Data Acquisition Project

Book Ratings Comparison

Amit Yadav, Alice Benziger, Rachan Bassi

11-OCT-2014

Problem Statement

The digital transformation of the way in which books are read, sold and published has only begun. A classic bibliophile is often found scrambling the Internet for reliable sources to find a good hold of a book's review. Reviews from retail websites like Amazon and eBay are good candidates. However, customer ratings are an integral part of marketing strategies employed by these retailers. It is a plausible question to wonder if their reviews are affected by the price of the book? Could cheaper books have inflated ratings or vice versa? On the other hand, one would not expect social cataloging websites like goodreads to have any bias based on price. However, goodreads users are passionate about reading. Can this factor affect their generosity while rating a book? As such, could their reviews be more critical in comparison to retail websites?

Motivated by these questions and the current absence of a single aggregator that provides a thorough consolidation of book reviews, we aim to build a book comparison system - a system that compiles reviews across different websites and gives the readers a convenient platform to compare reviews for a particular book.

Executive Summary

This report describes a system, which consolidates ratings of books across three different websites - eBay, Amazon and goodreads. We particularly chose these three websites for two reasons. Firstly, we wish to check if there is a bias in book ratings between retails websites and social cataloging websites and, secondly, for completeness as goodreads does not catalogue technical books and is predominantly a source for leisure reads. After scrapping ratings of books from Amazon, eBay and goodreads, we compared them for discrepancies. On average, we found that the ratings on eBay and Amazon were higher than those on goodreads. This is in line with our initial hypothesis, that ratings on goodreads possibly have a tendency of being slightly deflated in comparison to Amazon and eBay. This indicates there is a possibility that goodreads users have higher standards and thus rate books more harshly. However, more research needs to be carried out, using more data to further validate our hypothesis. Price will be an important factor to be considered in this study and must be extracted from Amazon and eBay.

Data Acquisition Process

The data acquisition process for our system starts with extracting book categories from eBay.com. These book categories are fed back into eBay's API. Based on the book category, we obtain the book title, ISBN and average rating for a particular book. The ISBN obtained from eBay API is fed as an input into goodreads and Amazon API. The amazon API gives us the average rating based on the ISBN we provide. Similarly, the goodreads API also provides the average rating for that ISBN. As a result, we get average rating from eBay.com, Amazon.com and Goodreads.com. See Figure 1 to understand how the information flows in our system. For more details on how we obtain information from each API, see appendix.

