Temporal Pattern of User Behavior in Micro-blog

Chunhong Zhang
Beijing University of Posts and Telecommunications, Beijing, China
Email: zhangch.bupt.001@gmail.com

Yaxi He and Yang Ji
Beijing University of Posts and Telecommunications, Beijing, China
Email: {heyaxi9005, ji.yang.0001}@gmail.com

***Abstract*—Temporal pattern provides a novel way to character user behavior in social network from the perspective of time. In this work, we study two types of temporal pattern of user behavior in Micro-blog: Long Term Pattern and Daily Pattern. Long Term Pattern stands for the overall trend of user behavior changes since one starts to use Micro-blog and it provides the global view of user behavior variations. Daily Pattern states about the everyday variation and it represents the microscopic regularity within a day. In order to find out temporal pattern, Wavelet Transformation and Dynamic time warping for K-Medoids algorithm (WT-DKM) is proposed to organize time series into dusters, each cluster corresponding to a pattern. Eventually, 4 long term patterns and 5 daily patterns are discovered. These patterns are various in many terms, which reveal the difference of regularity among users. Almost half users behave randomly without apparent regularity while the others' behaviors have obvious variation trend along with time. User group is often used to character user behavior according to their status, age, etc. So we study the relationship between temporal pattern and user group to discover whether users in the same user group are more likely to behave with same type of temporal pattern. It turns out that for some user groups temporal pattern of most members are identical, while other groups are not.**

***Index Terms*—temporal pattern, time series, SNS user behavior, clustering algorithms**

I. INTRODUCTION

As the development of Social Network Service (SNS), such as Micro-blog, the attitude of users is changing towards SNS, leading to changes of their behavior. Some may become more reliable on communication through this kind of online social platform, thus they post or retweet more and more microblogs, or comment more frequently. On the other hand, some users may get bored of SNS, thus less and less microblogs might be posted. The temporal variation of individual behavior ranges from random [16] to highly-correlated [9] and every user has his/her temporal pattern. However, similar pattern of individual behavior can be aggregated to create some representative patterns, which demonstrate how collective user behaviors change over time. Thus, temporal pattern provides a way to describe user behavior from the perspective of time. Mining temporal pattern of user behavior not only contributes to figure out valuable and active users, but also improve effectiveness of SNS by identifying infrequent ones and even proposing ideas to motivate such users. Recommendations can be improved by adding time factor if temporal patterns are distinguished.

However, uncovering patterns of human behavior is a difficult task because there are numerous factors that influence people's action, such as their mood, lifestyle or working environment, etc. Someone may post microblogs when he is happy while some others prefer to do this when he feels upset. Although the temporal pattern related to SNS has been conducted by some researchers, their attention is on the social content, like the popularity temporal pattern of a microblog or a video [11, 20]. Very few researchers have explained what the temporal pattern of user behavior is like in SNS.

Temporal pattern of user behavior is discovered from time series formed by users' online history. A time series is "a sequence of observed data, usually ordered in time" [2]. Time series can be abstracted from users' history online in many ways relevant to different practical implications, but the corresponding patterns exist and can be discovered. In this work, we determine to discover temporal pattern from two perspectives: Long Term Pattern (LTP) and Daily Pattern (DP). Long Term Pattern stands for the regularity of user behavior from the moment a user began to register SNS till the last time. Daily Pattern concerns about the regularity of users' daily behavior in SNS. Time series used for LTP discovery is called Long Time Series (LTS), and that used for DP discovery is called Short Time Series (STS). Fig. 1 and Fig. 2 depict the LTP and DP of two users in Micro-blog. The x-axis is the time period and y-axis is the number of microblogs during this period. Thus values of y-axis express active degree of a user where peak means high activeness and valley measures the dormancy. In Fig. 1, the curve of User2 is stable through the whole time, which means he posts microblogs all the time, while the curve of User1 tends to go up with fluctuations illustrating that he becomes more reliable on Micro-blog and post more microblogs over time. The two curves are not equal in length. In Fig. 2, User1 is vibrant at noon and at midnight while User2 tends to be active at any time all day along.

doi:10.4304/jsw.8.7.1707-1717

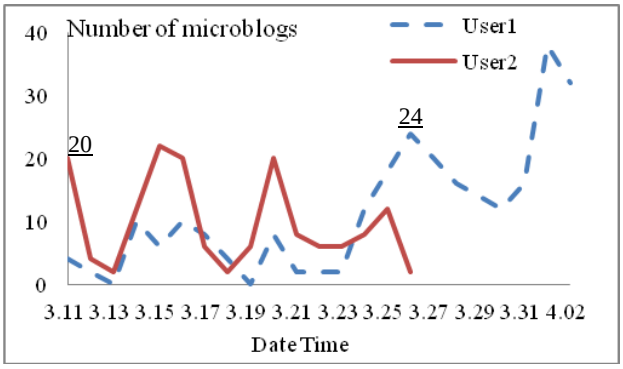

Figure 1. Long Time Series of two users. *User1 posted 20 microblogs in Mar.11, 2012, so the 1st value of LTS is 22. So as User2.*

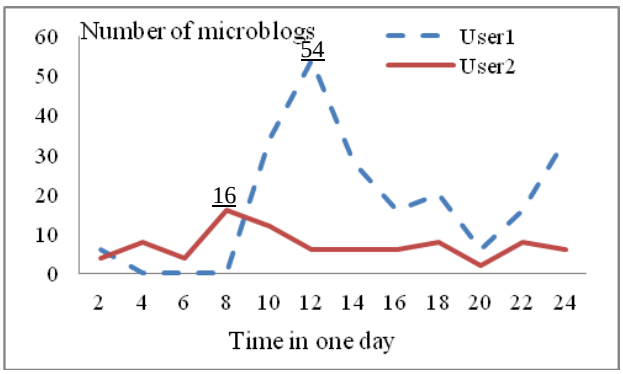

Figure 2. Short Time Series of two users. *User1 posted 54 microblogs in 10-12 o'clock, so the 12th value of STS is 54. So as User2.*

When studying user behavior, there are many perspectives to describe or study SNS user behavior except temporal pattern. Users are usually divided into user groups according to certain principles such as their age, motivation, social status, etc. So, would users in same user group be more likely to display identical temporal pattern is also what we want to know. This reveals the relevance of the two perspectives on user behavior study.

In this work, we aim to discover what temporal patterns of SNS user behavior hidden in time series and how different user group shape the temporal dynamics. We first define how time series is formed from the raw data of Micro-blog. Then, the unsupervised clustering algorithm, Wavelet Transformation and Dynamic time warping for K-Medoids algorithm (WT-DKM) is proposed to discover how many types of LTP and DP exist of SNS user behavior and what they are. This algorithm is able to get the overall trend variation of user behavior and find out time series of same shape even if they are of unequal length. Eventually, we analyze the temporal pattern of six user groups to obtain the temporal character for them. Through the experiment with real data from Sina Micro-blog, we analyze the potential relationship between Long Term Pattern and Daily Pattern.

