerous SES-related attributes by the customized factor functions. In addition, thanks to the implicit high-order mobility pattern relationships among users represented on the hypergraph structure, HyperFGM outperforms FGM (with a about 2-3% higher F1-macro score). Meanwhile, the recall and precision of HyperFGM also increase with the similar improvement. Furthermore, compared with the traditional hypergraph-based methods HL and EDRW, HyperFGM also increases

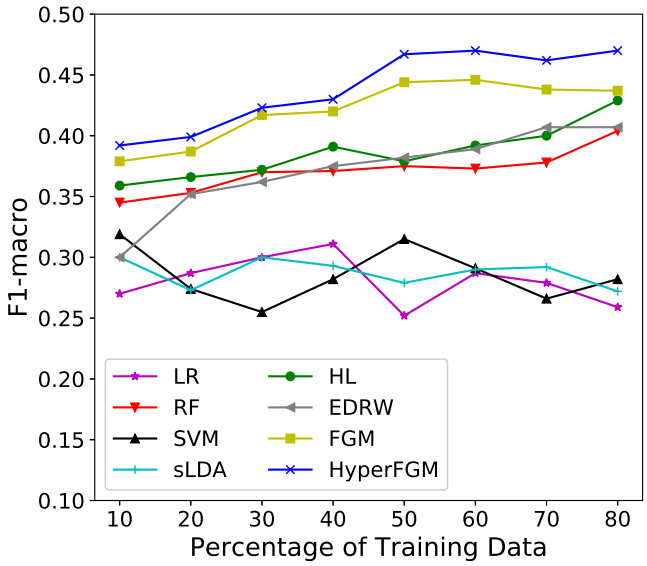


Figure 3.7: Performance (F1-macro) comparison of different methods with different percentages of training data.

2-8%, 3-9% and 3-9% in terms of precision, recall and F1-macro respectively. This is because traditional hypergraph-based methods is unable to directly represent the relations between various attributes of users and SES by the hyperedges, and they then convert numerous attributes into relationships among users, which leads to some performance loss.

Table 3.2: Performance of the prediction task for each SES level.

Models		LR	SVM	RF	sLDA	HL	EDRW	FGM	HyperFGM
Precision	A	0.188	0.153	0.264	0.146	0.270	0.243	0.357	0.396
	B	0.495	0.513	0.534	0.513	0.417	0.559	0.597	0.600
	C	0.316	0.316	0.419	0.369	0.325	0.352	0.388	0.401
	Avg	0.332	0.327	0.405	0.342	0.388	0.384	0.447	**0.466**
Recall	A	0.151	0.143	0.154	0.034	0.417	0.306	0.525	0.469
	B	0.528	0.546	0.738	0.918	0.450	0.483	0.446	0.548
	C	0.327	0.339	0.257	0.075	0.311	0.368	0.415	0.400
	Avg	0.335	0.342	0.382	0.342	0.392	0.385	0.462	**0.472**
F1-macro	A	0.115	0.127	0.192	0.054	0.324	0.210	0.423	0.428
	B	0.437	0.518	0.619	0.654	0.498	0.517	0.509	0.572
	C	0.202	0.299	0.315	0.122	0.315	0.359	0.400	0.400
	Avg	0.252	0.315	0.375	0.276	0.379	0.382	0.444	**0.467**

Performance of Each SES Level. Table 3.2 shows the prediction performance of different methods on the prediction task for each SES level. Due to the space limitation, here we only present the results in the context of taking 50% of users as training data and the rest for test. We observe that LR, SVM, RF, sLDA, HL and EDRW have much low performance on the prediction tasks of Level A and Level C while achieving relatively high performance on the Level B prediction task, which indicates that these methods may suffer from the label bias

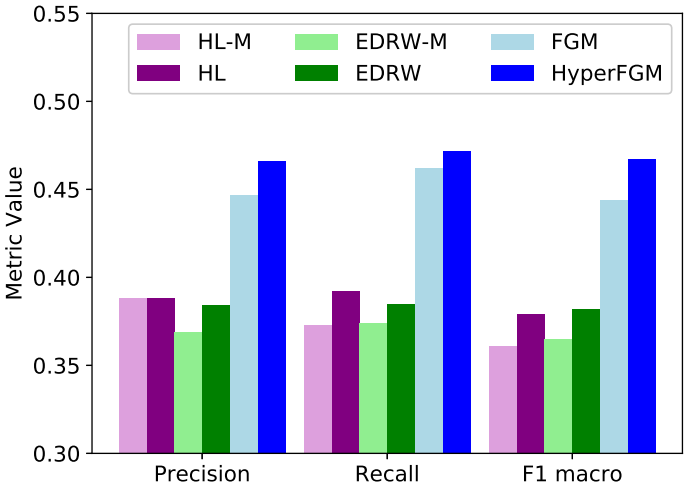


Figure 3.8: Mobility pattern relationship contribution analysis.

problem. On the contrary, FGM and HyperFGM have significantly higher performance with about 9-36% and 4-27% higher F1-macro scores in terms of Levels A and C, which shows that factor graph models handle the label imbalance problem much better. Furthermore, HyperFGM considers the attributes of users and exploits the implicit high-order relationships among users, thus achieving better performance than FGM in each SES level prediction.

Mobility Pattern Relationship Contribution Analysis. Figure 3.8 demonstrates the contribution of mobility pattern relationships in the graph-based models. Generally, the models considering the mobility pattern relationships among users mostly increase the prediction performance compared with their counterparts, i.e., HL-M, EDRW-M and FGM, which do not consider the mobility pattern relationships. Intuitively, from the social science perspective, the mobility pattern relationship factor improves the performance by bringing the prior knowledge that "the mobility patterns of users with a similar socioeconomic status tend to be similar". For example, users with similar SES have similar life style, i.e., they would work and live at the similar place areas during the similar time period. As a result, the results further prove this social science phenomenon.

Hyperedge Contribution Analysis. In this part, we evaluate the contribution of hyperedges in HyperFGM model. We implement four HyperFGM models, denoted as HyperFGM+k: HyperFGM+2 only considers pairwise (2-node) relationships; HyperFGM+3 considers 2-node and 3-node hyperedge relationships; HyperFGM+4 considers 2-node, 3-node and 4-node hyperedges; HyperFGM+5 considers more 5-node hyperedges than HyperFGM+4. We evaluate their prediction performance with the same experimental settings. We plot Figure 3.9 as an example to show the performance comparison of different versions of HyperFGM. The results show that HyperFGM+3 achieves the best performance. When considering higher-order hyperedges (i.e., $k = 4, 5$), the performance decreases; This may be because the discriminative ability of this hypergraph would be limited and even the hypergraph may confuse the correlations

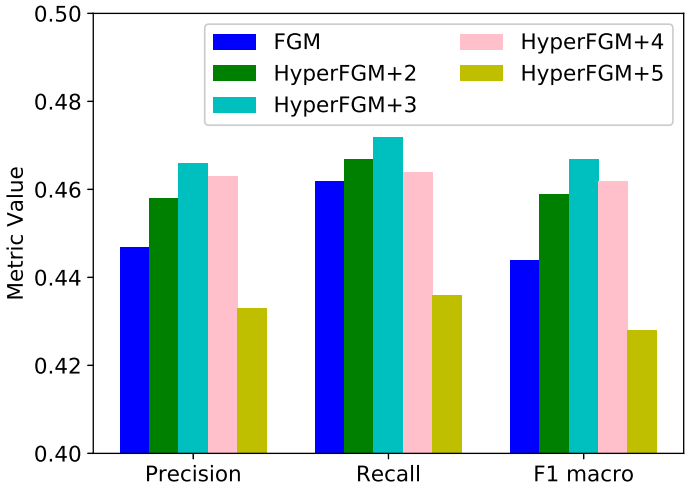


Figure 3.9: Hyperedge contribution analysis.

when each hyperedge connects to many vertices. Some work [108] has proven that the optimal k is data-dependent. This result shows the optimal k is 3 on our data. Therefore, when applying HyperFGM on other datasets, we first need to select the optimal k through grid search and use the model in other similar tasks. Compared with some other machine learning methods or deep learning methods which have many hyperparameters, our model only need to be tuned for searching one optimal hyperparameter, which simplifies the tuning procedure and decreases the tuning cost. Besides, according to previous hypergraph-based work [108], $k = 3$ always results in a good performance. Therefore, we could set $k = 3$ as default. In our future work, we plan to apply HyperFGM on different kinds of datasets to further investigate the influence of k and demonstrate the power of HyperFGM in other classification tasks.

3.5.3 Case Study

The proposed HyperFGM, as a general semi-supervised classification method, can be applied not only to the SES prediction problem but also to other similar tasks. For example, based on similar mobile phone data like CDR, with extracting related attributes and relationships this model can be utilized for mobile phone user profiling, such as occupation, income, gender, etc. Another typical use case is to infer user demographics based on their social media data. For instance, besides social media users' attributes, with customized factor functions HyperFGM can take into account various high-order relationships based on online behavior pattern similarity, e.g., following the similar users or mentioning the similar topics. Consequently, the proposed HyperFGM can be used in a classification problem, where each object has attributes while there exist explicit or implicit relationships among objects.

Compared with traditional method, e.g., demographic census, estimating individual socioeconomic status based on their own real-time mobile phone usage data provides a much more

real-time and cheaper method, which can benefit a wide range of applications. In order to further demonstrate the social and economic impact of this work, we take several specific case studies to show the practical value of our work.

From a commercial perspective, estimating users' SES in real time can assist in capturing each user's social and economic factors, such as income, wealth, education, health, which can improve many business applications. For example, [102] has shown that consumer's perceptions of food safety vary with socio-economic status and consumer may concern more about ingredient, ecology and food culture when purchasing food. Thus, food businesses can estimate the perception degrees of food safety of potential consumers according to related sociological achievements [102] and then recommend different kinds of food to different groups of persons by advertising. Another example may be that assessing individual SES can help banks and finance companies estimate users' credit risk index. In a word, companies can more efficiently recommend different levels of services and products to consumers with different SES. Furthermore, obtaining the personal SES distribution of each area or community can help companies select more suitable sites to start their business.

