An Improved Unicode Based Sorting Algorithm for Bengali Words

Partha Sarathi Kar[a], Shantanu Mandal[b], Labiba Jahan[c]
[a, b, c] Computer Science & Engineering
Metropolitan University
Sylhet, Bangladesh
partha@metrouni.edu.bd, shanto@metrouni.edu.bd, labiba@metrouni.edu.bd

Abstract. This paper describes an improved method of sorting Bengali words based on Unicode representation. There are different related works in this area but they are not efficient enough for a large dataset. Some works are based on ASCII representation where other are based on Unicode representation. We have chosen a Unicode based sorting algorithm for our work as Unicode produces a unique representation for each character of Bengali language. Here we have proposed and implemented an improved version of sorting algorithm for Bengali words where mapping is used to simplify the sorting procedure. We have improved the mapping system of the algorithm which is convenient for memory optimization. Moreover, the improved version of the algorithm is more efficient than the previous one.

Keywords: Word Sorting, Unicode representation, Modifier, Mapping.

1 INTRODUCTION

Bengali is the native language of Bangladesh and the Indian state of West Bengal. It is written using Bengali alphabets according to Bengali Grammar (Bengali Language on Wikipedia). Bengali is the national language in Bangladesh and second most spoken language in India (Olivet Nazarene University 20 April 2015). With about 250 million native and about 300 million total speakers worldwide, it is the seventh most spoken language in the world by total number of native speakers and the eleventh most spoken language by total number of speakers (Ethnologue 2012).

The process of sorting words into various categories plays an important role of word study. Young children make sense of words and patterns within words by categorizing different words. Word sorts combine both constructivist learning and teacher-directed instruction (Bear, D. et al. 2008). Word sorts place instructional emphasis on the exploration of patterns that can be detected in the sound, structure, and meaning features of words. It also contributes to orthographic or spelling knowledge, the engine that drives efficient reading as well as efficient writing (Templeton, P et al. January 1999).

As the Bengali language is a rich and widely used language, it must have some standardization such as Bengali keyboard layout, Bengali character recognition, voice synthesis etc. (Md Ruhul Amin et al. June 2011). Research works related Bengali language is improving day by day but a significant research field like Bengali word sorting is not that much standard till now. There are some works in this area but none of them is suitable for large database. In this paper, we focused on improving a Unicode based Bengali word sorting algorithm using mapping.

2 BENGALI ALPHABETS AND MODIFIERS

In the written form of Bengali alphabets, there are 11 vowels, 39 consonants and 11 vowel modifiers. There are also some consonant modifiers and joint letters in Bengali language.

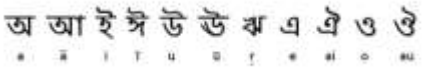

Fig. 1. Vowels

Fig. 2. Consonants

Fig. 3. Vowel modifiers

Word	Consonant Modifier	Example
র	র-ফলা	ক্র /ʤr/
য	য-ফলা	ন্য /ʤonno/
র	র-ফলা	ব্রি /tibro/

Fig. 4. Consonant modifiers

Fig. 5. Joint letters in Bengali Language

3 LITERATURE REVIEW

M. Shahidur Rahman et al. (Rahman, Md. Shahidur et al. 1998) have proposed an algorithm based on ASCII representation. According to their proposal a dummy character is placed after

the character, which does not have any modifier and there would be no dummy character between the constituent parts of a compound character. In this algorithm the modifiers are shifted after the character for the internal representation. In case of compound characters, they are decomposed into their constituent components and stored accordingly. This algorithm follow the following order to sort each word -Null modifier < Vowel Modifiers < Vowels < Consonants. As example the internal representation of "কুসুম" is "ক ু স ু ম".

According to Mafizul Haque Khan et al.'s "An Efficient and Correct Bangla Sorting Algorithm" (Mafizul Haque Khan et al. 2004) a character is represented with two digit unique number for every letter of Bengali alphabet along with the vowel modifiers and the consonant modifiers. For example the word "সা" is changed into number "6171". The consonant modifiers are having the same number as their original consonants.