Our algorithm provides proper solutions for the challenges. The remainder of this paper is structured as follows: Section II presents the related work. Section III describes our algorithm of WT-DKM used for pattern discovery. The experiment of our algorithm on pattern discovery of SNS user behavior in Sina Micro-blog is described in Section IV. We demonstrate applications of temporal pattern in Section V and conclude our work in Section VI.

II. Related Work and Challenges

Temporal pattern is a structure that makes a specific statement about the whole time series or just several data points. By analyzing time series, it is possible to discover hidden patterns, unexpected trends or other subtle facts in large data. Temporal pattern discovery has been researched in many areas such as financial prediction [13], weather forecasting [15] and cultural markets with the techniques such as Hidden Markov Models, time-delay neural networks or data mining methods [1]. For temporal pattern in SNS, there are something can be borrowed to find out collective behavior or large-scale trend. However, most of previous work focuses on the temporal pattern of social content. Ref. [19] proposed a framework to extract meaningful evolutionary patterns from group photo streams with a nonnegative joint matrix factorization approach to incorporate image content features and contextual information, such as associated tags, photo owners and post time and their prediction results outperform baseline methods. In [11] and [20], researchers computed the popularity patterns of online content such as hashtags in Twitter and videos in YouTube. Our attention is kind of like study in [17]. They found out the distribution of temporal behavior and activeness of users, in which more than half users displayed similar pattern and they are not active, for their average presence is 1.6 days over 16 days in Twitter. But they did not give what the overall trend is like over time. This is what we want to find out.

Although, there are amount of techniques inspired for temporal pattern discovery with time series, we believe that they can be divided into two different methods in practice.

1. Template matching, finding patterns that match predefined templates. Given a model of templates, each time series is made to match pattern by similarity search and templates are adjusted to become patterns. It is prerequisite to have a priori knowledge of the patterns or templates to be discovered.

2. Clustering, each cluster stands for a pattern. Through clustering the time series, patterns could be concluded by results, such as the averaging all members of a cluster as the pattern of this cluster.

With regard to templates matching methods, each time series is matched with predefined templates with related algorithms. Ref. [5] predefined the pattern model and took dynamic time warping technique to adjust time series and the predefined pattern templates by dynamic coding. Ref. [7] employed the time-delay embedding process to match time series generated form robot sensors with predefined templates. Keogh took piecewise function to represent template, and local features such as peaks, troughs and platueaus, were defined using a prior distribution in expected deformations from a basic template to match the templates with probabilistic method

[10]. However, the use of predefined templates completely prevents the achievement of the basic data mining goal of discovering useful, novel, and hidden temporal patterns. Besides, we do not have a priori knowledge of what temporal patterns may exist in the circumstance of SNS user behavior. Thus, clustering seems to be more practical.

The goal of clustering is to organize time series into homogeneous groups where the within-group-object dissimilarity is minimized and the between-group-object dissimilarity is maximized, which is often accomplished by raw data-based partition, feature-based methods or model-based methods [12]. K-means, K-medoids or fuzzy K-means, K-medoids are classical partition-based methods for clustering and there are lots of improved algorithms that fit the practical application [22, 23]. Ref. [18] introduced the Anytime K-Means algorithm based clustering time series to improve cluster quality by optimizing initial centroids by wavelet for K-means processing, which also dealt with equal length series. They performed this algorithm on several datasets and evaluated the performance compared with traditional K-means. Ref. [11] constructed a K-Spectral Centroid clustering algorithm which defined a distinctive distance calculation function for time series formed by online content's click rate, in order to get the popularity pattern. They came up with six distinct patterns. In [17], researchers proposed an example of clustering time series by hierarchical clustering algorithm based on K-means. Time series they got were formed by user behavior like regularity of access or activity, which was also equal in length, but they just gave the patterns' number and did not care about what they are like. However, in our work, the first aim is to discover temporal pattern of user behavior by time series that are of similar shape and unequal length, previous studies are unable to fulfill this requirement. Besides, they utilize the origin time series directly that contains too much detailed information, while we concentrate on their overall trend.

When discovering temporal pattern with clustering time series, there are some unique challenges to be dealt with. When the dimensionality of the time series is high, it has been proved that the distance of nearest time series is not different from others if calculated by measurement such as Euclidean distance [4], one reason is that consecutive values of a time series lead to the local fluctuation, which could lead to the result that distance between different series are the similar, damaging distinguishing ability of partition clustering algorithm. So, time series derived from raw data needs to be preprocessed. In addition, as our goal is to study the trend of behavior variation, we care about overall trend of a time series rather than its consecutive values. Thus, local fluctuations attached on the trend require to be removed. Wavelet Transformation (WT) can be utilized to smooth the Long Time Series. WT is used to transform data from time domain to anther domain in terms of separating the averages and differences of the original data [8]. The average component, i.e. the low frequency, stands for the trend and the difference component, i.e. the high frequency, represents the fluctuations of the original series. The advantage of WT is its multi-resolution representation of series which determines what degree of fluctuations is removed. Given the resolution, a series is approximated by neglecting all of fluctuations below. As noise often exists in high-frequency components, reduction of resolution contributes to remove detailed information and remain the overall trend of original time series.

In order to find temporal pattern of user behavior, we need search for time series of similar shape and organize them into same cluster. This could easily be done by human vision, but for machines this is a difficult task. Time series may have falls, rises or flats that occur in the same order, but they do not align in the time axis. This still displays the same pattern. Besides, since users did not register the SNS at the same time, the time series is not of equal length, while previous studies have ignored this problem and they take advantage of partial data with equal length. Dynamic Time Warping (DTW), which was introduced to data mining by Berndt and Clifford [3], is a method for similarity measurement of shapes between two series even if they are not aligned in the time axis. The ability makes it an essential similarity model in speech recognition, and financial sector. With DTW, series are considered similar if they have the same rise and fall patterns, neglecting the difference in time axis or in length.

III. METHODOLOGY

In order to find out hidden temporal pattern of SNS user behavior, we classify time series by a novel clustering algorithm, named Wavelet Transformation and Dynamic time warping for K-medoids clustering, which is abbreviated as WT-DKM. The first step of discovering Long Term Pattern is to achieve the trend of all time series by Wavelet Transformation because time series usually consists local fluctuations while the overall trend of time series is what actually in need. Then we find out similar series of the same shape by clustering to conclude temporal pattern. Dynamic time warping is used to measure similarity between time series. Then, we use reduced WT-DKM algorithm to obtain the Daily Pattern of user behavior, for the Short Time Series are created to be of equal length and short that WT is not necessary.

