Model Fusion Engine

Group F

Zhongjie Xu, 190622114

Miao Yu, 151008494

Maryanne Maira, 180857096

Jingjie Wang, 190371474

TABLE OF CONTENTS

I.

Introduction

1. Introduction

This is a search engine for movies and TV series hosted by Netflix. Users enter the query they want to search through the interactive interface (which can be the name, actors, directors and related content of the description) in order to query the database. The system will use multiple models to search the database and the results are integrated to give the best results.

2.

Related Tools

2. Related Tools

Elasticsearch

Kibana

3.

Engine Architecture

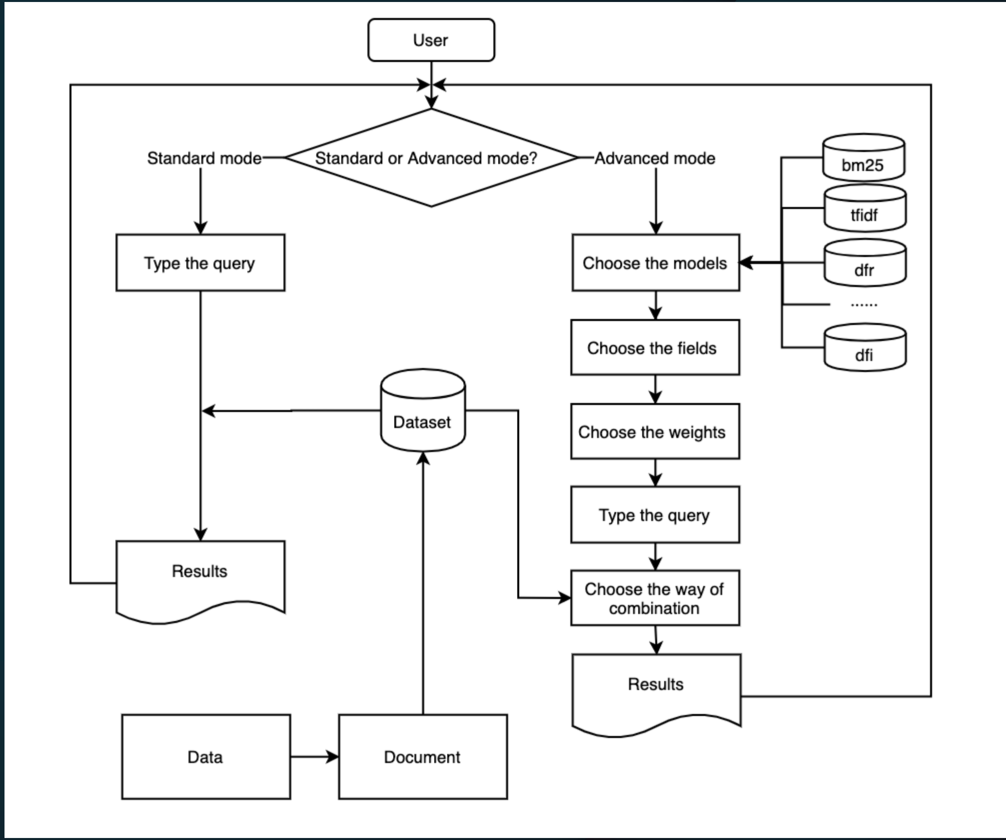
3.Engine Architecture

- Standard mode

    A default search method.

- Advanced mode

    A search method with multiple functions.



3.Engine Architecture

<u>Standard Engine:</u>

- According to paper “Fusion Analysis of Information Retrieval Models on Biomedical Collections”, it was demonstrated that the Score combination overall outperformed rank combination and three models outperformed two models.

- The Standard Engine encompasses score combination with three models, the Language Model Jelinek-Mercer Smoothing, weighted TF-IDF, and Information Based Model. These models are chosen as they have the highest precision and recall.

- ElasticSearch enables the weighting of different fields which is an advantage as our data encompasses structured documents instead of unstructured documents. Using the proposed method by Stephen Robertson, it is shown in the previous slide that only TF-IDF improves with weighted fields.

3.Engine Architecture

Advanced Engine:

- The Advanced Engine allows the combination of different models to your choice weighting different fields according to your choice.

- It utilises CFA algorithm proposed by D.F.Hsu, Y.S. Chung and B.S Kristal that combines the scores of different models and ranks the algorithms accordingly. The algorithm is discussed in the next slides.

- Models of the Advanced Engine：
  1. OKAPI BM25
  2. Bayesian smoothing using Dirichlet Priors (LMDirichletSimilarity)
  3. Divergence from randomness (DFR SImilarity)
  4. Divergence from Independence (DFI similarity)
  5. TF-IDF
  6. Language Model Jelinek-Mercer Smoothing
  7. Information-Based Model based on Stephane Clichant and Eric Gaussier

4.

