A Study on PSP Algorithm for Automatic Generation of Internet Traffic Signature

Baraka D. Sija, Kyu-Seok Shim, and Myung-Sup Kim
Dept. of Computer and Information Science, Korea Univ.
{Sijabarakajia25, kusuk007, tmskim}@korea.ac.kr.

Abstract

In this paper we propose an algorithm approach, so called PSP (Prefix tree for Sequential Patterns) for automatic Internet traffic signatures generation. In presenting PSP algorithm approach, we basically refer it to the GSP (Generalized Sequential Pattern), since PSP algorithm is an extension of GSP algorithm. Actually the two algorithms were originally proposed to deal with data mining problems, in searching of sequential data patterns. Therefore, the algorithms have been mostly applied in data mining related researches, we take a close study and a new approach of PSP algorithm so as to apply it in Network Management field for automatic generation of Internet traffic signature. In this paper we particularly apply PSP algorithm to generate sequential signatures patterns that will help us to identify applications such as Facebook, Youtube, Kakaotalk, BiTorent, Skype and HTTP which are daily interfaced by millions of people in the Internet across the global. In this paper we show that PSP algorithm is more efficient in generating effective signatures compared to both GSP and Apiori algorithms in reduction of candidates' numbers (signature patterns) to be examined in traffic identification, searching speed and reduction of memory usage.

I. Introduction

A huge number of Internet traffic are still highly encrypted, unidentified and unclassified. Since the Internet technology was allowed to be interfaced by the public millions of high-tech smart devices and high speed packets routing network devices have been either developed or discovered. This freedom has led to high traffic congestion in the Internet. As stated above most of these Internet traffic are still hard to identify and classify. Some of the traffic are even malicious to other traffic. Hereupon comes the need of researching on best methods for automatic generation of Internet traffic signature to identify and classify various traffic in the Internet

In order to counter this problem and manage daily rapid increase of Internet traffic a number of research approaches have been proposed. Won et al [1] present "Automated Application Signature Generation for Traffic Identification", Yoon et al [2] present "Behavior Signature for Fine-grained Traffic Identification traffic" and Park et al [3] present, "Towards Automated Application Signature Generation for Traffic Identification". For efficient traffic Identification, all of the mentioned and unmentioned research approaches still indicate weaknesses in terms generated signature patterns number, generated signature patterns searching speed and memory usage.

In this paper we propose PSP algorithm approach which has already been applied in data mining field researches and proved to be better and efficient [4-5]. Through a close study on the PSP algorithm we have found that the algorithm approach is better than previous algorithms bringing a solution to generated signature patterns searching speed, reducing generated candidates' numbers (signature patterns) to be examined in traffic identification and conserve memory usage. The three new features of PSP algorithm show much improvements in automatic signature generation.

The remainder of this paper is organized as follows. Section II, which is the body of this paper divides into two parts (II.1) and (II.2). In part (II.1), we briefly present GSP algorithm approach. In part (II.2), we present the proposed PSP algorithm approach and finally in section III we come to conclusion and brief our future work.

II. PSP for Automatic Signature Generation

II.1 GSP Algorithm Approach

GSP (Generalized Sequential Patterns) is a data mining APRIORI property based algorithm which requires multiple scans over the sequence database to determine candidate sequences and their supports [6]. Table 1 below shows a length-2 generation of sequential patterns under GSP algorithm. We apply the Table 2. generated signatures patterns throughout this paper.

Table 1. Generation of Length-2, signatures patterns

	<a>	<b>	<c>
<a>	<aa>	<ab>	<ac>
<b>	<ba>	<bb>	<bc>
<c>	<ca>	<cb>	<cc>

Table2. Orderly selection of signatures patterns

	<a>	<b>	<c>
<a>	~~<aa>~~	<ab>	<ac>
<b>	~~<ba>~~	~~<bb>~~	<bc>
<c>	~~<ca>~~	~~<cb>~~	~~<cc>~~

According to GSP algorithm, initially, every item in database is a candidate of length-1, generated pattern items (Length-2 candidates) are listed in their particular sub-sequences of each sequence orderly(alphabetically). For each level (i.e., sequences of length-k) the database is scanned to collect support counts for each candidate sequence and generate candidate up to length-(k+1) sequences from length-k frequent sequences using Apriori.

The process repeats until no frequent sequence or no candidate can be found in the database [4].

Table 2 indicates the phenomenon; whereby red color highlighted pattern items are first relisted orderly and then excluded in the Database Sequence table before further candidate (signature pattern) generation to length k+1. In so doing all unorderly listed sub-sequence pattern of signatures initially won't be used in the mining process. Sequence Database of orderly listed signature patterns is formed in Table 3. Both GSP and PSP algorithm use these formed signature patterns for extracting signature patterns of high frequency to a last phase (length k+1) of extraction (scanning) through certain designated support values.

