Responses to Reviewers Comments

June 15, 2007

1. Language and Formatting Changes

   • p.8, l.1,2 of the last paragraph: \( \sum_i^k \lambda p_i k_i \) should be \( \sum_i^k \lambda p_i S_i \). \( \sum_i^k \lambda S_i \) should be \( \sum_i S_i \). - p.9, l.5 from bottom: \( LS_i \) should be \( S_i / L \).
   • p.9, l.3 from bottom: \( \lambda p_i S_i > LS_i \) should be \( \lambda p_i S_i > S_i / L \).
   • p.10: in Alg 2 (which I would actually suggest to completely replace, see above) min, max are never initialized.
   • p.2, l.5: server -¿ servers
   • p.3., l.5: "related work to our own" sounds not so good. Just "related work" would be fine.
   • p.3, sec.2.1, l.9: space missing before [21]. Throughout the paper a couple of spaces are missing before "[", "(", and after ",". Please search for all opening brackets/citations/commas and check!
   • p.6, after the bullets, l.2: "it it" -¿ "it is"
   • p.11: I don’t like the wording "approximately optimal!"

We have gone through the paper again and corrected all grammar/formatting issues mentioned above and in our own review. We appreciate the ones that were pointed out to us and have given all of them another look.

2. I strongly urge *all* co-authors to go through the section carefully and improve on it! It still seems rather clumsy for such a simple algorithm. The correctness should be proved (in Algorithm 1 you state that the result is "optimal", you need to argue why) and the runtime can be improved. The expression you give for the runtime at the end of Section 4 is rather ugly and far too high: Instead of doing a binary search over \( L \) in Alg 1, you can do a binary search over \( k \). For each \( k \) one obtains an upper and a lower bound for \( L \) (from \( p_k \lambda L = 1 \) and \( p_{k+1} \lambda L = 1 \)). This replaces Alg 2. In lines 10-11 pullBW and pushBW are computed as now. Lines 12-16 are replaced by a check whether there is an \( L \) within the computed lower/upper bounds s.t. \( pushBW = sizeTotal_k/(2L) \). If yes, choose \( L \) as small as possible and then the algorithm is done. This of course needs a proof of correctness as well.... The described modification leads to a total runtime of \( O(logn) \) instead of the term you give, which even includes a 1/eps. How would you prove optimality, if the algorithm actually leads to a potential error of epsilon? By the way: it was already described in other words in my previous comments (comment number 7 in the list) how to obtain this improved runtime.

We have given much thought to both our current algorithm and the alternate algorithm described above and believe that our algorithm still provides a better split of both bandwidth and documents that the alternate algorithm proposed does. While the algorithm proposed would seem to provide the same division of documents, it does not guarantee an optimal split of the bandwidth. This is the purpose of our binary search over the latency. Doing this kind of search ensures not only that the documents are split correctly, but that the bandwidth is split correctly as well. The further explain our point, we have added an additional experiment into the paper which shows a brute force search for both the bandwidth and document division and the result of running our algorithm on the same data. We also believe our assessment of the runtime is correct based on the algorithm we have proposed.
3. Referring to comment number 10: I still do not believe that document classification is the bottleneck in the system. This is an $O(\log(n))$ algorithm!!! If it takes too long, you should update your implementation. For each request a document must be located on the disk to transfer it to the client(s). This surely on a standard computer takes orders of magnitude longer than a reasonable implementation of the described algorithm would, for any realistic value of n. When you argue about the advantages of the multicast pull channel other factors of course might still play an important role. E.g., the delay which is incurred by the report probabilities $s_i$. This should still leave enough room to see the advantage of the multicast pull channel. (again, see previous comment 10!)

We have rewritten the paragraph which mentioned the runtime being the bottleneck to include other factors that can delay the system from immediately adjusting to the changing in document popularities and thus providing a purpose for the multicast pull channel.

4. In Section 5 you should actually mention how these report probabilities $s_i$ are chosen. On page 8 a parameter $\beta$ is introduced which influences them. It is never mentioned again... you should give the choice of $\beta$ in Section 5 (perhaps additionally a chart with the resulting probabilities might be nice?).

We have added into the experiments section for both the regular experiments and the planet lab experiments a brief sentence about the report probabilities and how they were handled during the experiments presented in the paper.

5. Referring to comment number 9: you still draw the same questionable conclusions from figure 5. You even mention that this is ”the most interesting feature in Figure 5 (bottom of page 14). I do not see this from the figure at all!! This might be visible for even higher $\theta$, but not in the range you show.

We have changed the writeup about this figure to mention the fact that the tendency for improvement is as the theta increases, as we acknowledged in our previous response to comment 9.

6. Referring to comment number 8: you did not do experiments on real data sets. These are still randomly generated requests and not, e.g., web-server logs. But I certainly wouldn’t expect you to do this! The new experiments are interesting, although describing them as ”real world” experiments (caption of Section 6.4) is misleading.

We have changed the title to this subsection to be Simulated Real World Network... to help explain that these experiments are on a simulated real world network but are not, as the reviewer mentions, on real world data.

We would like to thank the reviewers for their time in reviewing this paper and for the very helpful comments which allowed us to greatly improve our paper.