From a social and economic perspective, previous sociological articles have investigated the social and economic value of predicting SES. Measuring SES can not only help capture and understand changes to the structure of a society, but also assist in investigating the relationship between other important social variables. In addition, predicting SES can assist in studying and making public polices in many fields, such as economic, education, health. For instance, regarding strong relationship between SES and health [3], assessing SES can help make sound policy decisions for health care.

3.6 Chapter Summary

In social science and public services, precisely assessing individual SES is very critical for informing public policy-making, which is yet very costly and challenging. With the advancement of AI techniques and availability of mobile phone data, existing work studied region/household-level SES assessment using mobile phone data. Compared with previous work, this chapter takes a new attempt to predict individual SES on mobile phone data, which aims to provide richer insight about the relations between SES and personal attributes and networking while also address the issues in existing work on SES prediction and direct applications of existing analytic methods. A semi-supervised Hypergraph-based Factor Graph Model (HyperFGM) is introduced to leverage customized factor functions on a hypergraph structure. It effectively captures the influence of user attributes and the implicit high-order mobility pattern relationships among users on SES. HyperFGM handles both labeled and unlabeled data in a semi-supervised way. HyperFGM is tested on a set of anonymized real-life mobile phone data and sociological domain knowledge for SES labeling. The extensive experiments demonstrate that HyperFGM provides more reasonable individual SES prediction results than all existing work on SES

prediction, and also achieves better performance than the state-of-the-art hypergraph-based methods and factor graph methods.

Chapter 4

Predicting Individual Socioeconomic Status based on Social Media Data

This chapter investigates the problem of predicting the socioeconomic status of social media users based on their social media content. The increasing popularity of social media, especially microblogging service, attracts billions of users, generating amounts of various user-generated data. These social media data record users' daily behaviors, which are becoming a bridge between the physical daily life and online behaviors. Regarding the rich information of social media data, some efforts have been made to predict SES-related information. Currently, most existing works leverage manually defined textual features and platform-based user level attributes that are extracted from social media content and then feed them into a machine learning based classifier for individual SES prediction. However, they ignore some key information of social media content, including the order and the hierarchical structure of social media text, and the relationships among user level attributes. To this end, this chapter proposes a novel coupled social media content representation model for individual SES prediction. The proposed model not only utilizes a hierarchical neural network to incorporate the order and the hierarchical structure of social media text but also employs a coupled attribute representation method to take into account intra-coupled and inter-coupled interaction relationships among platform-based user level attributes. To validate the efficiency and robustness of the proposed model, the experimental results demonstrate that the proposed model significantly outperforms other state-of-the-art models on a real Sina Weibo dataset.

Contents

4.1 Introduction

Predicting individual socioeconomic status (SES) from social media content recently has become an important research area. SES characterizes a person's economic and social position in relation to others, which is typically divided into three levels [84]. As an access to financial, social and human capital resources, inferring individual SES not only provides governments and research organizations with tools for studying and making public policies on a large scale population, but also helps promote online marketing and advertising by the analysis of user's purchasing power. It also benefits a wide range of other fields, such as education [103, 71], health [82] and public transportation [19]. With the worldwide ubiquity of online social media, especially microblogging platforms, online social media content generated by social media users has been used in recent research for population informatics in demographics [75, 15, 36], economics [11], social science [92, 55] and other research domains [24, 53, 54]. In consideration of the significance of SES, this work focuses on predicting SES of social media users based on their social media content.

Previous related work have looked into predicting individual socioeconomic information based on social media content, such as inferring occupation category [69], SES [55] and income [70] of social media users. In these works, they devote to manually design several kinds of user level attributes and textual features, such as n-grams, from social media content, and then feed all the features into a machine learning based classifier for prediction. However, the prediction performance of these models heavily depends on the extracted features, which need effective feature engineering. Furthermore, with extracted textual features, they ignore some important information for representing social media text (i.e., text in social media content), including the order of words and microblogs, and the hierarchical structure (words form microblogs, microblogs form social media text of a user). Besides, in the real world, attributes are more or less interacted and coupled via explicit or implicit relationships [96]. For example, business and social applications always see quantitative attributes coupled with each other [18]. However, previous work extract the user level attributes without considering relations among them, which leads to limited performance.

Motivated by the great success of deep learning in many fields, such as computer vision [52] and natural language processing [6], recent work use neural networks to learn text representation without any feature engineering and mostly achieve significantly higher performance compared with traditional machine learning methods. To this end, this chapter proposes a coupled social media content representation model for SES prediction, utilizing neural network for individual SES prediction from social media content, which is the first trial in this community as far as we know. Like previous related work, as shown in Figure 4.1, this work also regards social media text and platform-based user level attributes as social media content, which are ubiquitous in social media. Meanwhile, this work focuses microblogging platform as a use case study. To be more specific, first, in order to be able to consider the order of words and microblogs in social media text, we propose to employ Bidirectional Long Short-Term Memory (BiLSTM)

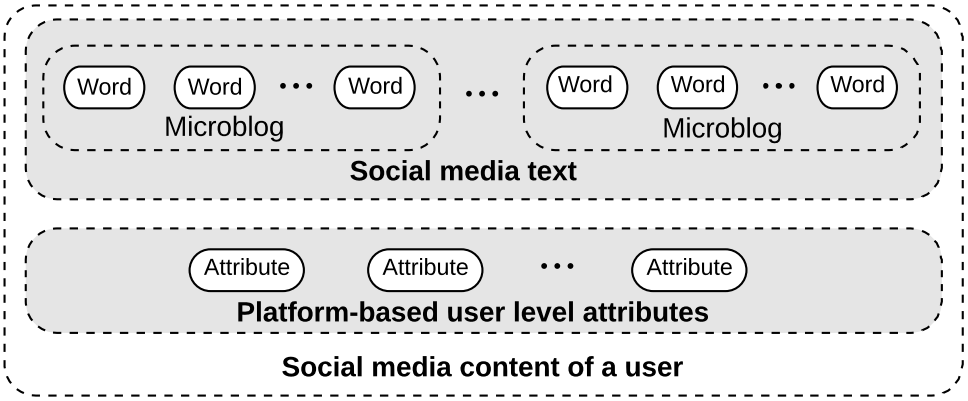


Figure 4.1: The architecture of social media content.

network, a variation of Recurrent neural networks (RNN), to represent social media text due to its representational power and effectiveness at capturing long-term dependencies of a sequence. Second, since social media text has a hierarchical structure, we likewise construct a social media text representation by first building representations of microblogs with corresponding words and then aggregating those into a social media text representation. Third, to consider the dependency of attributes, we devise an attribute coupled representation using intra-coupled interaction (i.e., the correlations between attributes and their own powers) and inter-coupled interaction (i.e., the correlations between attributes and the powers of others) [96]. Finally, we learn a joint social media content representation with aggregating social media text representation and platform-based user level attribute representation.

This work is applied to the microblogging platform of Sina Weibo [81], a Chinese microblogging website. We first build a new data set of Sina Weibo users with a SES label for each of them. To demonstrate the feasibility and efficiency of the proposed model on individual SES prediction, we apply the proposed model to the data set. Experimental results demonstrate that our proposed model significantly outperforms the baseline models in previous related work.

To sum, the main contributions of this work in Chapter 4 are as follows:

- We propose a novel coupled social media content representation approach for individual SES prediction, which utilizes neural network to integrate social media text and platform-based user level attributes. To our best knowledge, this is the first try in this community.
- We present a social media text representation method, which utilizes hierarchical recurrent neural network to take into account the order of words and microblogs as well as the hierarchical structure of social media text.
- We proposed to employ a coupled attribute representation method to analyze the intra-coupled and inter-coupled interaction among user level attributes, which can successfully capture the intrinsic couplings for SES prediction.

- We build a data set of Sina Weibo users with a SES label for each of them and demonstrate the power of the proposed model using this data set. Substantial experiments demonstrate that the proposed model significantly outperforms the state-of-the-art models.

The rest of this chapter is organized as follows. In Section 4.2, we briefly introduce the related work including socioeconomic-related information prediction based on social media and representation learning of social media content. Section 4.3 presents the proposed model in details. In Section 4.4, we introduce the data collection and preprocessing. The efficiency and robustness of our proposed model is demonstrated with experimental evaluation in Section 4.5. Finally, we conclude this chapter in Section 4.6.

4.2 Related Work

This section discusses two closely related work, including socioeconomic-related information prediction based on social media data and representation learning of social media content.

4.2.1 Socioeconomic-related Information Prediction based on Social Media Data

Socioeconomic attributes prediction from social media content has been studied in the past few years. [69] focuses on inferring the occupational class of Twitter user, in which they first extracted latent user level features and textual features such as word clusters and embeddings, and then employed a non-linear method Gaussian Process (GP) for classification. [70] presented a study if user behavior on Twitter can be used to build a predictive model of income. They designed different feature categories and used GPs for user income prediction, which achieves strong correlation between predicted and actual user income. The most similar work to ours is [55]. They also used a similar methodology, where they extracted several kinds of features from posted text and platform-related attributes in Twitter to represent each user and used a composite Gaussian Process model to infer the SES of Twitter users. However, these methods only consider the predefined features with feature engineering, which cannot capture the heterogeneous couplings among the social media text and platform-based user level attributes.

There are also several works based on additional sources, such as social networks [4] and geolocation information [12], which are different from our task since our work only focuses on social media content. A small body of research has focused on analyzing socioeconomic attributes based on some potential factors. For example, [39] first utilized a weakly supervised learning method to automatically identify the temporal orientation of tweets on Twitter and quantify a user's income based on overall temporal orientation.

4.2.2 Social Media Content Representation Learning

The main purpose of this work is to learn a good representation of social media content which contains text and user level attributes. There is a large body of existing work on text-based representation learning for various applications, such as sentiment classification, rumor detection, and user profiling. Early approaches devote to design effective features from text as representations and use machine learning algorithms to build classifiers with text features. Representative text features include word n-grams [98], text topic [32], bag-of-opinions [72], sentiment lexicon features [50]. However, this kind of methods are labor intensive and unable to extract the enough information from data for representation.