Shah Md. Emrul Islam et al. proposed a method to Sort Unicode Bengali Text Using Ancillary Maps (Shah Md. Emrul Islam et al. Buet, Dhaka). In this method, the Unicode characters are mapped and given a Sort Weight. For each word the mapped value are concatenated and a decimal point is added after two digits from the starting. Then it becomes a floating point number. By comparing all the floating point numbers, the list of words is sorted. For example, the word "কানকো" gets the value 25.0346002519.

Md. Ruhul Amin (Md Ruhul Amin et al. June 2011) et al. proposed a method where they mapped all the characters that are used in Bengali text according to Bangla Academy (Bangla Academy Bengali-English Dictionary 1994) sequence. In this method, an extra dummy character is added after the base letter which has no modifier. But if there is a vowel modifier with the base letter then that modifier is added in place of the dummy character. If there is no vowel modifier at the end of the last letter then the null modifier is not added. When we consider a compound character, then we get the base letter, a link character, and then the next character of the compound character and so on. In the next step, a string of digits is generated for each Bengali string using a map table to obtain a secondary representation. Then they sort these strings using MergeSort or any other efficient sorting algorithm using string comparison. Finally, the secondary representations is converted to its original Bengali strings. As example the internal representation of "আমার" is "আ ০ ম া র". We chose this method for further improvement because it is one of the efficient methods till now.

4 ANALYSIS ON PREVIOUS SOLUTION

We selected a Unicode based method for future analysis which is proposed by Md. Ruhul Amin et al. Here mapping is needed according a map table (Table 1) to sort words according to Bangla Academy standard because Unicode for Bengali characters are not in Bangla Academy dictionary order. The rules followed in this approach are-

- Any character without any modifier is considered as character followed by null modifier.
- Any character with vowel modifier is considered as character followed by vowel modifier.
- Any character with consonant modifier is considered as character followed by link character followed by consonant.
- Any compound character is considered as character followed by link character followed by character.

In this algorithm the precedence of the Bengali character is maintained using the following rule:

Dummy/Null Character < Vowel Modifier < Consonant Modifier < Vowel < Consonant

Table 1. Map value for each character

Character	Value
◌	01
◌া	02
◌ি	03
◌ী	04
◌ু	05
◌ূ	06
◌ৃ	07
◌ে	08
◌ৈ	09
◌ো	10
◌ৌ	11
◌্	12
অ	13
আ	14
ই	15
ঈ	16
উ	17
ঊ	18
ঋ	19
এ	20
ঐ	21
ও	22
ঔ	23
◌ং	24
◌ঃ	25
◌ঁ	26
ক	27
খ	28
গ	29
ঘ	30
ঙ	31
চ	32
ছ	33
জ	34
ঝ	35
ঞ	36
ট	37
ঠ	38
ড	39
ড়	40
ঢ	41
ঢ়	42
ণ	43
ৎ	44
ত	45
থ	46
দ	47
ধ	48
ন	49
প	50
ফ	51
ব	52
ভ	53
ম	54
য	55
য়	56
র	57
ল	58
শ	59
ষ	60
স	61
হ	62
◌়	63

Table 2. Internal representation for words

Word	Internal Representation	Representation with mapped value
আমার	আ ◌ম ◌া র	1401540257
কলম	ক ◌ল ◌ম	2701580154
আম	আ ◌ম	140154
ধন্য	ধ ◌ন ◌্ য	4801491255

5 PROPOSED METHOD

In our proposed solution, we mapped each character and their modifier together. Even we mapped all joint letter. As example we mapped unique value for each string like গ, ড, কা (ক + া), কি (ক + ি), ক্ক (ক + ্ + ক), গু (গ + ু + ড) etc. Then we generated a string of digits for each word according to their mapped value and apply ant sorting algorithm to sort words. Finally we can get all words sorted by converting digital representation to its Bengali form. There is no need to use a dummy character as we mapped each character including base letter, letter with modifier and joint letter using a map table (Table 3). In this algorithm we serialized all string according to the following order:
Base letter < Base Letter with Vowel Modifier < Base Letter with Consonant Modifier + Joint letter (according to order of each character)

