

大規模センサデータに対するインタラクティブ性を有する 可視化システムの開発

酒井 健[†] 山田 竜平^{††} 山本 幸生^{†††} 横山 昌平^{††††} 木谷 友哉[†]

[†] 静岡大学院総合科学技術研究科情報学専攻 〒432-8011 静岡県浜松市中区城北三丁目5-1

^{††} 会津大学コンピュータ理工学部 〒965-8580 福島県会津若松市一箕町上居合90 大字鶴賀字

^{†††} 宇宙航空研究開発機構宇宙科学研究所 〒252-5210 神奈川県相模原市中央区由野台3丁目1-1

^{††††} 首都大学東京システムデザイン学部 〒191-0065 東京都日野市旭が丘6丁目6

E-mail: [†]sakai.takeru.17@shizuoka.ac.jp, ^{††}ryamada@u-aizu.ac.jp, ^{†††}yukio@planeta.sci.isas.jaxa.jp,
^{††††}shohei@tmu.ac.jp, ^{†††††}tt-kitani@inf.shizuoka.ac.jp

あらまし 近年, IoT 機器の普及により, センサーデータを取得し, 分析する機会が増加している. そのため, 分析者がセンサーデータから波形をすばやく見つけ出すためのインタラクティブ性を有する可視化システムの必要性が高まっている. 本研究では, センサーデータをユーザが必要とする形で可視化するシステムを開発した. 本システムでは膨大な量の月地震データを用いたため, サーバとクライアント間でデータの送受信やクライアントでの操作に遅延が発生する問題がある. その問題を解決するために, 効率よくデータ量を削減する手法を提案した. また, 評価実験により従来の一般的な手法と処理速度と元データとの類似度を比較することでその有効性を示した.

キーワード SciDB, 時系列, 可視化

1 はじめに

近年 IoT 機器の普及により, 身近にセンサーデータを取得することが可能である. 膨大なセンサーデータを解析することは重要であり, センサーデータを解析する研究が増加している.

我々の研究では月に設置された地震計から取得したセンサーデータを扱っている. そのデータを解析することで月で起こる地震(月震)の頻度や要因, 震源, 月の活動度などが知見として得られる. そのため我々の研究では膨大な月震データの中から月震が起こった波形を探し出すことが必要不可欠となっている. しかしながら, 月震には地球における地震に用いられる解析手法を直接適用することが難しい. それは月と地球の地殻では物質や構造が大きく異なるためである. その結果, 月震データが取得されてから40年経った現在も未だに発見されていない月震や震源が特定されていない月震などが存在する. 膨大な月震データの中から月震の波形を探し出すことは非常に困難であり如何に効率よく波形データを閲覧できるかが課題となっている.

本研究では, 研究者が効率よく月震データを閲覧できるような Web システムを開発した. 研究者であるユーザには, 月震の波形を繰り返してユーザの指定した多くの区間を閲覧するニーズがある. したがって, ユーザが波形をすばやく見つけ出すためのインタラクティブ性を有した閲覧システムを構築する必要がある. また, インターネットを介した閲覧システムであるため, ユーザごとに異なる研究環境であったり, サーバとクライアント間の通信のタイムラグなどが問題として考えられる. これはサーバとクライアントで送受信されるデータ量を削減することで解決することができる. 画面に可視化できるな最低限の

データ数へと削減することで, ブラウザへの負荷も軽減することができる. 本研究では, このようにサーバから送信しクライアントで可視化するデータ量を削減することを”間引く”と定義する.

本研究では, サーバ上でデータを間引き開発した閲覧システムで波形の可視化を行う. データの間引きにおいて, 間引き前と間引き後の波形で波形の見た目を変えることなく可視化することが課題である. そこで本研究では, 元の波形と見た目が変わらないように近似させる間引き手法を提案する.

