Marvel Vs. DC: Predicting Consumer Similarity Using Latent Factor Models

Teofilo Erin Zosa IV
University of California, San Diego
Department of Computer Science and Engineering

Abstract— **Using data personally scraped from Amazon.com, I perform a sentiment analysis and generate a star-rating predictor from DC Comics and Marvel Comics graphic novel reviews. I show that sentiment for both publishers is essentially positive and that the latent factor model performance degrades when predicting user-item ratings for graphic novels from one publisher to another; a finding that suggests there is a qualitative difference between graphic novels produced by these two publishers.**

I. Introduction

Web Mining is a burgeoning field in which insights are drawn from data inherent in web pages, usually in the form of user-item or user-user interactions. Two popularly data mining methods that make use of this information are sentiment analysis and the generation of predictive models known as recommender systems. Recommender systems have been a boon for many areas such as e-commerce, entertainment media discovery, and adaptive systems, and are increasingly finding themselves in widespread use.

For assignment 2 I chose to perform a sentiment analysis on the 1200 most reviewed DC Comics and Marvel Comics hardcover, paperback, kindle, omnibus, and limited/collector's edition books [1], and use that data to create a star-rating predictor. The data was scraped from Amazon.com (hereafter referred to as "Amazon") directly and is accurate as of November 29, 2017.

A. Motivation

In the realm of graphic novels, DC and Marvel are the two major publishing houses together accounting for an astounding 66.96% retail market share (30.58% and 36.38% respectively) as of October 2017[1]. Given their popularity and history, there is much public opinion on who is the better publishing house in terms of quality of publications, with many arguing that the novels produced by one publisher are qualitatively different than those produced by the other. To test this assertion, I wanted to analyze the sentiment of users via their product reviews, train predictive models to generate product recommendations specific given either DC or Marvel product ratings, and use those models to test whether there is a significant difference in the relationship between DC novel consumers and Marvel novel consumers. Namely, if a recommender from one publisher can successfully recommend products for the other publisher. The hypothesis of this study was that, if intra-data performance was about equal for both models and there existed a large discrepancy between inter- and intra-data predictions, then we could surmise that there were different relationships in the data; relationships that the models learned and employed in generating their predictions.

II. Related literature

Although publicly available datasets related to Amazon product reviews exist[2], a cursory analysis showed that a majority of the ASINs (Amazon Standard Identification Numbers; i.e., Amazons unique inventory process) corresponding to the most reviewed graphic novels by DC or Marvel were either missing from those datasets, or had many fewer reviews than were currently displayed on Amazon[2]. Thus, generating a newer, complete dataset was a necessary next step.

Additionally, the authors of that study were focused on suggesting substitutable and complementary products, whereas this study is focused on using data mining to discover if products are similar, and if so, how and to what degree (i.e. via semantic analysis and cross-inferential performance).

[1]When returning a list of items sorted by filter (e.g. Most Reviewed), if the number of items exceeds 1200, Amazon only returns the first 1200 in that set.

[2]The authors did not specify how the user-item tuples were categorized, though presumably they disambiguated reviews based on the specific ASIN.

III. METHOD

A. Web Scraping Approach

An "overview" crawl was performed on the results of the most reviewed query in each category for both publishers (e.g. DC, Paperback). Each category-by-publisher combination query resulted in up to 20 pages of results with 60 unique products per page. This crawl retrieved the ASIN, number of reviews, and overall star rating (one to five, no fractional ratings) for each item (6,153 unique items).

Next, an "in-depth" crawl was performed on books retrieved from the overview crawl. Each book had a variable number of pages dictated by the number of reviews each book had (10 per page), ranging from 1 page to 833 pages. This crawl retrieved the user ID, review-specific ASIN [3], star rating, review title, review text, and date ("Month Day, Year" format) [4]

B. Avoiding Amazon's Bot Detection Algorithm

Initial crawling attempts were largely unsuccessful due to Amazons algorithm to check for automated traffic. Through trial and error, it was discovered that one could circumvent detection by varying digits in the header information (e.g. 'Safari/537.xx' where xx was a random number between 0 and 36) and utilizing a random time delay instead of a fixed delay. This enabled both a successful crawl as well as a multi-core map-reduce crawling paradigm, where each child thread would process the review of an individual book. The main thread would then aggregate their results back into a master data structure. This modification sped up the crawl by at least two orders of magnitude.

