

Local Spatial Obesity Analysis and Estimation Using Online Social Network Sensors

Qindong Sun^{1*}, Nan Wang¹, Shancang Li², Hongyi Zhou¹

1. Shaanxi Key Laboratory of Network Computing and Security, Xi'an University of Technology, 710048, China

2. Department of Computer Sciences and Creative Technologies, University of the West of England, Bristol BS16 1QY, UK

Abstract: Recently, the online social networks (OSNs) have received considerable attentions as a revolutionary platform to offer users massive social interaction among users that enables users to be more involved in their own healthcare. The OSNs have also promoted increasing interests in the generation of analytical, data models in health informatics. This paper aims at developing an obesity identification, analysis, and estimation model, in which each individual user is regarded as an online social network ‘sensor’ that can provide valuable health information. The OSN-based obesity analytic model requires each sensor node in an OSN to provide associated features, including dietary habit, physical activity, integral/incidental emotions, and self-consciousness. Based on the detailed measurements on the correlation of obesity and proposed features, the OSN obesity analytic model is able to estimate the obesity rate in certain urban areas and the experimental results demonstrate a high success estimation rate. The measurements and estimation experimental findings created by the proposed obesity analytic model show that the online social networks could be used in analyzing the local spatial obesity problems effectively.

Keywords: Online Social Networks, Internet of Things (IoT), Obesity, Health Informatics, Bioinformatics, Public health

1. Introduction

Advancements in social network, social media, and mobile technologies have significantly attracted tremendous attention from researchers and all kinds of users. In view of the features such as free exchange of information, users can share their interests, experiences and many other kinds of information through online social networks.

* Corresponding author E-mail: sqd@xaut.edu.cn

In healthcare applications, the online social network could be a very effective tool, this is because that online social networks have been the most popular web service and attracted lots of users to exchange information. According to the information, we could learn some characteristics about each user such as their hobbies, interests and their health condition. Therefore, online social networks user can be regarded as separate node releasing and receiving information like the node in Internet of things (IoT). The users in IoT are able to handle various real-world information spread in the network [1, 2]. Online social networks can be understood as the online sensor networks and each user, which can be regarded as a social sensor, has a unique identification, and can post information or spread information from others. The schematic diagram is outlined in Figure 1. As shown in the figure, the continuous information is poured into online social networks by users using various social networking sites. Based on data acquisition techniques and big data analytics, the information closely related to people's living habits and health condition can be gathered from the internet. The medical institutions and Centers for Disease Control and prevention can receive this kind of information in time which can aid in disease control and prevention.

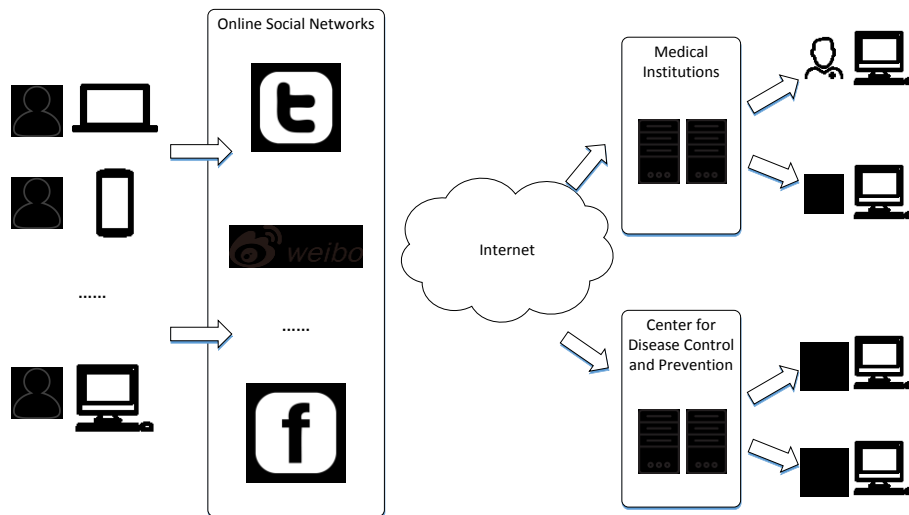


Figure 1. Example of Healthcare IoT-based online social networks

Users access the internet with various terminal devices and post or forward information in each online social network. User information can be gathered by area and provided to the medical personnel and disease control and prevention administrators after detailed and reasonable analysis and process. Compared with the traditional IoT solutions, we could carry out healthcare research

in certain area much easier with online social networks as their large information and user scales. For example, the data of Twitter users has been used to supervise the flu trend. In the surveillance process each user is regarded as a social sensor and the flu trend can be supervised by using temporal topic models weekly^[3]. This will also reduce the cost as we do not need a large number of physical sensor devices. Furthermore, we could achieve healthcare information from user's perspective of subjective.

Obesity is a key research issue in healthcare research field which can cause great negative impact on people's health and have been associated with many diseases such as diabetes, heart disease and high blood pressure^[5]. At present, obesity has become a major public health issue in the worldwide especially in developed countries. Directly report and sampling survey are the two main existing ways to obtain obesity status of population in certain location such as the data collected by centers for disease control and prevention (CDC). Obesity and overweight problems are closely related to people's dietary habit, exercise habit and other features of life-style which are hard to collect. By utilizing the user data in online social network platforms, researchers can extract obesity-related information based on user's posts, user profiles and their online behaviors. These kinds of data could enrich the data sources used in the existing studies and is help for the further research like obesity estimation, which can provide effective supplement for obesity control and prevention.

However, obesity study with online social networks is a challenging task due to several reasons. First, unlike the influenza and the other diseases, obesity has no periodicity and can hardly be characterized by abruptness therefore, it is difficult to accurately evaluate and predict the trend of obesity. In addition, obesity is related to many factors which should be carefully analyzed and selected according to the online social network data. Finally, the data scale in online social network is very huge which can make extra difficulties for obesity study.

By means of information acquisition technologies such as web crawler, we can collect and utilize the original and forwarded information posted by users. Based on these "social sensor", the online social network users the life-style and behavioral habit of people in certain area can be explore and monitored which gives us an opportunity to study the obesity problem with the online data.

In order to study the obesity problem with the online social network data in a more effectively way, we have analyzed the correlation between obesity and some features extracted from Twitter and

tried to predicate the trend in different location. The primary aims of this paper are finding out the features which are quantifiable and exactly related to obesity according to the Twitter data, and then estimating the obesity rate with these proposed features in certain areas.

