Digital Media Laboratory

Weekly Report December 12, 2021

Ali J. Alaee
alialaee98@gmail.com

Report No. 20
00/09/15- 00/09/21

Abstract

I experimented with my previous Search2Vec model on the AOL dataset to find the correct ground truth. I used the MRR metric because it is less complicated and is convenient for the first examinations. The results are reported, and I came across a few ambiguities, which I stated at the end.

Description

- **Experiments**

I did two experiments that fit the basic Search2Vec model on the AOL dataset. The remarkable thing is that this dataset has a long tail of queries. Most of the training data is ignored during the training, and the ranking results for most of the test data are not available (both because the vocabulary size is limited). Surprisingly, MRR for the default method (based on the default rankings available on the dataset) happened to be 0.72, which is much more than what is reported on the [1].
Some general information about both executions.

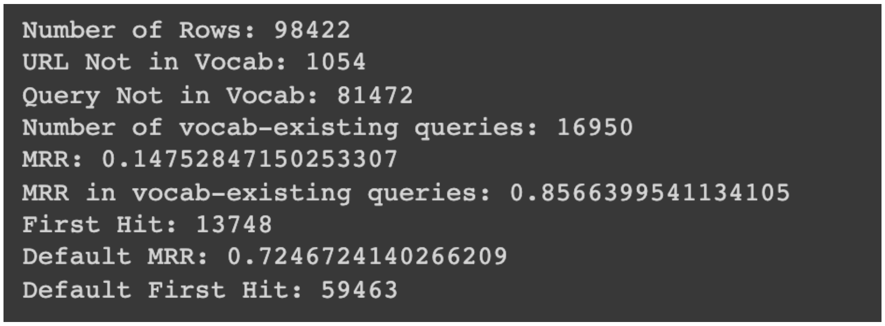
Training Samples	2,720,990
Training Samples after Omitting the queries not followed by clicks	1,475,996
Validation Samples	467,518
Test Samples	425,998
Test Samples after Just Keeping the First Clicks	98,422

Training, Validation, and Test data are split with a ratio of 6:1:1 based on the [1].
Results of each experiment is illustrated in the following section.

- **First Exp.**

Vocab Size	3007 items + 3080 queries
Approximate minimum frequency for vocab selection	item > 40, queries > 15

Results observed on the test set are shown in the following image.



```
Number of Rows: 98422
URL Not in Vocab: 1054
Query Not in Vocab: 81472
Number of vocab-existing queries: 16950
MRR: 0.14752847150253307
MRR in vocab-existing queries: 0.8566399541134105
First Hit: 13748
Default MRR: 0.7246724140266209
Default First Hit: 59463
```

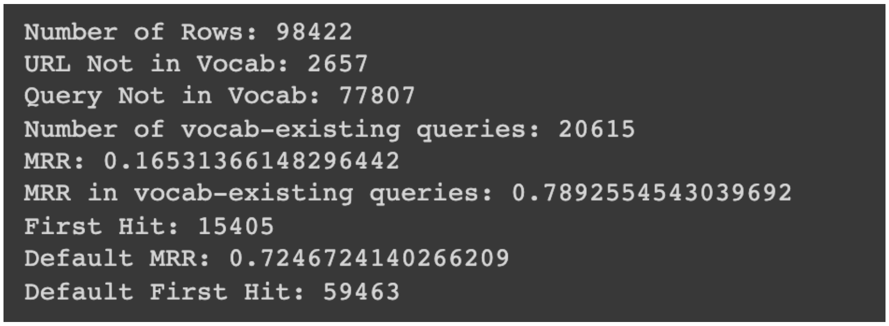
Remarkable point about this experiment is that around 83% of the queries are not a member of vocab. However, according to the MRR in vocab-existing queries (0.86), the model performs good when it comes to learned queries.

- **Seccond Exp.**

  In the second experiment, I enlarged the vocabulary, that resulted in better overall MRR but lower MRR for vocab-existing queries and slower execution.

Vocab Size	2,810 items + 12,495 queries
Approximate rule for vocab selection	item > 20, queries > 10
Training samples after applying the limmitations	288 k

Results observed on the test set are shown in the following image.



```
Number of Rows: 98422
URL Not in Vocab: 2657
Query Not in Vocab: 77807
Number of vocab-existing queries: 20615
MRR: 0.16531366148296442
MRR in vocab-existing queries: 0.7892554543039692
First Hit: 15405
Default MRR: 0.7246724140266209
Default First Hit: 59463
```

Although the size of vocabulary is doubled, still 79% of the test queries are from out of the vocabulary. Thus, the overall MRR (0.17) is yet far less than for the default method (0.72).

- **Questions**
  - Given the aim of our research, which is finding an effective method to retrieve rankings for head queries, is it right to first just keep the head queries (omit the tail entries), and then split the data to train, validation, and test? In this fashion, the MRR would be satisfying, and probably personalization could enhance the model.
  - In the [1] it is claimed that the MRR for the default method is 0.26. However, according to my investigation, it is way larger than this number (0.72). In addition, it is mentioned in the article that the BM25 algorithm generates the default rankings. However, the data itself contains all the default rankings! This is confusing.
  - I reimplemented the method for finding the cosine distances of points in the embedding space, which is now more efficient. However, I guess that a library should do this more efficiently. I will be delighted to find it because much of the time overhead is for these computations.

Next Week

- Finding the correct approach (mentioned in the questions) for finding our base-line, and then applying the first idea of personalization.

References

[1] Jing Yao et al. Employing personal word embeddings for personalized search. *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.