Motivated by the great success of deep learning in many fields, such as computer vision [52] and natural language processing [6], more recent work use neural networks to learn text-based social media content representation without any feature engineering and mostly achieve significantly higher performances compare with traditional machine learning methods. However, existing relevant social media content representation approaches can only lead to limited improvement for individual SES prediction as they only consider a part of complex couplings among such heterogeneous data. A large number of methods mainly focus on pure text representation. For instance, [90] designed a gated recurrent neural network to learn vector-based document representation in a unified and bottom-up fashion for sentiment classification. [106] proposed a hierarchical attention network for document classification inspired by the hierarchical structure of documents, which only capture the hierarchical couplings of textual data. Despite these approaches capture the heterogeneous couplings of textual data, they ignore the effects of social media text's attributes. Considering the potential effects of these attributes, a part of research work introduce several attributes into text data. For example, [47] proposed to use Recurrent Neural Network to fusing textual and social context features through directly concatenating textual embedding and numerical features. Like most of existing fusion methods, they do not consider couplings within and between them. Besides, several recent methods propose to fuse additional sources like image and video into social media content [47, 109, 29]. The key difference between our work and this strand of previous work lies in that we focus on learning the representation of general social media content, i.e., social media text and platform-based user level attributes.

4.3 The Proposed Model

This work aims at predicting individual SES based on their social media content over a given past period. For the generalization like previous related work [69, 55], this work regards the social media text and the user level platform-based attributes as social media content of a user since these data are ubiquitous in social media.

4.3.1 Problem Statement

Regarding the social media content, each user has social media text and platform-based user level attributes. Assume that a social media user $u \in U$ has a set of posted microblogs $B = \{b_1, b_2, ..., b_n\}$ and the i-th microblog $b_i \in B$ contains a sequence of words $\{w_1^i, w_2^i, ...w_{l_i}^i\}$, where l_i is the length of i-th microblog. Additionally, user u has a set of platform-based user level attributes $\{a_1, a_2, ..., a_m\}$, where m is the number of user level attributes. For each user, the proposed model aims at projecting the raw social media content into a vector representation, on which we build a classifier to perform individual SES prediction. In a word, the purpose of this work is to build a social media content representation model that can represent as much information as possible.

4.3.2 Coupled Social Media Content Representation Model

In this part, we first present the social media text representation method and coupled user level attribute representation method. Then, the social media text representation and platform-based user level attribute representation are aggregated into a vector representation of social media content. Finally, based on the social media content representation, we build a 3-way classifier to assign SES label to each social media user.

Social Media Text Representation. Long Short-Term Memory (LSTM) [43], a variation of RNN, is widely adopted for textual data modeling due to its excellent performance on sequence modeling. LSTM is able to consider long-term dependencies of a sequence through introducing a memory cell. To model the semantic representation of social media text and consider the order of text, we adopt BiLSTM (Bidirectional LSTM) to represent the social media text both from forward and backward, which can increase the amount of input information available to the network compared with LSTM. Besides, to take into account the hierarchical structure of social media text, inspired by the principle of compositionality [30], we model a social media user's text through a hierarchical structure composed of three levels, i.e., word-level, microblog-level and user-level.

As shown in Figure 4.2, in the word level, we first embed each word in a microblog b_i into a low dimensional semantic space, i.e., each word w_j^i is mapped to its embedding $\boldsymbol{w}_j^i \in \mathbb{R}^d$. The word embedding method and its settings will be described in Section 4.5.1. At each step, given

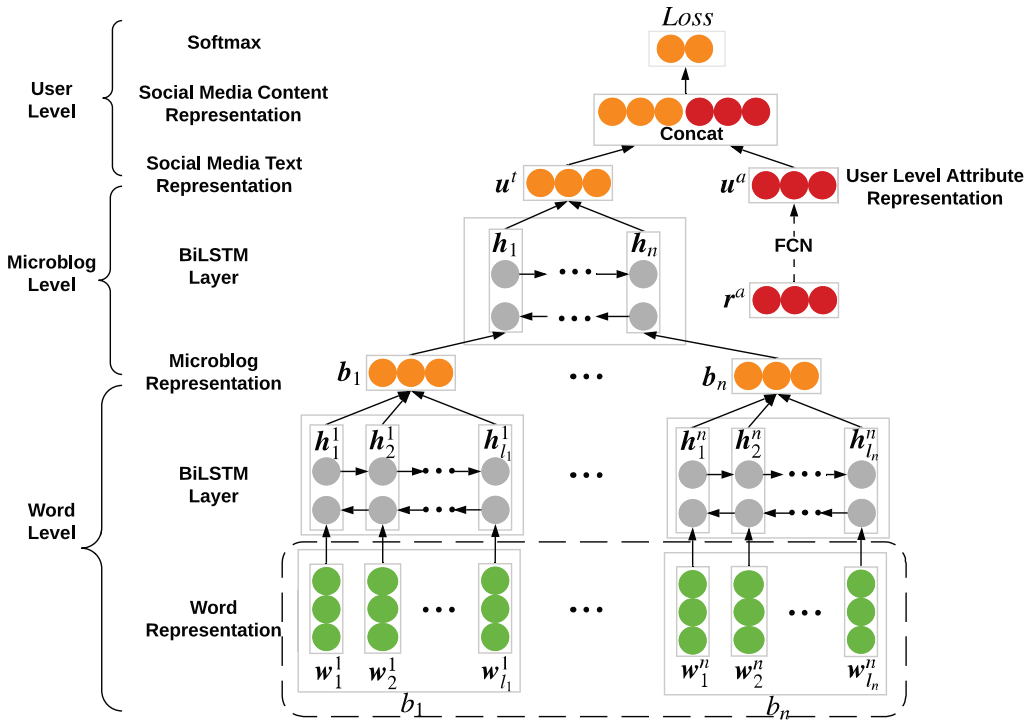


Figure 4.2: The architecture of the proposed model.

an input word embedding $\boldsymbol{w}_j^i$, the current cell state $\boldsymbol{c}_j^i$ and hidden state $\boldsymbol{h}_j^i$ can be updated with the previous cell state $\boldsymbol{c}_{j-1}^i$ and hidden state $\boldsymbol{h}_{j-1}^i$ as follows:

$$\begin{bmatrix} \boldsymbol{i}_j^i \\ \boldsymbol{f}_j^i \\ \boldsymbol{o}_j^i \end{bmatrix} = \begin{bmatrix} \boldsymbol{\sigma} \\ \boldsymbol{\sigma} \\ \boldsymbol{\sigma} \end{bmatrix} (\boldsymbol{W}[\boldsymbol{h}_{j-1}^i, \boldsymbol{w}_j^i] + \boldsymbol{b}), \tag{4.1}$$

$$\hat{\boldsymbol{c}}_j^i = \tanh(\boldsymbol{W}[\boldsymbol{h}_{j-1}^i, \boldsymbol{w}_j^i] + \boldsymbol{b}), \tag{4.2}$$

$$\boldsymbol{c}_j^i = \boldsymbol{f}_j^i \odot \boldsymbol{c}_{j-1}^i + \boldsymbol{i}_j^i \odot \hat{\boldsymbol{c}}_{j-1}^i, \tag{4.3}$$

$$\boldsymbol{h}_j^i = \boldsymbol{o}_j^i \odot \tanh(\boldsymbol{c}_j^i), \tag{4.4}$$

where $\boldsymbol{i}$, $\boldsymbol{f}$, $\boldsymbol{o}$ indicates gate activations, $\odot$ denotes element-wise multiplication, σ is the logistic sigmoid function and $\boldsymbol{W}$, $\boldsymbol{b}$ are the trainable parameters. Therefore, for a sequence of words $\{w_1^i, w_2^i, ..., w_{l_i}^i\}$, the forward LSTM reads the word sequence from w_1^i to $w_{l_i}^i$ and the backward LSTM reads the word sequence from $w_{l_i}^i$ to w_1^i. Then we concatenate the forward hidden state $\overrightarrow{\boldsymbol{h}_j^i}$ and the backward hidden state $\overleftarrow{\boldsymbol{h}_j^i}$, i.e., $\boldsymbol{h}_j^i = [\overrightarrow{\boldsymbol{h}_j^i}; \overleftarrow{\boldsymbol{h}_j^i}]$, where $[.;.]$ denotes the concatenation operation. In BiLSTM, the hidden state $\boldsymbol{h}_j^i$ denotes the information of the whole sequence centered around $\boldsymbol{w}_j^i$. As a result, the BiLSTM network receives $[\boldsymbol{w}_1^i, \boldsymbol{w}_2^i, ..., \boldsymbol{w}_{l_i}^i]$ and generates hidden states $[\boldsymbol{h}_1^i, \boldsymbol{h}_2^i, ..., \boldsymbol{h}_{l_i}^i]$. Then we feed the hidden states to an average pooling layer to obtain the microblog text representation $\boldsymbol{b}_i$ for microblog b_i.

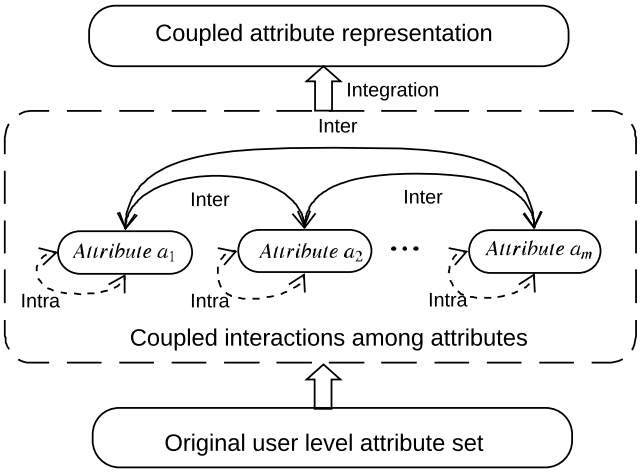


Figure 4.3: An overview of coupled user level attribute representation.

