Words and Topics: Content Representations for Book Recommendation

Larissa Kolesnichenko[1] (larkkinn@gmail.com), Pavel Braslavski[2,3] (pbras@yandex.ru)

[1] ITMO University, Saint Petersburg, Russia
[2] Higher School of Economics, Saint Petersburg, Russia
[3] JetBrains Research, Saint Petersburg, Russia

The paper describes an exploratory study on content-based book recommendation. We use a large dataset of book ratings along with book content. We experiment with several topic modeling variants and tf.idf representation. Predictions based on one of the topic modeling variants slightly outperform a simple baseline of averaged book scores. The obtained results suggest that content features can potentially improve hybrid book recommender systems.

Key words: book recommendation, content-based recommendation, topic modeling, tf.idf

Слова и темы: представление содержания книг в задачах рекомендации

Колесниченко Л.Ю.[1] (larkkinn@gmail.com), Браславский П. И.[2,3] (pbras@yandex.ru)

[1] Университет ИТМО, Санкт Петербург, Россия
[2] НИУ ВШЭ, Санкт Петербург, Россия
[3] JetBrains Research, Санкт Петербург, Россия

Работа описывает разведочное исследование, посвященное рекомендации книг на основе контента. В исследовании используется большой набор данных пользовательских оценок книг и тексты книг. Мы экспериментируем с несколькими вариантами тематических моделей и представлением текста с помощью весов tf.idf. Предсказание на базе вариантов тематического моделирования незначительно превосходит усредненные рейтинги книг. Полученные результаты демонстрируют, что использование признаков на основе содержания может улучшить качество гибридных подходов к рекомендации книг.

Ключевые слова: рекомендация книг, рекомендации на основе содержания, тематическое моделирование, tf.idf

1 Introduction

Nowadays people have a wide variety of options for entertaining content consumption. Despite this fact, traditional book reading is still very popular. As Pew Research Institute reported in 2016, 73% of Americans have read at least one book in 12 months.[1] According to a recent survey by Russian Public Opinion Research Center (VCIOM), only 4% of Russians reported in 2018 that they don't read books.[2] While the very process of reading remains the same, book production, promotion, and sales have changed significantly in recent years. Online stores and mobile reading apps are becoming a major channel for book/ebook distribution. At these spots with 'endless bookshelves' readers often struggle the agony of choice, which makes recommendations essential.

[1] https://pewrsr.ch/2cfA531
[2] https://wciom.ru/index.php?id=236&uid=9338

In this study, we tackle the problem of content-based book recommendation. Unlike many other recommendable items, e.g., videos, music, or products, books can be considered as content-rich items. Books are written in a natural language, they are structured, they are usually reasonably long, and can be characterized by topic, genre, style, acting characters, and plot line. Hence, the state-of-the-art text mining methods allow for modelling of the very content of books, in addition to the widely-used modelling of item metadata. Although content-based and hybrid recommenders are exploited in many domains, book recommendations based on their content have received limited attention in the research community, possibly due to the lack of publicly accessible datasets.

We experiment with a dataset containing users' ratings and content of approximately 45,000 books. We make use of topic modeling and tf.idf vectors as content representation approaches. For each of nearly 500,000 users in the dataset, we build a regression model that predicts book ratings based on training data and evaluate them on test data. One of the options outperforms a simple baseline based on the averaged rating over all users. This result suggests that book topic can be a useful signal for recommendation tasks. In the future, we plan to model other aspects of book content such as author's style and to explore hybrid methods combining content and collaborative information.

2 Related Work

As a recent comprehensive book recommendation survey [2] shows, there is a rather limited number of studies on content-based book recommendations. In an early work by Mooney and Roy [14] recommendations are made based on book description, bibliographic data, and reviews on the online store page, the actual book content is not considered. Givon and Lavrenko propose a method for predicting book tags based on their text and conduct an experiment on a small collection of 146 books [7]. Vaz *et al.* combine collaborative filtering and text-based features such as vocabulary richness, POS-tag bigrams, most frequent words, as well as LDA-based representation [21]. The results suggest that content features alone cannot outperform collaborative filtering, however, their combination improves the quality of book recommendations. McAuley and Leskovec combine users' ratings and topics extracted from their reviews to recommend a wide range of products, including books [12]. Related problems of textual item recommendations, e.g., scientific papers [8], have used a representation of the textual content as one of the inputs for the recommender.

