ACM SIGMOD Programming Contest 2017
Team: Mccree (Peking University)

Xupeng Li, Xuecan Yan, Yiru Chen
Advisor: Bin Cui

Task Overview

- An N-gram is a contiguous sequence of N words.
- In this contest, the task is to search documents and return N-grams from a given set, as quickly as possible.
- N-grams will be added into or deleted from the set during processing.
- Program should return all the N-grams occurring in a documents in the order of their first appearance.

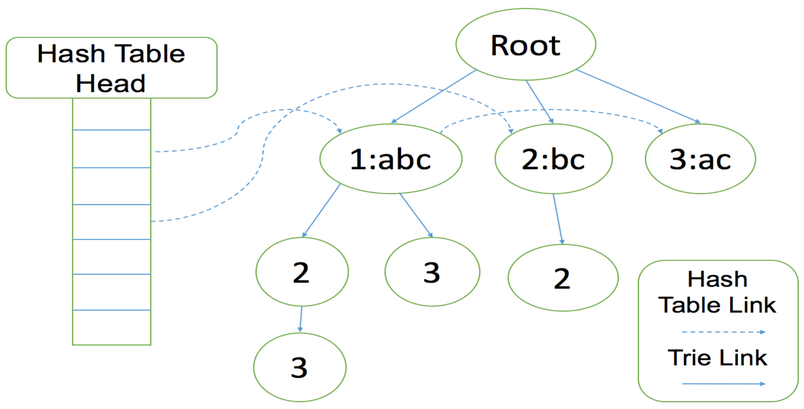
Our Strategies

- Store and retrieve N-grams within several Tries. Hash number of the first word of an N-gram is the identity of which Trie should this N-gram be stored.
- Treat ‘add’ and ‘delete’ operations in the same way: recording operation timestamps of each N-gram.
- Heuristic rules for fast searching based on the number of documents in a batch.
- Fast I/O with no-buffer implementation.

Data Structures

Integrating Hash-table and Trie into one data structure:

- Basic Method:
  - Mapping words to word-IDs by Hash-table
  - Building Trie for N-grams based on word-IDs.
- Our Method:
  - A Trie is also a Hash-table in a way that we consider the node-IDs in a Trie as word-IDs and use Hash-code to retrieve the nodes in the Trie’s first layer.
- With our method, when we get the word-ID of the first word of an N-gram, we also get the node of the first layer of a Trie. And in most time we can already find this N-gram is not exist, in only one Hash-table access. If we use basic method, we needs two -- a Hash-table access and a Trie access.

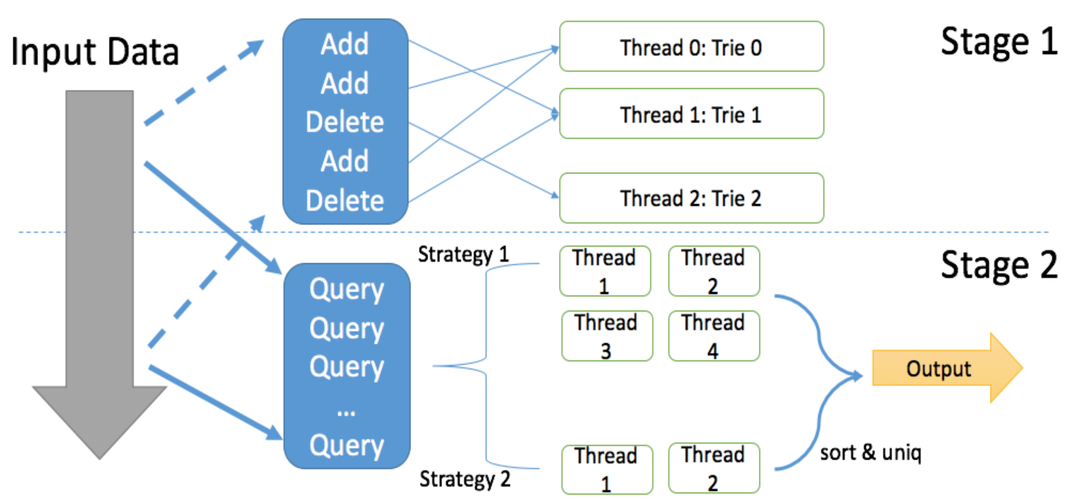


Methods of Parallelization

In each batch, we first deal with all add and delete operations, then process queries.

- Add and Delete N-grams:
  - We create several threads for dealing with adding and deleting N-grams, which are called add-delete threads.
  - For no-lock parallelization, we maintain a local Trie in each add-delete thread so that no data access between different threads.
  - When adding or deleting an N-gram, we first calculate the Hash-code of its first word, then decide which thread should hold it based on the Hash-code.
- Query Documents:
  - We create several threads for query, and use two strategies for different scenarios.
  - If the number of documents in a batch is small, e.g. less than thread number, we concatenate all the documents to one, cut it into the same length and assign the segments to the different threads.
  - If the number of documents in a batch is large, we simply assign different query to different threads.
  - We start from each word of all documents, identify which Trie we should search begin with the word, and search N-grams in the Trie with the following words.

Program Framework



Third Party Libraries

- Boost 1.58
- Intel TBB 4.4