2014 11th Web Information System and Application Conference

Quantifying Chinese Happiness via Large-Scale Microblogging Data

[1]Chong Kuang, [1]Zhiyuan Liu, [1]Maosong Sun, [2]Feng Yu, [3]Pengfei Ma
[1]State Key Laboratory of Intelligent Technology and Systems
[1]Tsinghua National Laboratory for Information Science and Technology
[1]Department of Computer Science and Technology, Tsinghua University,Beijing, China
[2]Department of Psychology, Tsinghua University, Bejing, China
[2]School of Economics and Management, Tsinghua University, Bejing, China
[3]Research Institute of Information Technology, Tsinghua University, Bejing, China
Email: {kuangchong07,lzy.thu}@gmail.com, sms@tsinghua.edu.cn, yufengfengfeng@gmail.com, mpf07@mails.tsinghua.edu.cn

***Abstract*—Happiness is an important indicator to measure our life satisfaction. Microblogging data can reflect users' living standards and psychological state. We build a large lexicon based on the PERMA lexicon and large-scale microblogging data on Sina Weibo by combining the PMI and distrabutional similarty method. Using this lexicon, we propose a method to calculate the happiness quantitatively based on PERMA theory. Experiments shows that our method achieve significant improvement compared with the baseline in terms of AP and Bpref respectively. After the verification of our method on manual annotation dataset, we perform an in-depth analysis on large-scale microblogging data using our method.**

***Keywords*—Microblogging; PERMA; Happiness.**

I. Introduction

Social media development is in full swing nowadays. It has become one of the most effective platforms for people to access information and communicate with each other. In China, the largest social media platform is Sina Weibo[1]. Apart from text messages, videos, URLs, and images are allowed as well. All these forms of messages are called *microblogs*. A large number of people use microblogs to share information and their statuses and express their moods. The large number of users on social media has led to the generation of massive information, which is of great potential value.

With the development of Chinese society, people start to pay increasing attention to the quality of life. Due to the rapid development, social media has become an important platform for the researches of social psychology recently. Happiness is an important research topic in positive psychology. Typical research work focuses on the demographic data. However, it's quite inefficient and is bound by statistical data. Recently, a lot of researches have been done on studying users' happiness via large-scale data in Twitter (see Section II). Compared with traditional methods, these methods can get the real time quantitative happiness results. But the measurement of happiness is still not settled yet in psychology field. Seligman et al.[1] proposed the PERMA theory, which contains five dimensions, to extend traditional happiness measurement that only takes positive emotions into account. The PEAMA theory gradually has become the primary measurement of happiness. The PERMA[1] lexicon contains five collections of words relating to five components of positive psychology: positive emotions, engagement, relationships, meanings, and achievement. Each component has positive dimension and negative dimension.

In this paper, a detailed and quantitative analysis of Chinese happiness is performed. We translate the happiness words lexicon accurately and use a series of methods to expand the lexicon in large microblogs dataset on Sina Weibo. We build our method for happiness computation based on our expand-PERMA lexicon and consider the impact of other words. Experiments show that our method can measure happiness quite well. Then a series of analysis have been conducted in different dimensions: dates, months, days and locations.

To sum up, our main contributions are: (1) we propose an effective method to expand the lexicon based on PERMA lexicon and large-scale data; (2) we propose an effective measurement to calculate the happiness in Chinese microblogging environment; (3) we preform an in-depth analysis about happiness on large-scale dataset.

II. Related Work

A large number of researches have been done to analyse Twitter and Chinese social media during the past few years. Java et al.[2] analyzed Twitter as early as in 2007. They described the social network of Twitter users and investigated the motivation of Twitter users. Bol-len et al.[3] studied the influence of moods on twitter on the stock market.