Latent Dirichlet Allocation (LDA) for topic modeling has become a widespread approach in various NLP tasks since its inception in 2003 [4]. The main idea of the method is to model text generation process. There is a distribution over topics and distributions of words over each topic. The generative process first picks a topic, then a word based on the distributions. For example, Jockers applied LDA to a large literary corpus to study book similarities [10]. In our research we use a similar topic modeling approach, implemented in the BigARTM library [22].

In this study, we use a subset of *Imhonet* dataset. The original dataset contains users' ratings on books, movies, video games, and perfumes. The dataset was used in several studies, for example addressing cross-domain recommendations [17].

3 Data

In this study, we use a large collection ratings from *Imhonet* users. *Imhonet* is a Russian recommendation service that was active in 2007–2017. The dataset consists of ratings filled in by users for different products: movies, games, books, and perfume. In the dataset we can

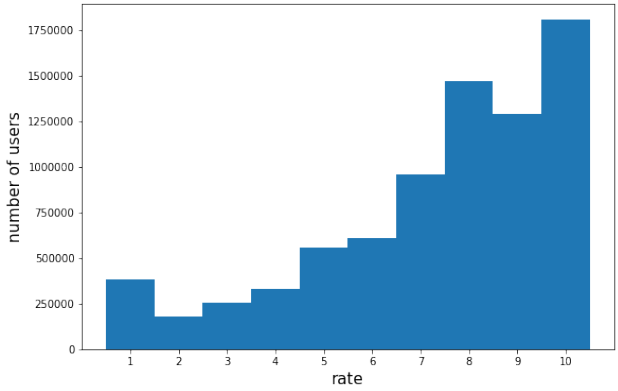

Figure 1: The distribution of book ratings in the *Imhonet* dataset.

see the ratings given by a user at a certain time for a certain product. Some users tend to rate the product multiple times, so the most recent rating is taken. Subsequently, we choose a subset of ratings for Russian books. The ratings are fully anonymized. Statistics of the original dataset are shown in the second column of the Table 1.

Table 1: *Imhonet* dataset statistics before and after matching of the books titles and filtering

	original	*content*
users	549 505	496 054
items	168 391	44 843
# records	23 007 804	8 009 408
density	0.000 24	0.000 36
avg. # per user	41	16
avg. # per item	136	179

We matched the original *Imhonet* dataset with a large online collection of ebooks based on author and title, which retain about one fourth of the books from the original datset.

Books are given ratings from 1 to 10, their distribution is present in Figure 1. Score 10 turns out to be the most popular one. As one can see from the histogram, the number of ratings increases almost linearly along with the rating value, though score one is quite popular. We excluded users who rated less than 10 and more than 1,000 books. Using content representation solves the cold start problem for books, but not for users, so we exclude these users too: 10 books per user is a widely used threshold [24], [3]. Statistics of the filtered dataset can be found in the third column of the Table 1.

Each user's ratings have been split into train:validation:test (70:20:10); so train, validation and test are balanced in terms of users. Constructed this way, train set contains $41,315$ unique books, validation set – $31,164$, and test set contains $24,558$ books.

For further usage, book text in fb2 format was cleared of tags, images, and links. Then, punctuation and stop-words were removed from the book: nltk.stopwords package [11] for Russian language was used for this purpose. In each book, we considered first 100,000 characters. Finally, every book was lemmatized using MyStem [20] with contextual disambiguation option.

4 Topic modeling

Topic modeling is a collection of methods that aim to represent documents and words in a low dimension space, usually by factorizing the Term Frequency (TF) matrix. In this study we adopt Additive Regularization for Topic Modeling (ARTM) [23] – a state-of-the-art topic modeling algorithm. ARTM is aimed to fix problems of two historically most widely used methods: LSA, Latent Semantic Analysis, [6] and LDA [5]. All three methods pick the decomposition of term frequencies matrix M into product of two matrices. LDA relies on prior assumptions on the distribution that produces matrices' columns. LSA, on the other hand, makes no such assumptions and optimizes the decompositons matrices directly, which makes the model more flexible than LDA at the cost of stability. ARTM also makes no prior assumptions on the distribution: instead, it uses several regularizers, which makes it significantly more stable than LSA, while the flexibility is achieved by applying different sets of regularizers. We perform topic modeling for 300 and 600 topics, as suggested in [10]. However, several options are taken into consideration: firstly, we use BigARTM over term frequency matrix. Topics' interpretablity has been evaluated manually, and a tendency of several topics consisting of names only or consisting of names mixed with words characteristic for a topic has been noticed. To avoid this case, as it potentially leads to less successful book recommendation based on BigARTM representation of content, we have removed a set of common names from the books.

