Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 26 (2023), No. 6, pp. 1705–1716
DOI : 10.47974/JDMSC-1617

A landscape view of news recommendation systems based on MIND dataset

Niran A. Abdulhussein *
Department of Computer Science
Faculty of Computer Science and Mathematics
University of Kufa
Najaf
Iraq

Ahmed J. Obaid †
Department of Computer Science
Faculty of Computer Science and Mathematics
University of Kufa
Najaf
Iraq

and

National University of Science and Technology
Thi-Qar
Iraq

Abstract

Nowadays, it's a very important way for researchers and all people to find their desired meaning instead of searching for a specific topic. Recommendation systems are strategies to solve the problems of search, finding or reducing the time of interest content for users under complex information environments. R.S. can show us the related results that are close to what we desire. In this paper, we list rates of the most used techniques applied over MIND dataset and show the results and comparisons among these techniques, we have proposed our analysis on that dataset, which has been collected from Microsoft in 2019, and

* *E-mail:* nirana.alasadi@student.uokufa.edu.iq (Corresponding Author)
† *E-mail:* ahmedj.aljanaby@uokufa.edu.iq

© Taru Publications

we proposed new techniques and explain the views of researchers in previous studies and the techniques in R.S. that depend on.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Recommendation system, Click behavior, MIND dataset.*

1. Introduction

In general, with the increasing amount of data on the internet every day, the internet has become an important part of our daily routine. Therefore, this data must be filtered to be selected correctly. The internet has become the best and fastest way to read news and reports online on (mobile, iPad, or laptop) [1]. R.S. is very important for online news platforms to help users reduce information overload of users and improve reading of news experience because there are enormous amounts of data on the internet, News recommendation systems allow us to filter news and present it to the users according to their interests. R.S. is widely used in many applications such as Product Recommendations, Movie Recommendations, News Articles, etc. [2].

R.S. is based on E-commerce view as a tool that helps users to search through knowledge records related to users' attention and preferences. R.S. was defined as means of helping and increasing the social process to use the recommendations of others to make choices [3]. R.S. techniques are: content-based filtering (CB) that based on features extracted from the content of news and user behaviors on news, collaborative filtering (CF) which is based on IDs of users and news, and hybrid filtering a hybrid of two previous methods that based on content-based collaborative filtering [4] as shown in Figure 1.

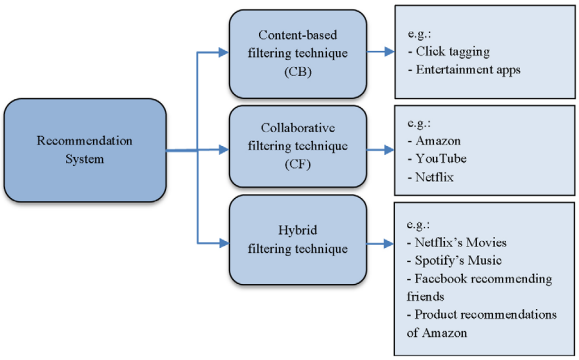

Figure 1

Recommendation techniques [3].

News recommendation systems are based on log behaviors and analysis impressions of users from previous clicks. So, our study in this paper is about MIcrosoft News Dataset (MIND[1]) [5], it's a large-scale public news recommendation dataset, this comprises the news impression records of 1 million Microsoft News readers over the course of six weeks (from 12 October to 22 November 2019). The samples from the last week are set aside for testing, while those from the first five weeks are used for training and validation. Table 1 displays the comprehensive statistics for this dataset.

Table 1

Detailed dataset statistics [5].

No. of Users 1,000,000	Avg. of No. of words 11-12
No. of News 1,61,013	No. of Click behaviors 24,155,470
No. of Impressions 15,777,377	No. of Topic categories 20
No. of Entities 3,299,687	Avg. of No. of entities 16-17

Each news article in the MIND dataset contains a news ID, a title, an abstract, a body, and a category label. As evaluation metrics, MIND is a metric in AUC, MRR, nDCG@5, and nDCG@10 [5].

2. Related Work

In this section, we review the opinions of researchers in previous studies about the relevant work related to R.S. research with the increasing data used on the internet, and a large number of recommendation algorithms used.