The main contributions of this work are as follows:

(1) Extending the range of online social networks application. There are already many studies about online social networks application such as using online social network data for public opinion monitoring, business marketing and infectious disease predication. Our study shows that the online social network data can be used to study the problem of chronic disease and potentially can be used many other research fields.

(2) An obesity estimation method based on online social network data and authoritative data. The experimental results also show that our method can be effectively to estimate the obesity rate in a certain area.

The remainder of the paper is structured as follows: Section 2 introduces the related works. Section 3 describes the obesity detection framework. Data processing is presented in Section 4. We show the experimental results in Section 5. In the end the discussion and conclusion are provided in Section 6.

2. Related works

Recently, with the online social networks grow in popularity, the public has uploaded tremendous data to online social networks, where the users can be regarded as social sensors. The behavior and information of online social network users can reflect conditions of the physical world ^[5]. Especially, the huge amount of online social network user data is very valuable for healthcare field ^[6].

There are already some diseases studies based on online social networks which can offer valuable experience for our research. Influenza trend is one of most-studied disease problems based on online social networks. Google flu trends (GFT) is a widely used research tool that uses the search engine data to predict the flu trends of the real world ^[7]. In many existing methods, Twitter or Facebook data are used to detect the outbreaks of Influenza and predict the trend. For instance, Aramaki et al. have proposed an influenza detecting method with positive and negative tweet information about influenza which outperforms the Google method ^[8]. Brennan et al. used the

geo-tagged tweets to infer the flow of individuals between cities and on this basis, they can predict the prevalence of flu or other similar disease ^[9]. Paul et al. studied the correlation between tweets content and surveillance result in many countries they found that Twitter data can be used to effectively supervise flu trend ^[10]. These studies have used online data to analyze and predict the flu trends and achieved good results, which gives good reference value for us to study the obesity problem. But unlike the infectious disease, obesity is a chronic disease, these existing flu-related research cannot be directly applied to the analysis or predication on chronic disease.

A few researches use online social networks to deal with other diseases. For instance, Eichstaedt et al. predicted the atherosclerotic heart disease by using the factors of social relationships, disengagement and emotion extracted from Twitter ^[11]. Larsen et al. proposed a “We Feel” system with analyzing the tweet information in Twitter which can reflect the emotion variation of people in certain area ^[12]. Yang et al. proposed an automatic method to detect depressed users based on their tweet information and GIS technology which can improve diagnosis techniques for depression ^[13].

For the obesity study, there are also some studies with the online social network data. For instance, Gore et al. analyzed the tweet content and measure the obesity rate in urban areas within the United States ^[14]. Chou et al. have analyzed the topic and conversations about obesity in online social networks but they did not carry out further research such as predication ^[15]. These studies show that online social networks can be used to explore the progression of chronic diseases such as obesity and provide some ways of problem solving such as area division method of online users. But there still exist some disadvantages in the existing studies, for example, the main works of some studies are only basic measurements and some studies have carried out obesity estimation or predication only based on few features and simple predicting method. These methods often ignore the statistical caliber, weighting analysis in effect factors and other important characteristics. The IoT for healthcare study gives us useful thought to the obesity study based on social sensors ^[16]. In IoT-enabled healthcare system, a lifelogging data validation model is proposed for eliminating irregular uncertainties and estimating data ^[17]. The large number of resource-constrained sensors in healthcare may create tremendous amount of data, which could be a challenge ^[18]. In recent, a number of IoT-enabled healthcare systems are proposed to make efficient diagnosis, and monitoring of health in these environments ^[19].

3. Framework of obesity detection

The obesity detection process involves two stages: data collection and obesity rate estimation process. Figure 2 shows the schema of the detection process. In the data collection process, the information that we need for obesity estimation can be extracted from each user in online social networks and from medical institutes or other authoritative sources. Then for the computational process we can find out the features closely related to obesity according to the acquired online social network data and authoritative medical data. The obesity status estimation can be handled with these features by proper estimation algorithms.

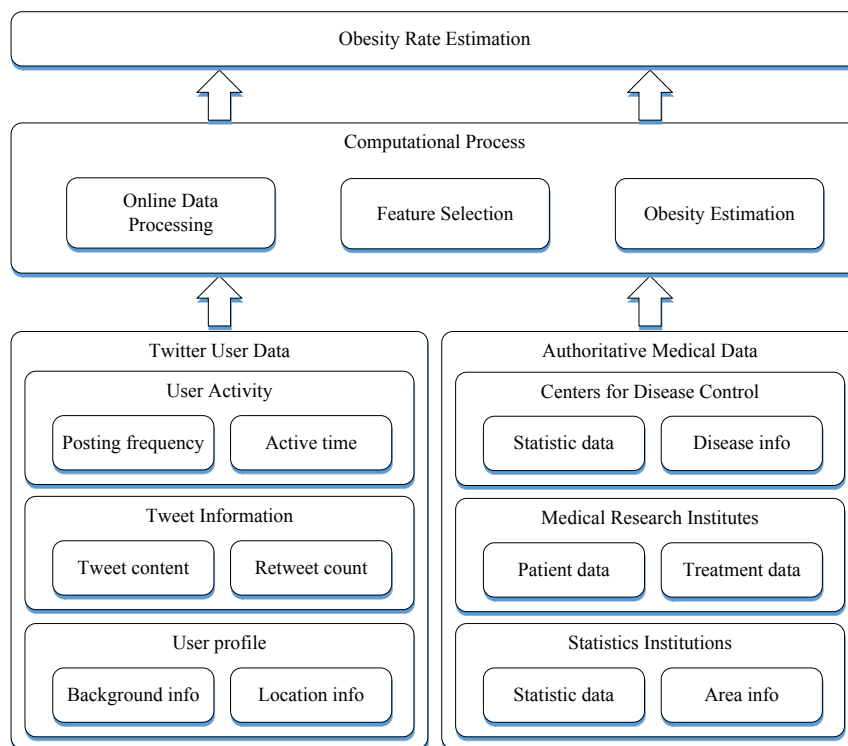


Figure 2. Framework of obesity detection

As shown in Figure 2, the data used in our paper include two categories: online social network data and authoritative medical data about obesity. The data of online social networks are crawled from Twitter which is one of the most popular online social network sites. we have crawled the user and tweet information with the geo-tagging information. In the collected dataset, there are 41 million tweets from 110 major cities which are posted from January 1, 2012 to December 31, 2013. The acquired Twitter data includes tweet ID, post time, tweet content, poster ID, and poster full name.