Bengali String	Value
অ	1
আ	2
ই	3
ঈ	4
উ	5
ঊ	6
ঋ	7
এ	8
ঐ	9
ও	10
ঔ	11
ং	12
ঃ	13
ঁ	14
ক	15
কা	16
কি	17
কী	18
কু	19
কূ	20
কৃ	21
কে	22
কৈ	23
কো	24
কৌ	25
List of All joint letters and consonant modifiers starting with 'ক' according to their order →	
ক্ক	26
ক্ষ	27
ক্য	28
.	.
.	.
.	.
খ	.
খা	.
খি	.
খী	.
.	.
.	.
.	.
Continued till strings related 'হ'	.
	.

6 OUTCOMES

We have implemented our proposal method and compared with the previous one. We have found that the previous system needs more memory than ours. As example, the word "আমাদের" needs to map 6 characters (আ, ম, া, দ, ে, র) according to previous system where we need to map only 4 (আ, মা, দে, র) characters. Moreover, we need not map any dummy character or link character to find out base letter and joint letter. We easily did that by forward checking and generated digital sequence according to map table having value for each possible string in

Bengali language. We used Merge sort algorithm to sort words and found satisfactory outcomes.

7 CONCLUSION

The method we proposed is an improved version of a previous system. We enhanced the mapping procedure which helped to make the system more easy and efficient. Moreover, we are successful to reduce memory for a large database. We also produced a simple structure of the whole algorithm by mapping each possible string in Bengali language. Though further research is needed in this field to make the process more efficient but till now we can consider it as a standard one. For large dataset the overall complexity of our proposed system is satisfactory.

References

https://en.wikipedia.org/wiki/Bengali_language

"*The Second Most Spoken Languages around the World*". Olivet Nazarene University. Retrieved 20 April 2015.

"*Statistical Summaries*". Ethnologue. 2012. Retrieved 2012-05-23.

Bear, D., Invernizzi, M., Templeton, S., Johnston, F. *Words Their Way: Word Study for Phonics, Vocabulary, and Spelling Instruction*. 4th ed. Upper Saddle River, NJ: Prentice Hall, 2008.

Templeton, S, & Morris, D (1999, January). *Questions teachers ask about spelling*. Reading Research Quarterly, 34(1), 102-112.

Md. Ruhul Amin, Asif Mohammed Samir, Madhusodan Chakraborty, Md. Mahfuzur Rahaman, *An Efficient Unicode based Sorting Algorithm for Bengali Words*, International Journal of Computer Applications (0975 – 8887),Volume 24– No.7, June 2011

Rahman, Md. Shahidur and Iqbal, Md. Zafar, "*Bangla Sorting Algorithm: A Linguistic Approach*". Proceedings of International Conference on Computer and Information Technology, Dhaka, 18-20 December 1998, pp. 204-208.

Mafizul Haque Khan, S M Rafizul Haque, Md. Sharif Uddin, Rahat Khan, A B M Tariqul Islam, "*An Efficient And Correct Bangla Sorting Algorithm*" 7th ICCIT, 2004, Page 125

Shah Md. Emrul Islam and Muhammad Masroor Ali "*An Approach to Sort Unicode Bengali Text Using Ancillary Maps*", BUET, Dhaka.

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

Partha Sarathi Kar is a Lecturer in Metropolitan University, Sylhet, Bangladesh. He received his BS in Computer Science & Engineering from Sylhet Engineering College, Sylhet, Bangladesh.

Shantanu Mandal is a Lecturer in Metropolitan University, Sylhet, Bangladesh. He received his BS in Computer Science & Engineering from Shahjalal University of Science & Technology, Sylhet, Bangladesh.

Labiba Jahan is a Lecturer in Metropolitan University, Sylhet, Bangladesh. She received her BS in Computer Science & Engineering from Shahjalal University of Science & Technology, Sylhet, Bangladesh.