本研究で扱うデータは NASA のアポロミッションによって設置された ALSEP(Apollo Lunar Surface Experiment Package)という観測装置群に含まれる地震計を用いた月面上での月震の連続観測実験 PSE(Passive Seismic Experiment)によって取得されている. その実験により約7年半(1969~1977)にわたる月震記録が得られ, 月震が初めて観測されたことにより月震の存在が明らかになった. これらのデータは長周期地震計の加速度 XYZ 軸成分や短周期地震計の加速度 Z 軸成分などによって構成されている. 観測から40年経った現在でも解析が行われている月震唯一のデータとなっている. 最近の研究でも新たに未発見の月震が確認された[8]. また, 月震波形には震源ごとに類似性があることが研究[9]によって明らかになっており, その類似性に着目して月震の震源を分類する研究などが存在する. 月震には分類された震源ごとに同じラベルが付与されており, 月震イベントカタログとして公開され現在も更新が続いている. このような分類結果の多くは専門家の目視によってラベル付与が行われている. もちろん機械学習などを用いた計算機によるラベル付与も行われているが, それらの教師データには専門家の目視による分類, ラベル付与結果が用いられている.

そのため今後、機械学習などによる分類をより高精度なものにするためにも専門家の目視による波形分類が必須である。本研究は専門家の目視による波形の閲覧をより効率よく行える環境を作ることを目的としている。

また、大規模なセンサーデータを格納しておくデータベースとして、本システムでは SciDB という配列型データベースを利用した。一般にセンサーデータなどの時系列データを関係データベースシステムで扱うのは困難である。関係データベースシステムでは時系列のようなデータの順序に意味のあるデータを順序通りに処理する保証はない。順序通りに処理するには ORDER BY を指定する必要があるが、大規模データに対し逐一 ORDER BY を指定するのは大きなコストとなる。近年ではこのような順序付けデータを専門に扱うデータベースシステムが多く開発されており、SciDB もそのひとつである。SciDB は Paradigm 社²によって開発されたアレイデータベースシステムであり、順序付けされているデータ、例えば遺伝子配列データ、センサーデータ、地理空間データなどの扱いを得意とするデータベースである。また、SciDB は分散処理や並列処理を行うアーキテクチャによってスケールアウトすることにより、より多くのデータを格納しアクセスできるという特徴がある。そこで本研究では、月震センサーデータを SciDB に格納することでデータへのアクセスの効率性を高める。

本論文は以下の構成に従う。2 章では本研究に関連する研究について述べる。3 章では提案システムについて述べる。4 章ではデータを間引くための提案手法について述べる。5 章では提案手法の有効性を示すための評価実験と考察について述べる。6 章ではまとめと今後の課題について述べる。

2 関連研究

月震の分類に関する研究として、Nakamura [6] の研究が挙げられる。月震波形に対して相互相関係数を用いた最短距離法による階層的クラスタリングを行い、その結果に基づいて人手による月震分類を行っている。この研究による分類が現在の月震研究の基準となっている。専門家の目視による月震波形の必要性が示唆されている。

月震センサーデータを可視化するシステムとして山田らが開発した moon seismic monitor [1]¹がある。moon seismic monitor はユーザが時刻を入力することで月震波形を閲覧、取得することのできるシステムである。しかしながら、このシステムは可視化された波形の拡大、縮小が困難な上、波形データも 1 時間、3 時間、6 時間といった一定間隔でしかデータを閲覧することができない。これは波形データが画像データとして可視化されているためである。また、データの取得においても 1 時間分のデータしかダウンロードすることができない。そこで本システムでは波形データの拡大、縮小をユーザが自由に操作可能にし、ユーザが指定した期間を全て表示、取得可能なインターフェースを開発する。また、moon seismic monitor で

は新たな波形データを要求後、直前に閲覧した波形を閲覧することができない。そのため本システムには直前に閲覧した波形を閲覧することができるような履歴を残す機能を搭載する。

