Big Data Architecture Design for Patterns of Social Life Logging

A. Hana Lee¹, B. Youseop Jo², C. Ayoung Cho¹, D. Hyunwoo Lee¹, E. Youngho Jo², and F. Mincheol Whang³

¹Department of Emotion Engineering, Sangmyung University, Seoul, Republic of Korea
²Team of Technology Development, Emotion Science Center, Seoul, Republic of Korea
³Team of Intelligent Engineering Informatics for Human, Sangmyung University, Seoul, Republic of Korea

Abstract - The era for data-driven decision-making has been growing tremendously due to the use of technology along with the need to know on what the users need. Collecting massive amount of data is not a limitation anymore, however, providing meaningful data is a limitation people face today. With the increasing number of social network platforms, there has been a development in new social environments. When people are networking through these social platforms, many emotional aspects are also being transferred, also known as emotional contagion. When people are going through these emotional phases they are experiencing synchronization. Therefore, this study is to develop a new data process to predict user’s needs based on the user’s synchronization patterns in a social network context.

Keywords: big data, architecture, social emotions, social networking, synchronization

1 Introduction

Determining user needs has been one of the main goals of marketing to ensure the satisfaction of the users. To determine the needs of the user, we need to determine who the user is. There are different ways to determine who the user is and one of them is to collect information about him or her, which is highly accessible in today’s world of technology. The new era for data-driven decision-making has been growing tremendously due to the use of technology along with the need to know on what the users need. Collecting massive amount of data is not a limitation anymore, however, providing meaningful data to this data is a limitation many people face today. To provide meaningful data, the data collected should have a significant collective meaning to the users and furthermore must be personalized [1]. Personalization will provide the optimal recommendation system because it is only for the distinct individual or user.

In life logging, personalization can be determined with four main factors as follows: the user’s physiological state, behavioral patterns, content consumption, and spatial factors. First, a person’s emotional state is a personal experience of events and having a physiological response to this event with an appropriate appraisal [2]. However, more recent theories claim that physiological responses and emotional appraisals happen almost simultaneously. Emotion recognition can be determined physiologically by measuring cardiovascular responses [3,4,5,6,7,8,9,10]. The body’s autonomic nervous system consists of two main divisions; the sympathetic nervous system (SNS) and parasympathetic nervous system (PNS). Both SNS and PNS have different responses to how the body responds during emotional stimulations. Many studies claim a stable, sine-wave like pattern in the heart rate variability waveform include physiological information. Coherent heart rhythm can therefore be defined as a relatively harmonic (sine-wave-like) signal with a very narrow, high amplitude peak in the LF region of the HRV power spectrum and no major peaks in the VLF or HF regions [5].

Although, cardiovascular responses mirror the involuntary behaviors of the human heart responses, behavioral responses reflect more on voluntary actions of individuals. Usually behavior can be reflected by a person’s action tendency. Before implying the action tendency, one must understand how to achieve human behavior, that person must go through an appraisal processes. This cognitive process claims that a person needs to determine or appraise the situation or event he or she is in to act accordingly. Therefore, the behavioral patterns of users are determined by measuring the user’s movement activity using his or her global positioning system (GPS) information on his or her mobile device. Studies have been conducted to determine the daily life patterns utilizing mobile devices by obtaining GPS data [11]. Daily life patterns through mobile devices show the users what kind of lifestyle the users reflect. Are they someone who moves in a routinely or spontaneously?

Due to the internet of things (IOT) we are constantly surrounded by contents that are easily accessible to users all around the world. With over millions of users signed up for a social network account, there are many studies on how the Internet is a new form of society that can affect our moods and emotions. When consumed, these contents can have an emotional impact on the user’s consuming them [12,13]. Whether it is a trending topic on Twitter or a Facebook article, users are surrounded by content where they can give
emotional feedback on the topic. Being exposed to contents that can have an emotional impact is a new form of consumption.

Finally, our spatial surrounding is the environment we surround ourselves in. This is another factor that can affect the user’s emotions. Spatial factors such as ambient noise and spatial complexity determines the loudness of our surroundings and the complexity of it can affect our emotions [14,15,16,17]. The ambient noise that exists in our surroundings has an impact on our emotions regardless of our consciousness.

