**Final Project Report**

Peng Chen 001098655

Fengyi Zhang 001564247

Shibo Lu 001548075

**Methodology**

The project aims to sort person names in Chinese with five sort algorithms, such as MSD radix sort, LSD radix sort, Dual-pivot Quicksort, Huskysort and Timsort.

Firstly, we wrote a Java class to read Chinese person names from a txt file, and stored the names as a String[].

Secondly, to sort Chinese names, we transferred Chinese words, a String[], to pinyins, a String[], by pinyin4j which is a Java package. In the meanwhile, we built a HashMap to save the relationship between the Chinese person names and the pinyins. The keys of HashMap are pinyins, and the values of HashMap are Chinese person names corresponding to the key, pinyin. Because of polyphone in Chinese, one pinyin may correspond to multiple Chinese names. Thus, we used String[] to store Chinese names as the value of HashMap, such as, HashMap = {“a1bin1”: [“阿斌”, “阿彬”]}.

Thirdly, we wrote MSD radix sort and Dual-pivot Quicksort methods, modified LSD radix sort method and directly referred Huskysort and Timsort. And then, we ran these sort algorithms with pinyins, a String[], as input. After the sorting step, we got sorted pinyins.

Finally, we can translate the sorted pinyins back to Chinese person names via the HashMap which we have gotten in second step. The Chinese person names are sorted.

We also wrote the first 1000 sorted Chinese name into five txt files.

**Implement**

When we were executing our methodology, we used pinyin without tone at first. However, there are many Chinese words share same pinyin, but different tones. So the pinyin without tone may cause inaccurate sort results. For example, without tone, both the pinyin of “阿琛” and the pinyin of “阿臣” are “achen”. With tone, the pinyin of “阿琛” is “a1chen1” and the pinyin of “阿臣” is “a1chen2”. Thus, we decided to transfer Chinese words to the pinyin with tone by pinyin4j, which made the accuracy of our sorting method get better.

**Result analysis**

According to analyzing the result of sorting, we found that the efficiency of single sorting algorithm is limited. By combining different sorting thoughts, and by having different sorting algorithms on applying to different data types could amplify the stability of sorting algorithms, and contribute to the algorithm efficiency as well.

In Husky sort, the fastest algorithm in our 5 different sorting algorithms, they reduce the number of expensive comparisons in the linearithmic phase as much as possible by using huskyCode to help compare the types of objects which are expensive to compare. There will be very few inversions (elements out of place) remaining after the dual-pivot quicksort phase. Then using timsort to make all the data sorted which is very useful to sort the part-sorted data.

In the Figure 1, we can see that the Huskysort extends the advantage both of dual-pivot quicksort and Timsort.

Finding the part which waste a lot of time and replace it by efficient way is a very useful way to improve the data sorting process.

Figure 1: Relative sort times for Huskysort, dual-pivot quicksort, system sort

表格

描述已自动生成

**Conclusion**

As the calculation time of different sorting methods, we found that times of all sorting methods increase linearly as the array length increases doubly. Husky sort uses least time to sort the array we passed in. Timsort, MSD radix sort, dual pivot quick sort get the 2nd, 3rd and 4th place of the sorting competition. LSD radix sort is the slowest sorting algorithm of these 5 algorithms set.

Unit tests:

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generatedGraphical user interface, text

Description automatically generated

Text

Description automatically generated

Banchmarks:

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated