www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 3 Issue 6 June, 2014 Page No. 6371-6375

Importance of String Matching in Real World Problems

Kapil Kumar Soni, Rohit Vyas, Amit Sinhal

TIT College Bhopal , Madhya Pradesh India

Email-id: kapilsonitit@gmail.com, rohitvyas83@gmail.com, amit_sinhal@rediffmail.com

Abstract: String Matching is the classical and existing problem, despite the fact that the real world aspects belonging to the research field of computer science. In this domain one or several strings called "Pattern" is to be searched within a well-built string or "Text". String matching strategies or algorithms provide key role in various real world problems or applications. A few of its imperative applications are Spell Checkers, Spam Filters, Intrusion Detection System, Search Engines, Plagiarism Detection, Bioinformatics, Digital Forensics and Information Retrieval Systems etc. This paper is inclusive of analyzing nutshells about string matching along with its long-ago contributory details in an assortment of real world applications.

Keywords: String Matching, Spell checkers, Spam Filter, Intrusion Detection System, Search Engines, Plagiarism Detection, Bioinformatics, Digital Forensics.

I. INTRODUCTION

String matching at-times call string searching and found to be conventional problem in the field of computer science. In string matching pattern strings are searched within a larger string or text. Let us assume that pattern string "p" and text string 'S'. The problem of string matching deals by finding whether a pattern set 'p' occurs in 'S' or not. And if 'p' occurs then the position of it should be reported in 'S' where 'p' occurs [1]. The string matching problem is described in figure 1.

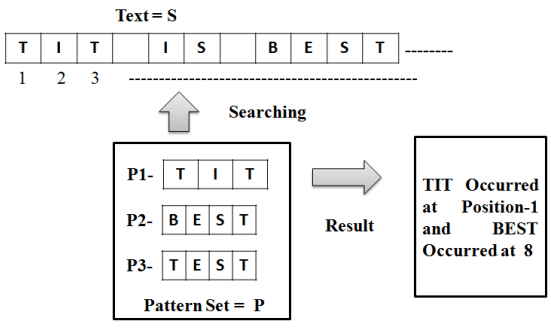

Figure 1 : String Matching Problem

There are various types and classifications of the string matching. There are two types of string matching Exact string matching and Approximate string matching. The search to be done on exact occurrence of the pattern comes underneath the category of Exact string matching. Approximate string matching allows inaccurate searching acceptance founded on specific applications. Based on the number of patterns, string matching has two classifications: Single Pattern string matching and Multiple Pattern string matching. In Single Pattern string matching a single pattern is to be searched in the text whereas in Multiple pattern string matching multiple patterns are searched in the text. Based on the order of searching string matching have four classifications i.e. left to right matching, right to left matching, specific order matching and no order matching[2]. The major classifications are described in the figure 2.

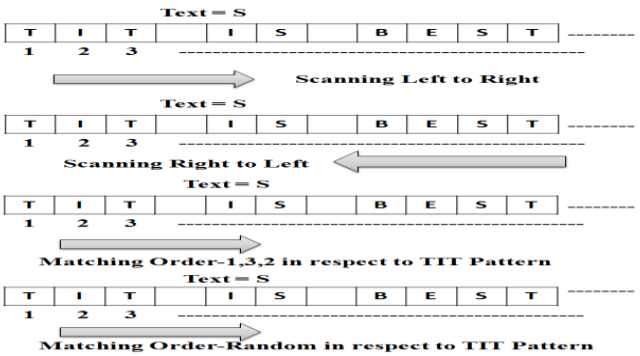

Figure 2 : String Matching Classification

There are many applications in which string matching plays important role. These applications are Spell Checkers [2], Spam Filters [3], Intrusion Detection System [4], Search Engines [5], Plagiarism Detection [6], Bioinformatics [7], Digital Forensics [8] and Information Retrieval Systems [9] and etc.

In this paper the role of string matching in above mentioned applications are being discussed. And brief descriptions of the string matching strategies or algorithms are defined as under:-

II. HISTORY OF STRING MATCHING

The very basic and conventional string matching strategy is Brute Force Algorithm which considers all possible cases and taking shifts only one place to right even match or mismatch condition occurs anywhere. This algorithm also known as Naives approach. [1]

In 1956 Kleene [2] proved the equivalence between finite automaton and regular expression which could be use to solve the string matching problem.

