IR : Assignment 1

Harshvardhan Kalra
2014043

Skip connections work as follows: given a set of keys, to search for the documents satisfying the Boolean query an user asks, we merge the sorted postings lists of each of the tokens in the query, and handle accordingly. This merge procedure takes time linear in the length of the postings lists in case the query is of ‘AND’ type, which can be reduced to amortized sublinear time by using skiplists.

Skiplists are lists with extra pointers for looking ahead more than one step. The query ‘reading and books’ has been considered for analysis as an example. The times and skips have been observed as follows, averaged over 5 runs:

Skip distance	Time (in seconds)	Number of skips
1 (no skipping)	0.01497197	0
3	0.01179289	915
5	0.01126098	369
10	0.01238393	78
50	0.01250386	0

The number of skipped elements is the highest for skip distance being 3, and the time to process increases as the skip distance is increased. This follows from our intuition that the postings list elements are relatively closely clustered, and the size of these clusters is small.

The top 100 unigrams were selected and used to evaluate the performances of the system. They can be found in ‘results’. The trend is similar to the example we saw previously, with there being a reduction in the processing time upto 5, and then an increase.