KDD on the SPL

-- karl yerkes 2015.02.12

What is Data Mining?

The term 'Data Mining' is overly broad and steeped in hype. Instead, consider the term 'Knowledge Discovery in Databases' (KDD). Testing a hypothesis with a query is useful, but we really want to learn things about our data that we don't even have a hypothesis for. We want a way to automatically expose patterns in our day. Start by reading From Data Mining to Knowledge Discovery in Databases (Fayyad 1996) and then look into the proceedings of ACM SIGKDD.

What is Frequency Pattern Growth (FPGrowth)?

The FPGrowth Algorithm is a well-researched and widely used tool for Association Rule Learning. Given a set of transactions involving items, it produces a set of associations between those items with "confidence" values. Such patterns are valuable to retailers: "If customer bought X, he/she is also likely to buy Y and Z." We see this sort of Affinity Analysis used by retailers such as Amazon. For more detailed technical information, read this paper on an implementation of FPGrowth.

How do we apply this to the SPL?

Consider this query which looks for all items checked out on my birthday in 2011:

```
select unix_timestamp(checkOut) as tid, count(itemNumber) as n, group_concat(itemNumber) as items
  from transactions where checkOut between '2011-06-26' and '2011-06-27'
  group by tid
  order by tid
```

This query reorganizes the check outs, formating them for input into FPGrowth. Now look carefully and convince yourself that the database only has a resolution of 1 minute. In practice, actual checkouts should take less than a minute, right? Yet we see here that a single minute might have hundreds of checkouts. This is because each minute represents the actions of N people where N may be greater than 1. Does this noise effect the results of FPGrowth?

I used this FPGrowth implementation to produce the data for this result but other implementations exist.

Where to go from here?

I recommend we segment the database and run FPGrowth on each segment. For instance, create results for each year: 2006, 2007, 2008, ... and compare them. Perhaps look at results within specific item type categories (i.e. only non-kids books).