ABSTRACT
We opensource under CC BY 4.0 license Lib-SibGMU – a university library circulation dataset – for a wide research community, and benchmark major algorithms for recommender systems on this dataset. For a recommender architecture that consists of a vectorizer that turns the history of the books borrowed into a vector, and a neighborhood-based recommender, trained separately, we show that using the fastText model as a vectorizer delivers competitive results.

CCS CONCEPTS
• Information systems → Recommender systems; Clustering and classification; • Computing methodologies → Information extraction.

KEYWORDS
natural language processing, short text classification, neural network, library circulation dataset

1 INTRODUCTION
Libraries seek their new place in the word. They have been the public source of information and knowledge for centuries and played a critical role in public education. With the advent of the internet the latter has become the information source of choice, "from the leisure of your home" or workplace. Libraries started to lose their audience.

Regaining the audience became critical for the libraries, both as internal desire and as a Key Performance Indicator (KPI) imposed by the funding authorities. On the other hand, decline in the audience led to the matching decline in the prestige of the librarian profession. Knowledgeable and trusted librarians become increasingly rare. Thus, the quality of advice by a librarian to a patron declines, leading to a vicious cycle. In this situation a good recommender system can both directly provide the library patron with quality book advice, and improve the advice by a librarian. Breaking the vicious circle helps libraries to re-gain the audience, at least, partially. In particular, the goal of an university library recommender system is to recommend relevant additional books to the library patron (student of faculty) and the librarian, with the KPI of more books lent. machine Learning (ML) based systems (such as recommender) research and development requires publicly available benchmarks so that the approaches and algorithms can be compared directly, apples to apples. To the best of our knowledge, no publicly available library circulation benchmark or dataset existed until now.

1.1 Our contribution
The contribution of our work is twofold. First, we opensource under CC BY 4.0 license Lib-SibGMU – an anonymized library circulation dataset – for a wide research community¹, so the algorithms for library recommender systems can be compared in a principled manner. Second, for a recommender architecture that consists of a vectorizer that turns the history of the books borrowed into a vector, and a neighborhood-based recommender, trained separately, we show that using the fastText model [11] as a vectorizer delivers competitive results for the Lib-SibGMU dataset.

2 RELATED WORK
2.1 Datasets
Some of the best known recommender system datasets are based on the explicit feedback, for example, MovieLens [20], Netflix Prize [1], Epinions [41], Flixter [28] and Amazon Fine Foods [36]. More recent HotelRec [9], KuaiRec [16] and SweetRS [30] also fall into this category. Implicit feedback / session based recommender system datasets span several important application areas, including location-based

¹The dataset is available at https://github.com/NTRLab/Lib-SibGMU-recommendation-dataset
user behavior (Gowalla [13] and Foursquare [54–56]), news (Microsoft News Dataset [51]), e-commerce (YOUCHOOSE [10], DIG-INETICA [5], RetailRocket [7], Tmall [6] and Ta Feng [3]), advertising (Criteo [2], Avazu [4] and iPinYou [34]) and music (LFM-1b [42, 43]). No public datasets with library circulation data are known to us.

2.2 Algorithms

The focus of work on recommender systems in the 1990s was on internet-related recommendations. Even before the massive adoption of WWW Goldberg et al. [19] and Resnick et al. [39] suggested recommender systems for email and internet discussion groups, respectively.

Although the recommender systems for librarians have been discussed at least since early 90s (see, for example, de la Peña McCook, Rolstad, and Gonsalves, [14]), it took to the early 2000s when the first implementations of recommender systems for libraries patrons took place. At that time Hwang et al. in China [27], Geyer-Schulz, Neumann and Theide in Germany [17] and Mooney and Roy [37], Heylighen and Bollen [24], Huang et al. [26] and Smeaton and Callan [44] in the US have implemented the first systems. Early work also include that of Karaulsh and Hasanov [21, 29], Liao et al. [35], Chen and Chen [12], Tsai and Chen [47], Yang and Li [57] and Spiering and Mönich [45].

