2H5-OS-15b-1

Twitter時系列における情報のカスケード

Information Cascade in Twitter time series

岡瑞起[*1]　池上高志[*2]
Mizuki Oka　Takashi Ikegami

[*1]筑波大学、[*2]東京大学
[*1]University of Tsukuba, [*2]the University of Tokyo

We crawled and analyzed the Twitter time series for one year, showing that Twitter dynamics is characterized by two modes; reactive and default modes. Without having any real world events, we see that Twitter can be driven by the inherent dynamics. We characterized it by computing transfer entropy between time series. We found that the information is transferred depending on the time scale. In this talk, the complexity of such information transfer among time series with different nouns will be reported.

1. Introduction

今回の発表では、Twitter を例にとって、ウェブの自律的な運動を特徴つける。Twitter 時系列を 1 年間にわたってクロールし、それを Twitter に含まれる時系列に関して分類すると、そこにはいくつかのパターンが見て取れる。ここでは、移送情報エントロピーを用いて、これらの時系列の間に流れる情報量を定量化し、Twitter が全体としてどのような内在的な運動を持っているかを明らかにする。

多くの研究では、Twitter の時系列を解析することで背後の人間の行動の特徴をあぶり出すような研究が多い。しかし、多くのメディアがそうであるようにメディアが巨大で複雑になるにしたがって、そこにはある自律性が創発されるだろうと期待される。例えば、Twitter 自身が典型的に持つ「短期記憶」の長さや、どういうことに反応するかといった「応答性」がそれである。ここでは、Twitter における「自律性」を、移送情報エントロピーを用いて、抽出しようという試みである。

2. Information transfer

ここでは、どのくらいの情報が 2 つの時系列間で移送されたかを計算する方法を提案する。つぎに、それを時系列を構成する最小時間単位と関係させて議論する。

2.1 Transfer entropy

あるパターン x の生成確率を $p(x)$ としたとき、そのシャノンエントロピーは以下の様に H(x) で定義される。ここで、たとえば x は部分時系列パターンとするのが普通である。

$$H(X) = -\Sigma_{x \subset X} p(x) log_2 p(x).$$

この表記をもちいて、2 つの時系列 X と Y のあいだの相互情報量 (MI) は次のように定義できる。

$$MI(X,Y) = H(Y) - H(Y|X),$$

ここで $H(Y)$ は時系列 Y のもつ不確定性であり、$H(Y|X)$ は X を知った時の Y のもつ不確定性である。相互情報量は、X と Y に関して対称であるが、X から Y への情報量 $TE(X \to Y)$ を次のように定義することで、時間的に非対称なエントロピーが定義できる。これを移送情報エントロピーという。

$$TE(X \to Y) = H(Y_{t+s}|Y_t^{(d')}) - H(Y_{t+s}|Y_t^{(d')}, X_t^{(d)}),$$

ここで $X_t^{(d)}$ と $Y_t^{(d')}$ はそれぞれ d と d' だけ過去に遡った時系列を与える。この $TE(X \to Y)$ を、異なる 2 つの単語を含んだ Twitter 時系列間で計算し、どのように情報が流れているかを計算する。

2.2 Transfer entropy and Δt

Twitter 時系列は、最小の時間単位 Δt を 1 分として時系列を構成するのであれば、ツイートされたかどうかの疎な時系列が得られる。興味深いのは、この最小単位の取り方によって、時系列のパターンの異なる特徴が際立って見えるということだ。ここで、次のようにして時系列 $x_1(n)$ を、もとの時系列 $x_0(s)$ から生成し、生成された時系列に関して情報エントロピーを計算する。

$$x_1(n) = \int_{(n-1)\Delta t}^{n\Delta t} x_0(s) ds.$$

具体的には、Δt は、1 分から 1024 分まで変化させて行う。

3. Results and Analysis

Oka and Ikegami [1]2 で示されているように、Twitter の時系列を単語ごとに分類し、それらの間の移送情報エントロピーを計測した。これから次の 2 つのことが分かった。

