ISSN 1727–6209

RESEARCH INSTITUTE FOR INTELLIGENT COMPUTER SYSTEMS
TERNOPIL NATIONAL ECONOMIC UNIVERSITY
and
V.M. GLUSHKOV INSTITUTE FOR CYBERNETICS,
NATIONAL ACADEMY OF SCIENCES, UKRAINE

International Journal of

Computing

Published since 2002

September 2013, Volume 12, Issue 3

Ternopil – 2013

International Journal of Computing

September 2013, Vol. 12, Issue 3

Research Institute for Intelligent Computer Systems, Ternopil National Economic University and V. M. Glushkov Institute for Cybernetics, National Academy of Sciences, Ukraine

Editorial Council

Address of the Journal

Research Institute for Intelligent Computer Systems
Ternopil National Economic University
3, Peremoga Square
Ternopil, 46020, Ukraine

Phone: +380 (352) 47-5050 ext. 12234
Fax: +380 (352) 47-5053 (24 hours)
computing@computingonline.net
computer.journal@gmail.com
www.computingonline.net

ISSN 1727–6209

НДІ ІНТЕЛЕКТУАЛЬНИХ КОМП'ЮТЕРНИХ СИСТЕМ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
і
ІНСТИТУТ КІБЕРНЕТИКИ ІМ. В.М. ГЛУШКОВА,
НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ

Міжнародний науковий журнал

Комп'ютинг

Видається з 2002 року

Вересень 2013, Том 12, Випуск 3

Постановою президії ВАК України № 1-05/3 від 14 квітня 2010 року міжнародний науковий журнал Комп'ютинг віднесено до переліку наукових фахових видань України, в яких можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора і кандидата технічних наук

Тернопіль – 2013

Міжнародний науковий журнал Комп'ютинг

Вересень 2013, том 12, випуск 3

НДІ інтелектуальних комп'ютерних систем, Тернопільський національний економічний університет і Інститут кібернетики ім. В. М. Глушкова, Національна Академія наук України

Зареєстрований Міністерством юстиції України. Свідоцтво про державну реєстрацію друкованого засобу масової інформації – серія КВ №17050-5820ПР від 15.07.2010.

Друкується за постановою вченої ради ТНЕУ, протокол № 1 від 30 серпня 2013 року

Редакційна рада

Адреса журналу

НДІ інтелектуальних комп'ютерних систем
Тернопільський національний економічний університет
площа Перемоги, 3
Тернопіль, 46020, Україна

Тел.: 0 (352) 47-50-50 внутр. 12234
Факс: 0 (352) 47-50-53
computing@computingonline.net
computer.journal@gmail.com
www.computingonline.net

CONTENTS

ЗМІСТ

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

THE CHALLENGE OF MANAGING AND ANALYZING BIG DATA

Hermann Heßling

Berlin University of Applied Sciences (HTW)
Wilhelminenhofstr. 75, D-12459 Berlin, Germany
hessling@htw-berlin.de

Abstract: The amounts of data produced in science are growing exponentially. Traditional methods for storing and maintaining the enormous flood of data seem to be no longer sufficient anymore. The complexity of the data that will be distributed more and more worldwide, is going to constitute a considerable challenge for their analysis. According to Alex Szalay there soon will be produced so many data that they cannot even be stored and maintained anymore. The data have to be analyzed in real time in order to extract the relevant information. An outline of the project Large Scale Management and Analysis (LSDMA) is given. The status of our research group on distributed real-time computing is reviewed. Finally, a novel approach to time-dependent image processing based on local thermodynamical methods is presented.

Keywords: Big Data, Real-time computing, Time-dependent image processing.

1. INTRODUCTION

Experiment and theory are pillars for scientific discoveries. With the emergence of computers new insights originated in simulations. In the meantime, data-intensive computing is considered as the fourth pillar for scientific discoveries [1].

In many scientific areas the resolution power of experimental devices is increasing strongly leading to a steadily increasing flood of data. The Large Hadron Collider (LHC) at CERN is used to explore the realm of elementary particles. The produced data rate of the order of 1 PB/s is reduced considerably in the subsequent workflow. Finally, of the order 25 PB are stored per year. The data are distributed worldwide and can be analyzed by thousands of scientists using the world largest grid computing system [2]. The radio telescope Square Kilometre Array (SKA) will allow to explore the Universe with thousands of antennas located in South Africa and Australia. First experimental results are expected in 2019. When finished in 2023 SKA will have to store of the order of 1 Exabytes per day [3].