Determining the user’s physiological state, behavioral patterns, content consumption, and spatial information is as personalized as it gets. However, the point of a social network is to have interpersonal connections with other users. Connections with other users can be made stronger and provide more meaning by determining synchronization. Synchronization is the operation or activity of two or more things at the same time or rate. This has been a phenomenon that focuses on the concept of two or more people in synchrony with physiological responses at the same time or rate. Synchronization exists in the interactions between humans. Whether they synchronize with cardiovascular responses, emotional, or even behavioral, the ability to synchronize shows that humans have the unique capability to empathize with each other [18]. When emotions are synchronized it can be inferred that two or more individuals are experiencing the same emotions, or at least somewhat similar. This usually occurs during the processes of emotional contagion [19,20]. Although quite controversial, due to privacy issues, social media platforms like Facebook have already started research regarding the contagiousness of emotions through the internet [21]. Having emotional interaction in a social network environment has been increasing due to the booming innovations of mobile applications and the Internet.

Social network platforms have impacted society’s definition on social emotions. Social situations have been studied for as long as people existed in group settings. Social groups are made up of more than one individual, and in these groups, there is known to be a collective emotion within the members [22]. This can also be done through emotional contagion [19]. However, the emotions of individuals in these social settings is important because of the impact it has on other individuals. That is why a social emotion model needs to be modelled after Russell’s two-dimensional model of emotions in a social network context [23, 24].

Therefore, this study is to develop a new data process to predict user’s needs based on the user’s synchronization patterns in a social network context. The proposed architectural design is believed to be more accurate than the previous data processing on determining user’s needs. The key to the new design is synchronization and personalization of data collection for each user and building a social network accordingly. This paper proposes the system architecture and structures of data to develop the optimal social network and recommendation system, however the accuracy of this system and user’s satisfaction is an ongoing process that will be analyzed in the future with experimental data.

2 Data Collection

2.1 Emotional Data

There are four main data features collected for this system. The first is cardiovascular features, behavioral features, content consumption features, and spatial features. The index for these features are defined in many previous studies and were utilized for this system to determine any possible feature that can affect a user’s emotional state. Emotional data has been collected through a series of subjective evaluations to label the user’s social emotions.

2.2 Cardiovascular Features

The cardiovascular features have been measured and analyzed to determine the physiological responses of individuals. These physiological features have been collected to determine the optimal index of cardiovascular responses in social conditions. Data for cardiovascular features were collected by minute intervals throughout the day. These features were measured with a wireless wearable photoplethysmography (PPG) sensor that was developed for this study, called the SociaL Band.

A total of 21 heart rate variability (HRV) is assessed in time or frequency domains. The time domain measurements are in beats per minute (BPM), standard deviation of normal to normal (SDNN), root mean square of successive differences (rMSSD), and pNN50. The frequency domain consists of very low frequency (VLF), low frequency (LF), high frequency (HF), total power, VLF percentage, LF percentage, HF percentage, natural logarithms of VLF, LF, and HF, VLF/HF ratio, LF/HF ratio, dominant power, dominant Hz, peak power, peak Hz and coherence ratio.

**Beats Per Minute.** BPM measures the average pulses the heart is beating per minute.

\[ \text{Mean (PPI)} = \frac{1}{N} \sum_{i=1}^{N} \text{PPI}_i \]

\[ \text{BPM} = \frac{60}{\text{Mean (PPI)}} \]  

(1)

**Standard Deviation of Normal to Normal.** SDNN measures the heart’s intrinsic ability to respond to hormonal influences and normal values range from 35-50Hz. When SDNN values show a pattern of increase, it implies the individual is feeling the cognitive load of stress or fatigue. On the contrary,
decrease in SDNN implies de-stressing patterns of cognitive load.

\[
SDNN = SD(PPI) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} [\text{Mean}(PPI) - PPI_i]^2}
\]

(2)

**Root Mean Square of Successive Differences.** RootMSSD estimates the parasympathetic regulation of the heart. The increase in values indicates arousal and decrease in values indicates relaxation.

\[
rMSSD = \sqrt{\frac{1}{N-2} \sum_{i=2}^{N} (PPI_i - PPI_{i-1})^2}
\]

(3)

**PNN50.** The pNN50 statistic is the NN50 count, defined as the mean number of times per hour in which the change in consecutive normal sinus (NN) intervals exceeds 50 milliseconds. The authors proposed this measure to help assess parasympathetic (vagal) activity from 24 hours of ECG recordings.

\[
pNN50 = \frac{\text{NN50 count}}{\text{total NN count}}
\]

(4)

**Very low frequency.** VLF is in the power range of 0.0033 to 0.04 Hz in the frequency domain, which usually indicates sympathetic modulation.

\[
\text{Power} = \text{FFT}(PPI)
\]

\[
df = \frac{\text{SamplingRate} \times \text{Length}(PPI)}{\text{Time}} = \frac{1}{\text{Time}}
\]