People's happiness is an important societal metric and typical measurement is through self-reporting. Happiness is hard to measure because it's a kind of people's subjective feelings. Psychologists commonly used questionnaires to measure people's happiness. It is a laborious process. Meanwhile, happiness has often been indirectly measured by economic indicators, such as GDP per capita. As the social media develops, happiness measurement based on large-scale data has become a new research topic. Peter Sheridan Dodds et al.[4] built their metric word list for happiness after analysing temporal patterns of happiness and information content for the very large data set generated by Twitter. Mitchell L et al.[5] investigated the correlations between average user happiness value and a wide range of emotional, geographic, demographic, and health characteristics. Morgan R. Frank et al.[6] found that

[1]http://weibo.com/

978-1-4799-5727-9/14 $31.00 © 2014 IEEE
DOI 10.1109/WISA.2014.48

expressed happiness increased logarithmically with distance from an individual's average location. H. Andrew Schwartz et al.[7] also built a lexicon and used it to analyse subjective well-being based on the geo-tagged tweets.

Most of the works focus on the English environment. But there is a big gap between Twitter and Sina Weibo, such as language convention and traditional culture. In China, few research has been done on happiness computation except some media and abovementioned methods on English environment can't be directly used in Chinese environment. Meanwhile, the quantitative measurement of happiness is still not settled yet in recent works. Our work aims to find a feasible way to calculate the happiness based on the data from Chinese social media.

III. Methodology

A. Framework

Many people post microblogs to express their living conditions on Sina Weibo. Our basic assumption is that people express their moods through their microblogs, intentionally or unintentionally. We only consider the original text microblogs and assume that the extent of happiness of each microblog is independent. We first compute the value of happiness produced by each microblog separately. then we can get the average happiness value of users in the set.

We'll give each microblog a formalized definition: weibo, originally defined as a text, would become a sequence of words after CWS (Chinese Word Segmentation) process:

$$m_i = \overrightarrow{t_i} =< t_{i1}, t_{i2}, \cdots, t_{ij}, \cdots, t_{in} >, m_i \in M \quad (1)$$

where n is the number of words in microblog m_i, t_{ij} is the j-th word in microblog m_i. The M represents the microblog sets, which are assigned with different meanings in different situations. For example, when we compute one person's happiness value, M represents all his original text microblogs. When we compute the happiness value on Sunday, M represents all original text microblogs posted on Sunday.

We use $h(m_i)$ to represent the happiness value of microblog m_i. Each word t_{ij} contributes to $h(m_i)$ of different weight. We use v_{ij} to represent the weight of t_{ij}. We compute the v_{ij} based on the Eq.2. Then we can get the weight vector $\overrightarrow{v_i}$ as : $\overrightarrow{v_i} = \overrightarrow{v_{m_i}} =< v_{i1}, v_{i2}, \cdots, v_{ij}, \cdots, v_{in} >$

The weight of each word could be affected by the other factors, such as privative words. So we consider all the factors, which affect the word t_ij, as w_ij. We compute the w_{ij} based on Eq.3. Then we get the impact vector $\overrightarrow{w_i}$ as follows: $\overrightarrow{w_i} = \overrightarrow{w_{m_i}} =< w_{i1}, w_{i2}, \cdots, w_{ij}, \cdots, w_{in} >$

Since we get the word's weight $\overrightarrow{v_i}$ and impact $\overrightarrow{w_i}$. We compute the happiness value via the inner product of $\overrightarrow{v_i}$ and $\overrightarrow{w_i}$: $h(m_i) = \overrightarrow{w_{m_i}} \cdot \overrightarrow{v_{m_i}} = \overrightarrow{w_i} \cdot \overrightarrow{v_i} = \sum_{j=1}^{n} w_{ij} \cdot v_{ij}$

Now we can compute each microblog's happiness value. We assume that each microblog's happiness is independent, so the overall happiness value in dataset is calculated as follows: $h(M) = \sum_{m_i \in M} h(m_i)/|M|$

B. PERMA Lexicon and Expand-PERMA Lexicon

Our lexicons are based on the lexicons that Schwartz et.al. [7] used for calculating happiness on Twitter. The lexicons combined PERMA lexicons and LIWC [8] (Linguistic Inquiry and Word Count). We first invite three Psychology's PhD candidates to translate the lexicons into Chinese language. Then we invite another three Psychology's PhD candidates to check the translation. We strict follow the translation and back-translation process to ensure the translation quality. Table I shows the statistics of our lexicon for happiness computation.

TABLE I. Statistics of the lexicons in PERMA.