"Soon we cannot even store the incoming data stream" (Alex Szalay [4]). Traditional approaches for analyzing and maintaining data seem to be no longer sufficient anymore. New ideas and tools are needed.

Data-intensive research is sometimes under attack. Nobel laureate Sidney Brenner criticizes most biology as low-input, high-throughput, no-output biology [5]. The LHC discovery of the Higgs particle showed what Big Data is capable of.

2. LSDMA

In the joint research and development (R&D) project "Large Scale Data Management and Analysis" (LSDMA) several Helmholtz centers and German universities are supporting scientists in analyzing, maintaining, and archiving their data [6]. Developing a one-size-fits-all solution is not feasible because of the heterogeneous requirements of the scientific communities. Therefore, the participating communities are divided into five "Data Life Cycle Labs" (DLCL).

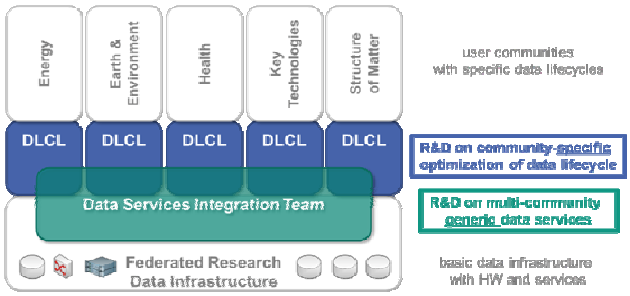


Fig. 1 – Organization of the LSDMA project.

The general goals of the DLCLs are supporting scientists in
- organizing their data and metadata,

- establishing easy access and use of local, national and international infrastructures for data storage, data processing and data archiving, and
- standardizing data management techniques in the scientific communities and their data life cycles.

The Data Services Integration Team (DSIT) develops generic services based on requests from the DLCLs:

- Federated identity management
- Federated data access
- Meta data catalogue
- Archive services
- Monitoring, modeling, optimization
- Data intensive computing

Due to lack of space it is not possible to review all LSDMA activities. The following selection is fairly subjective.

Based on the results of a DSIT workshop on AAI (Authentication, Authorization, Infrastructure) [7] it was decided to develop a distributed identity and authorization service based on Shibboleth.

dCache is a system for storing and retrieving huge amounts of data [8]. It manages approx. 50 % of the LHC data. Several access protocols are supported, e.g. GridFTP, NFS 4.1 (pNFS) and WebDAV. Cloud computing may provide a more flexible access to scientific data. For that reason, dCache is extended by the Cloud Data Management Interface (CDMI) [9]. As a cloud application an online storage is created called dBoX, to be used by the scientific community at DESY. The data are stored at DESY Hamburg and, thus, are subject to German law. The authentication is based on X.509 certificates and user/password. dBoX can be accessed via mobile devices [10].

A lot of research in the DLCLs is based on image processing. For example, the embryogenesis of vertebrates is studied with microscopes that are able to record 3D images with high spatial and temporal resolution. A tracking of the evolution of individual cells requires high data acquisition rates. By developing time-efficient algorithms and using high parallelization, data sets of the order of 10 TB from a single probe can be processed in less than a day [11], see Fig. 2.

3. DISTRIBUTED REAL-TIME COMPUTING

Data reduction in real-time will become increasingly important. For example, in photon science ultrashort X-ray flashes are used to explore nanostructures of probes in material sciences, chemistry and biology. Due to principal limitations in the experimental setup only a few percent of the data are of ~~a~~ sufficient quality for a later analysis [12]. In other words, only a fraction of the data should be stored and it is advantageous to identify useless data as early as possible. A natural strategy is to parallelize data analysis tools. In photon science diffraction images are taken, i.e. the relevant information may be smeared over the whole image. Consequently, parallelized image processing has to rely on the exchange of intermediate results. However, an efficient exchange of intermediate results is a challenging task, as large parallel systems are usually built of subsystems connected by heterogeneous networks and the memory may be distributed between the subsystems.