A. Time Series Creation

As SNS user behavior is the actions users take related to social network service and there are many kinds such as scanning and posting microblogs, or following others. Raw data are obtained from the OpenPlatform API provided by the SNS operator. We simply regard the action of posting microblogs as representative of user behavior because it is the most common action a user performs in Micro-blog. Besides, we can get the raw data of the actual posting time of a microblog through API.

In order to discover temporal pattern of user behavior, time series formation is the foundation. For each user, a time series is formed by the total number of microblogs during sampling period. Long Time Series and Short

Term Series differ in sampling period and the way to handle sampling results. Fig.3 depicts the LTS and STS creation process. LTS is formed from the moment that a user published his/her first microblog till now. We define SP_{LTS} as sampling period and the unit is day. Thus, every user has a LTS. The LTS of user i is defined as $X_i=[a_{1i},a_{2i},\ldots,a_{ij},\ldots,a_{n_i i}]$. a_{ij} is the number of microblogs in j^{th} sampling period and n_i is the length of LTS, which may be different for different user. With regard to Short Time Series, we define SP_{STS} as sampling period and the unit is hour. STS is formed by continuous values of the number of microblogs within one day from 0 o'clock to 24 o'clock. Then we sum up the values of the same period in T days to create the STS. The STS of user i is defined as $X_i'''=[a_{1i}''',a_{2i}''',\ldots,a_{ij}''',\ldots,a_{ni}''']$. a_{ij}''' is the number of microblogs in j^{th} sampling period and $n=24/SP_{STS}$, which is same for all users.

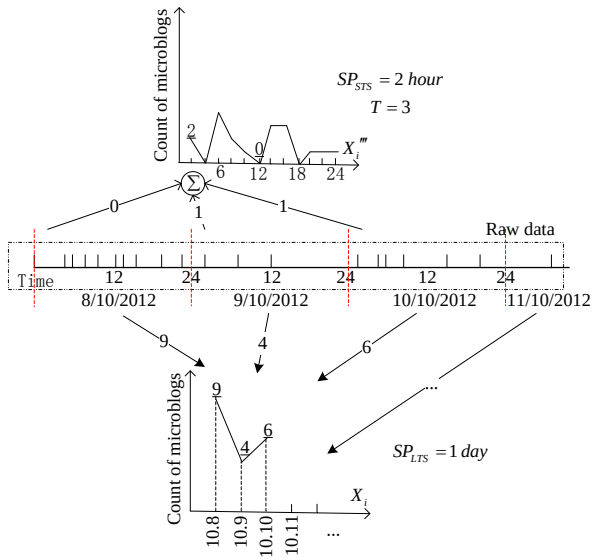

Figure 3. Time Series creation process. *Raw data is the actual posting time of microblogs. User i posted 9 microblogs in Oct. 8th, 2012, so the first value of X_i is 9. SP_{LTS}=1 day. User i posted 0, 1, 1 microblogs in 0 o'clock to 2 o'clock from Oct. 8th -10th, so the first value of X_i''' is 0+1+1=2. SP_{STS}=2 hour and T=3.*

B. Long Term Pattern Discovery

The procedure of long term pattern discovery consists of three parts. The first step is to preprocess original Long Time Series with Wavelet Transformation to obtain the overall trend. Then clustering algorithm is used on preprocessed LTS, and eventually Long Term Pattern can be represented by the clustering results.

1) Preprocessing LTS with WT

After long term time series is formed, WT is employed on LTS for preprocessing to obtain their overall trend.

Haar Wavelet, which is the simplest wavelet and works by averaging two adjacent values of the time series, is used. The averaging process is performed iteratively until the average value of the series is obtained. Scale is the index of iteration which determines the resolution of original series shown as Fig.3. Haar WT is employed on each time series to divide the detail component and the approximation component. Then the approximation component is used to reconstruct original time series with Inverse Wavelet Transformation (IWT) to obtain the trend of original series. In this way, local fluctuations are removed. Given a LTS $X_i=[a_{1i},a_{2i},\ldots,a_{n_i i}]$ and *scale* for WT, this transformation can be expressed as follows:

$$X_i^{'}=WT(Haar,X_i,scale)$$

$$X_i^{wt}=IWT(Haar,X_i^{'})$$

When performing WT on time series, we have to consider how to decide the *scale*. From Fig.4, we can know that high scale may result in over averaged series where the trend variation is also regarded as local fluctuation and then is erased. However, if the scale is low, time series may still contain a lot of local fluctuations leading to inefficacy of clustering algorithms. In our work, we determine the *scale* according to clustering results.

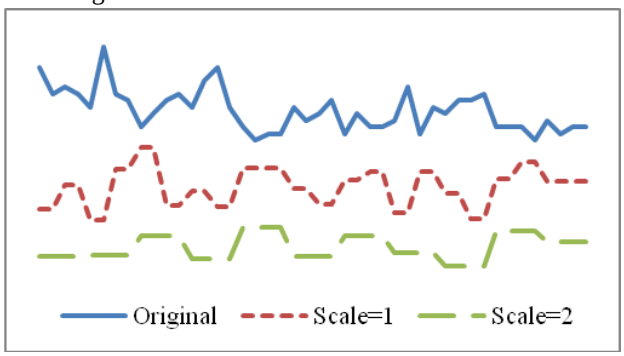

Figure 4. Original time series is approximated by Haar WT at different resolution. *High scale of WT results in low resolution of original series.*

2) Clustering with DKM

After each Long Time Series of a user is preprocessed by Wavelet Transformation, clustering algorithm, K-medoids, is performed for LTP discovery. Every cluster corresponds to a pattern. When performing the clustering, DTW distance is employed as distance measurement for time series. The K-medoids clustering with DTW distance is called DKM.

a) Clustering method

Clustering method is used to divide preprocessed LTS into several clusters and then to get corresponding patterns. We prefer K-medoids rather than K-means due to that the performance of K-means is sensitive to outlier. Even if an object is quite far away from cluster centroid, K-means still force it into a cluster and thus distorting the cluster quality. Besides, when computing the centroid, K-means calculates the means of all members in a cluster as the cluster centroid while the Long Time Series have unequal length and it is impossible to compute the means. But K-medoids utilizes real data located in the center of a cluster. This can avoid the effect of outliers and do not have to calculate means. Therefore, K-medoids is chosen for clustering.

b) Similarity measure

When measuring the similarity of time series, the distance calculation methods must cater for finding out similar shape time series with unequal length.

When creating Long Time Series from user's online history, we utilize a fixed sampling period, as stated above. As not all users begin to use SNS at the same time, the LTS of different users are not equal in length. Although other methods could be used to avoid unequal length such as using in variable sampling period, yet this will cause problem of determining a reasonable length for all time series. Short length would hide obvious trend variation and long length would produce lots of null value in the series. Behavior variation is expressed at different granularity for every user which makes it unable to reflect the real variations. Therefore, we prefer to handle distance measurement of unequal length series.

