

Your Name: Xiaoyun Ye

Your Andrew ID: xye2

Homework 2

Collaboration and Originality

Your report must include answers to the following questions:

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.

No.

If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

No.

If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

3. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

No.

4. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

Yes

If you answered No:

- a. identify the software that you did not write,
- b. explain where it came from, and
- c. explain why you used it.

5. Are you the author of every word of your report (Yes or No)?

Yes

If you answered No:

- a. identify the text that you did not write,
- b. explain where it came from, and
- c. explain why you used it.

Your Name: Xiaoyun Ye

Your Andrew ID: xye2

Homework 2

Instructions

1 Experiment 1: Baselines

| | Ranked Boolean AND (Exp-1.1a) | BM25 BOW (Exp-1.1b) | Indri BOW (Exp-1.1c) |
|-------------|--|------------------------------------|-------------------------------------|
| P@10 | 0.1100 | 0.5400 | 0.6100 |
| P@20 | 0.1150 | 0.4850 | 0.5750 |
| P@30 | 0.0967 | 0.4267 | 0.5100 |
| MAP | 0.0070 | 0.1380 | 0.1531 |

2 Experiment 2: Indri Parameter Adjustment

2.1 μ

| | μ (Note: $\lambda=0$) | | | | |
|-------------|--|---------------------------|----------------------------|-----------------------------|-------------------------|
| | 1500 (Exp-2.1a) | 500 (Exp-2.1b) | 3000 (Exp-2.1c) | 10000 (Exp-2.1d) | 0 (Exp-2.1e) |
| P@10 | 0.5600 | 0.5300 | 0.6200 | 0.5600 | 0.4600 |
| P@20 | 0.5650 | 0.4950 | 0.5650 | 0.5250 | 0.4400 |
| P@30 | 0.5033 | 0.4467 | 0.5133 | 0.4967 | 0.4400 |
| MAP | 0.1558 | 0.1424 | 0.1537 | 0.1403 | 0.1299 |

2.2 λ

| | λ (Note: $\mu=0$) | | | | |
|-------------|--|---------------------------|---------------------------|---------------------------|---------------------------|
| | 0.4 (Exp-2.2a) | 0.8 (Exp-2.2b) | 1.0 (Exp-2.2c) | 0.2 (Exp-2.2d) | 0.0 (Exp-2.2e) |
| P@10 | 0.4700 | 0.4700 | 0.1900 | 0.4600 | 0.4600 |
| P@20 | 0.4500 | 0.4650 | 0.2250 | 0.4400 | 0.4400 |
| P@30 | 0.4433 | 0.4367 | 0.2300 | 0.4433 | 0.4400 |
| MAP | 0.1307 | 0.1308 | 0.0451 | 0.1305 | 0.1299 |

2.3 Parameters

For μ , I chose to make some small changes to the value as well as some significant changes in order to compare whether the degree of value change will have any impact on the results.

For λ , I chose to make more uniform changes since the value range of λ is not as large as μ . And the uniform value change can show how the value ranges and whether larger or smaller values of λ contributes more to the retrieval precision.

For both parameters, I take into consideration the extreme case of zero.

2.4 Discussion

For μ , to change the value of μ is to change the smoothing of the term frequency. So significant changes in the values does not contribute much to the retrieval results, since it makes the term frequency either too minor or too important. But a slight increase of μ does improve the precision. And no smoothing of term frequency is not a good idea, judging from the statistics.

For λ , changing the value of λ means controlling the effects of idf, in order to avoid some misleadings of those words which occur often but exists in a lot of documents among the collection. The case of λ equalling one is the worst. The reason might be that we are experiments with rather short queries, in which case we don't have many query terms, so emphasizing too much on the rare words may not be the ideal way to improve the query results. Apart from that, the increase and the decrease of λ didn't show much significant changes on the results.

3 Experiment 3: Indri Representations

3.1 Results

| | Indri BOW (body) (Exp-3.1a) | 0.2 url 0.1 keywords 0.2 title 0.5 body (Exp-3.1b) | 0.25 url 0.25 keywords 0.25 title 0.25 body (Exp-3.1c) | 0.1 url 0.1 keywords 0.1 title 0.7 body (Exp-3.1d) | 0.2 url 0.5 keywords 0.2 title 0.1 body (Exp-3.1e) |
|-------------|--|---|---|---|---|
| P@10 | 0.6100 | 0.5500 | 0.5900 | 0.5500 | 0.5600 |
| P@20 | 0.5750 | 0.5050 | 0.5050 | 0.5100 | 0.5100 |
| P@30 | 0.5100 | 0.4633 | 0.4600 | 0.4567 | 0.4767 |
| MAP | 0.1531 | 0.1406 | 0.1416 | 0.1400 | 0.1454 |