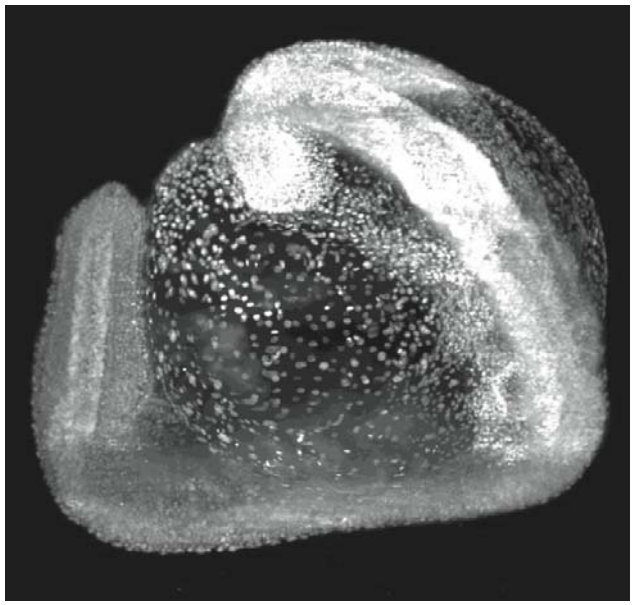


Fig. 2 – A Zebra fish embryo after 24 hours [11].

GriScha is a chess program running on distributed environments, e.g. a grid or cloud computing system [13]. It is an ideal test bed for exploring various aspects of distributed real-time computing. A Gatekeeper installs a pilot job on each participating worker node (WN), see Fig. 3. When running on a worker node the pilot job starts a chess engine client and, thereby, establishes a network connection to the external MasterNode. Worker nodes are protected by firewalls and cannot be accessed from the outside. Pilot jobs are used to overcome firewalls in that these do not block outbound connections, in general. The MasterNode distributes the game tree to the worker nodes which analyze their subtree for some period of time (default 15 s) and, then, send their best move back to the MasterNode. The MasterNode selects the best of all moves offered by the worker nodes. The chess engine on the worker nodes is very simple and evaluates a chess position like a beginner. No external intelligence is used, i.e. no data base of chess openings, no endgame tablespaces, no database of chess games. The essential question is:

can you make many beginners stronger than a master? The challenge is to establish a collective communication in huge distributed environments.

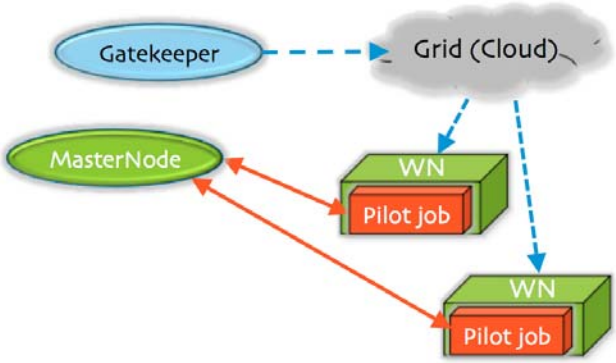


Fig. 3 – GriScha: chess in the grid.

The communication between the MasterNode and the worker nodes is based on SIMON, an extension of the RMI network protocol that allows a server to invoke methods on a client using the same socket connection [14].

In order to explore the capabilities of SIMON for exchanging large amounts of data, random message strings were send from a client over a LAN to a server. It turns out that there are jumps in the response time at some specific message lengths, see Fig. 4. A discontinuous response behavior is hardly acceptable in real-time computing.

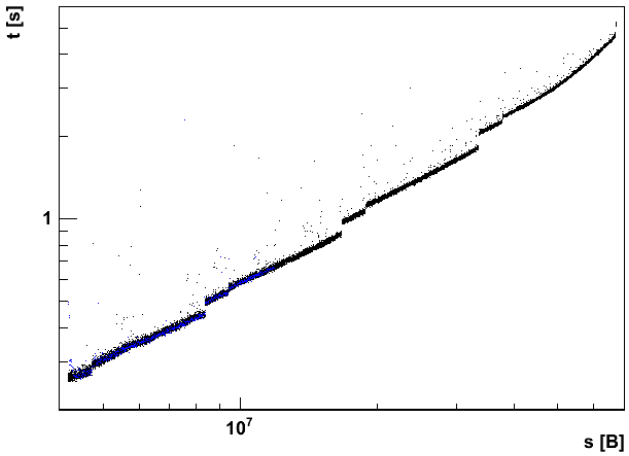


Fig. 4 – Message length s versus response time t.

