Info Retrieval: Project Proposal

Raj Patel
University of Utah
raj.patel@utah.edu

Blaze Kotsenburg
University of Utah
bkotsenburg@gmail.com

Brandon Ward
University of Utah
brandon.ward@utah.edu

ABSTRACT

Our project proposal is an exploratory project with the goal of developing a search engine based on what we have learned in class. We propose to develop a search engine for a dataset containing web page documents. In this exploratory project, we plan to include methods for ranking algorithms, clustering of web docs after a user query, developing a user interface (UI), and using several evaluation metrics to compare relevance of the queries.

KEYWORDS

search engine, learning to rank, clustering

1 INTRODUCTION

Our project proposal is an exploratory project with the goal of developing a search engine based on what we have learned in class. We propose to develop a search engine for a dataset containing web page documents. In this exploratory project, we plan to include methods for ranking algorithms, clustering of web docs after a user query, developing a user interface (UI), and using several evaluation metrics to compare relevance of the queries.

The proposed search engine project also gives us a great opportunity to combine our back-end learnings from the course with a front-end UI. We believe that this will be a great challenge as simplicity in UI design remains a crucial component for any search engine.

2 DATASET

The dataset that we are currently planning to work with contains Wikipedia webpage documents. The dataset was last updated in 2009, but it contains 10GB of web page document data. Since this dataset is so large, we were originally planning ons using it as our corpus for the search engine.

It was brought to our attention during our presentation that the Wikipedia dataset is already categorized making the post-query clustering a redundant feature. We are currently trying to find more reasonable datasets that would allow us to build a search engine that incorporates post-query clustering. We are looking into datasets from Twitter and will continue to explore existing datasets until we find one that is adequate for our project.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 RANKING AND SCORING

We want to perform various different ranking and scoring methods on the dataset so that we have a starting point for clustering the results of a query. We will first rank and score the documents from our dataset with common ranking functions such as the PageRank algorithm we learned in class.

We also want to experiment ranking our documents with machine learning techniques, such as Ranking SVM. From our research of learning to rank and Ranking SVM, we noticed that accuracy in results can be an issue. We've discovered that it is crucial during training to ensure that there is no bias towards queries with a large number of relevant documents. To avoid this issue, we will implement Ranking SVM in a way that optimizes the Hinge Loss function [1].

4 CLUSTERING

We want to incorporate clustering into our search engine in a way that it may still be meaningful for a user. We plan to cluster documents based on a user's query and display the clustered documents in a UI that would allow users to easily sift through different document topics.

Since we have not officially selected a dataset at this point, we don't know what to expect in our data. However, we will assume that the data will be in a random uniform distribution. With this assumption in mind, we will likely use some sort of K-Means clustering in our finished product. We will be testing different clustering models during our experiment and will choose the best k-clusters based on the point of diminishing returns.

5 EVALUATION METRICS

We would like to evaluate a few things within this project. First, depending on the size of the dataset, we would like to evaluate the speed of the different ranking and scoring algorithms. Second, we would like to evaluate the accuracy of the ranking algorithms using different evaluation metrics. Currently we have decided on using Mean Average Precision (MAP) and normalized Discounted Cumulative Gain (nDCG) for the evaluation metrics. Finally, for the clustering aspect of our project, we would like to evaluate similarities of the documents.

6 VISUALIZATION

For visualization, we want the ability for a user to search the dataset with a simple search bar within a webpage. Based on the query, a user will be able to compare and contrast the results of different ranking & scoring algorithms along with different evaluation metrics. In addition to this, the clustering of the results will also be available for visualization. To implement the UI, we are thinking of using the D3 library, Bootstrap, and Javascript.

7 CONTRIBUTION

The individual contributions to the project will be assigned as follows: Brandon will process the dataset and implement ranking functions; Blaze will develop a clustering model for post-query results; Raj will implement the front-end using D3/Bootstrap and collaborate with Brandon in implementing the ranking functions.

REFERENCES

[1] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. 2006. Adapting Ranking SVM to Document Retrieval. *Proc. ACM SIGIR Int. Conf. Information Retrieval (SIGIR'06)*, 186–193. https://doi.org/10.1145/1148170.1148205