ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2009. No. 5(93)
ELEKTRONIKA IR ELEKTROTECHNIKA

TELECOMMUNICATIONS ENGINEERING
T180
TELEKOMUNIKACIJŲ INŽINERIJA

Detection of Trends of Internet Traffic using Sequential Patterns

J. Jeļinskis, G. Lauks
Institute of Telecommunication, Faculty of Electronics and Telecommunication, Riga Technical University
Āzenes str. 12, LV-51368 Riga, Latvia, phone: +371 7089201; e-mail: lauks@rsf.rtu.lv, jans.jelinskis@etf.rtu.lv

Introduction

The modern traffic engineering has become more and more information-based. Given the widespread use of information technology a large number of data are collected in on-line real-time communication environments, which results in massive amounts of data. Such time-ordered information typically can be aggregated with an appropriate time interval, yielding a large volume of equally spaced time series data. Those can be used in many fields of traffic engineering such as adaptive fraud detection, intrusion prevention, active bandwidth allocation, throughput prediction etc.

Trend detection is a very important topic in data traffic research and is ubiquitous in the communication system characterization. Trends are usually estimated using simple linear regression. When taking into account the complexity of the communication systems, trends are expected to have various features such as global and local characters, and this leads to the more complex evaluation where the trend line is locally linear but with change points where the slope and intercept change.

Withal sequential pattern analysis is used to identify frequently observed sequential occurrence of items across ordered transactions over time. Such a frequently observed sequential occurrence of items (called a sequential pattern) must satisfy a user-specified minimum support. Understanding long-term process behavior is the goal of the sequential pattern analysis.

Application areas for sequential patterns include analysis of telecommunication systems, discovering frequent buying patterns, analysis of patients' medical records, as well as Web mining application development etc.

Time series analysis is often associated with the discovery and use of patterns (such as periodicity, seasonality, or cycles), and prediction of future values (specifically termed forecasting in the time series context). This leads to the necessity of large dataset mining and especially data classification for it effective use in control in traffic engineering field.

Large time-series mining for sequential pattern detection is motivated by the decision support problem faced by most traffic engineering tasks. A number of related methods have been proposed based on the well-known learning models like decision tree or neural network. [4, 5].

For all that, kinds of classification methods mentioned above may not perform well in mining time sequence datasets [6] like time-series data that are widely used when analyzing communication network performance.

It is therefore important to develop methods that permit a systematic decomposition of traffic data into different sequence patterns that thereafter can be used for trend detection in data time-series.

In this paper we verify the possibility of usage of trend detection method, using sequential pattern analysis (Mining Sequential Patterns) with throughput time-series. Experimental data are based on measured data traces, where Riga Technical University campus traffic was used. Throughput time-series data traces of 100 millions packets were used.

The main methodology of this method is to integrate the sequential pattern mining with the pattern comparison and matching, and the trend detection in the analyzed time-series trace data of the throughput of internet connection.

Statement of the problem

The subjacent problem of majority of traffic engineering tasks is an appropriate decision making under uncertain conditions. Any possible way of prediction of future events or, to some extent, trend detection of some process, can be the priceless assistant when making a decision.

We assume that specific mix of traffic in determined switching environment can be characterized not only by seasonal trends, but also with frequently observed local trends, that my vary by duration and growth rate. We also assume that the decision making in the underlying control layer of the specific network area should not be dependent only on the local traffic characteristics but also should be a self-instructional with the possibility to adapt to the changing behavior of the mix of the analyzed traffic. Such control system can be developed if the trend detection of the stochastically changing data, that is in the essential variable in the decision making process, can be done in different time scales and the results can be stored as a knowledge using if-then rules.

In this research we propose to divide the decision making process activities into 3 groups:

1) Long term decisions – hours, days, weeks;
2) Short term decisions – minutes;
3) Fast online decisions – seconds.

The detection of trends in stochastically changing traffic data should be than divided accordingly to the assumed separation of the decision making terms.

The long term decisions can be easily made based on seasonal trends that can be mined out of the large time-series data sets. We do not pay a lot of attention to this aspect in this research, as the fuzzy pattern recognition is very effective in this case and with certainty finds the coincidence.