The origin of the discontinuities is unclear. They depend neither on details of the network, operating system, versions of the Java virtual machine and SIMON [15], nor on the garbage collection in Java [16].

Communication tools on huge networks are based on a decentralized organization and use Peer-to-Peer (P2P) protocols. As an alternative to SIMON the open source P2P protocol JXTA was implemented in GriScha [17]. However, the future of JXTA is unclear as Oracle announced its withdrawal from this project.

After some moves a previously obtained game position may reappear again. To improve the playing strength of GriScha moves already analyzed should not be reanalyzed a second time. Instead, they should be stored in a shared memory accessible by worker nodes. When a worker node realizes that a certain move is noted it can cut out a subtree and analyze in more depth the remaining part of the game tree. JuxMem is a data-sharing grid service based on JXTA [18]. However, the future of JuxMem is also unclear. In Ref. [19] a design of a grid-compatible shared memory is suggested that is currently being realized.

The communication protocol XMPP is used to exchange messages in near real-time. The time needed for simultaneously sending messages from clients to a server is increasing linearly with the number of clients even for very short messages of only one Byte and is significantly longer compared to SIMON [20, 21]. Therefore, XMPP is of limited attractiveness for large distributed real-time systems.

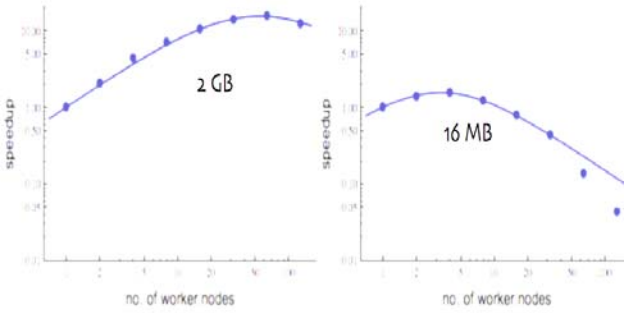


Fig. 5 – Speedup for message exchange: total message size N = 2 GB (left), N = 16 MB (right). The curve is a least-square fit to the data using Amdahl's law (1).

Speedup is an essential quantity in parallel computing. It is defined by the ratio $S(p)=T(1)/T(p)$ where $T(p)$ is the execution time of a tool running in parallel on p devices. According to Amdahl's law the speedup is limited by the sequential part of a tool since $T(p) = T_{seq} + T_{par}/p$. This formula has to be modified if latencies due to message exchange cannot be neglected,

$$T(p) = T_{seq} + T_{par}/p + (p-1)\, T_{msg}. \qquad (1)$$

Let us assume that the exchange of intermediate results needs a transport of N/n Bytes of data from n worker nodes to an external server. In Fig. 5 speedup measurement results for the grid system DECH with up to 128 worker nodes are shown. The worker nodes of DECH are located at research institutes and universities in Germany and Switzerland. The results

indicate that in the case of data exchange one should be aware of an upper bound for parallelization. Beyond this bound the system becomes inefficient as the speedup is going down.

4. TIME-DEPENDENT IMAGE PROCESSING

4D imaging is applied in various sectors. In medicine, for example, it is used to improve the success of a radiation therapy by aligning a radiation source along the motion of a tumor [22]. The upcoming European XFEL will produce up to 27,000 light pulses per second and new high-resolution detectors will open new research areas in photon science. It will be possible to take movies not only from chemical reactions but also from the motion of quantum objects like electrons [23].

In the following we present a novel method for identifying "regions of interest" in dynamically deformed objects that is based on statistical physics. For simplicity we concentrate on one-dimensional systems.

The left picture of Fig. 6 shows the motion of an object of initial length L. The object is compressed and after some time t it is decompressed. From the evolution of the "texture" of the object trajectories can be determined. In this case, the trajectories have the form

$$x_t = x_0 \cosh(gt) + p_o/(gm) \sinh(gt).$$

The momentum trajectories are given by $p_t = m\mathrm{d}x_t/\mathrm{d}t = p_o \cosh(gt) + mgx_0 \sinh(gt)$. The initial momentum is proportional to the initial position, $p_o = ax_o$.