The obesity data used in this paper are collected from the Gallup Healthways Wellbeing Survey

and the Data, Maps and Trends of Overweight & Obesity (CDC, centers for disease control and prevention, USA) ^[20, 21]. These data resources provide geographic obesity rates with the provincial or city level during 2012~2013.

4. Data processing

In online social networks, each node (online social network user) may post lots of information, thus we need filter the useless information to select proper features to estimate the obesity rate in certain areas. The criterions for the feature selection include the following aspects:

- (a) Features must be associated with obesity,
- (b) Features must be measurable.

Based on our analysis on online social network data and the medical data, we will explore various features and demonstrate their effectiveness on obesity study. The proposed features can be categorized as follows:

- **Dietary habit.** The dietary habit has a close correlation with the changes in body weight. The bad eating habit such as eating large amount of high calorie foods and fast foods can present huge problems to weight management and is contributing to the current obesity problem in the world wide.
- **Physical Activity.** Physical activity is important factor closely related obesity. People who have little physical activity or simply not, often suffered from the problem of obesity and the diseases caused by obesity.
- **Emotion features.** Negative emotions like anxiety, pressure and depression are also key factors of the obesity problem. The negative emotion can cause listlessness and mood swings which may affect hormone balance and dietary, therefore leading to obesity.
- **Self-consciousness.** People who care about the body shape or the weight problem may pay more attention to their eating habit and living style. Therefore, this feature could be also related to obesity.

We collected the above four kinds of features according to the information posted by each user in online social networks. These features give us an opportunity to study obesity problem based on the online social networks. Then we will show a more detailed presentation about each feature and explain how these features are selected.

4.1. Dietary habit

Food is always one of the most popular topics in online social networks. Users often post content about what they are eating or talk about their favorite foods from which we can infer the dietary habits of different users. From the point of food energy and food types we propose three features to characterize the dietary habit.

4.1.1 High Calorie Foods

To quantify the dietary content of this feature we identify the users related to the high calorie foods which are selected according to the USDA National Nutrient Database. Instead of using only the traditional keyword-based method, the sentiment orientation is also involved during the measurement process. For example, ‘like’ is one of these words which can indicate user’s attitude. So we can find out more accurate search results about these users who are interested in the high calorie foods. Furthermore, due to the differences on urban scale, the number of Twitter user could be different for each area. Hence, the absolute numbers of related tweet and users scale will be higher in these locations with large scale, and this problem can cause inaccurate analysis results. As there is no existing statistics for the user number of each urban area we have proposed a method to estimate the user number at a certain area. The process is to calculate how many users contained in all the tweets with a specific geo-tag collected over a time period. The result will approximate the number of users in the geo-tag area.

For a certain urban area, the high calorie food feature can be calculated as follows:

$$f_{hcf}^l = \frac{\sum_{i=1}^n D_h(i)}{P(l)} \quad (1)$$

where f_{hcf}^l means the high calorie food density in the certain area l . As different urban area has different online social network users, more users may cause more tweets. So in order to get rid of the problem caused by the quantitative differences of online social network users, there should be a normalization process. $P(l)$ is the normalized user number of l , i denotes different kinds of high calorie food, $D_h(i)$ is the number of users who like the i th high calorie food.

4.1.2 Fast Foods

Fast foods are always high in calories and have a large number of customers. Hence it will be measured as part of the dietary habit in our experiments. In order to quantify this feature, we

choose the top 10 fast food chains in America as the analysis objects, for example McDonald's and Burger King. For each urban area, we intend to find out users who like or eat fast food frequently. To do this the tweet selection process is similar with the high calorie foods, the tweet contains the fast food name and the co-occurrence word such as 'like'. According to the above analysis the fast food eating habit in a certain area can be calculated as follows:

$$f_{fd}^l = \frac{\sum_{i=1}^n D_f(i)}{P(l)} \quad (2)$$

where f_{fd}^l is the fast food density in the certain area l , $P(l)$ is the normalized user number of l , i denotes the types of fast food restaurants, $D_f(i)$ denotes the number of users which are considered as loving and often eating fast food.

4.2. Physical activity

There are lots of tweets related to physical activity and sports on this basis, we analyze the relationship between obesity and physical activity from two aspects: (a) popular professional sports, (b) fitness exercises.

4.2.1 Popular professional sports

Playing basketball and football can have contribution to keep fit and obesity control. Professional sports attract lots of attention from various people and there are so many users talking about these sports. We intend to investigate that compared with the fitness sports how is relationship between these sports and obesity based on the Twitter data. The selected professional sports include basketball, football, baseball, tennis, etc. With the analysis on the users who like and usually take part in these sports according to the tweet content about these sports, we can get the results about the relationship between popular sports and obesity. The calculation process is as follows:

$$f_{ps}^l = \frac{\sum_{i=1}^n (S_p(i) + S'_p(i))}{P(l)} \quad (3)$$

where f_{ps}^l is the popular professional sports density in the certain area l , i denotes the number of popular sports which are already mentioned, $S_p(i)$ denotes the number of users who have expressed their favorite feelings on these sports. As there are too many tweets related with the selected sports we have analyzed the users, which have a positive attitude to these sports such as

“Time for basketball now”. Hence, $S_p'(i)$ denotes the number of this kind of user. $P(l)$ is the normalized user number of l .

4.2.2 Fitness sports

For the fitness sports, the related user selection is similar with the previous features measurement. If a tweet from a user contains the specific sports and with positive sentiment tendency, it will be marked and put in the targeted collection. All these fitness sports are selected based on the obesity presentation guideline in CDC such as jogging, yoga, aerobics, etc. The fitness sports density in a certain area can be calculated as follows:

$$f_{fs}^l = \frac{\sum_{i=1}^n S_{fit}(i)}{P(l)} \quad (4)$$

where f_{fs}^l is the fitness sports density in area l , $P(l)$ is the normalized user number of l , $S_{fit}(i)$ denotes the number of users who like and often do fitness sports.

4.3. Emotion features

Stress is another important factor which can result in obesity. Online social networks are important platforms for users to express their emotions. Therefore, along with the words related to the negative emotions such as depression and stress expressed in Twitter we have analyzed the overall condition of emotion in a certain place. Then the relationship between obesity and the emotion condition expressed from online social networks can be analyzed. The users with the negative emotion problem are measured by their tweets: If the tweets contain the complaint or statement with the depression, stress and other fifteen negative emotion words, the user will be labeled. The calculation is as follows:

$$f_{ef}^l = \frac{\sum_{i=1}^n E(i)}{P(l)} \quad (5)$$

where f_{ef}^l is the stress and depress emotion density in area l , $P(l)$ is the normalized user number of l , $E(i)$ denotes the number of users who has shown their negative emotion such as stress and depression in Twitter.

