University of
Helsinki
Department of
Computer Science

Петрозаводский
государственный университет
Кафедра информатики
и математического обеспечения

Ежегодный международный научный семинар

Vuosittainen kansainvälinen seminaari

Annual International Workshop

AMICT'2007

Передовые методы
информационных и коммуникационных технологий

Tieto- ja viestintätekniikan edistyksellisiä menetelmiä

Advances in Methods of Information and Communication Technology

Abstracts of the Talks

Contents

Experimental Study of Wireless Network Performance

Alexandr V. Borodin (Petrozavodsk State University, Russia, aborod@cs.karelia.ru)

In this investigation we have provided a series of network performance measurements under various workload scenarios. Experimental results were processed and estimations of network performance parameters were obtained. These results were compared with analytical ones.

Regeneration Cycles Dependence in Confidence Estimation by Splitting

Alexandra V. Borodina (Karelian Research Centre of Russian Academy of Science, Russia, musen@cs.karelia.ru)

The problem of confidence estimation of a high load probability γ in M/G/1 and GI/G/1 systems is considered.

In the previous research a special speed-up simulation technique based on combination of splitting and regenerative approach has been proposed for the consistent (point) estimation of the probability γ.

Since the queue-size process in a M/G/1 system (and the workload process is GI/G/1 queue) is non-Markovian, then a dependence between regeneration cycles obtained by splitting appears.

It follows from the splitting procedure, that the regeneration cycles turns out to be at most D-dependent, where constant D is defined as $D := R_0 R_1 \dots R_M$ and R_i is the number of split trajectories at the i-th threshold. (M is the number of thresholds.)

Therefore, we may use a Central Limit Theorem for D-dependable random variables to construct confidence interval for γ. Furthermore, the dependence of on the shape (and width) of the confidence interval on constant D is investigated.

Support Vector Machine in the Task of Voice Activity Detection

Evgenia Chernenko (University of Joensuu, Finland, echernen@cs.joensuu.fi)
Tomi Kinnunen (Institute for Infocomm Research, Singapore, ktomi@i2r.a-star.edu.sg)
Marko Tuononen (University of Joensuu, Finland, mtuonon@cs.joensuu.fi)
Pasi Franti (University of Joensuu, Finland, franti@cs.joensuu.fi)
Haizhou Li (Institute for Infocomm Research, Singapore, hli@i2r.a-star.edu.sg)

We define voice activity detection (VAD) as a binary classification problem and solve it using the support vector machine (SVM). Challenges in SVM-based approach include selection of representative training segments, selection of features, normalization of the features, and post-processing of the frame-level decisions. We propose to construct SVM-VAD system using MFCC features because they capture the most relevant information of speech, and they are widely used in speech and speaker recognition making the proposed method easy to integrate with existing applications. Practical usability is our driving motivation: the proposed SVM-VAD should be easily adapted into new conditions.

Voice activity detection (VAD) aims at classifying a given sound frame as a speech or non-speech. It is needed as a front-end component in voice-based applications such as speech recognition, speech enhancement, variable frame-rate speech coding, and speaker recognition. Furthermore, VAD is an important tool for a forensic analyst to locate the speech-only parts from large audio collections which can consists of tens of hours of data [1]. A large number of methods have been proposed. Simple methods are based on comparing the frame energy, zero crossing rate, periodicity measure, or spectral entropy with a detection threshold to make the speech/non-speech decision. More advanced models include statistical hypothesis testing [2], long-term spectral divergence measure [3, 4], amplitude probability distribution [5], and low-variance spectrum estimation [6]. The common property in these methods is that they include estimation of the background noise levels and/or noise suppression as a part of the process. The methods usually have a large number of control parameters, which are more or less tuned to a specific application. As an example, in [1] it was reported that the accuracy of the long-term spectral divergence VAD [3] depends much on the selection of the seven control parameters of the method.

The idea of this work is to extract the standard mel-frequency cepstral coefficients (MFCC) with delta and double delta coefficients and train a binary classifier using training files with speech/non-speech annotation. The VAD then labels each test utterance frame by using the trained classifier. We use the support vector machine (SVM) as the classifier since this has shown excellent performance in other classification tasks, e.g. speaker verification [7]. An advantage of this supervised learning is that it can be easily adapted to new operating conditions by providing representative training examples for the new condition. In this way, optimization of the parameters is absorbed to the training algorithm of the SVM whereas optimizing the parameters of conventional VADs, on the other hand, is more difficult.