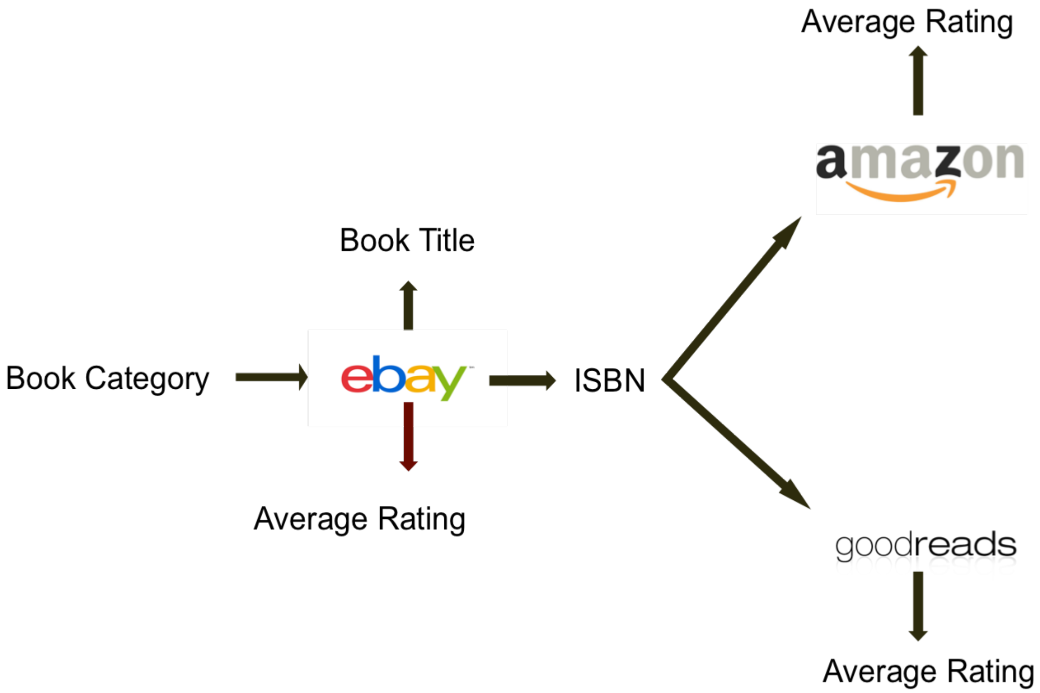


Figure 1: Information Flow

Data Challenges and Solutions

1. The biggest challenge in acquiring data is the throttle limit imposed by all the three APIs. eBay's API has a throttle limit of maximum 5000 queries per day. Goodreads does not allow more than 1000 ISBNs per request, and Amazon has an additional constraint of 10 requests per second. To overcome the throttle limitations, we split the data acquisition process over multiple days. We also generated random wait times to avoid being detected on Amazon's website. However, with over 3 billion books available, even splitting it across multiple days could be a challenge. So to make the concept viable, the solution would be to gain premium access to the data from these APIs.
2. Apart from the throttle limits, another challenge was to find a repository of ISBNs. Our initial idea was to gather ISBNs from ISBN.org. However, we found their API very unreliable with multiple site crashes. So, we decided to use eBay to collect the ISBN numbers. This was not straightforward either as multiple API methods had to be navigated to obtain the ISBNs.
3. Amazon's API did not have a method that could give the ratings directly. To overcome this, we grabbed the URL to each book's Amazon webpage and did an HTML parsing to obtain the ratings.
4. Not all the websites had ratings corresponding to all the ISBNs. Goodreads did not have ratings for a lot of technical books.

Limitations

The biggest limitation would be to keep the data in sync. New books are added every day. The current system does not include a push mechanism to update ratings pertaining to new books. This means re-running the entire process periodically. This is time consuming and tedious due to throttle limitations.

Amazon did not have an API method that could give the ratings directly. This issue was overcome by parsing the book's Amazon webpage using Beautiful Soup. This means that the code fetching the ratings must be rewritten every time a change is made on Amazon's website.

Data Storage and Representation

The data obtained from the three websites is stored in the form of a dictionary of dictionaries. A sample form of the dictionary is shown below.

```
{ Category_ID :
   { ISBN:
        {Book_Title:  'XYZ', ebay_avg_rating: '4.3', amazon_avg_rating: '4.5',      goodreads_avg_rating:
        '5'}}}
```

Using this storage, we can obtain the data in tabular form and do a comparison. A sample table is shown below:

ISBN	Book Title	Book Category	ebay Average Rating	Amazon Average Rating	GoodReads Average Rating
0375842209	The Book Thief	Fiction & Literature	5	4.5	4.36
0385349947	Lean In: Women, Work, and the Will to Lead	Non-Fiction	4.5	4.5	3.87
0439708184	Harry Potter and the Sorcerer's Stone	Children & young adults	4.5	4.7	4.38

Recommendations

We extracted data for about 2000 books from three categories namely, Fiction and Literature, Non-Fiction and Children and young adults. Based on the data collected, we have found ratings reported by goodreads.com to be lower than those at eBay and Amazon. To further study price bias or the level of criticality, the next step would be to extract the corresponding prices from both eBay and Amazon and incorporate those into the analysis.

Increasing the number of categories that are extracted and check for category based favoritism can further enhance the comparison. Publishing houses could benefit from this while shelving books.

