Unicode Searching Algorithm Using Multilevel Binary Tree Applied on Bangla Unicode

Md. Akhtaruzzaman
Department of Computer Science and Engineering, International Islamic University Chittagong Dhaka Campus
akhter.mail@gmail.com, akhter900@yahoo.co.uk

***Abstract*- Unicode Searching Algorithm using multilevel binary tree is proposed to search the Unicode in efficient way. The algorithm is applied on Bangla Unicode searching to convert Bijoy string into Unicode string. Firs, the algorithm build a multilevel binary tree based on ACII code with its corresponding Unicode. The tree is build from a multilevel binary sorted data containing ASCII code and its corresponding Unicode. The data must be sorted based on ASCII code. The algorithm takes Bangla Bijoy string as input value and output the same string in Unicode format. The input Bijoy string must be in Unicode readable format**

I. INTRODUCTION

Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. These encoding systems also conflict with one another. That is, two encodings can use the same number for two different characters, or use different numbers for the same character. Any given computer (especially servers) needs to support many different encodings; yet whenever data is passed between different encodings or platforms, that data always runs the risk of corruption. Unicode is changing these all. Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.

Bangla is the mother tongue of Bangladeshi people and the second most popular language in India. It is a reach language and more than 10% people speak in Bangla in the world. But in field of Computer Science the research on this language is not good enough. There exists some Bangla writing software like Bijoy, Avro, Akkhor etc. Bijoy is the most popular and oldest software which is used to write the plain test only. Avro uses Unicode to write Bangla sentences. Now a day Unicode format is used to write any language and sometimes it is necessary to convert the plain text into Unicode format. So searching Unicode is necessary. In Unicode, there are 65,535 distinct characters that cover all modern languages of the world. So Unicode searching must be efficient and reliable for all languages. In some Unicode searching algorithm, especially in Bangla, there exist only 'if – else' condition. Some one uses the 'Hash Table' to develop Unicode searching method. Here a Multi Level Tree based Unicode searching algorithm is proposed which will be more efficient for Unicode searching on different languages.

II. PRELIMINARY STUDIES

A. Bijoy String to Unicode Readable Format String

There exist 11 independent characters (vowel) and 39 dependent characters (consonant) [6] in Bangla literature. There also exist some independent and dependent character symbols called 'Kar' and 'Fala' respectively. These symbols must be used with a character. A large number of Complex Characters (combination of two of more characters) exist in Bangle language. A single Complex Character may contain two or more independent or dependent characters, must joined with a symbol named 'Hasanta' (◌্).

In plain text all the characters and symbols may placed independently anywhere in a sentence. Bijoy follows this rule. But Unicode maintains a unique format to use the symbols with a character. In Unicode the symbols must be used after a character with no gap between them i.e. "character + symbol". But in some cases like 'Chandrabindu' or 'Ref' the placement is different. 'Chandrabindu' must be used after a character if no symbols are exists with that character i.e. "character + Chandrabindu". If any symbol is exist with a character then the 'Chandrabindu' is used after the symbol i.e. "character + symbol + Chandrabindu". 'Ref' must be placed before character.

Figure 1 shows some examples, representing Bijoy plain text and its Unicode readable format of Bangla sentences. Figure 1(a) contains a complex character 'ম্ম' that forms with two ম s, Bangla dependent characters (consonant), i.e. ম + ◌্ + ম = ম্ম .

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 321–326.