Avoiding numerous comparisons in brute force algorithm, In 1970 Morris and Pratt [11] algorithm was proposed which has linear behaviour. This algorithm is based on pre-processing of pattern and compares character from left to right and if mismatch occurs, it skips some character based on pre-processing phase. In 1977 Knuth Morris Pratt [12] introduced an algorithm having a choice of improvements in Morris and Pratt algorithm. KMP has same time complexity as Morris and Pratt algorithm but searching performance found to be much better than Morris and Pratt algorithm.

In 1977 Boyer and Moore [13] also proposed algorithm which compares character from right to left.

There are so many multiple pattern string matching algorithms has already been proposed in past decades such as: In 1975 Aho-Corasick algorithm [14] was presented by Alfred V. Aho and Margaret J. Corasick, which constructs automata for patterns in pre-processing phase. Commentz Walter [2] proposed an algorithm which was based on Aho-Corasick and Boyer-Moore algorithm, Rabin Karp algorithm [15] is also used to search multiple patterns.

An assortment of algorithms based on different methodologies has already been suggested in the past decades, historical listing of various important string matching algorithms is being described in the figure 3: String Matching History [1, 2, and 17].

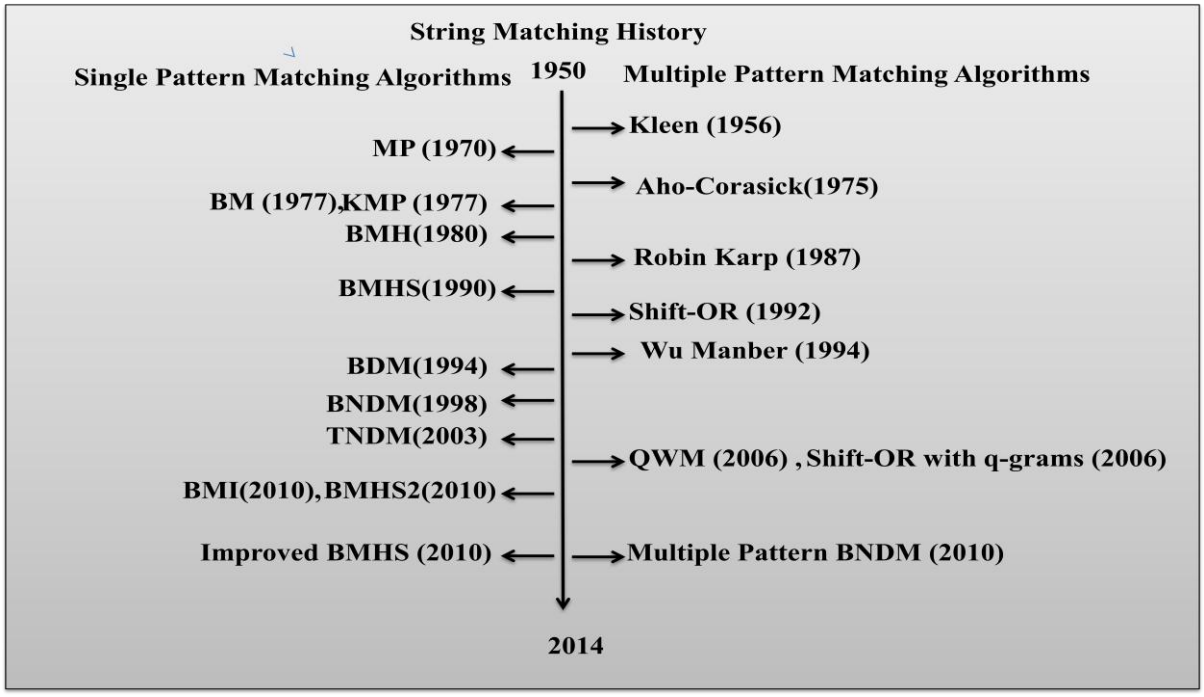

Figure 3: String Matching History [1, 2, 17]

III. STRING MATCHING APPLICATIONS

In perspective to the real world problems string matching is having several applications, few of which are being described here.