また、後藤ら [7] による SOM を用いた月震の可視化システムでは従来の月震波形の相互相関係数ではなく周波数成分を特徴量として教師なし学習のひとつである SOM による深発月震の分類を可視化し、月震解析基盤には分散処理などを用いて RDBMS の高速化を図った。また、この研究によりこれまで人手によって行われていた分類は機械学習を用いた分類によってさらに容易になり、震源の分類基準の改良や新たな月震の発見を可能にした。機械学習には正解データが必要であり、その正解データは専門家の目視による結果を利用しているため、今後機械学習による解析の精度をあげるためにも可視化システムを開発することは重要である。

Philippe ら [2] は SciDB と関係データベースシステムのベンチマークを比較した。Philippe らは SS-DB という新たなベンチマークテストを提案し、SciDB と一般的な関係データベースシステムである MySQL の性能を比較した。99GB と 999GB の 2 次元配列データをデータセットとし比較を行った結果、MySQL よりも SciDB が処理にかかった時間が短く、配列データには SciDB が適していることを示した。そのため、月震センサーデータのような大規模なセンサーデータを SciDB で扱うことでデータへのアクセスの効率性を高めることができる。

これらの関連研究から、本研究では、SciDB を用いたセンサーデータ可視化システムを開発する。

また、山田らの研究 [3] によって新たな月震観測の可能性が示唆されている。この研究では、月探査用ペネトレータ搭載用として開発した短周期地震計に対し 0.1 Hz 以上で高感度を実現するための広帯域化を実施している。これにより、月震波形中の P 波や S 波などの識別がより容易になる。この研究により、今後も月震観測データが増加することが考えられる。したがって、センサーデータとは別に予め間引きを行った結果を保存しておく方法では今後増え続けるセンサーデータの容量に対応できない可能性がある。そのため本システムでは、ユーザが月震データを要求した際にリアルタイムに間引きの処理を行う設計とする。

3 提案システム

本研究では、研究者の目視による時系列データの解析を行うためのファーストステップとしての可視化システムを提案する。図 1 に本研究で提案するシステム構成とシステムの処理の流れを示す。

サーバサイドでは、大規模な時系列データを効率よく扱うために配列型データベースである SciDB を用いる。SciDB は分散処理、並列処理を行うアーキテクチャであり、一般的な関係データベースシステムと比べて配列データに特化したデータベースシステムである。特に SciDB は R や Python といった汎用の解析ツールとも相性がよく、R や Python にて SciDB を操作することができる。本システムはブラウザでの閲覧を

2 : <http://www.paradigm4.com/>

1 : <http://www.darts.isas.jaxa.jp/planet/seismology/apollo/app/>

行うことのできる Web システムにするため Python を使用する。特に、Web フレームワークに加えて機械学習や統計解析のライブラリが豊富であることが Python を使用するひとつの理由である。本システムはサーバサイドを Python で、クライアントサイドを JavaScript にて実装している。波形の描画には JavaScript の highcharts.js というライブラリを使用している。サーバサイドとクライアントサイドは Websocket によって双方向通信が確立され、データの送受信が行われる。

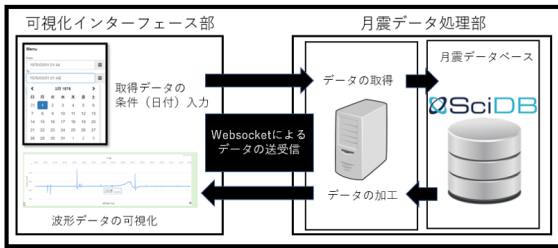


図 1 システム構成と処理の流れ

図 2 は本システムのインターフェースとなっている。highcharts.js の標準機能により、波形をマウスオーバーすることで図 2 のように座標 (加速度 Z 軸成分の値と時刻) を表示する。