U Kang, 2020 [6] the author views a summary of the work on MIND competition which was organized by Microsoft News and Microsoft Research Asia team. The proposed method NRMS+ NAML+ Glove (NNG) and NRMS+ NAML+ BERT (NNB). They select two NNG models and four NNB models as final models to generate ensembles. Final model was achieved AUC 0.7114, MRR 0.3568, nDCG@5 0.3916, nDCG@10 0.4485.

Danyang Liu and et. al., 2020 [7] presented the Knowledge-aware Representation Enhancement model for news Documents (KRED). The advantages of this framework are three-fold. The authors got the overall evaluation is AUC 0.6914, nDCG@10 0.2684.

1 https://msnews.github.io/

Chuhan Wu and et. al., 2021 [8] reviewed Distance-Aware Transformer (DA-Transformer), this model depends on a Transformer that achieved great success in Natural Language Processing (NLP) field. They got results of the experiment are AUC 68.32, MRR 33.36, nDCG@5 36.34, nDCG@10 42.07.

Chuhan Wu and et. al., 2021 [9] submitted Unified method for Recall and Ranking in news recommendations (UniRec). In Ranking use Multi-head Self-attention networks, and in Recall take user embedding for ranking. The evaluation of Ranking performance is AUC 68.41, MRR 33.50, nDCG@5 36.47, nDCG@10 42.26.

Chuhan Wu and et. al., 2021 [10] Pre-trained Language Models (PLMs) are powerful for Natural Language Understanding. They used negative sampling techniques from impression logs. They got results in NRMS-UniLM method is AUC 70.64, MRR 35.39, nDCG@5 38.71, nDCG@10 44.38.

Tao Qi and et. al., 2021 [11] suggested Knowledge-aware Interactive Matching (KIM) framework method. They got AUC 67.13, MRR 32.08, nDCG@5 35.49, and nDCG@10 41.79.

Tao Qi and et. al., 2021 [12] conducted Hierarchical user interest modeling for Personalized News Recommendation (HieRec). The result is AUC 67.95, MRR 32.87, nDCG@5 36.36, nDCG@10 42.53.

Yu Tian and et. al., 2021 [13] suggested a knowledge pruning based recurrent graph convolutional network (KOPRA). KOPRA is devised to identify the relevant entities from both users clicked history and knowledge graph. They got results of the model overall were AUC 68.80, MRR 34.64, nDCG@5 41.59, nDCG@10 44.89.

Qi Zhang and et. al., 2021 [14] proposed User-News matching BERT (UNBERT), they explored the use of BERT. This experiment was conducted on two versions of the dataset. They got results in MIND-small is AUC 0.6762, MRR 0.3172, nDCG@5 0.3475, nDCG@10 0.4102, and in MIND-large is AUC 0.7068, MRR 0.3568, nDCG@5 0.3913, nDCG@10 0.4478.

Chuhan Wu and et. al., 2021 [15] discussed communication efficiency federated learning method based on Knowledge Distillation (FedKD). The goal of FedKD is to learn a strong model with less communication cost. The results are AUC 71.0, MRR 35.6, nDCG@5 38.9, and nDCG@10 44.8.

Chuhan Wu and et. al., 2021 [16] presented News-BERT, PLMs like BERT have made great progress in NLP. The goal is to learn the light-

weight student model that can keep most of the teacher's performance. They got results in AUC 70.31, MRR 34.89, nDCG@5 38.32, nDCG@10 43.95.

Chuhan Wu and et. al., 2021 [17] proposed Fastformer, a transformer is a powerful model for text understanding. An efficient Transformer model based on additive attention. Its complexity is quadratic to the input sequence length and the computational complexity can be reduced to linearity. The results are Accuracy 82.34, Macro-F 63.89, AUC 69.11, MRR 34.25, nDCG@5 37.26, nDCG@10 43.38.

Chuhan Wu and et. al., 2021 [18] presented the User-as-Graph user modelling method (UaG). Each user in UaG is represented as a personalized heterogeneous graph based on their behaviour. They got results AUC 69.23, MRR 34.14, nDCG@5 37.21, nDCG@10 43.04.

