CS298 Progress Report

Improving the BM25F algorithm for use with OPIC based crawlers.

Introduction

This project aims at improving the BM25F algorithm for use with OPIC-based crawlers. The BM25F ranking function calculates the page relevance by assigning weights to document fields and the anchor field. The title and body of a document are termed as document fields. The anchor field of a document refers to all the anchor text in the collection pointing to a particular document. Thus if a lot of unimportant links are pointing to a document they can increase the page relevance of an important web page for unimportant word searches that are not relevant to it. Hence the goal is to implement a modified BM25F by combining page rank computed by OPIC algorithm while computing weight for anchor field associated with a web page. The open source search engine YIOOP is being used as a case study for the project. The modified BM25F will provide a better ranking estimate for documents crawled by YIOOP. For the documents crawled in YIOOP, a posting list is set of all documents that contain a word in the index. This posting list is very large and needs to be trimmed to get the most relevant documents.

Work Done in Spring 2011

The spring semester was spent on running experiments on how far we should go in the posting list and decide on an optimum cutoff point for scanning posting list. For comparison and statistical analysis we first downloaded the TREC software from the TREC website. The software was run using a test dataset and the results were studied to understand its functioning. The TREC software requires two files “trec_top_file” and the “trec_rel_file”. “trec_rel_file” contains the relevance judgements, and “trec_top_file” contains the results that need to be evaluated. To generate these files 10 keywords were chosen for the SJSU domain. Then each keyword was searched using 5 popular search engines and top 10 results were retrieved with a total of 500 results. Finally top 10 results were finalized for each keyword using the mean score. These entries were stored in the “trec_rel_file”. After this all 10 keywords were searched using the YIOOP open source search engine and top 10 results for each of them were retrieved. These entries were stored in the “trec_top_file”. Using the generated files the TREC software was run and obtained results were studied.

The next task was to study the group iterator in Yioop to get an understanding about how the posting list is generated and scanned. The group iterator scans the posting list for a fixed number of results.

Conclusion and Roadmap for finishing the project

The next task is to study variations that can be made to the posting list so that the TREC score improves for the same set of keywords. Next semester the results of this study will be finalized to decide upon an optimum cutoff. Later we will come up with an optimal weighting scheme for the BM25F to improve the algorithm.

The spring semester was spent working on the first two deliverables of the CS298. A considerable time was spent on Deliverable 1 as a lot of computational work was involved. Therefore I could not complete all the work that was planned. The remaining deliverables will be completed next semester.