Algorithms

1- Basic flow of Program:

- Initial seed is added manually to the database.

- Crawler crawls them and keeps expanding seed and pushes crawled pages to ranks for re-crawling.

- Indexer gets crawled pages HTMLs and parses them, thus filling the database with useful information to be retrieved when needed.

- Re-crawler crawls ALREADY crawled pages to check for changes, It does partition pages into subgroups each corresponds to a specific frequency.

- Re-crawler crawls each subgroup as frequent as specified by the frequency of the subgroup.

- Query search gets input from the user, parses it as needed, looks for relevant pages and displays them.

2- Seeds Scheduling Algorithm:

- Crawler crawls pages highest priority first.

- Each page has initial priority based on its domain (.com,.net,.gov,.edu).

- If a page gets referenced by multiple pages, its priority increases.

3- Dividing into Ranks Algorithm:

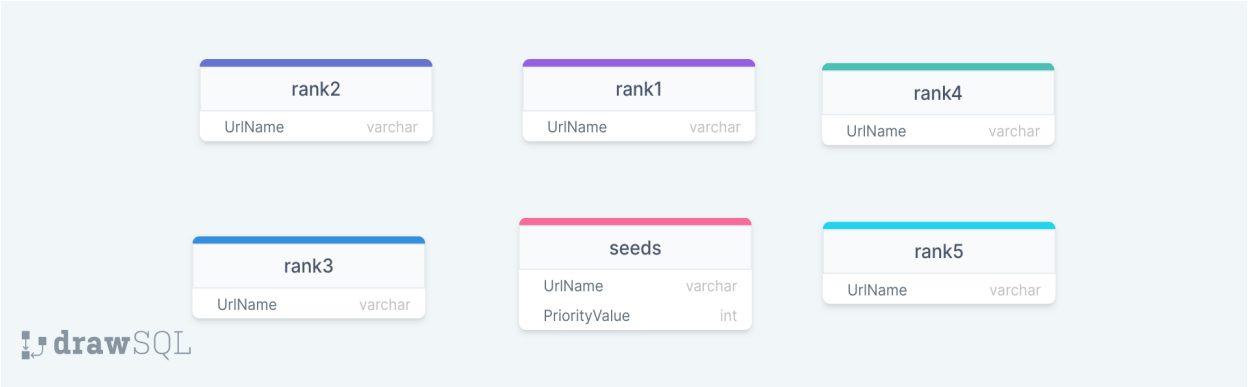
- Each rank represents a subgroup of ALREADY crawled pages which share a frequency of change.

- Each rank gets RE-CRAWLED as frequent as the specified subgroup frequency.

- If a page is NOT-CHANGED then it is pushed to the next rank, which has lower frequency.

- if a page is CHANGED, then it stays at the same rank, and gets scheduled to be indexed.

4- Database Schema(Crawler):



5- Database Schema(indexer):