In the microblog level, given the microblog representation vectors of a user $\{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_n\}$, we also utilize BiLSTM to encode the microblogs as follows:

$$\overrightarrow{\boldsymbol{h}_i} = \overrightarrow{LSTM(\boldsymbol{b}_i)}, \tag{4.5}$$

$$\overleftarrow{\boldsymbol{h}_i} = \overleftarrow{LSTM(\boldsymbol{b}_i)}, \tag{4.6}$$

We then concatenate the forward hidden state $\overrightarrow{\boldsymbol{h}_i}$ and the backward hidden state $\overleftarrow{\boldsymbol{h}_i}$, i.e., $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}_i}; \overleftarrow{\boldsymbol{h}_i}]$. $\boldsymbol{h}_i$ summarizes the neighbor microblogs around the i-th microblog but still focus on the i-th microblog. Then we feed the hidden states to an average pooling layer to obtain the final social media text representation $\boldsymbol{u}^t$ for user u.

Coupled User Level Attribute Representation. Besides social media text, each social media user generally has platform-based user level attributes. For example, some attributes like the number of followees indicate platform impact, some like the number of microblogs indicate platform behaviors. Like previous related work, we assume that these user level attributes could make a contribution to the representation of social media content for individual SES prediction. To our best knowledge, most previous works only leverage original user level attributes without considering relations among attributes. However, inspired by previous work [18, 96], in the real word, attributes are more or less coupled via explicit or implicit relationships. Therefore, it is natural to hypothesize that the user level attributes are related to each other in some way. To this end, this work proposes to employ a coupled representation method [96] to represent user level attributes, which is able to capture such latent relations among attributes.

To be more specific, as illustrated in Figure 4.3, we consider two kinds of interaction relations among platform-based user level attributes: the intra-coupled interaction within an attribute with the correlations between every attribute and its own powers, and the inter-coupled interaction among different attributes with the correlations between each attribute and the powers of other attributes.

Firstly, we map the original attribute space to an expanded space for incorporating linear and nonlinear information by means of a power expansion as follows:

$$\{\langle a_1\rangle^1, \langle a_1\rangle^2, ..., \langle a_1\rangle^L, \langle a_2\rangle^1, \langle a_2\rangle^2, ..., \langle a_2\rangle^L, ..., \langle a_m\rangle^1, \langle a_m\rangle^2, ..., \langle a_m\rangle^L\} \tag{4.7}$$

where $\langle a_j\rangle^p (1 \leq p \leq L, p \in \mathbb{Z}, 1 \leq j \leq m)$ denotes the p-th power of the corresponding value of attribute a_j.

Leveraging the power expansion, the intra-coupled interaction within attribute a_j^n is defined as an $L \times L$ matrix $\boldsymbol{M}_{Ia}(a_j)$, with considering the correlations between attribute a_j and its own powers $\langle a_j\rangle^p$.

$$\boldsymbol{M}_{Ia}(a_j) = \begin{pmatrix} \theta_{11}(j) & \theta_{12}(j) & \cdots & \theta_{1L}(j) \\ \theta_{21}(j) & \theta_{22}(j) & \cdots & \theta_{2L}(j) \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{L1}(j) & \theta_{L2}(j) & \cdots & \theta_{LL}(j) \end{pmatrix}, \tag{4.8}$$

where $\theta_{pq}(j)$ denotes the Pearson's product-moment correlation coefficient between $\langle a_j\rangle^p$ and $\langle a_j\rangle^q$. Here, we use the revised correlation coefficient by taking account of the p-values for testing the hypothesis of no correlation between attributes, i.e., if p-value is no less than 0.05, the correlation coefficient is set to 0.

Besides, the inter-coupled interaction between numerical attribute a_j and other attributes a_k $(k \neq j)$ is defined as an $L \times L \cdot (m-1)$ matrix $\boldsymbol{M}_{Ie}(a_j|\{a_k\}_{k\neq j})$.

$$\boldsymbol{M}_{Ie}(a_j|\{a_k\}_{k\neq j}) = \begin{pmatrix} \delta_{11}(j,k_1) & \cdots & \delta_{1L}(j,k_1) & \cdots & \delta_{11}(j,k_{m-1}) & \cdots & \delta_{1L}(j,k_{m-1}) \\ \delta_{21}(j,k_1) & \cdots & \delta_{2L}(j,k_1) & \cdots & \delta_{21}(j,k_{m-1}) & \cdots & \delta_{2L}(j,k_{m-1}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \delta_{L1}(j,k_1) & \cdots & \delta_{LL}(j,k_1) & \cdots & \delta_{L1}(j,k_{m-1}) & \cdots & \delta_{LL}(j,k_{m-1}) \end{pmatrix}, \tag{4.9}$$

where $\delta_{pq}(j,k_i)$ denotes the Pearson's product-moment correlation coefficient between $\langle a_j\rangle^p$ and $\langle a_{k_i}\rangle^q$, and $\{a_k\}_{k\neq j} = \{a_{k_1}, ..., a_{k_{m-1}}\}$ is the set of attributes other than a_j.

For each user object u_i, the attribute values of a_j and its powers are presented as a vector:

$$\widetilde{\boldsymbol{z}}_{\boldsymbol{i}}(a_j) = [\langle v_{ij}\rangle^1, \langle v_{ij}\rangle^2, ..., \langle v_{ij}\rangle^L], \tag{4.10}$$

while the attribute values of other attributes $\{a_k\}_{k\neq j}$ and their powers are denoted as another vector:

$$\widetilde{\boldsymbol{z}}_{\boldsymbol{i}}(\{a_k\}_{k\neq j}) = [\langle v_{ik_1}\rangle^1, \langle v_{ik_1}\rangle^2, ..., \langle v_{ik_1}\rangle^L, ..., \langle v_{ik_{m-1}}\rangle^1, \langle v_{ik_{m-1}}\rangle^2, ..., \langle v_{ik_{m-1}}\rangle^L]. \tag{4.11}$$

Here, the attribute value of user u_i on attribute a_j is v_{ij}. We incorporate the intra-coupled interaction and the inter-coupled interaction into a new coupled attribute representation, a $1 \times L$ vector $\boldsymbol{r}_i(a_j)$, for user object u_i on the numerical attribute a_j as follows:

$$\begin{aligned}\boldsymbol{r}_i(a_j) =& \widetilde{\boldsymbol{z}}_{\boldsymbol{i}}(a_j) \odot \boldsymbol{w} \otimes [\boldsymbol{M}_{Ia}^n(a_j)]^T \\ &+ \widetilde{\boldsymbol{z}}_{\boldsymbol{i}}(\{a_k\}_{k \neq j}) \odot \underbrace{[\boldsymbol{w}, \boldsymbol{w}, \cdots, \boldsymbol{w}]}_{m-1} \otimes [\boldsymbol{M}_{Ie}^n(a_j | \{a_k\}_{k \neq j})]^T,\end{aligned} \tag{4.12}$$

where $\boldsymbol{w} = [1/(1!), 1/(2!), \cdots, 1/(L!)]$, $\odot$ denotes the Hadamard product and $\otimes$ indicates the matrix multiplication. After considering all the d_n original numerical attributes, we obtain the final coupled user level attribute representation for the user object u_i as follows:

$$\boldsymbol{r}_i^a = [\boldsymbol{r}_i(a_1), \boldsymbol{r}_i(a_2), \cdots, \boldsymbol{r}_i(a_m)] \in \mathbb{R}^{L \cdot m} \tag{4.13}$$

Before fusing the user level attributes, to capture the latent relationships between high level features, we link the raw attribute vector $\boldsymbol{r}^a$ to the k-length representation vector $\boldsymbol{u}^a$ in terms of a fully connected network as follows:

$$\boldsymbol{u}^a = \boldsymbol{r}^a \cdot \boldsymbol{W}_a \tag{4.14}$$

where the weight $\boldsymbol{W}_a$ encodes the interaction strength over attributes in the fully-connected layer.

Consequently, in the user level, we aggregate user level attributes and social media text into a representation vector. More specifically, we concatenate the social media text representation and the coupled user level attribute representation to obtain the social media content representation $\boldsymbol{u} = [\boldsymbol{u}^t; \boldsymbol{u}^a]$.

Individual SES Prediction based on Social Media Content. Given the high level representation of social media content, we employ a linear layer and a softmax layer to project the social media content representation $\boldsymbol{u}$ into SES distribution of C classes as follows:

$$p_c = softmax(\boldsymbol{W}\boldsymbol{u} + b). \tag{4.15}$$

where p_c is the predicted probability of SES label c. In this model, the cross-entropy error between ground truth SES level distribution and predicted SES level distribution is defined as loss function for optimization when training:

$$L = -\sum_{u \in U} \sum_{c=1}^{C} p_c^g(u) \cdot \log(p_c(u)), \tag{4.16}$$

where p_c^g denotes the gold probability of SES label c with ground truth being 1 and others being 0, and U represents the training social media users.

Figure 4.4: A demonstration of user search function in Sina Weibo.

4.4 Data Collection and Preprocessing

The work in this chapter aims at predicting individual SES from personal social media content. To this end, we need to create a data set which contain social media users' content and convincing SES labels for social media users for this task. This section presents the data collection and preprocessing in details.

4.4.1 Data Collection

In the field of sociology, many articles have shown that socioeconomic index is highly associate with occupational status [8, 93, 41] and there exist some mapping between SES and occupations like the Standard Occupation Classification hierarchy attached to socioeconomic categorisations in conjunction with the National Statistics Socio-Economic Classification [27, 79].

To create the data set that can be used for predicting SES of social media users, according to the China Occupation Classification, for each major occupation we queried Sina Weibo's search API to retrieve a maximum of 500 user accounts whose certificated person card best matched the occupation keywords. As shown in Figure 4.4, after we search users using the occupation keyword "CEO", the best matched users whose certified person card contain the keyword are listed. To remove potential ambiguity in the raw user set, we manually inspect

Figure 4.5: A sample of social media content in Sina Weibo.

accounts and filtered out those users who belong to companies and other occupations. After that, we collected the rest users' microblogs posted from February 2017 to February 2018 and their platform-based attributes. Figure 4.5 demonstrates a sample of social media content of a user in Sina Weibo. Finally, we extracted active users who published more than ten microblogs during this given period. In total, about 50% of the accounts were removed after this filtering process. As a result, the final data set consists of 20452 users from 73 occupations and 6893746 unique microblogs. To obtain a SES label for each user, we invited several sociologists who study the socioeconomic index of occupations in China to assign a high (level A), middle (level B) or low (level C) SES to each user in our data set. The distribution of users across classes is 3974 users with level A, 12451 users with level B and 4027 users with level C. Finally, after undersampling for ensuring data set balance, the final distribution of users across classes is 3974 users with level A, 4004 users with level B and 4027 users with level C.

