Custom Web Search Engine Project

Members

Fatih Hafızoğlu, 1746049, fatih.hafizoglu@ceng.metu.edu.tr
Emre Can Küçükoğlu, 1746239, emre.kucukoglu@ceng.metu.edu.tr
Yiğit İlgüner, 1560762, yigit.ilguner@ceng.metu.edu.tr

Summary

We will design simple web search engine. Engine settings can be modified by user, thus for an enthusiastic searchers, precision and recall values can be increased. Engine takes a website address as an argument and starts crawling until reaching the depth which is also given as argument. Before the crawling operation, user can set

- Initial web address for initiation of crawling
- Depth limit
- Stemming,
- Ranking,
- Compression technique,
- Ignored websites list,
- tba

Retrieved websites automatically indexed and according to user settings, dictionary stored on a server. On the client side, user can search in corpus that is created. User can select as an option in addition to keyword query:

- Pivot website address, (use my query and this website's relation)
- tba

Besides the techniques that are described above Java is used for server-side implementation. Furthermore, we will benefit from the information retrieval functionalities of Apache Lucene.

Sample user story

User always searches 'linux operating systems' based queries. Thus, user sets initial web address as https://www.kernel.org/ which is base for linux kernel. User can set depth limit as 5, and with the help of other options (they will be developed while implementing), can start crawling the web. Since almost all crawled websites are related with linux, corpus's recall and precision values are expected to be higher. After crawling phase, user can set a pivot website. According to query, cosine distance of pivot website will help engine to get intended results. Ranked search results are shown with or without pivot site.