Mining Exact Information from Overlapped Pattern

Sanil Shanker KP

Department of Computer Science, Farook College
Kozhikode, India

Abstract: This paper explains a technique to mine exact information from overlapped pattern. The technique is based on the concept of Logical Matching Strategy. The paper illustrates the method in detail.

Keywords: String Matching, Logical Matching Strategy Overlapped Pattern.

I. INTRODUCTION
Symbols or characters are considered to be the basic elemental building blocks in a string. A string is a sequence of instances of symbols or characters over a finite alphabet Σ. An alphabet is a specific set of symbols and is usually a finite set. For instance, Σ = {a, b, c, d, e} is an alphabet containing symbols a, b, c, d and e. The length of the string S is the number of instances of the symbols or number of characters in the string. The string S may be represented as S = s₁s₂s₃…….sₘ, where sᵢ is an instance of a symbol, or character and m is the length of the string S and is represented as |S|.

Consider a pattern P = p₁p₂…pₘ of length m and a text T = t₁t₂…tₙ of length n are two strings formed over the same finite alphabet Σ such that m < n. Here the pattern p occurs in the text T at the beginning of the text location q if 1 ≤ q ≤ (n-m) and tᵢ₋₁ = pᵢ for 1 ≤ i ≤ m. The string matching problem is to find all the occurrences of pattern in the text locations [1].

For example, let’s consider a string T = M A T H E M A T I C S and a pattern P = M A T over the finite alphabet Σ = {M, A, T, H, E, I, C, S} as shown.

Σ = {M, A, T, H, E, I, C, S}

Pattern (p): M A T

The pattern ‘M A T’ occurs in the text locations 1 and 6. Sometimes characters/symbols in the pattern may overlap.

II. EXACT INFORMATION FROM OVERLAPPED PATTERN

Consider the string t = A M M A M A M M A M A and the pattern p = A M A over a finite alphabet Σ, where Σ = {A, M}.

Pattern (p): A M M A

Position: 1 2 3 4 5 6 7 8 9 10

Text (t): A M M A M A M M A M A

A M M A

Match at 1

A M M A

Match at 7

A M M A

Match at 4

Mining exact information from overlapped pattern is a difficult task. It is possible to solve this problem by using Logical Matching Strategy [1, 2].

III. MINING OVERLAPPED PATTERN USING LOGICAL MATCHING STRATEGY

Let the Text and Pattern be AMMAMMAMMA and AMMA respectively. Shift the Text and Pattern from left to right.

Phase I. Generate indices of Text and Pattern

Text => <A (1, 4, 7, 10);

Position: 1 2 3 4 5 6 7 8 9 10 11

Text (t): M A T H E M A T I C S

↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

M A T M A T

Match Match

ISSN: 2231-2803 http://www.ijcttjournal.org Page 103
M (2, 3, 5, 6, 8, 9) >

Pattern => < A (1, 4);
M (2, 3) >

Phase II. Match the indices of Pattern with the indices of Text. Select the lowest index value alphabet in the Pattern and then the indices corresponding to the same alphabet in the Text have to arrange in a row.

That is,

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>1</th>
<th>4</th>
<th>7</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>M</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

The Pattern AMMA repeats in the Text AMMAMMAMMA in the locations of (1, 2, 3, 4), (4, 5, 6, 7), (7, 8, 9, 10).

IV. SUMMARY

The paper presents the solution to the problem of mining exact information from overlapped pattern. The mining technique explained is based on the concept of Logical Matching Strategy. This technique finds application in data mining to locate overlapped pattern.

REFERENCES