Review of using regular expression for efficient classification in various applications

Dinesh D. Puri[1], Dr. G. K. Patnaik[2]
[1]*Research Scholar, SSBT's COET*
[2]*Professor, SSBT's COET*
[1]*ddpuri@gmail.com*, [2]*girishpat2001@yahoo.com*

Abstract:
Regular expressions are playing key role in various applications such as network security, network traffic classification, healthcare, bioinformatics etc. Here the main focus is on review using regular expression for pattern matching and classification. Methods such as RED + ALIGN, RED + SVM, SVM are reviewed for better performance. The data is represented in terms of patterns specially in terms of string. Now days various methods are invented to handle multiple patterns designed in regular expression which is used for string matching. For fast regular expression matching memory efficient DFA based approach is also discussed

Keywords: Regular expression, pattern matching, DFA

1. Introduction

A regular expression is a sequence of characters used to describe a pattern of text. It is a standard technique supported by most programming languages. Usually this pattern is used by string searching algorithms or for input validation. The natural language processing in very important domain in which pattern matching is key issue. With the help of regular expression NLP related challenges can be solved. For that purpose regular expression can be generated related to that domain. Regular expression can be generated manually or with help of tool. But there is not standard method by which regular expression can be generated. The expansion and maintenance is also big challenge. The regular expression can be used with top down approach or bottom up approach according to the need of application.

The use of regular expression can be done to build classifier to construct effective decision tree. Classifier classifies the input data to appropriate class with the help of pattern mathing. Traditional classifiers are already used for this but the combination of such traditional classifiers with regular expression classifiers will be more effective.

This paper is presenting literature survey on use of regular expression in various domains such as natural language processing, data mining and machine learning. In next section discussion on literature survey is done and author stated his view. The paper is concluded with conclusion section.

2. Literature Survey

J. Duy Duc An Bui [1] described Regular Expression Discovery (RED) which is used to generate regular expression. In this paper two RED based classifiers are implemented which is RED +ALIGN classifier and RED + SVM classifier. Two clinical dataset were used in this project obtained from UD department of VA. The evaluation results suggests that the two classifiers using Regular Expression Discovery patterns having better performance than SVM performance. To improve performance of classifier it is suggested that RED + ALIGN classifier can be combined with SVM and other classifiers.

Mowbray, Miranda; et al. [2] addressed challenges to compile large set of patterns it may be in different data structure. It helps runtime to match input string to appropriate patterns effectively. Some time input string matches more than one pattern. This is main challenge in efficient classification of strings. To tackle with above problem sequential approach can be used but main problem with this approach is it will take more time as pattern increases. This paper also suggested the refinements so that decision tree compilation algorithm will be time effective.

Sailesh Kumar et al.[3] This paper, introduced a new representation for regular expressions, That new representation is called as the Delayed Input DFA, which is very memory effective.

To construct D2FA some changes should be made in DFA so that instead of several and multiple transitions one transition can be used it is called as single default transition with this solutions number state will be reduced and it will save the time. Basically using DFA itself saves time and if D2FA is used it reduces the state and number computations will be reduced. This approach also save space and memory.

Now a days very fast networks are designed such as backbone networks. To apply sedcurity constraints at such high speed with traditional method and to achieve high accuracy is not so easy. High speed lan where propagation delay is greater than transmission delay is very difficult to handle packet scanning. In such case this approach used very effective which reduces the transisions more than 90%.

Vinoth George et al.[4] In todays network traffic classification of network is key problem. Traffic classification identifies what applications are used by the end users. The network contains various protocol and each protocol is having various packet format. While general traffic classification is performed it is done through checking the payload as well as header. The main approach is use of signature with each type of protocol and packet format. While packet is injected in the network, the signature is attached with the packet header and it is transmitted in the network. At the receiver side that signature is decoded and packet will be label to appropriate class. Main issue with the signature generation was it use to generate manually by checking packet's protocol type with the help of packet filter but manually signature generation was having its own drawbacks. This paper described how the signature can be generated with the help of regular expression. The main steps to generate signatures are preprocessing which discard some packets based on some defined rules. The information about packet can be extracted by any filter such as wire shark .Next step is substring extraction, which can extracted with information collected with previous stage. Next step is signature generation. The signature should update periodically to increase the reliability.

Fung yu et al. [5] Network security is the main concern in secure communication. In todays high speed network scanning packet content is become extremely important. Each packet is having two important fields header and payload. For secure communication header scanning as well as payload scanning is essential. This paper presenting payload scanning for secure communication. In payload scanning use of regular expression work as catalyst. For regular expression matching special syntax is also used. XML filtering also one of the way for payload scanning. This paper also comparing the performance of both approach i.e regular expression matching and XML filtering. Some important notations are use for regular expression matching such as "^" indicates that string starts with some suffix for Example ^ab means start of input is from ab. If the pattern is without this symbol it can be matched anywhere. The notation | indicates the or operation in the pattern. It selects on of the string and still pattern will be successfully match. The . will work as wild character. The * quantifier denotes zero or more appearance. The {} denotes repeating substring such as p{50} denotes 50ps

Utkarsha P. Pisolkar et al[6] This paper also discusses the security constraint in the network. The intrusion detection and prevention is main issue. For security check use of regular expression can be done memory size is big issue. The regular expression can be represented in DFA. If the DFA state is reduced the space effective solution will be provided.

3. Discussion

In literature survey various issues are focused such as regular expression generation ,running time in pattern matching, fast and memory efficient regular expression matching. One thing came to know that in comparison with another classifiers, regular expression base classifier is having better efficiency. To increase the performance regular expression based classifier can be combined with other classifier like SVM to improve classification performance.[1]

The classifiers are trained by training data which is also called as training set and then test of performance of classifier is taken by testing data. If data is in form of string and new input string come in to the system the string can be considered to create update training set to enable time to time repeatedly compilation of the decision tree to include and consider more recent input data.[2]

4. Conclusion

Regular expression can be used more effectively for classification and string matching purpose. If traditional classification methods are combined with regular expression based classifier it will produce better results. Some memory related issues can be addressed for better performance.

5. References

[1] Qing Zeng-Treitler, Learning regular expressions for clinical text Classification, Bui DDA, et al. J Am Med Inform Assoc 2014,page number 850–857. Year 2013.

[2] Mowbray, Miranda; Horne, William; Rao, Prasad, Efficient classification of strings using regular expressions Hewlett Packard Labs HPE-2017-03

[3] Sailesh Kumar ,"Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection" SIGCOMM'06, 11th -15th september, 2006 Italy.

[4] Vinodh Ewards, "Efficient Regular Expression Signature Generation for Network Traffic Classification", International Journal of Science and Research , ISSN: 2319-7064, Volume 2 Issue 3, March 2013.

[5] T. V. Lakshman, " Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection" ANCS 2006, December 3–5, 2006 USA.

[6] T. J. Green, "ProcessingXML Streams with Deterministic Automata and Stream Indexes," ACM, vol. 29, 2004.

[7] Y. Diao, M. Altinel "Path Sharing and Predicate Evaluation for High-Performance XML Filtering, "ACM, 2003.

[8] Utkarsha P. Pisolkar " A Memory Efficient Regular Expression Matching by Compressing Deterministic Finite Automata" IJoCA Page number 0975 – 8887) Vol. 122 – No.20, July 2015.

[9] Ken ompson, "Pro rammin tec niques: re ular expression searc al orit m", in Communications of the ACM, Pages 419-422 , Volume 11 Issue 6, June 1968

[10] Jeffrey D. Ullman, "Introduction to Automata Theory Computation", year1979.