4.4.2 Data Preprocessing

For the data set, we need to do some data preprocessing for each user's social media text and user level attributes. With regard to the textual data, we remove punctuation, non-Chinese words, digits, and specific symbols meanwhile we convert all the traditional Chinese words in the social media text into simplified Chinese words. Then we choose to leverage a Chinese Language Technology Platform (LTP) [22], an integrated Chinese processing platform which includes a suite of high performance natural language processing (NLP) modules and relevant corpora, to segment Chinese text of each microblog into a sequence of Chinese words. Besides,

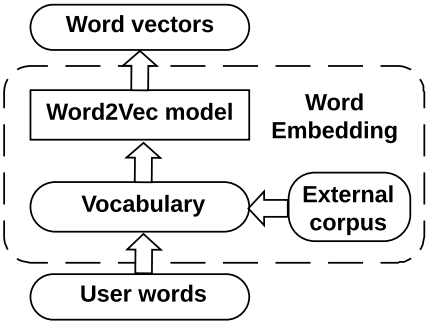


Figure 4.6: The procedure of word embedding.

a separate Chinese Wikipedia data set [23] is used as a reference corpus in order to build the word embedding representations. Regarding the platform-based user level attribute, in this work, we extract seven user level attributes, i.e., the number of followers, followees, microblogs, proportion of forwarded microblogs, the average number of favorites, forwarded, comments per microblog.

4.5 Experiments and Evaluation

In this section, we conduct extensive experiments on our crawled Sina Weibo dataset to demonstrate the efficiency and robustness of the proposed model.

4.5.1 Experimental Settings

For the textual data in the social media content, we employ the distributed representation for words [64] as shown in Figure 4.6. We only retain words appearing more than 5 times in building the vocabulary with our whole textual dataset and a separate Chinese Wikipedia dataset. Then we pre-train the Word2Vec model with the vocabulary in an unsupervised fashion with default parameter settings. Finally, we obtain a 50-dimensional word embedding vector for each word in the dataset. The word embedding is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, and so on.

Regarding the user-level attributes, like the word embedding method, we introduce the coupled attribute representation for the user level attributes. To take advantage of information from testing objects, we use both the training and testing objects' attributes in the coupled attribute representation within an unsupervised manner. For the user level attributes, the original attribute dimension is 7. We set the power expansion value $L = 6$ so that the extracted coupled attribute dimension is 42.

In the experiments, we set the dimension of the hidden states in LSTM cell to be 32 so that a combination of forward and backward LSTM gives us 64 dimensions for microblog and social media content annotation. In order to speed up training, we limit that the maximum length of every microblog is 40 words and a social media user has 50 microblogs at most. We use

Table 4.1: SES prediction performance for the baseline models and the proposed model.

Models	Accuracy	Precision	Recall	F1-score
Feature engineering based Models				
LR	0.3972	0.3461	0.3973	0.2951
SVM	0.4832	0.4484	0.4825	0.4330
GP	0.5563	0.5544	0.5560	0.5519
Neural network based Models				
RNN	0.6215	0.6280	0.6212	0.6168
HRNN	0.6323	0.6471	0.6319	0.6274
AHRNN	0.6498	0.6777	0.6493	0.6413
CAHRNN	**0.6689**	**0.6880**	**0.6684**	**0.6611**

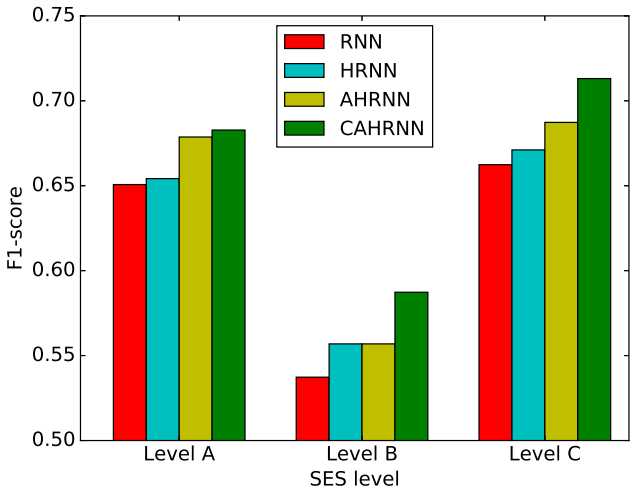


Figure 4.7: Performance comparison for each each SES level.

Adam [49] to update parameters with setting initial learning rate as 0.0001 when training. We use 80% of the data for training and the remaining 20% for testing.

4.5.2 Performance Comparison

To validate the proposed model on individual SES prediction based on their social media content, we compare the proposed model with two groups of state-of-the-art methods. The first group consists of feature engineering based models. To be more specific, we compare it with previous machine learning based methods [55, 69, 70]. These methods first extract several kinds of features, which contains platform-based user level attributes and textual features extracted from social media text (i.e., the frequency of the 1-grams and the frequency distribution across latent topics represented by clusters of 1-grams [55]). Then, they apply common machine learning methods, containing logistic regression (**LR**) with Elastic Net regularization, Support Vector Machine (**SVM**), and Gaussian Process (**GP**).

Table 4.2: Performance of the proposed model for each SES level.

	Level A	**Level B**	**Level C**
Precision	0.7192	0.7400	0.6050
Recall	0.6500	0.4868	0.8684
F1-score	0.6828	0.5873	0.7131

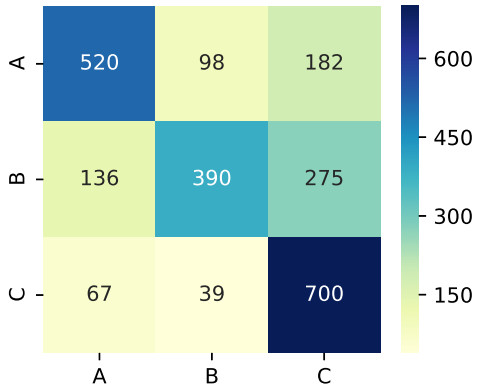


Figure 4.8: The confusion matrix for SES prediction. Rows represent the actual SES level (A, B, C) and columns represent the predicted SES level.

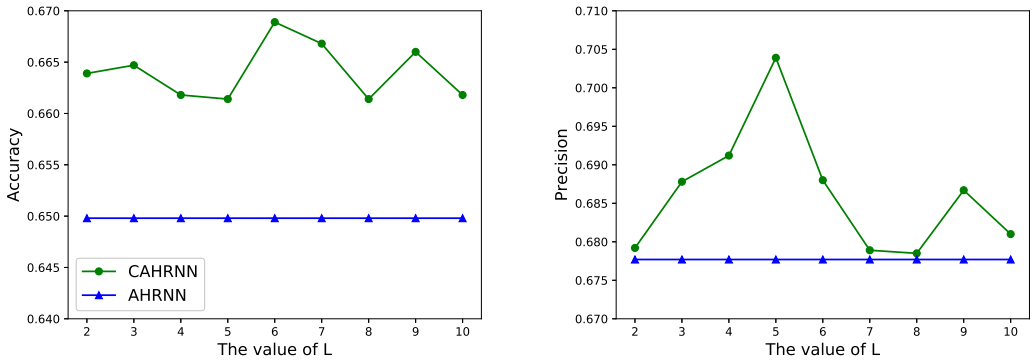
The other group is composed of neural network based methods, which are widely leveraged in recent text classification related work. As we know, there have been many kinds of neural networks proposed for text-based classification. In this work, we focus on the coupling methods used in these works not the neural network itself. Hence, we chose the following methods as baselines:

RNN represents each word with the word embedding vector and feeds each user's word embedding vectors into the Recurrent Neural Network (RNN) [111]. Afterwards, the hidden vectors of RNN are averaged to obtain social media content representation for individual SES prediction.

HRNN considers the hierarchical structure of social media content following the Hierarchical Attention Network (HAN) [106]. We first likewise construct a user level social media text representation by first building representation of microblogs with word embedding and then aggregating those into a user-level representation.

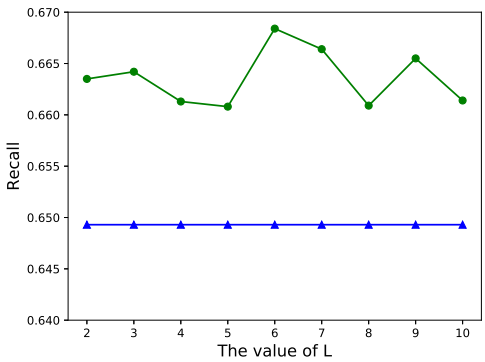
AHRNN leverages HRNN to represent the user level social media text and combines extracted platform-based user level attributes to represent the social media content for the individual SES prediction task.

