EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Relationship discovery in public opinion and actual behavior for social media stock data space

Yunlan Xue[1], Lingyu Xu[2*], Bingcheng Qiu[1], Lei Wang[2] and Gaowei Zhang[2]

Abstract

With the rise of social data media, the cyber world nearly parallels to the real world. The trajectory of a hot event is reflected in social media by Public Opinion Data Space (OS) and Actual Behavior Data Space (BS). However, the relationships with a variety of mechanisms in each space or between them are often unknown. To solve the above issues, the traditional methods for inferring relationship are by performing a statistical similarity analysis of time sequence from dynamic elements. In specially, the research of clustering nonlinear correlation data object is rare, so we propose Matrix Similarity Clustering Algorithm (MSCA) based on random matrix theory and combined with sliding window technology to cluster the similarity of multidimensional time sequences. This method is effective to detect the trend relationship of time sequences with multiple dynamic elements. In addition, we construct a knowledge map to analyse the relationships in OS and BS.

Keywords: Big data, Relationship, Sliding window, Matrix clustering, Similarity

1 Introduction

With social media showing increase popularity such as Twitter, Micro Blog, the data complex and value sparse features have caused much attention in Internet big data mining research. Cyberspace almost parallels with the real world with the increasing socialization interaction between online and offline. Thousands of Internet users compose a network system, who interact each other with a variety of mechanisms. A system of Public Opinion Data Space and Actual Behavior Data Space (OBS) is a comprehensive computing, cyber, and physical environment for the multidimensional complex systems. It is a network physical system with a controllable, credible, and scalable function on the basis of environmental perception, depth fusion calculation, communication, and control ability. It realizes depth fusion while it increases or extends new function with real-time interaction by the feedback of calculation process and physical process. It controls a physical system in the form of safe, reliable, efficient and real-time.

Recent years, due to potential benefits to society, economy, and the environment, more and more scholars take attention on the research related to BS. BS as the next generation of engineered systems require tight integration of computing, communication, and control technologies in many application domains [1]. Sometimes, how to know the relationship among elements in the system is impossible due to technical limitations or the nature of the system itself. It is important to address the problem of how to optimally infer the relationship of system from the observable elements [2, 3]. A happening event causes changes in Public Opinion Data Space (OS) and in Actual Behavior Data Space (BS) [4]. For example, clicks, posts, replies, etc. of the event on some forums change in OS. And deal, prices, and stock trade related to the event change in BS also. The event is mapped into OS and BS, and we use the data mapping of OS and BS to analyze the event.

In this paper, the main elements are the form of multifactor time sequences in OS and BS. Our main job is to find the relationships between the main member and others of event in OS or BS and the relationship between OS and BS. Because our study object is multifactor time sequences and there are no simple linear relationships, we

* Correspondence: xly@shu.edu.cn
[2]School of Computer Engineering and Science, Shanghai University, Shanghai, China
Full list of author information is available at the end of the article

propose Matrix Similarity Clustering Algorithm (MSCA) based on random matrix theory combined with sliding window technology to cluster the similarity of multidimensional time sequence.

The rest of the paper is organized as follows. Section 2 gives the related work in the literature. We introduce preliminary concepts in Section 3. Section 4 is about the relationship discovery in Public Opinion and Actual Behavior of social media stock data space description. Extensive experimental results are presented in Section 5. We conclude the paper in Section 6.

2 Related work

In recent years, the relationship of a system from the observable dynamics elements has raised many scholars' interests, especially in finance field. Finance presents a variety of collective behavior [5, 6]. Scholars focus on statistical cross correlation between individual stocks that not only reveal the complex structure of finance system, but also have important practical values for risk control to asset allocation and portfolio assessment [7, 8]. The probability of price returns generally obey power-law distribution, and the properties in different markets are stable [9].

It is an important challenging task to explore the spatial structure of finance system. For example, stocks hierarchy of financial market is constructed by the minimum spanning tree and its derivative method [10]. In particular, plate structure and topology association are detected effectively by random matrix theory in finance market [11, 12]. Securities market has a large number of real observable data, but the relationship of stocks lack precise form of internal interaction. For solving this drawback, researchers use random matrix theory to study cross correlation properties of stock. Someone obtain special forms of non-random internal interaction between the securities market stocks [13, 14] through the correlation matrix contrast to the nature of random correlation matrix.

The industry sector structure of mature market has been systematically studied, such as New York stock exchange and South Korea stock exchange [15]. Recently, the random matrix theory is used to explore eigenmode of industrial production index fluctuation in dissipative structure [16]. In particular, Shen and Zheng study the interaction of stock in Chinese securities market and associated structures by random matrix theory. Ren and Zhou research the cross correlation dynamics properties of Chinese securities market [17].

In this paper, we study the event role with multiple attributes, and most of them have no direct linear relationship. We present a random matrix to express these multiple attributes with time sequence. And then, we propose the MSCA method to evaluate similarity of the attribute time sequence with a sliding window technique which greatly improves the accuracy.

3 Preliminary concepts

3.1 Event definition in social media stock data space

An event is a snapshot of perceived experience at one moment in time which can be defined as a collection or a tuple of attributes.

The so-called event occurs in a particular time or place by one or more roles, which is composed of one or more actions; it means an action or state change. Event is the unit of people understanding and experiencing about the world, and it meets people's normal cognition rule. Event e is defined as Formula 1 in this paper.

$$e = \langle A, M, T, S \rangle \tag{1}$$

where A is element of the action, and it means the change process and its characteristics of event, which is the degree of movement, description of the way, method, and so on. $A = \langle a_1, a_2, \dots, a_k \rangle$. Each attribute a_i is defined as a tuple such that $a_i = \langle a_1^i, a_2^i, \dots, a_l^i \rangle$ and a_j^i is normalized real value $a_j^i \in [0, 1]$. M is the object element that includes all roles participating in the event. T is the time element, and it means the time from the beginning to the end of the event. S is the environmental element, and it means the place and its environmental characteristics of event.

However, information presents polymorphism and complexity under the status of the rapid development of Internet big data. When an event happens, in different S, the form of A is presented diversely. For example, RMB join in the SDR event, the attention, participation, and interaction of a user present different forms of A. Someone discuss the event in forums (OS), while others change their investment strategies (BS), such as buy or sell related product or stock which are affected by the event. It can be seen that only one perspective cannot accurately describe an event. Therefore, we study event from a new perspective, that is, from OS and BS. In this paper, the event is mapped into OS and BS, and we can use the mapping data of the event from OS and BS to analyze trajectory of the event.

3.2 The definition of Public Opinion Data Space (OS) and Actual Behavior Data Space (BS)

3.2.1 Public Opinion Data Space (OS) in social media stock data space

With the rise of Internet media, people are used to comment an event on Internet carriers, while Internet carriers record information of people's lives, works, and studies. The current popular network media are microblog, Twitter, Facebook, forums, etc., which have thousands of users. When a hot event happens, a lot of views and comments have diffusion fast on network to form a powerful network influence. The research and analysis of network events emerge in endlessly [18, 19]. In particularly, investors publish their views and emotions when a good or a bad event happens in the finance field. The form of A in Formula (1) is diversified in OS, such as click, reply, content and size of post, etc. in forums. Time and A are not limited in OS. For example, stock forum users can post, click, reply, etc. at any time, and the number of post, click, and reply are not limited. Therefore, we build ⟨click, post, reply⟩ as a set of A to measure the action of event.

Firstly, click reflects the degree of user attention on an event. On the other hand, click rate can measure how many roles are participating in the event. The role of an event by click is a light degree, only showing the instinct of roles.

