Paper ID: CSEIT 37

STUDY OF DIFFERENT PATTERN MATCHING ALGORITHMS AND IT'S APPLICATIONS

Miss. Ketaki H. Pangu
Department of Computer Science and Engineering Rajarambapu Institute of Technology
Rajaramnagar, Sakhrale, Sangli
Mr. Rahul R. Date
Department of Computer Science and Engineering Rajarambapu Institute of Technology
Rajaramnagar, Sakhrale, Sangli

Abstract— Pattern matching algorithms are used to find one or more patterns or sequence of patterns from huge data or symbols. Pattern matching algorithms works for many purpose in engineering as well as in bioinformatics. Mostly pattern matching algorithms are used for information retrieval, intrusion detection, data compression,content filtering, bioinformatics for analyzing DNA/RNA and other genome sequences.In this paper different pattern matching algorithms and their applications are discussed.

I. INTRODUCTION

Pattern matching algorithms are one of the important areas in computer science field. Pattern matching are used to find out different characters or sentences or set of symbols from large datasets . Traditional pattern matching algorithms and enhanced pattern matching algorithms are mainly catagorised for different technical as well as non-technical applications. If we want to apply pattern matching algorithms then it is based on the applications, whether it is approximate or exact pattern matching algorithms.[1] Depending upon applications where we want to use pattern matching algorithms that are again caractorised by single string pattern matching and multistring pattern matching algorithms. The main objective of the pattern matching algorithm is to reduce the execution time as well as increase memory efficiency by reducing number of comparisons between characters and given texts. In section II different pattern matching algorithms found in literature are discussed. Section III is about different problems solved using pattern matching algorithms. Finally in section IV we have concluded and some future research directions given.

II. PATTERN MATCHING ALGORITHMS

In [2] Ziad A.A. Alqadi1et al have proposed MSMPMA i.e. Multiple Skip Multiple Pattern Matching Algorithm that is multi string pattern

Matching algorithm which is used in different applications where search as well as matching is required. It also used in applications where indexing and sorting of data is required and it works in different datasets of medium sizes. In this study the proposed algorithm is works on unlimited file size. The time complexity of the algorithm is O (n*m), but because of the skips it moves to around O (n).

In [3] Hossein Gharaee et al have proposed a survey of pattern matching algorithm where Brute Force (BF) algorithm is traditionally used for the intrusion detection system for network security purpose. In which the list of the keywords are given in which the particular keyword want to found for tracking the malicious activities. But now a days this algorithm is not used in practice due to low speed of pattern preceding in NIDS. If there are n elements through the set, time of implementing this algorithm for matching operation is of linear position and equal to O(n).

In [4] Alexander Gee have proposed Naïve string search Algorithm where it is used for single pattern search mainly used to find a string or a phrase from the text and mainly used in internet related applications.it is linear search algorithm in which time of algorithm implementation in the linear position and equal to O(m+n) where m is the length of the text and n is the size of the phrase to be searched. The worst case time complexity is O(mn).

In [5] Knuth- Moris –Pratt string matching algorithm is worked on both short and long strings and determine the pattern appears somewhere in a text. It is linear time string matching algorithm and it keeps the information which is wasted during Naïve algorithm work. After avoiding the wasted information it achives a running time of O (n+m) which is optimal in worst case. So, in the worst case KMP algorithm examines all characters in the text and pattern at least one time. The running time of KMP algorithm is directly proportional to the time needed to read the characters in the patterns.(kent.edu)

In [6] Boyer- Moore is single pattern matching algorithm where it performs the comparisons from left to right between texts. It is most efficient algorithm in string matching. The algorithm scans the the characters in right to left beginning with rightmost one. At the time of mismatch it uses the functions like good-suffixshift and bad suffix shift. In the processing phase th time and space complexity is O(m+n) and searching phase it is O(mn). The best performance of the algorithm is O(n/m).(igm)

In paper [7] Aho-Corasick is exact multiple string matching algorithms gererally used for bioinformatics where large amount of data with various sequences need to analyze simultaneously. In th algorithm it locates the occurences of any pattern of a set P= {p1, p2…,pk}in the target T[1…..m]. It works on three main functions i.e. goto function that is used to construct and match the patterns with tri like structure of automata. The failure function is called when the pattern is not matched and it call goto again and output function is used to give the output when exact pattern is found. It works in time O (n+m+z) where z is number of pattern occurences in text T.

In paper [8] Parallel Failureless-AC Algorithm has a multiple pattern match structure and it is an advanced version of AC algorithm having less time complexity. The parallel structure here are used which gives three times as fast as older algorithm.This algorithm has a multiple pattern match structure and is an advanced version of AC algorithm with less time complexity. Also short memory stage should be used is one of the advantages and AC algorithm running time is O (n) + O (m+k) where the complexity of PFAC is 3 times faster that is O (n) + O (m+k/3) .