4.4. Self-consciousness

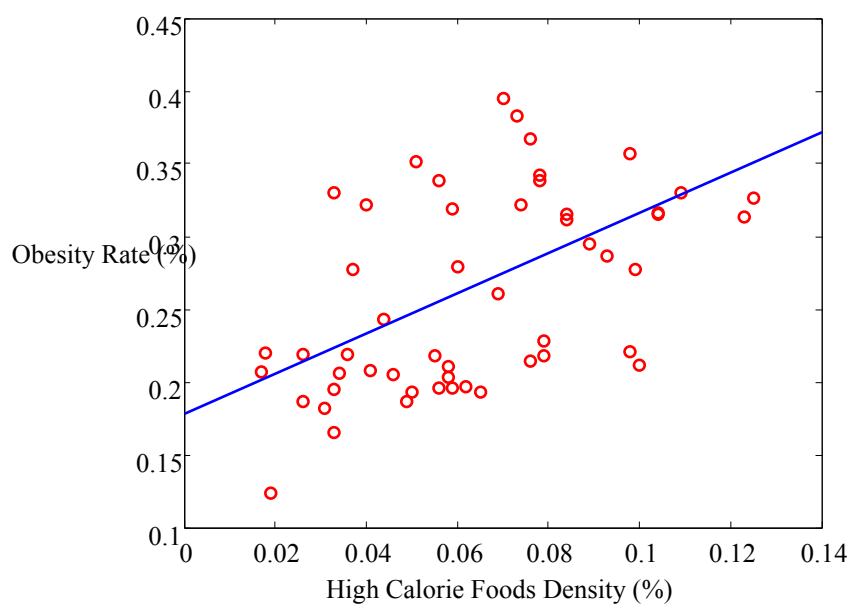
With the growing concerning on obesity, users have been paying more attention to their weight

and figure. Users who are not satisfied with their body may post the tweets with the content as complaining gaining weight in the Twitter. We call this feature as self-consciousness and intend to analyze whether this kind of information have correlation with obesity. To measure this feature, the users which have posted tweets about expressing the willing of lose weight will be labeled. The calculation can be as follows:

$$f_{sc}^l = \frac{\sum_{i=1}^n C(i)}{P(l)} \quad (6)$$

where f_{sc}^l is the self-consciousness density in area l , $P(l)$ is the normalized user number of l , $C(i)$ denotes the number of users who have post tweet about losing weight or keeping fit.

5. Experimental results



5.1. Feature measurement results

Figure 3. Correlation of high calorie foods and obesity rate

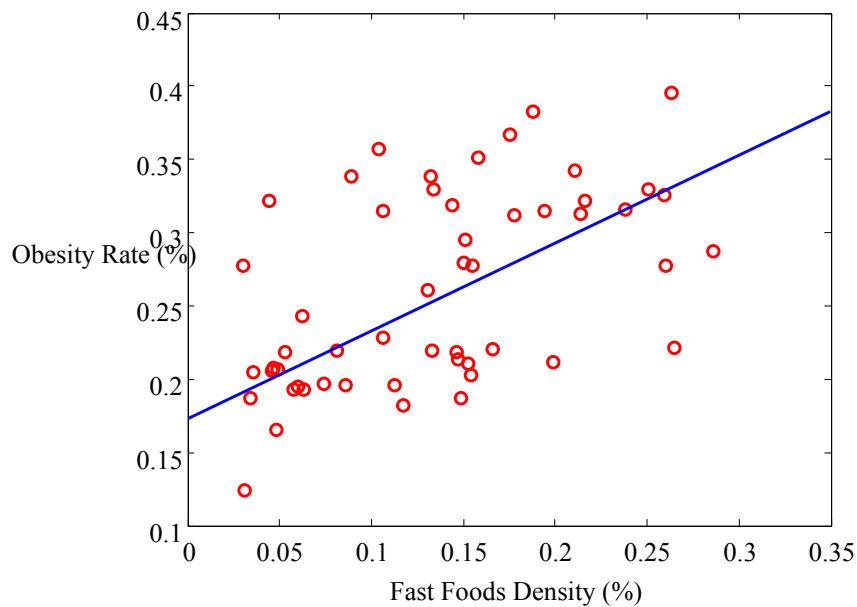
The purpose of the measurement experiments is to attempt to find out the proper relationship between the proposed features and obesity rate. According to the dataset, we have analyzed the correlation of obesity rate and each feature in different urban areas.

(1) Dietary habit and obesity

Figure 3 shows the correlation of high calorie foods and obesity rate in different urban areas. As the population sizes in some areas are too different we have removed these places with small population and Twitter data. In addition, for the cities with close geographic position, we only keep the area with the largest population.

It is obvious that obesity rate has positive correlation with high calorie food density. For example, the city Boulder, CO has the lowest obesity rate and the lowest high calorie food density as well. On the contrary, Mobile, AL is the city with high obesity rate and high rate of high calorie food density.

Fast foods are one of the causes of the obesity hence usually eating fast food can make the obesity



rate into a high level. Then in order to further investigate the relationship between dietary habit and obesity, the correlation of fast food density and obesity rate is also being measured. The results are shown in figure 4. According to the results it is clear that the high obesity rate areas are more likely to have high fast foods density.

Figure 4. Correlation of fast foods and obesity rate

From the above measurement results, the dietary habit extracted from Twitter can be regarded as correlation with obesity rate. The measurement results show proper correlation of dietary habit and obesity which is to say that although our target user selection process abandon lots of users but the results are well enough to support the analysis experiment.