To make the experimental results more convincing, we employ BiLSTM in the above baseline methods. The hyperparameters of BiLSTM in the baseline models are same as our proposed model. In the experiments, we refer to our proposed model as **CAHRNN** for

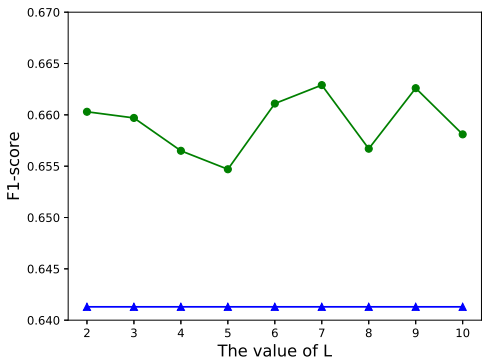


(a) Accuracy

(b) Precision



(c) Recall

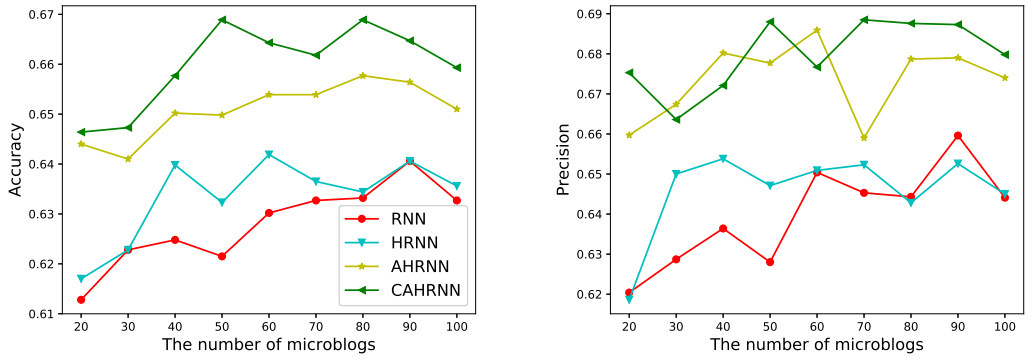


(d) F1-score

Figure 4.9: Performance over various power expansion value L.

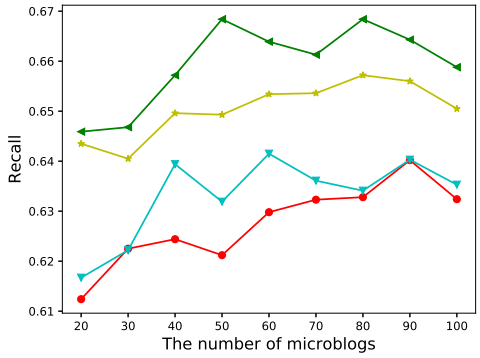
convenience. We report experimental results using all methods in terms of accuracy, precision, recall and F1-score. Particularly, accuracy is calculated as the number of correctly predicted testing samples divided by the total number of testing samples. For the 3-way classification, precision, recall and F1-score are macro-averaged, which take into account the skewed class label distributions by weighting each class uniformly.

As illustrated in Table 4.1, we can observe that the proposed model **CAHRNN** greatly outperforms the baseline models in terms of all metrics. Compared with neural network based methods, the three machine learning based methods in previous work have much lower performance, which indicates that the extracted user level features and textual features cannot represent social media content very well. This is because the traditional feature engineering methods is unable to capture some important information of social media content, i.e., the order and structure of social media text and relations among user level attributes. On the contrary, although only considering social media text representation, **RNN** significantly outperforms these machine learning based methods with about 6-13%, 7-28%, 7-23% and 6-32% higher performance score in terms of accuracy, precision, recall and F1-score respectively. This implies that **RNN** can learn text representation much better with neural networks compared with predefined textual features owing to considering the order of word sequence. Due to

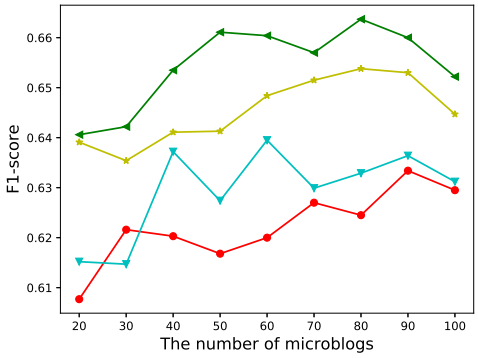


(a) Accuracy

(b) Precision



(c) Recall



(d) F1-score

Figure 4.10: Performance over various microblog number.

considering the structure of social media text, **HRNN** has higher performance than **RNN**. In addition, compared with **HRNN**, **AHRNN** enhances the prediction performance with about 1.4-3% higher performance score, which proves that appropriately fusing user level attributes can improve the ability of representing social media content. Furthermore, through considering the linear and nonlinear relationships among user level attributes, the proposed **CAHRNN** can take into account the couplings in the social media text as well as various couplings among user level attributes, which significantly improve the prediction performance compared with baseline models.

In addition, in order to compare the prediction performance of these models for each SES level, we plot Figure 4.7 for demonstrating the prediction performance for each SES level in terms of F1-score, which takes into account both precision and recall. From Figure 4.7, we can observe that the proposed **CAHRNN** has the highest F1-score on each SES level prediction task. Furthermore, with adding more information, i.e., order, structure of social media text and relations among user level attributes, the prediction performance for each SES level can be enhanced, which further validates the effect of these three key information of social media content on the individual SES prediction.

Table 4.2 shows the prediction performance of the proposed model **CAHRNN** for each SES level. We can observe that relatively more users are wrongly assigned as Level C and about half Level B users are assigned as the other SES levels. In terms of F1-score, we can observe that it is more difficult to correctly classify users from the Level B class (lowest F1 score). Figure 4.8 illustrates the confusion matrix for the SES prediction results of the proposed model. Intuitively, we can also observe that Level B users are more likely to be predicted as Level C, which may be because there exist some similar platform behaviors between some Level B user and Level C users. In the future, we will further investigate to fully understand the nature of these errors in the model.

4.5.3 Coupled Attribute Representation Analysis

To further validate the advantage of the coupled attribute representation, we compare the performance of **CAHRNN** and **AHRNN** by varying the power expansion value L from 2 to 10. In Figure 4.9, we present the performance changes of the two methods over different power expansion values in terms of accuracy, precision, recall and F1-score.

For all the evaluation metrics, the proposed **CAHRNN**, considering the coupled user level attribute representation, outperforms **AHRNN** no matter what the L value is, which validate the efficiency and robustness of the coupled attribute representation method. To be more specific, the proposed **CAHRNN** improves by 1.1%-1.9% (Accuracy), 0.2%-2.6% (Precision), 1.2%-1.9% (Recall), and 1.3%-2.2% (F1-score) for the individual SES prediction. That is to say, fusing the coupled user level attribute representation can assist in enhancing the performance of individual SES prediction.

4.5.4 Performance Comparison over Microblog Numbers

To further investigate the performance and robustness of the proposed model over social media content with various microblog numbers, we compare the performance of the proposed model and other three neural network based baseline models under different microblog number settings (i.e., maximum microblog number parameter). Figure 4.10 shows the performance of individual SES prediction generated by **RNN**, **HRNN**, **AHRNN**, and the proposed **CAHRNN** with respect to input microblog numbers in a social media content for each user.

As shown in Figure 4.10, we can observe the changing performance of four models over different microblog number in terms of accuracy, precision, recall and F1-score. Particularly, we can observe that the proposed model **CAHRNN** with considering coupled social media content representation consistently outperforms other baseline models for all input microblog numbers in terms of accuracy, recall and F1-score. For the precision metric, the proposed model mostly has better performance than other models. It indicates the robustness and flexibility of our model **CAHRNN** on dataset of different scales.

4.6 Chapter Summary

Recently, there has been a great interest in predicting individual SES from social media content, which is useful for a range of applications in enabling related organizations for economic and social policy-making. Previous related work utilize a machine learning based classifier with manually defined textual features and user level attributes from social media content for SES-related information prediction. Nevertheless, regarding the social media text in social media content, they ignore the information about the order and the hierarchical structure. For the platform-based user level attributes, the latent relationships among these attributes are omitted.

In this chapter, we propose a novel coupled social media content representation model for the individual SES prediction. On one hand, it utilizes a hierarchical recurrent neural network to incorporate the order and the hierarchical structure of social media text. On the other hand, it employs a coupled attribute representation method to take into account intra-coupled and inter-coupled interaction relationships among platform-based user level attributes. From extensive experiments on the built Sina Weibo dataset, we validate the efficiency and robustness of the proposed model by comparing with other state-of-the-art models.

Chapter 5

Conclusion

This chapter concludes the thesis by giving a summary of three specific works on the identification of online users' social status via mining user-generated data in this thesis and looking to the future work.

5.1 Summary

This thesis studies some specific issues on the identification of online users' social status via mining user-generated data. More specifically, we focus on three specific works in terms of different data sources and scenarios, which address the corresponding challenges through proposing and implementing novel effective methods respectively.

In the first work, the purpose is to identify topical opinion leaders in social community question answering sites. Most existing works either focus on using knowledge expertise to find experts for improving the quality of answers, or aim at measuring user influence to identify influential users. To identify the true topical opinion leaders, we propose a novel topical opinion leader identification framework called QALeaderRank, taking into account both the topic-sensitive influence and the topical knowledge expertise. To be more specific, on one hand, to measure the topic-sensitive influence of each user, we design a novel influence measure algorithm, which simultaneously takes into account the social network structure, the topical similarity between users and the knowledge authority. On the other hand, to infer the topic-relevant knowledge expertise of each user, we design three topic-relevant metrics, which are knowledge capacity, knowledge satisfaction and knowledge contribution. In order to evaluate the performance of the proposed QALeaderRank, extensive experiments are conducted on a set of real data that were crawled from Zhihu. The experimental results and an online user study demonstrate the efficiency of the proposed model compared with the state-of-the-art methods. Moreover, we further analyze the topic interest change behaviors of users over time and examine the predictability of user topic interest through experiments.

In the second work, we study a new problem of predicting individual socioeconomic status from mobile phone data. Most existing work on mobile phone data leverage classic supervised machine learning methods to predict regional or household SES. Compared with previous work, this work studies the SES prediction at an individual level. The new task of predicting individual SES on mobile phone data also proposes some new challenges, including sparse individual records, scarce explicit relationships and limited labeled samples. To address these issues, a semi-supervised Hypergraph-based Factor Graph Model (HyperFGM) for individual SES prediction is proposed. To handle the individual record sparsity, HyperFGM leverages customized factor functions to efficiently capture the associations between SES and individual mobile phone records. For handling the scarce explicit relationships, HyperFGM models implicit high-order relationships among users on the hypergraph structure built based on mobility pattern. In addition, HyperFGM explores the limited labeled data and unlabeled data in a semi-supervised way. Experimental results corroborate HyperFGM is efficient and greatly outperforms the state-of-the-art methods on a set of anonymized real mobile phone data.

