BFM: a forward backward string matching algorithm with improved shifting for information retrieval

ASSIGNMENT1:

Abstract:

In recent years, the use of a string matching algorithm in text mining has grown in popularity. While searching, these algorithms must perform fewer character comparisons and pattern changes. In this article, we propose a new algorithm called forward and backward. BFM is a faster algorithm that matches patterns that can be seen from both the forward and backward directions.

Introduction:

- A fast pattern matching algorithm is an essential component of page ranking in search engines and digital libraries, as well as checking syntax and spelling errors, detecting network breaches, and a variety of other applications. It is also required for bioinformatics, DNA sequence matching, and behavioural analysis.
- String matching algorithms are divided into two phases: preprocessing and searching . In general, the pattern is preprocessed that determines where the pattern needs to be shifted in order to find a mismatch. Then searching phase, comparisons between pattern and text characters are made from right to left, left to right, or in specific ways to find all occurrences of exact pattern match.
- In this article we present BFM, a new string matching algorithm for finding all occurrences of exact patterns or a string in a text.
- The goals of these algorithms are to reduce the number of character comparisons and increase the duration of changes and minimizing the total number of shifts.
- In contrast to other algorithms, this algorithm has a lot less character comparisons and a lot more change lengths.

ASSIGNMENT 2

Related Work:

1. The Boyer Moore algorithm:

- used pattern matching algorithm
- determines the maximum shifts of the pattern in case of mismatch
- based on two rules. The good suffix and the bad character
- produce two different arrays known as the preprocessing table
- Boyer Moore determines best of the two rules

2. The Quick Search algorithm:

- To search small pattern in large character set

3. Quick-Skip Search algorithm:

- combination of Quick Search and the Skip Search algorithm

4. The Knuth–Morris– Pratt (KMP) algorithm:

- works like a naive algorithm, but uses the degenerative property of the searched string
- instead of checking all the characters after each shift, construct a table that helps skipping comparisons of those pattern characters that had already matched

5. The Rabin Carp string searching algorithm :

- search both single and multiple patterns in a string

ASSIGNMRNT3

Methodology :

- The algorithm compares characters from both left and right side in each attempt while searching
- The first step: pre-processing table is prepared that stores the positions in the text, where the first and last characters of the pattern matches with the text.
- Next step: search a pattern, the algorithm checks for the pattern only at the positions stored in the preprocessing table.
- BFM algorithm finds out all the possible positions in the text, where the first and last characters of the pattern are matched with the text.

Result analysis:

- Along with our algorithm BFM, we evaluated both BMH and KMP algorithms to compare our algorithm's performance on both files.
- The number of shifts, comparison and execution time of our algorithm BFM are very much lower than BMH and KMP.

Conclusion:

- BFM algorithm is pre-processing a text by applying forward and backward matching to match both the first and last characters of the pattern.
- This preprocessing task enhances the searching by enabling the algorithm to search only on the indexed positions
- Decreasing both the number of shifting and character comparisons.
- If the text and pattern both are small, then BFM's performance degrades, as there is a preprocessing phase to complete before searching.

- This algorithm would certainly an efficient choice in case of the tasks that require huge amount of text searches,
- We found the performance of BFM in respect to execution time, number of shifts and comparisons to be exceptionally high in contrast to KMP and BMH.