Secondly, the number of post on the Internet also reflects user attention on an event, but the strength of a post is larger than the strength of a click to reflect the participation of an event. Posting for an event with their comments and analysis fully illustrate the subjective initiative of a user, which reflects user's hope to enhance or inhibit the spread of event influence through their own comments.

Again, the strength of a reply to an event is the largest than a post and a click on the network, which embodies the role of reply which is given to the interaction of an event.

$$A = \langle \text{click},\ \text{post},\ \text{reply} \rangle \tag{2}$$

where *click* is when a user takes attention to an event by a click, *post* is the number of user posting the comments on forums and *reply* is the number of user's reply to some post.

3.2.2 Actual Behavior Data Space (BS) in social media stock data space

Nowadays, the life of people is alternative in online and offline, which have been officially formed as two different living environments for people. People tend to leave their trajectories online through video, audio, picture, text, etc. However, in real life or when offline, it is difficult to capture the information with its mass scale and few value features. Especially, capturing the group behavior after the event happens is more difficult, so the data can be used extremely limitedly. On the other hand, some who carries such as Stock Exchange centre and electric business platform can record human behaviors in the real world. For instance, on November 11, 2015, the most large-scale commercial activities happened on Chinese Internet again: Taobao website sales total to 91.2 billion yuan. The Taobao can cope with such a great event because they comprehensively analyze A of M that contain past trade volume and super-high concurrency requirements, especially consumption, search, and browsing habits of users [20]. Therefore, some special carrier is a good choice to observe human behavior in the real world.

Due to mapping relationship of the special carriers and the real behavior, we can analyze real behavior in reality through the mapping data from carriers. For example, trajectories of stock trading volume, price, and change rate of price would be recorded by Stock Exchange. In particular, if an event happens in the finance field, investors will buy or sell their stocks or futures and other investment products by the good or bad event. The form of A is multitudinous in BS, such as volume, turnover, change, price, etc. The domains of time and A have their specific limitation. The trade time in stock exchange is during 9:30–11:30 and 13:00–15:00 on work days in China, while the maximum change of rise and drop is 10 % in stock market, and trade stock needs an interval of 1 day in China, that is, $T + 1$ model different from other countries. The investors react to event mainly through buying and selling stocks in the real world. Therefore, we select the original action of stock affected by an event as A.

Firstly, A can reflect the degree of participation or the group game mainly through %Turnover, Vol, Turnover. One day, the volume of a stock blew up that illustrates buying and selling are more active, and group M makes a clear judgment for a potential stock trend. A illustrates an event that occurs or will occur soon, so quantitative analyzing the participation degree of the real social M is an important indicator to study the event.

Secondly, with occurrence, development, decay, and death of an event, the price of one stock will change in the event's life cycle. If it is a good event, the price will be gradually pushed up from the beginning to the end of the event. The price trend maps out the development trend of the events, such as Close, High, Low, Open, and Pre Close.

Lastly, with the development of an event, the change of stock price reflects the heat and expectation of M, which are others important indicators to measure and analyze an event, such as *Change* and %*Chg*.

$$\begin{aligned} A = \langle &\text{Close},\ \text{High},\ \text{Low},\ \text{Open},\ \text{Pre Close},\ \text{Change}, \\ &\%\text{Chg},\ \%\text{Turnover},\ \text{Vol},\ \text{Turnover},\ \text{Market Cap},\ \text{CSV} \rangle \end{aligned} \tag{3}$$

The A of stock contains Close, High, Low, Open, Pre Close, Change,%Chg, %Turnover, Vol, Turnover, Market Cap, CSV in BS. Close is the closing price in the end of the trading time. High is the highest price during the trading time. Low is the lowest price during the trading time. Open is the opening price in the beginning of the trading time. Pre Close is the yesterday's Close. Change is the amount money of rise or fall. %Chg equals to

Change divided by *Pre Close*, which is the price change rate. %Turnover equals to percentage of the result of Vol divided by the total number of circulation shares. Vol is the amount of trade shares. Turnover equals to Vol multiply by traded shares, which is the amount money of trade shares. Market Cap equals to Close multiply by the total shares, which is the total market price. CSV equals to Close multiply by the circulation shares, which is the circulation price. In this paper, all A of stock is included to analyze the real behaviors of an event.

3.2.3 Analysis of OS and BS in social media stock data space

In the paper, we describe an event by both OS and BS comprehensively. And, there are some common features and characteristics of OS and BS, which are as follows.

1. The common features of OS and BS:
   (a) Being recordable/storable: investor sentiment has both professional characteristics of financial information and semantic characteristics of short-text information. The comments or real trading behaviors on the Internet of investors once formed will leave trajectories on the carriers of OS and BS, and they can be recorded and stored.
   (b) Spontaneity: A of investors from any one space is subjective evaluation and judgment of stock financial market and company's information. And such behavior is spontaneously generated, spread, and accepted by investors that enable us to get more pure investor sentiment without other information interference.
   (c) Interaction: the network evaluation information of OS includes both objective description about all kinds of information and investor's subjective judgment. The interaction processing is more on orientation, that is, the network evaluation information interacted with the attention of investors, click, and reply. This interaction brings unprecedented influence on investor group behavior due to the interactivity of investor's attention. Similarly, A of BS not only has subjective judgment of investors but also change the trading behaviors of interaction by Turnover, %Chg, prices, etc.
   (d) Being representative: information view of OS is generally trusted, because information is held by a publisher himself or organization, so information represents the specific emotions tend view of different investors. Although investor sentiment revealed in OS is divergent, investor focus is concentrated. The focus contains a recent performance of the listed companies and the future development and other kinds of information; thus, we are able to extract the performance characteristics of the investor sentiment under OS.
2. The characteristic properties of OS and BS:
   Because the comments on the Internet are subjective, it contains a large number of loosely and redundant information. However, all data of BS are real data, and there is no false information.
   The comments of OS are only from user's ideology or psychological changes. However, they cannot real reflect trading behavior. Therefore, we make up the incompleteness of OS through BS's trading behaviors.

3.3 The description of data in social media stock data space

Because the data of our research is multidimensional time sequence in OS and BS, we give the relative definition and description of time sequence.

3.3.1 The definition of multidimensional time sequence

Intuitively speaking, time sequence is a set of numerical value collection according to chronological order. Figure 1 is a typical time sequence in social media stock data space. Figure 1 is the Shanghai Composite Index from 1991 to 2015, which has the feature of multidimensional time sequence.

Time sequence: a time sequence $T = t_1, t_2, \ldots, t_n$ is an order set of values, $1, 2, \ldots, n$ is the time node with the length of n. Time Subsequence: given a time sequence T with the length of n, subsequence C is a sample of consecutive T with the length of m, $C = t_k, t_{k+1}, \ldots, t_{k+m+1}$, where k is the beginning of the sample, which satisfies the condition of $1 < k < k - m + 1$. Multidimensional time sequence: time sequence can be divided into one-dimensional time sequence and multidimensional time sequence according to the numbers of independent variable of the time sequence. One-dimensional time sequence is only one variable changing over time in numerical sequence; and multidimensional time sequence is multiple variables changing over time in multidimensional numerical sequence.

A multidimensional time sequence with a size of V_{ji} is a collection of time sequences that is limit by d variable numbers and n length. Multidimensional time sequence can be expressed as Fig. 2: where V_{ji} is the value of j on the time of t_i. $d = 1$, that is, one-dimensional time sequence. $d \geq 2$, that is, multidimensional time sequence. In real application, multivariate time sequence has become more and more important.