SciDB には部分配列を取り出すために *subarray* という関数がある。配列 *array* から *i* 番目の配列要素から *j* 番目の配列要素を取り出したい場合は *subarray(array, i, j);* というクエリを発行することで *i* 番目から *j* 番目までの部分配列を取り出すことができる。ここでインターフェースではユーザーが時刻を入力するため、式 (1) を用いて時刻を配列番号に変換する。時系列データは任意のサンプリングレートにより等間隔に並んでいるため、検索時刻 $time[i]$ と初期時間 $time[0]$ との差をサンプリングレート *samplingrate* で割ることにより配列番号 *index* を算出できる。

$$index = \frac{|time[i] - time[0]|}{samplingrate} \quad (1)$$



図 2 インターフェース：波形

また、波形をドラッグすることで波形の拡大が可能となっている。拡大された波形は拡大前の波形を上書きすることなく拡大前の波形の上部に追加される。これにより、拡大前の波形と拡大後の波形を比較可能となっている。これは既存のシステムで課題となっていた拡大を実現したものであり、一度閲覧した波形はブラウザを更新しない限り履歴として残り続ける。既存シ

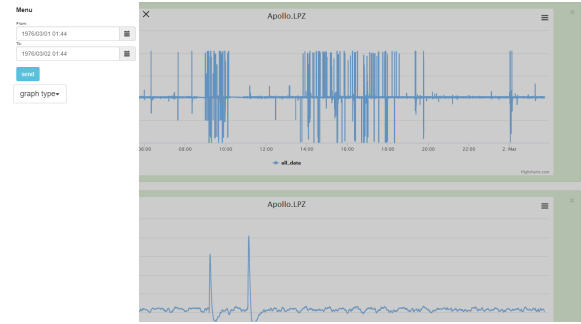


図 3 インターフェース：メニュー画面

ステムのように逐一波形をサーバに要求する必要がなくなる。左上のメニューバー (図 2) を選択することで入力フォーム (図 3) が出現し、必要な期間を入力することでその期間のデータをサーバへ要求することができる。

4 提案手法

波形の特徴点を維持するために、任意の 2 点とその間の 1 点との垂直距離に着目した。

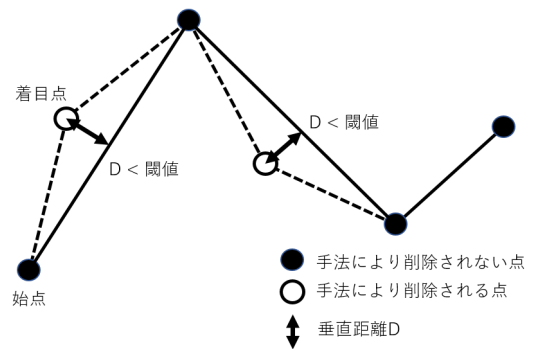


図 4 提案手法：垂直距離による間引き

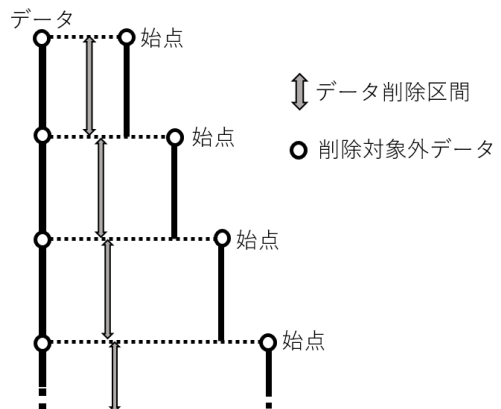


図 5 提案手法：間引きの繰り返し図

本手法は、図 4 のような始点と着目点と着目点の次点の 3 点に関して、始点と着目点の次点と、着目点の垂直距離を計算し、その距離が閾値未満であれば着目点を次点へと移し再度計算す