In the third work, we study predicting the socioeconomic status of social media users based on their social media content. Previous related work leverage machine learning based classifiers with manually defined features extracted from social media content, which ignore the order and the hierarchical structure of social media text as well as the relationships among user level attributes. To this end, we propose a novel coupled social media content representation model for individual SES prediction. The proposed model utilizes a hierarchical neural network to incorporate the order and the hierarchical structure of social media text. Meanwhile, with employing a coupled attribute representation, the model can take into account intra-coupled and inter-coupled interaction relationships among platform-based user level attributes. Through extensive experiments on a set of Sina Weibo data, we validate the efficiency and robustness of the proposed model, which can achieve significant gain over other stat-of-the-art models.

5.2 Future Work

This section discusses some potential extension directions for the three specific works in the future.

For the issue of topical opinion leader identification in SCQA sites, we plan to improve the proposed model in some directions. First, regarding measuring topic-sensitive influence, besides the votes and following link structure, we will explore to incorporate the network structure based on question answering and the comments on answers. Second, due to the dynamic change of knowledge and topics in SCQA sites, in the next step, we will take into account the time factor as a weight to identify the current influence and knowledge expertise. Third, as discussed in Section 2.5.3, we plan to leverage the idea of learning to rank to improve the identification performance. Furthermore, to enhance the prediction performance of the user topic change, we plan to consider more features, such as the number of votes, the number of comments, and employ attention mechanism to select informative factors for the sequence.

For the problem of predicting individual SES based on mobile phone data, there are some potential future directions of this work. First, in order to predict finer grained SES value of each user, some other methods can be further explored and utilized such as ranking method and regression method. For example, this work could be regarded as a ranking problem. The goal of the new ranking problem is to optimally sort the users in terms of SES, which would be a more challenging and interesting problem. Next, it is interesting to study how to further explore more implicit relationships, e.g., involving mobile Internet behavior of each user. In addition, to further verify the feasibility and reusability of the proposed model, we plan to apply HyperFGM on different kinds of datasets to demonstrate the power of HyperFGM in other classification tasks.

For the issue of SES prediction of social media users, we will explore more information from social media content. First, considering the potential effect of microblog level attributes, we will explore to incorporate microblog level attributes to improve the social media content representation. Next, we will take into account coupling information between attributes and social media text to improve our model. Third, as most attributes contain categorical and numerical ones, we plan to study the embedding representation of categorical attributes and the method of capturing the couplings between categorical and numerical attributes. Finally, we plan to apply the proposed model to different datasets, such as Twitter, Quora and Zhihu, which can further verify its efficiency and robustness.

Finally, there are also some potential new issues of online users' social status identification that need to be addressed. For example, considering people almost use many applications and services every day, identifying social status based on cross-platform data sources become a promising research work. Another issue is to fuse multi-modal data, such as video, image, audio, text, etc, for identifying social status of users.

Bibliography

[1] http://uk.businessinsider.com/.

[2] http://www.woshipm.com/evaluating/2491866.html.

[3] Nancy E Adler, Thomas Boyce, Margaret A Chesney, et al. "Socioeconomic status and health: the challenge of the gradient." In: *American psychologist* 49.1 (1994), p. 15.

[4] Nikolaos Aletras and Benjamin Paul Chamberlain. "Predicting twitter user socioeconomic attributes with network and language information". In: *Proceedings of the 29th on Hypertext and Social Media*. ACM. 2018, pp. 20–24.

[5] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. "Everyone's an influencer: quantifying influence on twitter". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 65–74.

[6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model". In: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155.

[7] Peter M Blau and Otis Dudley Duncan. "The American occupational structure." In: (1967).

[8] Peter Michael Blau and Otis Dudley Duncan. "The American occupational structure." In: *American Journal of Sociology* 33.2 (1967), p. 296.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[10] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata". In: *Science* 350.6264 (2015), pp. 1073–1076.

[11] Johan Bollen, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market". In: *Journal of computational science* 2.1 (2011), pp. 1–8.

[12] Guilherme R Borges, Jussara M Almeida, Gisele L Pappa, et al. "Inferring user social class in online social networks". In: *Proceedings of the 8th Workshop on Social Network Mining and Analysis*. ACM. 2014, p. 10.

[13] Mohamed Bouguessa and Lotfi Ben Romdhane. "Identifying authorities in online communities". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 6.3 (2015), p. 30.

[14] Mohamed Bouguessa, Benoît Dumoulin, and Shengrui Wang. "Identifying authoritative actors in question-answering forums: the case of yahoo! answers". In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 866–874.

[15] John D Burger, John Henderson, George Kim, and Guido Zarrella. "Discriminating gender on Twitter". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, pp. 1301–1309.

[16] Chris Burges, Tal Shaked, Erin Renshaw, et al. "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 89–96.

[17] Ronald S Burt. “The social capital of opinion leaders”. In: *The Annals of the American Academy of Political and Social Science* 566.1 (1999), pp. 37–54.

[18] Longbing Cao, Yuming Ou, and S Yu Philip. “Coupled behavior analysis with applications”. In: *IEEE Transactions on Knowledge and Data Engineering* 24.8 (2011), pp. 1378–1392.

[19] Annika Carlsson-Kanyama and Anna-Lisa Linden. “Travel patterns and environmental effects now and in the future:: implications of differences in energy consumption among socio-economic groups”. In: *Ecological Economics* 30.3 (1999), pp. 405–417.

[20] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. “Measuring user influence in twitter: The million follower fallacy.” In: *Icwsm* 10.10-17 (2010), p. 30.

[21] Kenny K Chan and Shekhar Misra. “Characteristics of the opinion leader: A new dimension”. In: *Journal of advertising* 19.3 (1990), pp. 53–60.

[22] Wanxiang Che, Zhenghua Li, and Ting Liu. “LTP: A Chinese Language Technology Platform”. In: *Journal of Chinese Information Processing* 2.6 (2010), pp. 13–16.

[23] *Chinese Wikipedia Data Set*. `https://dumps.wikimedia.org/zhwiki/`.

[24] Aron Culotta. “Towards detecting influenza epidemics by analyzing Twitter messages”. In: *Proceedings of the first workshop on social media analytics*. acm. 2010, pp. 115–122.

[25] Ido Dagan, Lillian Lee, and Fernando Pereira. “Similarity-based methods for word sense disambiguation”. In: *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 1997, pp. 56–63.

[26] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. “Rank aggregation methods for the web”. In: *Proceedings of the 10th international conference on World Wide Web*. ACM. 2001, pp. 613–622.

[27] Peter Elias and Margaret Birch. “SOC2010: revision of the Standard Occupational Classification”. In: *Economic & Labour Market Review* 4.7 (2010), pp. 48–55.

[28] Dominik Maria Endres and Johannes E Schindelin. “A new metric for probability distributions”. In: *IEEE Transactions on Information theory* 49.7 (2003), pp. 1858–1860.

[29] Golnoosh Farnadi, Jie Tang, Martine De Cock, and Marie-Francine Moens. “User profiling through deep multimodal fusion”. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 171–179.

[30] Gottlob Frege. “On sense and reference”. In: *oversatt av Max Black, i J. Guitérrez-Rexach (red.): Semantics: Crictical concepts in linguistics* 1 (2003), pp. 7–25.

[31] Vanessa Frias-Martinez and Jesus Virseda. “On the relationship between socio-economic factors and cell phone usage”. In: *Proceedings of the fifth international conference on information and communication technologies and development*. ACM. 2012, pp. 76–84.

[32] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. “Beyond the stars: improving rating predictions using review text content.” In: *WebDB*. Vol. 9. Citeseer. 2009, pp. 1–6.

[33] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, and Qionghai Dai. “3-D object retrieval and recognition with hypergraph analysis”. In: *IEEE Transactions on Image Processing* 21.9 (2012), pp. 4290–4303.

[34] Darren George. *SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e*. Pearson Education India, 2011.

[35] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. “Cognos: crowdsourcing search for topic experts in microblogs”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 575–590.

[36] Jennifer Golbeck, Cristina Robles, and Karen Turner. “Predicting personality with social media”. In: *CHI’11 extended abstracts on human factors in computing systems*. ACM. 2011, pp. 253–262.

[37] Delaney Granizo-Mackenzie and Jason H Moore. “Multiple Threshold Spatially Uniform ReliefF for the Genetic Analysis of Complex Human Diseases.” In: *EvoBIO*. Springer. 2013, pp. 1–10.

[38] S Grin and Lawrence Page. “The anatomy of a large-scale hypertextual Web search engine”. In: *Computer networks and ISDN systems* 30.1-7 (1998), pp. 107–117.

[39] Mohammed Hasanuzzaman, Sabyasachi Kamila, Mandeep Kaur, Sriparna Saha, and Asif Ekbal. “Temporal orientation of tweets for predicting income of users”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017, pp. 659–665.

[40] Robert M Hauser and John Robert Warren. “Socioeconomic indexes for occupations: A review, update, and critique”. In: *Sociological methodology* 27.1 (1997), pp. 177–298.

[41] Robert M. Hauser and John Robert Warren. “Socioeconomic Indexes for Occupations: A Review, Update, and Critique”. In: *Sociological Methodology* 27.1 (2010), pp. 177–298.

[42] Taher H Haveliwala. “Topic-sensitive pagerank”. In: *Proceedings of the 11th international conference on World Wide Web*. ACM. 2002, pp. 517–526.

[43] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[44] Lingzi Hong, Enrique Frias-Martinez, and Vanessa Frias-Martinez. “Topic Models to Infer Socio-Economic Maps.” In: *AAAI*. 2016, pp. 3835–3841.

[45] Qunying Huang and David WS Wong. “Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?” In: *International Journal of Geographical Information Science* 30.9 (2016), pp. 1873–1898.

[46] Yuchi Huang, Qingshan Liu, Shaoting Zhang, and Dimitris N Metaxas. “Image retrieval via probabilistic hypergraph ranking”. In: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE. 2010, pp. 3376–3383.

[47] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. “Multimodal fusion with recurrent neural networks for rumor detection on microblogs”. In: *Proceedings of the 25th ACM international conference on Multimedia*. ACM. 2017, pp. 795–816.

[48] Simon Kemp. *DIGITAL 2019: GLOBAL INTERNET USE ACCELERATES*. `https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates`. 2019.

