Physica A 448 (2016) 196–204

Contents lists available at ScienceDirect

Physica A

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

The scaling of attention networks

Cheng-Jun Wang [a], Lingfei Wu [b,*]

[a] School of Journalism and Communication, Nanjing University, Nanjing 210093, PR China
[b] Center for Behavior, Institutions and the Environment, Arizona State University, Tempe, AZ 85281, United States

HIGHLIGHTS

- The flow of attention on Web forums can be modeled by networks, which explain the scaling relationships between clicks, users, and threads.
- Attention networks preserve self-similarity quantified by a new form of Zipf's law containing a parameter that changes with network size.
- Larger forums are more productive, as an average user will generate more clicks.
- Forum users seek for online interactions with off-line consequences.

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form 8 September 2015
Available online 31 December 2015

Keywords:
Attention dynamics
Complex networks
Online behavior
Scaling
Power law

ABSTRACT

We use clicks as a proxy of collective attention and construct networks to study the temporal dynamics of attention. In particular we collect the browsing records of millions of users on 1000 Web forums in two months. In the constructed networks, nodes are threads and edges represent the switch of users between threads in an hour. The investigated network properties include the number of threads N, the number of users UV, and the number of clicks, PV. We find scaling functions $PV \sim UV^{\theta_1}$, $PV \sim N^{\theta_3}$, and $UV \sim N^{\theta_2}$, in which the scaling exponents are always greater than 1. This means that (1) the studied networks maintain a self-similar flow structure in time, i.e., large networks are simply the scale-up versions of small networks; and (2) large networks are more "productive", in the sense that an average user would generate more clicks in the larger systems. We propose a revised version of Zipf's law to quantify the time-invariant flow structure of attention networks and relate it to the observed scaling properties. We also demonstrate the applied consequences of our research: forum-classification based on scaling properties.

1. Introduction

Attention economics was firstly proposed by H. A. Simon et al. in 1971 to address the conflict between the scarcity of human attention and the explosive growth of information [1]. In the past decades, with the development of media technology such as the World Wide Web and smartphone, it is increasingly easier to copy and distribute information, thus the task of developing a quantitative framework for attention dynamics becomes more urgent. Fortunately, the availability of online data sets allows scholars to investigate the dynamics of attention empirically. For example, studies have been conducted to investigate the competition for user attention between emails [2], news [3], tags [4], and tweets [5–7].

* Corresponding author.
E-mail address: wlf850927@gmail.com (L. Wu).

http://dx.doi.org/10.1016/j.physa.2015.12.081
0378-4371/

A conclusion already drawn is that the competition usually leads to a long-tail distribution of attention allocated on information resources [8,2]. However, how this distribution evolves in time and how the evolution shapes the growth of online information systems remain unknown.

To address these questions, we collect clickstream data from a group of Web forums and use clicks as a proxy to study the dynamics of attention. We construct networks for each of the top 1000 forums in Baidu Tieba, a very large Chinese online system containing many forums of diverse topics. In these networks, nodes are threads and edges represent "user-flow", i.e., the switch of users between threads in an hour. We call these networks "attention network" because they show how the collective attention of users is transported between threads within a forum. Several limitations of our study should be noted. Firstly, using clicks as an approximation of attention helps us focus on the most essential and traceable activity in the cyber-world, but the information on the duration of visits is lost. Therefore, we have to assume that all threads that receive the same number of clicks are equally popular. Another limitation is that we only study the clickstreams between threads in forums, which are not constrained by the tree structure of posts within threads, but are still strongly affected by the displaying order of threads on the homepage of forums.

We find that these networks exhibit the same scaling properties as discovered in real-world flow systems including rivers [9] and circulatory systems [10]. These scaling properties are usually viewed as a signature of the time-invariant flow structure [11]. We propose a revised version of Zipf's law to describe the effect of system size on distribution function in flow systems. We find that the discussed scaling properties can be derived from such a function, which means that the discovered hidden connection between size and structure may widely exist in various flow systems.

These findings help us understand why we may spend more time on larger forums: most of the studied attention networks preserve a self-similar flow structure such that when it becomes larger, it is more difficult for a random walker to leave from the system. This pattern is not only interesting by itself, but also has strong applied consequences. As a demonstration we group the studied forums into 20 categories according to their topics and try to relate the discussed self-similar flow structure with the topics of forums. It turns out that the forums in the "Cities" category are the most attractive, whereas the "Art" forums are the least attractive. Our clickstream-based analysis not only provides development suggestions for the Tieba system, but also presents a general method that can be easily applied to many other online social systems.

2. Materials and methods

2.1. Data source

Baidu (http://www.baidu.com/), a leading Chinese language search engine, is the fourth largest site in the world according to Alexa (http://www.alexa.com/topsites). Tieba (http://tieba.baidu.com/) is a big forum systems supported by Baidu. In this system, users search a forum by typing a keyword within the system, and if the forum has not been created yet, it will be created upon the search. This novel design successfully kept search traffic within the website and effectively motivated users to contribute. As of 2014 more than eight million forums are created (mostly by fans). The topics of these forums cover popular stars, movies, comics or books. More than one billion threads are created in these forums.

In our research we collect the browsing activities of users on the top 1000 forums, whose daily traffic (the number of unique visitors) varies from thousands to millions. For each forum, we construct 1440 successive hourly-based attention networks in two months (from Feb. 27, 2013 to Apr. 27, 2013). See Fig. 1 for the construction of attention networks in detail.

2.2. Clickstream networks and key variables

Fig. 1 presents an example Baidu Tieba clickstream network, whose nodes are threads and edges represent the switch of users between threads. In the annotation of Fig. 1 we introduce how to construct clickstream networks from users' browsing records. We firstly divide the entire data set into hourly pieces and then sort each piece by cookies (the unique and permanent labels used by a website to identify users). After that, we select all successive pairs of threads visited by the same user and add corresponding edges in the clickstream network. Sorting data by cookies guarantees that a user would not be repeatedly counted even if he/she entered and left the system more than once during an hour.

After two artificial nodes "source" and "sink", which represent the "environment", are added to the clickstream networks, they satisfy "flow conservation", i.e., inflow equals outflow both on the network level and also on the node level. As a result, the network properties PV (the total number of clicks) and UV (the total number of users) can also be expressed as the sum of flow passing through node $i(T_i)$ and the sum of flow dissipated by node i to "sink" (D_i), respectively:

$$UV = \sum_{i=1}^{N} D_i, \tag{1}$$

$$PV = \sum_{i=1}^{N} T_i. \tag{2}$$