COMP4321 Group 1 Project

BY ZHANG ZHE, KONG TSZ YUI, GUPTA, HARSH VARDHAN

Database Design

The database totally contains 10 tables, these tables are:

1. id_url

   This table contains two column, which are page_id and url. This is the mapping table for page_id and url.

2. page_info

   This table contains 4 columns, which are page_id, size, last_mod_date and title. This gives the basic element of an website, which are page id, size, last modification date and title of the website. last_mod_date are stored in terms of a linux timestamp.

3. relation

   This table contains 2 columns, which are child_id and parent_id. This gives the relation of parent and child, which in later can be used to construct parent/child link relation.

4. word_id_word

   This table contains 2 column, which are word_id and word. This is the mapping table for word_id and word.

5. 6. 7. 8.

   4 tables of page_id_word

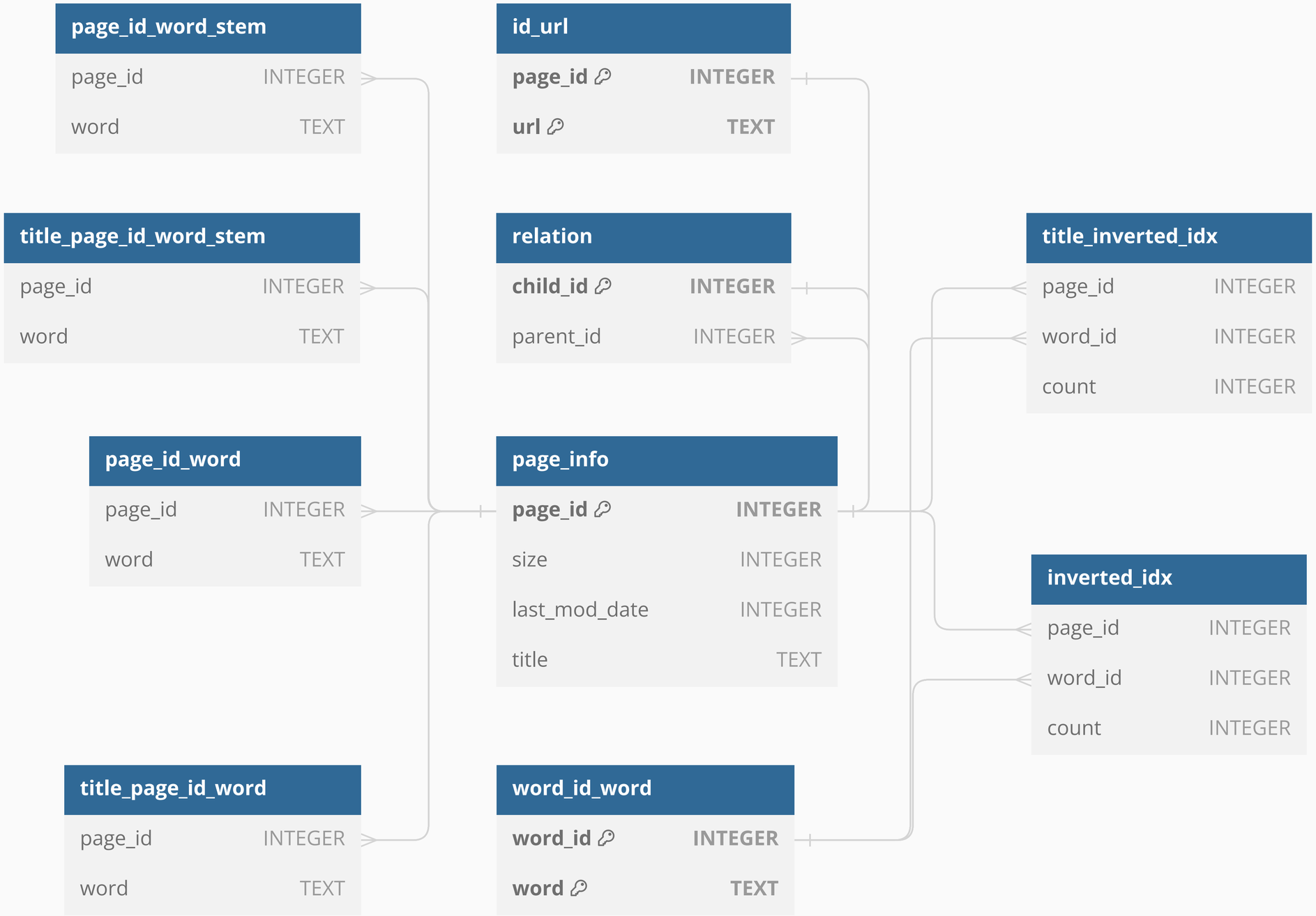
   This table contains 2 column, which are page_id and word. The table gives all the words in a pages. However, those with title_ at front implies this table contains words from title, otherwise it contains words from body. Those with _stem at behind implies the stopwords are removed and words are being stemmed. This table can be treated as forward index.