A. Spell Checkers: In spell checkers [2] we build a "trie" of pre-defined set of patterns. This trie is used for the string matching means if any such pattern occurs then it shows the occurrence by reaching to its final states. Spell Checkers basic module is shown in figure 4.

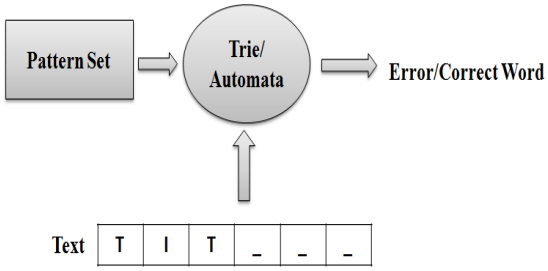

Figure 4: Spell Checker

B. Spam Filters / Spam Detection Systems: Unsolicited and unwanted emails called spam that engages lots of network bandwidth. This will causes great financial loses. All spam filters use the concept of string matching to identify and discard the spam. Spam filter searches suspected signature patterns in the content of email by applying string matching. All content based filters are worked on string matching [3]. Spam filter basic structure is shown in figure 5.

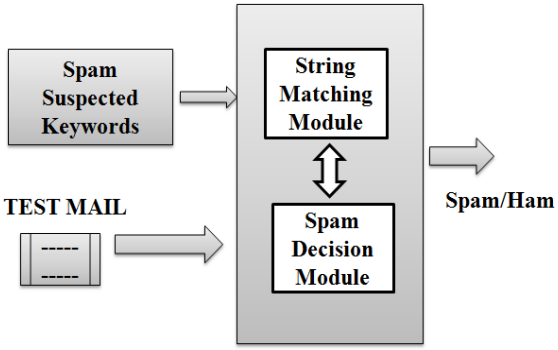

Figure 5 : Spam Filter

C. Intrusion Detection System: In Intrusion Detection System [4] data packets that contain intrusion related keywords are found by applying string matching strategy. All the malicious code is stored in the database and every incoming data is compared with stored data. If match found then alarm is generated. It is based on exact string matching algorithms where we have to capture each and every intruded packet and they must be detected. The Intrusion detection system modal is shown in figure 6.

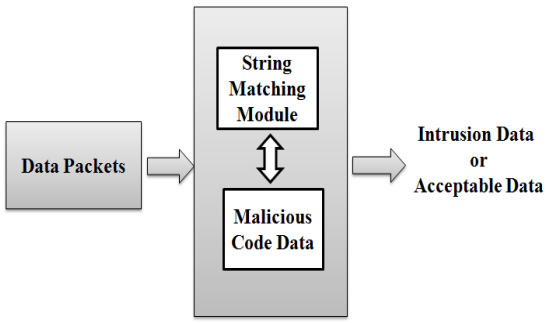

Figure 6 : Intrusion Detection System Model

D. Search Engines / Content Searching in Large Databases: Most of the data are available on internet in the form of textual data. Due to the large quantity of uncategorized text data, it becomes really difficult to search a particular content. Web search engines help us to solve this problem by organizing the required text / data as efficiently as possible. To categorize these data string matching algorithms are used. Categorization is done on the basis of search keywords [5]. Figure 7 shows the basic model of Search Engine.

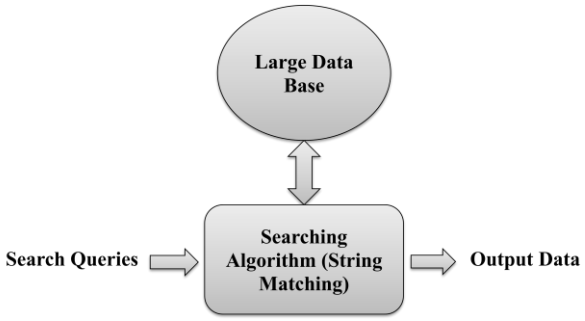

Figure 7 : Search Engine Module

E. Plagiarism Detection: Copy someone's work and claim it as own is called as Plagiarism. So with the use of string matching we can compare the texts and detect the similarities between them. On the basis of these similarities declare whether it is original work or taken from somewhere else. Figure 8 shows the Plagiarism detection technique [7].

