Reviewer 1

Review Report Form

English language and style

( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors

The authors present a new data structure called multiset-trie, which is used to store and process the set of multisets. The manuscript extends the ideas of trie and capture the requirement of the operations that need to be handled in the collection of multisets. The novel part of the manuscript is mathematical modeling, mathematical analysis, and empirical evaluation. The manuscript reads well and is of interest to the algorithms community. I recommend acceptance.

Submission Date

10 December 2022

Date of this review

10 Jan 2023 00:01:43

Reviewer 2

Review Report Form

English language and style

( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	( )	(x)	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	( )	( )	(x)	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	( )	( )	(x)	( )

Comments and Suggestions for Authors

The paper proposes a data structure that stores multisets and can answer some natural queries on it. Each multiset is represented as a vector of natural numbers, and such vectors are stored in a trie. The queries are answered using naive search on the trie. The model is evaluated in a random scenario, in which each multiset is equally probable.

All in all, I think it would be difficult to improve this paper so that it would deliver results that deserve publishing.

I do not see any novelty or importance in those results.

The data structure is naive and offers no improvements over a naive approach.

The evaluated model is strange: an a-priori bound on the cardinality is artificial and not justified.

Uniform distribution over all possible multisets is completely unrealistic.

No particular application is given. In a fixed application one could discuss number of attributes, the distribution of values for an item etc.

Important implementation details are not described: for a fixed node the way the children are accessed matters. Using a table makes large overhead. Using a search tree adds time complexity. Using hash tables makes accessing the next node difficult. This is not addressed at all.

The experiments are run on randomly created data with uniform distribution. This gives little insight into real life performance; in fact probably could be calculated.

A natural approach to answer your queries is the range orthogonal queries on vectors of natural numbers. This problem in particular has well analyzed and implemented data structures, say k-d trees and others. They literally appear in textbooks. You should at least compare to them.

Submission Date

10 December 2022

Date of this review

16 Dec 2022 13:10:28

Reviewer 3

Review Report Form

English language and style

( ) English very difficult to understand/incomprehensible
(x) Extensive editing of English language and style required
( ) Moderate English changes required
( ) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	( )	(x)	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	( )	(x)	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	( )	(x)	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors

Some parts of the text are difficult to read. Try to clarify your definitions.

And Appendix of definitions would be very useful.

Also, in the experiments is necessary to clarify how is build the input data.

And tell us a little bit more about the C++ implementation and how the structure is coded.

And finally, regarding the performance, is it possible to compare it with another data structure commonly used?

Submission Date

10 December 2022

Date of this review

11 Jan 2023 17:14:35

Reviewer 4

Review Report Form

English language and style

( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I don't feel qualified to judge about the English language and style

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	( )	(x)	( )	( )
Are the methods adequately described?	( )	( )	(x)	( )
Are the results clearly presented?	( )	(x)	( )	( )
Are the conclusions supported by the results?	( )	(x)	( )	( )

Comments and Suggestions for Authors

The authors present a trie-based data structure for indexing a set of multisets, supporting, among common trie operations,

containment queries to report multisets that are supersets or subsets of the queried multiset.

The manuscript is well-structured, and language mistakes are few in numbers.

The topic and the results meet the scope of the journal Algorithms.

The data structure is well-explained, and evaluated under various angles.

Nevertheless, I am not satisfied with the mathematical presentation proving the expectancy of various characteristics of the proposed data structure.

Also, the experiments are lacking any comparison with existing solutions, which are only briefly sketched in one of the last sections of the manuscript.

I would think positively about accepting the paper if

(1) the authors can make the proofs in Section 4 better understandable, and

(2) augment the experiments in Section 5 with a thorough comparison with existing data structures.

Detailed comments follow, where numbers at the beginning are the line numbers printed on the right side of the manuscript pages.

4: "such as sub-multiset and super-multiset": these are not operations but objects.

8: What precisely are the time and space complexities of your data structure?

13: "a particular" : I think any element is allowed to have duplicates in a multiset, not just a particular element.

62/63: What are the merits for being height-balanced for this use case? Any trie that is "full" forms a complete n-ary tree.

73: What is the "time complexity space"?

74: empirical analyses, -> empirical, analyses

100: Multiset -> A multiset

104: What is n? Is n a user-defined constant?