(2) Physical activity and obesity

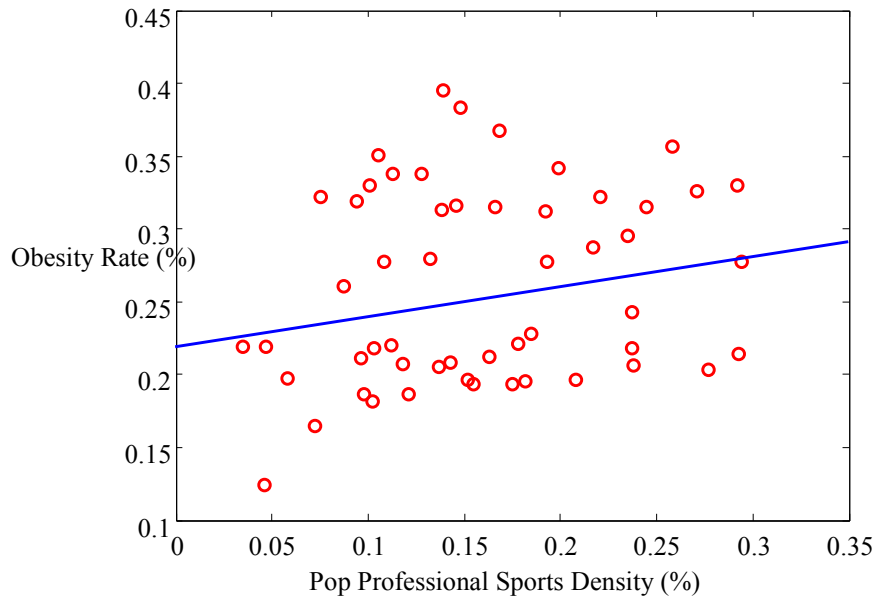
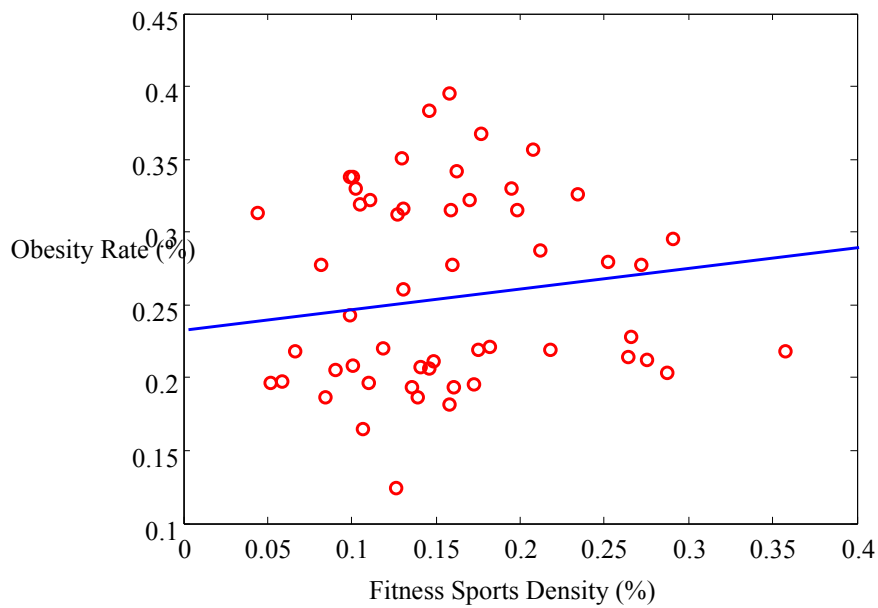


Figure 5. Correlation of popular professional sports and obesity rate

Figure 5 shows the correlation of the popular professional sports and obesity rate. The results are not as significant as the features of dietary habit. And then we have carried out the measurement about the correlation of fitness sports and obesity rate as shown in figure 6. The distribution is similar with the professional sports or even worse. These results are unexpected as sports and physical activity are always one of the hot topics in online social networks. The previous research



on the physical activity is different from ours.

Figure 6. Correlation of fitness sports and obesity rate

As shown in section 4, for the measurement on each proposed feature, we consider not just the numbers extracted from the data and meanwhile the user scale in the certain area are also considered. Furthermore, the user selection course is not just based on using physical activity key words directly. Although our measurement method could have some shortage but after multiple experiments the results are valid.

The causes of the results can be that: Sports especially the professional sports like football and basketball have attracted many attentions and people all like to watch or discussion issue related sports no matter them really do exercise or not. Therefore, in Twitter the corresponding distribution regularity of sports related tweets will be not obvious. The insufficient data may another reason for the results as the obesity data in different areas are list annually. If we can get more detailed authoritative data, the results will be more accuracy.

(3) Emotion and obesity

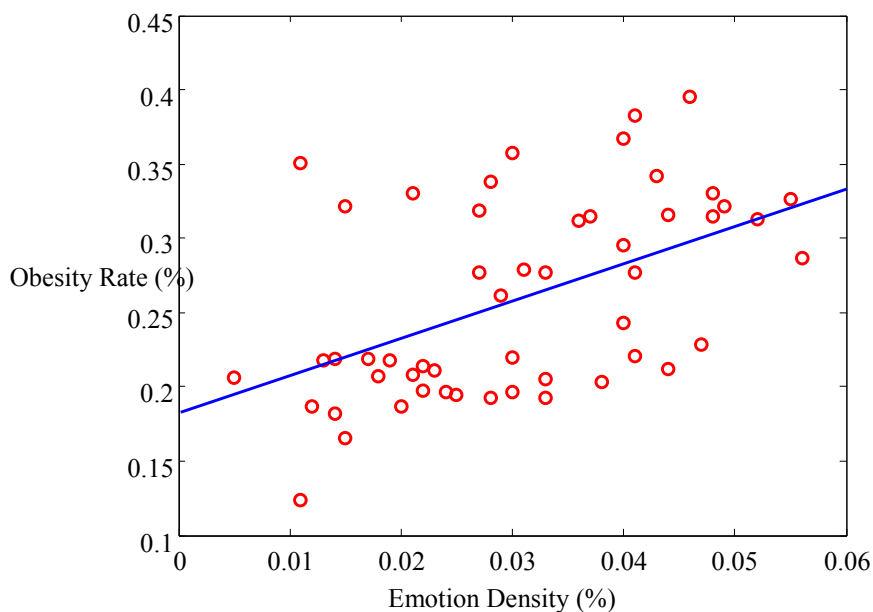


Figure 7. Correlation of emotion and obesity rate

Figure 7 shows that there is a statistically relationship between emotion and obesity rate. As we expect the emotion such as stress and depression are higher in high obesity areas. Nowadays, the development of social economy is so fast which speeds up people's living rhythm and have brought various pressures. Users post their complaint or sentiment in Twitter as online social networks are satisfactory platform for people to express their emotion.

High stress and depression are identified as the causes of obesity for medical researchers. The experimental results have also confirmed the conclusion. For example, with our dataset Memphis

with the high obesity does also have high stress and depression scores.