In order to evaluate the correctness of topic modelling, the topic interpretability is introduced. In other words, topic interpretability means how integral and significative a topic is. 300 topics modelled by BigARTM have been evaluated manually and, as a result, labels for each topic have been elaborated. It has been of a help, as at first a considerable percentage of the observed 300 topics consisted of names only. After introducing into the text processing pipeline common names extraction, the topics quality increased.

There is a big variety of formal criteria aimed to evaluate topic coherence: [15, 13, 1]. We have chosen a state-of-the-art method described by Lau *et al.* [9]. The metric's purpose is to automatically measure the semantic interpretability of a topic by calculating similarity scores between the words within it. A number (usually 6) of most probable words for a topic is taken, then pairwise similarity scores between these words are calculated and the final coherence of the topic is a median score. Lau *et al.* [9] used normalized pointwise mutual information (NPMI) based on words co-occurrence within a sliding window over the text corpus as similarity score, and we have employed this method as the most conventional one. The results of this evaluation have not only proved what had been observed by a manual assessment (that the most effective BigARTM topic modeling is achieved when we have 300 topics with names extracted) but have been also used for cleansing the topics set.

The common practice is to sort topics by observed coherence and to filter out the least coherent topics to leave only top N. One can even use N (or the coherence threshold) as an optimization parameter and vary it. Currently, three variants are tested: recommendation using an entire set of topics, recommendation with $N = 100$ and $N = 50$.

TF.IDF is a classical method of representing text document content [18] with word frequencies normalised to their inverse frequencies in the corpus. In our explorative study, we use tf.idf representation with subsequent singular value decomposition [19] into 300 components.

5 Results

A comparative study of two book representation methods can be observed in Tables 2 and 3, where the books most close to "Angels and Demons" by Dan Brown and "Winter Queen" by

Boris Akunin are present. For both books the nearest ones (in terms of BigARTM proximity) include the books written by same author: "The Da Vinci Code" and "The Lost Symbol" for Dan Brown, and "Planet Water" for Boris Akunin. The second interesting observation is that for the book "The Winter Queen" among the books close by BigARTM proximity there are items "The Dangerous Surname" and "The Holy Poison" written by Anton Chizh, who also writes historical detective stories.

Table 2: Books close to "Angels and Demons" by Dan Brown in different text representations

BigARTM	*tf.idf*
The Da Vinci Code **(Dan Brown)**	The Last Fashion. Gilyarovsky and Lamanova (Andrey Dobrov)
The Lost Symbol **(Dan Brown)**	The Border Lord and Lady (Bertrice Small)
Orion (Ben Bova)	Secrets of a Wedding Night (Valerie Bowman)
The Starlight (Michael Puhov)	The Summer With Mary-Lou (Stefan Casta)
The Blind Geometer (Kim Stanley Robinson)	A Rope for Fenrir (Dmitry Kazakov)
The Key to Irunium (Kenneth Bulmer)	The Legal History (O. Omelchenko)
The Mind Net (Herbert Werner Franke)	La Bataille de Poitiers (J. Deviosse)
Skeletons In the History's Cupboard (Anatoly Wasserman)	Peter God (James Curwood)
The Engineer Graves' Secret (Yuri Tupitsin)	Scouting Is Our Bread (Albert Baikalov)

Table 3: Books close to "The Winter Queen" by Boris Akunin in different text representations

BigARTM	*tf.idf*
A Corpse on an English Lawn (Yulia Oleynikova)	Hard To Be a Student (Margarita Blinova)
John the Baptist's Car (Elena Basmanova)	In Kislovodsk (Vasily Grossman)
Planet Water **(Boris Akunin)**	One Night in Paris (Kayla Perrin)
The Dangerous Surname (Anton Chizh)	She read lips (A. and S. Litvinov)
The Holy Poison (Anton Chizh)	Beast in the Ocean (Andrey Dashkov)
Finita la Comedia (Irina Melnikova)	Silvant (Veniamin Kaverin)
Cars in the Russian Empire (Ivan Kramer)	Forever mine (Charlene Raddon)
Moscow Scenes (Mikhail Bulgakov)	Tilly Trotter (Catherine Cookson)
Before the hour (Irina Glebova)	Infernal Merzenarius (Alexey Yadov)

