Prof. Jian Pei,
Editor-in-Chief,
IEEE Transactions on Knowledge and Data Engineering

Dear Prof. Pei,

Attached please find a revised version of our submission to IEEE Transactions on Knowledge and Data Engineering, *K Nearest Neighbour Joins for Big Data on MapReduce: a Theoretical and Experimental Analysis*.

Thank you for your constructive comments on our manuscript. The paper has been substantially revised according to the referees' comments. Before addressing them more precisely, we would like to address the general comments.

(1) Depth of the empirical and theoretical analysis:
We added a theoretical analysis about the load balance for RankReduce in Section 4.1.
We extended the theoretical analysis of complexity of each methods in Section 4.3, and added the analysis of communication overhead, explaining better how the three factors (jobs, tasks and candidates) affect the performance.
Except for the first benchmark where we compute the adequate number of nodes, we re-ran all the benchmarks to have the communication overhead metric in the experiments, with a corresponding analysis added in Section 5.1.3 and in Section 5.2.3.
We added the recall and precision metrics for the approximate methods (H-zkNNJ and RankReduce), and discuss them through Section 5.

(2) Some important methods have not been included in the analysis:
We cited two papers suggested by the reviewers, one as a related work [21], the other one as a possible local improvement for all methods. However, regarding other mentioned papers, for example the discussion about the accuracy of z-value and LSH methods, we think that they are out of the scope of this survey.
We added the missing discussion suggested by the reviewers about theoretical complexity, recall and precision.
Finally, we added a discussion about some open issues and future work.

(3) Presentation of the work and results:
We extended our introduction with possible applications and challenges.
We added a description for each figure describing a method, in order to help the readers to get a general idea of their principles.
We re-organized and extended the discussion in Section 4 with more details in order to relate better the theoretical analysis and experimental evaluation.
We re-wrote the description of the different setup for the experiments to make it more readable and easier to reproduce.
We changed legend symbols for the figures to make them consistent.

We will now address the specific comments of each referee.

Response to the comments of Referee 1.

[R1C1] *section 2.1, definition 3 for aknn(r,S), "where s^k is the exact kth nearest neighbor of r in S." Here, s^{k*} should be the exact kth nearest neighbor of r in S. this typo also appeared in its conference version, which should be corrected in this journal version.*
[Response] Thanks. We have corrected this and put a footnote for the erratum to be applied in the conference

version.

[R1C2] *section 3.2.1, in paper[27], the upper bound of the partition is computed with respect to any r $\in R_i$ (R_i is a cell of R) instead of the pivot of the cell R_i. Therefore, the author may need to address their statement correspondingly.*
[Response] This has been fixed, thank you.

[R1C3] *section 3.2, the authors give a few of figures for various method without further explain the figure in detail. I'll suggest the authors walk through the figures and help the readers to get a clear idea of each method.*
[Response] For a more comprehensive description, we have updated the three figures picturing the three partitioning strategies in Sections 3.2.1 and 3.2.2. We added a detailed explanation for each of them after they are cited in those Sections. Thanks for this suggestion.

[R1C4] *section 5, a detailed dataset setup is missing. To make the experiments more readable and easy to repeat by the others, I suggest the authors to add a dataset setup section.*
[Response] We re-wrote and re-organized the setup part of the experiments in the introduction of Section 5. The datasets used for the experiments are much more detailed. Also, we specify the metrics that are going to be evaluated and we expand their formal and informal definitions. Then, for each experimental subsection, we give the exact parameters used in the experiments.

[R1C5] *Starting from Fig 10, the legend symbol for various method are different from Fig 6 to Fig9 which may confuse the reader. I'll suggest the authors to use consistent notation across the paper.*
[Response] The figures are now consistent. Moreover, some figures have been improved to achieve better readabilityy.

Response to the comments of Referee 2.

[R2.W1] *The paper tends to deliver a survey type paper, however, there are some related methods that it does not consider. For example,*

[1] Qinsheng Du and Xiongfei Li, A novel knn join algorithms based on hilbert r-tree in mapreduce, in Computer Science and Network Technology (ICCSNT), 2013 3rd International Conference on, Oct 2013, pp. 417 420.