(4) Self-consciousness and obesity

During the initial analysis of the crawled dataset, we conceive that this feature is related with the obesity rate and we want validate our thought through the experiments based on large scale actual data. Figure 8 shows measurement results, although the correlation is not more significant that dietary and emotion features, the results can illustrate the correlation of obesity rate and self-consciousness. In the high obesity area users are often more likely to talk about losing weight. Given all these observations above, the measurement results show that the proposed features (except the physical activity features) are correlated with obesity rate.

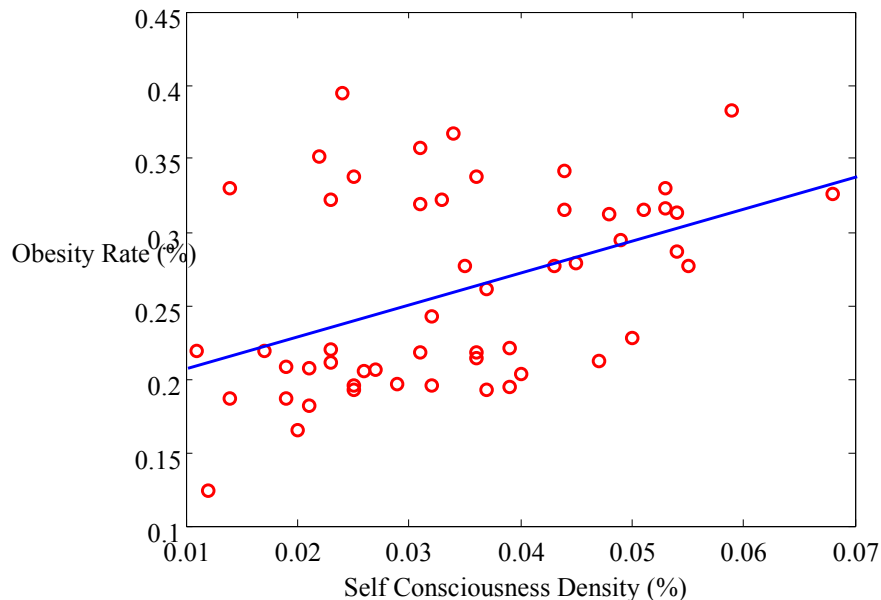


Figure 8. Correlation of self-consciousness and obesity rate

5.2. Obesity rate estimation

Epidemic model is usually used to estimate the condition of disease diffusion such as flu. Obesity is a kind of typical chronic disease which do not contain obvious stages like the infectious disease. In order to predicate the trend of obesity in certain area we propose an obesity trend estimation method based on the epidemic model idea. People can be divided into different groups and according to the number change of each group the obesity dynamic model in certain area can be constructed. The obesity status can be analyzed by the group size variation and related influence factors. The method is described as follows and related state transition diagram is shown in figure

9.

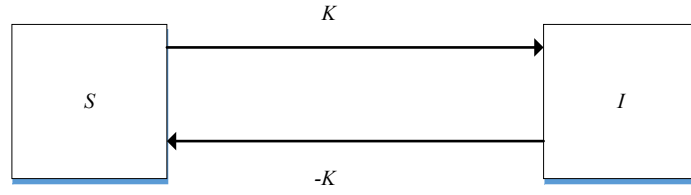


Figure 9. The figure of transferring from obese person to normal person

In a certain area, supposing that $I(t)$ denotes the number of people who are obese at time t , K denotes the coefficient of obesity, $S(t)$ denotes the number of people who are normal at time t .

$I(0)=I_0$ means that initially there are I_0 people who are obese.

Assumption 1. Without regard to the population mobility, death rate or other population dynamic factors, the total number of people from each area is a constant value which denotes N , $S(t) + I(t) = N$.

Assumption 2. The coefficient of variation K is a variable.

Assumption 3. When a person has gained weight, who will keep the obesity condition of body for a long time.

According to the assumptions above, the differential equation related obesity can be shown as follows:

$$\begin{cases} I(t) + S(t) = N \\ \frac{dI(t)}{dt} = KS(t) + I(t) \\ I(0) = I_0 \end{cases} \quad (7)$$

In our model the variation coefficient can be positive or negative. The value means that with the change of influencing factors obesity can adjust correspondingly. There are plenty of factors related to obesity and some factors are hard to measure in practice. In the previous sections, we have found that online social network can be used to reveal the life style. Therefore, according to user data the information related to obesity in online social network could be achieved and the coefficient K can be calculated with these various online data.

How to use online user features effectively is essential to obesity population status estimation. In this paper, we adopt multiple linear regression model to analyze the relationship between different online user features and the obesity status. At first the information features need to be filtered further by the Pearson correlation coefficient calculation process. The results are shown in table 1 ($p=0.01$).

Table 1. The Pearson correlation coefficients of different features and obesity rate

Features	Pearson Correlation Coefficient
High Calorie foods	0.548
Fast Foods	0.531
Emotion	0.516
Self-Consciousness	0.475
Popular Sports	0.146
Fitness Sports	0.057

As shown in table 1, it is clear that features like dietary habit and emotion are more related to obesity. For the physical activity, the results show that the correlations between this feature and obesity is relatively low. Therefore, during the analysis process of coefficient K the physical activity will not be involved. As for the rest of the features, we utilize Mallet to calculate the value of K with the multiple linear regression model. The concrete results are as following

$$K_l = \beta_0 + \beta_1 f + \beta_2 m + \beta_3 s \quad (8)$$

where K_l is the value of K based on multiple user features in certain area l . During the process, the Mallet has decoupled several features and eventually the dietary habit f , emotion m and self-consciousness s are variates for the regression model. β_0 is the random error. And $\beta_1=0.328$, $\beta_2=1.137$, $\beta_3=0.208$. After achieving the value of each parameter, we can obtain the obesity rate of certain area by input the feature value. For the parameters, the values are computed based on the data of each feature. If we obtain more user information, each parameter will be more accurate, and the obesity rate can be closer to the real value. K_l is a positive variate its value can be different as the habits of users reflected in the online social network are changed.

With the regression model, the obesity variation coefficient can be obtained by different related user features. During the experiment, we used a city as the tweet information search scope, and the cities are selected based on the Gallup wellbeing dataset which contains the obesity rate in some large scale American cities of 2012. We used the city data to represent the data of state which the city is locating in. Furthermore, as little tweet information is gathered from some cities so finally

we selected 40 states to carry out the experiments (The eliminated states are HI, AK, MT, ID, ND, SD, NE, WY, NM, NV and PR). Among our dataset the state with the most Twitter user is California (CA) which contains 61409 users, and the state with the fewest users is Washington (WA) which contains 2619 validated users. The concrete way is as follows: using the data official data of each state from CDC as the initial value, using the variation of yl with adjacent years as the value of coefficient K , and at last the obesity status of certain areas at time t can be obtained. The obesity estimation results are shown figure 10 and figure 11.