Figure 1: I would write the stored sets as leaf-labels. Further, I see no advantage for writing c_j instead of just j.

156: For what is the variable dev used? Can you give an intuition behind it?

You described up to Section 3.3 standard trie operations, which you could put into the appendix since there is nothing new to see for those readers familiar with trie data structures.

190: tire -> trie

I think Section 3.5 is identical to Section 3.4 due to the symmetry of the problem.

254: What do you mean with a "quite" precise upper bound?

255: "appears" : Is this not clear by symmetry?

269: Since you need $\Omega(n\sigma)$ time to transform a given input multiset into your type Multiset representation, I am not convinced with your O(1) time.

279: The selection of the same letter M and $\mathcal{M}$ might be confusing. Maybe you can select a different letter for one of the two entities?

Lemma 1 : I think you should find Lemma 1 in literature for the standard trie data structure, which you can cite here.

Lemma 2: What is a "generating function"?

Page 10:

I could not follow the proofs since it seems that some notations have not been introduced in a sufficiently detailed way:

- What does $\tilde$ mean?

- What is $\mathcal{B}_0$?

- What is $Bernoulli()$?

- Referring to the book of Gardiner [27] to possibly omit crucial definitions and keep the proof shortly is in stark contrast to the elaborative description of standard trie operations in the previous section. A more detailed description in how the proof works is more than welcome.

- What is a probability generating function?

- What is 1^-?

- What is G'_X?

Definition 1: What are the intuitions behind α and β?

321: previous Definition 1 -> Definition 1

336: Here, the time bound of $O(|M|)$ is fine with me since this is also the number of elements we report, so we are output-sensitive for this query.

However, if we have a query like ${1^n, 2^n, ..., \sigma^0}$, then the time stays the same with fewer up to none elements to report, which is not optimal.

340: Can you give some informal conclusion on what the cumbersomely long equations for the time complexities mean?

It would be also good to make a distinction between expected and worst case scenarios. Currently, the analysis seems to mix both worlds.

345: on the -> the

399: the relation : with respect to what?

Figure 2: I have not understood what probability is addressed in this figure.

Figure 3: I would rename the y-axis to the ratio you compute.

398: Note that analyzing -> Analyzing

(15): What is p in this equation? It seems to be only used in the limit?

Figure 8 caption: Exsitence -> Existence

638: Why is your time complexity constant?

677-679: Can you explain in detail where the problem was and how it got solved?

680-691: Are you aware of the Leapfrog Triejoin algorithm, which seems to be closely related to the problem addressed here.

Todd L. Veldhuizen: Incremental Maintenance for Leapfrog Triejoin. CoRR abs/1303.5313 (2013)

695: The performance also depends on what type of queries you want to optimize.

707: How can the space complexity of your data structure be O(|M|) when storing a constant number of elements (|M| = O(1)) already costs you $\Omega(\sigma)$ nodes in the trie?

715-719: Specifying the implementation of your data structure as you did here would make sense to me if you would have practically evaluated the time/space of your data structure. It seems that such an experiment is missing, in particular with comparison to other approaches.

Finally, I could not compile the code (https://github.com/nick-ak96/mstrie) since it seems that some header files are not included such as <memory> needed for std::shared_ptr<>.

Maybe you can check whether you can make your code more portable?

g++ -DHAVE_CONFIG_H -I. -I.. -std=c++14 -Ilib -g -O2 -MT core/mstrie.o -MD -MP -MF $depbase.Tpo -c -o core/mstrie.o core/mstrie.cpp &&\

mv -f $depbase.Tpo $depbase.Po

In file included from core/mstrie.cpp:17:

core/mstrie.hpp:68:7: error: ‘shared_ptr’ in namespace ‘std’ does not name a template type

```
   68 |  std::shared_ptr<std::vector<std::shared_ptr<MstrieNode>>> mult_switch;
      |       ^~~~~~~~~~
```

core/mstrie.hpp:13:1: note: ‘std::shared_ptr’ is defined in header ‘<memory>’; did you forget to ‘#include <memory>’?

```
   12 | #include <queue>
  +++ |+#include <memory>
```

Submission Date

10 December 2022

Date of this review

16 Jan 2023 07:59:48