In our experiments, we use three datasets with a varying degree of difficulty. The first dataset is a subset of the NIST2005 speaker recognition evaluation corpus, consisting of conversational telephone-quality speech having a sampling rate of 8 kHz. We selected 15 files for our purposes, all from different speakers and having duration of 5 minutes per file. The second data set consists of timetable system dialogues recorded in 8 kHz sampling rate. The material consists of human speech commands that are mainly very short, and synthesized speech that provides rather long explanations about bus schedules. Finally, the third data set consists of a one long continuous recording from the lounge of our laboratory in 44.1 kHz. The goal of the material was to simulate wiretapping material collected by the detectives.

We compare the proposed method with existing ones based on energy levels, long-term spectral information, and Gaussian mixture modeling, and provide comparative results on three described datasets. The method works excellently when small false speech acceptance rate is desired, which is the case in text-independent speaker verification, for example. Main advantage of the SVM-based VAD is that it works consistently in the same manner with different corpora. The other methods were more prone to the change of data set and variations of their parameters. Our main conclusion is that, according to our experiments, SVM is easier to adapt to the new data sets than conventional methods as long as we have a short training audio sample from the recording environment.

References

[1] M. Tuononen, R. González Hautamäki, P. Fränti, "Applicability and Performance Evaluation of Voice Activity Detection", submitted to IEEE Trans. on Information Forensic and Security.

[2] J.-H. Chang, N.S. Kim and S.K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models", IEEE Trans. Signal Processing, 54(6), June 2006, pp. 1965–1976.

[3] J. Ramirez, J.C Segura, C. Benitez, A. de la Torre, A. Rubio (2004), "Efficient voice activity detection algorithms using long-term speech information". Speech Comm. 42, pp. 271–287.

[4] J. Ramirez, P. Yelamos, J.M. Gorriz, J.C. Segura (2006), "SVM-based speech endpoint detection using contextual speech features". Elec. Letters 42(7), 2006.

[5] S.G. Tanyer and H. Özer, "Voice Activity Detection in Nonstationary Noise". IEEE Trans. Speech and Audio Processing, 8(4), July 2000.

[6] A. Davis, S. Nordholm, R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold", 14(2), March 2006.

[7] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition", Computer Speech and Language 20(2-3), pp. 210–229, April 2006.

Adaptive Content Management in Structured P2P Communities

Jussi Kangasharju (University of Helsinki, Finland,

Jussi.Kangasharju@cs.helsinki.fi)

A fundamental paradigm in P2P is that of a large community of intermittently-connected nodes that cooperate to share files. Because nodes are intermittently connected, the P2P community must replicate and replace files as a function of their popularity to achieve satisfactory performance. We develop a suite of distributed, adaptive algorithms for replicating and replacing content in a P2P community. We do this for structured P2P communities, in which a distributed hash table (DHT) overlay is available for locating the node responsible for a key. In particular, we develop the Top-K MFR replication and replacement algorithm, which can be layered on top of a DHT overlay, and in addition adaptively converges to a nearly-optimal replication profile. Furthermore, we evaluate the file transfer load caused by the adaptive algorithms on each peer, and present two approaches for achieving a better load balance. Our evaluation shows that with our two algorithms, an arbitrary load distribution is possible, hence allowing each peer to serve requests at the rate it wishes.

Improving Speaker Verification by Periodicity Based Voice Activity Detection

Ville Hautamaki (University of Joensuu, Finland, villeh@cs.joensuu.fi)
Marko Tuononen (University of Joensuu, Finland, mtuonon@cs.joensuu.fi)
Tuija Niemi-Laitinen (National Bureau of Investigation, Finland, tuija.niemi-laitinen@poliisi.fi)
Pasi Franti (University of Joensuu, Finland, franti@cs.joensuu.fi)

Voice Activity Detection (VAD) aims at classifying a given sound frame as a speech or non-speech. It is often used as a front-end component in voice-based applications such as automatic speech recognition, speech enhancement and voice biometric. A common property of these applications is that only human sounds (typically only speech) are of interest to the system, and it should therefore be separated from the background. Although VAD is widely needed and reasonably welldefined, existing solutions do not satisfy all the user requirements. The methods either must be trained for the particular application and conditions, or they can be highly unreliable when the conditions change. Demands for working solutions, however, are frequently requested by practitioners.