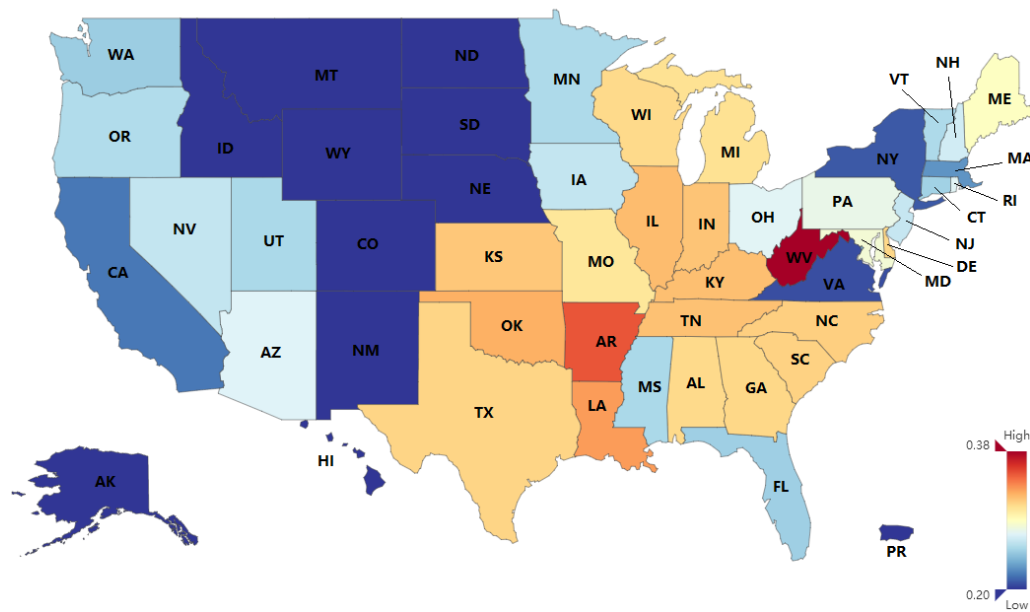


Figure 10. USA obesity status (2013) calculated by our method

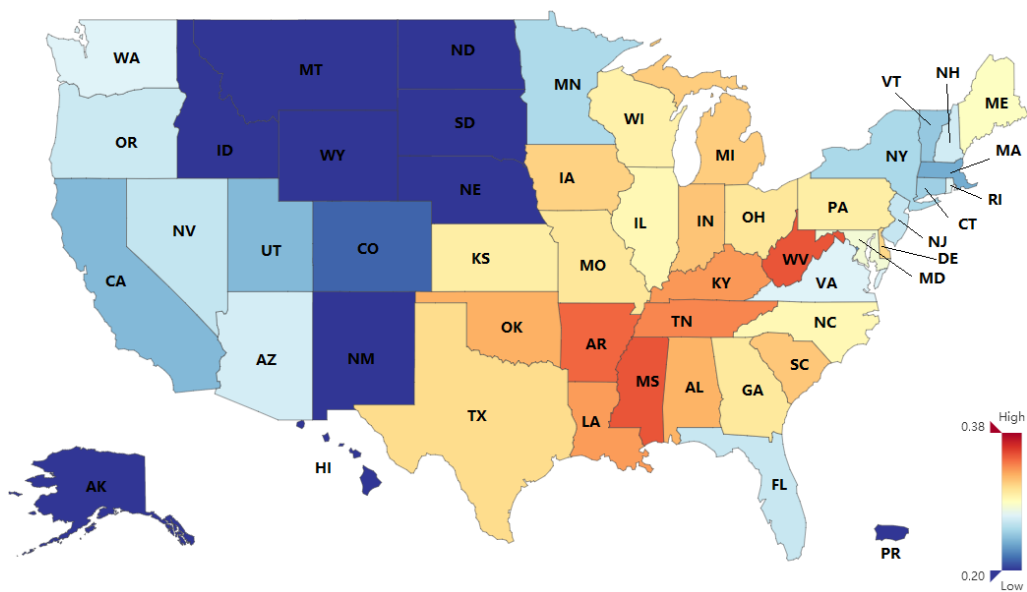


Figure 11. Official USA obesity status data (2013)

As shown in Figure 10 and Figure 11, different colors represent different obesity degree. Some state such as Alaska and New Hampshire are lack of enough online social network data therefore, the obesity data are not calculated in the experiment process and their colors are set to dark blue. Although in the result there are some calculation errors the whole trend calculated by our method is similar to the CDC data. To explain the effectiveness of the results further, we have calculated the different values between our results and the official data as shown in figure 12.

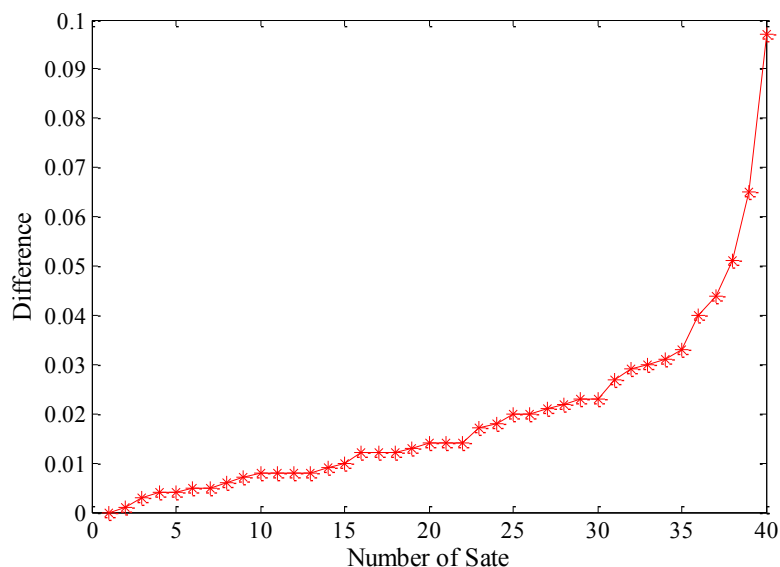


Figure 12. Different Values Between Our Results and The Official Data

According to the results shown in Figure 12, there are 19 states whose difference is lower than 0.01, there are 22 states with differences lower than 0.02. Compared with the disease study achievement such as on flu, the accuracy of the estimation on obesity is not at very high level but for the chronic disease the result can show the obesity trend to a certain degree. The reason of large error is due to the poor online social network data quality of these area, even the trend is similar but the difference is too large. Overall, the results are acceptable.

