Mining Recurrent Activities: Fourier Analysis of Change Events

Abram Hindle, Michael Godfrey, Ric Holt
University Of Waterloo, Waterloo, Canada

Developers create artifacts

Source Control Repository (CVS/SVN/GIT)

Developers create various kinds of artifacts when they make changes:
source code, test suites, bug reports, documentation, mailing list messages, etc

Software development has recurrent behaviour

I regularly make source code changes
My job is to test and break these changes

I hold regular inspection meetings
I am master of this current iteration.

Within an iteration, there are recognizable repeating patterns.
Iterations themselves are also a recurrent and repeating behaviour.

How do we discover recurrent behaviour?

With Time-series analysis
A signal, multiple years in length

You must choose a period to use time-series analysis!

A day
A week
A year
But which is correct or useful?

What period should I analyze this signal by? If only I had a tool to tell me what to do!

A solution: Fourier transform

We can convert from a time/amplitude view to...

Recompose the signal from sinewaves

... a frequency/magnitude view which shows us dominating frequencies and periods

How can we apply the Fourier transform to software change and software related data?

We have a signal

We apply the Fourier transform.

Investigation of each frequency bin shows that 2 authors are individually responsible for 2 bins. Their frequencies imply their periodicities.

We discover that:
* blue commits many changes
* green tests blue’s commits
* The other developers follow their own schedule.

Convert discrete events into signals

First, take an event stream, such as, revisions to a project, log events, mailing list posts, etc.

Then we take these events and aggregate them. In this example, we put the changes into buckets of a day. Then we get a signal from these buckets which we can analyze via Time-series Analysis or the Fourier Transform.

So what? Does it work?

We can find periodicities of software. Here’s the Fourier transform of MaxDB 7.500

Spectrograms of software change over time

Spectrograms are Fourier transforms of short periods, shown side by side. The x-axis is time, y-axis frequency and color is magnitude.

Horizontal smears show recurrent behaviour.

Note the highlighted peaks, these are the dominant frequencies, which you could use for intervals in Time-series analysis.

Future Work

Who’s responsible for this behaviour? Semantic Slicing

A behaviour can be explained by a combination of slices.

Other uses of the Fourier Transform: Dynamic Analysis via call depth

Partition development time via Self Similarity