Traditionally, VAD research has been driven by telecommunications and voice-coding applications, in which VAD has to operate with as small delay as possible and all the speech frames should be detected. Typically, these voice activity detectors work by modeling the noise signal statistics. Initial noise estimates are usually obtained from the beginning of the signal, which is updated as VAD makes non-speech decisions.

Even though realtime VAD for telecommunications application is maybe the most common, other applications also exist. For those applications, design criteria are usually different than telecom VAD.s. In speech segmentation application, realtime operation is not as important as the goal is to process the input speech file (usually hours or even days long) in a background process and then the output will be used for the retrieval or the diarization tasks. Segmentation of the forensic wiretapping is especially difficult task is. In those recordings, no close talking microphone is used and the noise does not possess properties assumed in the telecom VAD.s (long-term stationarity).

In this study, our primary goal is to use VAD as a preprocessor for the realtime speaker verification. It is clear that including non-speech frames in the modeling process would bias the resulting model, especially if the number of non-speech frame is significant. It is also known that not all speech frames have equal discriminative power as it is well known that voiced phonemes are more discriminative than unvoiced phonemes,and therefore, it can be beneficial to drop out unvoiced frames to increase recognition. This is in contrast to telecom or speech recognition application where speech should be accurately detected.

In realtime speaker verification, we need to start processing speech frames with as small delay as possible. New recorded speech frame is pushed to signal processing subsystem, and to scoring subsystem while subject keeps speaking to the microphone. Not only is this useful feature in the standard applications, but it is essential in low power mobile phones, where non-realtime realtime processing is not even usable. Good results in speaker verification can be achived by a simple energy based VAD, but the method needs analyze the whole utterance before it can start to make speech/non-speech decisions.

Our proposed system is based on the detecting the periodicity of the given frame. In contrast to the telecom voice activity detectors we do not model noise, but we base our decisions on the feature that is known to be short term stationary. Performance of the proposed method is compared against two existing methods: a realtime method based on long-term spectral divergence (LTSD) and a simple energy based method, which needs two passes on the data. Summary of the speaker verification results with the corresponding VAD thresholds is shown in Table 1. Best parameters found in tuning with NIST 2001 corpus have been used for the NIST 2006 experiments. We obtained significantly higher error rates for NIST 2006 than NIST 2001, as expected. The periodicity-based method clearly outperforms the other realtime method (LTSD), and performs comparably with the energy-based method when applied for NIST 2001 and 2006 speaker recognition evaluation corpora. The method is also tested for segmenting surveillance recordings, voice dialog and forensic applications.

Table 1: Summary of speaker verification results (% EER).

	NIST 2001				NIST 2006
	model 512		model 64		model 512
	EER	Thr	EER	Thr	EER
No VAD	14	-	16	-	44
LTSD	12	40	14	45	36
Energy	9	36	10	35	17
Periodicity	8	0.61	10	0.66	17

Sparse Networks: Balance of Processing and Communication

Ville Leppänen (University of Turku, Finland, ville.leppanen@it.utu.fi)

Martti Penttonen (University of Kuopio, Finland, penttonen@cs.uku.fi)

In this talk we discuss the balance of processing and communication, in particular the need of bandwidth, and emphasize the concept of *sparseness*.

It was pointed out by Vitányi that in 3-dimensional world the diameter ϕ of any network is $\Omega(\sqrt[3]{n})$, where n is the number of nodes, routers or processors. Hence, if processors are allowed to send each others messages without target or frequency restriction, the bandwidth per processor must be $\Omega(\phi)$. If p is the number of processors and each of the n nodes is connected with a constant number d of nodes, by condition $p\phi \in O(dn)$ we see that not all nodes can be processors. An interesting case is s-sided cube, where $n = s^3, p = s^2, \phi = 3s, d = 3$. This is called a *sparse mesh* or *torus*, depending whether circular paths are allowed. In this kind of a structure, it is possible that all processors send and receive at every step and the routing time is still $O(1)$ per packet.

It may seem that sparseness is waste — in our example $s^3 - s^2$ nodes are routers and only s^2 nodes are processors. That is the cost of the bandwidth — they are necessary if the communication of the processors is not restricted. Denser networks may work, if communication is restricted, for example if

- communication is sparse. If only every kth step of the processor is communication, the condition becomes $p\phi/k \in O(dn)$.
- communication is local. In this case $\Sigma_i p_i \phi_i \in O(dn)$, where ϕ_i is the routing distance. If for example, 1/2 of communication is to processor itself, 1/4 to neighbor, etc, the condition becomes $p \in O(dn)$. So, average routing time can be $O(1)$ per packet even if every node is a processor.

