

Project 1: Data Query

“New” Books

It's a common situation that library buys new books every several months or each year, so does Seattle Public Library. The library has its own collection code defined. With the call number together with the collection code, it's easy to know whether the book is a new one bought this year or not. Therefore the question is how does these “new” book attract readers over time. People might guess that those new books won't be as popular as they arrived. But is that the truth? How fast does the frequency pattern drop? Is the category of books affects as well? These are the concepts I try to explore.

SQL Query

```
SELECT
  right(callNumber, 4) as versionYear,
  CASE
    WHEN length(deweyClass) > 0 THEN 'non-fiction' ELSE 'fiction' END as isFiction,
  SUM(CASE
    WHEN Year(checkOut) = 2006 THEN 1 ELSE 0 END) AS '2006',
  SUM(CASE
    WHEN Year(checkOut) = 2007 THEN 1 ELSE 0 END) AS '2007',
  SUM(CASE
    WHEN Year(checkOut) = 2008 THEN 1 ELSE 0 END) AS '2008',
  SUM(CASE
    WHEN Year(checkOut) = 2009 THEN 1 ELSE 0 END) AS '2009',
  SUM(CASE
    WHEN Year(checkOut) = 2010 THEN 1 ELSE 0 END) AS '2010',
  SUM(CASE
    WHEN Year(checkOut) = 2011 THEN 1 ELSE 0 END) AS '2011',
  SUM(CASE
    WHEN Year(checkOut) = 2012 THEN 1 ELSE 0 END) AS '2012',
  SUM(CASE
    WHEN Year(checkOut) = 2013 THEN 1 ELSE 0 END) AS '2013',
  SUM(CASE
    WHEN Year(checkOut) = 2014 THEN 1 ELSE 0 END) AS '2014',
  SUM(CASE
    WHEN Year(checkOut) = 2015 THEN 1 ELSE 0 END) AS '2015'
FROM
  spl3._rawXmlDataCheckOuts
WHERE
  substring(collectionCode from 3 for 3) = 'new'
  AND right(callNumber, 4) > 2005
  AND right(callNumber, 4) < 2016
  AND itemType = 'acbk'
GROUP BY versionYear , isFiction
ORDER BY versionYear , isFiction
```

Duration: 56.480 sec
Fetch Time: 0.013 sec

Explanation

There are so many types of items in the library. To make the result more clear and meaningful, I only chose itemType to be 'acbk', which means adult book. And this is done in the where clause. The tricky part is how to select those books which are new, not only new to the library, but also a new published item. I use collectionCode together with callNumber to find those suitable books. And I display the result according to their category (fiction or non-fiction) and published year (version) in time dimension.

Result

Please see CSV file in attachment.

Commentary

The result is quite interesting. As I expect, those new books are more likely to be checked out within their first two years, and the checkout frequency keeps dropping as time goes by. Comparing fiction to non-fiction books, the decrease speed for novels are much more dramatic. Even though there is no way to know the exact amount of new books for each type of the book that library imported each year, it's still clearly shown from the trend of the statistics that new fictions attract reader more than new non-fictions; but this interest disappears faster among new novels. Another interesting question I discovered from this result is that there are always some borrow records even before the book was published. And this phenomenon happened every year. The possible reasons could be either library has early access to some new books before their publication or there are errors in the system every year, or both.