Reviewer’s report

Title: Efficient Algorithms for Fast Integration on Large Data from Multiple Sources

Version: 1 Date: 15 February 2012

Reviewer: Kevin Livingston

Reviewer’s report:

(The formatting makes identifying paragraphs difficult as new paragraphs are not indented or otherwise specially separated. I did my best.)

Major:

1. Methods: is there a complete list of features used? The work cannot be replicated or validated otherwise.

2. In “the basic algorithm” how are thresholds identified? Much later you say you use a sample of the test data, but how do we know you are not over-fitting? Cross-fold validation would have provided confidence in this. There is discussion of this near the end of the Discussion section, however as described, the reader is left concerned that there is over-fitting.

3. Step 2 of the algorithm is “pairwise compare all records”, yet you spent the entire introduction setting up a straw man argument against that? That is the “traditional” method is it not?

4. There is no evaluation against the standard method of merging two sources at a time. The authors claim it is slower but given 3 sets of records: m, n, p. The authors method seems to start by making (m+n+p)^2 comparisons. Which doesn’t seem like it could be more efficient?

The traditional method should be compared to (and implemented if necessary) for complete understanding of the contribution.

5. The basic algorithm section references table 1, but without discussion as to how to interpret it. As is true of the other times the tables are referenced. The tables are extremely dense, and the work is difficult to impossible to evaluate completely without knowing how to interpret the tables.

6. The authors talk of the importance of sometimes switching the first and last name values, but then in section 4 they state that only the last name field is used. What is lost by this choice?

7. The results on simulated data only go to 10,000 records, but the introduction clearly states that the benchmark is millions of records. Comparisons to larger sets should be made. (Ideally to orders of magnitude larger than the benchmark so that it can be understood if the methods break down.)
8. Results on real data. What’s the rate at which the data used to construct the gold-standard results is missing? That is, what is the extent of the problem you are solving in the real world?

9. The authors suggest they have shown scaling at the end of results. However, the simulated data was only take to 3k records. “<1s” is a weird result to report, you don’t have a specific number? 999ms is <1s.

Minor:

1. The background starts off discussing the multi-source record-merging problem, but would benefit greatly from more concrete examples of why this is an issue.

2. Background paragraph 4 has a list of reasons the “traditional” method is deficient. The items in the list do not always seem to be of the same kind, and are poorly worded. Some are just asserted with absolutely no effort to back up their claim, e.g., 4, 5. Are there no existing methods for resolving 2?

3. Background p5, can the old method not incorporate new things with out recomputing?

4. Background section is long and rambling. The following is awkward; “been used to estimate to draw better decisions rules”.

5. End of background if you are going to mention distance metrics it would benefit from mentioning where they are discussed. This is a writing ‘quirk’ throughout, things are mentioned as if they won’t be discussed immediately and then in a few sentences they are discussed, but in the mean time the reader is left annoyed wondering when they will be discussed.

6. Improved algorithms: the algorithm doesn’t run out of memory, the machine does, and from your later notes it sounds like you are giving the JVM extremely small resource limitations, for undocumented reasons. If you say it runs our memory you need to document at that place what the cap was.

7. The time complexity of Algo. 3 is omitted.

8. The section “Faster Computation of the Edit Distance” is a complete waste of reader’s time. It should be replaced with the single sentence, “The methods used in [52] are inapplicable due to the short length of the strings involved.”

9. The authors define “accuracy” and “correctness”, these are, I believe non-standard names if not metrics. Precision and recall are more standard for evaluation of information retrieval types of methods.

10. Results on Simulated Data paragraph 3 says “dramatic improvement” what was the improvement and was it significant? What was the error rate?

11. Same paragraph, why is time a critical evaluation function for this task?
Presumably it’s done once in a huge batch?

12. End of results: best and worst are interesting, although median might be more informative, or median and worst.

Discretionary:

1. The use of the circled plus leaves me looking for a definition, which turns out to just be addition, why not just use the plus sign, and then say other methods of combination could be considered as future work. The circled plus just obscures the readability.

2. It is extremely difficult to interpret the data in the tables, plots would go a long way, and make the contribution easier to understand. (arguably this should be mandatory?)

3. End of Results on Real Data, I don’t understand the use of “necessary” and “sufficient” here, for definitions in logic systems these have precise meanings that seem to differ from how you use these words, they seem to be a distraction here. More clear, simple, English descriptions would help.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

no competing interests