Examples of sparse networks are

- *sparse meshes*. We have studied *coated* meshes, where a cube of routers is coated with processors, and *sparse tori*, where on any path of an s-sided d-dimensional torus every sth node is a processor.
- *sparse butterflies*, *sparse fat trees* etc. In SBPRAM (Saarbrücken PRAM) n processors are connected by a $\log n$ layered butterfly of router nodes.

Also dynamic networks must be sparse, unless communication is sparse. In principle, sparseness could be implemented by embedding in one of the above mentioned structures. For example, Gupta and Kumar (1999) say that large scale wireless networks can at best have $\Theta(1/\sqrt{n})$ throughput per node. Hence, to provide constant throughput per node, wireless (sensor) networks of n processor nodes should be made sparse by adding n^2 transmission media.

GIS Technology in Forest Harvesting Planning

Viktor M. Lukashevich (Petrozavodsk State University, Russia, lvm-dov@mail.ru)

Ludmila V. Shchegoleva (Petrozavodsk State University, Russia, schegoleva@psu.karelia.ru)

Pavel O. Schukin (Petrozavodsk State University, Russia)

In this talk the use of GIS technologies for decision-making for the organization of spadework on timber cuttings, a choice of the complete set of machines and definition of road-building actions is considered.

Equilibrium in a P2P system

Vladimir V. Mazalov (Karelian Research Centre of Russian Academy of Science, Russia, vmazalov@krc.karelia.ru)

Igor A. Falko (Karelian Research Centre of Russian Academy of Science, Russia, ifalco@krc.karelia.ru)

Andrei V. Gurtov (Helsinki Institute for Information Technology, Finland, gurtov@hiit.fi)

Andrey A. Pechnikov (Karelian Research Centre of Russian Academy of Science, Russia, pechnikov@krc.karelia.ru)

We consider P2P-system with two kinds of users: ”sponsors” and ”free-riders”. Sponsors give information in random manner with known distribution and free-riders have to maximize incomes trying to guess sponsor's behavior. Equilibrium is found in such a system.

Heavy-tailed distributions with applications to broadband communication systems traffic

Evsey V. Morozov (Karelian Research Centre of Russian Academy of Science, Russia, emorozov@karelia.ru)

Michele Pagano (University of Pisa, Italy, m.pagano@iet.unipi.it)

Alexandr S. Rumyantsev (Petrozavodsk State University, Russia, _ar@list.ru)

In the 90's statistical studies of high-resolution traffic measurements in different communication networks scenarios (from LANs to WANs) have highlighted properties of long memory of actual traffic flows. Later the asymptotic self-similarity characteristic of measured aggregated traffic has been explained in terms of heavy-tailed, infinite variance phenomena at the level of individual network connections and of the interactions between the human user and the network.

These experimental evidences, in strong contrast with conventional Markovian models and light-tailed distributions, have a deep impact on analytical techniques as well as on simulation of real networks. For instance, one of the basic results is that if network traffic components have infinite variance (e.g. transmission time) then the corresponding limit variable has infinite expectation, that implies a drastic change of the simulation methodology (as far as sample generation and confidence interval estimation are concerned).

In spite of a great progress in this new area, many problems are still open and continue to attract attention of researchers.

In this work, we focus on the basic properties of heavy-tailed distributions and discuss their applicability to describe some important new features of the network traffic. Moreover, we also consider the related long-range dependence property of the network traffic.

Most of the work is a survey of the existing literature on this topic, and an important goal is to present the material in a unified way, taking into account the most recent contributions on international journals, suitable for advanced study on the general topic of network performance. Also it is assumed to present a comparison between different methods of estimation of Hurst parameter and their application to real-life network traffic.

Service Migration Using Virtualization

Tiina Niklander (University of Helsinki, Finland,
Tiina.Niklander@cs.helsinki.fi)

With the current distributed world and increasing number of clustered environment, it has become more and more important to be able to move running services from one machine to another. Virtualization provides a feasible mechanism to break the connection between the running process and the hardware. Placing the process in virtual environment makes it possible to give a virtual view to the process of its environment. This view can them be mapped to different actual resources (memory, names, identifiers) in different machines.

