ACM SIGMOD 2013 Programming Contest

Peking University, China

Team phoenix: Jinyu Yao

Task: Streaming Document Filtering

The general idea of the task is to filter a stream of documents using a dynamic set of exact and approximate continuous keyword match.

Specifically, the goal is to minimize the latency with which documents are disseminated to active queries. Queries can be dynamically added to and removed from the system. Whenever a new document arrives, the system must quickly determine all active queries satisfied by this document. Queries and documents are represented as a set of words. For a document to satisfy a query it should contain all the words in the query.

Three types of keyword matching must be supported: exact matches, approximate matches under an edit distance constraint, and approximate matches under a Hamming distance constraint.

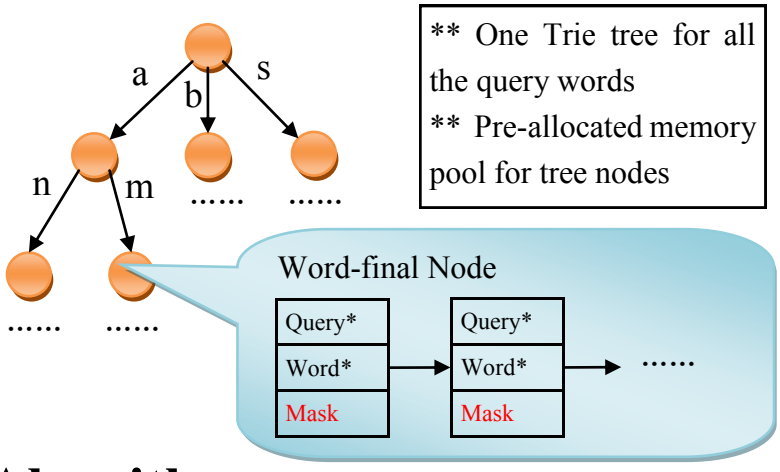
Evaluation Machine

- Processor: 2.67GHz, 12cores
- Memory: 96GB

Final Test Data

- ~500,000 Concurrently active queries
- ~400,000 Documents

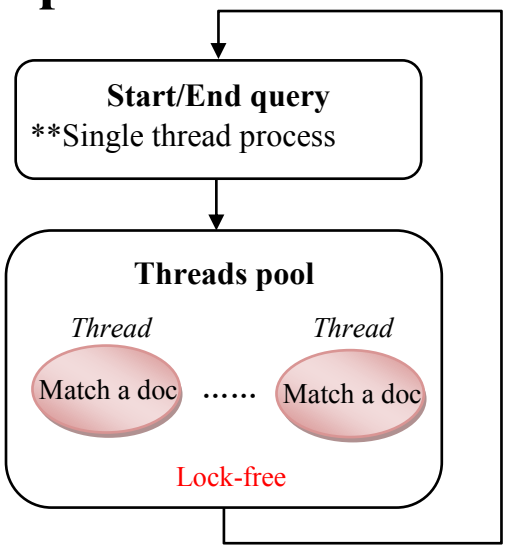
Data Structure



** One Trie tree for all the query words

** Pre-allocated memory pool for tree nodes

Implementation



Algorithm

- **Method #1: edit match**

Manipulate only on the document side. Insert the original query words, enumerate all the changes of document word and then look up them in Trie tree.

Q-Word: p h o e n i x
D-Word: p **n** o **e** n i x y

** Light query operation

** Heavy matching operation (use cache to accelerate the process)

- **Method #2: delete match**

Delete on both query and document sides. Enumerate all the delete changes of query words and insert them into Trie. For matching operation, also enumerate all the delete changes of document words and look up them in Trie. When a match happens, calculate the minimum edit distance based on the positions of deletion.

Q-Word: p **h** o **e** n i x
mask: 0101000
D-Word: p **n** o n i x **y**
mask: 0100001

** Balance the complexity of operations and leverage the huge memory

** The calculation of minimum edit distance is not related to real word. Only the positions of deletion (mask) matter. We used dynamic programming to calculate mask matching during initialization. In total there are 5489 masks.