Related Technologies

4.Related technology： CFA

- CFA definition one: Given two score functions sa and sb, the score function of the score combined function is defined as ½[sa + sb] for every similar object d. If greater than two models chosen this would then be the average score combinations.
- CFA definition two: Given two rank functions ra and rb, the score function of the rank combined function is defined as ½[ra + rb] for every similar object d. If greater than three models chosen this would then be the average rank combinations.
- CFA definition three [Diversity measure]: Given two rank functions ra and rb the diversity between them can be calculated using Kendall’s Tau distance, equation shown below:

$$\tau = \frac{C - D}{N(N+1)/2},$$

x	1	2	3	4	5	6	7	8	9	10
$r_A(x)$	d_2	d_8	d_5	d_6	d_3	d_1	d_4	d_7	d_{10}	d_9
$s_A(d)$	10	7	6.4	6.2	4.2	4	3	2	1	0

(a) Ranked list A

x	1	2	3	4	5	6	7	8	9	10
$r_B(x)$	d_5	d_9	d_6	d_2	d_8	d_7	d_1	d_3	d_{10}	d_4
$s_B(d)$	10	9	8	7	6	5	4	3	2	1

(b) Ranked list B

d	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
$g_{AB}(d)$	6.5	2.5	6.5	8.5	2	3.5	7	3.5	6	9

(c) Combinations of A and B by rank

x	1	2	3	4	5	6	7	8	9	10
d	d_5	d_2	d_6	d_8	d_9	d_1	d_3	d_7	d_4	d_{10}
$s_g(d){:}f_g(x)$	2	2.5	3.5	3.5	6	6.5	6.5	7	8.5	9
$r_C(x)$	d_5	d_2	d_6	d_8	d_9	d_1	d_3	d_7	d_4	d_{10}

(d) Sorted scores $g_{AB}(d)$ into $s_g(d)$ with document indices

d	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	d_{10}
$h_{AB}(d)$	4.0	8.5	3.6	2.0	8.2	7.1	3.5	5	4.5	1.5

(e) Combinations of A and B by score

x	1	2	3	4	5	6	7	8	9	10
d	d_2	d_5	d_6	d_8	d_9	d_1	d_3	d_7	d_4	d_{10}
$s_h(d){:}f_h(x)$	8.5	8.2	7.1	5.0	4.5	4.0	3.6	3.5	2.0	1.5
$r_D(x)$	d_2	d_5	d_6	d_8	d_9	d_1	d_3	d_7	d_4	d_{10}

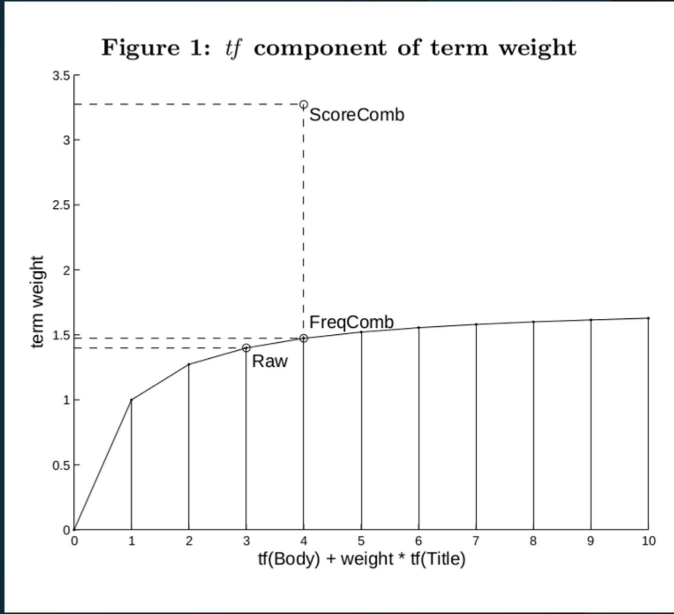
(f) Sorted scores $h_{AB}(d)$ into $s_h(d)$ with document indices

4.Related Algorithm Weighted Model

- Stephen Robertson et al. discuss extending BM25 for structured documents to create an alternative linear combination of term frequencies. The argument given is that most weighting functions based on tf are non-linear and to preserve the nonlinearity it is better to eight separate fields according to their importance.

- The netflix data was weighted according to the proposed method, the description of the tv/movie having the highest weighting, the title having the second highest and the rest having a default weighting.

  Title x 2
  Description x 3
  Cast x 1
  Country x 1



5.

Evaluation

5. Evaluation

$$\text{Recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of documents retrieved}}$$

$$\text{Precision} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents}}$$

$$\text{Average Recall} = \frac{\text{Recall1}+\text{Recall2}+\cdots+\text{Recall(i)}}{10}$$

$$\text{Average Precision} = \frac{\text{Precision1}+\text{Precision2}+\cdots+\text{Precision(i)}}{10}$$