3.4 Data feature in social media stock data space

(1) The data source of OS
    The occurrence environment of an event is OS. And the source of OS is very broad, as long as

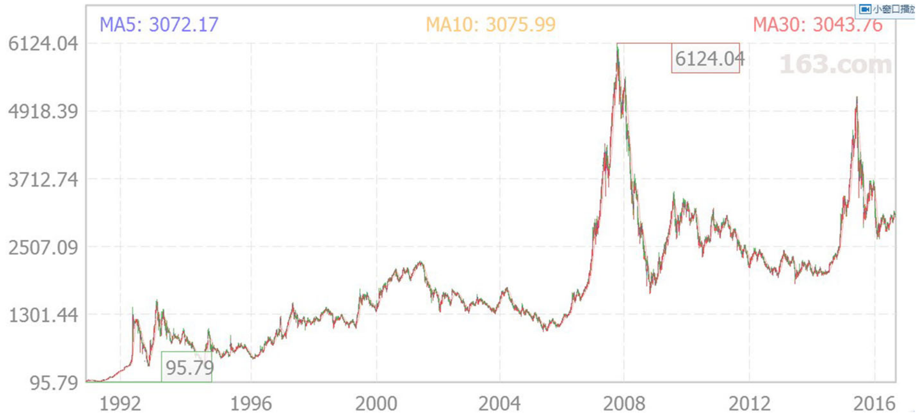


Fig. 1 A time sequence of Shanghai Composite Index

investor comments on Internet carriers which all belong to the category of OS. Due to Chinese investors mostly focus on Eastmoney and Sina, two big authoritative websites, we crawl the post as a sample from Eastmoney and Sina forums about stock 600519 during January 2013 and May 2014. The samples of two websites have a strong correlation according to the result of analysis and comparison. The Pearson correlation coefficient is 0.877. We select the data of Eastmoney forum as a mapping data of OS to analyze event feature. The sample data contrast figure of Eastmoney and Sina as Fig. 3.

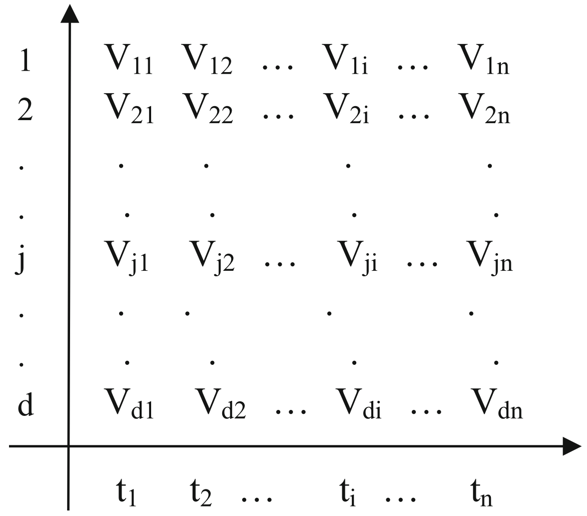
(2) The data source of BS:
Unlike OS, the data of BS has its uniqueness, and data is the same no matter from which trading software. We adopt stock trading data from the 163 website as data source in BS. They are mainly from the Shanghai Stock Exchange and Shenzhen Stock Exchange. The original transaction data of each stock include 12 dimensions, such as Close, High, Low, Open, Pre Close, Change, %Chg, %Turnover, Vol, Turnover, Market Cap, and CSV. Because one dimension variable is calculated by several other dimension variables, there are linear relationships between them. In order to reduce the complexity of BS to improve efficiency, principal component analysis (PCA) can be used in dimension reduction in this paper.

(3) Data set feature normalization
Data standardization (normalization) processing is a basic work of data mining. Different evaluation indices have different dimensions and dimensional units. And, it will affect the result of data analysis. In order to eliminate the dimension influence between indicators, we make data normalization processing to solve the comparability among data index. The index of original data in the same order of magnitude after dealing with data normalization is suitable for a comprehensive comparative evaluation.

For different data sets, the number of A and data unit of A are different. It means that we are unable to compare the data set vectors of different data sets directly. In order to address this problem, a common statistical method, namely, the min-max normalization of data set, is extended and used to approximate original vectors. It is a linear transformation of original data, the result value mapping between [0, 1].

3.4.1 Feature extraction of data in social media stock data space

The structure of time sequence is often complex and has multiple dimensions. And there are a lot of noises. If we mine data directly on time sequence, not only high efficiency but also high accuracy and reliability of the mining results cannot be obtained. In addition, the analysis of the follow-up work will be affected. How to reduce the dimension of time sequence, while reserving the main information, and correctly measure the similarity of two sequences is the key to improve efficiency. Also, it is the basis of



$$\begin{array}{c|ccccccc}
1 & V_{11} & V_{12} & \dots & V_{1i} & \dots & V_{1n} \\
2 & V_{21} & V_{22} & \dots & V_{2i} & \dots & V_{2n} \\
. & . & . & & . & & . \\
. & . & . & & . & & . \\
j & V_{j1} & V_{j2} & \dots & V_{ji} & \dots & V_{jn} \\
. & . & . & & . & & . \\
. & . & . & & . & & . \\
d & V_{d1} & V_{d2} & \dots & V_{di} & \dots & V_{dn} \\
\hline
 & t_1 & t_2 & \dots & t_i & \dots & t_n
\end{array}$$

Fig. 2 Multidimensional time sequence in social media stock data space

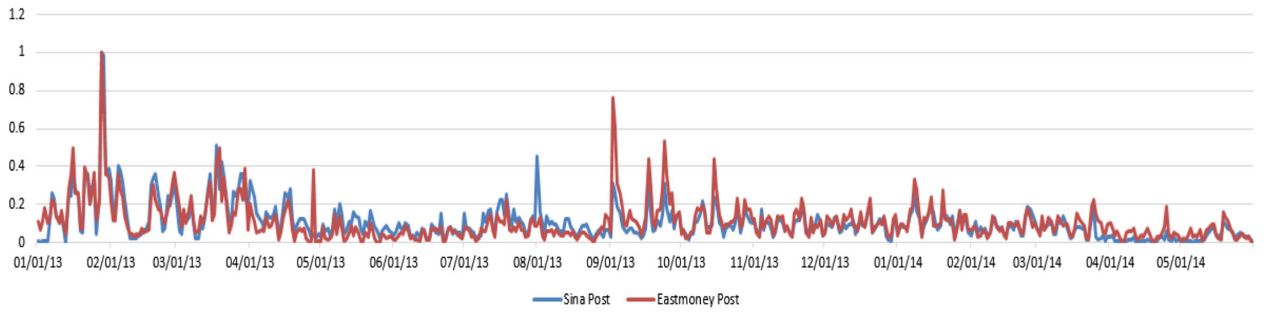


Fig. 3 Stock data of Eastmoney and Sina in Public Opinion Data Space (OS)

classification and clustering, correlation analysis, prediction, anomaly detection, and cycle mode discovery in further. Therefore, we need to express again the original time sequence from a higher level before an experiment, which is abstraction and generalization.

Faloutsosetal (1994) [21] laid the foundation for data mining research of time sequence, who pointed out how to improve the efficiency of distance measure (similarity search) by dimension reduction techniques. The original data space is known as the real space, and dimension reduction space is known as the search space in the method of high-dimensional data space which is mapped to a low-dimensional space. A new distance measure is defined in search space, and the low-dimensional data is being used to improve the efficiency of calculation for data mining tasks after dimension reduction. The dimension reduction technique is PCA in this paper.

The goal of PCA (Principal Component Analysis) (JolliffFe, 2005) [22] is to reduce variables by linear transformation. And, there are a lot of applications in feature selection. Based on PCA, Lu et al. puts forward principal feature analysis (PFA) [23] and Yoon et la. put forward the characteristics scoring method of the multivariate time sequence on the basis of the correlation between information according to the common principal component analysis (CPCA) [24].