The transformation defined by

$$X(x,p,t) = [x + t\, p/m]\cosh(gt) - [p/(gm) + gtx]\sinh(gt)$$
$$P(x,p,t) = p\cosh(gt) - mgx\sinh(gt).$$

is canonical.

Proof (sketch). The assertion follows from the fact that there is a generating function F(X,P,t) such that the Poincaré-Cartan 1-form

$$p\,\mathrm{d}x - H\,\mathrm{d}t = P\,\mathrm{d}X - P^2/(2m)\,\mathrm{d}t + \mathrm{d}F(X,P,t)$$

is fulfilled where $H = p^2/(2m) - (m/2)\, g^2 x^2$ is the Hamilton function.

From the Poincaré-Cartan 1-form follows that the trajectories with respect to the new coordinates are straight lines $X_t = x_0 + (p_0/m)\, t$.

The entropy of a 1-dim. ideal gas consisting of N particles enclosed in a volume V reads (W. Pauli)

$$S = k_B N\, [\, ½ \ln(E/N) - \ln(N/V) + s0\,]$$

where $E = ½\, N\, k_B\, T$ is the internal energy and T the temperature. For a local description we introduce the densities $s=S/V$, $n=N/V$, and $e=E/V$. Then the temperature

$$T(e,n) = (2/k_B)\; e/n$$

and the entropy density

$$s(e,n) = k_B\, n\, [\, ½ \ln(e/n) - \ln(n) + s0\,] \qquad (2)$$

can be represented in terms of the particle density n and energy density e.

The texture of an object is given by the initial gray values of an image of the object. We interpret the texture as a state function $\rho(x_0, p_0)$ describing the statistical distribution of the initial values x_0, p_0. The state function is positive, $\rho(x_0,p_0) \geq 0$, and normalized, $\int \rho(x_0, p_0)\, \mathrm{d}x_0\, \mathrm{d}p_0 = 1$. The state function of the left picture of Fig. 6 has the form

$$\rho(x_0, p_0) = f(x_0)\; \theta(x_0)\,\theta(L-x_0)\; \delta(p_o - a\, x_o)$$

where $\theta(x)$ is the Heaviside step function and $\delta(x)$ the Dirac function. The shape of the texture is given by

$$f(x_0) = [(3/2)^4 + x_0(L-x_0)\,(x_0 - L/2)^2]/[L(3/2)^4 + L^5/120].$$

In statistical physics the time evolution of the mean value of an observable $A(x,p,t)$ can be determined in the Heisenberg picture

$$\langle A_t \rangle = \int A(x_t,p_t,t)\; \rho(x_0, p_0)\; \mathrm{d}x_0\; \mathrm{d}p_0.$$

Consider the point-like observables

$$n_{x^*}(x) = \delta(x - x^*),\; e_{x^*}(x,p,t) = 1/(2m) P^2(x,p,t)\, \delta(x - x^*).$$

Their expectation values $e_{x^*,t}$ and $n_{x^*,t}$ are inserted into Eq. (2) in order to obtain the local entropy $s_{x^*,t} = s(e_{x^*,t}\,,\, n_{x^*,t}\,)$.

The two local maxima of the particle density follow the trajectories and are most prominent when the object is maximally compressed, see Fig. 6 (left panel). The maximum at the bottom is larger than the local maximum at the top. The local entropy shows also two maxima, see the yellow regions in Fig. 6 (right panel, the blue colors represent smaller values). However, the entropy is maximal in the top region where the curvature of the trajectories is stronger. In other words, the local entropy may be useful for rating textures.

Fig. 7 (left panel) shows that the total entropy is not constant. As long as the object is compressed the entropy is decreasing.

The energy density $e_{x^*,t}$ and the particle density $n_{x^*,t}$ are conserved. The resulting continuity equations can be used to derive a balance equation for the local entropy that, in turn, allows to extract the entropy production σ [24]. As can be seen in Fig. 7 (right panel) the entropy production changes its sign in certain subregions of the system and in the course of time.

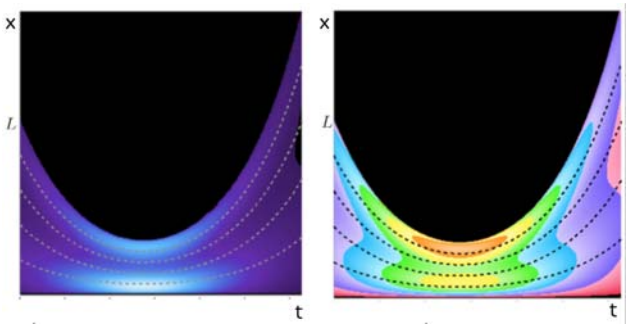


Fig. 6 – Particle density (left) and local entropy.

