An Infrastructure to Store and Analyse Seismic Data as Suffix Trees

MSc Data Analytics - Project Proposal

Tom Taylor

April 2016

This report is substantially the result of my own work except where explicitly indicated in the text. I give my permission for it to be submitted to the JISC Plagiarism Detection Service. I have read and understood the sections on plagiarism in the Programme Handbook and the College website.

Contents

1 Abstract

The purpose of this project is to develop an infrastructure and tool set for converting raw seismic time series data in to a searchable string using SAX (**S**ymbolic **A**ggregate appro**X**imation) and then to store this data as a suffix tree for fast searching and analysis. An interface will then be developed to enable the searching of these suffix trees and provide visualisation of the data. The code should be re-usable in a project to receive live streaming data from many stations.

2 Background

2.1 Seismic Waves

Seismic waves take on two main forms, body waves and surface waves. Body waves are those that travel through the interior of the earth and are the fastest travelling. The body waves are comprised of **P** (primary) waves which are compressional waves and arrive first. **S** (secondary) waves are shear waves and travel more slowly, sometimes minutes behind. The surface waves travel only along the earth's crust and will normally arrive much later than the body waves.

A seismic station records movement over three axis: vertical (**z**) alongside horizontal in terms of north-south (**n**) and east-west (**e**). P waves are primarily observed on the z axis and so a P wave can be measured as a simple metric of displacement along the Z axis. S waves are primarily observed as movement across both the n and e axis. The geometry of the fault tends to have a bearing on the orientation of the displacement so the two axis of movement cannot be easily combined in to a single metric for time series analysis.

2.2 Symbolic Aggregate Approximation (SAX)

SAX (**S**ymbolic **A**ggregate appro**X**imation) (Lin et al., 2007) is a technique where by a single dimension of a time series is reduced to a string of symbols for pattern matching. The technique involves first transforming the normalised time-series in to a Piecewise Aggregate Approximation which is then represented by a fixed number of symbols.

The data is first divided in to equal sized frames, then the mean deviation from zero of each frame is calculated. An appropriate number of breakpoints symmetrical along the x-axis are created so that they follow a Gaussian distribution and a symbol assigned to each range between the breakpoints. Then for each frame, a symbol is assigned based on which range the mean falls in to. The symbols assigned to each frame

are then concatenated in to a string and it is this string that gives the SAX representation of that data. The width of the time frame and the number of discrete regions would be two parameters passed to this process alongside the data.

2.3 Suffix Trees and the related infrastructure at Birkbeck

3 Objectives

1. Interpretation of raw data from seismic stations and storage as time-series data in a time-series database such as OpenTSDB for easy access during conversion to SAX and for later rendering during interactions with the data.

2. Conversion of the data to SAX and storage in a suffix tree.

3. An interface for searching and viewing the raw data.

4 Approach

As laid out in the objectives, the development will produce three separate components.

4.1 Interpretation and Storage as time-series data

The raw data received from the stations is accessed as files in the SAC (Seismic Analysis Code) binary format. These will be read using the ObsPy python library which can read both the headers and return the raw data as a Python object.

This data will be batch processed and imported in to a time series database. Initially OpenTSDB is being considered for performant reasons but there may be scope for this to change depending on how difficult it is to implement this on the Universities hardware. Other options could be InfluxDB or Graphite.

4.2 Conversion of data to SAX and building of the Suffix Tree

4.3 Development of an interface

The interface to query and view the data will be web based as to ensure maximum compatibility with clients. Many open source graph renderers are in existence such as

Grafana and Cubism.js and components of both are likely to be used for viewing the raw seismic data alongside the SAX data.

5 Plan

As both ObsPy and the existing infrastructure for Suffix Tree storage are written in Python, this seems the logical choice. Specifically Python 3.5 will be used for all components.

Development will take the approach of Test Driven Development (TDD) where unit tests will be written at a class level before the classes are coded.

References

Lin, J., E. Keogh, L. Wei, and S. Lonardi
2007. Experiencing sax: a novel symbolic representation of time series.