数理解析研究所講究録 第2102巻 2019年 87-90

The Analysis of Big Data and Applications of Wavelets

Kiyoshi Mizohata

Abstract. The amount of social media data is now growing exponentially. Such data is now called Big Data. In this paper, we shall show several interesting results obtained by the wavelet analysis of Nico Nico Douga (famous social media in Japan) which is a typical example of Big Data, using Hadoop distributed file system.

Keywords. Wavelets, Big Data.

1. Introduction

We are now living in the world where the amount of the data set is increasing exponentially. Such data set is now called "Big Data". In this paper we first show how to deal with Japanese Big Data. Because the size of Big Data is very big, we must use Hadoop distributed system to analyze it. But in this case, to deal with data written in Japanese, we must be careful because Hadoop distributed system can't understand Japanese sentences. Next, we shall explain several interesting results obtained by the wavelet analysis.

2. Pre-processing of Japanese Data

In this investigation, we analyze comments of Nico Nico Douga (famous social media in Japan). Nico Nico Douga is famous video sharing website in Japan managed by Dwango. Users can upload video clips. Many comments can be overlaid directly onto the video by many viewers. So comments of video clips are one of the most famous Big Data in Japan.

To investigate this Big Data (300GB) by Hadoop distributed system, pre-processing of Japanese data is required since Hadoop system can not analyze Japanese words directly. Japanese words must be reformed by Mecab, an famous open source morphological analyzer for Japanese nouns, verbs and adjectives. By using Mecab, we can do pre-processing of Japanese data by Hadoop system.

3. Results by Hadoop distributed system

Investigations of the number of comments of this Big Data by using Hadoop distributed system lead us to very interesting results. In this paper we show one typical example, concerning to the musician A. A is now one of the most famous musician in Japan. (A is, of course, pseudonym) We want to know a turning point of A's life by the comments of Nico Nico Douga.

Let us find the number of comments related to the musician A. By counting comments with Hadoop distributed system ([1]) after pre-processing, we obtain the following comment data. See Figure 1.

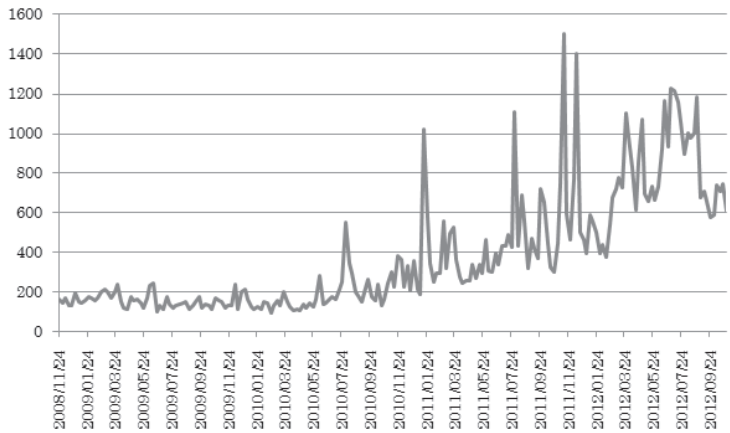

Figure 1. The number of comments by week.

We decompose this data using D_2 wavelets.([2]) Denote by H_1, high frequency part of the data and by L_1, low frequency part of the data.

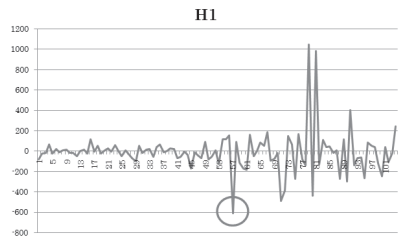

Figure 2. H_1 data.

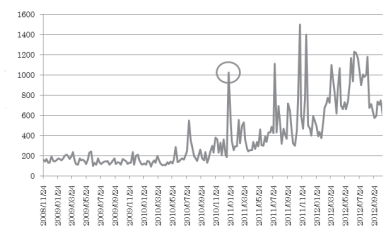

Figure 3. Turning point of data by H_1.

The lowest value of H_1 data corresponds to a turning point. See a small circle in Figure 2. In Figure 3, we also put a circle to a turning point. More interesting results can be found by decomposing L_1 to H_2 and L_2.