Many of the later works treat library recommender systems as systems with explicit feedback, where user ratings are available for some of the books read by a user. For example, Vaz et al. [49] propose a hybrid system that uses ratings for both books and authors. Xiao and Gao [52] study approaches to building a book recommender system based on explicit feedback. Ghadling et al. [18] build a hybrid recommendation system that combines individual recommendation algorithms. Anoop and Ayush Ubale [8] in their recommender system design only recommend the top-rated books, completely ignoring the well-known diversity problem in ranking items.

On the implicit feedback side of the spectrum, Fu, Zhang and Seiminn in a relatively early work [15] advocate the use of user-based collaborative filtering for a library recommender system. Xu [53] designs a student library recommender system based on SVD of the interaction matrix. Ziegler and Shrape [58] use circulation data collected through the collection-management software package, Aeon, to automate recommendations. Knyazeva et al. [31, 32] study several approaches to collaborative filtering for a university library recommender system based on implicit feedback data. Morawski et al. [38] propose a hybridization of collaborative filtering with a content filter using a fuzzy taste vector for a consortium of rural libraries where the item sets for recommendations are limited by supply rather than by readers’ interests.

In a more distant case, Rhanoui et al. [40] proposed using a hybrid recommender system for patron driven library acquisition and weeding.

The works most similar to ours algorithmically are Valcarce et al. [48] and Zubchuk et al. [59]. Valcarce et al. [48] have first suggested using a vector space with language models in a neighborhood-based recommender setting. Zubchuk et al. [59] have recently developed a successful practical recommender using fastText embeddings.

![Figure 1: The distribution of users interactions with items by date](image)

From a mathematical standpoint, library recommender systems with implicit feedback are similar to session-based and sequential recommender systems (see a review by Wang et al. [50]).

3 THE DATASET

The dataset was obtained from the circulation history database of a medical school (Siberian State Medical University) library. The fields containing descriptions of documents, the time when the document was lent and returned, the encrypted (for anonymization purposes) user identifier, and the inventory number of the document issued to the user were unloaded. The history of circulation spans from September 2013 to December 2021. During this time, the users of the medical school repeatedly received books and other resources both for use in the educational process, and for conducting scientific research and self-development. The database of users includes students pursuing bachelors, masters, and graduate degrees, employees and professors of the medical school, as well as professionals from other medical institutions who had access to the resources of this medical library.

The dataset file – medlib – contains the records of users’ interactions with the library. The file is in XML format. Each interaction has a root tag <record> with the fields <field tag=X>, where X can be:

- "ID" - a unique identifier of user;
- "ID book" - a unique identifier of book;
- "name book" - string that contains book’s author, name and brief description;
- "category reader" - category of a reader, e.g. student, professor, etc.;
- "birth YYYY" - user’s birth year;
- "date circulation YYYYMMDD" - date in yyyy/mm/dd format when the event took place;

The textual field data are in Russian, encoded in UTF-8. The dataset contains 153,364 entries related to 7,149 users and 2,664 books. Share of students among the interaction data is about 97%. The distribution of users’ interactions with items in the dataset is presented on the Figure 1 below.
About 20\% of library visits are uninformative data (e.g. consultation, book returns, exhibition visits). The average number of books borrowed by the students is 18. The average number of books borrowed by the other user types is much smaller - about 7. The distribution of the number of books borrowed by students is presented on the Figure 2 below.

As the dataset represents past activities of the students active at the time the dataset was exported, events long past only relate to activities of the students who have been admitted long ago, then took a break in their studies - medical, military or otherwise, and then returned to studies. The distribution of the students’ activities in the dataset by term and year is represented on Figure 3.

The main peculiarities of the dataset are:
- high repeatability - same instances (specific user picks specific book) appear often, as users take a book several times (e.g. to prepare for exams);
- term dependence - the majority of events falls to the period from 2018 to 2022, and, which is more important, to a specific time - the last month of the academic semester (except for 2020, when visiting the library was limited due to COVID-19).