Jiahao Xun and et. al., 2021 [19] suggested a visual IMpression-aware multi-modal news Recommendation framework (IMRec) model for multi-modal features. They used two versions of the dataset. They achieved higher evaluations in FIM-IM method was in IM-MIND-Small AUC 0.6661, nDCG@5 0.3526, nDCG@10 0.4146, MRR 0.3199, and in IM-MIND-Large AUC 0.6912, nDCG@5 0.3725, nDCG@10 0.4364, MRR 0.3364.

Peitian Zhang and et. al., 2021 [20] discussed a Selective Fine-grained Interaction (SFI) framework with a learning-to-select mechanism. The biggest challenge of SFI is to select the informative history news sparsely, precisely, and efficiently. They got results in AUC 69.95, MRR 0.3503, nDCG@5 38.31, nDCG@10 43.97. SFI can improve the AUC by 2.17% over the state-of-the-art models with four times faster speed.

3. Results Discussion

In this section, we will discuss the results reached by the researchers, which we reviewed in the 2nd section. They reached the highest value of the proposed models are AUC 0.7304, MRR 0.3770, nDGC@5 0.4180, and nDGC@10 0.4718. So, we will review a table of the related work discussed in this paper, and discuss the methods and techniques that used, also the results that reached as shown in Table 2.

Table 2

Brief description of MIND News Recommendation System for previous studies.

No.	Paper info.	Year of study	Model	Dataset used	Results				Notes
					AUC	MRR	nDCG@5	nDCG@10	
1	U Kang [6]	2020	two NNG and four NNB	MIND	0.7114	0.3568	0.3916	0.4485	Power of ensembles selected two NNG and four NNB models as final models.
2	Danyang Liu and et. al. [7]	2020	KRED		0.6914			0.2684	Advantages are three-fold; It has no constraint on which specific document representation method should be used, It can utilize all the data, and It is fast.
3	Chuhan Wu and et. al. [9]	2021	DA-Transformer	AG[1], Amazon[2], SST, SNLI, MIND	68.32	33.36	36.34	42.07	This model depends on Transformer that achieved great success in NLP.
4	Chuhan Wu and et. al. [10]	2021	UniRec	MIND	68.41	33.50	36.47	42.26	Powerful of this model they used Multi-head Self-attention networks.
5	Chuhan Wu and et. al. [11]	2021	PLMs	MIND, Multilingual	70.64	35.39	38.71	44.38	They used PLMs that are powerful models for Natural Language Understanding.
6	Tao Qi and et. al. [12]	2021	KIM	MIND-Small, feeds	67.13	32.08	35.49	41.79	They applied this model on part of MIND that led to these results, but if they apply it on all MIND may be led to higher results.
7	Tao Qi and et. al. [13]	2021	HieRec	MIND, feeds	67.95	32.87	36.36	42.53	HieRec can capture diverse and multi-grained user interest.
8	Yu Tian and et. al. [14]	2021	KOPRA	MIND, Adressa	68.80	34.64	41.59	44.89	KOPRA devised to identify the relevant entities from both users history and knowledge graph to derive the user representation.

Contd...

9	Qi Zhang and et. al. [15]	2021	UNBE-RT	MIND-Small	0.6762	0.3172	0.3475	0.4102	When they apply this model on all MIND, they got a great result.
				MIND-Large	0.7068	0.3568	0.3913	0.4478	
10	Chuhan Wu and et. al. [16]	2021	FedKD	MIND, SMM4H[3]	71.0	35.6	38.9	44.8	FedKD aims to learn a strong model in a privacy-preserving way with less communication cost.
11	Chuhan Wu and et. al. [17]	2021	News-BERT	MIND, NewsRetrieval	70.31	34.89	38.32	43.95	BERT made great progress in NLP. KD techniques can compress a large PLM into a smaller one and meanwhile keep good performance.
12	Chuhan Wu and et. al. [18]	2021	Fastfor-mer	Amazon, IMDB[4], MIND, CNN/DM, PubMed	69.11	34.25	37.26	43.38	The core of Transformer model is self-attention mechanism, and its complexity is quadratic to the input sequence length and the computational complexity can be reduced to linearity.
13	Chuhan Wu and et. al. [19]	2021	UaG	MIND	69.23	34.14	37.21	43.04	Each user is represented as a personalized heterogeneous graph based on their actions.
14	Jiahao Xun and et. al. [20]	2021	IMRec	IM-MIND- Small	0.6661	0.3199	0.3526	0.4146	They got good results in FIM-IM method on part of dataset, but they got better results when used all dataset.
				IM-MIND- Large	0.6912	0.3364	0.3725	0.4364	
15	Peitian Zhang and et. al. [21]	2021	SFI	MIND	69.95	0.3503	38.31	43.97	SFI improve the AUC by 2.17% over the state-of-the-art models with four times faster speed. Despite of effectiveness improvement, the model is slow.

