Physica A 413 (2014) 212–219

Contents lists available at ScienceDirect

Physica A

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

Human behavior during Flash Crowd in web surfing

Junshan Pan [a,b,*], Hanping Hu [a], Ying Liu [a]

[a] Key Laboratory of Image Processing & Intelligent Control of Education Ministry, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

[b] School of Computer and Information Science, Hubei Engineering University, Xiaogan 432100, China

HIGHLIGHTS

- We focus on human behavior during a special stage called Flash Crowd.
- Collective behavior in the FC stage tends to be more consistent and has self-similarity.
- The change of dynamic mechanism of behavior may be the origin of FC formation.

ARTICLE INFO

Article history:
Received 23 April 2014
Available online 8 July 2014

Keywords:
Human dynamics
Flash Crowd
Collective behavior
Individual behavior
Multiscale entropy

ABSTRACT

This paper focuses on human behavior in web surfing during the special stage called Flash Crowd (FC) period. Some statistical properties of human behavior are investigated. A moving approximate entropy (ApEn) method is provided to precisely locate the FC stage at first. Then the multiscale entropy (MSE) method is applied to study the difference of behaviors between the FC stage and the Normal stage. The lower entropy value may imply that collective behavior in the FC stage tends to be more consistent and follows a process with self-similarity. Further investigation by MSE and interval time distribution on the collective level and the individual level reveals that the origin of FC formation is not due to the increasing number of users, but more likely the change of dynamic mechanism of individual behavior.

1. Introduction

Revealing human behavior dynamics from big data of human activity has attracted increasing interests in the past decade. Evidences from various human behavior, ranging from E-mail communications [1,2], mobile communications [3,4], Web surfing [5–7] to MicroBlog visiting [8,9] have shown that human dynamics are non-Poissonian, with bursts of active behavior separated by long periods of inactivity, which leads to a power law heavy tail of interval time between two consecutive behaviors [10]. Previous researches were conducted either on the collective level or the individual level, but the behavior dynamics during different active periods are still not well understood.

In this paper, we focus on the human behavior during a special period called Flash Crowd (FC) period. FC is a large surge in traffic to a particular website causing a dramatic increase in server load [11]. Unlike network attacks, FC is an unintended phenomenon occurring as a consequence of collective reaction to a hot event. However, traffic fluctuation caused by FC is similar as that by distributed denial of service (DDoS) attack. Therefore, revealing human behavior dynamics in the FC period has important practical significance to network management.

* Corresponding author at: Key Laboratory of Image Processing & Intelligent Control of Education Ministry, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China. Tel.: +86 13886388791.
E-mail addresses: panjunshan@gmail.com, 2312798270@qq.com (J. Pan).

http://dx.doi.org/10.1016/j.physa.2014.06.085
0378-4371/

To uncover the underlying mechanism of human behavior during the FC period, we use data selected from World Cup 98 dataset which contains an FC. First, we provide approximate entropy (ApEn) to locate the FC stage from the request number sequence precisely. Then the multiscale entropy (MSE) method is used to analyze the difference of behaviors between the FC and Normal stages. Variation of entropy value between these stages may confirm the distinction of behavior dynamics. To reveal the origin of FC formation, we try to discuss the possible causes from the collective level and the individual level. After dividing the individuals into three groups according to their requests number of two stages, the distinction of behavior of the three groups in the FC stage is studied by MSE. We find that behavior in Group 2 contributes most to forming FC. Interval time distributions of request behavior from all individuals are further investigated. The majority of individual's interval time distributions in Group 2 are more likely to follow Generalized Pareto (GP) compared to the Generalized Extreme Value (GEV) in Group 3. Finally, we study the changing of individual's interval time distribution from the FC to the Normal stage in Group 2.

The paper is organized as follows. In Section 2, we describe the dataset used in our empirical analysis. The methods we used are introduced in Section 3. The empirical results of behavior dynamics in the FC stage are presented in Section 4. In Section 5, we discuss the possible causes of FC formation, and the final section gives the conclusion.

2. Data Description

Our data is collected from World Cup 98 dataset, which consists of all the requests made to the 1998 World Cup Website between April 30, 1998, and July 26, 1998, from 33 different World Cup HTTP servers at four locations. The website received 1,352,804,107 requests during 88 days [12]. The data we choose is from day 74 to 75 because there was a semi-final which conducted an FC traffic.

Since a webpage contains a lot of objects such as text, photos, videos, etc., the requests of all the objects in this webpage will be send to servers when a user clicks the link, which is, the logs record them all as long as a user clicks. Thus, we only take into account the logs with HTML file type as a valid request behavior. Interval time in our analysis is the time between two consecutive valid request behaviors. As a result, 4,893,381 requests performed by 239,705 users over a period of 48 h are selected as our empirical data. The resolution of the time is in seconds.

3. Methods

Approximate entropy. Approximate Entropy (ApEn) [13] which measuring the regularity of the signal can be useful to track qualitative changes in time series patterns, without precisely characterizing the generating system [14]. It has been applied to research fields including climatic revolution [15], clinical application [16], mechanical equipment fault diagnosis [17], etc. Here, we briefly describe the ApEn method [14]:

(1) Given a one-dimensional discrete time series $\{x_1, x_2, \ldots, x_N\}$, where N is the length of time series. $u_m(i)$ considered as the m-length vectors is reconstructed as

$$u_m(i) = \{x_i, x_{i+1}, \ldots, x_{i+m-1}\}, \quad 1 \le i \le N - m + 1.$$

(2) Define distance $d[u_m(i), u_m(j)]$ for vector $u_m(i)$ and $u_m(j)$ as

$$d[u_m(i), u_m(j)] = \max_{k=1,2,\ldots,m} (|x(i + k - 1) - x(j + k - 1)|).$$

(3) Let $n_i^m(r)$ represent the number of vectors that satisfy $d[u_m(i), u_m(j)] \le r$.

(4) $C_i^m(r) = n_i^m(r)/(N - m + 1),\ 1 \le i \le N - m + 1$

represents the probability that any vector $u_m(j)$ is close to vector $u_m(i)$.

Define $\Phi^m(r) = (N - m + 1)^{-1} \sum_{i=1}^{N-m+1} \ln C_i^m(r)$, represents the average of the natural logarithm of the probability.

(5) ApEn is estimated by the statistics,

$$\mathrm{ApEn}(m, r, N) = [\Phi^m(r) - \Phi^{m+1}(r)],$$

measures the likelihood that two vectors which are close remains close on next incremental comparisons.

Multiscale entropy. Costa et al. [18] introduced the multiscale entropy (MSE) measurement, which indicated the regularity of signal at multiple time scales. It was applied to measure the complexity of biologic systems using the time series of heat rates [19], human gaits [20] and flow rates of river [21]. The procedure of MSE is described as follows [18]:

(1) Given a one-dimensional discrete time series $\{x_1, x_2, \ldots, x_N\}$, consecutive coarse-grained time series $\{y^{(\tau)}\}$ is constructed determined by the scale factor τ. The original time series is divided into non-overlapping windows of length τ. Each element of a coarse-grained time series is calculated according to the equation,

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x_i, \quad 1 \le j \le \frac{N}{\tau}.$$

For scale factor one ($\tau = 1$), the time series $\{y^{(1)}\}$ is simply the original one.

(2) The sample entropy measure (S_E) is calculated for each coarse-grained time series $\{y^{(\tau)}\}$ as a function of scale factor τ.