Virtualization can be done either on process level or on the system level. When virtualizing the environment for the process. The virtual container only has the process (or process group). This container can be migrated to a machine with similar hardware and operating system. If the operating system must also be allowed to change, then the virtualization is done on the system level and the virtual machine contains both the running processes and the operating system serving them. In this case the whole virtual machine needs to be migrated.

Benchmarking Telecommunication Systems

Kimmo Raatikainen (University of Helsinki, Finland,
Kimmo.Raatikainen@cs.helsinki.fi)

The world is full of benchmarks. In January 2006 Google provided 2.9 million hits on phrase ”computer performance benchmark”. IEEE Xplore found 426 articles published since 1.1.2000 in that category. ACM Digital Library had 11,519 entries in that category. In this paper we review some existing benchmarks that are useful in analyzing performance of telecommunication sytems. The focus is on networking and database benchmarks, but also other benchmarks are briefly summarized.

Telecommunication systems of today are quite complicated networked and distributed computing and communication environments. A telecommunication system involves networking devices and protocols, databases, network management and control software, operations support systems, accounting and charging facilities, billing systems and various kinds on servers. It is evident that a single benchmark cannot give enough insight to understand performance of a telecommunication system.

In telecommunications we divide the functionality into user (or data), control, and management planes. The user plane concerns transmitting data between the end-points. The control plane is involved in establishing, maintaining, and tearing down connectivity between the end-points. The management plane covers management operations of the systems. Efficiency of packet handling is the primary factor in performance of user plane functionality. On the control-plane we have timeliness requirements of setting-up a service session, preparing accounting and charging, executing handoff, reserving and releasing resources, etc. The performance of control-plane operations is affected by databases, authentication, communication between network elemnts, among others. Although processing efficiency is important, performance is not so critical on the management plane. The main concerns include correctness, reliability, security, and availability.

Application of Modern Information Technologies to the Study of Karelian Petroglyphs

Ksenia A. Rogova (Petrozavodsk State University, Russia,
rogova@cs.karelia.ru)
Konstantin N. Spiridonov (Petrozavodsk State University, Russia,
spiridonov@psu.karelia.ru)
Maksim U. Bystrov (Petrozavodsk State University, Russia,
bystrov@cs.karelia.ru)

Modern informational computer technologies allow not only to carry out scientific researches on a new level but also to put into practice the results of the research which we could not get before. Computer visualization of historical materials gives new unique opportunities for researches. As an example, let us take a look at a program system called ”Information retrieval system of Karelian petroglyphs” or ”PIRS”. Development of this system is supported with a grant RGNF no. 05-01-12118v (manager N. V. Lobanova).

Computer support for data base is divided into two programming modules according to user’s functions (for simple users and for professionals). Programming module for simple users is a form of website and it can be seen in Internet by http://smalt.karelia.ru/~petroglyphs. The program for specialists includes data base, searching system according to features and characteristics, catalogue of the petroglyphs, information about the program. Data base is the most important part of the information retrieval system. Nowadays it is the only data base of petroglyphs, their characteristics and descriptions. It includes such functions as searching, adding and changing information. This base is necessary for specialists working in the field of scientific researches. Searching systems allow a user to find similar petroglyphs or petroglyphs with some definite features. In addition this module includes various algorithms, which implement mathematical methods for petroglyphs’ analysis such as cluster analysis, method correlation galaxies and multi fractal's image parametrization. For example, multi fractals image characteristics allows to define an invariant of a petroglyph which may define the sequence of petroglyph's appearance on a rock in common parts. On top of that this method enables to fulfill a binary segmentation of petroglyph's images.

Multidimensional Index Structures for NetFlow Record Processing

Alexander V. Sherikov (Petrozavodsk State University, Russia, sherikov@cs.karelia.ru)

Yury A. Bogoyavlensky (Petrozavodsk State University, Russia, ybgv@cs.karelia.ru)

The talk is devoted to multidimensional index structures in context of NetFlow record processing. We consider records to be points in multidimensional space. We investigated appropriate index structures. Their classification, main properties and problems are stated in first chapter of article. Second one contains detailed overview of A-tree structure and explains why it is the most interesting for us. Finally we convey our modifications of A-tree, that were proposed to suit our needs better. Innovations of our work consist in using index structures to process NetFlow records and adaptation of A-tree for this task.