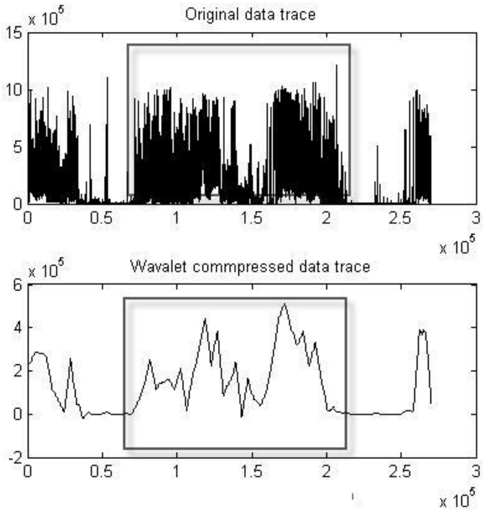

Fig. 1. Large seasonal trend of daytime data traffic throughput – effective for fuzzy pattern recognition

Problems should emerge when seasonal trend are not so typically marked - effective fuzzy trend detection then becomes a difficulty while mining through the time-series traffic data.

Step 1 – Template pattern discovery in time-series data

In this research we have used the denoised 100 million packet data throughput time-series traces, which after normalization became an appropriate data for template pattern discovery.

The process of wavelet transform (WT) was used to decompose the Time-Series Data with the low and the high pass filters. When WT is used for original time-series data denoising and pattern detection than the high frequency component is filtered out [3].

The search for patterns was differentiated upon the length analyzed data chunk. Time-series line segments starting form 20 seconds and finishing at 10 minutes were preferred for forthcoming pattern comparison in order to detect trends in analyzed long term data. In Fig. 2 and 3 the examples of highlighted template patterns are shown.

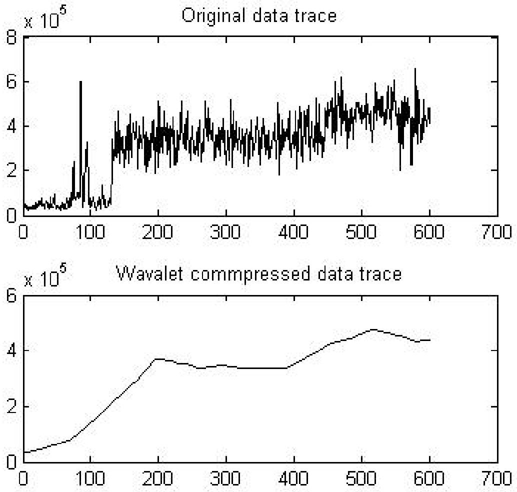

Fig. 2. Pattern of 10 minute data throughput trace

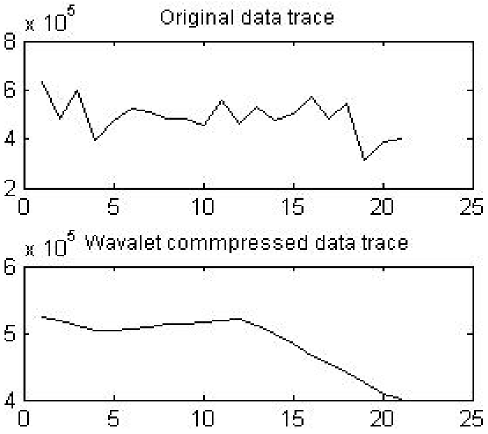

Fig. 3. Subpattern of 10 minute data throughput trace – 20 second cut

Step 2 – Pattern comparison and matching for sequential pattern analysis

As the main instrument in the sequential pattern analysis the fuzzy pattern recognition was used in this investigation. Pattern comparison and matching is the keystone operation for effective operating of the sequential pattern detection.

A time-series sequence is a sequence of real numbers, each of which represents a measure at a time point. Those sequences are used as patterns when analyzing time-series data for pattern detection. Given a collection of time-series sequences and a query sequence Q, time-series similarity analysis is to find those sequences that are similar to Q (called whole matching), or to identify the sequences that contain subsequences similar to Q, that is called subsequence matching (See Fig. 4).

For pattern matching the Fuzzy c-means algorithm was used that is frequently used in pattern recognition. This method was developed by Dunn in 1973 [8] and improved by Bezdek in 1981 [9].