The facts presented above allow us to conclude that in creating a recommender system, the main attention should be paid to students during the end of the semester and preparation for exams period, and conclusions about the most suitable books can be made based on the previous groups of students who have already passed the presented course.

4 ALGORITHMS AND BENCHMARKS

We have benchmarked classical algorithms of Matrix Factorization: LightFM [33] and Alternating Least Squares (ALS - implementation based on [25, 46] from https://github.com/benfred/implicit) on the dataset. However, peculiarities of the dataset make these approaches less effective than expected, and HitRate@10 for these algorithms did not exceed 0.17, while MAP@10 was under 0.042 (see Table 1). We have also benchmarked well-known Neural Collaborative Filtering (NCF) [23] with almost twice better, but still imperfect outcomes.

We then hypothesized that a neighborhood-based algorithm would be a better for this dataset. We assumed that the users can be split into clusters that represent students with specific specialization and at the specific year of studies. Thus we have benchmarked K-Means and DBSCANS clusterization to split the users into some clusters and then recommend items that other users from this cluster interacted with the most at the same semester. We have first implemented this idea using one-hot encoding of users’ circulation history. To choose the optimum number of clusters, we used the argmax of the silhouette coefficient \((b - a)/\max(a, b)\), where \(a\) is the average intra-cluster distance and \(b\) is the average distance to the nearest cluster. Unfortunately, both K-Means and DBSCAN performed only marginally better than LightFM and worse than NCF (see Table 1, lines KMeans+OHV and DBSCAN+OHV).

This prompted us to explore neighborhood-based models that contain two parts (see Zubchuk et al. [59]): a vectorizer that transforms the user circulation history into a vector in some semantically meaningful vector space, and a neighborhood-based recommender, operating with the vectors from the previous step. The core idea of this approach is to find a vector space that describes the degree of similarity of the users much better than one-hot encoding. To implement this idea we have used two important insights. First, since the dataset contains different editions of the same book, it is important to treat these editions as the same book despite different ISBNs. However, in the basic one-hot data structure, these editions are essentially two different entities with different ISBNs, resulting in orthogonal one-hot vectors. Actually, while the number of books with two or more editions is just 14\%, the number of interactions with such books is 35.6\%, meaning that books with two or more editions are more then twice as popular (see also Figure 4 and Figure 5 for comparison of distributions of books and their interactions by number of editions).

Second, the same words in the book titles mean that the books are probably related. This may be less true for fiction (William Golding’s "Lord Of The Flies" and J.R.R. Tolkien’s "The Lord Of The Rings" are completely unrelated despite having the majority
The use of a language model in the vectorizer allows us to address both insights above. We trained fastText in an unsupervised manner using all the available circulation history and collapsing the titles of all the books borrowed by a single person into a single line. Using K-Means for unsupervised clustering on vectorized embeddings provided by the language model allowed us to achieve HitRate@10 of 0.374 and MAP@10 of 0.190. To further disentangle the influence of embedding from the recommendation head we have trained a simple neural network – a MultiLayer Perceptron (MLP) – for classification of fastText embedding vectors. The results were somewhat better than K-Means, demonstrating the superiority of the embedding chosen.

Finally, we have decided to use kNN to achieve more fine-grained user similarity. This has allowed us to improve the core metrics by further 50% with respect to K-Means, demonstrating the superiority of the embedding chosen.

5 CONCLUSIONS

In this paper, we presented the Lib-SiBGMU – an anonymized university library circulation dataset. The novel nature of this publicly available dataset poses new challenges as the classical algorithm performance is suboptimal and requires development of novel efficient algorithms. One of such algorithms is presented in this article: a recommender that consists of a fastText vectorizer that turns the history of the books borrowed into a vector, and a neighborhood-based recommender, trained separately, achieving HitRate@10 of 0.544 and MAP@10 of 0.269.