[49] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[50] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. “Sentiment analysis of short informal texts”. In: *Journal of Artificial Intelligence Research* 50 (2014), pp. 723–762.

[51] Jon M Kleinberg. “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[53] Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. “A user-centric model of voting intention from Social Media”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 993–1003.

[54] Vasileios Lampos, Andrew C Miller, Steve Crossan, and Christian Stefansen. “Advances in nowcasting influenza-like illness rates using search query logs”. In: *Scientific reports* 5 (2015), p. 12760.

[55] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. “Inferring the socioeconomic status of social media users based on behaviour and language”. In: *European Conference on Information Retrieval*. Springer. 2016, pp. 689–695.

[56] Paul Felix Lazarsfeld, Bernard Berelson, and Hazel Gaudet. “The peoples choice: how the voter makes up his mind in a presidential campaign.” In: (1968).

[57] Changhyun Lee, Haewoon Kwak, Hosung Park, and Sue Moon. “Finding influentials based on the temporal order of information adoption in twitter”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 1137–1138.

[58] Feng Li and Timon C Du. “Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs”. In: *Decision Support Systems* 51.1 (2011), pp. 190–197.

[59] Laura Lotero, Rafael G Hurtado, Luis Mario Floría, and Jesús Gómez-Gardeñes. “Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes”. In: *Royal Society open science* 3.10 (2016), p. 150654.

[60] Huina Mao, Xin Shuai, Yong-Yeol Ahn, and Johan Bollen. “Quantifying socio-economic indicators in developing countries from mobile phone communication data: applications to Côte d’Ivoire”. In: *EPJ Data Science* 4.1 (2015), p. 15.

[61] Bernard Marr. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*. `https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/`. Forbes, 2018.

[62] Miller McPherson, Lynn Smith-Lovin, and James M Cook. “Birds of a feather: Homophily in social networks”. In: *Annual review of sociology* 27.1 (2001), pp. 415–444.

[63] Qingliang Miao, Shu Zhang, Yao Meng, and Hao Yu. “Domain-sensitive opinion leader mining from online review communities”. In: *Proceedings of the 22nd International Conference on World Wide Web*. ACM. 2013, pp. 187–188.

[64] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[65] Kevin P Murphy, Yair Weiss, and Michael I Jordan. “Loopy belief propagation for approximate inference: An empirical study”. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. 1999, pp. 467–475.

[66] James H Myers and Thomas S Robertson. “Dimensions of opinion leadership”. In: *Journal of marketing research* 9.1 (1972), pp. 41–46.

[67] Adi Omari, David Carmel, Oleg Rokhlenko, and Idan Szpektor. “Novelty based ranking of human answers for community questions”. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM. 2016, pp. 215–224.

[68] Aditya Pal and Joseph A Konstan. “Expert identification in community question answering: exploring question selection bias”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM. 2010, pp. 1505–1508.

[69] Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. “An analysis of the user occupational class through Twitter content”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1. 2015, pp. 1754–1764.

[70] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. “Studying user income through language, behaviour and affect in social media”. In: *PloS one* 10.9 (2015), e0138717.

[71] Carol Propper, Michael Damiani, George Leckie, and Jennifer Dixon. “Impact of patients’ socioeconomic status on the distance travelled for hospital admission in the English National Health Service”. In: *Journal of Health Services Research & Policy* 12.3 (2007), pp. 153–159.

[72] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. “The bag-of-opinions method for review rating prediction from sparse text patterns”. In: *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics. 2010, pp. 913–921.

[73] *Quora*. https://www.quora.com/.

[74] Michael O Rabin and Dana Scott. “Finite automata and their decision problems”. In: *IBM journal of research and development* 3.2 (1959), pp. 114–125.

[75] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. “Classifying latent user attributes in twitter”. In: *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM. 2010, pp. 37–44.

[76] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. “Finding expert users in community question answering”. In: *Proceedings of the 21st International Conference on World Wide Web*. ACM. 2012, pp. 791–798.

[77] Everett M Rogers and David G Cartano. “Methods of measuring opinion leadership”. In: *Public Opinion Quarterly* (1962), pp. 435–441.

[78] David Rose and David Pevalin. “Re-basing the NS-SEC on SOC2010”. In: (2010).

[79] David Rose and David Pevalin. “Re-basing the NS-SEC on SOC2010: a report to ONS”. In: *Techincal report, University of Essex* (2010).

[80] Sai Nageswar Satchidanand, Harini Ananthapadmanaban, and Balaraman Ravindran. “Extended Discriminative Random Walk: A Hypergraph Approach to Multi-View Multi-Relational Transductive Learning.” In: *IJCAI*. 2015, pp. 3791–3797.

[81] *Sina Weibo*. https://www.weibo.com/.

[82] Selcuk R Sirin. “Socioeconomic status and academic achievement: A meta-analytic review of research”. In: *Review of educational research* 75.3 (2005), pp. 417–453.

[83] Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. “Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 511–520.

[84] *Socioeconomic Status*. https://en.wikipedia.org/wiki/Socioeconomic_status.

[85] Siqi Song, Ye Tian, Wenwen Han, Xirong Que, and Wendong Wang. “Leading users detecting model in professional community question answering services”. In: *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE. 2013, pp. 1302–1307.

[86] Xiaodan Song, Yun Chi, Koji Hino, and Belle Tseng. “Identifying opinion leaders in the blogosphere”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 971–974.

[87] Victor Soto, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. “Prediction of socioeconomic levels using cell phone records”. In: *User modeling, adaption and personalization* (2011), pp. 377–388.

[88] *Stack Overflow*. https://stackoverflow.com/.

[89] Lifan Su, Yue Gao, Xibin Zhao, et al. “Vertex-Weighted Hypergraph Learning for Multi-View Object Classification.” In: *IJCAI*. 2017, pp. 2779–2785.

[90] Duyu Tang, Bing Qin, and Ting Liu. “Document modeling with gated recurrent neural network for sentiment classification”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 1422–1432.

[91] Wenbin Tang, Honglei Zhuang, and Jie Tang. “Learning to infer social ties in large networks”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2011, pp. 381–397.

[92] Simo Editha Tchokni, Diarmuid O Séaghdha, and Daniele Quercia. “Emoticons and phrases: Status symbols in social media”. In: *Eighth International AAAI Conference on Weblogs and Social Media*. 2014.

[93] Dj Treiman. “Index - Occupational Prestige in Comparative Perspective”. In: *American Journal of Sociology* 85.3 (1977), 511–514.

[94] Ryan J. Urbanowicz, Randal S. Olson, Peter Schmitt, Melissa Meeker, and Jason H. Moore. *Benchmarking Relief-Based Feature Selection Methods*. arXiv e-print. https://arxiv.org/abs/1711.08477. 2017.

[95] Yogatheesan Varatharajah, Min Jin Chong, Krishnakant Saboo, et al. “EEG-GRAPH: A Factor-Graph-Based Model for Capturing Spatial, Temporal, and Observational Relationships in Electroencephalograms”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 5377–5386.

[96] Can Wang, Zhong She, and Longbing Cao. “Coupled attribute analysis on numerical data”. In: *Twenty-third international joint conference on artificial intelligence*. 2013.

[97] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y Zhao. “Wisdom in the social crowd: an analysis of quora”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 1341–1352.

[98] Sida Wang and Christopher D Manning. “Baselines and bigrams: Simple, good sentiment and topic classification”. In: *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*. Association for Computational Linguistics. 2012, pp. 90–94.

[99] *WeChat*. `https://www.wechat.com/`.

[100] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. “Twitterrank: finding topic-sensitive influential twitterers”. In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM. 2010, pp. 261–270.

[101] *WhatsApp*. `https://www.whatsapp.com/`.

[102] Anne Wilcock, Maria Pun, Joseph Khanona, and May Aung. “Consumer attitudes, knowledge and behaviour: a review of food safety issues”. In: *Trends in Food Science & Technology* 15.2 (2004), pp. 56–66.

[103] Marilyn A Winkleby, Darius E Jatulis, Erica Frank, and Stephen P Fortmann. “Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease.” In: *American journal of public health* 82.6 (1992), pp. 816–820.

[104] *Yahoo!Answer*. `https://answers.yahoo.com/`.

[105] Yang Yang, Walter Luyten, Lu Liu, et al. “Forecasting Potential Diabetes Complications.” In: *AAAI*. 2014, pp. 313–319.

[106] Zichao Yang, Diyi Yang, Chris Dyer, et al. “Hierarchical attention networks for document classification”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 1480–1489.

[107] Yang Ye, Yu Zheng, Yukun Chen, Jianhua Feng, and Xing Xie. “Mining individual life pattern based on location history”. In: *Mobile Data Management: Systems, Services and Middleware, 2009. MDM’09. Tenth International Conference on*. IEEE. 2009, pp. 1–10.

[108] Jun Yu, Dacheng Tao, and Meng Wang. “Adaptive hypergraph learning and its application in image classification”. In: *IEEE Transactions on Image Processing* 21.7 (2012), pp. 3262–3272.

[109] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. “Tensor fusion network for multimodal sentiment analysis”. In: *arXiv preprint arXiv:1707.07250* (2017).

[110] Zhongwu Zhai, Hua Xu, and Peifa Jia. “Identifying opinion leaders in BBS”. In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*. IEEE Computer Society. 2008, pp. 398–401.

[111] Dong Zhang, Shoushan Li, Hongling Wang, and Guodong Zhou. “User classification with multiple textual perspectives”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 2112–2121.

[112] Jun Zhang, Mark S Ackerman, and Lada Adamic. “Expertise networks in online communities: structure and algorithms”. In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, pp. 221–230.

[113] Zhou Zhao, Lijun Zhang, Xiaofei He, and Wilfred Ng. “Expert finding for question answering via graph regularized matrix completion”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.4 (2015), pp. 993–1004.

[114] *Zhihu*. `https://www.zhihu.com/`.

[115] Denny Zhou, Jiayuan Huang, and Bernhard Schölkopf. “Learning with hypergraphs: Clustering, classification, and embedding”. In: *Advances in neural information processing systems*. 2007, pp. 1601–1608.

List of Acronyms

List of Figures

List of Tables