In [9] Karp-rabin algorithm are discussed which is used on hash functions. The use of hash function is to reduce quadratic number of character comparisons so, instead of checking at each character it checks the patterns only it looks like the pattern which we want to search. The hash function is used to resemblance between two words. The time complexity of preprocessing phase is O (m) with constatnt space complexity and searching phase it is O (mn). The expected running time complexity is O (n+m).

In [10] KMP skip search algorithm is single search algorithm and it is used linear using the two shift tables of Morris-Pratt and Knuth-Morris-Pratt. It is the improvement in the skip search algorithm andit uses buckets of positions for each character of the alphabet for comparison purpose. For the preprocessing phase the time and space complexity is O (m+n) and for searching phase it is O (n).

Following are some rarely used pattern matching algorithms[11]:

1. Deterministic Finite Automaton.
2. Shift Or algorithm.
3. Simon algorithm.
4. Colussi algorithm.
5. Galil-Giancarlo algorithm.
6. Apostolico-Crochemore algorithm.
7. Turbo BM algorithm.
8. Smith algorithm.
9. Quick Search algorithm.
10. Reverse Factor algorithm.
11. Turbo Reverse Factor algorithm.
12. Two Way algorithm.
13. Alpha Skip Search algorithm.
14. Raita Algorithm.
15. Forward Dawg Matching algorithm.

III PATTERN MATCHING ALGORITHM APPLICATIONS

Pattern matching algorithms are having a various applications in technical as well as different real time fields. There are different applications like intrusion detection, information security, data mining, bioinformatics fields to analyze DNA/RNA as well as protine sequences, dictionary searching etc. The following table shows algorithms and applications of that algorithm

Sr. no.	Algorithm Name	Applications
1	MSMPA	Intrusion detection, DNA string search
2	Broute-Force	Network reconfiguration, Online/offline signature forgery test
3	Naïve-String search	Keyword search in large datasets, network security
4	Knuth Moris pratt	High speed multi stream packet scanning
5	Boyer Moore	Medical sequence analysis, Dictionary search
6	Aho-Corasick	Information Security, DNA/RNA Sequence analysis
7	Parallel Failureless AC	Protine sequence analysis
8	Karp Rabin	Higher dimentional pattern matching problems
9	KMP-SKIP	Bioinformatics

So, the above table describe the different pattern matching algorithms with its applications in various fields.

IV SUMMARY

Pattern matching algorithms are very useful in different technical as well as bioinformatc purpose. Approximate, exact, single as well as multiple different pattern matching algorithms are used as per their area of applications. So, the survey above give detail of 9 different pattern matching algorithms with their applications and the list of rarely used algorithms.

REFERENCES

[1] Bhukya, Raju, and D. V. L. N. Somayajulu. "Exact multiple pattern matching algorithm using DNA sequence and pattern pair." International Journal of Computer Applications 17.8 (2011): 32-38.

[2] Alqadi, Ziad AA, Musbah Aqel, and Ibrahiem MM El Emary. "Multiple skip Multiple pattern matching algorithm (MSMPMA)." IAENG International Journal of Computer Science 34.2 (2007): 14-20.

[3] Zoebisch, Frank, and Claus Vielhauer. "A test tool to support brute-force online and offline signature forgery tests on mobile devices." Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on. Vol. 3. IEEE, 2003.

[4] haraee, Hossein, Shokoufeh Seifi, and Nima Monsefan. "A survey of pattern matching algorithm in intrusion detection system." Telecommunications (IST), 2014 7th International Symposium on. IEEE, 2014.

[6] Galil, Zvi. "On improving the worst case running time of the Boyer-Moore string matching algorithm." Communications of the ACM 22.9 (1979): 505-508.

[7] Lin, Cheng-Hung, et al. "Accelerating pattern matching using a novel parallel algorithm on gpus." Computers, IEEE Transactions on 62.10 (2013): 1906-1916

[8] Zha, Xinyan, and Shashank Sahni. "GPU-to-GPU and Host-to-Host Multipattern String Matching on a GPU." Computers, IEEE Transactions on 62.6 (2013): 1156-1169.

[9] Karp, Richard M., and Michael O. Rabin. "Efficient randomized pattern-matching algorithms." IBM Journal of Research and Development 31.2 (1987): 249-260..

[10] Rafiq, ANM Ehtesham, M. Watheq El-Kharashi, and Fayez Gebali. "A fast string search algorithm for deep packet classification." Computer Communications 27.15 (2004): 1524-1538.

[11] http://www-igm.univ-mlv.fr