Since we are aiming to find time series with similar shape even if they have different length, the normally used Euclidean distance or Cosine Distance is not appropriate. Dynamic Time Warping distance is utilized for similarity measurement of LTS in K-medoids for Long Term Pattern discovery. If two series have similar overall shape but they are not aligned at time axis, DTW is able to provide intuitive similarity measurement. Given two preprocessed LTS, $X_i^{wt}=[a_{1i},a_{2i},\ldots,a_{pi},\ldots,a_{n_i i}]$ and $X_j^{wt}=[b_{1j},b_{2j},\ldots,b_{qj},\ldots,b_{n_j j}]$, Euclidean distance assumes the i[th] value in one series is aligned directly with corresponding i[th] value in the other and takes the squared sum of aligned points' dissimilarity as the overall difference. The Euclidean distance of X_i^{wt} and X_j^{wt} is

$$D_{Euc}(X_i^{wt},X_j^{wt})=\sqrt{\sum_{m=1}^{n_i}(a_{mi}-b_{mj})^2}\ , \textit{only when}\ n_i=n_j .$$

DTW takes a nonlinear alignment that is more sophisticated and does not require $n_i=n_j$. It aligns the mountains and valleys of two time series as much as possible by expanding and compressing the time axis, in order to minimize the distance. DTW calculation begins by constructing an $n_i \times n_j$ warping matrix. The (p,q) element in matrix corresponds to an ailment between a_{pi} and b_{qj}, calculated as $d(a_{pi},b_{qj})=(a_{pi}-b_{qj})^2$ (Euclidean distance). A warping path W is a contiguous set of matrix elements that defines a mapping between X_i^{wt} and X_j^{wt}, and the l^{th} element is defined as $w_l=(p,q)_l$, thus a path can be expressed as $W=w_1,w_2,\ldots,w_l,\ldots,w_L$, $\max(n_i,n_j)\le L\le n_i+n_j-1$. The path that minimizes the warping cost is what we are looking for, so the DTW distance of X_i^{wt} and X_j^{wt} is defined as

$$D_{DTW}(X_i^{wt},X_j^{wt})=\min \sum_{l=1}^{L} w_l \Big/ L\ ,$$

$$\max(n_i,n_j)\le L\le n_i+n_j-1 .$$

L is used to compensate for the influence of length of warping path. Dynamic programming is used to find this path by calculating the cumulative distance $d_{cum}(p,q)$, which is defined as an iterative formula,

$$d_{cum}(p,q)=d(a_{pi},b_{qj})+\min\{d_{cum}(p-1,q-1),d_{cum}(p-1,q),d_{cum}(p,q-1)\}$$

Thus, $D_{DTW}(X_i^{wt},X_j^{wt})=d_{cum}(n_i,n_j)$.

Suppose there are three time series shown as Fig. 5,

$$X_1^{wt}=[7,7,7,7,7,6,5,4,4,4,4,4,4]\ ,$$

$$X_2^{wt}=[5,5,5,5,4,3,3,3,3,3,3,3,3]\ ,$$

$$X_3^{wt}=[4,4,4,4,4,4,5,5,5,5,5,5,5] .$$

X_2^{wt} is more similar to X_1^{wt} than X_3^{wt} estimated by human vision, however, when measuring the similarity with Euclidean Distance, we get

$$\{D_{Euc}(X_1^{wt},X_2^{wt})=6.55\}>\{D_{Euc}(X_2^{wt},X_3^{wt})=5.39\} .$$

When calculating the distance with DTW, we have

$$\{D_{DTW}(X_1^{wt},X_2^{wt})=5.29\}<\{D_{DTW}(X_2^{wt},X_3^{wt})=5.39\} .$$

So, DTW Distance is more suitable for measuring similarity between time series by shape.

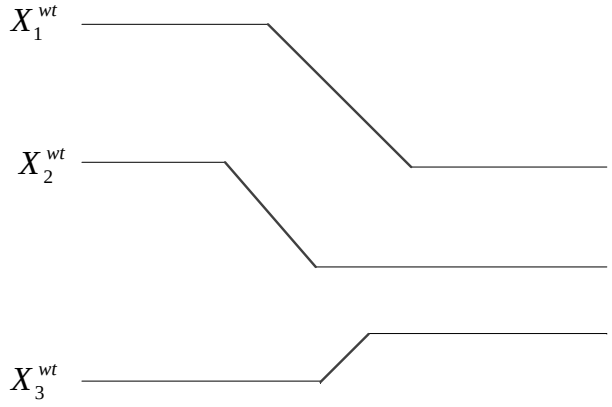

Figure 5. Comparison of Euclidean distance and DTW distance

c) Clustering process

The objective of DKM algorithm is to discover K patterns from a set of M series, $X_i^{wt}=[a_{1i},a_{2i},\ldots,a_{pi}\ldots,a_{n_i i}]$, $i=1,2,\ldots M$. DKM starts by selecting K initial cluster centroids, μ_k^{wt}, $1\le k\le K$, which are determined randomly. Then, each series, X_i^{wt}, is assigned to the nearest cluster by calculating the distance between X_i^{wt} and μ_k^{wt} by $D_{DWT}(X_i^{wt},\mu_k^{wt})$. Then we update the cluster centroid with all time series. The final cluster centroids $\hat{\mu}_k^{wt}$ must ensure that the distance sum of $D_{DWT}(X_i^{wt},\mu_k^{wt})$ over all X_i^{wt} in cluster C_k should be minimal, that is

$$\hat{\mu}_k^{wt} \propto \min(\sum_{k=1}^{K} \sum_{X_i^{wt} \in C_k} D_{DWT}(X_i^{wt}, \mu_k^{wt})^2), \ k \in [1,K].$$

As the clustering results of K-medoids may lead to local optimization caused by the initial clustering centroid, although there are ways for optimization [22], we perform multiple runs of DKM, each with a different set of randomly chosen initial controids and then select the set of minimum distance sum.

When performing DKM, the number of cluster, K, is corresponding to the number of patterns. K guarantees obtaining quality and effective patterns that reflects the variation of SNS user behavior and must be given before DKM starts. In order to decide the optimal value of K, we utilize two criterions, Silhouette Coefficient (SC) and Hartigan Index (HI), as the evaluation criterion [14, 15, 21]. These two criterions are able to judge the clustering result quality by measuring cohesion within the same cluster and the separation between different clusters. When SC(K) is maximum or HI(K) is the minimum value of the successive difference, the corresponding K is the optimal, indicating the derived clusters are able to distinguish time series.

3) Pattern representation

As we regard each cluster stands for a temporal pattern, we obtain K long term patterns. It is complicated on how to represent temporal pattern through clustering results. It is common practice to calculate the means of all members within the same cluster as representative of the cluster, or, add up all members with weight. However, in our work, the length of LTS is unequal, making averaging or adding difficult. It is the centroid that is the core of all members within a cluster and contains unique characters, so we believe centroids, $\hat{\mu}_k^{wt}$, are capable to be representatives of clusters. Thus, the centroid of each cluster is a temporal pattern.