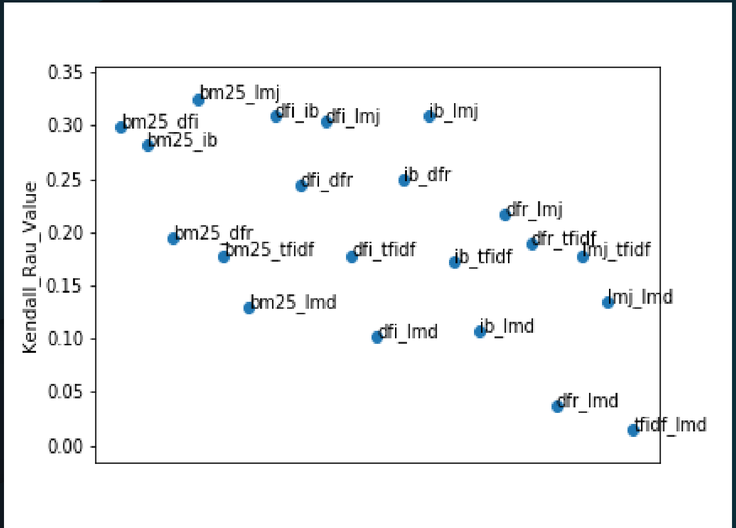
(i: the number of documents)

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}\right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

5. Evaluation of Individual Models

Index	Average Precision	Average Recall	Harmonic Mean	Weighted Precision	Weighted Recall	Weighted Harmonic Mean
Okapi BM25	0.676	0.43	0.496	0.6	0.37	0.14
DFI	0.66	0.42	0.483	0.56	0.35	0.14
Information Based Model	0.71	0.45	0.52	0.59	0.36	0.14
DFR	0.62	0.39	0.45	0.54	0.33	0.14
Language Model Jelinek-Mercer Smoothing	0.7	0.45	0.517	0.61	0.38	0.14
TF-IDF	0.667	0.42	0.488	0.747	0.48	0.16
Bayesian smoothing using Dirichlet Priors	0.52	0.3	0.357	0.55	0.33	0.14

5. Evaluation of Models using Kendall-Tau Distance



- The Kendall Tau states that if the value is high, the models have a similar rank and when the value is low the models have a dissimilar rank.

Index	Average Precision	Average Recall	Average_f_ measure
tfidf_lmd_rank	0.667	0.42	0.486
tfidf_lmd_score	0.667	0.42	0.486
bm25_lmj_rank	0.677	0.43	0.50
bm25_lmj_score	0.677	0.43	0.50

5. Evaluation of Standard Engine

- The standard engine was evaluated for 10 queries. The score combination and rank combination MAP was calculated, the score combination generating the same results as rank combination.According to paper “Fusion Analysis of Information Retrieval Models on Biomedical Collections”, it was demonstrated that the Score combination outperformed rank combination on some occasions and three models outperformed two models.

Index	**Average Precision**	**Average Recall**	**Harmonic Mean**
CFA score	0.667	0.42	0.485
CFA rank	0.667	0.42	0.485

- The Combination Model which included the three models and score combination shows that the highest precision and recall is lower than the one derived with Kendall-Tau Rank correlation. The average precision, recall and harmonic mean were not higher than weighted TF-IDF. This could suggest that perhaps the indexed data’s structure pairs well with one model which is the weighted TF-IDF.

6.

Demonstration

7.

Conclusion and Future Work

7. Conclusion

- The CFA model has limitations in terms of its scope and power. It is deemed effective in systems which have a high overlap of relevant documents and low overlap of non-relevant documents.
- Future work would involve using a combination model that is differentiable as it would make a likely candidate for use in gradient optimization procedures which could be used by neural nets etc.
- Example of combination model to be used:

$$\rho(w_1, w_2, x, q) = w_1\rho_1(x, q) + w_2\rho_2(x, q) \quad \longrightarrow \quad \rho(w, x, q) = \sin(w)\rho_1(x, q) + \cos(w)\rho_2(x, q)$$

7. Future Work

- Test the CFA on completely different set of structured documents. The description of netflix data which is the longest field of the data only had a maximum of around 500 characters.
- Have more diverse queries and more queries.
- Extend the CFA to Neural Nets
- Develop visual operation interface. And can return pictures of the corresponding results

8. Reference

[1] R. Kumar, S. K. Singh, and V. Kumar, 'A heuristic approach for search engine selection in meta-search engine', in International Conference on Computing, Communication & Automation, Greater Noida, India, 2015, pp. 865–869, doi: 10.1109/CCAA.2015.7148496.

[2] D. Liu, X. Xu, and Y. Long, 'On member search engine selection using artificial neural network in meta search engine', in 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), Wuhan, China, 2017, pp. 865–868, doi: 10.1109/ICIS.2017.7960113.

[3] W. Meng, C. Yu, and K.-L. Liu, 'Building efficient and effective metasearch engines', ACM Comput. Surv. CSUR, vol. 34, no. 1, pp. 48–89, Mar. 2002, doi: 10.1145/505282.505284.

[4] Yankun Li, Nigtao Shi, and D. Frank Hsu, 'Fusion Analysis of Information Retrieval Models on Biomedical Applications..

[5] D. Frank Hsu, Yun-Sheng Chung, Bruce S. Kristal, 'Combinatory FUsion Analysis: Methods and Practices of Combining Multiple Scoring Systems.

[6] S. Robertson, H. Zaragoza, and M. Taylor, CIKM, 'Simple BM25 extension to multiple weighted fields'

THANKS!