Sz-Rung Shiang

Your Andrew sshiang

Homework 4

Collaboration and Originality

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

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

   No

3. 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

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

   Yes

Sz-Rung Shiang

Your Andrew sshiang

Homework 4

Instruction

1 Experiment: Baselines

	BM25	Indri BOW	Indri SDM
P@10	0.3200	0.2000	0.2800
P@20	0.2400	0.1940	0.2460
P@30	0.2173	0.1907	0.2213
MAP	0.1250	0.0865	0.1294

Parameters for BM25:

- BM25:k_1=1.2
- BM25:b=0.75
- BM25:k_3=0

Parameters for Indri:

- Indri:mu=2500
- Indri:lambda=0.4

Paramters for SDM model:

- AND: 0.6
- NEAR/1: 0.2
- WINDOW/8: 0.2

2 Custom Features

f17: Length of the document (body field).

Although the BM25 feature already considers the normalization effect of term frequency and document frequency, length of document might be another informative feature. For document which is too long or too short, it might reflect the quality of the document; for example, if the document is too short, there might not be enough information to the users, while if the document is too long, there might be some trash information or concept of the document is not concise to the users' need. I set this feature as the length of the body field of each document. Note that the document with length too long or too short is not good; therefore, kernel trick imposing non-linear feature might be more useful for this feature.

f18: Range of the term frequency of the query terms.

In this feature I used the range of term frequency of query terms. If the query is a phrase, it's better that the term frequency of each term is roughly the same as the term frequency of other query terms. In addition, if a query is composed of two terms, and there is term not appearing in the document but the other term appearing with a lot of times, it might not be a good fit for our search engine. Even though Indri has the effect normalizing the difference in term frequencies of query terms, it may still be dominated by the terms with very high term frequency. I calculate the minimum and maximum term frequency of every terms in the query, and then I set the score as max_value – min_value. This feature may be risky that it depends on what type of query we use in our system. For example, if the query is "Los Angeles" which is a phrase, then the range is the smaller the better. On the other hand, if the query is "Living in India", then the frequency range doesn't help much.

3 Experiment: Learning to Rank

	IR Fusion	Content-Based	Base	All
P@10	0.2920	0.3080	0.3080	0.2960
P@20	0.2280	0.2420	0.2420	0.2440
P@30	0.2147	0.2227	0.2253	0.2240
MAP	0.1050	0.1147	0.1165	0.1117

As shown in the table, the base model (f1-f16) performs the best among these models, while IR fusion performs the worst. It makes sense that IR fusion (f5, f6, f8, f9, f11, f12, f14, f15) uses the least features and it gets the worst result. However, all of them perform worse than BM25 model for only body field. According to the previous homework and experiments using WSUM for Indri operator, when we increase the weights for other fields (such as inlink and url), it performs worse. We have the similar results that adding BM25 model and Indri model for other fields gets worse results. It might be because for some relevant document, it's not necessary to contain query terms in these comparably short fields, and therefore the scores for some relevant document are low. Moreover, the body field has good ability to distinguish classes, and therefore models using other fields may be redundant.

Content-based approach (adding overlap) achieves improvement over IR fusion model. It infers that overlap feature can bring additional information that is not covered by BM25 and Indri models. The overlap feature capture how much does the document match each query term, and it can help to distinguish between the document with all query terms matched but with low term frequency and the document with only a few terms matched and with high frequency for these matching. The information is blurred in BM25 and Indri model, and therefore overlap feature is not redundant.

Adding other features only brings a little improvement on P@30 and MAP. Although these features are not covered by content-based approaches; however, they are not related to the query and thus they might don't have the ability to distinguish classes. For example, it's not necessary for all the relevant documents to be related to Wikipedia pages, and therefore adding this feature might not be helpful.

The all model (with custom features) performs worse than base model (without custom features) in most of the evaluation metrics. Adding more features doesn't give us better result. It might be because of overfitting when we train the model or the feature is not informative. I also use the training data as the testing data, and the result stays the same adding f17 feature and degrades adding f18 feature. It means that these two features don't provide additional information to distinguish the ranking, and f18 (range of term frequency) even impose noise to the model. For f18 feature, it depends on the query. For example, if the query is "Los Angeles" which is a phrase, then the range is the smaller the better. On the other hand, if the query is "Living in India", then the frequency range doesn't help much. Due to the dynamics of the query, this feature may impose noise into our SVM model.