(5)

\[
\text{VLF} = \sum_{i=1}^{\frac{0.04}{df}} \text{df} \text{ Power}_i
\]

**Low Frequency.** LF is in the power range of 0.04 to 0.15 Hz and is usually the indicator of both sympathetic and parasympathetic modulation.

\[
\text{LF} = \sum_{i=1}^{\frac{0.15}{df}} \text{df} \text{ Power}_i
\]

(6)

**High Frequency.** HF is in the power range of 0.15 to 0.4 Hz, which can contain the rhythms that regulate the parasympathetic activity.

\[
\text{HF} = \sum_{i=1}^{\frac{0.4}{df}} \text{df} \text{ Power}_i
\]

(7)

**Total power.** It is a short-term estimate of the total power of power spectral density in the range of frequencies between 0.0033 and 0.4 Hz. Total power measures reflect the overall autonomic nervous activities, where sympathetic activity is a primary contributor.

\[
\text{TotalPower} = \sum_{i=\frac{0.0033}{df}}^{\frac{0.4}{df}} \text{df} \text{ Power}_i
\]

(8)

**Very Low Frequency percentage.** VLF percentage indicates the division of VLF by the total power range.

\[
\text{VLFp} = \frac{\text{VLF}}{\text{TotalPower}}
\]

(9)

**Low Frequency percentage.** LF percentage indicates the division of LF by the total power range.

\[
\text{LFp} = \frac{\text{LF}}{\text{TotalPower}}
\]

(10)

**High Frequency percentage.** HF percentage indicates the division of HF by the total power range.

\[
\text{HFp} = \frac{\text{HF}}{\text{TotalPower}}
\]

(11)

**Natural Logarithm of Very Low Frequency.** LnVLF indicates the natural logarithm of the very low frequency. This allows any exponential values of VLF to be natural.

\[
\ln\text{VLF} = \ln(\text{VLF})
\]

(12)

**Natural Logarithm of Low Frequency.** LnLF indicates the natural logarithm of low frequency rates. This allows any exponential values of LF to be natural.

\[
\ln\text{LF} = \ln(\text{LF})
\]

(13)

**Natural Logarithm of High Frequency.** LnHF indicates the natural logarithm of high frequency rates. This allows any exponential values of HF to be natural.

\[
\ln\text{HF} = \ln(\text{HF})
\]

(14)

**Very Low Frequency/High Frequency ratio.** VLF/HF ratio measures the sympathetic responses and high ratio numbers indicate sympathetic responses akin to agitation.

\[
\text{VLF/HF} = \frac{\text{VLF}}{\text{HF}}
\]

(15)

**Low Frequency/High Frequency ratio.** LF/HF ratio measures the overall balance between sympathetic and parasympathetic
systems. When the ratio is high it indicates the domination of the sympathetic system and oppositely, when the ratio is low it indicates domination of the parasympathetic system.

\[
\frac{LF}{HF} = \frac{LF}{HF}
\]

(16)

**Dominant Power.** Dominant power indicates the power of the highest peak in the power spectrum.

\[
\text{Dominant Power} = \text{Power}_{\arg \max (\text{Power})}
\]

(17)

**Dominant Hz.** Dominant Hz modulates both the sympathetic and parasympathetic activities. A high dominant Hz value indicates relaxation and a low value indicates arousal.

\[
\text{Dominant Hz} = \arg \max (\text{Power}) \times df
\]

(18)

**Peak Power.** Peak power indicates the band of power spectrum range between -0.015 and 0.015 Hz based on Peak Hz.

\[
\text{Peak Power} = \sum_{i=\text{Peak Hz}+0.015}^{\text{Peak Hz}+0.015} \text{Power}_i
\]

(19)

**Peak Hz.** Peak Hz resembles the highest peak in the power spectrum range between 0.04 and 0.26 Hz. The values indicate the frequency range in which coherence and entrainment can occur.

\[
\text{Peak Hz} = \arg \max (\text{Power}_i) \times df
\]

(20)

**Location variance.** Location variance measures the variability in a user’s GPS location. Latitude and longitude are components of the algorithm to calculate the location variance.

\[
\text{Location Variance} = \log (\sigma_{\text{lat}}^2 + \sigma_{\text{lon}}^2)
\]

(22)

**Number of Clusters.** The number of clusters represents the number of location clusters found by the K-means algorithm in the preprocessing stage.

