Digital Media Laboratory

Weekly Report December 19, 2021

Ali J. Alaee Report No. 21

alialaee98@gmail.com 00/09/24- 00/09/28

Abstract

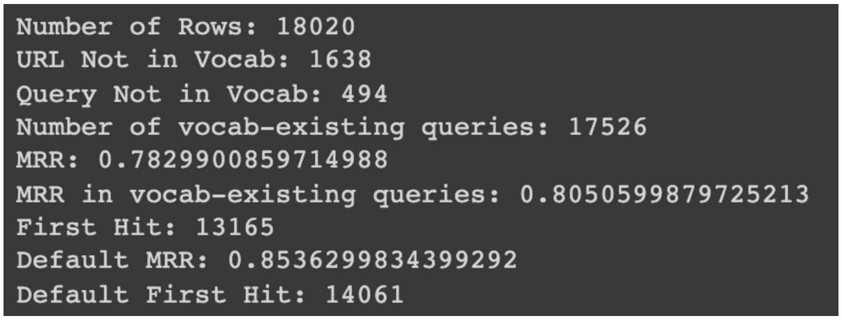
- Finding the baseline according to the wednesday discussion
- Setting up the user similarity matrices according to [1]

Description

- **Baseline**

After setting the minimum frequency of queries to 35 and applying this condition to the whole data, the size of the data was reduced to 813 k entries.
Details of the baseline experiment can be observed below.

Training Samples	813,190
Validation Samples	108 k
Test Samples	138 k
Training Samples after Applying the min URL frequency > 10	248 k
Number of Unique Queries in the Training Set	4,980
Number of Unique URLs in the Training Set	2,915
Vocab Size	7,896
Test Set after Only Considering the First Clicks for Each Query	18 k

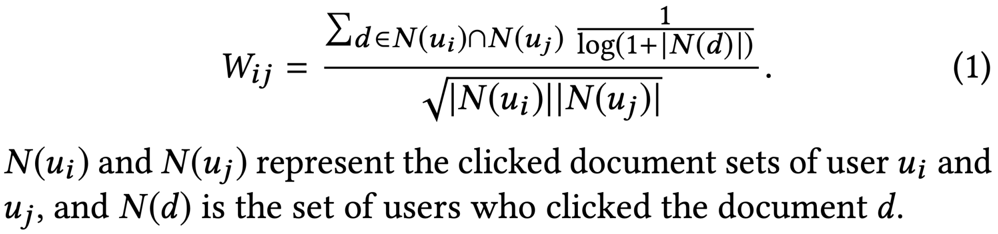


```
Number of Rows: 18020
URL Not in Vocab: 1638
Query Not in Vocab: 494
Number of vocab-existing queries: 17526
MRR: 0.7829900859714988
MRR in vocab-existing queries: 0.8050599879725213
First Hit: 13165
Default MRR: 0.8536299834399292
Default First Hit: 14061
```

The above screenshot indicates the results for the baseline on the test set. The observed MRR is 0.783, considering that the MRR for the Default Ranking method on the data set is 0.854. Although all the tail queries are removed before splitting the data, still less than 2.5 percent of the test data are not in vocab; Because there are some queries repeated only after a period of time, which is allocated for the test set.

- **User Similarity Matrix**

Based on the [1] proposed method, which is presented in the illustration, the W matrix is formed. The implementation is available in the AOL_User_Classification.ipynb, in the AOL_Implementation folder. Users with more than 5 issued queries are conserved, so the total number of considered users is 18,781.



$$W_{ij} = \frac{\sum_{d \in N(u_i) \cap N(u_j)} \frac{1}{\log(1+|N(d)|)}}{\sqrt{|N(u_i)||N(u_j)|}}. \tag{1}$$

$N(u_i)$ and $N(u_j)$ represent the clicked document sets of user u_i and u_j, and $N(d)$ is the set of users who clicked the document d.

Given the number of users, W matrix has occupied nearly 2GB of memory. This similarity matrix is based on the clicked documents(URLs) of users. Furthermore, classifications of users can take place with this matrix

Next Week

- Finding the best way to make effective clusters out of the User Similarity Matrix.
- Is there a more effective way to calculate all cosine distances in a multi-dimensional space?

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.