**Analyzing Persistent In-Memory Text Inversion**

**High Priority Tasks**

1. mergeable-indexer.C : What this indexer does is use a fixed size hash table in PM. Then when the table reaches a specific size, converts it into an unsorted string table (UST). The result is several USTs in a \*single\* file on PM. The format of the UST is a little different from the UST format we have now. Specifically, there is room for linking the different segments in a final step called merge().

Please check ds.h and merger.C in the SPIRIT directory. Also check the pseudocode of merge algorithm in the summer project repository.

2. Implement query search functions for two terms (words). Two approaches: (1) Use a direct pointer to postings in PM and copy the matches in a DRAM buffer (2) Copy the postings in DRAM buffers and then do the matching.

**Read Section 2 of book,** Information Retrieval: Implementing and Evaluating Search Engines

Also, in the **query directory** of **psearchy**, read the code and.h and and.C

**Measurement-Related Tasks**

1. What is the relationship between search latency (QPS) and BLOCKSIZE? What happens to QPS as we increase the BLOCKSIZE?

2. For a given BLOCKSIZE, quantify the internal fragmentation or waste. For example, if the BLOCKSIZE Is 4096, what % of the blocks are not fully utilized? What is the total memory unused (i.e., where a block is used by a bucket but only partially, add for all such blocks and plot a sensitivity graph)?

3. Store the per-query latency in a per-thread array of floating points. And print the arrays at the end of the main function. (Tail latency plots without doing printf())

5. Analyze the CMAP indexer and query evaluator.

**Other Tasks (future, advanced, mainly here to remember)**

1. Concurrently index and search the hash table.

2. Break down the search latency into different components : (1) copy (2) lookup

3. Lazy UST construction

4. Sorted hash table (advanced)

5. Persistent hash table and UST

6. Think of optimizing the hash table so that it performs the same as the UST at search time.

7. Huge pages (advanced)

7. Postings cache