C. Preprocessing

1) Filtering: Books with no reviews (which occurred primarily in the omnibus and limited/collector's editions categories) were discarded from the dataset.

Next, books with a large amount of reviews were filtered to remove reviews for unrelated products. Amazons pooling of reviews of closely related items resulted in specious reviews for specific books, especially for derivative novels. For example, the Stephen King graphic novels produced by Marvel include reviews for the original Stephen King novels, which were not produced by Marvel, and share the reviews between each other (presumably due to the fact that they are from the same series or share main titles). This results in an inordinate amount of erroneous reviews. To counter this phenomenon, books with over 800 reviews were filtered such that only reviews that corresponded to the ASINs of the graphic novels in question were retained. This filtered out 62,239 reviews. This was only done for a handful of books (given in the Appendix) that were large outliers in review count as the discovery of the ASINs was done by hand. The remainder of the dataset consisted mostly of original source material. Presumably, any other specious reviews of this type would be so few in number as to only trivially affect any analysis or models based on that data.

In addition to the specious reviews, duplicate reviews were also removed. Large amounts of duplicate reviews in the dataset were another consequence of Amazons review pooling. For the purposes of this project, the original pooled reviews were left untouched in the first binding category they were discovered and removed from later categories. For example, if a product was first discovered in the paperback category but also had hardcover and kindle editions, the reviews in the paperback category were retained while the reviews in the hardcover and kindle versions were removed. This approach was chosen to avoid the need to cross-reference book editions as well as for practical purposes related to the exploratory analysis and generation of the predictive models (see the following section).

D. Dataset characteristics

For both the sentiment analysis and the predictive model training, only reviews for books with at least one paperback edition were chosen. This was based on feedback from experts from one of the publishing houses (DC Comics). According to those experts, the publisher's most representative works always have a paperback edition, with content printed solely in other formats typically catering to niche crowds; crowds that are by definition small and biased. Their inclusion in the dataset would have either had a negligent effect, biased the analysis, or more likely, both.

A dataset of this form also obviated disambiguating books by their specific identifier number, an approach that would have resulted in an explosion in the size and sparseness of the set of the user/item matrix for the predictive task and loss of information (i.e. books with identical content but variant covers or printing editions

[3] Amazon pools reviews for related editions (e.g. if the book was produced in paperback and hardcover, the reviews on the product pages of paperback and hardcover editions are identical) but retains the ASIN for which the particular review was written in the review information.

[4] Which was later converted into a Unix timestamp.

TABLE I
DATASET CHARACTERISTICS

	DC	Marvel	Total
Users	62,867	44,358	107,180
Books (unique)	3,002	3,071	6,070
Books (pooled)	1,188	1,193	2,381
Average Rating (out of 5)	4.35	4.24	4.30
Total Reviews	**81,941**	**64,188**	**146,129**

are completely different).

IV. Exploratory Analysis: Analyzing Sentiment

Two methods were used to determine product sentiment: Rapid Automatic Keyword Extraction (RAKE)[3] and TextRank, a graph-based ranking model derived from Google's PageRank algorithm[4], [5]. These methods were performed on text from review titles as well as text from review bodies.

V. Predictive Tasks

A. Dataset Partitioning

Using the dataset consisting only of reviews for books with at least one paperback edition, the data was split into six separate partitions: DC-only training data (70% of the DC data), DC-only validation data (20% of the DC data), DC-only testing data (10% of the DC Data), Marvel-only training data (70% of the Marvel data), Marvel-only validation data (20% of the Marvel data), and Marvel-only testing data (10% of the Marvel Data). see I for a more in-depth overview of the data split and cardinality.

B. Models Used

Three latent-factor models were trained to predict the star ratings a user would give a particular book given just a user ID and a product ID. One model was trained on DC graphic novels, another model was trained on Marvel graphic novels, and a third model was trained on graphic novels from both publishing houses[5]. The DC and Marvel-specific models were also tested on data from the opposing partition(i.e. the DC-LFM on Marvel data and the Marvel-LFM on DC data) to test if the models learned qualitatively different relationships in the data. Model hyperparameters were the same across the three models: $\lambda = 0.2$, learning rate $(\eta) = 0.005$, and number of features $= 1$.

