WUT21 at the NTCIR-16 Data Search 2 Task

Lin Li
Wuhan University of Technology
cathylilin@whut.edu.cn

Xinyu Chen
Wuhan University of Technology
268595@whut.edu.cn

Sijie Long
Wuhan University of Technology
sijie.long@whut.edu.cn

Abstract

- We choose a simple base model for the IR subtask and using a document-based storage method.
- Formulating a retrieval strategy with Elastic Search.
- This strategy uses embedded retrieval algorithms to retrieve topics and calculate text similarity.

Introduction

- A common strategy for text-based information retrieval is to use a ranking function to rank all texts according to search terms and select the top n.
- This report describes and discusses our results using different textual similarities for topics in the IR subtask to calculate how well topics match documents and returning a sorted list.

Methods

- LM Jelinek Mercer Similarity algorithm:

  Under the query-likelihood approach, language models for IR try to estimate for each document the probability that the query Q was generated by the underlying language model. If it is assumed that terms occur independently, then the probability becomes the product of the individual query terms given the document mode.

Experiments

- Statistical analysing

Table 2: The number of topics with L2 labels.

	Number of topics
L2 label	69
sum	192

Table 1: The number of L2 labels in the training set.

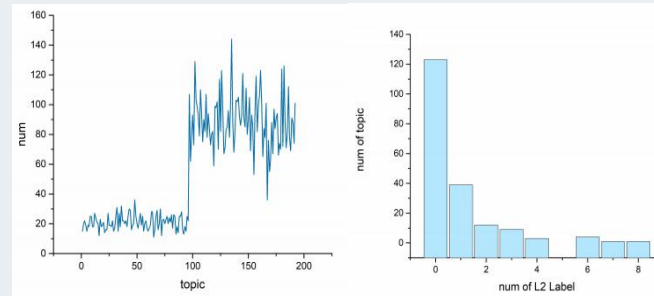|	training set
L2 label	141
sum	10536

Table 1 records the number of L2 labels in the training set and the size of the training set.
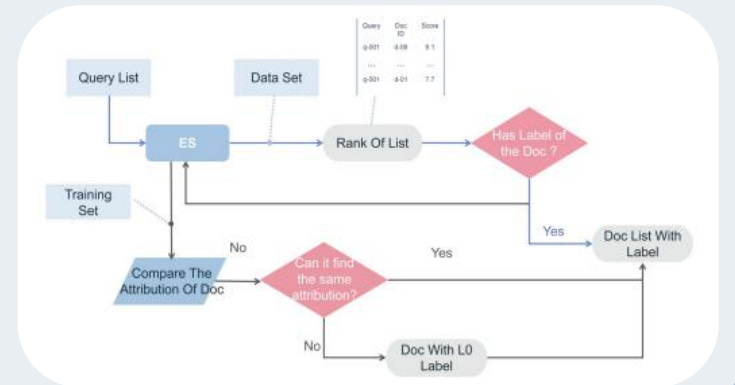Table 2 records the number of topics with L2 labels and the total number of topics in the training set.
Figure 1 counts the number of document entries in the training set corresponding to each topic.
Figure 2 counts the number distribution of L2 label in topics.

Experiments

- Search process

Conclusion

- In the final performance results, the effect presented by our team is moderate in the overall performance

Reference

- Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2022. Overview of the NTCIR-16 Data Search 2 Task. In Proceedings of the NTCIR-16 Conference
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.ISBN:978-1-4503-0000-0/18/06 [J]