Your Name Xueting Hu

Your Andrew ID xuetingh

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

Your Name Xueting Hu

Your Andrew ID xuetingh

Homework 4

1. Experiment: Baselines

	BM25	Indri BOW	Indri SDM
P@10	0.3280	0.2520	0.3600
P@20	0.3120	0.2720	0.3600
P@30	0.3040	0.2920	0.3667
MAP	0.1789	0.1662	0.2047
time	60s	56s	69s

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

Indri: mu=2500, lambda=0.4

Sequential Dependency Model: 0.5 AND 0.25 NEAR 0.25WINDOW

2. Custom Features

f17: Min window size of all terms occurs in a document, normalized by document length.

If not all terms show up in this document ,the score is 1, basically, the score ranges in (0, 1] , the smaller the better, because smaller value means all terms appears closer to each other. This feature works on improving Relevance. This feature is time consuming, the time complexity of finding smallest range is of O(N), N is the document length, and disabling this feature takes 72s and not disabling this feature takes 79s for 25 test queries.

f18: Number of letters in url / url length

This feature works on judging the authority of this document. I think authority document tend to have more letters in url than numbers and punctuation. I tried a binary feature to note if the rawUrl string contains “https”, but as I went through documents in qid 74, 125 and 145, I found no url with https, so I came up with this way to describe authority. This feature is not time consuming, doesn’t cause obvious increase in total ranking time.

3. Experiment: Learning to Rank

	IR Fusion	Content-Based	Base	All
P@10	0.2600	0.2760	0.3000	0.2920
P@20	0.2800	0.2760	0.2840	0.2820
P@30	0.2773	0.2827	0.2853	0.2880
MAP	0.1626	0.1609	0.1646	0.1650

Compare “Base” with “Content-Based”, all measures are higher in “Base”. “Base” utilizes additional features like spam score and url depth of document, which are query independent and based only on the extent of authority of a document, that is how well do we believe the content is. Adding authority features improves P@N by ranking better documents higher.

Compare “Content-Based” with “IR Fusion”, the former one has better MAP and slightly better P@N. “Content-Based” utilizes overlap score besides BM25 and Indri score; overlap features (f7,10,13,16) is a little bit like boolean method, matching a term is better than not matching, matching other terms many time will never make up for the score you lose for not matching a certain term; while best-match features (f_{5}, f_{6}, f_{8}, f_{9}, f_{11}, f_{12}, f_{14}, f_{15}) do not focus on matching more terms in query, matching other terms many times may make up for the score you lose for not matching a certain term. With the help of overlap features, Precision is improved thus MAP is improved.

Compare “Base” with “All”, the latter one has slightly better measures. Looking into the model files, I find weights of f1-f16 changes very little after adding f17-18, and the newly added two features has weight of 0.02 and 0.10, quite small, not so influential, and as a result, improved measures a little bit. I construct f17 to improve relevance and f18 to describe authority. f17 measures how close terms appears in documents, like a #Window operator, it finds distance and normalize it by document length. f18 calculates the percentage of letters in whole url, under the assumption that more letters, more authority.

As for the effectiveness of custom features, I calculate the MAP win/loss/equal, which is 13/10/2. qid of loss are: 8, 9, 16, 20, 26, 27, 29, 40, 47, 49 and I fail to find what is common for these queries, some queries are one-word, others are multiple-word. Maybe because f18 is not so valid or not all these terms appears only in “body” field.

4. Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	Comb$_1$	Comb$_2$	Comb$_3$	Comb$_4$
P@10	0.2920	0.2880	0.3040	0.3200	0.3080
P@20	0.2820	0.2860	0.2940	0.2920	0.3000
P@30	0.2880	0.2893	0.2933	0.2987	0.2933
MAP	0.1650	0.1667	0.1664	0.1712	0.1672

Comb1-f1, f2, f3, f4, f5, f8, f11, f14, f18: 4 BM25 features and all document attributes, 70s

As I think BM25 and Indri are correlated, I pick BM 25, without overlap features it still beats baseline. The reason might be the authority features together with relevance features work well.

Comb2-f1, f3, f5, f6, f7, f8, f10, f13, f16: Top 9 features with larger weight, 73s

Less features but more time, the additional time was used by SVM rank. Different features contribute to ranking with different influence, those features with higher weight could still guide ranking.

Comb3-f1, f2, f3, f4, f5, f6, f17, f18: body and and all document attributes, 58s

Body is the largest field in document, and custom feature 17 also focus on this field, so I tried this set of feature. This combination has the highest measures, this might indicate that the body field is the most important factor when learning to rank.

Comb4-f1, f2, f3, f4, f5, f6, f8, f9, f10, f17, f18: BM25 and Indri features in body, title field and all document attributes, 58s

Comb3 beats baseline using only "body" field, and title field has a large weight in model file when using all 18 features, so I add title field to comb3 and get comb4. Comb4 beats baseline, too, but fails to beat comb3, the reason might be best-match score for title field is not a good feature. But as "body" field features are good enough, with all authority features, comb4 is able to beat baseline.

5. Analysis

f5 BM25(body) has the largest weight 0.63, and f6 Indri(body) has the third largest 0.37, the second large is f1 spam score 0.40. Actually I think f5 and f6 are correlated for both are best-match method (thus have similar heuristic) in "body" field, so I expect to see a large weight and a small one for these two, however, they each has a large weight, **so how well a query match a document's body field is a vital factor in ranking,** as shown in experiment 3, column comb3**.** Another important factor is f1 spam score, I think it correlated with PageRank score, both are authority measures, and due to the correlation relationship, PageRank has a relatively small weight. As for other BM25 and Indri features, "title" has an large influence, too; while "inlink" field has very small weight thus not a high-quality factor in ranking. Before looking at the model file, I expect a large weight for feature 17, which is normalized window size of all terms, this one is the most time consuming and focuses on “body” field and adding it improves MAP as shown in experiment 2, column “All”, so theoretically it should be a good feature, however it has a tiny weight of 0.02, so maybe because this one is correlated with f5, f6 (all three focus on body field), and f7(f7 is overlap score with weight of 0.32, calculate how many terms occur in a document, the higher the better; f17 is the window size, for document whose overlap score is not 100%, its f17 is 1).

Before looking at model file, I expect BM25 and Indri for same field should be highly correlated, however, in those 4 fields, the weights are (5:0.63 6:0.37), (8:0.26 9:0.05), (11:0.14 12:0.19), (14:0.03 15:-0.05). Some are similar, some are not. Both Indri and BM25 are best-match measures and give scores based on tf and idf, so the difference may be default score of Indri has influence in harder-to-match fields like “url” and “inlink”, so Indri has more influence in these two field, while in “body” and “title”, Indri got lower weight due to correlation.