Polarity	P	E	R	M	A	Sum
+	140	61	139	90	163	593
-	374	87	237	46	71	715
+&-	514	148	376	136	234	1408

Although we adopt the strict translation process, the lexicons are designed for the happiness computation in English environment. There still exists a gap between the lexicons and our microblogs happiness computation scene. To adjust the lexicons according to Chinese language habits and Sina Weibo's characteristics, we propose a series of word expansion methods to expand our lexicons.

Since the dataset is quite large, we only consider the words with term frequency greater than 100. Then we get a candidate vocabulary which contains 374312 words.

1) PMI-based Word Expansion: PMI (pointwise mutual information) is a classic measurement of word association.We compute all the PMI value between our PERMA lexicons' words and the vobabulary which we obtained from the dataset.

2) Distributional Similarity based Word Expansion: We use the Word2Vec[2] tools to learn the word representation of each word.We compute the vectors of every word in the vocabulary in the whole dataset. Each word is represented by 50-dimensions continuous real value vector.

We get the each lexicon word's top 100 PMI candidate similar words and top 100 Distributional Similarity candidate similar words. We identify the overlapping words in these two groups. We also consider the candidate words' TF, D-F(document frequency) and part-of-speech to filter out the noisy words. Finally, we expand the PERMA lexicon to a larger lexicon which we call it expand-PERMA. The statistics of the expand-PERMA is shown in Table II.

As in Section III-A, we compute the v_{ij} based on our lexicon. We deem that big lexicon which contains more words has more term occurrences than the small lexicon. So we set the v_{ij} as the inversely related to the lexicon size. The v_{ij} is formalized as follows:

$$v_{ij} = v_{t_{ij}} = \begin{cases} \frac{(-1)^{I(P(L)is-)}}{|L|} & \text{if } t_{ij} \in L \\ 0 & \text{else} \end{cases} \quad (2)$$

[2]https://code.google.com/p/word2vec/

TABLE II. Statistics of the lexicons in Expand-PERMA.

Polarity	P	E	R	M	A	Sum
+	328	135	297	170	288	1218
-	691	259	387	116	141	1594
+&-	1019	394	684	286	429	2812

TABLE III. STATISTICS OF THE ANNOTATION MICROBLOGS DATASET.

Happiness	Unhappiness	Objective	Sum	Kappa Coefficient
2047	4304	3649	10000	0.720

Here $L \in \{P+, P-, E+, E-, R+, R-, M+, M-, A+, A-\}$, means that L represents any Lexicon in PERMA or Expand-PERMA. $|L|$ is the size of Lexicon L. And $P(L)$ means the polarity of Lexicon(See Table I). If L is $P-$, $P(L)$ results $-$. $I(x)$ is the indicator function.So $(-1)^{I(P(L)is-)}$ leads to the result of -1 if L's polarity is $-$, otherwise it leads to the result of 1.

C. Impact Factors on Lexicon Word

Take the following sentence as an example, "I'm not happy". Here 'happy' can be regarded as a happiness word. But we will get a completely wrong result if we don't consider the impact of the word 'not'. Another example can be given, "I'm very happy". Here 'very' can strengthen the happiness. So we must consider the word impact on other words. Here we take the Previous Word factors into account. For simplicity, we consider the last word's impact on current word, such as privative or adverb of degree. We get a negative word lexicon L_N and a degree word lexicon L_D. Each degree word has a degree value $d(t)$ varying from -0.5 to 0.75. The positive degree value means that the degree word strengthen the next word t's weight. And the negative degree value means that the degree word recede the next word t's weight. Our negative word lexicon and degree word lexicon have 35 and 143 words respectively. As shown in the framework, the impact weight $\overrightarrow{w_i}$ can be computed as :

$$w_{ij} = w_{t_{ij}} = \begin{cases} 1 & \text{if } j = 1 \\ 1 + d(t_{ij-1}) & \text{else if } t_{ij-1} \in L_D \\ -1 & \text{else if } t_{ij-1} \in L_N \\ 1 & \text{else} \end{cases} \quad (3)$$

Note that the first word's w_{i1} in a sentence is set as 1.