In Table 3, discuss the statistical analysis of most dataset used which is MIcrosoft News Dataset (MIND) [2], a large-scale public news recommendation dataset containing the news impression logs of 1 million Microsoft News users over the course of six weeks (from 12 October to 22 November 2019). MIND has two versions of the dataset, and it has 8 labels: News ID, Category, SubCategory, Title, Abstract, URL, Title Entities, and Abstract Entities.

Table 3

MIND dataset statistics analysis.

Dataset Characteristics	Training set	Testing set	Total
No. of Instances	101,527 records	120,959 records	222,486 records
No. of Definite article	345,712	412,704	758,416
No. of Auxiliary verbs	196,547	234,325	430,872
No. of pronouns	237,340	283,137	520,477
No. of prepositions	720,058	860,446	1,580,504
No. of words	14,090,156	12,177,718	26,267,874
Size	530.2 MB	605 MB	1135.2 MB

4. Dataset Representation:

Based on our study, we have seen that MIND dataset could be represented as a vector space model by using a number of columns and rows as the following table. Where rows represent the number of instances in the dataset which can be represented as I= $\{I_1, I_2, I_3, \ldots I_m\}$, where m is the total number of instances. And the columns represent the number of features extracted from the dataset after excluding No. of Definite article, No. of Auxiliary verbs, No. of pronouns, and No. of prepositions which can be represented as F= $\{F_1, F_2, F_3, \ldots F_n\}$, where n is the total number of features. So, our data in each entity in our dataset can be represented as $\{I_m F_n\}$. As shown in Table 4.

Table 4

Representation of MIND datasets.

Dataset	F_1	F_2	F_3	…	F_n
I_1	$I_1 F_1$	$I_1 F_2$	$I_1 F_3$	…	$I_1 F_n$
I_2	$I_2 F_1$	$I_2 F_2$	$I_2 F_3$	…	$I_2 F_n$
I_3	$I_3 F_1$	$I_3 F_2$	$I_3 F_3$	…	$I_3 F_n$
…	…	…	…	…	…
I_m	$I_m F_1$	$I_m F_2$	$I_m F_3$	…	$I_m F_n$

By selecting some instances which have been chosen from MIND dataset, we suggested this method to represent our data, as shown in Table 5.

Table 5

Representation of MIND datasets by selecting some instances.

News ID	Shop	Apple's	harmless	bags	fraud	good	very	device
N88753	1	0	0	0	0	0	0	0
N45436	0	1	0	0	0	0	0	0
N23144	0	0	1	0	0	1	0	0
N86255	0	0	0	1	0	0	0	0
N93187	0	0	0	0	1	0	0	0
N75236	0	0	1	0	0	1	1	0
N99744	0	0	0	0	0	0	1	0
N5771	0	0	0	0	0	0	0	1

5. Conclusion

In this paper, we are listing a brief introduction about recommendation systems and MIND dataset that was collected. We find that the dataset can be analyzed in many ways and we are discussing the recent study that applied to the MIND dataset in the related work section. Our analysis in the discussion of the results is important to represent the data by using two-dimensional presentation, we suggest some analysis ways which can give better performance and high speed with low-cost complexity of the dataset. Lastly, we have covered the related work in this survey so that the readers can get a vision of the current research practices and may start to help and develop them.

References

[1] Shaina Raza and Chen Ding, "News recommender system: A review of recent progress, challenges, and opportunities", Published online: 21 July 2021, Springer, https://doi.org/10.1007/s10462-021-10043-x, (2021).

