Overall Design of the System

In this project, a **web-based search engine** is built.

On the web interface, user can submit request of fetching pages given a starting url and number of pages to fetch. User can also check the existing database, and clear the database. On the other hand, user can search by submitting a search query, even with quotation marks on phrases to boost ranking of pages containing the phrases. Pages will be ranked by descending scores, and at most 50 pages will be displayed. Page Title, url, last modified date, size of page, top 5 keywords and frequency, and parent and child links will be displayed.

Class Spider, Indexer, Database, StopStem, Porter, SearchEngine are used to develop the search engine with several jsp and html files for supporting the web interface.

Class Spider will fetch pages recursively given a starting url and number of pages to fetch. When fetching a page, Spider will call class Indexer to extract words, titles, bigrams, and trigrams, and insert them to the database through class Database. Dealing with the words, stop words will be removed and then stemmed by class StopStem and Porter. After crawling the pages, the tfxidf/max(tf) matrix will be calculated for each Page and Word, ready for the search process. Class SearchEngine given a query will then calculate scores for each page with favour in matches in title, and matches of phrases, and get the top 50 scores pages.

JDBM Database Scheme of the Indexer

**1. PageID – URL Mapping Tables:**

***PageIDtoURL(PageID, URL), URLtoPageID(URL, PageID)***

For simplicity of storing the URL in other databases, an ID is given to the page and recorded on this mapping table. Both directions are recorded, for efficiency of retrieval.

**2. Table for title of each web page:**

**PageIDtoTitle(PageID, Title)**

The Table is for recording the title of the fetched page.

**3. Table for LastModifiedTime of each web page:**

***PageIDtoTime(PageID, LastModifiedTime)***

The LastModifiedTime is for recording the last modified time of the pages fetched. When the Spider has fetched that page before, and the page is not modified after that, the last modified time from the page and from the table would be the same. Then the Spider will skip this page this time, as it is not modified. This can help handle the cyclic links.

**4. Table for Length of each web page:**

***PageIDtoLength(PageID, Length)***

The length is for recording the length of the pages fetched. All three Content-Length, HTML Length, and Number of Words will be recorded as some may be missing fetching a page.

**5. WordID – Word Mapping Tables:**

***WordIDtoWord(WordID, Word), WordtoWordID(Word, WordID)***

For simplicity of storing the word in other databases, an ID is given to the word and recorded on this mapping table. Both directions are recorded, for efficiency of retrieval.

**6. Forward and Inverted Indexes for storing Parent – Child relationship:**

***ParenttoChild(ParentPageID, ChildPageID), ChildtoParent(ChildPageID, ParentPageID)***

These tables store the Parent – Child relationships. Both directions of retrieval are recorded, for efficiency.

**7. Forward and Inverted Indexes for page and title words on page:**

***PageIDtoTitleWordID(PageID, WordID),*** ***TitleWordIDtoPageID(WordID, PageID)***

These tables store the title words crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**8. Forward and Inverted Indexes for page and word on page:**

***PageIDtoWordID(PageID, WordID), WordIDtoPageID(WordID, PageID)***

These tables store the words crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**9. Forward Index for page and top five word on page:**

***PageIDtoTopFiveWordID(PageID, WordID)***

The table store the top five words crawled from the web pages.

**10. Forward and Inverted Indexes for page and title word bigrams on page:**

***PageIDtoBiTitleWordID(PageID, WordID), BiTitleWordIDtoPageID(WordID, PageID)***

These tables store the title word bigrams crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**11. Forward and Inverted Indexes for page and title word trigrams on page:**

***PageIDtoTriTitleWordID(PageID, WordID),TriTitleWordIDtoPageID(WordID, PageID)***

These tables store the title word trigrams crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**12. Forward and Inverted Indexes for page and word bigrams on page:**

***PageIDtoBiWordID(PageID, WordID), BiWordIDtoPageID(WordID, PageID)***

These tables store the word bigrams crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**13. Forward and Inverted Indexes for page and word trigrams on page:**

***PageIDtoTriWordID(PageID, WordID), TriWordIDtoPageID(WordID, PageID)***

These tables store the word Trigrams crawled from the web pages. Both directions of retrieval are recorded, for efficiency.

**14. Table for TFxIDF:**

***PageIDtoTFxIDF(PageID, TFxIDF)***

The table stores the TFxIDF weights of web pages and words.

**15. Table for SearchHistory:**

***SearchHistory(Index, SearchHistory)***

The table stores the past search history inputted by the user.

Algorithms

**1. Crawl**

If a page is not fetched before, it will not be in the database URLtoPageID, then it will be fetched. If a page is fetched before, it will be in the database URLtoPageID and last modified time will be extracted from the database PageIDtoTime. If the new last modified time is earlier than or same as the one in database, then it will be ignored, and next page in pages\_queue will be tried to be fetched. If new last modified time is later, then it will be fetched.

Fetching a page, it will be added to visited\_pages (for this run) to avoid circular loop. Titles, words, bigrams, trigrams will be extracted and stored after stop word removal and stemming. At the same time, the Indexer will also keep track of the top five frequent words. Links will be stored and added to pages\_queue, if not visited.

After fetching the required number of pages, TFxIDF/max(TF) matrix will be calculated and stored to the database.

**2. Search**

After extracting phrases from the query, and performing stop word removal and stemming, scores of each page will be calculated using the formula below.

**For each matching of query word and title word,**

is the number of documents

is the number of documents with title containing word i

**For each matching of query word and word in page,**

is the frequency of term j in document i

is the max term frequency in document i

is the number of documents with title containing word j

As tf/max(tf) is upper bounded by 1, the score added from matching in title would usually be larger (if similar df). Also the scores from matching in title is considered as a bonus besides of the score from tfxidf/max(tf), which gives favours in matching in title.

**For each matching of query phrase and the title word/bigram/trigram,**

is the number of documents

is the number of documents with title containing phrase i

**For each matching of query phrase and word/bigram/trigram in page,**

is the frequency of phrase j in document i

is the max phrase frequency in document i

is the number of documents with title containing phrase j

Similar to the score from matching in title, the scores from matching phrases in title/page is considered as a bonus besides of the score from tfxidf/max(tf), which gives favours in matching in phrases. The coefficient 2 is to weight more on phrase match than normal title/page match.

Finally, after calculating the scores for each page, pages will be sorted in descending order of scores. Top 50 Pages with non-zero scores will be normalized to 0 to 100 and printed on the web interface.

Installation Procedure

Download the files on Github and unzip it

Add environment var:

*CATALINA\_HOME* = "your tomcat path (e.g. /Users/ABC/apache-tomcat-9.0.74)"

*JAVA\_HOME* = "your jdk path (e.g. /Users/ABC/jdk-20.0.1.jdk/Contents/Home)"

Go to $CATALINA\_HOME/webapps and create a folder "Search"

Put the files under $CATALINA\_HOME/webapps/Search

Then build the program by:

> javac -cp combined.jar WEB-INF/classes/project/\*.java

Then start it by:

> $CATALINA\_HOME/bin/startup.sh

Open in your browser: \*\*\*http://localhost:8080/Search/Search.html\*\*\* to test the program

Shut it down by:

> $CATALINA\_HOME/bin/shutdown.sh