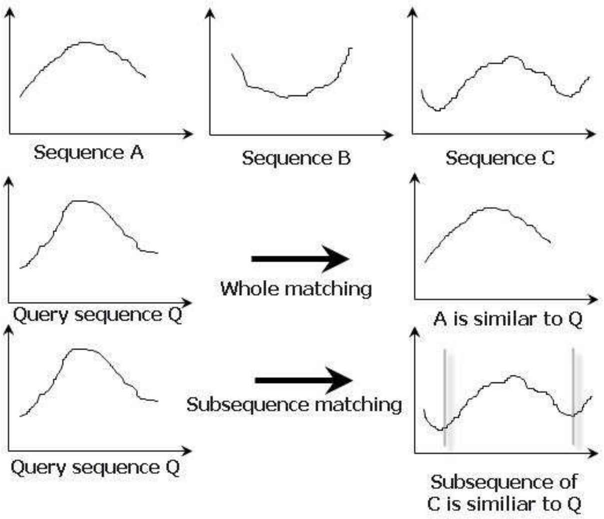

Fig. 4. Whole matching and subsequence matching in the pattern recognition process

With the increasing number of template patterns, the performance of pattern recognition with the fuzzy c-means decreases (See Fig. 5.). The growing number of clusters means the more precise mapping of the possible sub-trends, but at the other hand the degree of the analyzed object's membership to classes decreases, which emerges with a worse pattern matching.

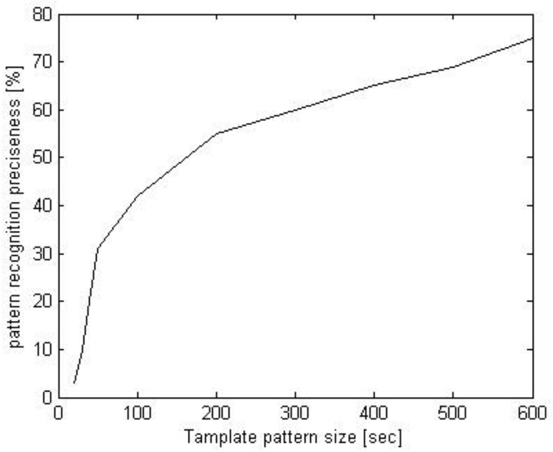

Fig. 5. Pattern recognition preciseness dependence on the template pattern size

As it seen form the results (See Fig. 5.), we can more or less rely on the recognized patterns credibility when using template patterns of the size of 300 and more seconds. This leads to the conclusion that using this pattern recognition method and this specific mix of traffic the decision making process can be done in the time frontier of minutes and hours, but not seconds, which excludes effective fast online decision making possibility.

As it was mentioned before, we assume that every specific mix of traffic can be characterized by the different template pattern characters. This leads to different organization of decision making time ambit and should be done upon a continuous analysis of network performance.

Step 3 – Trend detection with sequential patterns mining and knowledge generation

The problem of discovering sequential patterns is to find inter-transaction patterns such that the presence of a set of items is followed by another item in the time-stamp ordered transaction set [1, 2].

In Web server logs, for example, a visit by a client is recorded over a period of time. The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user navigation patterns and helps in targeting advertising aimed at groups of users based on these patterns [7].

By finding sequential patterns in the data throughput time-series traces we can treat them as trends in internet traffic and searching through the different time scales generate the knowledge by the if-then rules.

The analyzed data was mined to find sequentially occurring patterns that were nominated as templates. Using fuzzy pattern recognition the patterns starting form 600 seconds were mined. Sequential patterns were detected more accurately with the increase of number of the template patterns, but less frequently, using the same minimum support (See Fig. 6.).

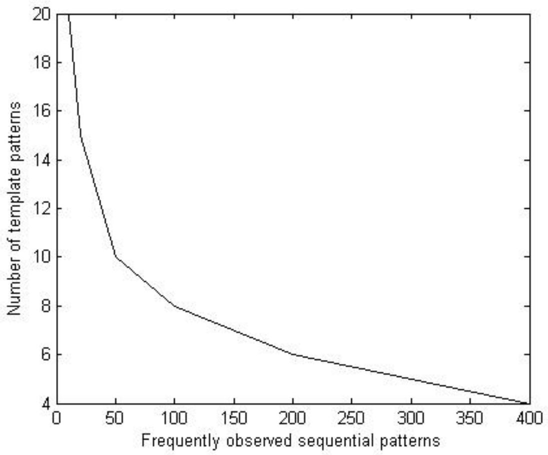

Fig. 6. Sequential patterns detection dependence on the template pattern number

Conclusions and future research

In this paper we propose a method of possible trend detection in the internet traffic by using sequential pattern analysis. We can conclude, that the basic step of this method, which is based on fuzzy pattern detection is dominant in the performance of the proposed approach. It defines the time resolution of the decision making possibilities even if the mining for patterns is done correctly and fast enough.