[2] Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman, "Ch9: Recommendation Systems", Mining of Massive Datasets, Published online by Cambridge University Press, pp. 292–324 (05 December 2014), DOI: https://doi.org/10.1017/CBO9781139924801.010.

[3] F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh, "*Recommendation systems: Principles, methods and evaluation*" (2015).

[4] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie, "Personalized News Recommendation: A Survey". 1, 1 (July 2021), 40 pages, arXiv:2106.08934v2. https://doi.org/10.1145/nnnnnnn.nnnnnnn.

[5] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et. al., "MIND: A Large-scale Dataset for News Recommendation". In ACL. 3597-3606 (2020).

[6] U Kang, "MIND News Recommendation Competition" (2020).

[7] Danyang Liu, Ying Qiao, Jiun-Hung Chen, Jianxun Lian, Guangzhong Sun, Shiyin Wang, and Xing Xie, "KRED: Knowledge-Aware Document Representation for News Recommendations", Virtual Event, Brazil 2020 Association for Computing Machinery, arXiv:1910.11494v3 (2020). https://doi.org/10.1145/3383313.3412237.

[8] Chuhan Wu, Fangzhao Wu, and Yong Huang, "DA-Transformer: Distance-aware Transformer", In National Natural Science Foundation of China. arXiv:2010.06925v2 (2021).

[9] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. "Two Birds with One Stone: Unified Model Learning for Both Recall and Ranking in News Recommendation". ACM, New York, NY, USA, Article 4, 5 pages (2021). https://doi.org/10.475/123_4.

[10] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. "Empowering News Recommendation with Pre-trained Language Models". ACM, New York, NY, USA, Article 4, 5 pages (2021). https://doi.org/10.475/123_4.

[11] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. "Personalized News Recommendation with Knowledge-aware Interactive Matching", Virtual Event, Canada. ACM, New York, NY, USA, 10 pages (July 11–15, 2021). https://doi.org/10.1145/3404835.3462861.

[12] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang, "HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation", *National Natural Science Foundation of China,* arXiv:2106.04408v1 (2021).

[13] Yu Tian, Yuhao Yang, Xudong Ren, Pengfei Wang, Fangzhao Wu, Qian Wang, and Chenliang Li, 2021. "Joint Knowledge Pruning and Recurrent Graph Convolution for News Recommendation", Virtual Event, Canada. ACM, New York, NY, USA, 10 pages (July 11–15, 2021). https://doi.org/10.1145/3404835.3462912.

[14] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang and Xiuqiang He, "UNBERT: User-News Matching BERT for News Recommendation", Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), (2021).

[15] Chuhan Wu, Fangzhao Wu, Ruixuan Liu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie, "FedKD: Communication Efficient Federated Learning via Knowledge Distillation", arXiv:2108.13323v1 (2021).

[16] Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu, "NewsBERT: Distilling Pre-trained Language Model for Intelligent News Application", arXiv:2102.04887v2, National Natural Science Foundation of China under Grant numbers 82090053, 61862002, and Tsinghua-Toyota Research Funds 20213930033 (2021).

[17] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie, "Fastformer: Additive Attention Can Be All You Need", arXiv:2108.09084v6 (2021).

[18] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie, "User-as-Graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation", Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21). National Natural Science Foundation of China (NSFC) (2021).

[19] Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu, 2021. "Why Do We Click: Visual Impression-aware News Recommendation",

Virtual Event, China. ACM, New York, NY, USA, 10 pages (October 20–24, 2021). https://doi.org/10.1145/3474085.3475514.

[20] Peitian Zhang, Zhicheng Dou, and Jing Yao, "Learning to Select Historical News Articles for Interaction based Neural News Recommendation", Conference'17, July 2017, Washington, DC, USA, 2021 Association for Computing Machinery. ACM ISBN 978-1-4503-XXXX-X/18/06. arXiv:2110.06459v1 (2021). https://doi.org/10.1145/1122445.1122456.

[21] Shubham Sharma & Ahmed J. Obaid. Mathematical modelling, analysis and design of fuzzy logic controller for the control of ventilation systems using MATLAB fuzzy logic toolbox, *Journal of Interdisciplinary Mathematics*, 23:4, 843-849 (2020), DOI: 10.1080/09720502.2020.1727611.

Received March, 2022

Revised April, 2022