る。計算結果が閾値以上である場合、その点を配列へ格納し、図5のようにその点を始点として上記の処理を繰り返す。始点を更新することで一度計算した点へ戻ることなく、計算量を $O(n)$ に抑えることができる。この手法により図4のような垂直距離が閾値未満となる点を削除しつつ特徴点を残すことができる。ここで3点 $A(x_1, y_1)$, $B(x_2, y_2)$, $C(x_3, y_3)$ において直線 AB から点 C までの垂直距離 L は式(2), (3), (4), (5)によって求める。ただし、 $A \rightarrow B$ のベクトルを u , $A \rightarrow C$ のベクトルを v とし、 u と v の角度を θ と定める。また Cr は u と v の外積である。

$$Cr = |u||v|\sin \theta \quad (2)$$

$$|v|\sin \theta = Cr/|u| \quad (3)$$

$$L = |v|\sin \theta \quad (4)$$

$$L = Cr/|u| \quad (5)$$

5 実験

本手法によるデータ数を削減した結果を他の処理手法による結果と比較し評価する。評価指標は実行時間と特徴点を維持しているかの2点である。特徴点を維持しているかどうかの評価方法として動的時間伸縮法(DTW) [4]を用いる。DTWによって、データ数を削減していない波形と各手法を用いた波形の類似度を求める。DTWは時系列データ間の類似度合いを計測する手法である。DTWによる波形同士の距離は式(6)のように定義される。

$$DTW(x, y) = \min \sum_{i=0}^n |x - y| \quad (6)$$

ここで x, y は距離を算出したい波形である。DTWは波形同士の距離を算出するため小さい値であるほど類似度が高いといえる。本実験では、手法を適用していない波形、つまりデータ数を削減していない波形データと各手法を適用した波形データのDTWによる比較を行う。また、実行時間は各手法を10回試行した平均実行時間を結果とする。

本手法と比較する手法としてダグラスポーカーアルゴリズム [5] とリサンプリング法の2つを用いる。ダグラスポーカーアルゴリズムは時系列データの始点と終点からの最遠点を算出し、その点で2つのデータ配列に分割し、それぞれ最遠点を求める。これを再帰を用いて全データに対して繰り返す。ダグラスポーカーアルゴリズムは地図などに使われる手法であり地図の輪郭を少ない頂点数で近似することができる。リサンプリング法にはPythonのpandas.DataFrame.resampleを用いる。各手法により、同一のデータ数へと間引きする際の実行時間と波形の類似度を比較するために、本実験ではそれぞれの実行後のデータ数が1000に近くなるように各閾値を定める。各手法で用いる閾値と手法適用後のデータ数を表5に示す。

本実験で用いる時系列データは、Apollo14号によって月に

表1 各手法で使用する閾値と手法適用後のデータ数

手法	閾値	データ数
垂直距離に着目した手法	1.06	1003
ダグラスポーカーアルゴリズム	0.9931	1003
リサンプリング	3.6S	1001

設置された地震計から得られた加速度Z軸成分データのうち、1971-04-18 08:40:00~09:40:00の期間のデータ(23,851レコード)を用いる。

5.1 実験結果と考察

手法を適用していない波形グラフと各手法適用後の波形グラフをそれぞれ図6, 図7, 図8, および図9にそれぞれ示す。波形グラフの縦軸は加速度Z軸成分の値であり横軸は時刻となっている。また、実験結果を表5.1, および表5.1にそれぞれ示す。

まず、各手法の実行時間であるが最も処理時間が短かったのはリサンプリングであり、最も処理時間が長かったのはダグラスポーカーアルゴリズムであった。本手法はリサンプリングの処理時間には劣るものの、ダグラスポーカーアルゴリズムよりは高速に処理が行われた。ダグラスポーカーアルゴリズムは最大で $O(n^2)$ の計算量となるが、本手法は $O(n)$ で計算することができる。また、リサンプリングが最も処理が速かった理由として、データを単に閾値毎に間引いているからである。つまり、特徴点であるかどうかは無視していることになる。それはリサンプリング後の波形とリサンプリング前の波形を比較すれば明らかである。