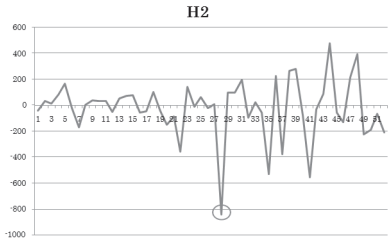

Figure 4. H_2 data.

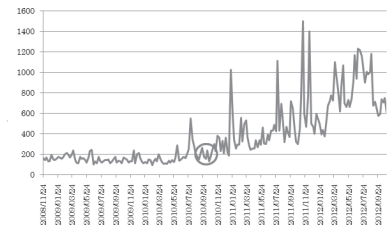

Figure 5. Turning point of data by H_2.

It is obvious that a turning point of data by H_2 data (circle in Figure 5) is important. This is a turning point of A's life. By analyzing A's turning point more precisely, we can find very interesting result. By analyzing other famous people's data, we can find more interesting results. Let us consider the number of comments related to the musician B.

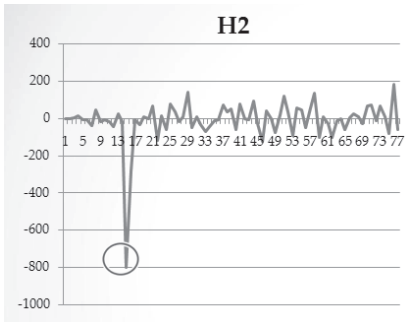

Figure 6. H_2 data of B.

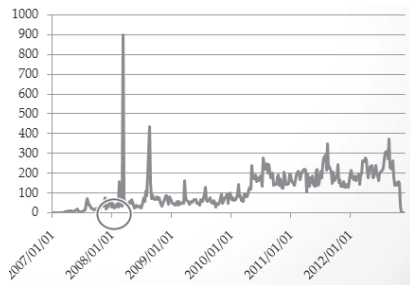

Figure 7. Turning point of B by H_2.

According to H2 data of B, their turning point is January 2008. Indeed, it was a time when their CD was released and their popularity boomed. Next, we examine famous actors and actress

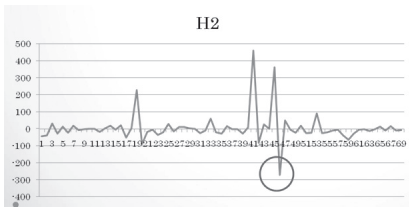

Figure 8. H_2 data of C.

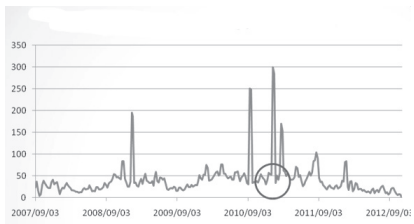

Figure 9. Turning point of C by H_2.

By analyzing H2 data of the actor C, the lowest point is February 2011. At that time, he appeared in "Doraemon" (the most famous Japanese animation).

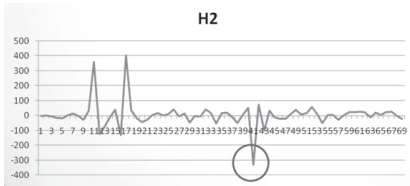

Figure 10. H_2 data of D.

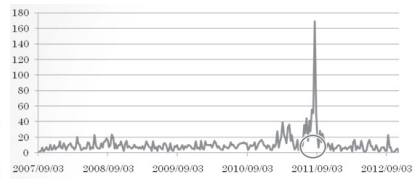

Figure 11. Turning point of D by H_2.

The lowest point of H2-data of the actress D is about October 2011. At that time, she played the lead in a major TV drama and became the center of attention.

4. Conclusions

These results show that wavelets are strong tools to analyze Big Data. Using wavelets, we can detect important edges of data, and also turning points of a person's life. But for several cases, we can't detect so called "turning points". More careful analysis of comments must be required. A more difficulty is that It spends a lot of time to analyze Big Data by Hadoop system. More efficient algorithm must be also required.

References

[1] Tom White, *Hadoop: The Definitive Guide.* O'Reilly Media, 4th edition, 2015.

[2] Ingrid Daubechies, *Ten Lectures in Wavelets.* SIAM, 1992.

Kiyoshi Mizohata
Miyakodani
Doshisha University
kyoutanabe-shi Kyouto-hu
610-0294 Japan
e-mail: kmizoha@mail.doshisha.ac.jp