Information & Management 52 (2015) 777–788

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Emotion recognition and affective computing on vocal social media

Weihui Dai [a,*], Dongmei Han [b,c], Yonghui Dai [b], Dongrong Xu [d]

[a] Department of Information Management and Information Systems, School of Management, Fudan University, Shanghai 200433, China
[b] School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China
[c] Shanghai Financial Information Technology Key Research Laboratory, Shanghai 200433, China
[d] Psychiatry Department, Columbia University/New York State Psychiatric Institute, New York City, NY 10032, United States

ARTICLE INFO

Article history:
Received 7 September 2014
Received in revised form 17 December 2014
Accepted 14 February 2015
Available online 25 February 2015

Keywords:
Social media
Social network
Voice instant messaging
Vocal data mining
Emotion recognition
Affective computing

ABSTRACT

Vocal media has become a popular method of communication in today's social networks. While conveying semantic information, vocal messages usually also contain abundant emotional information; this emotional information represents a new focus for data mining in social media analytics. This paper proposes a computational method for emotion recognition and affective computing on vocal social media to estimate complex emotion as well as its dynamic changes in a three-dimensional PAD (Position–Arousal–Dominance) space; furthermore, this paper analyzes the propagation characteristics of emotions on the vocal social media site WeChat.

1. Introduction

In today's social networks, communication methods are undergoing a change due to emerging vocal social media such as WeChat, QQ (China), ICQ, WhatsApp (U.S.), Line (Japan) and various other tools for instant voice messaging. While facilitating the conveyance of semantic information, vocal social media can also transmit abundant emotional information. This ability has resulted in significant influence not only in terms of improving the user's experience and sense of belonging to particular social groups and thereby enhancing their continuance intentions toward these groups [31,60] but also in terms of strengthening the interpersonal relationships between the members of these groups and the community's cohesion and cognitive consistence in the social network [23,57,58].

In the relevant literature, we can find a wide range of applications for social networks due to the rapid development of social media analytics [1,2,53], which provide an effective methodology for unearthing business value from social networks [3,13,51]. Recently, the strong influence of social networks on psychological cognition and social behaviors has generated much attention [5,9,22,27,48]. In our previous research findings [23], the propagation effects of social networks on emergent events affect the community through a process that contains five interactional layers, which mostly depend on group cognitions: information, emotion, attitude, behavior and culture. Among these, emotion takes the role of inducing the group's primary recognitions; this leads to a consistent attitude and behavioral reactions in the "small world" through the member's close social relationships, trust and empathic effects. Recent neural and behavioral research work has also indicated the theoretical basis for this phenomenon [32]. The effects of vocal social media on interpersonal relationships, group cognitions and, particularly, emotion propagation endows social networks with some new prominent features and social functions worthy of further research.

In recent years, the emotional impact of social media on society has been confirmed by an increasing number of research findings and empirical cases and thus has drawn great attention from a variety of areas such as Internet marketing research, service comments analysis, social mood monitoring, and emergent event management [5,8,9,22,27]. Social media is considered to be a sensor that perceives and predicts society's behaviors in the real world through data mining techniques targeting emotional information [5,27]. Since 1997, when Professor R.W. Picard at

* Corresponding author. Tel.: +86 2125011241; fax: +86 21 65644783.
E-mail addresses: whdai@fudan.edu.cn (W. Dai), handongmei19610320@gmail.com (D. Han), dyh822@163.com (Y. Dai), dx2013@columbia.edu (D. Xu).

http://dx.doi.org/10.1016/j.im.2015.02.003
0378-7206/

the Massachusetts Institute of Technology proposed in his book *Affective Computing* that a computer could capture, process and reproduce human emotions [40], this issue of human emotion recognition and computing has been explored by machine intelligence technologies. Among these efforts, emotion recognition attempts to identify the possible types of emotions based on signals, which can be regarded as a pattern recognition task. Affective computing usually requires, in addition, a quantitative measurement of the emotion, which is commonly related to the issue of value estimation based on a trained model.

To date, researchers have proposed a series of computational models for analyzing emotional information from social media data [14,19,26,48]. Due to the widespread application of voice instant messaging tools, emotion recognition and computing on emerging vocal media has become a hot new area for research in social media analytics and data mining. Vocal emotion recognition has made great progress in recent decades, from speaker-dependent, template-matching recognition based on simple vocabulary to today's speaker-independent statistical model-based recognition, which can process continuous speech [19,29]. However, there are still barriers to addressing vocal social media. The speech signal in vocal social media is human conversation using natural language and, in most cases, contains mixed emotions embedded with dynamic changes. This signal cannot be simply recognized as a specific typical emotion by the existing methods. Precisely computing such complex and dynamic emotions is more technically difficult. Therefore, this issue must be studied in more depth. This paper aims to develop an effective computational method for processing complex and dynamic emotions from the speech signals of vocal social media, so that the propagation effects of emotions may be analyzed in a meticulous and in-depth way.

