An Easily Comprehendible Unicode based Sorting Algorithm for Bangla Words

Aamira Shabnam, Debakar Shamanta Piklu
Department of Computer Science and Engineering,
Shahjalal University of Science and Technology,
Sylhet-3114, Bangladesh.

ABSTRACT
This paper is based on an easily comprehendible algorithm to sort Bangla words in the linguistic order. Though a few works has been done on this project, only a few served the purpose. Since Unicode serves as a standard and a fixed representation for all characters of most of the language, Unicode representation of Bangla characters has been used to sort thousands of words in a linguistic order in this method. This method which requires a mapping can sort any Bangla word without keyboard dependency. This method is open for future modification so that changing the sorting algorithm to be used does not affect the whole algorithm.

Keywords:
Bengali Word Sorting, Bangla Word Sorting, Unicode Bangla Sorting, Bangla Dictionary Sorting

1. INTRODUCTION:
Since Bangla is one of the most rich and widely used language and about 10% of the total world's population speak in Bangla[1], its computerization and standardization such as Bangla Keyboard Layout, Bangla Character Recognition, Bangla Database System etc. has become a necessity. Sorting is a part and parcel for algorithmic analysis, Bangla data manipulation and for the development of Bangla Database System. For this reason, it is necessary to have an easy, efficient and versatile algorithm to sort Bangla words.

There are quite a few papers on this topic[2][3][4][5][6] but none of them can be said easy for non-professionals. In this paper, the analysis of the previous best algorithm is shown and a graph has been given to show the limitation of the proposed algorithm. In this paper, an algorithm based on Unicode which satisfies the dictionary standard to sort Bangla words given by Bangla Academy[7][8] has been proposed.

2. BANGLA ALPHABETS [9]
2.1 Base letters
There are 11 vowels (স্বরবর্ণ) and 39 consonants (ব্যঞ্জনবর্ণ) in Bangla alphabets called base letters.

2.2 Modifiers
2.2.1 Vowel modifiers:
There are 10 vowel modifiers known as কার.

Table 01 : Vowel Modifiers with example

Vowel modifiers	Example
া	কামার
ি	দিক
ী	নদী
ু	অসুবিধা
ূ	অনসূয়া
ৃ	সৃষ্টি
ে	আমাদের
ৈ	কৈ
ো	কোমর
ৌ	নৌকা

2.2.2 Consonant Modifiers:
There are about 6 consonant modifiers known as – ফলা. Some of them are given below,

Table 2 : Consonant Modifiers with example

Consonant Modifiers	Example
ব ফলা	স্বর
র ফলা	শুভ্র

2.3 Compound characters:
Two or more consonant characters of Bangla alphabet used together makes a compound character. There are about 270 compound characters in Bangla. Three of them are given below:

Table 03 : Compound Characters

Word	Compound Character	Decompressed Form
আনন্দ	ন্দ	ন+দ
যুক্ত	ক্ত	ক+ত
নষ্ট	ষ্ট	ষ+ট

3. SORTING DIFFICULTIES:

- Bangla words should be sorted according to the Bangla Academy Dictionary[10] order but Unicode representation of Bangla words does not follow this order. That is why, mapping is required.
- Compound characters should be considered accordingly.
- In computing, vowel modifiers must *follow* a character, not precede.
- Unicode characters ঢ়, ড়, য় and vowel modifiers ো , ৌ can be written in two ways either as a single character or a compound of two characters.

4. ANALYSIS ON PREVIOUS SOLUTION:

The previous solution was given by Md. Ruhul Amin, Asif Mohammed Samir, Madhusodan Chakraborty and Md. Mahfuzur rahman called "An Efficient Unicode Based Sorting Algorithm for Bengali Words ".The main idea of their process is given below in several steps,

- At the first stage of word processing, an extra dummy character is added after the base letter which has no modifier. If there is a vowel modifier, then it is placed after the base letter.
- If there is no vowel modifier at the end of the last letter then null modifier is not added.
- In considering a compound character, a link character (্) is added after the base character.
- Then with proper mapping according to Bangla Academy a string is generated with the mapped values. These mapped values are of two digits. Their mapping table is given below,

Table 04: Mapping (Previous Solution)

Unicode	Character	Value
09f9		01
09be	া	02
09bf	ি	03
.	.	.
.	.	.
.	.	.
0985	অ	13
0986	আ	14
.	.	.
.	.	.
.	.	.
09bc	়	63

- The mapped representations are sorted with an efficient algorithm.
- The Bangla words are then retrieved by reverse mapping process.
- The precedence of the Bangla characters:

  Dummy/Null character < Vowel Modifier < Consonant Modifier < Vowel < Consonant

4.1 Complexity Analysis:

Mapping and reverse Mapping is done in a linear time i.e. **O(N)** where N= Number of Words

If merge sort is used the total complexity becomes **O(NlgN).**

4.2 Limitations:

- Dummy characters can be avoided
- Reverse Mapping makes it more complex.