次に、各手法適用後と適用前の波形の類似度であるが、最も類似していなかったのはリサンプリングであり、先述したように図6と図9を目視で比較してもその違いは明らかである。また、本手法とダグラスポーカーアルゴリズムの結果は本手法の方が類似度が高い結果となった。

以上の2つの実験結果から、実行時間を重視する場合はリサンプリングを用い、波形の類似度を重視する場合は本手法またはダグラスポーカーアルゴリズムを用いることが望ましい。また実行時間と波形の類似度を2つを重視する場合は本手法を用いることが望ましい。本システムでは、研究者が指定した範囲の波形をより高速により近似された波形を可視化することを重視している。そのため、処理時間の大きいダグラスポーカーアルゴリズムや波形の類似度が小さいリサンプリング手法は不適切である。本システムでは本手法を用いることにより、データ数の多さから引き起こされるブラウザへのメモリの負荷やサーバへデータを要求してからブラウザへ表示するまでの速度を低下させないような可視化が可能である。

表2 各手法の実行時間

手法	実行時間[s]
提案手法	1.425
ダグラスポーカーアルゴリズム	80.212
リサンプリング	0.097

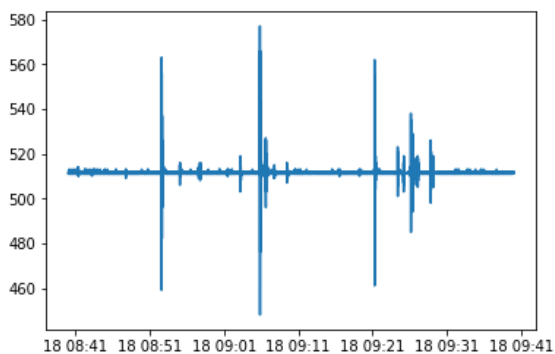


図 6 手法を適用していない場合

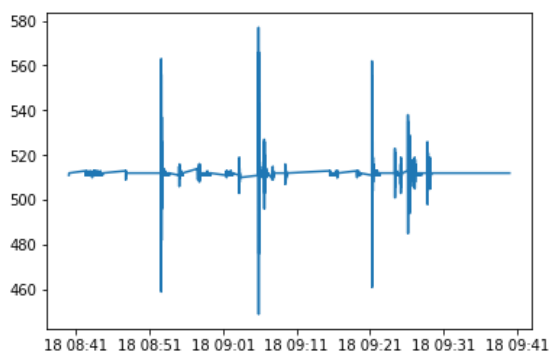


図 7 垂直距離に着目した手法

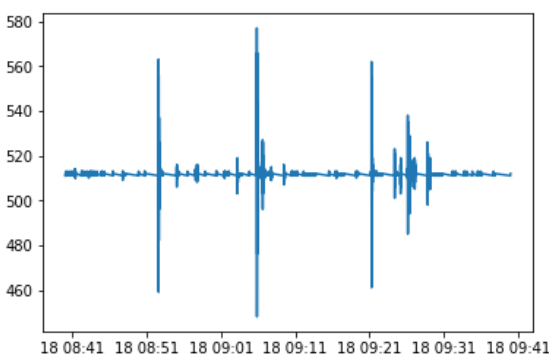


図 8 ダグラスポーカーアルゴリズム

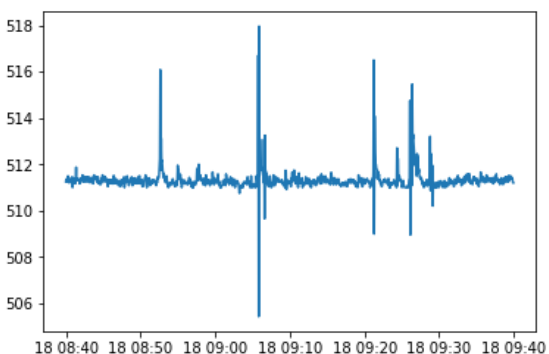


図 9 リサンプリング法

表 3 DTW による類似度

手法	類似度
提案手法	7343
ダグラスポーカーアルゴリズム	7462
リサンプリング	12087.4

6 まとめと今後の課題

本論文では、時系列データを Web ブラウザ上に可視化するシステムを提案し、サーバとクライアント間のデータ送信量を削減するために任意の 2 点とその間の 1 点との垂直距離に着目した手法を提案した。各手法は既存の手法と比較した。本手法の処理時間はダグラスポーカーアルゴリズムより短く、リサンプリング法よりも長い結果となった。元データとの類似度比較はリサンプリング法よりも本手法とダグラスポーカーアルゴリズムが類似度が高い結果となった。波形の可視化時間の速さと元波形との類似度を重視する本システムでは、類似度の低いリサンプリング法と実行時間が大きいダグラスポーカーアルゴリズムと比べて本手法の有効性を示した。

今後の課題として、本実験で用いた月の地震計加速度データ以外の時系列データの可視化にも本システムが有効であることを確認し、より汎用的なシステムへと改良することがあげられる。また提案手法は本実験において、間引き後のデータ数が約 1000 になるように閾値を定めたが、実際にシステムを運用するにあたりユーザの指定する波形の大きさにより、閾値は変動する。最適なデータ数になるまで提案手法を繰り返し実行し閾値を定めることは容易に検討できるが、閾値の最適解を見つけ出すために実行時間が増加しては本研究の論点から逸脱していると考えられる。そのため、閾値の最適解を高速に定めることが今後システムを実用する上での大きな課題となる。

文 献

- [1] 山田竜平, 山本幸生, 桑村潤, 中村吉雄. アポロ月地震データ公開システムの開発. 宇宙科学情報解析論文誌宇宙航空研究開発機構研究開発報告, 1 (2012), 121-131
