Major Technical Project
Report1 v1.2

Title of MTP

Building a Search Engine

Abstract of MTP Proposal

A web search engine is basically a tool which is widely used to search information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. The motivation in our case is to build a search engine that provides you with the list of professors from premium institutes when queried against a specific specialisation.

Description

The primary task in order to implement the search engine will be crawling a large dataset of webpages corresponding to the various Faculty, Staff, Deans, Professors and Director of IIT's, NIT's and various other premium Institutes spanning across India. After we have the data and links of webpages, the next step will be to focus on document classification based on the specialisation among the predefined list we will be having. Depending on the specialisation, individuals(webpages) will be indexed corresponding to their respective specialisation. After having a list of pages for a specific specialisation we will be ranking them using standard techniques like hubs and authorities according to their relevance and will output the result in decreasing order of priority.

Keywords : Search Engine, PageRank, Spider, Crawling, Indexing, Big Data, Map Reduce, Hubs and Authorities

Course of Action for MTP and Timeline

The Entire project will be divided into three subparts which will eventually lead to the formation of search engine. The subproblems are :

1. Web Crawling
2. Indexing
3. Searching

For the Implementation, decomposing the above mentioned into important parts of our version of search engine gives us the following points:

- **Crawler**

  Crawler browses the web, downloads documents(pages) and saves them. Crawler task will be to begin from a given page and extend in such a manner that all the webpages related to the professors and teachers from all the prestigious institutes of India are crawled. All the pages will then be downloaded and saved for further processing.

Deadline : 11th October 2014

- **Parser**

  After Crawler is done with crawling all the relevant contents in an unorganised manner, parser will convert the content into a format which will be suitable for Indexing.

  Deadline : 4th November 2014

- **Indexer**

  Based on the predefined list of the specialisation we are interested in, structures will be created by the Indexer after processing the documents passed to it by the crawler. After processing we will be having a dictionary kind of thing where key will be specialisation and value will be the list of documents with the respective key given prime importance.

  Some of the specialisation we will be looking at are namely::

Specialization
Image Processing
Information Theory
Wireless Communication
Computer Networks
Machine Learning
Pattern Recognition
Artificial Intelligence
Human Computer Interaction
Software Technology
Distributed Software Systems

Deadline :

- **Ranker**

  Ranker is the part that arranges documents in some particular order. The goal is to put the most relevant documents on top of a results page. Usually machine learning is used here, but there is the difficulty of obtaining data-sets(for test and training). One option is to use BM25 or some other text ranking function.

Deadline :

- **Snippets Generator**

  In order to identify the page, some of the contents needs to be showed along with the link. Snippet generator does this thing by extracting short extracts from the document and then display them beneath the link.

  Deadline :

- **Searcher**

  Searcher takes queries entered by users, transforms them somehow, parses them and finds documents containing words of query, gets all the necessary stuff like annotations and finally generates and presents results pages.

  Deadline :

Work Carried out till now

After studying about the working and model of crawler, started working on building it in order to fetch the required content.
Studied and Implemented various ranking algorithms on Big Data that could be used to rank various pages after we are done with indexing part. Algorithm includes Hubs and Authorities and Page Rank and Topic Sensitive Page Rank.

Team

Name : Sahil Mutneja
Roll No : B11031
Email : sahil_mutneja@students.iitmandi.ac.in
Major : Computer Science and Engineering

Signature

Faculty supervisor name and signature

Dr. Dileep A.D.
Assistant Professor
School of Computing and Electrical Engineering

Signature