Waveform feature extraction and similar waveform detection by Locality Sensitive Hashing

*KOMAGATA Ryota[1], Shiro Hirano[1], Hironori Kawakata[1], Makoto Naoi[2]

1. Ritsumeikan University, 2. Kyoto University

Even tiny waveforms due to uncataloged earthquakes inform us of the characteristics of an earthquake source, seismic activity, and underground structures. A frequently used method to detect them is the template matching based on a correlation coefficient (CC) among known templates and continuous waveforms (e.g., Doi & Kawakata, 2012; Kato *et al.*, 2012), which is nothing but to find lookalikes of known waveforms. However, if possible, a similarity analysis among all pieces out of the continuous waveforms may enable us to detect more seismic waveforms due to small events.

In this study, we focused on a hashing method without templates to detect similar waveform pairs quickly. Yoon *et al.* (2015) proposed a similar waveform detection method called Fingerprint and Similarity Thresholding (FAST) using Fingerprinting, which is a kind of Locality Sensitive Hashing (LSH) and based on Waveprint proposed for voice search by Baluja & Covell (2008). FAST consists mainly of two components: extraction of waveform features based on an LSH hash function and similarity search based on approximation of the Jaccard coefficient, which is the similarity between the hash values of two waveforms. By using Fingerprinting, many pairs of similar and uncataloged small waveforms can be found. However, FAST needs complicated processing such as spectrogram calculation and a wavelet transform. Thus we proposed two new hash functions with fewer procedures and detected similar waveform pairs by using them and FAST. We compared the performance focusing on the relationship between the similarity of the detected pairs and the CC value and runtime.

Considering that the amplitude of the seismic waveform generally has higher amplitudes than that of the surrounding microtremors, we designed two time-domain hash functions: 2bit-aHash, modified from aHash for image detection by Fei *et al.* (2015) and *k*Hash, a novel method we proposed in this study. 2bit-aHash assigns binaries of "10" and "01" to each sample of a waveform above and below 1 σ, respectively, deviating from the moving averaged continuous waveform , and assigns "00" to the others. *k* Hash assigns binaries "10" and "01" to the top-*k*% amplitudes of the absolute value in the 5-seconds waveform window depending on the sign (i.e., "10" and "01" for the positive and negative outliers, respectively) and assigns "00" to the others.

We analyzed a 24-hours continuous velocity seismogram of the Hi-net Matsumoto Wada station, Japan, of June 29-30, 2011, where an *M*5.4 earthquake and its fore- and aftershocks occurred. As a result, the employment of 2bit-aHash and *k*Hash succeeded in detecting uncataloged similar seismic signal-like waveform pairs. For 2bit-aHash and *k*Hash, each pair showed high CC values, and we observed a good correlation between the similarity and the CC values, which was absent in the case of FAST despite its complicated processing. Also, the overall runtime of both 2bit-aHash and *k*Hash was about 4 to 5 times shorter than FAST, which is not only because their feature extraction is straightforward but also because the similarity search based on them was several tens times faster than FAST. The runtime of similarity search approaches $O(n^2)$ when the data length n becomes very long, so 2bit-aHash and *k*Hash may contribute to a significant improvement in the runtime of similarity searches. In the similarity search, 2bit-aHash and *k*Hash yields extremely low Jaccard coefficients for pairs including noise, which enables to delete dissimilar pairs efficiently from the database that holds many candidates of pairs based on the

Jaccard coefficients, while FAST still holds an enormous amount of noise-noise and signal-noise pairs as candidates. Given the above features, the novel two functions, 2bit-aHash and *k*Hash, have shorter runtime than FAST not only for feature extraction but also for similarity search and have higher CC values between the detected waveform pairs.

Keywords: Earthquake event detection, Waveform similarity, Similarity search, Hash method