- [2] P. Cudre-Mauroux, H. Kimura, K. -T. Lim, J. Rogers, S. Madden, M. Stonebraker, S. B. Zdonik, and P. G. Brown: SS-DB: A Standard Science DBMS Benchmark. http://www-conf.slac.stanford.edu/xldb10/docs/ssdb_benchmark.pdf
- [3] Ryuhei Yamada, Tanguy Nebut, Hiroaki Shiraishi, Philippe Lognonne, Naoki Kobayashi, Satoshi Tanaka: "Frequency band enlargement of the penetrator seismometer and its application to moonquake observation" Advances in Space Research 56 (2015) pp.341-354
- [4] Senin, Pavel. "Dynamic time warping algorithm review." Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA 855 (2008): 1-23.
- [5] David H. Douglas and Thomas K. Peucker, "Algorithms for the reduction of the number of required to represent a digitized line or its caricature," Canadian Cartographer Vol 10 No 2 December 1973, pp. 112-122.
- [6] Nakamura, Y. New identification of deep moonquakes in the apollo lunar seismic data. Physics of the Earth and Planetary Interiors 139, 3 (2003), 197-205.

- [7] 後藤康路, 山田竜平, 山本幸生, 横山昌平, and 石川博. 階層型 SOM に基づいた大規模月地震波形の可視化システム. 第 6 回データ工学と情報マネジメントに関するフォーラム, E1-4 (2014).
- [8] Taichi Kawamura, Naoki Kobayashi, Satoshi Tanaka and Philippe Longnonné. Lunar Surface Gravimeter as a lunar seismometer: Investigation of a new source of seismic information on the Moon. *Journal of Geophysical Research: Planets*, Vol. 120, No. 2, pp. 343-358, 2015.
- [9] R. C. Bulow, C. L. Johnson, and PM Shearer. New events discovered in the apollo lunar seismic data. *Journal of Geophysical Research: Planets* (1991-2012), Vol. 110, No. E10, 2005.