1. Δt に対し、移送情報エントロピーはあるところで極大値を持つ傾向にある（図 1)。
2. 頻度ランク順位で上位の単語の時系列は、ほかの時系列に対して「情報の上流」となっていることがわかる（図 2)。

1) に関し、たとえば、図 1 に示すように「今日」という単語は、ランダムなランクから選んだ 46 個の単語に関し、ほとんどの場合 Δt が 60 分前後にピークを持ち、逆に、それらの単語から「今日」という単語へは、2〜4 分あたりにピークを持っている。しかし、ランクが下がるにしたがって、この傾向は逆転する。また、流れるエントロピーの量も上位から下位に流れるほうが、その逆よりも大きいことが見て取れる。

連絡先: 岡瑞起: mizuki@cs.tsukuba.ac.jp

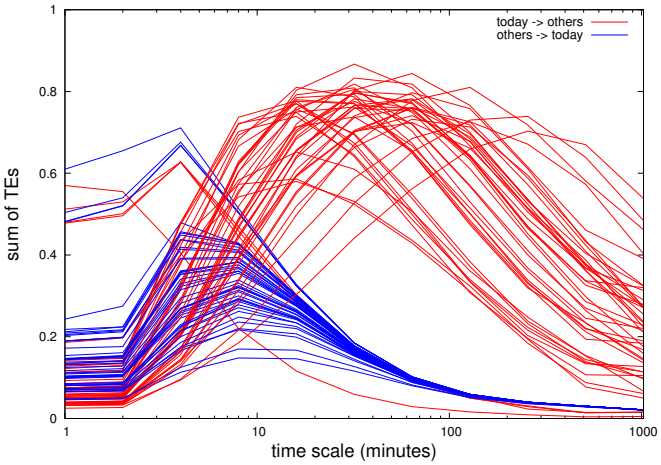

図 1: **Integrated TE of the keyword "today" to/from the other 45 keywords by varying the time resolution Δt from $2^0(=1)$ minute to $2^{10}(=1024)$ minutes.**

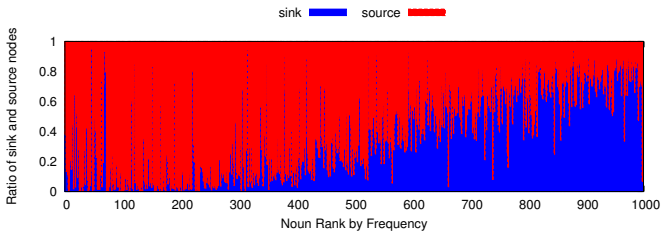

図 2: **The role of each keyword.** (Top) The ratio of keywords becoming sources and sinks shown as a function of keyword frequency over time. Red shows the source ratio and blue shows the sink ratio, as a function of keyword frequency over time. The frequent keywords tend to become source nodes and infrequent keywords tend to become sink nodes.

2) そこで Δt を 60 分として、1000 個の単語間のに関し、相互に移送情報エントロピーを計算する。このとき書く単語を、source（他の単語に対して情報の湧き出しとなっている）と sink（他の単語に対して情報の沈み点となっている）というタグつけを行う。その結果、図 2 が示すようにランクの高い単語ほど source になりやすいことがわかる。

4. Discussion

乱流現象を、情報の source と sink のネットワークで特徴つけられると書いたのは、Robert Shaw [2] である。ここでは Twitter の乱流状態を、移送情報エントロピーの source と sink のネットワークとしてとらえた。また、ランク順位の高い単語から低い単語に情報が流れてゆく、情報のカスケードがみられるのではないか。ただし、ランク分布の中間領域では順序は確定せず、複雑なカスケード構造にみえる。こうした情報の乱流、カスケードなどの概念をもとに、今後の研究では Twitter に代表されるウェブのダイナミクスを自律的なダイナミクスを特徴付けていきたい。

参考文献

[1] Mizuki Oka and Takashi Ikegami, "Exploring Default Mode and Information Flow on the Web", PLoS One (in press), 2013.

[2] Robert Shaw, "Strange Attractors, Chaotic Behaviour and Information Flow", Zeitschrift fur Naturforschung, 36A, pp.80-112, 1981.