In conclusion, WT-DKM algorithm for discovering Long Term Pattern is concluded by the pseudo-code followed:

Input: a set of LTS, X_i, where i=1,2…,M,
Output: pattern number: $\hat{K}$
pattern membership
pattern representation(PR)
Step1: Wavelet Transformation
for i=1 to M
$X_i^{'}$=WT (X_i, scale)
X_i^{wt}=IWT ($X_i^{'}$, scale)
end
Step2: Clustering Process
for K=2 to $\sqrt{M}$
[centroid(K), cluster_member(K)]=DKM (K, X^{wt})
end
$\hat{K}$ = SC_HI(centroid, cluster_member)
Step3: Pattern Representation
PR=centroid($\hat{K}$)

C. Daily Pattern Discovery

When discovering the Daily Pattern with a set of Short Time Series, DKM is employed directly. Wavelet Transformation is skipped because the length of STS is much shorter than most LTS. Although the length of STS is same, we still utilize DTW Distance to measure similarity of time series. The process of determining initial cluster centroid and optimal pattern number, K, is similar to that of Long Term Pattern discovery. Eventually, daily patterns of users' daily behavior are obtained in terms of cluster centroids.

IV. EXPERIMENT

A. Data Collection

We conducted the experiment in Sina Micro-blog, which is a kind of social network service like Twitter that users can post microblogs and follow others. The data are obtained through Sina Micro-blog OpenPlatform API. We select 3000 users and spend 15 days collecting all microblogs of them. In order to study whether temporal pattern is related to user groups, six user groups are selected based on priori knowledge shown in Fig. 6. Celebrities are famous persons that have great influence in particular areas in real life such as famous singers, professors and CEOs, which are confirmed by Sina. Enterprises or organizations take advantage of Micro-blog to publish information or communicate with customers. This kind of user is the official representative of the company and organization and they are expressed as Officials in the experiment. Seniors and Juniors are ordinary users. Seniors we selected are users more than fifty-year old and Juniors are university students. Grass-roots are such users who publish a lot of interesting or useful microblogs that attract a great number of followers, and they become famous with the help of Micro-blog. Else consists of users that we choose randomly. After removing the time series the length of which is longer than 200 or shorter than 5, eventually, we got 792 valid Long Time Series and Short Time Series separately from 792 users to find out hidden temporal pattern.

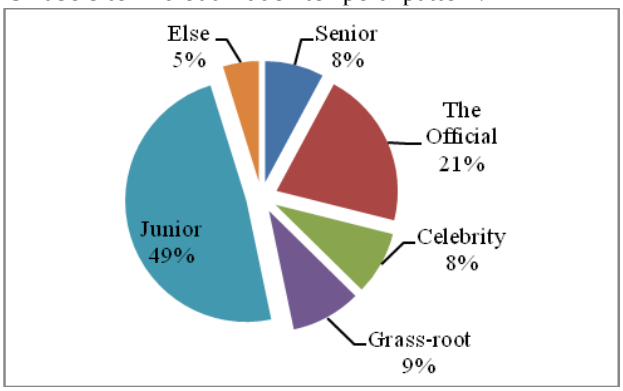

Figure 6. Proportion of each user group

We set sampling period $SP_{LTS} = 7\ days$ to form Long Time Series and $SP_{STS} = 2\ hours$, $T = 30$ to form Short Time Series. Fig. 7 illustrates the Cumulative Distribution Function (CDF) of microblog count of all

users. For most users, the total count of microblogs is distributed from 100 to 2000, demonstrating that most users are not very active in Micro-blog, and very few users post more than 10000 microblogs. Fig. 8 depicts the CDF of the length of LTS of all. LTS length of most users is distributed between 40 and 100 and only a few users utilize Micro-blog for a long time. The Pearson Correlation Coefficient of the two parameters, microblog count and LTS length, is 0.161, demonstrating that the relationship between the number of microblogs and participation time of users is not as high-correlated as that have been thought.

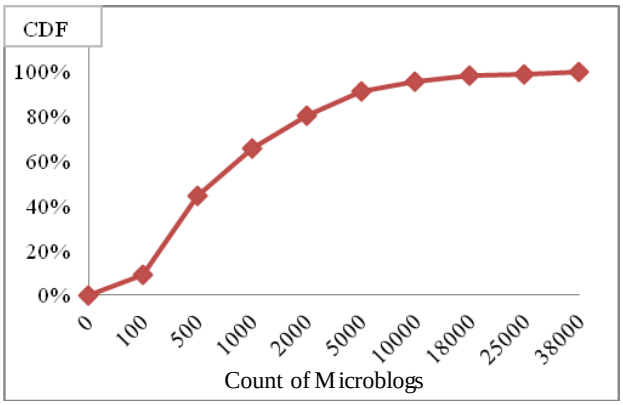

Figure 7. Distribution of microblog count of all users

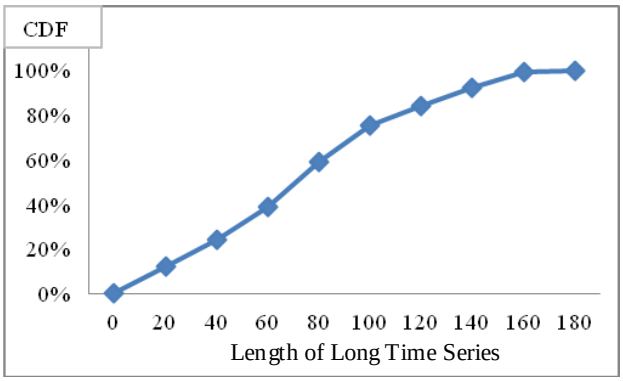

Figure 8. Distribution of the length of all LTS

B. Results of Long Term Pattern Discovery

1) Long Term Pattern

WT-DKM algorithm is employed on 792 Long Time Series to discover the Long Term Pattern of SNS user behavior. 4 patterns are derived shown as full curves in Fig. 9 and we regard them as the Templates of Long Term Patterns (T-LTP). The dotted curve represents one example member of this pattern. We can see the contour of these four patterns is various, but two curves in each subplot have similar overall shape, even enough they are unequal in length. The valley and peak of curves demonstrate the activeness of user participation. T-LTP1 displays random changes and there are dramatic ups and downs over all time. T-LTP2 goes down to the bottom in the former period and then rises up. T-LTP3 goes up to the peak in the middle while before and after remains flat with slight fluctuations. T-LTP4 rises up gradually and then keeps the flateau to the end.

