11/6/2015

Avantika Borikar (BT12CSE045)

Sarvesh Parab (BT12CSE054)

Prakhar Sethi (BT12CSE059)

Shaikh Mohammed Danish (BT12CSE072)

Information Retrieval

BM25 Ranking Algorithm

Avantika Borikar (BT12CSE045)

Sarvesh Parab (BT12CSE054)

Prakhar Sethi (BT12CSE059)

Shaikh Mohammed Danish (BT12CSE072)

**Problem Statement:**

Simulate a document partitioned search engine using BM25 ranking function. Build an index for 100%,50%,25%,12.5% of a corpus. Choose any corpus available on internet freely. For each index size measure the average time per query. What do you observe? How does this affect the scalability of document partitioning?

**Technical details:**

Corpus:

We have used ANC’s MASC corpus which has 53 documents with 36236 distinct words after stop-word elimination.

<http://www.anc.org/MASC/download/masc_wordsense.zip>

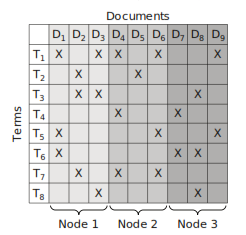
Stop-word list:

a about above after again against all am an and any are aren't as at be because been before being below between both but by can't cannot could couldn't did didn't do does doesn't doing don't down during each few for from further had hadn't has hasn't have haven't having he he'd he'll he's her here here's hers herself him himself his how how's I I’d I’ll I’m I’ve if in into is isn't it it's its itself lets me more most mustn't my myself no nor not of off on once only or other ought our ours ourselves out over own same shan't she she'd she'll she's should shouldn't so some such than that that's the their theirs them themselves then there there's these they they'd they'll they're they've this those through to too under until up very was wasn't we we'd we'll we're we've were weren't what what's when when's where where's which while who who's whom why why's with won't would wouldn't you you'd you'll you're you've your yours yourself yourselves

**Assumptions:**

1. Words are considered case insensitive.
2. Threads are used to mimic parallelism. Every partition is run on a separate thread concurrently.

**Advantages of document partitioning:**

1. Enables fast parallel processing of queries.
2. The main advantage of the document-partitioned approach is its simplicity. Because all index servers operate independently of each other; no additional complexity needs to be introduced into the low-level query processing routines.

**Conclusions:**

1. Document partitioning reduces the query time.
2. Fine-grain or coarse grain partitioning doesn’t show sufficient improvement in performance as compared to intermediate grain partitioning.