In order to evaluate the quality of obtained text representations, we adopt the ratings prediction task. Our approach is based on regression, i.e. learning to predict ratings that users give to items in the test dataset. We adopt mean average absolute error (MAAE) as a quality metrics. The prediction procedure is ultimately simple: for each user we fit an individual linear regression on their train subset and then use this regressor and clipping to the range from 1 to 10 to predict rates on the user's test dataset (via scikit-learn package [16]). The goal of this work is not to construct a state-of-the art recommendation system that would achieve the best result on the *Imhonet* dataset, but to investigate which representation of the content best suits the recommendation tasks and to gain insights into the dataset structure. Thus, we deliberately choose the most primitive method for recommendation, for the distinction between using different content representations not to be masked by some fluctuations of a more sophisticated recommendation algorithm. The results are shown in the Table 4. As one can see, BigARTM outperforms tf.idf, and filtering topics with the aid of observed coherence further improves prediction accuracy, so that the regression on BigARTM with 100 highly coherent topics outperforms the baseline, which is a trivial collaborative filtering algorithm: each book is always rated with its mean rate (calculated on the train dataset). On the other hand, if we filter topics too strictly, the accuracy deteriorates.

Table 4: Mean average absolute error (MAAE) for recommendations by predicting ratings with clipped linear regression using different content representations.

Method	*content*
baseline	2.094
TF.IDF (+SVD)	2.185
BigARTM	2.163
BigARTM + top100 coherence	**2.089**
BigARTM + top50 coherence	2.122

6 Conclusions

In this exploratory study, we conduct an experiment on content-based book recommendation using a large collection of users' rating and actual book texts. We have found out that topic modeling accurately represents books' content and thus provides a valuable signal for book recommendations. In particular, coherence-based topic filtering allows to outperform a baseline of averaged users' rating. We will continue to explore other facets of books' content, for instance – authors' style. We will also address combination of collaborative and content-based approach in a hybrid recommender.

References

[1] Nikolaos Aletras and Mark Stevenson. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 13–22, 03 2013.

[2] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. A survey of book recommender systems. *Journal of Intelligent Information Systems*, 51(1):139–160, 2018.

[3] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz. Authorship identification for literary book recommendations. In *COLING*, 2018.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003, 2003.

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[7] Sharon Givon and Victor Lavrenko. Predicting social-tags for cold start book recommendations. In *Proceedings of the third ACM conference on Recommender systems*, pages 333–336, 2009.

[8] Prem K Gopalan, Laurent Charlin, and David Blei. Content-based recommendations with poisson factorization. In *Advances in Neural Information Processing Systems*, pages 3176–3184, 2014.

[9] Jey Han Lau, David Newman, and Timothy Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pages 530–539, 01 2014.

[10] Matthew L Jockers. *Macroanalysis: Digital methods and literary history*. University of Illinois Press, 2013.

[11] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. *CoRR*, cs.CL/0205028, 2002.

[12] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[13] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

[14] Raymond J. Mooney and Loriene Roy. Content-based Book Recommending Using Learning for Text Categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 195–204, 2000.

[15] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic Evaluation of Topic Coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 100–108, 01 2010.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Shaghayegh Sahebi and Peter Brusilovsky. Cross-domain collaborative recommendation in a cold-start context: The impact of user profile size on the quality of recommendation. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 289–295. Springer, 2013.

[18] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[19] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, Minnesota Univ Minneapolis Dept of Computer Science, 2000.

[20] Ilya Segalovich. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. In *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications*, Las Vegas, Nevada, USA, 2003.

[21] Paula Cristina Vaz, Ricardo Ribeiro, and David Martins de Matos. Book Recommender Prototype Based on Author's Writing Style. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR'13, pages 227–228, 2013.

[22] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[23] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323, 2015.

[24] Yiwen Wang, Natalia Stash, Lora Aroyo, Laura Hollink, and Guus Schreiber. Using semantic relations for content-based recommender systems in cultural heritage. In *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516*, pages 16–28. CEUR-WS. org, 2009.