5. PROPOSED SOLUTION:

5.1 Steps:

- All words are copied in the rows of a two dimensional array.
- A single word is divided into characters accordingly and is represented by a string based on the map and written in the second column of the rows. Modifiers are followed by a character and while considering a compound character, a link character (্) is added after the base character.

  For example, বালক (ব + া + ল + ক) is represented by 5005625. যুক্ত (য + ু + ক + ্ +ত) is represented by 533251043.

  The array is sorted according to its second column which means on the basis of the mapped string in a lexicographic order.
- The precedence of a Bangla character maintained using the following rule :

  Vowel Modifier < Consonant Modifier < Vowel < Consonant

5.2 Mapping:

Table 05 : Mapping (Proposed Solution)

Unicode	Character	Value
09be	া	0
09bf	ি	1
09c0	ী	2
09c1	ু	3
09c2	ূ	4
09c3	ৃ	5
09c7	ে	6
09c8	ৈ	7
09cb	ো	8
09cc	ৌ	9
09cd	্	10
0985	অ	11
0986	আ	12
0987	ই	13
0988	ঈ	14
0989	উ	15
098a	ঊ	16
098b	ঋ	17
098f	এ	18
0990	ঐ	19
0993	ও	20
0994	ঔ	21
0982	ং	22
0983	ঃ	23
0981	ঁ	24
0995	ক	25
0996	খ	26
0997	গ	27
0998	ঘ	28
0999	ঙ	29
099a	চ	30
099b	ছ	31
099c	জ	32
099d	ঝ	33
099e	ঞ	34
099f	ট	35
09a0	ঠ	36
09a1	ড	37
09dc	ড়	38
09a2	ঢ	39
09dd	ঢ়	40
09a3	ণ	41
09ce	ৎ	42
09a4	ত	43
09a5	থ	44
09a6	দ	45
09a7	ধ	46
09a8	ন	47
09aa	প	48
09ab	ফ	49
09ac	ব	50
09ad	ভ	51
09ae	ম	52
09af	য	53
09df	য়	54
09b0	র	55
09b2	ল	56

09b6	শ	57
09b7	ষ	58
09b8	স	59
09b9	হ	60
09bc	়	61
09d7	ৗ	62

5.3 Algorithm

- N := Total number of words
- For i := 1 to N
- Arrayi1 = Wordsi
- For i := 1 to N
  Arrayi2 = For each word GetMappedString
- Sort Array according to the MappedString

5.4 Complexity

N = total number of words

K = the maximum length of a word

Copying words into a two dimensional array has been done in linear time which means **O(N).**

Mapping is done in a linear time which also means **O(N).**

Only where K>=N which is not realistic the complexity of mapping becomes **$O(N^2)$.**

Merge sort is used in sorting which forces the sorting complexity into **O(NlgN).**

Total Complexity

Best and Average Case: O(NlgN)

Worst Case: $O(N^2)$

Since the worst case is unrealistic, we can consider the complexity to be

O(NlgN) which is quite satisfying.