[2] Changqing Ji, Tingting Dong, Yu Li, Yanming Shen, Keqiu Li, Wenming Qiu, Wenyu Qu, and Minyi Guo, Inverted grid-based knn query processing with mapreduce, in Seventh ChinaGrid Annual Conference, ChinaGrid 2012, Beijing, China, September 20-23, 2012, 2012, pp. 2532.

[Response] We have cited Paper [1] in section 3.3.2 as a technique, aside the classical R-Tree technique, in order to index locally the data of each block. This is indeed a possible local improvement. We did not go into much details though because the focus of this paper is on distributed techniques. Also, we cited Paper [2] as a possible distance-based partitioning method and an alternative to the presented distance-based partitioning with Voronoi diagram. However, we did not implement it because of its close resemblance to the Voronoi diagram technique for two dimensional data.

[R2.W2] *The main problem of the paper is lack of significance.*
[Response] The kNN problem receives a lot of attention in the database community, so we think it is worthwhile to compare kNN techniques and discuss which one is adapted (or not) to which case.

To the best of our knowledge, no published paper has ever compared the surveyed methods all together:
1) Both theoretically and experimentally
2) In the same environment with the same hardware and settings
3) On the same sets of data
To make the reader aware of those challenges, we have added the aforementioned arguments in the introduction

of the paper (Section 1), that we have modified in this sense. Indeed, by just comparing the published paper at a glance, one cannot conclude on which method performs the best. Published experiments cannot be compared because they use different dataset (type, size, dimension, distribution), and different parameters of each underlying technique. Only by performing these experiments were we able to give insight on how well an algorithm that was tailored for a particular use case work in general. We believe this is an invaluable feedback.

For all those reasons, we think that this paper has a high relevance and we will do our best to correct the problems that remain in order to make the community benefit from our analysis.

[R2.W3] *The authors attempt to present theoretical analysis but it could not give a very clear and complete results. For example, the load balance analysis for RankReduce algorithm is missing, and there are no direct theoretical conclusions for approximate algorithms in the Accuracy subsection at all.*
[Response] In this revision we have put a lot of effort in making the theoretical analysis section (Section 4) more complete and informative. We have added one paragraph in Section 4.1 (last paragraph) to discuss the load balance of RankReduce. We have also added elements of discussion for approximate algorithms in Section 4.2 in order to give the readers a general direction. However, as the accuracy depends on many external factors, such as data distribution, no general conclusion can be established.

Please check the next answer for the complementary changes in the theoretical analysis section (Section 4).

[R2.W4] *The authors propose some additional performance factors such as the number of MapReduce jobs, the number of Map/Reduce tasks, and the number of distances to compute and to sort, however, how can these factors affect the performance is not clearly presented in the experimental results.*
[Response] Section 4.3 has been extensively reworked. The theoretical analysis of the communication overhead and the computation overhead are not simple combination of the three mentioned factors, this is why it is difficult to relate them. In Section 4.3, we have added more discussion on the communication overhead of the algorithms. In particular, precise complexity are given in the text and summarized in Table 1. We have added a complexity of the final candidates found by the algorithms. In the text, we infer some conclusion from the candidate set complexity, which, we think, gives a insight on overall computation. Now, the paper exposes clearly which algorithms are expected to be efficient, given the complexity analysis. This motivates the experimental evaluation section (Section 5).

[R2.W5] *This paper also does not present very important performance metrics such as communication overhead and recalls for approximate solutions.*
[Response] We took in consideration the two pointed out metrics and re-ran all benchmarks.

In particular, we computed the communication overhead of all the algorithms for the Geo and SURF datasets (Sections 5.1 and 5.2) taking into account the variation of the input records and the variation of k. The results can be found on Figure 8 (Geo dataset) and Figure 11 (Surf dataset). We added an analysis of the communication overhead respectively in Section 5.1.3 and Section 5.2.3.