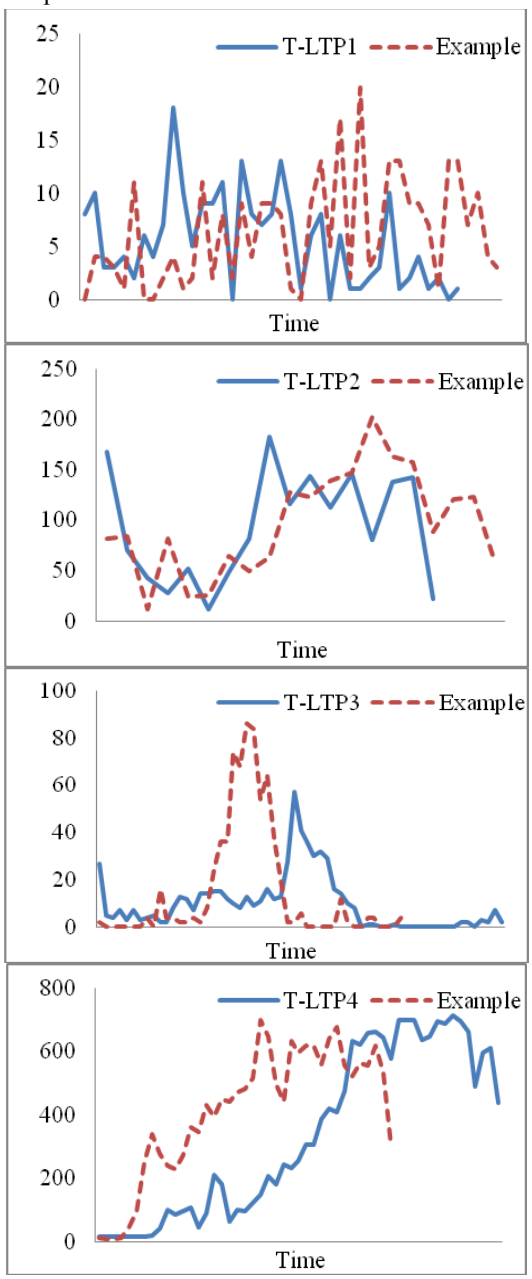

Figure 9. Long Term Pattern of Micro-blog user behavior. SP_{LTS}=7 *days*.

The two criterions, Hartigan Index and Silhouette Coefficient is used to decide how many patterns, K, there are when performing WT-DKM. However, they are not necessarily consistent as in Fig. 10, thus a compromise is required to select optimal K. We determine the optimal K is 4.

2) User group and LTP

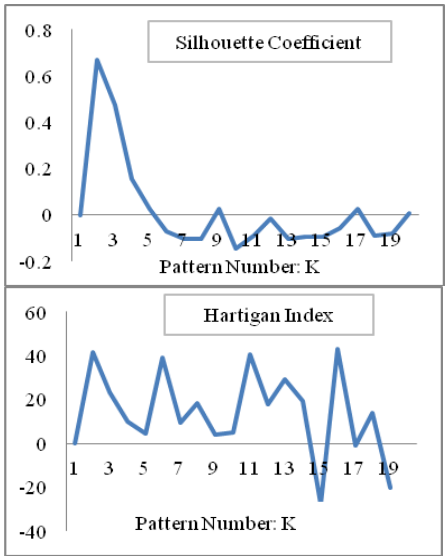

Figure 10. SC and HIwith varying K

We analyze how users of each user group are distributed in four patterns in Table I. The column is user group and the row is four long term patterns. Most Seniors performs as T-LTP1, which states random behavior without apparent overall trend, so as Juniors, Officials and Else. The magnitude of T-LTP1 is less than others, declaring that most average users are not very active. It is surprising that 48% Celebrities act as T-LTP3, for this pattern stands for users who have interest in Micro-blog in the intermediate period but do not pay attention before and after that. This may be caused by the events related to themselves, such as an actor advertising his new movies, inducing him to speak out. It is expected that quite many Grass-roots behave like T-LTP2 and T-LTP4, where the magnitude of pattern template is larger than others. Grass-root users are extremely active by posting numerous microblogs, because they are created to attract followers and then for further economic profits. Their microblogs are often funny, useful and retweeted by many followers. Once they become famous in Micro-blog, they will keep stay active, just as the variation trend of T-LTP2 and T-LTP4. For Official, it is a little complicated. 27% of Officials display variation as T-LTP2 that take advantage of Micro-blog to publish product or promote information and communicate with their customers. They may succeed to achieve their goals and then posting more and more microblogs. While 29% of Officials behaves like T-LTP3, these Officials are not very well-known. After some active time, they gave up Micro-blog as a media probably due to ineffectiveness of Micro-blog in advertising themselves.

TABLE I. PROPORTION IN EVERY PATTERN OF EACH USER GROUP

T-LTP	Senior	Junior	Celebrity	Grass-root	Official	Else
1	*0.63*	*0.84*	0.37	0.23	*0.41*	*0.84*
2	0.29	0.01	0.15	*0.31*	0.27	0.11
3	0.06	0.15	*0.48*	0.16	0.29	0.05
4	0.02	0.00	0.00	0.30	0.02	0.00

3) Wavelet scale and pattern number

The relationship of WT scale and the optimal pattern number is excavated indicating by SC and HI. When preprocessing time series with a high scale, we can get low resolution series, since more details are ignored and series tend to have similar overall shape with higher possibility. This may result a smaller optimal $\hat{K}$. We practice two cases with a high scale and a low scale separately. The results are shown as Fig. 11. It illustrates that the optimal $\hat{K}$ does not change much. SC falls down gradually with slight fluctuations while HI changes with dramatic rise and fall along with the increase of K. When K =4 and with a low scale, series tend to scatter more dispersedly as shown in Fig. 12, less series locate in the biggest cluster than with a high scale.

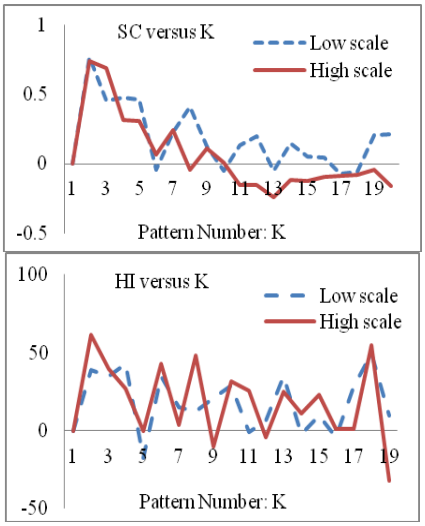

Figure 11. WT scales and corresponding clustering criterion index

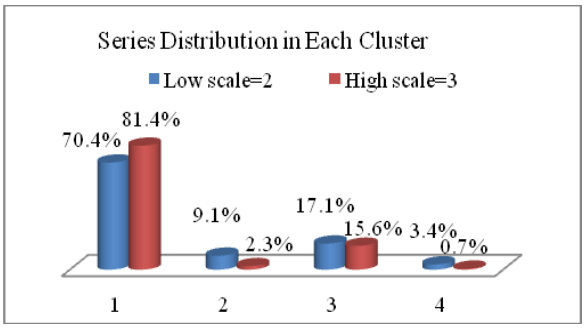

Figure 12. Series distribution with different WT scale. *Series tend to scatter more dispersedly when scale is low.*

