MIND: A Large-scale Dataset for News Recommendation

(Summary)

Abstract

Constructed from the user click logs of Microsoft News, MIND contains 1 million users and more than 160k English news articles, each of which has rich textual content such as title, abstract and body.

Keywords: news recommendation

1 Introduction

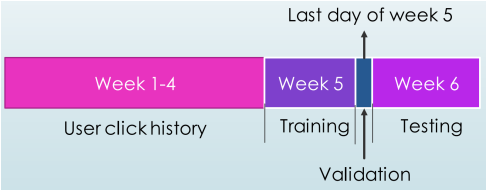
We implement many state-of-the-art news recommendation methods originally developed on different proprietary datasets, and compare their performance on the MIND dataset to provide a benchmark for news recommendation research.
The experimental results show that a deep understanding of news articles through NLP techniques is very important for news recommendation. Both effective text representation methods and pre-trained language models can contribute to the performance improvement of news recommendation. In addition, appropriate modeling of user interest is also useful.

2 Related Work

3 MIND Dataset

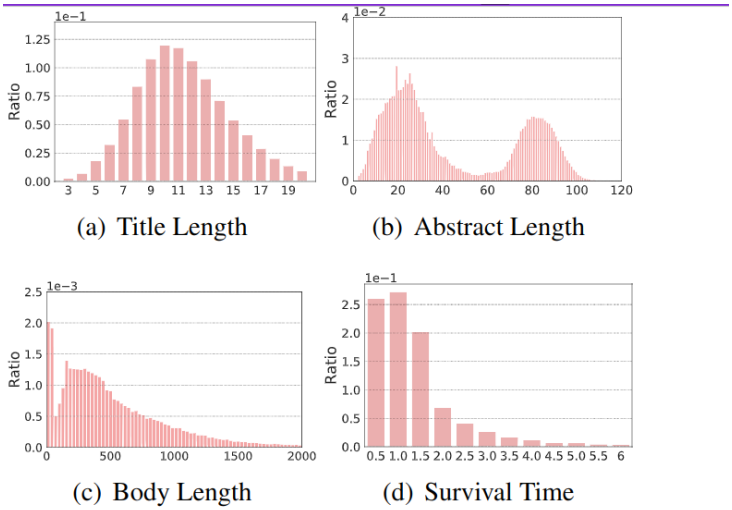
3.1 Dataset Construction

It was collected from the user behavior logs of Microsoft News. We randomly sampled 1 million users who had at least 5 news click records during 6 weeks from October 12 to November 22, 2019. An impression log records the news articles displayed to a user when she visits the news website homepage at a specific time, and her click behaviors on these news articles.
The format of each labeled sample is [ImpID, uID, t, ClickHist, ImpLog], where uID is the anonymous ID of a user, and t is the timestamp of this impression. ClickHist is an ID list of the news articles previously clicked by this user (sorted by click time). ImpLog contains the IDs of the news articles displayed in this impression and the labels indicating whether they are clicked, i.e., [(nID1,label1), (nID2,label2), ...] where nID is new article ID and label is the click label (1 for click and 0 for non-click).
We used the samples in the last week for test, and the samples in the fifth week for training. For samples in training set, we used the click behaviors in the first four weeks to construct the news click history. For samples in test set, the time period for news click history extraction is the first five weeks. We only kept the samples with non-empty news click history. Among the training data, we used the samples in the last day of the fifth week as validation set.



Each news article in the MIND dataset contains a news ID, category, subcategory, title, abstract, url, title entities, abstract entities.

3.2 Dataset Analysis



(a) Title Length (b) Abstract Length

(c) Body Length (d) Survival Time

Figure 2: Key statistics of the MIND dataset.

# News	161,013	# Users	1,000,000
# News category	20	# Impression	15,777,377
# Entity	3,299,687	# Click behavior	24,155,470
Avg. title len.	11.52	Avg. abstract len.	43.00
Avg. body len.	585.05		

Table 2: Detailed statistics of the MIND dataset.

4 Method

Chu Dinh Duc
31/1/2023