F. Bioinformatics / DNA Sequencing: Bioinformatics is the application of information technology and computer science to biological problems, in perspective to the issues involving genetic sequences and in order to find the DNA patterns, string matching module and DNA analyser both works with collaboration for finding the occurrence of the pattern set [7]. Figure 9 shows the Bioinformatics DNA Sequencing Module.

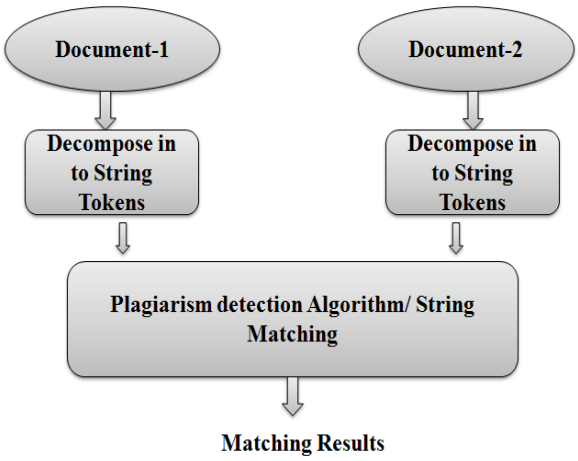

Figure 8: Plagiarism Detection System

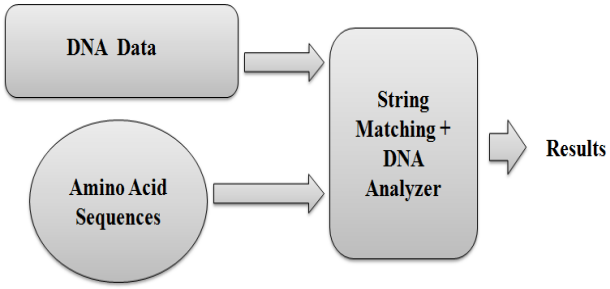

Figure 9: DNA Sequencing Module

G. Digital Forensics: Digital forensics refers to the recovery and investigation of material found in digital devices. In digital forensic text string searches are designed to search every byte of digital evidence, at the physical level, to locate specific text strings of interest to the investigation. Figure 10 describe the basic model where string matching is used.

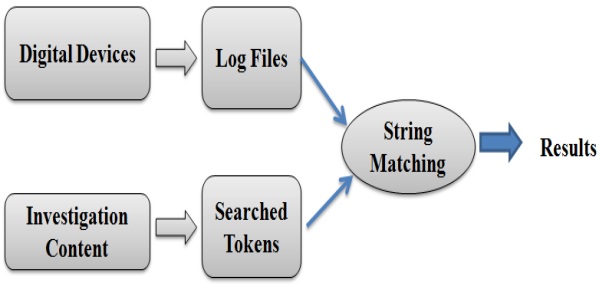

Figure 10: Digital Forensic Results

H. Information Retrieval: In text mining task designed to extract previously unknown information by analysing large quantities of text. String matching plays very vital role here like as information extraction, topic tracking, question answering etc. Figure 11 shows the basic structure of information retrieval system.

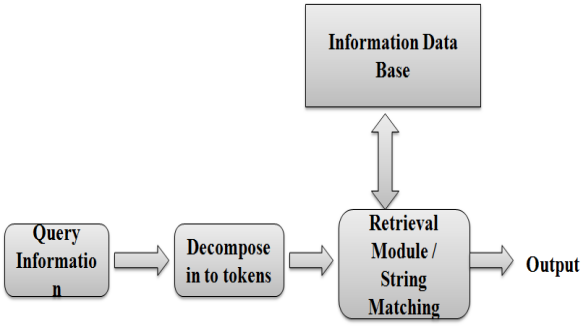

Figure 11: Information Retrieval Modal

IV. CONCLUSION

String matching has greatly influenced the field of computer science and will play an important role in various real world problems. As the time grows more and more efficient string matching algorithms will be used. Since 1950 lots of single and multiple patterns string matching algorithms has been suggested. There are many more possible areas in which string matching can play a key role for excelling. Improvement and creative activities in string matching can provide the major role for getting time proficient performance in various domains of computer science. Its application area is wide and demand is expected to be increased in the future. Efficient and innovative searching algorithms will be the major research areas for the future perspective.