We also worked on the recall metric. In the submitted version, what we called accuracy was in fact the recall metric you mentioned, given by the formula: $\wp = \frac{|A(v) \bigcap I(v)|}{|I(v)|}$. We updated accordingly the paper wherever this misnaming appeared. We have also added the precision metric (introduction of Section 5). All these metrics are presented in Figure 6c, Figure 7c, Figure 9c, and Figure 10c, Figure 12b and Figure 13b and thoroughly discussed.

[R2.W6] *The high-dimensional datasets for experiments are rather small, thus the experimental conclusions are less convincing for BIG data.*
[Response] We agree that the size of all considered datasets is rather small. We actually did not expect that the runtime for all algorithms would be so high, even on modest datasets. Nevertheless, with the considered datasets, the algorithms are already strained enough to draw informative conclusions. We don't see much point

in pushing further the size of the datasets because we clearly see on the figures when an algorithm reaches its peak performance. Also, it might be worth noticing that we went up to data in dimension 386, which is larger than any published experiment we are aware of.

[R2.W7] *The experimental figures could be further improved and should be more consistent.*
[Response] The figures are more consistent now. Moreover, some figures have been improved to achieve higher quality.

Response to the comments of Referee 3.

[R3C1] *I see that the authors mentioned challenges in k-NN in the introduction section. Can they put more challenges in k-NN and make a small subsection. It may help readers to see.*
[Response] The introduction has been reworked to make clearer the challenges in kNN as early as possible in the paper. More precisely, we expose the complexity of the naive solution to understand the benefits of more advanced algorithms.

[R3C2] *Also, one application of k-NN can be described in detail, and all the other applications can be summarized in the same subsection. By following the points 1 and 2, the introduction part will look more efficient and readable.*
[Response] We have added possible fields of application in the second paragraph of the introduction together with the associated references. After that, we have explained a more precise kNN application that consist in finding similar images using feature descriptors (data in dimension 128 – later called SURF data in the paper), which we use later in the experimental evaluation. In addition, this justifies the analysis of kNN for data in high dimensions.

[R3C3] *The authors reviewed and classified the paper in a well manner. However, what are the problems of the papers? This point is missing in all the reviewed papers. If they can add 2-3 lines about the problems for each paper, it will be good. For example, one approach is described at lines 14-23 on page 4 without any disadvantages/problems in that.*
[Response] In the surveyed papers some aspects were missing in the expriments. We have pointed out those issues for RankReduce and PGB in Section 3.1 We also completed the comments for each method in the experimental evaluation section. In particular, the conclusion for PGBJ has been updated in Section 5.4.3, the one for H-zkNNJ has been updated in Section 5.4.4, and the one for RankReduce in Section 5.4.5.

[R3C4] *In Section 4.3, the authors may also consider the replication cost and its impact on the data transfer between the map and the reduce phase.*
[Response] We have completely reworked Section 4 to have a better complexity analysis. In particular, the second part of Section 4.3 takes a particular interest in the communication overhead. We did not expand there the replication cost because it would be difficult to compare it between surveyed methods (no point of comparison) and thus would not been relevant here. However the replication cost is intrinsically taken into account in the communication overhead analysis, since replicated data are also present in the shuffle phase.

[R3C5] *It may be good to write a tradeoff between one round and multi-rounds MapReduce-based k-NN finding.*
[Response] We have enriched our discussion in the beginning of Section 4.3, discussing more about balancing the number of jobs and tasks, and how the three factors can affect the performance.

[R3C6] *The authors are wrapping each section in a nice manner. But, I am unable to see any open issue. I would like that the authors provide some open issues and future directions in the field with a nice description or a small subsection about it.*
[Response] We have added open issues and future directions in the conclusion. Basically, the biggest general open issue is the inability to efficiently treat data in high dimensions. This is the case for all the algorithms,

but for different reasons: exact algorithms take too much time whereas approximate solutions exhibit a poor precision. Also, the replication is inherent to all methods and has a substantial cost, as seen in the experiment. Reducing the replication cost at minimum constitutes a possible future direction. Finally, the case of dynamic R is not addressed in the existing solutions.

Your sincerely,

Ge Song, Justine Rochas, Lea El Beze, Fabrice Huet, Frédéric Magoulès