March 11, 2014

I carefully reviewed your proposal and I found that it is well done. There is no major issue in your proposal. The followings are my minor questions:

(1) In "PERFECT FILTER", there is a statement, "More importantly, it only requires computing two hash functions. Thus, it has a significant computational advantage over Bloom filters", while in "DISADVANTAGES", there is another sentence, "The time to construct a Pf is slower compared to a Bloom filer, although it is still linear in the number of members". These two sentences sound antagonistic to each other. Did I misunderstand anything?
(2) I try not to complicate your thesis, but do you plan to cover relevance searches for phrases that consist of multiple words? I believe relevance searches for phrases are easy to implement and scoring and ranking documents based on relevance searches for phrases should be straight forward (please correct me if I am wrong for this assumption).

For experimental designs, can you come up with your experimental designs? By saying "experimental design", I mean the definitions of:

(1) What are the metrics we will measure?

(2) What are the parameters we will input?

(3) What values will we change to study their effects to the metrics we defined for (1)?

It will be the best, if we can name each experiment with well-defined input parameters. For example, if the primary input parameter that will be changed is "block size" in the number of distinct words, we name the experiment as "Block Size Experiment" and enumerate the block sizes that will be used for the experiment.

I usually developed multiple "experiments", one for each input parameter that will be studied. Those multiple experiments were derived as "the base experiment" as the core. The core experiment will be the "base case" for all experiments. For example, if block size is, say 50 words for the base experiment, "block size experiments" changed the block size to 100, 200, 400, 600, 800 words and so on. I applied the same strategy to other experiments.

One of the questions for this strategy is "how to define the base experiment"? I usually tried some arbitrary numbers to see if the set of numbers can be a good starting point for any interesting outcomes. For example, if the block size is set to "10 words" in the base experiment, and by increasing the size by 10 words for each experiment (e.g., 10, 20, 30, 40, 50 60 and so on) did not make any interesting effect to any metrics I should observe, I stared over the same for 50, 100, 150, 200, and so on). I repeat this for other input parameters until I figure out where is the best place to start to show interesting outcomes (by the way, I hope you keep the outcomes from doing this, since some of them may be used as the outputs from your real experiments).

Please feel free to let me know if there is anything I confused you.