6. Discussion and conclusion

Users in online social networks can be regarded as different IOT nodes and various kinds of researches can be carried out according to their information. In this paper, we have extracted some features related obesity with online social network data crawled from online social networks and carried out some research on obesity detection in certain areas. The measurement on each feature is not just based on the simple statistics of key words on certain topic. User preference and their

emotion are also considered during the analysis process which is more detailed and reasonable.

According to the measurement results, we have found that the obesity rate is close related with dietary habit, emotion and self-consciousness of users. In addition, based on the experiment results we have found that the relation between physical activity and obesity rate is difference with some existing methods although there is something different about the measurement method. With the analysis on the features related to obesity, we have tried to estimate obesity rate in certain urban areas. The results show that the features we proposed in this paper have correlation with the obesity rate. Furthermore, although the obesity rate estimation results are not 100 percent accurate but the results can also validate the effectiveness of this work. This research outcome can be applied many other IoT healthcare cases as long as there are sufficient information in online social networks. Users can be seen as the social sensors and based on various user information we can obtain the current situation of each kind of disease among online users. Accordingly, we can implement relevant measures to the specific disease based on the online social network sensors.

However, there are still some issues needed to be addressed: (1) As obesity is a kind of chronic which is very different with influenza the periodic is not obvious. It is hardly to study the obesity problem based on Twitter information with detailed time information. (2) Obesity is related to many factors but for the subjectivity users may not directly post corresponding information in online social networks which may cause the data scale is not enough to carry out more detailed experiments. (3) The geographic information is important but only a few users present their geo information which can cause the bias to the dataset, which could be fixed by applying more accurate localization IoT systems^{[22][23]}. In future studies, above problems will be fixed and carry out more detailed study on obesity analysis and prediction. Specifically, by using more accurate localization.

Acknowledgement

The research presented in this paper is supported in part by the National Natural Science Foundation (No.: 61571360), Shaanxi Science & Technology Co-ordination & Innovation Project (No.: 2016KTZDGY05-09), and the Innovation Project of Shaanxi Provincial Department of Education (No.: 17JF019)

References

- [1] C. P. Dowling, J. J. Harrison, A. V. Sathanur, et al. Social sensor analytics: Making sense of network models in social media. *Intelligence and Security Informatics (ISI)*, 2015 IEEE International Conference on. IEEE, 2015.
- [2] S. Diplaris, S. Papadopoulos, I. Kompatsiaris, et al. Social Sensor: sensing user generated input for improved media discovery and experience. *Proceedings of the 21st international conference companion on World Wide Web*. ACM, 2012.
- [3] L. Chen, K. S. T. Hossain, P. Butler, et al. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data Mining & Knowledge Discovery*, 2016, 30(3): 681-710.
- [4] Executive summary: Guidelines (2013) for the management of overweight and obesity in adults. *Obesity*, 2014, 22(S2): S5–S39.
- [5] Abdelzaher T, Wang S, Giridhar P, et al. *Social Network Signal Processing for Cyber-Physical Systems*. Cyber-Physical Systems. 2017.
- [6] Lv Z, Song H, Basanta-Val P, et al. Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics. *IEEE Transactions on Industrial Informatics*, 2017, PP(99):1-1.
- [7] L. David, K. Ryan, K. Gary, V. Alessandro. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 2014, 343(6176): 1203-5.
- [8] E. Aramaki, S. Maskawa, M. Morita. Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011: 1568-1576.
- [9] S. Brennan, A. Sadilek, H. Kautz. Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions. *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*. 2013: 2783-2789.
- [10] M. J. Paul, M. Dredze, D. A. Broniatowski, Worldwide Influenza Surveillance through Twitter. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015: 6-11.
- [11] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, et al. Psychological Language on Twitter

Predicts County-Level Heart Disease Mortality. *Psychological Science*, vol. 26 no. 2
2015:159-169

- [12] M. E. Larsen, T. W. Boonstra, P. J. Batterham, et al. We Feel: Mapping Emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4), 2015: 1246-1252
- [13] Y. Wei, M. Lan. GIS analysis of depression among Twitter users. *Applied Geography*, 2015, 60: 217-223.
- [14] R. J. Gore, S. Diallo, J. Padilla. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. *PloS one*, 2015, 10(9): e0133505.
- [15] W. Y. S. Chou, A. Prestin, S. Kunath. Obesity in social media: a mixed methods analysis. *Translational Behavioral Medicine*, 2014, 4(3): 314-323.
- [16] Qi J, Yang P, Min G, et al. Advanced internet of things for personalised healthcare system: A survey[J]. *Pervasive & Mobile Computing*, 2017, 41.
- [17] Po Yang, Dainius Stankevicius, Vaidotas Marozas, Zhikun Deng, Enjie Liu, Arunas Lukosevicius, Feng Dong, Lida Xu, Geyong Min: Lifelogging Data Validation Model for Internet of Things Enabled Personalized Healthcare. *IEEE Trans. Systems, Man, and Cybernetics: Systems* 48(1): 50-64 (2018)
- [18] Li, S., Xu, L. and Wang, X. (2013) Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Transactions On Industrial Informatics*, 9 (4). pp. 2177-2186.
- [19] Shancang Li, Lida Xu, Xinheng Wang (2013): A Continuous Biomedical Signal Acquisition System Based on Compressed Sensing in Body Sensor Networks. *IEEE Trans. Industrial Informatics* 9(3): 1764-1771.
- [20] Centers for Disease Control and Prevention, Data, Trends and Maps Database, [Available online on 2nd Feb 2018] <http://www.cdc.gov/obesity/data/index.html>
- [21] Gallup news, Gallup-Sharecare Well-Being Index, [Available online on 2nd Feb 2018] <http://www.gallup.com/poll/106756/galluphealthways-wellbeing-index.aspx>
- [22] P. Yang, W. Wu, "Efficient Particle Filter Localisation Algorithm in Dense Passive RFID Tags Environment", *IEEE Transactions on Industry Electronics*, Vol 61, Issue 10, pp.5641 - 5651, Oct, 2014.
- [23] P. Yang, "Experimental Evaluation and Analysis of Achieving High Accuracy and Precision

in Passive HF RFID Location Systems”, IEEE Internet of Things Journal, (IF: 7.596), Vol 2, Issue, pp.159-167, 2015.