The feature size was chosen to generate models that were reliant on the most highly predictive feature of their particular dataset to see if the models were qualitatively different from one another. If so, this may imply that there is a distinct difference in between the different publishers or the users who review those products. A baseline model which produced a star rating based on the a linear combination of the users ratings average and a products rating average was also used. The choice of models and baseline were inspired by the rating prediction task in assignment 1 wherein the baseline model achieved performance on par with the complex latent factor models (up to 2,000 features) trained for that assignment.

The strength of the baseline model is that it is very easy to train, but a glaring weakness is that it is unable to model any complex relationships in the data, whereas latent-factor models have the opposite problem, being able to model complex, non-linear relationships but at a cost of increased computational time and resources.

Though the state-of-the-art has moved more towards neural network-based models, the use of these models would necessitate feature selection. As I wanted to use a latent factor model to select the single best feature for each model across the different datasets, a latent-factor model was more appropriate.

C. Model Performance Evaluation

Model performance was evaluated via RMSE.

VI. Results

A. Sentiment Analysis

True to its name, RAKE did produce rapid keyword extraction that seemed to indicate positive vague positive sentiment when applied to review titles (see II). However the results were returned for the text reviews were not very informative, and selected for novel word-phrases in the reviews over potentially more informative keywords (the top ten are given in the Appendix). No major insights were gleaned from RAKE as title keywords between DC and Marvel were nearly identical and non-specific (save for one superhero(s) reference in each category).

TextRank produced much more relevant keywords (see III for the keywords extracted from titles and the Appendix for keywords extracted from reviews), even picking out the two most popular characters in the DC universe (viz. Superman and Batman), but did so at the

[5]with another 70%/20%/10% train/validation/test split randomly sampled from the entire dataset (i.e. not the same data partition as the DC and Marvel-only models)

TABLE II
TOP 10 TITLE KEYWORDS (RAKE)

	DC	Marvel
1	stars	stars
2	great story	great story
3	great read	great read
4	great book	graphic sf reader
5	great	great book
6	good	good
7	graphic sf reader	good read
8	good read	x-men
9	batman	great
10	read	good story

TABLE III
TOP 10 TITLE KEYWORDS (TEXTRANK)

	DC	Marvel
1	story	story
2	great story	good story
3	batman	fantastic stories
4	series batman	best comics
5	batman graphic novel	favorite comics
6	superman	comic history
7	superman/batman	classic 90s comic
8	comic books	single comic book
9	great comic	marvel
10	great book, amazing story	marvel crossover

expense of computational resources; analysis took two days on a machine with two E5-2678W processors (8 physical cores each) and 128 GB of RAM.

Overall, it seems that the only information gleaned is that, generally, both publishing houses have positive sentiment associated with their products.

B. Predictive Tasks

Inter-data model performance is shown in table IV and intra-data model performance is shown in table VI. As we can see, the LFMs were slightly better than their respective baseline models V. Additionally, cross-prediction RMSE is indeed much higher for both the LFMs and the baseline models. Similar results were achieved for LFMs with up to 100 features (data not shown).

VII. LIMITATIONS

As data consisted only of Amazon product reviews, findings may not generalize to other populations not adequately represented in this dataset (i.e. fans who do not write reviews on Amazon).

TABLE IV
LATENT FACTOR MODELS PREDICTION ERROR (RMSE)

	DC	Marvel	All
Train	0.48	0.47	0.48
Test	0.83	0.82	0.84
Validation	0.84	0.83	0.84

TABLE V
BASELINE MODELS PREDICTION ERROR (RMSE)

	DC	Marvel	All
Train	0.62	0.65	0.64
Test	0.87	0.88	0.88
Validation	0.88	0.89	0.88

Additionally, though great care was taken to reduce noise and bias in the data, it was retrieved independently and not with the express approval or help of Amazon or its partners. As such, there is a possibility that the data may be imperfect or incomplete.

Finally, the semantic analyses employed were designed to run on single documents. Aggregating the reviews into one linear document ignores the temporal aspect of each review, potentially negatively biasing the results of the algorithms. For instance, TextRank consistently produced salient keywords and high-quality summaries when applied to single reviews, but produced fairly vague results when applied to the entire corpus.

