Index Construction

Database Model,
Index Construction Process

Johannes Gosda, Gerald Töpper

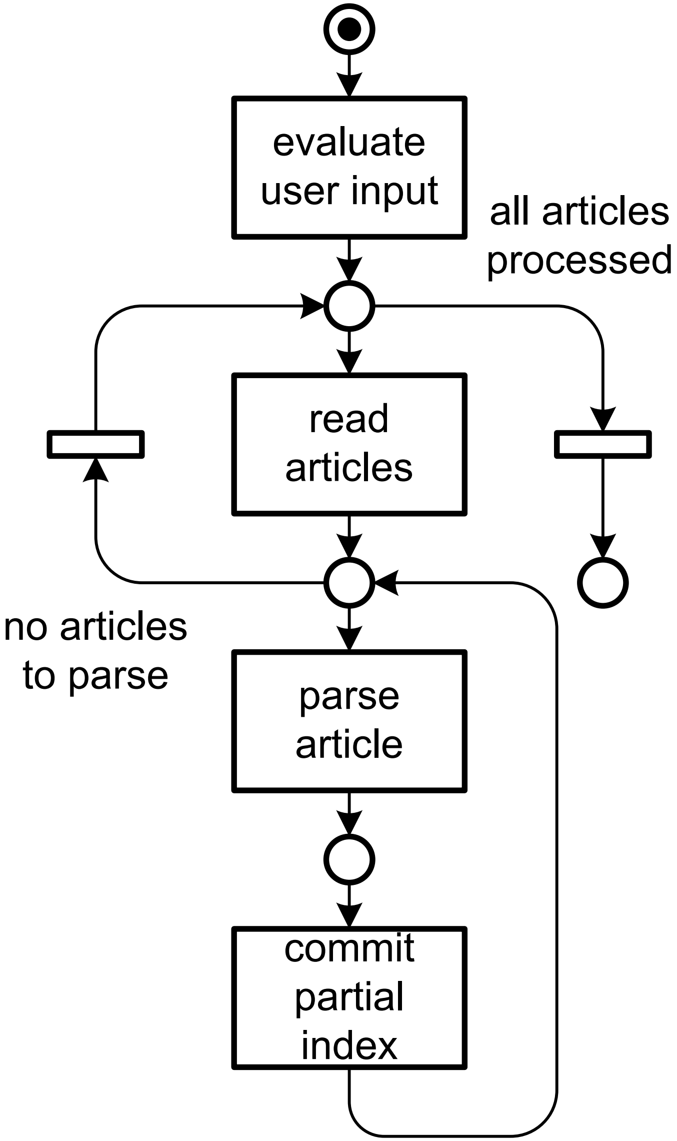
Index

inverted_index_0					
word	articleid	frequency_title	frequency_header	frequency_body	positions
abraham	307	1	0	12	1, 262, 322, 346, ...
lincoln	307	1	1	151	2, 31, 43, 66, ..., 263, ...
lincoln	620	0	0	1	1001

- Inverted index table
- Additional statistics table (e.g. avarage word count)

Index Constructor

- Concurrency
  - □ Program
  - □ Threads
- Parsing
  - □ Apache Lucene Wikipedia Tokenizer
  - □ No markup, no footnotes
  - □ 'a-zA-Z' & '0-9' only
  - □ Stop word removal
  - □ Porter stemming algorithm
- Multiple index tables



Demo

Configurable parameters	
Start article	Range of articles to index
Article count	
Articles per cycle	Number of articles that should be fetched from database at once (per thread)
Thread count	Number of threads
Table count	Number of tables the index is distributed on

- Low main memory usage
- 1 million articles
  - index with > 140 million records
  - 4,5 GiB Data
  - 17h creating time

Future prospects

- Better distribution of articles (to threads)
- Make articles per cycle depended from a given main memory value
- Dependable parser
- Individual stop word lists
- Support updates without complete new indexing