2. Literature review

2.1. Emotion recognition of vocal signals

Vocal emotion recognition involves the issues of emotion classification, signal pre-processing, feature extraction, and pattern recognition. The classification and description of human emotions continues to be a controversial issue. The classification of emotions falls into two categories: discrete form and continuous form. The discrete form only provides emotion types, such as the "big six": anger, disgust, fear, joy, sadness, and surprise [12]. The continuous form describes the emotional state in a continuous space with different dimensions. Among these, the single dimension only classifies positive or negative emotions and determines their strength. The 2-D spaces are usually based on Hidenori and Fukuda's Emotional Space [20], where the emotional state is represented in a unit circle with two opposing coordinates: peace vs. excitement, happiness vs. sadness. The 3-D space has different models presented by Wundt [54], Schlosberg [42], Izard [24], and Osgood [39]. Based on comprehensive psychological research, Mehrabian demonstrated that any type of emotional state can be well described by three nearly independent continuous dimensions: Pleasure–Displeasure (P), Arousal–Non-arousal (A), and Dominance–Submissiveness (D). Based on this, he therefore proposed the famous PAD model [35,36]. This model provides an effective means for evaluating complex emotions, and has been successfully applied to manual subjective measurements in a variety of areas [25,33,49]. Fig. 1 shows the continuous form of emotions in different dimensions.

In the signal pre-processing procedure, the initial speech signal is transformed so that its acoustic feature parameters can be extracted. Generally speaking, this procedure includes three steps: signal sampling and quantizing; pre-emphasis; and framing and windowing [50]. In this regard, one of the most important contributions was made in the 1970s, when DTW (Dynamic Time Warp) and VQ (Vector Quantification) were presented to address the problems arising from the different lengths of speech signals [41]. The extraction finds appropriate and effective parameters that can be used to identify the emotions from the vocal signal. The commonly used acoustic parameters are divided into three categories [19]: (1) prosody parameters such as the duration, pitch and energy of a vocal signal; (2) spectral parameters such as the LPC (Linear Predictor Coefficient), OSALPC (One-Sided Auto-correlation Linear Predictor Coefficient), LFPC (Log-Frequency Power Coefficient), LPCC (Linear Predictor Cepstral Coefficient), and MFCC (Mel-Frequency Cepstral Coefficient); and (3) sound quality parameters such as format frequency, bandwidth, jitter, shimmer, and glottal parameter. In the above categories, prosody parameters are the basic parameters for vocal emotion recognition. The human auditory system is a special nonlinear system that responds selectively to different frequency signals. The MFCC is based on known variations of the human ear's bandwidths and includes frequency characteristics linearly below 1000 Hz and logarithmically above 1000 Hz, which matches well with the auditory characteristics of human speech signals. Experience shows that the performance of the MFCC parameters is usually better than the performance of other spectral parameters [17,37]. Recent experiments have found that sound quality parameters play an important role in differentiating the emotions associated with attitudes and intentions; therefore, combined parameters applying all three categories to the feature extraction may be the new trend [16,56].

In pattern recognition, methods are usually based on the Hidden Markov Model (HMM), Artificial Neural Networks (ANN), the Gauss Mixture Model (GMM), the Support Vector Machine (SVM) and Bayesian Classification. In their paper *Speech Emotion Recognition Using Hidden Markov Models*, Nwe et al. reported that the six typical emotions, anger, disgust, fear, joy, sadness and surprise, can be recognized at an accuracy rate of 78% [38]. In *Toward Detecting Emotions in Spoken Dialogs*, Lee and Narayanan correctly recognized "positive" and "negative" emotions from the dialogues of call voice using combined information from the speech and its converted text [28]. Using HMM and GMM, Schuller

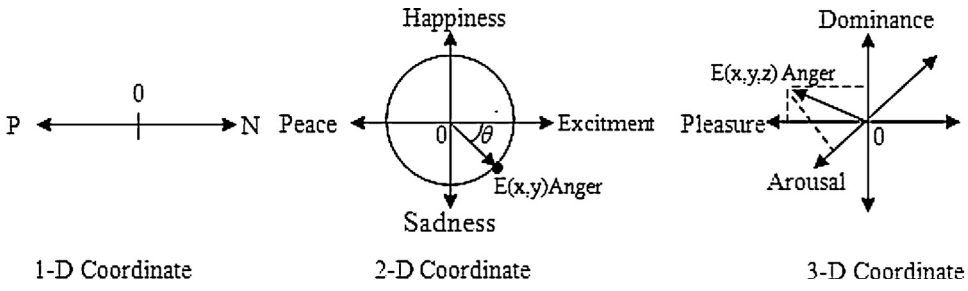

Fig. 1. Continuous form of emotions in different dimensions.