The eigenvectors with the larger eigenvalues are selected to reconstruct a new data set to achieve dimension reduction according to the PCA.

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{n-1} \tag{4}$$

$$[\text{eigenvectors}, \text{eigenvalues}] = \text{eig}(\text{cov}) \tag{5}$$

where X, Y are two different vectors and cov(X, Y) is a covariance of X, Y. *Cov* is a covariance matrix of ($x_1, x_2, \ldots, x_i, \ldots, x_n$), and eigenvectors are the feature vectors of Cov. eigenvalues is the value of eigenvectors. Bigger eigenvalues mean greater contribution of original data. Therefore, top n eigenvectors will be selected as new data of BS.

4 Relationship discovery in public opinion and actual behavior of social media stock data space

Because the time sequence is an observable phenomenon of the real system, similar time sequence often presents a variety of deformation, such as amplitude deformation, the overall shape linear drift, discontinuous points, containing noise, and timeline deformation. So similarity measure time sequence is not completely tight; so the research of similarity is a very basic and difficult problem [25–27]. Similarity metric methods in time sequence data mining mainly adopt the Euclidean distance [28], dynamic time warping distance [29, 30], irrelevant distance with complexity, etc.

4.1 The description of Matrix Similarity Clustering Algorithm (MSCA)

MSCA of this paper is implemented using the sliding window technology, which set a fixed window and slide on the time sequence. The size of the sliding window is a key factor, we determine the window size as 3 according to the experience data, and experiments show that it can lead to a better experimental effect.

1) Data sequence: DS is the flow of data sequence $\langle A_1, A_2, \ldots, A_m, \ldots \rangle$, $A_m = \{D_1, D_2, \ldots, D_i, \ldots\}$, and $A_m (m = 1, 2, \ldots)$ is the action attribute set. DS data flow is to be sectioned, and each section corresponds to a data flow subsequence and a certain number of attributes, which a data block is a basic window as W.
2) Sliding window: a sliding window SW corresponds to a continuous sequence of basic window $\langle W_1, W_2, \ldots, W_k \rangle$, which contains the number of basic window with a fixed value k, as Fig. 4. With new data coming, the sliding window update with basic

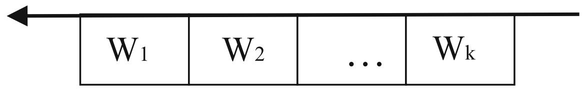


Fig. 4 Structure of a sliding window

window. The oldest basic window is deleted when data comes into a new basic window, while the sliding window will update. Therefore, the data of sliding window change and update.

Random matrix theory is introduced to the study of finance field in recent years, and it provides a good lesson for quantitatively study of stock interactions. According to random matrix theory, if there are n random time sequences with the length of l and they are not related, then random matrix P can be built by the time sequence.

Because this paper studies the similarity problem of multidimensional time sequence, we need to calculate matrix similarity as Formula 6.

$$\mathrm{Sim}_{P,Q} = \frac{t_{\mathrm{r}}(PQ)}{\sqrt{|P^T P||Q^T Q|}}, \qquad t\mathrm{r} = \sum_{i=1}^{n} \lambda_i, P = |A_i, A_{i+1}, \ldots, A_{i+n}| \tag{6}$$

where P and Q are the vector matrix of A. When $|PQ - \lambda E| = 0$, we obtain eigenvalue λ_i of matrix PQ. t_r is the sum of λ_i. Then, we calculate $\mathrm{Sim}_{P,Q}$ based on the sliding window according to P, Q, and t_r. We first select slide window k on random matrix, calculate similarity between k multidimensional vector and $k+1$ multidimensional vector, then move back a sliding window till the end of the sliding window. So, it is similarity of matrix and vector.

$$\begin{cases} \mathrm{granularity}_{\mathrm{BS}}^{\mathrm{threshold}}\left(\mathrm{Sim}_{P,Q}\right) = \mathrm{cluster}_{\mathrm{threshold}}\left(\mathrm{Sim}_{P,Q}(\mathrm{BS})\right) \\ \mathrm{granularity}_{\mathrm{OS}}^{\mathrm{threshold}}\left(\mathrm{Sim}_{P,Q}\right) = \mathrm{cluster}_{\mathrm{threshold}}\left(\mathrm{Sim}_{P,Q}(\mathrm{OS})\right) \end{cases} \tag{7}$$

where $\mathrm{granularity}_{\mathrm{BS}}(\mathrm{Sim}_{P,Q})$ is the granularity of $\mathrm{Sim}_{P,Q}$ in BS and means the set of some M with the close $\mathrm{Sim}_{P,Q}$ in BS. $\mathrm{granularity}_{\mathrm{OS}}(\mathrm{Sim}_{P,Q})$ is the granularity of $\mathrm{Sim}_{P,Q}$ in OS and means the set of some M with the close $\mathrm{Sim}_{P,Q}$ in OS. $\mathrm{cluster}(\mathrm{Sim}_{P,Q}(\mathrm{BS}))$ is cluster method about $\mathrm{Sim}_{P,Q}(\mathrm{BS})$, and $\mathrm{Sim}_{P,Q}(\mathrm{BS})$ is the time sequence similarity of two M in BS.

4.2 Algorithm description

Data set features are noteworthy factors that affect the performance of cluster algorithm. If different data sets are described by their own features, the relationship between data set features and cluster algorithms' performance is obtained.

If the event that happened is a good event, this paper defines that the corresponding emotion of the event is *positive*. If the event that happened is a bad event, this paper defines that the corresponding emotion of the event is *negative*.

ALGORITHM: Matrix Similarity Clustering Algorithm (MSCA)

```
Input：Every $A$;
Output: $granularity_{BS}^{threshold}\left(Sim_{P,Q}\right)$ and $granularity_{OS}^{threshold}\left(Sim_{P,Q}\right)$
Scan: data sources from Eastmoney Website and Sina Website;
    Select data source (OS) from Eastmoney Website by $\rho_{X,Y} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2 \Sigma(Y-\overline{Y})^2}}$;
    Compress ( $A$ ) by equation 4 and equation 5;
        For (i=1; i<=t)
     Select (max( $eigenvalues$ )) ;
        End
    Scan: $e$ ;
        If ( $e$  is negative)
        $A = A - A(positive)$;
         Else $A = A - A(negative)$;
        End
    For i<length of $M$
       Scan: P, Q;
           Generate  $Sim_{P,Q}$  by equation 6;
Generate $granularity_{BS}\left(Sim_{P,Q}\right)$ by clustering ( $Sim_{P,Q}$ );
    End
```

5 Experiment and analysis

5.1 Knowledge map construction

We extract entity, concept, and relationship from the crawled posts to construct wine knowledge system. The knowledge system contains wine knowledge with the Chinese wine stocks as background, and these knowledge linked to each other are beneficial to observe the relationships of data. However, such knowledge system without regular data structure keeps the original connection between knowledge and does not reflect more granularity and dimension characteristic of wine knowledge. The next step is mining these entities and concepts in the field of multigranularity and multidimensional relations according to the background.