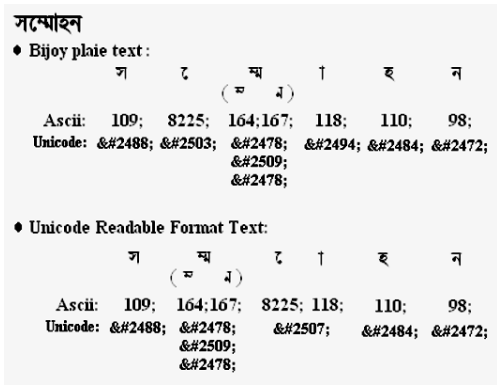

Fig. 1 (a)

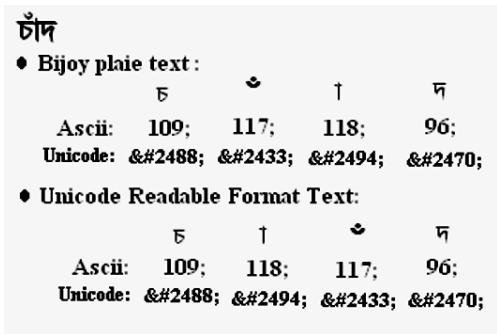

Fig. 1 (b)

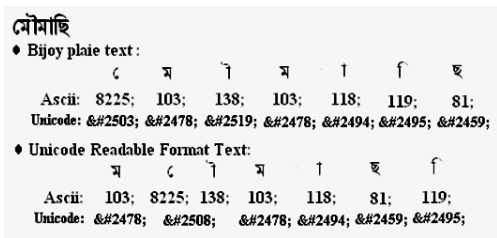

Fig. 1 (c)

Fig. 1. Bijoy plain text and its Unicode readable formatted text showing ASCII and Unicode of corresponding character.

B. Multi Level Binary Sort

Binary search is a well known and efficient search algorithm in the field of Computer Science. To apply this algorithm on a list it must be sorted in increasing or decreasing order. There exist a lot of sorting algorithms which are used to sort data in non-increasing or non-decreasing order. Here the term 'Binary Sort' is used because the method follows the technique of Binary Search algorithm and rearrange the data into a in order tree format where its first value indicates the middle value of the data list, second value indicates the middle value of the first dividing part of the previous data list and so on.

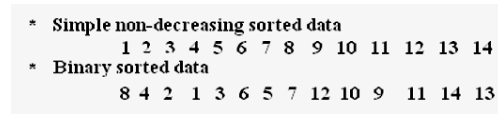

Fig. 2. Binary Sorted data generated from a simple non-decreasing sorted list.

Figure 2 shows the Binary Sorted data generated from a non-decreasing sorted list. Here in Binary sorted data the first value '8' indicates the middle value of Simple sorted data. Second value '4' is the middle value of the first division (1 to 7) and third value '2' is the middle value of the first division (1 to 3) of the previous first division of Simple sorted data.

Here another term 'Multi Level' is used with the term 'Binary Sort' because this method is applied on multilevel sorted data. Figure 3 shows the Multi Level Binary Sorted Data formation.

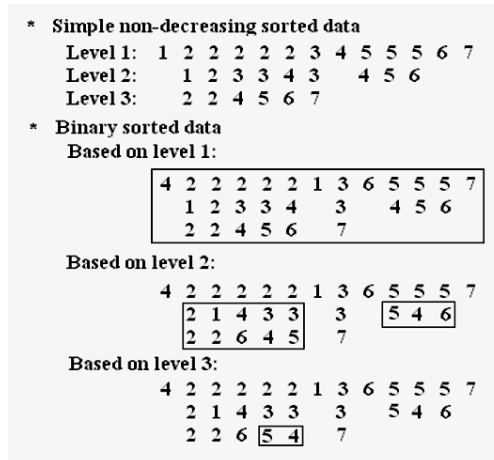

Fig. 3. Three level Binary Sorted data generation from three level non-decreasing sorted data.

Here in the first level of the Simple non-decreasing data has 5 twos and 3 fives which are treated as single integer each. So applying Binary Sort on the basis of first level the sequence of 2 and 5 will remain unchanged. Now come to the second level. Here exist 2 threes for common value '2' of the first level and is treated as single three. Applying Binary Sort on the values of common two and five of first level the values will be rearranged where the sequence of three of second level will remain unchanged. In the third level the sort will be applied on the values of the common three of second level. At the end of the process the three level non-decreasing data will become as the last format shown on the figure 3.