Search Problem of the Reels Cutting Optimal Sequence

Alexey Smoliy (Petrozavodsk State University, Russia, alexey.smoliy@metsopartners.com)

The primary production of pulp and paper industry factories is paper making. In accordance with process flowsheet, paper web, the length of which is limited from above and below, is wound around iron spools. Further, that paper web (is called also reel) is cut on winder (machine which is used to cut the paper/board on a reel) into great number of rolls with certain lengths of generatrixes are called formats. The number of rolls, which are born at the same time as the result of unwinding and cutting paper web, is called set of rolls. The important characteristic of all rolls in one set is its radius (length). Reel length should correspond to the all sets, which are forming its cutting. Such reels are so called standart reels. If sets are identical inside one reel, then reel length is multiple of set length.

Not all produced reels could be considered as standart reels, because paper web break could happen during production process, otherwise paper could have bad quality or damages. Such reels are called non-standart reels, after cutting which some paper is left on the spool. For the purpose of paper waste minimization and increasing the efficiency of production that paper web remainders are spliced together in such a way that obtained paper web could be used in the production.

The initial data for the task are information about sets to produce, as well as about available standart and non-standart reels. Let us to consider that sequence of spliced reels is one logical reel. After cutting such long logical reel it is necessary to take into account locations of splicing points that are formed as the result of paper web splices. Ideally it would coincide with the points of sets superpositions. For some sets, splices are allowed at the beginning, in the middle and at the end of roll.

The problem is concluded into synchronization of two flows, reels and products (sets), with taking into account the limitations for splices locations. The complexity of synchronization with minimum waste formation is that during generation of permutations, it is necessary to consider specific characteristics of the production line and equipment, including well-defined sequence of paper processing machines for one separately taken reel, limitations for possible reel dimensions, which can be processed on that machines, handling of parallel machines and others.

Inductive Constructed a Set of PROSPECTOR-like Rules for Classification of Literary Works

Tatyana G. Surovtsova (Petrozavodsk State University, Russia, tsurovceva@psu.karelia.ru)

Sergey P. Chistjakov (Karelian Research Centre of Russian Academy of Science, Russia, chistiakov@krc.karelia.ru)

The method based on inductive constructed a set of PROSPECTOR-like rules we used for classification of literary works. Description of syntactic and morphological features of texts according to grammar of Russian calculated by means of expert system, an integrated component of information retrieval system ”SMALT”.

Joint analysis of Squid and Netflow log files using http client port information

Alexandr S. Volkov (Petrozavodsk State University, Russia, avolkov@cs.karelia.ru)

Yury A. Bogoyavlensky (Petrozavodsk State University, Russia, ybgv@cs.karelia.ru)

Recently, some authors pay attention to joint analysis of traffic data (log files) from different levels of Internet protocols. This kind of analysis allows to obtain new information about network and to raise the accuracy of traffic analysis algorithms due to additional information obtained from compound traffic data.

This report contains key features of algorithm for joint analysis of Netflow and access.log log file of proxy server Squid. The proposed approach for joint analysis provide next new features:

- Netflow data aggregation by uniting all flows corresponding to one primary web request;
- essentially higher accuracy of router's web-based workload characterization;
- higher accuracy methods for cache performance quality analysis;
- interrelation between primary users' http-requests and secondary requests for objects.

On the whole, the proposed approach allows exposing and analyzing mechanisms of correlations between working processes of network and application levels of Internet.

Since any Netflow can be identified by next seven attributes: IP source address, IP destination address, Source port, Destination port, Layer 3 protocol type, Class of Service, Router or switch interface; the presence of respective data from Squid's log files is an important requirement for univocal comparison of traffic data. We need to draw attention to obligatory presence of client's port numbers from which clients request objects (there are no such data in native format of access.log log file). Otherwise, we need to solve the problem of distribution a number of Netflows between several log file's records. In some cases, it is possible to obtain correspondences between some Netflows and records of access.log without any information about client's port, for example, when computing estimation for mistiming between router's and proxy server's system clocks.

According to our pilot analysis of traffic data, there are mistiming between router's and proxy server's system clocks. We use an algorithm based on rules for expiring NetFlow cache entries. Using these rules we unambiguously determine correspondences between certain Netflows and access.log's records. This pairs of flows and records allow us to evaluate system time divergences (an estimate of mistiming) in certain points of time. As we learned, there was a linear decrease of divergence's estimates on the tested set of traffic data. So, we plotted functional time dependence for divergences of system clocks under reviewed period of time using the mean-square approximation.

