the stock market in China, the dynamic analysis on relationship between investor sentiment and stock market is proposed using Thermal Optimal Path (TOP) method. The structure of this paper is presented as follows: Section 1 is the introduction including literature review that studied the relationship between investor sentiment and stock market. Then construction of the investor sentiment indicators is introduced in Section 2. In Section 3, using TOP method, the dynamic lead-lag analysis is proposed between sentiment indicator and stock index returns such as the whole market and some classical industries. Section 4 is the conclusion.

2. Construction of Investor Sentiment Indicator

2.1 Original data

The original data of the investor sentiment is from the social networking site of China stock market called Xueqiu which means snowball in English. The sense of snowball is that collecting the thinking of individual investors. Website of Xueqiu is a representative vertical type financial community in China that set up in 2011. It provide comprehensive financial service such as real-time quotes, news, investment strategy, transaction service and so on. The users of Xiuqiu are individual investors who have a certain professional knowledge of stock market. So that, the comments from these users can better reflect the sentiment of individual investors. Further, the users' comments data is contains in the homepage of each stock, so it is easy to do subject analysis as the website structure is clear.

2.2 Procedure of investor sentiment data

In order to obtain the investor sentiment data based on the Chinese social networking site, five steps will be proposed.

Step 1. Web Crawling: The web crawler is developed based on Python and using Pycharm as integrated development environment. The data is collected by calling the browser kernel to send requests through Webdriver.

Step 2. Text Cleaning: The crawled data from Xueqiu is fragment of HTML including a large number of HTML labels. In order to facilitate subsequent processing and get relatively clean data, the labels are all removed using some special script commands.

Step 3. Topic Extraction: The topic here is the name of each stock. The stock name in Xueqiu website has some fixed rules such as "$stock name$" or "$stock name (stock code)$". Through these rules, the information of stock name can be extracted from the text. So we can identify certain stock for each text.

Step 4. Semantic Analysis: Semantic analysis is the most important step to obtain investor sentiment data. In this step, the Chinese word segmentation is first be done to get useful words of each comment text. Then the sentiment classification model is constructed based on a Chinese sentiment corpus from Renmin University in China. The classification model used here is Logistic model which is a commonly used machine learning methods to estimate the possibility of a certain thing. Through this step, each user comment has a certain sentiment, positive, negative or neutral.

Step 5. Construction of Sentiment Index： As each user comment has one or more topics, that is

to say, one comments is corresponding to one or more stocks. And also it has a certain sentiment. So the positive sentiment index of one stock in a certain day is the number of the positive comments that referred to this stock in that day. The construction of the negative and neutral sentiment index is the same. Meanwhile, we can also obtain the attention index which is the frequency of occurrence of each stock in a certain day. As each stock is belong to a certain industry, the industry sentiment index can be calculated by adding up all the sentiment index of the stock contained in this industry. In this way, the sentiment index of the whole market also can be obtained.

3. Dynamic Lead-lag Analysis based on TOP Method

3.1 Thermal Optimal Path (TOP) method

The thermal optimal path (TOP) method was proposed to identify and quantify the time-varying lead-lag structure between two time series by Sornette and Zhou in 2005 [17]. With the globalization of financial market and the prevalence of quantitative trading, more and more features of complex system are emerged in stock market, such as nonlinear, dynamic, self-organization and so on. Traditional linear econometric models are not suitable to study the complex system. The past literatures which studied the relationship between investor sentiment and stock market often obtained very different conclusions due to time varying characteristics of the relationship in complex system. Therefore, it is important to employ some time varying model to study this issue. The TOP method has been already widely and successfully applied to several economic and financial cases [18-22]. The method is briefly introduced as follows.

Consider two standardized time series X and Y. The matrix $E_{X,Y}$ of distance between them is defined as $\varepsilon(t_1,t_2)=|X(t_1)-Y(t_2)|$. The element $\varepsilon(t_1,t_2)$ of the matrix $E_{X,Y}$ thus compares the realization $X(t_1)$ of X at time t_1 with the realization $Y(t_2)$ of Y at time t_2. The value $|X(t_1)-Y(t_2)|^2$ defines the distance between the realizations of the first time series at time t_1 and the second time series at time t_2. The $\mathrm{N}\times\mathrm{N}$ matrix $E_{X,Y}$ thus embodies all possible point-wise pairwise comparisons between the two time series. Once the matrix $E_{X,Y}$ is obtained, an optimal path is determined that quantifies the lead-lag dependence between the two time series. It is convenient to use the rotated coordinate system (x,t) such that

$$\begin{cases} t_1 = 1+(t-x)/2 \\ t_2 = 1+(t+x)/2 \end{cases}$$

where t is in the main diagonal direction of the (t_2,t_1) system and x is perpendicular to t. Then, $x=t_2-t_1$ and paths which have $x(t)\neq 0$ define varying lead-lag patterns. A positive x corresponds to $t_2>t_1$, which by definition of the optimal thermal path below means that the second time series $Y(t_2)$ lags behind the first time series $X(t_1)$, or equivalently $X(t_1)$ leads