We divide entity and concept set of the equivalence2016 class according to the characteristics of corpus, and the relationship set is the attribute set which can be used as the basis for equivalence class division. Entity and concept sets with the relationship tag are divided into the equivalence class set. The entity and concept set are divided into many equivalence classes, and different equivalence classes are made up of different color labels. In order to distinguish different equivalence class, the same equivalence

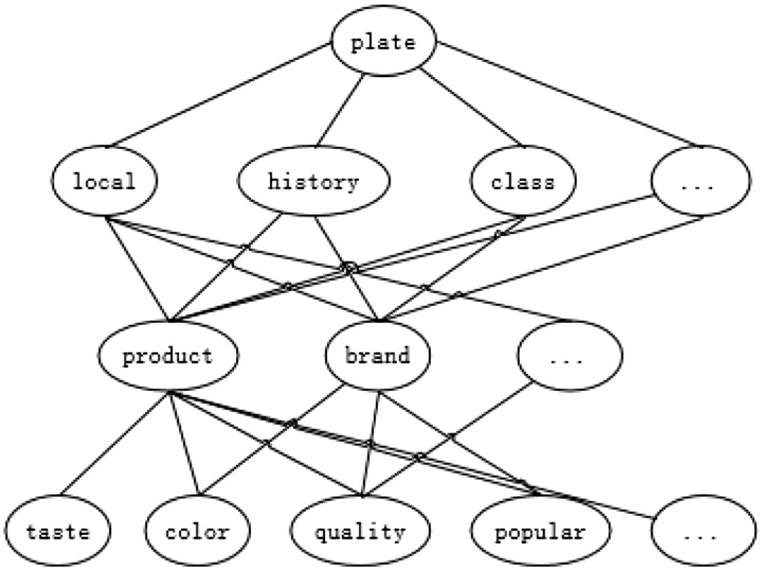


Fig. 5 Multidimensional multigranularity data model

class entity or concept has the same color mark, and each equivalence class has its own subject and mark color (Fig. 5).

5.2 Feature extraction

A of BS has diversity characteristics, and there are multiple relationships between A. In order to reduce redundancy of data and improve effective of operation, we need to reduce dimension. We gain the important variables through linear transformation based on PCA.

The data of 000858 from October 30, 2015 to November 27, 2015 is used as sample data in this paper. We delete the data during holidays because Chinese stock are suspended in that time. The 12 dimensional vectors of BS include Close, High, Low, Open, Pre Close, Change, %Chg, %Turnover, Vol, Turnover, Market Cap, and CSV. We analyze the 12 dimensional vectors based on PCA, the data form the original matrix $[X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}]$. Eigenvalue = (7.2802, 2.6827, 1.8468, 7.2802, 2.6827, 0.0339, 0.0032, 0.0002, 0.0001, 0.0000, 0.0000, 0.0000).

We reconstruct a new data set by a feature vector with larger eigenvalues according to PCA. We select the first three which are 7.2802, 2.6827, and 1.8468 eigenvalues of the eigenvector Y_1, Y_2, and Y_3 to reconstruct a new data set. For example, based on the value of eigenvector (Y_1), we get $Y_1 = 0.3571\,X_1 + (-0.3525)X_2 + \ldots + (-0.3547)X_{12}$. We get Y_2 and Y_3 and use the same method according to the eigenvector (Y_2) and eigenvector

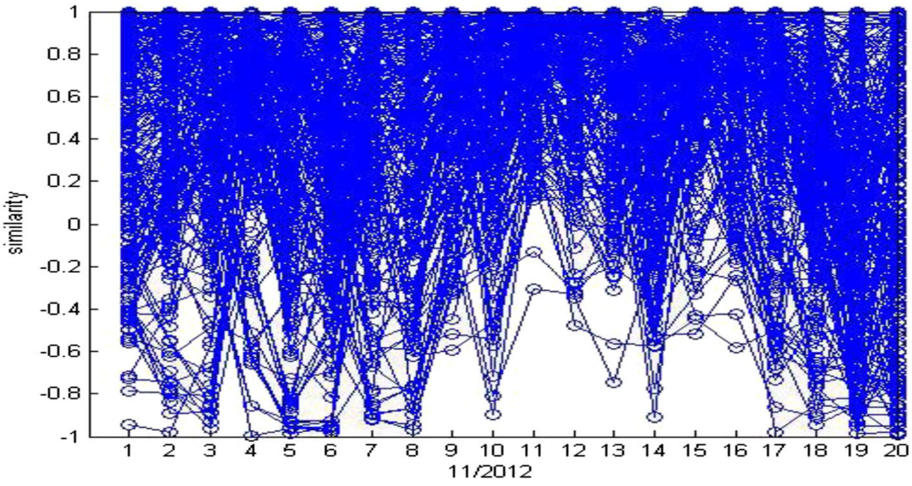


Fig. 6 Similarity of random two M in tea, beverage, and wine plate

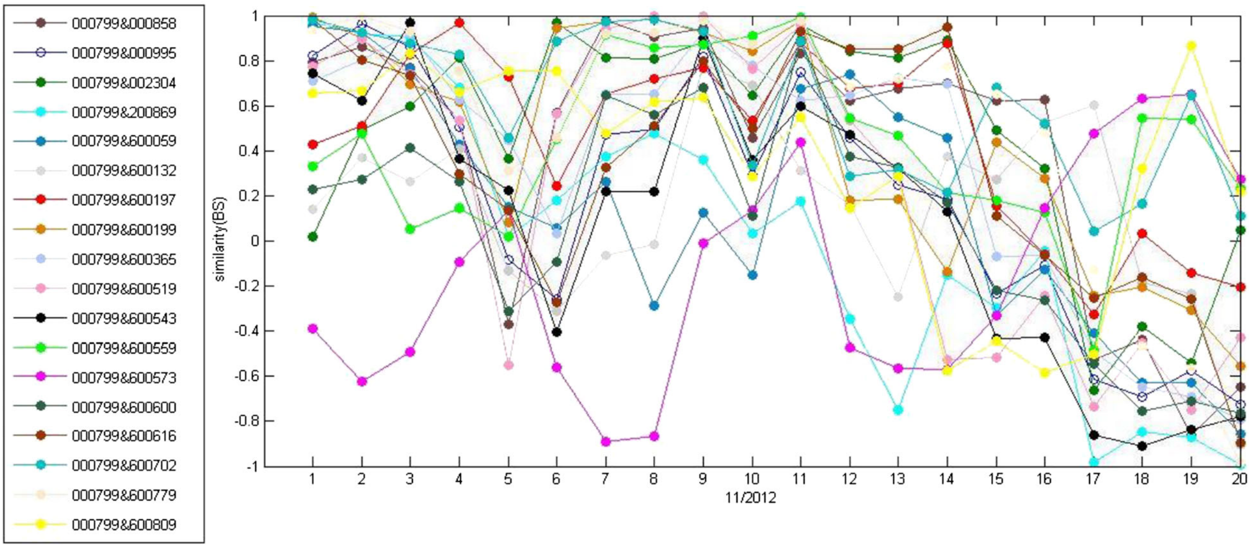


Fig. 7 Similarity of main *M* and others of tea, beverage, wine plate in OS

(Y_3). So the data of BS become three-dimensional vectors based on PCA, which matches to the three-dimensional vectors of OS.