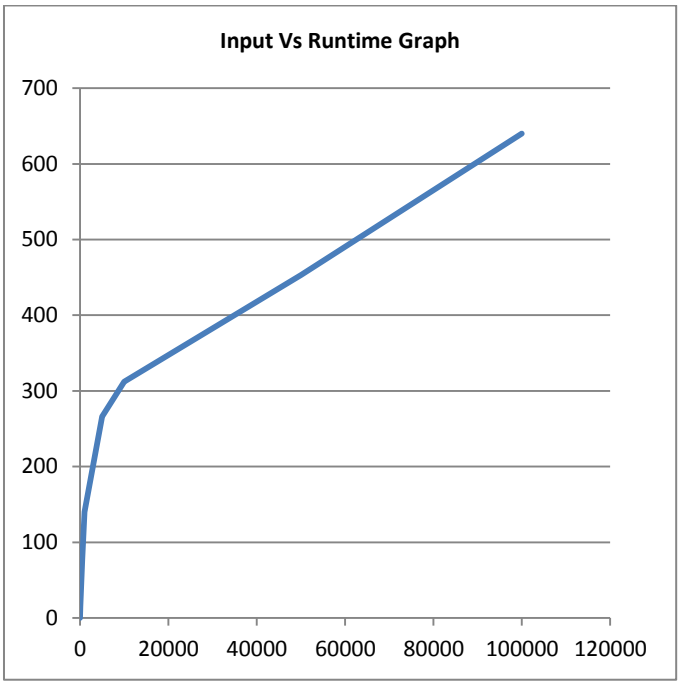
5.5 Input vs Runtime Computation:

Tested on: Platform Java SE 7

OS: Windows 8 64 bit

Processor: Intel core i5

Input	Runtime in ms
10	0
50	0
100	15
500	78
1000	140
5000	266
10000	312
50000	453
100000	640



5.6 Uniqueness:

- No use of dummy characters.
- No use of reverse mapping which simplifies getting back Unicode represented text in its given form.
- Using a two dimensional array for simplification.
- Simpler and easily comprehendible algorithm.

5.7 Limitations:

From the input versus runtime graph, we can see that the time required by this algorithm needs the more time as per the previous algorithms when the input size is low. But as the input grows larger, it works better.

6. CONCLUSION:

An easier and comprehendible solution has been proposed in this paper. It assures to sort Bangla strings with the proper linguistic order and proper structure of Bangla words. The main effort was given to sort by the standard set by Bangla Academy. More than 1lac random words from the "Rabindra Rachanabali"[11] site has been used for this purpose and checked if it maintains the dictionary order of sorting. This algorithm has the potential to be considered as the easiest standard procedure to sort Bangla strings based on Unicode characters. It has been made easier for any non-professional people with very little knowledge of coding. . Since it works

inefficiently for smaller set of strings, improvements are still in progress.

7. ACKNOWLEDGMENTS

Heartfelt gratitude to Mr. Debakar Shamanta Piklu (Associate Professor, Shahjalal University of Science and Technology, Sylhet-3114) for his support and supervision of the project. Special thanks to Aqib Ashef for his contribution to debug the code.

8. REFERENCES:

[1] http://en.wikipedia.org/wiki/Bengali_language retrieved 2013-08-31

[2] Rahman, Md. Shahidur and Iqbal, Md. Zafar , "Bangla Sorting Algorithm : A Linguistic Approach".

[3] Mafizul Haque Khan, S M Rafizul Haque, Md. Sharif Uddin, Rahat Khan, A B M Tariqul Islam, "An Efficient and Correct Bangla Sorting Algorithm ".

[4] Shah Md. Emrul Islam and Muhammad Masroor Ali, "An Approach to Sort Unicode bengali Text Using Ancillary Maps".

[5] Md. Ruhul Amin, Asif Mohammed Samir, Madhusodan Chakraborty, Md. Mahfuzur Rahman, "An Efficient Unicode based Sorting Algorithm for Bengali Words ".

[6] Minhaz Fahim Zibran, Arif Tanvir, Rajiullah Shammi, Md. Abdus Sattar, "Computer Represenation of Bangla Characters and Sorting of Bangla Words".

[7] http://en.wikipedia.org/wiki/Bangla_Academy

[8] http://www.banglaacademy.org.bd/

[9]http://en.wikipedia.org/wiki/Bengali_alphabet retrieved 2013-08-31

10] Bangla Academy Bengali-English Dictionary, First Edition June, 1994, Bangla Academy, Dhaka, Bangladesh.

[11] www.rabindra-rachanabali.nltr.org/

[12] Cormen, Thomas and Leiserson, Charles and Rivest, Ronald, "Introduction to Algorithm", Prentice – Hall of India Private Limited, 1999.

[13] The Unicode Standard 4.0, copyright 1991-2003, Unicode, Inc.

[14] Mohammad, Kazi Din, "Adhunik Bangla Byakoron o Rochona"

[15] Deitel and Assosiates, "Java How to Program".