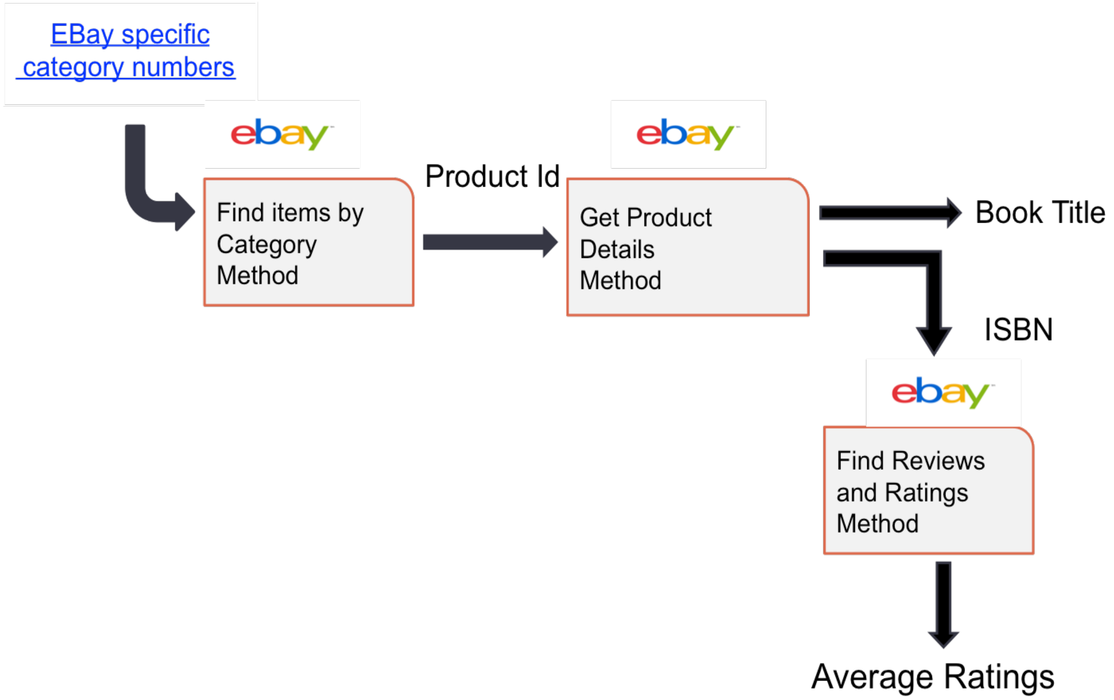
The scope of the model should also be increased to include other book retailers like Barnes and Noble, Powell's books etc. as well as international retailers.

References

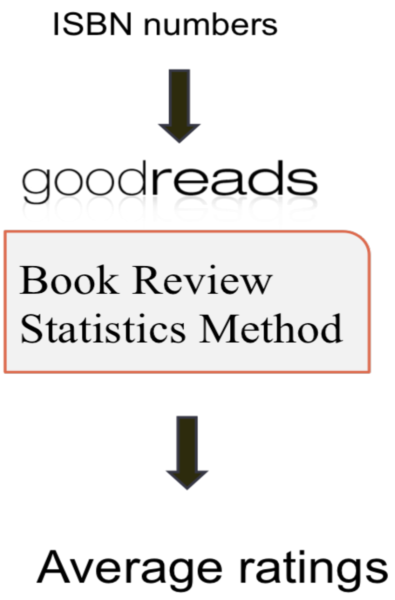
http://go.developer.ebay.com/developers/ebay
https://developer.amazon.com
https://www.goodreads.com/api

Appendix

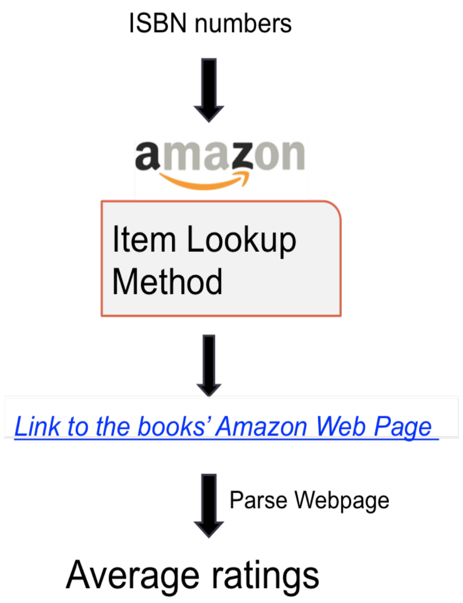
Data extraction Process: Ebay API


EBay specific category numbers
Find items by Category Method
Product Id
Get Product Details Method
Book Title
ISBN
Find Reviews and Ratings Method
Average Ratings

Data extraction Process: Goodreads API



Data Extraction Process: Amazon API


ISBN numbers
amazon
Item Lookup Method
Link to the books' Amazon Web Page
Parse Webpage
Average ratings