6 ACKNOWLEDGMENTS

The authors are grateful to

- colleagues at NTR Labs Machine Learning Research group for the discussions and support, especially to Sergey Men'shchikov, who traditionally managed our industry implementations of recommender engines and Vlad Bolshakov;
- Prof. Sergey Orlov and Prof. Oleg Zmeev for the computing facilities provided.

REFERENCES

[1] 2006. The Netflix Prize. https://www.kaggle.com/datasets/netflix-inci-netflix-prize-data
[2] 2014. Display Advertising Challenge. https://ailab.criteo.com/resources/
[3] 2014. Ta Feng Grocery Dataset. https://www.kaggle.com/datasets/charanjivdan09/ta-feng-grocery-dataset
[4] 2015. Click-Through Rate Prediction. https://www.kaggle.com/c/avazu-ctr-prediction/data
[5] 2016. CTRM Cup 2016 Track 2: Personalized E-Commerce Search Challenge. https://competitions.codalab.org/competitions/11161
[6] 2016. User Behavior Data on Taobao/Tmall IJCAI16 Contest. https://tianchi.aliyun.com/dataset/dataDetail?dataId=53
[7] 2017. Retailrocket recommender system dataset. https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset
[8] A Anoop and N Ayush Ubale. 2020. Cloud Based Collaborative Filtering Algorithm for Library Book Recommendation System. In 2020 Third International Conference...
Zubchuk et al.

[50] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z. Sheng, Mehmet A. Orgun, and Defu Lian. 2021. A Survey on Session-Based Recommender Systems. ACM Comput. Surv. 54, 7, Article 154 (Jul 2021), 38 pages. https://doi.org/10.1145/3465401

[51] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 3597–3606. https://doi.org/10.18653/v1/2020.acl-main.331

[52] Jingshan Xiao and Wenli Gao. 2020. Connecting the Dots: Reader Ratings, Bibliographic Data, and Machine-Learning Algorithms for Monograph Selection. The Serials Librarian 78, 1-4 (2020), 117–122. https://doi.org/10.1080/0361526X.2020.1707599

[53] Chunyan Xu. 2017. A Personalized Recommender System Based on Library Database. International Journal of Emerging Technologies in Learning (iJET) 12, 12 (Dec. 2017), pp. 134–141. https://doi.org/10.3991/ijet.v12i12.7964

[54] Dingqi Yang, Daqing Zhang, and Bingqing Qu. 2016. Participatory Cultural Mapping Based on Collective Behavior Data in Location-Based Social Networks. ACM Trans. Intell. Syst. Technol. 7, 3, Article 30 (Jan 2016), 23 pages. https://doi.org/10.1145/2814575

[55] Dingqi Yang, Daqing Zhang, Zhiyong Yu, and Zhiwen Yu. 2013. Fine-Grained Preference-Aware Location Search Leveraging Crowdsourced Digital Footprints from LBSNs. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Zurich, Switzerland) (UbiComp ’13). Association for Computing Machinery, New York, NY, USA, 479–488. https://doi.org/10.1145/2493452.2493464

[56] Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. 2015. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Transactions on Systems, Man, and Cybernetics: Systems 45, 1 (2015), 129–142. https://doi.org/10.1109/TSMC.2014.2327053

[57] Yan Yang and Jian Zhong Li. 2005. Interest-based Recommendation in Digital Library. Journal of Computer Science 1, 1 (Mar. 2005), 40–46. https://doi.org/10.3844/jcssp.2005.40.46

[58] Scott Ziegler and Richard Shrake. 2018. PAL: Toward a Recommendation System for Manuscripts. Information Technology and Libraries 37, 3 (Sep. 2018), 84–98. https://doi.org/10.6017/ital.v37i3.10357

[59] Eduard Zubchuk, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2022. Using a Language Model in a Kiosk Recommender System at Fast-Food Restaurants. https://arxiv.org/abs/2202.04145 (02 2022). Issue 2202.04145.

[60] Eduard Zubchuk, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. Efficiency of short text classifiers for payment classification. In 2021 International Conference on Information Technology and Nanotechnology (ITNT). IEEE, 1–4.