C. Results of Daily Pattern Discovery

1) Daily Pattern and user group

DKM algorithm is employed on 792 Short Time Series to discover the Daily Pattern of Micro-blog user behavior. 5 patterns are derived and shown in Fig. 13 and we express them as the Templates of Daily Patterns (T-DP). The overall trend of T-DP1 and T-DP2 are almost same where most microblogs are posted in the later period of a day, except the difference in magnitude. T-DP3 fluctuates dramatically, stating that this kind of user is posting microblogs all day along. The curve of T-DP4 goes down

to the bottom in the middle but rises up before and after, indicating this kind of user is active in the afternoon and at night. T-DP5 displays a gentle fall and rise in the middle of a day, and the magnitude of the curve is the largest of the five patterns. All of patterns show significant activeness from afternoon to midnight.

How users in the same user group are distributed in each pattern is analyzed in Table II. Most users in every group act as T-DP2, especially Seniors, Juniors and Else, where the total number of microblogs published per day is the least. This explains that most users are not so dependable on Micro-blog that they only post several microblogs within 30 days. 28% of Seniors display like T-DP1 where microblogs count falls down after 20 o'clock, only second to T-DP2, consistent with the fact that aged people are used to sleep early. 28% of Celebrities and Officials behave like T-DP3, demonstrating that they post microblogs during the whole day. This result complies with the fact that Micro-blog is taken by such users to raise fame or publish information. As for Grass-roots, 23% of them act like T-DP5, extreme active during the night. They publish a huge number of microblogs per day, for the reason that they are created to attract followers by transferring and sharing entertaining information.

2) Relationship between Long Term and Daily Pattern of user behavior

We analyze how the relationship is between Long Term Pattern and Daily Pattern by calculating proportion of users who display one LTP and one DP simultaneously. It reveals that 51.7% of all users display as T-DP2 and T-LTP1 simultaneously, which explains that most users are behavior randomly and do not appear in Micro-blog frequently. It is interesting to find out that no users of T-DP5 display as T-LTP1 T-LTP3 for long term pattern. Although DP and LTP can be retrieved with same techniques, the practical meaning hidden behind is completely different and patterns are distinct. They are the two aspects expressing the human behavior and not necessarily correlated.

TABLE II. PROPORTION IN EVERY PATTERN OF EACH USER GROUP

T-DP	Senior	Junior	Celebrity	Grass-root	Official	Else
1	*0.23*	0.02	0.19	0.12	0.17	0.03
2	0.48	0.83	0.36	0.34	0.36	0.74
3	0.19	*0.15*	*0.28*	0.11	*0.28*	*0.18*
4	0.06	0.00	0.12	0.20	0.12	0.05
5	0.03	0.00	0.04	*0.23*	0.07	0.00

D. Future Work

We perform WT-DKM on time series derived from Micro-blog, however, this algorithm can be employed on other social network to discover temporal pattern of user behavior. The problem of WT-DKM is time and space complexity. The time and space complexity of WT is $O(n)$, while that of DTW is as high as $O(m*n)$, makes it not appropriate in large database. So, when implementing WT-DKM in realtime system, the complexity needs to be reduced. And there are a whole bunch of methods can be employed, such as constrained DTW and lower bounding measures [6].

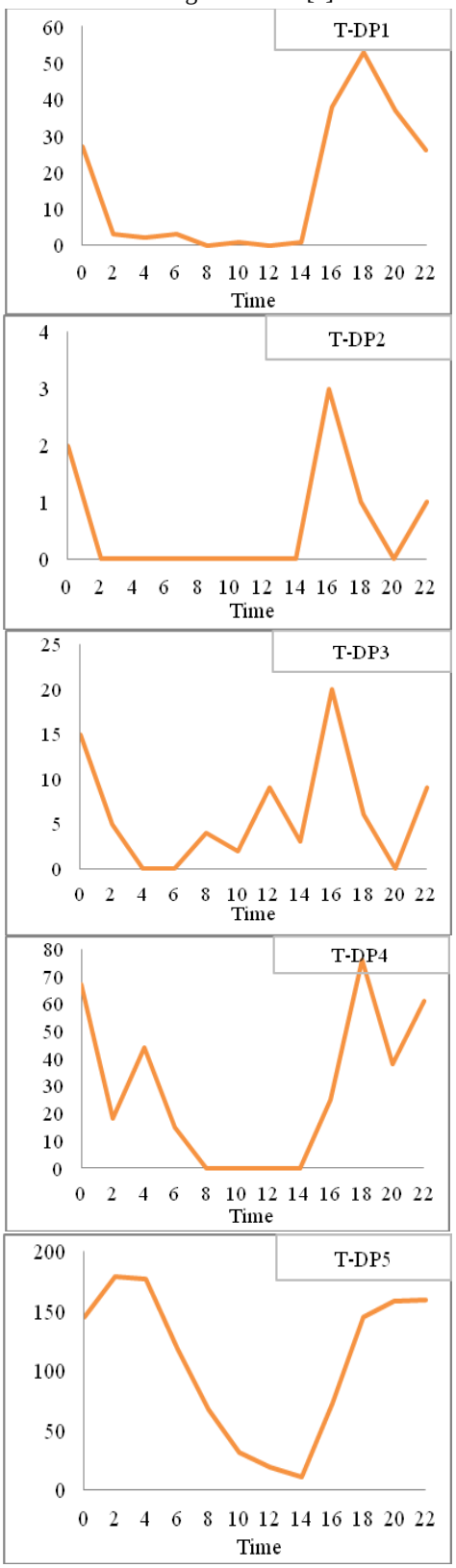

Figure 13. Daily pattern of Micro-blog user behavior. SP_{STS}=*2 hour* and T=30.

V. APPLICATION

Long Term Pattern stands for the rule of user behavior variation for a long time, which can also reflect users' attitude towards social network over time. If this service serves people well, users may spend more time and energy on it, such as posting more microblogs. Therefore, temporal pattern discovery is able to reveal the maturity and development of Micro-blog. For example, as we find that some Official users tend to abandon Micro-blog after some active time, it can be reasoned that this kind of service does not serve them well, thus suggestions can be proposed for Micro-blog operators.

Daily pattern of user behavior is more like the variation regularity related to users' living habit and lifestyle, such as when designing recommendation, researchers can take temporal factor into account, specifically, suggesting restaurant if one is used to checking Micro-blog at mealtime. This can make the recommendation more effective and accurate.

VI. CONCLUSION

In this paper, we study the Long Term Patten and Daily Pattern discovery of Micro-blog user behavior by Long Time Series and Short Time Series respectively.