REFERENCES

[1]. Christian Charras and Thierry Lecroq," Handbook of Exact String_Matching Algorithms", Published in King's college publication, Feb 2004.

[2]. Alberto Apostolico and ZviGalil," Pattern Matching Algorithms" Published in Oxford University Press, USA, 1st edition, May 29, 1997.

[3] Ching-Tung Wu, Kwang-Ting Cheng, Qiang Zhu and Yi-Leh Wu, "Using Visual Feature For Anti-Spam Filtering", In the proc. of IEEE International Conference on Image Processing (ICIP2005), pp. 509–512, 2005.

[4]. Hyunjin Kim, Hong-Sik Kim and Sungho Kang,"A Memory-Efficient Bit-Split Parallel String Matching Using Pattern Dividing for Intrusion Detection Systems" IEEE Transactions on Parallel and Distributed Systems, Volume:22 , Issue: 11, pp. 1904-1911, Nov 2011.

[5]. Sanchez D., Martin-Bautista M.J., Blanco I. and Torre C.," Text Knowledge Mining: An Alternative to Text Data Mining", In the proc. of IEEE International Conference on Data Mining Workshops, ICDMW '08, pp. 664-672, 15-19Dec. 2008.

[6]. Ramazan S. Aygün "structural-to-syntactic matching similar documents", Journal Knowledge and Information Systems, ACM Digital Library, Volume 16 Issue 3, pages 303-329, Aug 2008.

[7]. Lok-Lam Cheng, David W. Cheung and Siu-Ming Yiu," Approximate String Matching in DNA Sequences", In Proceedings of the Eighth International Conference on Database Systems for Advanced Applications (DASFAA'03), pp. 303-310, 26-28 March 2003.

[8]. Jooyoung Lee, Sungkyung Un, and Dowon Hong," Improving Performance in Digital Forensics: A Case using pattern matching board", In the Proc. of International Conference on Availability, Reliability and Security(ARES), pp. 1001-1005, 16-19 March 2009.

[9]. Mei-Chen Yeh and Kwang-Ting Cheng," Fast Visual Retrieval Using Accelerated Sequence Matching", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 13, NO. 2, APRIL 2011.

[10]. MORRIS JR, J. H. AND PRAT, V. R.,"A linear pattern-matching algorithm", Technical Report 40, University of California, Berkeley, 1970.

[11]. KNUTH, D. E, MORRIS JR J. H AND PRATT V. R,"Fast pattern matching in strings", In the procd. Of SIAM J.Comput.Vol. 6, 1, pp. 323–350, 1977.

[12]. KNUTH, D. E, MORRIS JR J. H AND PRATT V. R,"Fast pattern matching in strings", In the procd. Of SIAM J.Comput.Vol. 6, 1, pp. 323–350, 1977.

[13]. BOYER, R. S. AND MOORE, J. S,"A fast string searching algorithm", Communication of ACM 20, Vol. 10, pp. 762–772, 1977.

[14].Alfred v. aho and margaret j. corasick,"efficient string matching: an aid to bibliographic search" communication of acm, vol. 18, june 1975.

[15]. Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford (2001-09-01). "The Rabin–Karp algorithm". Introduction to Algorithms (2nd ed.). Cambridge, Massachusetts: MIT Press. pp. 911–916.

[16]. Lok-Lam Cheng, David W. Cheung and Siu-Ming Yiu," Approximate String Matching in DNA Sequences", In Proceedings of the Eighth International Conference on Database Systems for Advanced Applications (DASFAA'03), pp. 303-310, 26-28 March 2003.

[17]. Ali Peiravi, "Application of string matching in Internet Security and Reliability", Marsland Press Journal of American Science 2010, 6(1): 25-33.

[18]. Jingbo Yuan, Jisen Zheng, Shunli Ding, "An Improved Pattern Matching Algorithm", In the Proc. of Third International Symposium Intelligent Information Technology and Security Informatics, pp. 599-603, 2-4 April 2010.