Table 2. Sequence Database of signatures patterns

S-Id	Sequence of signatures patterns
01	<(ab)(b)(b)(ab)(b)(ac)>
02	<(ab)(bc)(bc)>
03	<(b)(ab)>
04	<(b)(b)(bc)>
05	<(ab)(ab)(ab)(a)(bc)>

II.2 The Proposed PSP Algorithm Approach

PSP (Prefix tree Sequential Patterns) algorithm is an order sequential patterns with the root of a null sequence • [6]. Each child node stores a sequential pattern, its support value and the flag to determine sequential patterns [6]. At level 1, each node is set with a single frequent item; at level k, each node is set with a k-pattern sequence. Recursively, there are nodes at the next level (k+1) after a k-pattern sequence is extended with a single frequent item. [6].

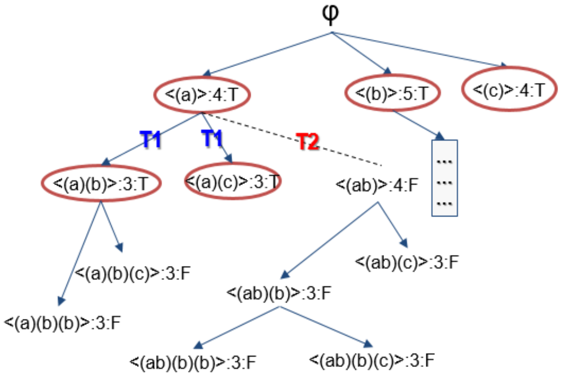

Fig.1. PSP structure with sequential patterns from Table 3.

Fig. 1 shows the prefix tree representative of sequential signatures patterns generation of the sequence database in Table 3. The generated sequential signatures patterns are highlighted in the ellipses, where minSup = 50%. Sequences <(a)(b)> and <(a)(c)> are sequence-extended sequences of <(a)> and <(ab)> is an itemset-extended sequence of <(a)>. Sequence <(a)> is a prefix of all sequences in T1 and an incomplete prefix of all sequences in T2. [6]. The same procedure of generating sequential signature patterns and extending its sequence and itemset continues to sequence <(b)>. Due to space limitations for clarity we have shown only sequence <a>'s extended signature itemsets and sequences.

Similarly, the process for sequence <(b)> has sequence-extended sequences <(b)(a)>, <(b)(b)>, and <(b)(c)> and itemset-extended sequence <(bc)>. Sequence <(b)> is a prefix of all sequences in T3 and an incomplete prefix of all sequences in T4. T3 is a region similar to T1 (indicating completeness in prefix for sequence <(b)>) and T4 is a region similar to T1 (indicating incompleteness in prefix for sequence <(b)>). In a given minimum support threshold (MinSup) and a sequence database generated signature patterns, the problem of extracting sequentially generated patterns takes place. PSP algorithm is an orderly Prefix-tree sequential patterns generating method. From this fact you can clearly see from Fig 1. that sequence <c> extends neither sequences nor itemsets. You can also clearly see from Table 3, from which Fig 1's tree structure was formed. Thus, as long as sequence <c> is completely not a prefix of any sequence hence none of its either extended sequences or itemsets are formed.

III. Conclusion and Future Work

In this paper we propose a new approach on PSP (Prefix tree Sequential Patterns) algorithm to generate signatures for Internet traffic identification. We have presented and applied PSP algorithm similarly as it is applied in data mining field. Major strengths of PSP algorithm are reduction of number of candidates (signature patterns) to be examined in traffic identification, high speed signature patterns searching and reduction memory usage size.

Our future work is to apply the algorithm in real time environments as well and see whether the stated PSP features are practically superior to both GSP and Apriori algorithms or not. We will also keep learning the current PSP mechanism which applies to only ordered signatures patterns so as to generate and apply both ordered and unordered signatures patterns, which we believe will be much more efficient in Internet traffic identification, Making it much more suitable and efficient approach in automatic generation of Internet traffic signature.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A3A01018057) and KREONET-Emulab Testbed of Korea Institute of Science and Technology Information (K-16-L01-C02-S03).

References

[1] Y.J. Won, S.C Hong, B.C Park, and J.W. Hon "Automated Application Signature Generation for Traffic Identification", POSTECH, Korea, Aug. 16, 2008.

[2] S.H. Yoon, J.S Park and M.S Kim, "Behavior Signature for Fine-grained Traffic Identication" Korea Univ. 1 Apr. 2015

[3] B.C Park, Y.J. Won, M.S Kim and J.W. Hong, "Towards Automated Application Signature Generation for Traffic Identification" IEEE Xplore R

[4] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. EDBT'96.

[5] F. Masseglia, F. Cathala. and P. Poncelet, "The PSP Approach for Mining Sequential Patterns" France.

[6] Thi-Thiet Pham, "Efficiently Mining Sequential Generator Patterns Using Prefix Trees" DOI 10.3233/FI-2015-1217, IOS Press.