Eigenvector (Y_1) = [−0.3571, −0.3525, −0.334, −0.309, −0.285, −0.1502, −0.1506, −0.2295, −0.2295, −0.2401, −0.3547, −0.3547]

Eigenvector (Y_2) = [−0.0778, −0.0654, −0.2204, −0.2923, 0.3637, 0.3936, 0.394, 0.3698, 0.3698, 0.3571, −0.0782, −0.0782]

Eigenvector (Y_3) = [−0.166, 0.1102, −0.1216, 0.0988, 0.1547, −0.4756, −0.4749, 0.3669, 0.3669, 0.3585, −0.1764, −0.1764]

5.3 Similarity calculation

Forty stocks of tea, beverage, and wine plate have been our test data, the data of OS from Eastmoney website, and the data of BS from Shanghai Stock Exchange and Shenzhen Stock Exchange. According to the difference of OS and BS, there are comments every day in OS. However, the Chinese stock was not traded in holidays, suspension time in BS. Therefore, in order to cooperate with data features of BS, we choose the data of OS where data about the corresponding holidays and suspension days is removed . For group event, there are more than one *M*. So if there are one or several stocks suspension, in

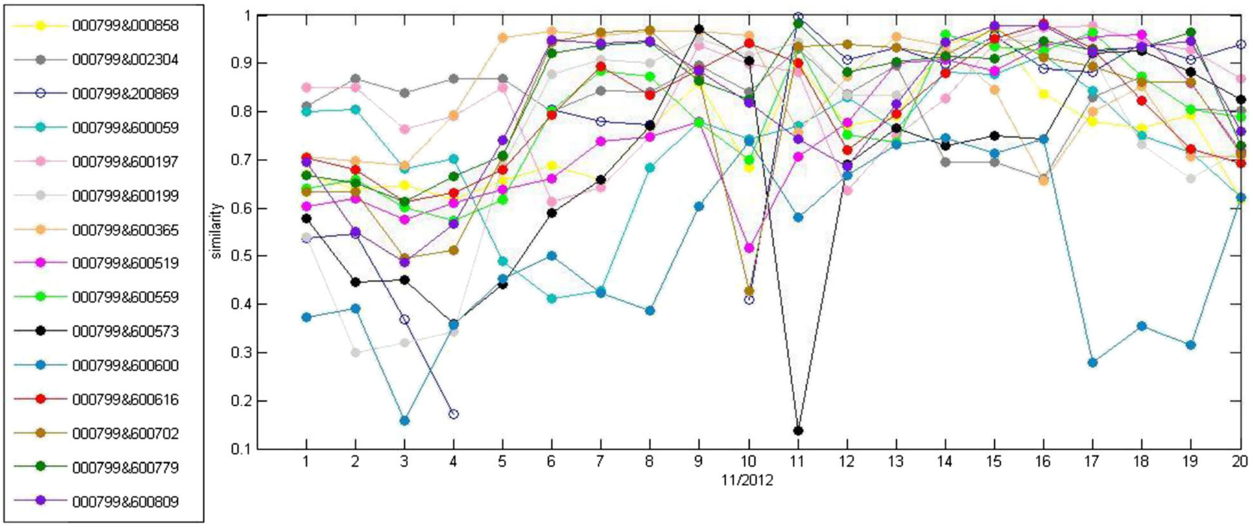


Fig. 8 Similarity of main *M* and others of tea, beverage, wine plate in BS

Table 1 Similarity clustering result in OS when the threshold = 0.8

	9–13	12–14	13–15	14–16	15–19	16–20	19–21	20–22	21–23	23–26	24–27	26–28	27–29	28–30
000858			0.85		0.89			0.93	0.96	0.83				
002304	0.81	0.8	0.88		0.95	0.84	0.81	0.88			0.83	0.87	0.8	0.8
200869					0.99	0.9	0.93	0.89	0.95	0.88	0.88	0.94	0.9	0.94
600059						0.82		0.88	0.87	0.91	0.84			
600197			0.93	0.89	0.88			0.82	0.94	0.97	0.97	0.94	0.92	0.86
600199	0.9	0.9	0.95	0.91	0.94	0.83	0.83	0.92	0.95	0.94	0.91			
600365	0.95	0.96	0.96	0.95		0.87	0.95	0.93	0.84			0.85		
600519							0.9	0.9	0.88	0.93	0.95	0.96	0.86	
600559	0.88	0.87			0.93			0.95	0.93	0.92	0.96	0.87	0.8	
600573			0.97	0.9							0.92	0.92	0.88	0.82
600616	0.89	0.83	0.88	0.94	0.9			0.87	0.95	0.98	0.92	0.82		
600702	0.96	0.96	0.87		0.93	0.93	0.93	0.91	0.97	0.91	0.89	0.86	0.86	
600779	0.93	0.94	0.86	0.82	0.98	0.88	0.9	0.91	0.91	0.94	0.92	0.93	0.96	
600809	0.94	0.94	0.88	0.81			0.81	0.94	0.97	0.97	0.92	0.93	0.94	

order to put the vectors of multiple M in one coordinate, we move forward data after suspension of the life cycle of an event in order to make up the leakage data. Previous studies have shown that this approach does not lead to deviation [19].

Figure 6 shows the similarity of random two M from 43 stocks in tea, beverage, and wine plate during 22 days from November 2012. Their similarity calculation is based on MSCA. We choose window $k = 3$ according to the experience value that calculation effect is better. The multidimensional vector of OS are selected as ⟨click, post, reply⟩, and $\langle Y_1, Y_2, Y_3 \rangle$ in BS. Y_1, Y_2, and Y_3 is constructed by Close, High, Low, Open, Pre Close, Change, %Chg, %Turnover, Vol, Turnover, Market Cap, and CSV, the details is shown in the first section of the experiment.

The plasticizer event in Jiuguijiu on November 19, 2012 is selected in this paper. On November 19, 2012, plasticizer was detected in Jiuguijiu in official news, which has pasticizer more 2.6 times than normal level. The plasticizer is a kind of material additives, and this kind of material is added in the plastic processing which is legally used for industrial purposes. Plasticizer can make the wine become stronger viscosity, longer taste, and look up higher grade and quality. This event is selected because it is very widespread and has a long life cycle, which fit to the purpose of the relationship between the two spaces studied.

There are main multidimensional vector that are constructed by the main M of the event—Jiuguijiu (000799), the similarities of 000799 and each M of the sentiment is the same as the event in test data are calculated respectively and other wine stocks in November 2012, according to the MSCM, and window $k = 3$. Figure 7 shows the similarity based on multidimensional vector random matrix ⟨click, post, reply⟩ of OS. It shows the similarities of 000799 and 000858, 000799, 000858, 600059, 600197, 600199, 600365, 600519, 600559, 600573, 600600, 600616, 600702, 600779, 600809, in November 2012. We can clearly see that similarity change trend of some M is very high, especially on the event that occurs on the day of November 19, 2012 and after the next few days. From Fig. 6, the similarity relationship is more close between members on the beginning time of the Plasticizer event, and the similarity relationship is more loose between members in other time. Except 600059, 600600,600573, 600197, 000858,600519, and 000799, the similarity between others and 000799 are very high. After 15 days, their overall similarities are all decreased which illustrate that the influence of the event become weaker in OS.

Figure 8 shows the similarity based on multidimensional vectors random matrix $\langle Y_1, Y_2, Y_3 \rangle$ of BS. It shows the similarities of 000799 and 000858, 000799, 000858, 200869, 600059, 600132, 600197, 600199, 600365, 600519,

Table 2 Similarity clustering result in BS when the threshold = 0.8

	9–13	12–14	13–15	14–16	15–19	16–20	19–21	20–22
000858	0.97	0.9	0.94		0.83			
002304	0.81	0.8	0.88		0.95	0.84	0.81	0.88
600197					0.87			0.87
600199	0.97	0.98	0.92	0.83	0.97			
600519	0.93	0.99	0.99		0.95			
600559	0.91	0.85	0.86	0.91	0.99			
000799			0.92		0.92	0.85	0.84	0.95
600702	0.97	0.98	0.92		0.88			
600779	0.91	0.92	0.97	0.87	0.97			