**Entropy.** Entropy is the measurement of the variability with the time a user spent at the location of clusters. It was developed based on the concept of entropy from information theory. High entropy implies the time spent more uniformly across the different location clusters and oppositely, low entropy indicates greater inequality in the time spent across the cluster. For example, if a user spends 85% of his/her time at home, and 15% of his/her time at work that user’s entropy would be lower than if he/she spent 50% of time at home and 50% at work (higher entropy).

\[
\text{Entropy} = -\sum_{i=1}^{N} p_i \log p_i
\]

(23)

**Transition Time.** Transition time is the percentage of time during which a user was in a non-stationary state. The percentage is calculated by dividing the number of GPS location samples in transition states by the total number of samples.

**Total Distance.** Total distance, in kilometers, was taken by a user and it is calculated by accumulating the distance between the location samples.

\[
\text{Total Distance} = \sum_{i}^{N} \left( \frac{\text{dist}_i}{\text{total distance}} \right) \text{dist}_i
\]

(24)

**Circadian Movement.** Circadian movement indicates the temporal information of the location data. In other words, this measurement indicates the patterns of movement and to what extent a user’s sequence of locations followed a circadian rhythm. For example, if a user went to school and came home at the same time everyday then the circadian movement would be high. On the other hand, a user with more irregularity in patterns of moving between locations had lower circadian movement.

\[
E = \sum_{i=1}^{N} \text{psd} (f_i) / (i_{1} - i_{N})
\]

(25)

2.3 Behavioral Features

The behavioral features measured were latitude, longitude, location variance, number of clusters, entropy, transition time, total distance, and circadian movements. These features were modeled after the study referencing the measurements of user behaviors through smartphone devices [11]. The behavioral data collected had a set range of 1 kilometer/hour to determine stationary and transitional states of the users.
2.4 Content Consumption Features

The content consumption features were measured by the frequency and duration of the how often and how long a content was viewed, respectively. This would measure how focused or interested the user was while viewing the content. Later, the data collected would use a natural-language processing (NLP) to determine the emotional aspects of the content viewed. The content was separated into 15 different categories such as politics, administration, society, economy, business, IT/science, sports, culture, game, car, events, weather, health, lifestyle, and self-improvement.

2.5 Spatial Environment Features

There were 9 spatial features that measured time complexity, vertical edge, horizontal edge, hue, saturation, intensity, contrast, amplitude, and sound frequency. These were measured with the mobile device by getting the user’s consent of collecting the sound frequency and asking them to make 5 second recordings of his or her surroundings. The video footage would be analyzed for the spatial complexity and the audio files would be analyzed for noise complexity of the environment.

3 Data Architecture

The design is divided into four main parts; the first is sensing and device, the second is analysis and server, the third is archiving and database, and the fourth is service and client. Sensing is measured with a PPG sensor and mobile device. All features are measured with these two devices as shown in Fig 1. The measured data is then wirelessly streamed in real time into the database, which is consists of three levels of data layering.

Primary data is raw data collected by the wireless devices. This data is then recognized through the Context Recognition Module that consists of an index that recognizes the different features of cardiovascular responses, behavioral activity, spatial context, and content consumption of the users. This data then becomes the secondary data.

The secondary data consists of the recognized features and inserts the data into the Social Emotion Recognition Module, which extracts the significant features from each individual and determines the user’s social emotions through statistical analysis. After, with the significant features are extracted, synchronization patterns are determined. The secondary data then becomes the tertiary data.

The tertiary data consists of synchronization patterns of all the different contexts such as PPG, GPS, spatial, and contents and is inserted into the Social Emotion Group Pattern Analysis Module. This module analyses the synchronized context patterns along with the user’s activity analysis to determine the group’s pattern. The activity analysis is the user’s social activity patterns that represents the amount of posts and social activity through the Social Application. With the tertiary data, the user’s synchronization group (network) is determined and the group’s pattern and trend is analyzed to determine the
optimal recommendation for the user. This recommendation system determines the user's needs by making activity recommendations the user needs according to their life-logging data.

4 Preliminary Results

After all the filtering and analyzation of the data, an experiment is conducted to stimulate the appropriate social emotions based on social networking context. Followed by the experiment, the amount of data analyzation is deemed to be voluminous. The interrelationships of these specific features will provide a pattern of how physiological, psychological, content consumption, and spatial environments can represent the full aspects of our daily life patterns. That is why it is crucial to detect these patterns for each user and provide a network for users to be able to make a deeper connection through synchronization. As a result, the proposed big data structure is expected to have a tremendous impact in the fields of artificial intelligence, big data mining, and affective computing.

5 References