Using access.log log file and Netflow data, and also being aware of system clock's divergences at any moment, we can unambiguously find basic Netflows (Netflows corresponding to data transfer between a client and proxy server) corresponding to any access.log's items. Being founded on access.log data, we can obviously find all collateral Netflows which correspond to retrievals and loads of a requested object by proxy server. The number of generated Netflows is within the limits of 2 and $2m + 2n + 6$ and depends on router's, local proxy server's and dns-server's settings, depends on network topology and on local and sibling proxy servers' caches' conditions (m and n is the numbers of icq- and dns-requests from local proxy and dns-servers to sibling proxies and official dns-servers correspondingly, which passing through the router).

There are some cases of data comparison then one Netflow corresponds to several access.log's records, and vice versa, several Netflows correspond to one record. The first one is possible because of using HTTP/1.1 protocol, which allow to request and load several objects trough one TCP connection. With all this going on this TCP connection is represented only by two Netflows. The second case, then some Netflows, for example then transmitting data from client to proxy (primary http-request), correspond to one access.log's record, is possible because of rules for expiring NetFlow cache entries.

Because of stable connections using HTTP/1.1 some of the data flows are presented only by couple of Netflows, and logically we can joint such access.log's records into more common units of traffic - web seance flows. Generally web seance flow is a number of access.log's records and Netflows corresponding to one full html page including all inner objects' requests and loads. Using such traffic units we can look differently at the web user's activity. Hereafter we plan to use even more common units for traffic except web seance flows - web session flows. Web seance flow is a number of web seance flows (for fixed client) that are very close to each over in time. There should be long pauses between such two groups of web seance flows, which actually enable us to divide the consecution of this web seance flows to web session flows. Logically, one web session flow presents user's activity of searching and processing some piece of information at that long pauses between two web session flows correspond to the process of pondering over founded (loaded) information (reading text for example).

The final aim of the work is raw Netflow and access.log's data transformation into mentioned units of web traffic, and subsequent logical analysis of obtained data.

A New Regenerative Estimator for Effective Bandwidth Prediction

Irina S. Vorobieva (Petrozavodsk State University, Russia, vesnik@sampo.ru)
Evsey V. Morozov (Karelian Research Centre of Russian Academy of Science, Russia, emorozov@karelia.ru)
Michele Pagano (University of Pisa, Italy, m.pagano@iet.unipi.it)
Gregorio Procissi (University of Pisa, Italy, g.procissi@iet.unipi.it)

The main purpose of the effective bandwidth (EB) theory is to guarantee a required Quality of Service (QoS) (a performance guarantee) for a wide class of communication networks (for instance, ATM, Ethernet). At that performance guarantee has a loss-ratio form, being a fraction of the lost arrivals (it is so-called overflow probability). It follows from the Large Deviation Theory (LDT) that an overflow probability (rare event probability) for a wide class of the buffered systems decreases exponentially fast as buffer size increases. Moreover, the LDT describes a rate-function, which in turn allows to calculate the required exponent. Thus, calculation of EB is reduced to calculation of the scaled cumulant generating function (SCGF) of the arrival process.

In a recent research, an approximation of the SCGF (to find an overflow probability) has been developed relaying on an assumption that input data form a stationary, mixing sequence. At that, a partition of the given input sequence onto blocks of a fixed size B is proposed to construct a sample mean estimate of SCGF. In other words, the batch-mean method is used for the estimation which assumes that the blocks constitute i.i.d variables, if size B is large enough.

In this work, instead we present a refined EB approximation which is based on a regenerative structure of the input sequence. We consider a tandem network with two single server stations, renewal input to 1st station, and a constant service rate at the second station. It follows that the 2nd station is fed by a regenerative input (the output from 1st station) and we can construct classical regenerations of the second station which occur when an arriving customer sees totally empty network. This allows us to use partition of the input sequence on the i.i.d.

blocks of random length (instead of fixed length B above) which coincide with the boundaries of regenerative cycles of the input.

Using regenerative simulation we calculate the effective (constant) service rate s as a function of given QoS probability Γ_0 and the buffer-size b. It is assumed to discuss the quality of new estimator and verify the effectiveness of the two different approximations (with fixed and random block lengths, respectively) comparing results with a Crude Monte-Carlo simulation.