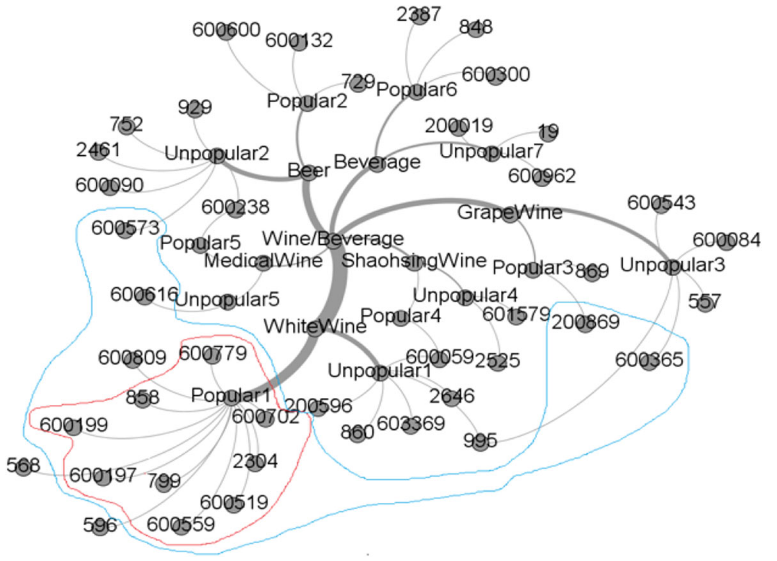


Fig. 9 The relationship set of OS and BS when threshold = 0.8

600543, 600559, 600573, 600600, 600616, 600702, 600779, and 600809, in November 2012. We can clearly see that the similarity change trend of some *M* are very high, especially on the event that occurs on the day of November 19, 2012 and after the next few days. High similarity shows members affected by event in the cycle life of event and other times which are affected by plate and the influence of composite index. Except 600702, 600132, 200869, 600059, and 000799, the similarity between others and 000799 are very high. The changing trend of similarity shows that their relationship are very close and highly consistent with the Plasticizer event. After 10 days, their overall similarities are all decreased which illustrate that the influence of the event become weaker in BS. Contrast to the OS, the influence cycle of the plasticizer event in BS is shorter than the influence cycle of the plasticizer event in BS.

5.4 Relationship discovery

5.4.1 Threshold = 0.8

The tea, beverage, and wine plate can be divided into beer, medical wine, beverage, Shaohsing wine, grape wine, and white wine according to ingredient granularity. White wine = <600238, 000729, 000752, 000929, 002461, 600090, 600132, 600573, 600600>. Beverage = <000019, 200019, 600300, 000848, 002387, 600962 >. Grape wine = <000869, 000557, 200869, 600084, 600365, 600543, 000995>.

Table 3 Similarity clustering result in OS when the threshold = 0.9

	9–13	12–14	13–15	14–16	15–19	16–20	19–21	20–22	21–23	23–26	24–27	26–28	27–29	28–30
000858								0.93	0.96					
002304					0.95									
200869					0.99	0.9	0.93		0.95			0.94	0.9	0.94
600059										0.91				
600197			0.93						0.94	0.97	0.97	0.94	0.92	
600199	0.9	0.9	0.95	0.91	0.94			0.92	0.95	0.94	0.91			
600365	0.95	0.96	0.96	0.95			0.95	0.93						
600519							0.9	0.9		0.93	0.95	0.96		
600559					0.93			0.95	0.93	0.92	0.96			
600573			0.97	0.9							0.92	0.92		
600616				0.94	0.9				0.95	0.98	0.92			
600702	0.96	0.96			0.93	0.93	0.93	0.91	0.97	0.91				
600779	0.93	0.94			0.98		0.9	0.91	0.91	0.94	0.92	0.93	0.96	
600809	0.94	0.94						0.94	0.97	0.97	0.92	0.93	0.94	

Table 4 Similarity clustering result in BS when the threshold = 0.9

	9–13	12–14	13–15	14–16	15–19	16–20	19–21	20–22
000858	0.97	0.9	0.94					
002304					0.95			
600199	0.97	0.98	0.92	0.83	0.97			
000799			0.92		0.92			0.95
600519	0.93	0.99	0.99		0.95			
600559	0.91	0.85	0.86	0.91	0.99			
600702	0.97	0.98	0.92					
600779	0.91	0.92	0.97		0.97			

Shaohsing wine = <601579, 002525, 600059>. Medical wine = <600616, 600238>. Table 1 is the similarity clustering result in OS when the threshold = 0.8. Similarity of 00 0799 and others are widely distributed in Table 1; there are beer, medical wine, white wine, and grape wine.

Table 2 is the similarity clustering result in BS when the threshold = 0.8. However, higher similarity of 000799 and others are relative concentrate in Table 2, and they are only in the popular set of white wine.

In order to better analyze the relationship between the main member and others, we construct a knowledge map of the tea, beverage, and wine plate. The knowledge map of the tea, beverage, and wine plate is mainly composed of beer, medical wine, beverage, Shaohsing wine, grape wine, and white wine according to the ingredient, and each node is divided into two different subsets of popular and unpopular. We further discover the latent relationship between the OS and BS, and the latent relationship between OS and BS with MSCA combined the knowledge map technology.

From Fig. 9, the relationship set of BS is 600779, 000858, 600199, 000799, 600559, 600519, 002304, 600702, and 600197; the relationship set of OS is 600779, 000858, 600199, 000799, 600559, 600519, 002304, 600702, 600809, 600197, 600573, 600616, 200869, and 600365. According to the experimental result, the relationship between BS and OS in the plasticizer event is that the relationship of lead M and others in BS is more concentrate in one set, while the relationship of lead M and others in OS is more disperse in more sets, and the relationship set of BS is a subset of the relationship set of OS. The experimental analysis result is conformity with the industry step of wine, and plasticizer is mainly used to join in white wine to make it stronger viscosity, longer taste, higher grade and quality. Because 000799 belongs to the popular set, the greatest influence M are the nearest granularity. Figure 9 shows the granularity containing relation of OS and BS that is very obvious. $\text{granularity}^{0.8}_{\text{BS}}(\text{Sim})\subseteq\text{granularity}^{0.8}_{\text{OS}}(\text{Sim})$.

Tables 3 and 4 shows the similarity clustering result in BS when the threshold = 0.9. However, the higher similarity of 000799 and others are relative concentrate in Table 4, and they are only in the popular set of white wine. And $\text{granularity}^{0.9}_{\text{BS}}(\text{Sim}_{P,Q}) = \text{granularity}^{0.8}_{\text{BS}}(\text{Sim}_{P,Q})-(600197)$; when threshold = 0.9, $\text{granularity}^{0.9}_{\text{BS}}(\text{Sim})\subseteq\text{granularity}^{0.9}_{\text{OS}}(\text{Sim})$.

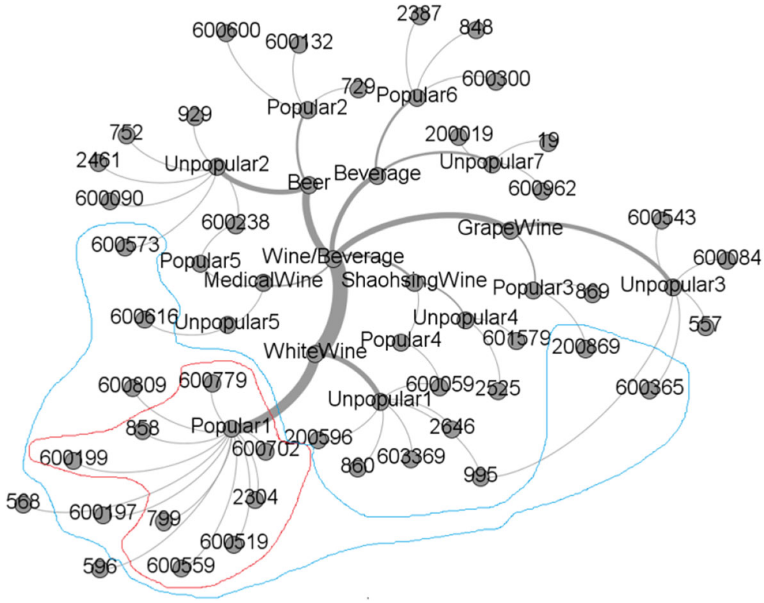


Fig. 10 The relationship set of OS and BS when threshold = 0.9

5.4.2 Threshold = 0.9

Table 3 shows the similarity clustering result in BS when the threshold = 0.9. The higher similarity of 000799 and others are widely distributed in Table 3, there are beer, medical wine, white wine, and grape wine. And $\text{granularity}_{\text{OS}}^{0.9}\left(\text{Sim}_{P,Q}\right) = \text{granularity}_{\text{OS}}^{0.8}\left(\text{Sim}_{P,Q}\right)$, and only the period of similarity have an obvious change. $\text{granularity}_{\text{BS}}^{0.9}(\text{Sim})\subseteq\text{granularity}_{\text{OS}}^{0.9}(\text{Sim})$