IV. EXPERIMENT ON MANUAL ANNOTATION DATASET

A. Experiment Setup

In order to verify our method's validity, we evaluated our method by using an annotation dataset. Since there was no existing data set to use, we had to construct our own dataset to evaluate our method. We randomly sampled 10000 microblogs from the whole dataset. The microblogs was original and consisted of merely text content. We asked 7 microblogging users to annotate the microblogs. Each microblog was annotated by two users at least. If two users annotated different labels at one microblog, we asked another user to label it and adopted the majority answer. The label was designed for three types: -1,0,1. Label -1 represented the users felt the microblog was expressing unhappiness. Label 0 represented the microblog was irrelevant to happiness or was an objective microblog. Label 1 represented the users felt the microblog was expressing happiness. We calculated the Kappa Coefficient to measure our annotation consistency. The Kappa Coefficient was 0.720 which showed that our annotation process had high consistency. The statistics of the annotation dataset was shown in Table III.

TABLE IV. THE EXPERIMENTS RESULTS ON ANNOTATION DATA.

method	Baseline	expand-P+PWF
AP	0.335	**0.339**
Bpref	0.530	**0.543**

B. Experiment Result

1) Evaluation Methods: Our method is an unsupervised method and the happiness value also makes sense when used for intercomparison. We compute all the microblogs' happiness value and then rank the microblogs by happiness value. By ranking a set of microblogs, if microblogs with happiness are in the highest ranks compared to other microblogs, it will prove that our ranking method is effective. We adopt the binary preference measure (Bpref)[9] and the Average Precision to evaluate our method.

2) Baselines: We use the PERMA lexicon and take no account of Previous Word factors (P+noPWF) as our baseline. Under a simplified situation, we consider that all words are mutually independent, that means all the words are not affected by each other. So the impact weight vector $\overrightarrow{w_i}$ can be set as $< 1, 1, \cdots, 1 >$. Our method is expand-P+PWF (expand-PERMA+Previous Word factors).

3) Result: Table IV shows the results based on the manual annotation data. From the results our lexicon expand methods achieved significant improvement both in Bpref and AP. Relatively, the Previous Word factors improves little because the microblogs which contains negative situation or degree words are still a few cases. Please note that the AP values are generally low, which is because that our ranking list is much longer than in the usual situation. So even a small increase in the value of AP has significant meaning.

V. STATISTICS AND ANALYSIS ON LARGE-SCALE DATASET

We have verified our method on the manual annotation dataset. The results prove our method is a effective measure for happiness. In order to further analyse the Chinese people's happiness, we perform an in-depth analysis on the large-scale dataset based on our method. Our dataset contains 1,486,777 users and 1,943,470,481 microblogs, which includes all microblogs and user profiles updated before November 28, 2011.

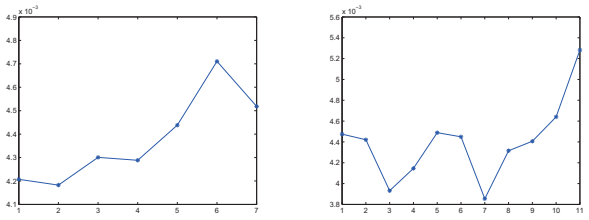
A. Time-based Statistics and Analysis

We split the microblogging data into 3 kinds of classifications: by dates (from 2011-01-01 to 2011-11-28), by months (from 1 to 11) and by days (from 1 to 7). Table V lists ten dates in 2011 year. Four of top five in the left tab are eastern and western festivals. It's easy to understand: people always send blessings to each other on festivals. Note that "11-11" ranks top 2 among all dates. Similar to Christmas sale in Western countries, Taobao[3](China's largest e-business platform) held the largest online shopping festival on that day. In contrast, the right table lists dates when people feel most sad. There are two distinctive social events. WenZhou highway accident happened at 9 p.m. on July 23rd. The accident drew most users' attention and people expressed their anger to the government and sorrow for the victims. Japan's 9.0 magnitude earthquake took place

[3]http://www.taobao.com/

TABLE V. PARTIAL DATES' RESULT IN AVERAGE HAPPINESS VALUE IN 2011.