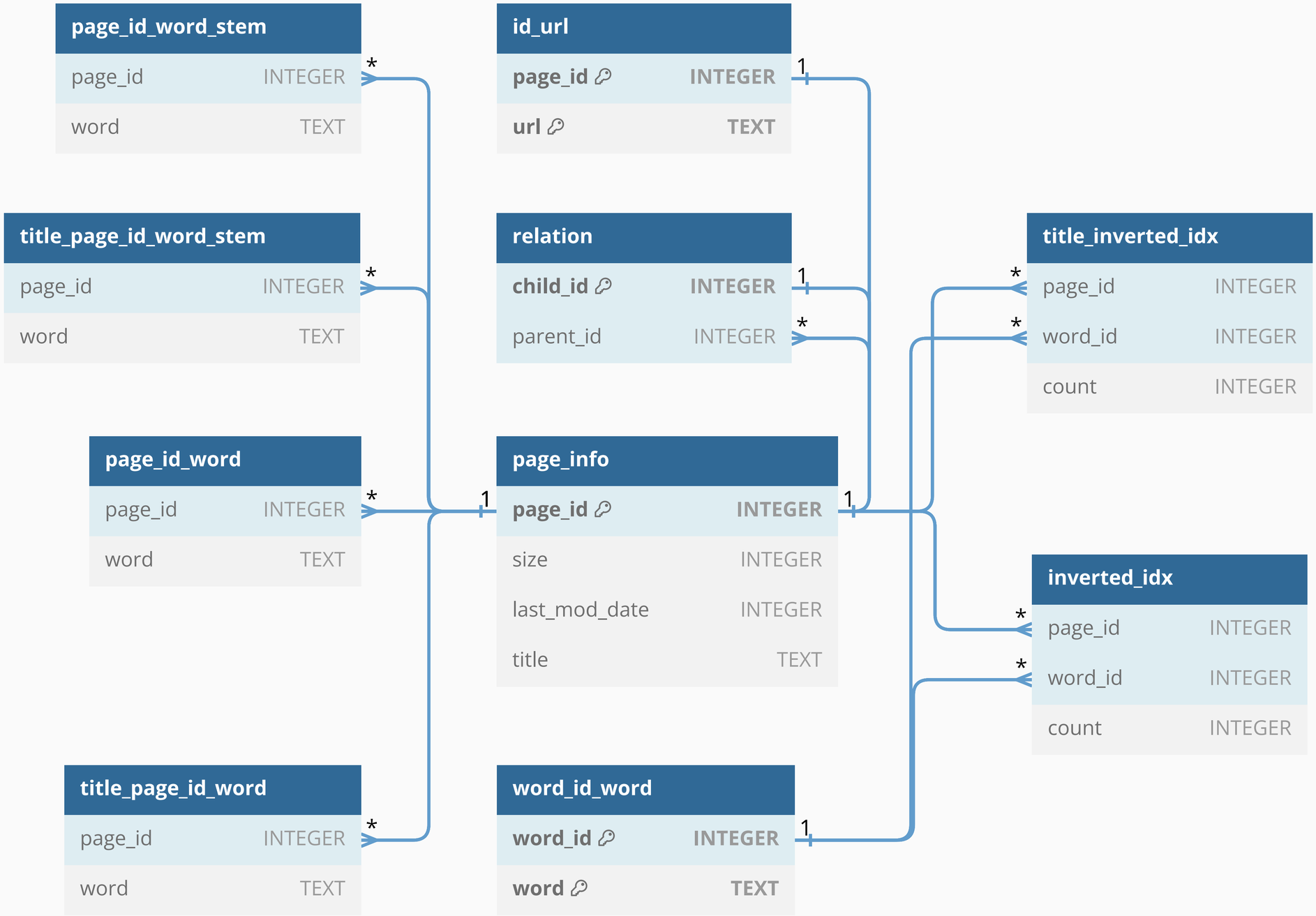
9. 10.

   2 tables of inverted_idx

   This table contains 3 columns, which are page_id, word_id and count. This is used as the inverted index. Those with title_ at front implies this table contains words from title, otherwise it contains words from body.

For more details about the table, one may refer to following pages.




page_id_word_stem
page_id INTEGER
word TEXT
id_url
page_id INTEGER
url TEXT
title_page_id_word_stem
page_id INTEGER
word TEXT
relation
child_id INTEGER
parent_id INTEGER
title_inverted_idx
page_id INTEGER
word_id INTEGER
count INTEGER
page_id_word
page_id INTEGER
word TEXT
page_info
page_id INTEGER
size INTEGER
last_mod_date INTEGER
title TEXT
inverted_idx
page_id INTEGER
word_id INTEGER
count INTEGER
title_page_id_word
page_id INTEGER
word TEXT
word_id_word
word_id INTEGER
word TEXT
*
1
dbdiagram.io