4 Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	f5,f6,f7,f8 ,f9,f10	f1,f2,f5, f6,f7,f8, f9,f10	f1,f2,f3,f4, f5,f6,f7,f8, f9,f10	f1,f2, f5,f6,f7,	f1,f2, f5,f7,
P@10	0.3080	0.3000	0.3040	0.3040	0.3120	0.3200
P@20	0.2420	0.2480	0.2640	0.2500	0.2640	0.2600
P@30	0.2253	0.2253	0.2253	0.2280	0.2427	0.2240
MAP	0.1165	0.1126	0.1133	0.1159	0.1282	0.1324

According to the previous experiments, BM25 and Indri for other fields has worse results than BM25 for body field. In addition, BM25 performs better than Indri model according to the result in section 1. Consequence, I use the strategy that select only body field and BM25 model, and add some non-redundant features that may be useful. In the first set, I remove all the feature set related to other fields and the query irrelevant features, and it performs worse than the baseline model but it achieves comparable results. When adding the spam score and url score, all of the evaluation metrics increase and P@20 is even higher than baseline model; however, when adding pagerank score and wiki score, the performance degrades. When removing title filed related feature, the performance increases due to the redundancy of the feature set to the body filed related feature set. In the final selection, I choose spam score, url depth score, BM25 score for body field and overlap score for body field. This feature set achieves the best performance that exceeds the baseline model and the BM25 model.

Using smaller set of features can increase both the effectiveness and efficiency. To use feature set with the ability to distinguish classes and reduce the number of redundant features can prevent the model from overfitting. In addition, it may reduce the risk that features are noisy for classification and may deteriorate the model. Another benefit is that using less features can definitely improve the efficiency for both reducing the time for feature extraction and also the time to train and test the model.

In conclusion, we should use features that have information with diversity which cover most of the dynamics of the document set, and we should reduce the redundancy in the feature set that represents similar information. Also, we should choose the feature set that has ability to distinguish different classes, which might be able to check using entropy or dimension reduction.

5 Analysis

Weight in SVM:

1:0.16407286	2:0.036731094	3:-0.16391207	4:-0.12884209	5:0.47309396
6:0.23992337	7:0.50707108	8:0.30458361	9:0.1640749	10:0.39498276
11:0.019954907	12:-0.0058421907	13:0.064427376	14:0.049949471	15:-0.027705675
16:0.090754151	17:0.17548569	18:0.224676		

The large weight in SVM means this dimension of feature is important to distinguish the classes and it has the positive contribute to the model, the small weight (negative) means that the feature has negative effects (the higher the feature, the less important this example), and the weight close to 0 mean the weight is not useful to distinguish the classes.

In the SVM model file, we can see that f7 (overlap in body), f5 (BM25 in body), f10 (overlap in title), f8 (BM25 for title) and f6 (indri for body) get the highest value. It makes sense because if we use only BM25 and Indri model in body field, we can get comparable results (Actually using BM25 model is the best model in this report).

f3 (wiki score), f4 (pagerank score), f12 (indri in url) and f15 (indri in inlink) get the negative weights. For wiki score, number of documents with wiki url is only small amount in the whole document corpus, therefore this feature may not bring useful information for most of the documents. Most of the relevant documents are not in the Wikipedia webpages, anad therefore SVM can easily learn a negative weight. In addition, Wikipedia can't provide the information that meets the users' needs in our query set, such as "uplift at yellowstone national park" (intention to find out the official website), "equal opportunity employer" and "interview thank you" (informational webpages but not Wikipedia). As a result, it's reasonable the value of this dimension is negative. For PageRank score, it may not be a good feature because this dimension is not query related, and therefore it may not bring much information to distinguish classes.

F12 (indri for url), f11(BM25 for url) and f15(indri for inlink) get the weights close to zero. According to the previous homework and experiments, url and inlink fields are not as helpful as body or title fields. They have overlap with BM25 or Indri for body field, and they even perform worse than these models; therefore, these features are redundant for our SVM model to learn to distinguish the classes.