From the Fig. 10, the relationship set of BS is 600779, 000858, 600199, 000799, 600559, 600519, 002304, and 600702, and the relationship set of OS is 600779, 000858, 600199, 000799, 600559, 600519, 002304, 600702, 600809, 600197, 600573, 600616, 200869, and 600365. Figure 10 shows the granularity containing relation of OS and BS that is very obvious.

$$\text{granularity}_{\text{BS}}^{0.9}(\text{Sim})\subseteq\text{granularity}_{\text{OS}}^{0.9}(\text{Sim})$$

6 Conclusions

In this paper, we calculated the event relationship in a new method and found the relationships between Public Opinion Data Space (OS) and Actual Behavior Data Space (BS) in the life cycle of an event. We calculated the similarity of multidimensional time sequence combined with the sliding window technology based on random matrix theory. We not only calculated the similarity of lead member and others in OS and BS but also found the relationship between OS and BS. Furthermore, we have constructed a knowledge map to analyze the relationship of public opinion and actual behavior. Experimental results showed that MSCM is a typical scheme using both OS and BS and found that the relationship set of OS is more wider and the relationship set of BS is more concentrate, while the relationship set of BS is a subset of the relationship set of OS.

Acknowledgements

This work is supported by The Ocean Public Welfare Project of the Ministry of Science and Technology (No. 201105033) and The National Natural Science Foundation of China (No. 40976108).

Competing interests

The authors declare that we have no competing interests.

Author details

[1]Department of Computer Science, Guangdong Polytechnic Institute, Guangzhou, China. [2]School of Computer Engineering and Science, Shanghai University, Shanghai, China.

Received: 4 March 2016 Accepted: 9 August 2016
Published online: 09 September 2016

References

1. R Rajkumar, I Lee, L Sha, J Stankovic, Cyber-physical systems: the next computing revolution [C]. Strasbourg **14**(6), 731–736 (2010)
2. Z Levnajić, A Pikovsky, Network reconstruction from random phase resetting [J]. Phys. Rev. Lett. **107**, 034101 (2011)
3. M Timme, J Casadiego, Revealing networks from dynamics: an introduction [J]. Phys. A **47**, 343001 (2014)
4. Y Xue, L Xu, J Yu, L Wang, G Zhang, *An Estimate Method of Event Influence Scope based on Special Field, 2015 International Conference on Identification, Information & Knowledge in the Internet of Things*, pp. 39–44 (2015)
5. B Podobnik, D Wang, D Horvatic, I Grosse, HE Stanley, Time-lag cross-correlations in collective phenomena [J]. EPL **90**, 68001 (2010)
6. H Meng, WJ Xie, ZQ Jiang, B Podobnik, WX Zhou, HE Stanley, Systemic risk and spatiotemporal dynamics of the housing market [J]. Sci. Rep. **4**, 3655 (2014)
7. RN Mantegna, HE Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance [M]* (Cambridge University Press, England, 2000)
8. JP Bouchaud, M Potters, *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management [M]* (Cambridge University Press, England, 2003)
9. RK Pan, S Sdsfha, Self-organization of price fluctuation distribution involving markets [J]. Eutrophys. Lett. **77**, 58004 (2007)
10. T Aste, W Shaw, TD Matteo, Correlation structure and dynamics in volatile markets [J]. New J. Phys. **12**, 085009 (2010)
11. V Plerou, P Gopikrishnan, B Rosenow, LAN Amaral, T Guhr, HE Stanley, Random matrix approach to cross correlations in financial data [J]. Phys. Rev. **65**, 066126 (2002)
12. A Utsugi, K Ino, M Oshikawa, Random matrix theory analysis of cross correlations in financial markets [J]. Phys. Rev. **E70**, 026110 (2004)
13. J Shen, B Zheng, Cross-correlation in financial dynamics [J]. Europhys. Lett. **86**, 48005 (2009)
14. XR Jiang, B Zheng, Anti-correlation and subsector structure in financial systems [J]. EPL **97**, 48006 (2012)
15. G Oh, C Eom, F Wang, WS Jung, HE Stanley, S Kim, Statistical properties of cross-correlation in the Korean stock market [J]. Euro. Phys. J. B **79**, 55 (2011)
16. H Yetomi, Y Nakayama, H Aoyama, Y Fujiwara, Y Ikeda, W Souma, Fluctuation-dissipation theory of input-output inter industrial relations [J]. Phys. Rev. E **83**, 016103 (2011)
17. F Ren, WX Zhou, Dynamic evolution of cross-correlations in the Chinese stock market [J]. PLoS One **9**, e97711 (2014)
18. CC Aggarwaland, K Subbian, Event detection in social streams [C], in *SDM'12*, 2012, pp. 624–635
19. R Li, KH Lei, R Khadiwala, KC-C Chang, Tedas: a twitter-based event detection and analysis system [C]. Data Eng. Int. Conf. on **0**, 1273–1276 (2012)
20. J Holland, *Emergence: from chaos to order [M]* (Addison-Wesley, Redwood City, 1997)
21. C Faloutsos, M Ranganathan, Y Manolopoulos, Fast subsequence matching in time-series databases[C]. Proceeding of the ACM SIG-MOD International Conference on Management of Data, 1994. pp. 419–429.
22. I Jolliffe, *Principal component analysis [M]*. (Wiley, 2005)
23. Y Lu, I Cohen, XS Zhou, Q Tian, Feature selection using principal feature analysis. *In Proceedings of the 15th international conference on multimedia*. (ACM, 2007), pp. 301-304
24. H Yoon, K Yang, C Shahabi, Feature subset selection and feature ranking for multivariate time series [J]. IEEE Trans. Knowl. Data Eng. **17**(9), 1186–1198 (2005)
25. P Esling, C Agon, Time-series data mining. ACM Compute Survey. [J] **45**(1), 1–34 (2012)
26. GEAPA Batista, X.Wang, EJ Keogh, *A Complexity-Invariant Distance Measure for Time Series [C] // Proceedings of the Eleventh SIAM International Conference on Data Mining* (SIAM, Mesa, 2011), pp. 699–710
27. B Boucheham, Reduced data similarity-based matching for time series patterns alignment. Pattern Recogn. Lett. [J] **31**(7), 629–638 (2010)
28. B-K YI, C Faloutsos, *Fast Time Sequence Indexing for Arbitrary Lp Norms [C] //Proceedings of the 26th International Conference on Very Large Data Bases (VLDB), Morgan Kaufinann Publishers Inc.* (Cairo, Egypt, 2000), pp. 385–394.
29. H Ding, G Trajcevski, P Scheuermann et al., Querying and mining of time series data: experimental comparison of representations and distance measures. Proc. VLDB Endow. [J] **1**(2), 1542–1552 (2008)
30. E Keogh, S Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min. Knowl. Disc. [J] **7**(4), 349–371 (2003)