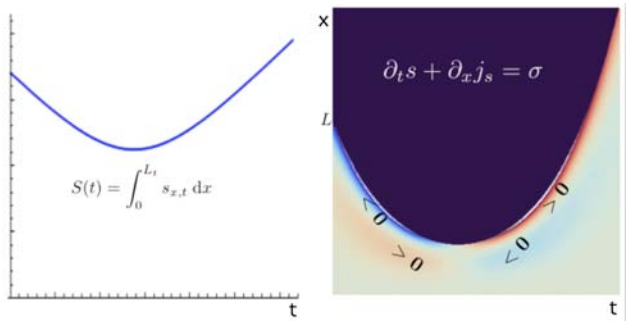


Fig. 7 – Total entropy (left) and entropy production.

5. CONCLUSION

LSDMA is a project on Big Data in science. It follows a dual approach by developing community-specific and generic services. To handle the rapidly growing deluge of data real-time analysis of Big Data will become increasingly important. In distributed real-time computing data exchange may limit the degree of parallelization.

A novel method for specifying "regions of interest" in time-dependent image processing is presented which uses methods of statistical physics. It is applied to a simple 1-dimensional example. A generalization to higher dimensions and more involved dynamics with time-dependent Hamilton functions is feasible.

6. ACKNOWLEGEMENTS

Stimulating discussions with many members of the LSDMA project are gratefully acknowledged.

7. REFERENCES

[1] T. Hey, S. Tansley, and K. Tolle (Eds.), *The fourth paradigm, data-intensive scientific discovery*, Microsoft Cooperation, 2009. Available at http://research.microsoft.com/en-us/collaboration/ fourthparadigm.

[2] http://wlcg.web.cern.ch

[3] P. E. Dewdney, *SKA1 system baseline design*, SKA-Tel-SKO-DD-001 (2013).

[4] A. Szalay, *Extreme data-intensive computing in science*, at: 1st International LSDMA Symposium *The Challenge of Big Data in Science*, Karlsruhe, Germany (25 September 2012).

[5] E. C. Friedberg, an Interview with Sydney Brenner, *Nature Reviews Molecular Cell Biology*, (9) (2008), pp. 8-9.

[6] http://www.helmholtz-lsdma.de.

[7] LSDMA IDM-Workshop, DESY, Hamburg, Germany (March 11, 2013).

[8] http://www.dCache.org

[9] J. Weschenfelder, *CDMI for dCache*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German)

[10] L. Blöcher, and T. Schubert, private communication, 2013.

[11] A. Kobitskiy, G. U. Nienhaus, J. C. Otte, M. Takamiya, U. Strähle, J. Stegmaier, and R. Mikut, *Light Sheet Microscope (LSM)*, in: R. Stotzka (Ed.), *Data Life Cycle Lab. Key Technologies, Big Data in Science, Status 2013*. Available at http://digbib.ubka.uni-karlsruhe.de/volltexte/1000037134.

[12] A. Barty, The coming deluge of data from XFEL sources, *LSDMA Workshop*, DESY, Hamburg, Germany, March 12, 2013.

[13] H. Heßling, Real-time grid computing: recent results on a pilot job approach, DESY Computing Seminar, University Hamburg, Germany, February 7, 2011.

[14] SIMON (Simple Invocation of Methods over Networks), http://dev.root1.de/projects/simon.

[15] P. Eckert, *Optimization in Java: Client-Server Communication*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[16] P. Eckert, *Garbage Collection in Java*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[17] C. Lehmann, *Studies on Using Peer-to-Peer Techniques in Grid Computing*, Master Thesis, University of Applied Sciences (HTW), Berlin, 2013. (in German).

[18] http://juxmem.gorge.inria.fr.