Most Happiness Date	$h_{ave} \times 10^{-3}$	Remark	Most Unhappiness Date	$h_{ave} \times 10^{-3}$	Remark
11-24	6.849	Thanksgiving Day	07-25	0.989	7.23 highway accident
11-11	6.804	Single's Day	07-24	1.772	7.23 highway accident
05-08	6.687	Mother's Day	07-26	2.148	7.23 highway accident
01-01	6.552	New Year	07-27	2.317	7.23 highway accident
09-12	6.513	Mid-autumn festival	03-11	2.504	Japan's 3.11 earthquake



(a) Days' happiness. (b) Months' happiness.

Fig. 1. The happiness value varies with days and months respectively.

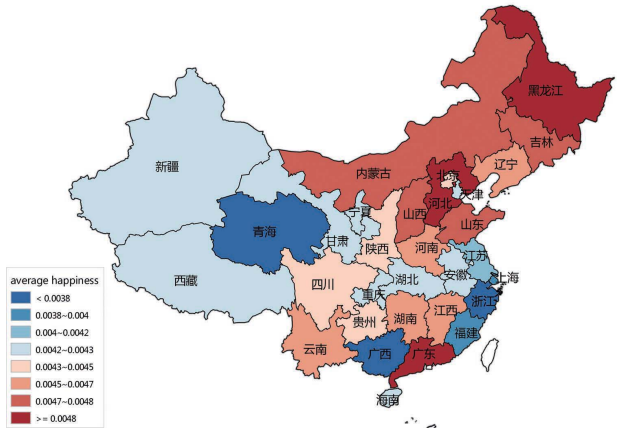


Fig. 2. The Chinese happiness map. Each province is colored according to their average happiness value. Warm color indicates a larger happiness value while cold color indicates relatively low happiness value.

on March 11th. From the result we can conclude that people feel happy on festivals and sad about tragic events.

Fig.1(a) shows the happiness variation with days. In line with our intuition, people feel happy during weekends. Tuesday is the most unhappy day. Fig.1(b) shows happiness variation with months. March and July are the most unhappy months in all months. This phenomenon is due to the two big tragic events inferred from day's classification analysis.

B. Location-based Statistics and Analysis

The Fig.2 shows the Chinese happiness map. From the Chinese happiness map, the regions in the east have higher happiness value relative to those in the west, and the northeast part of China is of particularly high happiness value.

VI. CONCLUSION

In this paper, we propose a method to calculate happiness based on PERMA theory. We first translate the English PERMA lexicon and then we combine a series of word expansion method to expand the lexicon based on the large dataset on Sina Weibo. Experiments on manual annotation dataset show that our method can measure the happiness of microblogs quite well than baseline. Then we perform a comprehensive statistical analysis on about 2.0 billion microblogs from 1.5 million users on Chinese Sina Weibo. The results prove that happiness is closely related to consumption and economic development. We consider the following work as the future research directions: (1) Establish our measurement of happiness by involving other characteristics of Sina Weibo. For example, take the repost and comment behavior into account. (2) Consider both the geographic information and the users' profiles. Happiness analysis on migrants and permanent residents is an interesting research topic.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No.61202140 and No.61170196), and the Major Program of the National Social Science Foundation of China (Grant No.13&ZD190).

REFERENCES

[1] M. E. Seligman, *Flourish: A visionary new understanding of happiness and well-being*. Simon and Schuster, 2012.

[2] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," pp. 56–65, 2007.

[3] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[4] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PloS one*, vol. 6, no. 12, p. e26752, 2011.

[5] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth, "The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place," *PloS one*, vol. 8, no. 5, p. e64417, 2013.

[6] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, "Happiness and the patterns of life: A study of geolocated tweets," *arXiv preprint arXiv:1304.1296*, 2013.

[7] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. P. Seligman, and L. H. Ungar, "Characterizing geographic variation in well-being using tweets." in *ICWSM*. The AAAI Press, 2013.

[8] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of liwc2007," *Austin, TX, LIWC. Net*, 2007.

[9] C. Buckley and E. M. Voorhees, "Retrieval evaluation with incomplete information," in *SIGIR*. ACM, 2004, pp. 25–32.