The next limitation is the number of template patterns which are used in the real data search for sequential patterns. Sequential patterns were detected more accurately with the increase of number of the template patterns, but less frequently, using the same minimum support. This aspect requires the optimization procedures and is the object of the future research.

We assume that every specific mix of traffic can be described by the different template pattern characters. This leads to different organization of decision making time ambit and should be done upon a continuous analysis of used parameters, such as throughput. Thus, this method is suitable only for one specific blend of traffic and becomes a strictly heuristic as soon as it is not modified accordingly to the continuously changing traffic characteristics.

References

1. **Mannila H., Toivonen H., and Verkamo A. I.** Discovering frequent episodes in sequences // In Proceedings of the First International Conference on Knowledge Discovery and Data Mining. – Montreal, Quebec. – 1995. – P. 210–215.
2. **Srikant R., Agrawal R.** Mining sequential patterns: Generalizations and performance improvements // Proceedings of the Fifth International Conference on Extending Database Technology. – Avignon, France. – 1996.
3. **Percival D., Walden A.** Wavelet Methods fot Time-Series Analysis // Cambridge University Press. – 2000.
4. **Rouming Jin, Gagan Agrawal**. Efficient decision tree construction on streaming data // Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. – Washington. – 2003. – P. 571–576
5. **Liang B. and Austin J**. A neural network for mining large volumes of time series data. // IEEE International Conference on Industrial Technology (ICIT). – New York. – 2005. – P. 688–693.
6. **Vincent S. M., Tseng Chao-Hui Lee.** CBS: A New Classification Method by Using Sequential Patterns // Proceedings of the SIAM International Data Mining Conference. – California, USA. – 2005. – P. 596–600.
7. **Xiaozhe Wanga, Ajith Abrahamb, Kate A. Smith**. Intelligent web traffic mining and analysis // Journal of Network and Computer Applications 28. – 2005. P. 147–165.
8. **Dunn J. C.** A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics 3. – 1973. – P. 32–57.
9. **Bezdek J. C.** Pattern Recognition with Fuzzy Objective Function Algoritms. – New York: Plenum Press, 1981.

J. Jeļinskis, G. Lauks. Detection of Trends of Internet Traffic using Sequential Patterns // Electronics and Electrical Engineering. – Kaunas: Technologija, 2009 – No. 5(93). – P . 3–6.

We propose a new method of possible trend detection in the internet traffic by using sequential pattern analysis. The main steps of this method, such as the template pattern discovery in time-series data and pattern comparison and matching for sequential pattern analysis, are analyzed and the possible constraints and problems are pointed. Experimental data are depicted and future research subjects are described. Ill. 6, bibl. 9 (in English; summaries in English, Russian and Lithuanian).

Я. Елинскис, Г. Лаукс. Определение тенденций потока данных интернета путем использования секвенцных шаблонов // Электроника и электротехника. – Каунас: Технология, 2009. – № 5(93). – С. 3–6.

Предлагается новый способ определения тенденций потока данных интернета, используя миповые секвенцные шаблоны. Анализируются основные этапы метода: обнаружение типовой последовательности в данных, изменяющихся во времени, их сравнение при выполнении анализа. Указываются возможные ограничения и потенциальные проблемы. Приведены экспериментальные данные и описаны направления предстоящих исследований. Ил. 6, библ. 9 (на английском языке; рефераты на английском, русском и литовском яз.).

J. Jeļinskis, G. Lauks. Interneto duomenų srauto tendencijų nustatymas naudojant tipines nuosekliąsias sekas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2009. – Nr. 5(93). – P. 3–6.

Siūlomas naujas būdas interneto duomenų srauto tendencijoms nustatyti naudojant tipines nuosekliąsias sekas. Analizuojami pagrindiniai metodo etapai – tipinės sekos aptikimas laikui bėgant kintančiuose duomenyse, jų palyginimas ir sutapatinimas atliekant nuosekliųjų sekų analizę. Nurodomi galimi apribojimai ir potencialios problemos. Pateikti eksperimentiniai duomenys ir aprašomos ateities tyrimų kryptys. Il. 6, bibl. 9 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).