

PA4 Report

Name(s): Emmanuel S Thomas & Sumit Gupta

Due Date: 06/01/2014

Class: CS276

Professors: Chris Manning & Pandu Nayak

Overview

For this assignment we implemented extended algorithms for ranking documents. As compared to last assignment we learned and hand tuned ranking algorithms – Cosine Similarity, BM25, Smallest Window and BIM. In this assignment we used Machine Learning techniques to automatically learn weights to predict ranking of documents with respect to the query.

Task 1

We implemented **Point-wise** learning with Linear Regression learning model using Weka to learn weights and came up with a ranking function, which is a linear equation (model) to predict score of the test document.

Key points for task 1:

- 1) We used TF for document features and IDF (with la-place smoothing) for query terms to compute TF-IDF score for a Query-Document.
- 2) LinearRegression Model from Weka library used.

Score: The highest NDCG score we were able to achieve on the development set was: **0.87**

Task 2

We learned how to implement Pair-wise learning SVM techniques (Linear and RBF) to come up with different model using pairwise approach where we keep the difference of feature vectors of 2 documents in the current pair.

Key Points:

1) For a given query, we computed all difference feature vectors for every document pair we have for the query. We tried computing reverse order too (($d_2 - d_1 = -1$) for ($d_1 - d_2 = +1$)) to have more data-points in our training sets but it did not help much in boosting any score.

2) Used Standardization for dataset before computing difference of feature vectors. Then finally standardize the difference feature vector too.

Linear Kernel:

Score: The highest NDCG score we were able to achieve was: **0.86**

Non-Linear SVM Kernel:

The highest NDCG score we were able to achieve was **0.87**. Based on our experiments, it appears that the best values for the cost and gamma are close to 2. Below is a table with the different power of 2 used for the **cost** and **gamma**.

C (Cost)	G (Gamma)	NDCG Score
1	1	0.8720
1	-1	0.8708
0	0	0.8699
2	2	0.8685
-1	1	0.8623
-3	3	0.7824
-5	5	0.7543

Task 3

We added more features to SVM we implemented in task 2, namely: BM25 scores, Smallest Window Score and Page Rank and tried tuning the parameter K for Smallest Window. Below are the findings:

K (Smallest Window)	NDCG Score
70	0.8573
20	0.8573

1	0.8569
---	--------

Error Analysis for different Features we used in Task 3:

We found that using more features like BM25, Smallest Window and PageRank did not help much in terms of boosting relevance score. We think relevance could be boosted if we could experiment with other features like user clicks and other user behavior related features instead of using more and more document similarity score other than Cosine similarity we used with the original features. Also, here in pairwise learning we are marking a document better than other but we do not have any relative ranking between 2 documents in the pair to check how similar these 2 documents are.