3.2 Example Query

755:#AND(#WSUM(0.2 Scottish.url 0.1 Scottish.keywords 0.2 Scottish.title 0.5 Scottish.body)
#WSUM(0.2 Highland.url 0.1 Highland.keywords 0.2 Highland.title 0.5 Highland.body) #WSUM(0.2
Games.url 0.1 Games.keywords 0.2 Games.title 0.5 Games.body))

755:#AND(#WSUM(0.25 Scottish.url 0.25 Scottish.keywords 0.25 Scottish.title 0.25 Scottish.body)
#WSUM(0.25 Highland.url 0.25 Highland.keywords 0.25 Highland.title 0.25 Highland.body)
#WSUM(0.25 Games.url 0.25 Games.keywords 0.25 Games.title 0.25 Games.body))

755:#AND(#WSUM(0.1 Scottish.url 0.1 Scottish.keywords 0.1 Scottish.title 0.7 Scottish.body)
#WSUM(0.1 Highland.url 0.1 Highland.keywords 0.1 Highland.title 0.7 Highland.body) #WSUM(0.1
Games.url 0.1 Games.keywords 0.1 Games.title 0.7 Games.body))

755:#AND(#WSUM(0.2 Scottish.url 0.5 Scottish.keywords 0.2 Scottish.title 0.1 Scottish.body)
#WSUM(0.2 Highland.url 0.5 Highland.keywords 0.2 Highland.title 0.1 Highland.body) #WSUM(0.2
Games.url 0.5 Games.keywords 0.2 Games.title 0.1 Games.body))

3.3 Weights

The sets of weights tries to figure out whether the weights assigned to different fields does affect the retrieval results much. We tested the even case(all weights equals) as well as emphasizing some fields: slightly preferring body field, significantly preferring body field and slightly preferring keywords field.

3.4 Discussion

The results does not show much improvements after we assigned different weights for different weights. But among those cases, even weights and slightly keyword emphasis performs better than other cases. This somewhat confusing result may be partially due to the characteristics of the malfunction of my system. Another possibility is that when the scenarios of queries vary much, the best way is to pay equal attention to every field or slightly more focused on body and keywords.

4 Experiment 4: Sequential dependency models

4.1 Results

| | Indri BOW (body) (Exp-4.1a) | 0.5 AND 0.25 NEAR 0.25 WINDOW (Exp-4.1b) | 0.5 AND 0.4 NEAR 0.1 WINDOW (Exp-4.1c) | 0.5 AND 0.1 NEAR 0.4 WINDOW (Exp-4.1d) | 0.4 AND 0.3 NEAR 0.3 WINDOW (Exp-4.1e) |
|-------------|--|---|---|---|---|
| P@10 | 0.6100 | 0.6600 | 0.6500 | 0.6600 | 0.6500 |
| P@20 | 0.5750 | 0.6300 | 0.6350 | 0.6300 | 0.6150 |
| P@30 | 0.5100 | 0.5667 | 0.5700 | 0.5567 | 0.5567 |
| MAP | 0.1531 | 0.2091 | 0.2094 | 0.2057 | 0.2041 |

4.2 Example Query

755:#wand(0.5 #and(Scottish Highland Games) 0.25 #and(#near/1(Highland Games) #near/1(Scottish Highland)) 0.25 #and(#window/8(Highland Games) #window/8(Scottish Highland)))

755:#wand(0.5 #and(Scottish Highland Games) 0.4 #and(#near/1(Highland Games) #near/1(Scottish Highland)) 0.1 #and(#window/8(Highland Games) #window/8(Scottish Highland)))

755:#wand(0.5 #and(Scottish Highland Games) 0.1 #and(#near/1(Highland Games) #near/1(Scottish Highland)) 0.4 #and(#window/8(Highland Games) #window/8(Scottish Highland)))

755:#wand(0.4 #and(Scottish Highland Games) 0.3 #and(#near/1(Highland Games) #near/1(Scottish Highland)) 0.3 #and(#window/8(Highland Games) #window/8(Scottish Highland)))

4.3 Weights

Those sets of weight choices represent the following scenarios: emphasizing more on AND and ignore the differences between WINDOW and NEAR(equal weights on both operators), generally equal weights on all three operators, slight preference on either NEAR or WINDOW.

4.4 Discussion

By assigning weights to the operators and constructing the sequential dependence model, the retrieval results do improve. And assigning equal weights to NEAR and WINDOW or even preferring WINDOW slightly seems to be better among all the results. The reason might be that when the query terms have relations or connections but do not show significant sequential order, it is better to choose WINDOW since it does not strictly follow the order of terms in the query.