We described how to create LTS and STS from social content. LTS records SNS user behavior for the moment they entered Micro-blog till the last action time. And STS captures microscopic variation of user behavior within a day. Then WT-DKM algorithm is proposed to divide LTS and STS into different clusters which correspond to long term and daily patterns. With the data derived from Sina Micro-blog, 4 Long Term Patterns and 5 Short Term Patterns are discovered. They have obvious overall trends, indicating the temporal regularity of user behavior.

Compared with studying user behavior from temporal aspect, we select users in six user groups to figure out how different user group shape the temporal pattern. From the perspective of Long Term Pattern, Grass-roots and Officials are becoming more and more active with longer active time and more microblogs than most average individuals, that are Junior and Senior and Celebrity users, who behave randomly without obvious regularity. Yet, quite a number of Official users stop posting microblogs after an active time. For Daily Pattern of Micro-blog user behavior, the active time stays from afternoon till midnight. Most individual users are not active and they post only several microblogs per day, such as Seniors and Juniors, while Officials and Celebrities tend to use Micro-blog all day along. Grass-roots are most active users in Micro-blog who post a huge number of microblogs per day.

The relationship between long term pattern and daily pattern is also analyzed. It reveals that most users behave randomly without apparent variation trend, and they are not dependable on Micro-blog and only post several microblogs everyday.

ACKNOWLEDGMENT

This work was supported in part by a grant from Nokia Research Center in China and Mobile Life & New Media. The authors wish to thank Ying Gao and Cheng Cheng for insightful discussions and suggestions.

REFERENCES

[1] Srivatsan Laxman, P. S. Sastry, "A survey of temporal data mining" in *Sadhana*, vol. 31, Part 2, April 2006, pp. 173-198.

[2] Sudhakar M. Pandit, Shien-Ming Wu, *Time series and system analysis, with applications*. New York: Wiley, 1983.

[3] Berndt, D. & Clifford, J. Using dynamic time warping to find patterns in time series. *AAAI-94 Workshop on Knowledge Discovery in Databases*. Seattle, Washington, 1994, pp: 359-370.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful," *in Proc. 7th Int. Conf. Database Theory*, 1999, pp. 217–235.

[5] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy. Finding patterns in time series: A dynamic programming approach. Advances in knowledge discovery and data mining. *American Association for Artificial Intelligence Menlo Park*, CA, USA, pp:229-248.

[6] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 2004, 7 (3), pp: 358-386.

[7] Michael T. Rosenstein and Paul R. Cohen. Continuous categories for a mobile robot. *AAAI '99/IAAI '99 Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, 1999, pp: 634-640.

[8] Ivan Popivanov and Renee J. Miller. Similarity search over time-series data using wavelets. *Proceedings 18th International Conference on Data Engineering*, Toronto, Canada, Feb., 2002, pp:121-221.

[9] Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435, May 12, 2005, pp: 207-211.

[10] E. Keogh and P. Smyth, A probabilistic approach to fast pattern matching in time series databases. *Proceedings of Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, California, 1997.

[11] Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. *Proceedings of the 4th ACM international conference on Web search and data mining*. New York, NY, USA, 2011, pp:177-186.

[12] T.Warren Liao. Clustering of time series data—a survey. *Pattern Recognition*, vol 38, Issue 11. Nov. 2005, pp: 1857-1874.

[13] Xiaoxi Du, Ruoming Jin, Liang Ding, Victor E. Lee and John H. Thornton Jr. Migration Motif: A spatial-temporal pattern mining approach for financial markets. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA, 2009, pp:1135-1144.

[14] Kaufman L, Rousseeuw P J. Finding groups in data: an introduction to cluster analysis. *Wiley series in Probability and Statistics*, November, 1990, New York.

[15] Xiang Li. Storm clustering for data-driven weather forecasting. *24th Conference on IIPS*.

[16] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. A. N. Amaral. A poissonian explanation for heavy tails in e-mail communication. *PNAS*, 105(47):18153–18158, 2008.

[17] Lipika Dey and Bhakti Gaonkar. Wavelet-based clustering of social-network users using temporal and activity profiles. *Proceedings of the 4th international conference on Pattern recognition and machine intelligence*. Springer-Verlag Berlin, Heidelberg, 2011, pp: 60-65.

[18] Michail Vlachos, Jessica Lin, Eamonn Keogh, Dimitrios Gunopulos. A wavelet-based anytime algorithm for K-

means clustering of time series. *SIAM International Conference on Data Mining*, 2003.

[19] Yu-Ru Lin, Sundaram H., De Choudhury M.,Kelliher, A. temporal patterns in social media streams: theme discovery and evolution using joint analysis of content and context. *IEEE International Conference on Multimedia and Expo*, Tempe, AZ, USA, Jun.,2009, pp: 1456-1459.

[20] Gabor Szabo, Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, Vol.53, Issue 8. New York, USA. Aug, 2010, pp:80-88.

[21] Hamidreza Zaboli, Mohammad Rahmati, Abdolreza Mirzaei. Shape Recognition by Clustering and Matching of Skeletons. Journal of Computer, Vol. 3, NO.5. May, 2008, pp:24-33.

[22] Juanying Xie, Shuai Jiang, Weixin Xie, Xinbo Gao. An efficient global K-means clustering algorithm. Journal of Computers, Vol.6, NO.2. Feb, 2011, pp: 271-279.

[23] Yu Zong, Ping Jin, Dongguan Xu, Rong Pan. A clustering algorithm based on local accumulative knowledge. Journal of Computers, Vol.8, NO. 2. Feb, 2013, pp: 365-371.

Chunhong Zhang, born in February, 1971, received her master and bachelor degree majored in Information and Communication System of Beijing University of Posts and Telecommunications in China in 1996 and 1993.

She is a lecturer in Beijing University of Posts and Telecommunications, China, since 1996. She has published three books such as Social Network (Beijing: 2011), Internet of Things (Beijing: 2010), and more than ten academic papers, 5 patents. She has engaged in three national programs and several enterprise projects. Her research interests include social network, data mining, network science and distributed system, etc.

Yaxi He, born in May, 1990, is a master student majored in Information and Communication System of Beijing University of Posts and Telecommunications since 2011, China. She received the Bachelor Degree of Communication Engineering from Harbin Engineering University, China in 2011.

Her paper "Principle Features for Tie Strength Estimation in Micro-blog Social Network" was published in the Proceedings of the 12th IEEE International Conference on Computer and Information Technology. Her research interests include data mining and social network service (SNS).

Yang Ji, born in April, 1972, received his doctor degree of majored in Information and Communication System of Beijing University of Posts and Telecommunications in China in 2002.

As a professor in Beijing University of Posts and Telecommunications in China, he has published dozens of articles in high-level conferences and journals. He has engaged in international programs, national projects and nature science foundation, etc. His research interest include mobile internet, ubiquitous network technology and internet of things, etc.