VIII. CONCLUSION

DC Comics and Marvel Comics seem to share much in common while indeed differing in some way. As both publishers account for over two-thirds market share, with the next highest publisher only possessing a paltry 11.46% market share [1], the fact that sentiment was mostly positive for both publishers is not an unexpected finding. Users that purchase graphic novels from DC and Marvel are clearly numerous and generally feel that these novels are worth the money, time, and brand devotion. This devotion may also come with the assertion that DC and Marvel comics are qualitatively dissimilar; a claim that seems to be supported by the cross-predictive performance of our inferential models. The latent-factor models in this study achieved similar performance on their own datasets relative to one another (approximately 0.83 RMSE) and similar degradations in performance on their opposing datasets (approximately 1.07 RMSE) indicating an inability to

TABLE VI
SINGLE PUBLISHER MODELS CROSS-PREDICTION ERROR (RMSE)

LFM		Baseline	
DC-Only	Marvel-Only	DC-Only	Marvel-Only
1.11	1.04	1.12	1.04

completely generalize their predictions. In other words, although the basic products are the same, the most important relationship learned from the user-item interactions from one publisher was not fully applicable when predicting user-item interaction for the other.

Going forward, it would be interesting to perform this same study on a specific subset of graphic novels from each publisher. For example, for a dataset consisting solely of work centered on their canonical characters, we can see whether the differences are exacerbated and if any specific, more insightful sentiment can be derived from that data. Additionally, using other state-of-the-art approaches (e.g. neural-network based methods) to predict preferences as well as analyze sentiment may prove to be more successful.

APPENDIX

A. ASIN-Specific Filtered Books

1) Dark Tower: The Gunslinger - Last Shots
2) Dark Tower: The Gunslinger - The Battle of Tull
3) Dark Tower: The Gunslinger - The Little Sisters of Eluria
4) Dark Tower: The Gunslinger - The Man in Black
5) Dark Tower: The Gunslinger - The Way Station
6) Dark Tower: The Gunslinger Born
7) Ender's Game Graphic Novel
8) Ender's Game: Speaker for the Dead
9) Oz: The Wonderful Wizard of Oz
10) Pride and Prejudice (Marvel Illustrated)

B. RAKE Keywords From Review Text (Top 10)

DC Comics

1) evil-warrior-princess-turned-good-and...
2) red haired grunge-boy-alice-in-chains...
3) let-me-tell-you-the-epic-story-of-how...
4) you-got-chocolate-in-my-peanut-butter...
5) usual laboratory explosion/dump-the...
6) usual good-guy-fights-bad-guy-and...
7) jim-lee-big-breasted-over-the-top...
8) over-the-top alien-tourist-causes...
9) horror/myth/fantasy all-sorts-of...
10) else-is-the-only-way-i-can-show...

Marvel Comics

1) james-and-natalia-fighting-demons-of-their...
2) glorified monster-getting-one-final-spook...
3) the-memory-of-a-superhero-has-mysteriously...
4) fantastic four-spider-man-spider-woman...
5) well-reproduced-on-the-right-kind-of...
6) pleasant maybe-he-will-grow-as-a-god...
7) 6-issues-so-we-can-make-it-into-a...
8) full ripping-out-your-heart-and-trampling...
9) arm-in-arm-in-arm-in-arm-in-arm-in-arm
10) blood-colossus-mowing-down-every-dead-...

C. TextRank Keywords From Review Text (Top 10)

DC Comics

1) are mentions
2) dark is rising novel
3) adult human being
4) has wonder girl
5) the wonderbolts from friendship is magic
6) story
7) back story
8) story line
9) favorite story
10) nemoalan quartermainthe invisible manit was ok

Marvel Comics

1) the year is 1916
2) book
3) final book ties
4) bad book
5) book love
6) story
7) great story
8) good read
9) great read
10) fantastic read

ACKNOWLEDGMENT

The author wishes to thank his collaborators at DC Comics for all their input and suggestions, especially Monique F. Narboneta for all the days and nights spent edifying the author on the DC universe.

REFERENCES

[1] D. Comics, "Publisher Market Shares: February 2017," 2017. [Online]. Available: https://www.diamondcomics.com/Home/1/1/3/237?articleID=201973

[2] J. McAuley, R. Pandey, and J. Leskovec, "Inferring Networks of Substitutable and Complementary Products," 2015. [Online]. Available: http://arxiv.org/abs/1506.08839

[3] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction," *Text Mining: Applications and Theory*, pp. 1—-277, 2010.

[4] P. Nathan, "Pytextrank, a python implementation of textrank for text document nlp parsing and summarization," https://github.com/ceteri/pytextrank/, 2016.

[5] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," *Proceedings of EMNLP*, vol. 85, pp. 404–411, 2004. [Online]. Available: http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf