Design of the RocksDB database

|  |  |  |  |
| --- | --- | --- | --- |
| **Database Name** | **Key** | **Content** | **Special Key** |
| url | URL | page\_id | max\_id (store max page\_id) |
| word | word | term\_id | max\_id (store max term\_id) |
| forward | page\_id | term\_id frequency, |  |
| term\_weight | page\_id | term\_id term\_weight, |  |
| inv | term\_id | page\_id position, |  |
| title\_inv | term\_id | page\_id position, |  |
| parent\_child\_relation | parent\_page\_id | child\_page\_id, |  |
| child\_parent | child\_page\_id | parent\_page\_id, |  |
| page\_info | page\_id | title, date, size |  |
| idf | term\_id | idf(=log2(N/df)) |  |
| history | query | page\_id, |  |
| inv\_url | page\_id | URL max\_tf |D| |  |
| inv\_word | term\_id | word df |  |

**Explanations on the database design:**

**Key and value in databases**

url contains unique page\_id for each URL crawled

word contains unique term\_id for each word encountered

forward contains every word (represented by term\_id) in a page (represented by page\_id) and its corresponding term frequency (represented by an integer following term\_id)

term\_weight contains the term weight (=tf x idf/i(tf)) of each word in a page

inv contains in which pages the word in content is in and its corresponding position

title\_inv contains in which pages the word in tile is in and its corresponding position

parent\_child\_relation contains parent page id and its corresponding child page id

page\_info contains title, date and size of a page

idf contains idf value for each word (for the purpose of saving computation time during search)

history contains recent 50 queries and their corresponding results (for the purpose of speeding up common search queries)

inv\_url contains maximum term frequency (max\_tf) in a page and the length of the document vector (|D|)

inv\_word is the reverse of word with extra document frequency of each word

**Usage of databases**

url, word, inv\_url, inv\_word is for mapping

idf, history, inv\_url, inv\_word, term\_weight is for pre-computing and speeding up searching time.

parent\_child\_relation, child\_parent is for crawling and retrieving

forward, page\_info is for finding and displaying page content

inv, title\_inv is for indexing and matching the query and getting position of word

**Remarks:**

If there are multiple values in each entry of the content, the values would be separated by space. (expect for the page\_info database, it would be separated by a comma)

If there are multiple entries in the content of the key, the entry would be separated by comma.

**Example:**

If the term\_id 2 keyword appears in page\_id 2 with position 45 and in page\_id 3 with position 88, then the content of inv database with key 2 would be 2 45,3 88.