[19] K. Kochan, *Distributed Tree Search in GriScha*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[20] L. Bortfeld, *Real-time Communication in Grid Computing based on XMPP*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[21] P. Stewart, *Real-time Communication in Grid Computing based on XMPP*, Internal report, University of Applied Sciences (HTW), Berlin, 2013. (in German).
[22] J. Erhard, and C. Lorenz (Eds.), *4D Modeling and Estimation of Respiratory Motion for Radiation Therapy*, Springer, Heidelberg, 2013.
[23] G. Dixit, J. M. Slowik, and R. Santra, *Proposed imaging of the ultrafast electronic motion in samples using X-ray phase constrast*, Physical Review Letters, (110) 13 (2013), 137403.
[24] D. Kondepudi, and I. Prigogine, *Modern Thermodynamics. From Heat Engines to Dissipative Structures*, John Wiley, 1998.

Hermann Heßling studied Physics at the Universities of Münster, Göttingen and Hamburg. He received the Ph.D. (Dr. rer. nat) in Theoretical Physics and was appointed a postdoctoral research fellow at Deutsches Elektronen-Synchrotron (DESY), Hamburg (1993-1996).

Subsequently, he continued his work with a computer communications and networking company and accepted in 1999 an offer from the University of Applied Sciences Hof as a Professor of Operating Systems. Since 2000 he has been Professor of Applied Informatics at the University of Applied Sciences HTW Berlin. His research areas include distributed real-time computing and Big Data.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

SELF-ORGANIZING NEURAL GROVE: EFFICIENT NEURAL NETWORK ENSEMBLES USING PRUNED SELF-GENERATIONG NEURAL TREES

Hirotaka Inoue [1], Kyoshiro Sugiyama [2]

[1] Kure National College of Technology,
2-2-11 Agaminami, Kure, Hiroshima 737-8506 Japan, e-mail: hiro@kure-nct.ac.jp
[2] Advanced Engineering Faculty, Kure National College of Technology,
2-2-11 Agaminami, Kure, Hiroshima 737-8506 Japan, e-mail: s201212@sd.kure-nct.ac.jp

Abstract: Recently, multiple classifier systems have been used for practical applications to improve classification accuracy. Self-generating neural networks are one of the most suitable base-classifiers for multiple classifier systems because of their simple settings and fast learning ability. However, the computation cost of the multiple classifier system based on self-generating neural networks increases in proportion to the numbers of self-generating neural networks. In this paper, we propose a novel pruning method for efficient classification and we call this model a self-organizing neural grove. Experiments have been conducted to compare the self-organizing neural grove with bagging and the self-organizing neural grove with boosting, and support vector machine. The results show that the self-organizing neural grove can improve its classification accuracy as well as reducing the computation cost.

Keywords: neural network ensembles; self-organization; improving generalization capability; bagging; boosting.

1. INTRODUCTION

Classifiers need to find hidden information within a large amount of given data effectively and classify unknown data as accurately as possible [1]. Recently, to improve the classification accuracy, multiple classifier systems such as neural network ensembles, bagging, and boosting have been used for practical data mining [2-5]. In general, base classifiers of multiple classifier systems use traditional models such as neural networks (back-propagation network and radial basis function network) [6] and decision trees (CART and C4.5) [7].

Neural networks have great advantages such as adaptability, flexibility, and universal nonlinear input-output mapping capability. However, to apply these neural networks, it is necessary for the network structure and some parameters to be determined by human experts, and it is quite difficult to choose the right network structure suitable for a particular application at hand. Moreover, they require a long training time to learn the input-output relation of the given data. These drawbacks prevent neural networks from being the base classifier of multiple classifier systems for practical applications.

Self-generating neural networks (SGNN) [8] have a simple network design and high-speed learning. SGNN are an extension of the self-organizing maps (SOM) of Kohonen [9] and utilize the competitive learning that is implemented as a self-generating neural tree (SGNT). The abilities of SGNN make it suitable for the base classifier of multiple classifier systems. In order to improve in the accuracy of SGNN, we proposed ensemble self-generating neural networks (ESGNN) for classification [10] as one of multiple classifier systems. Although the accuracy of ESGNN improves by using various SGNN, the computation cost, that is the computation time and the memory capacity increases in proportion to the increase in numbers of SGNN in multiple classifier systems.

In an earlier paper [11], we proposed a pruning method for the structure of the SGNN in multiple classifier systems to reduce the computation cost. In this paper, we propose a novel pruning method for more effective processing and we call this model a self-organizing neural grove (SONG) [12]. This pruning method is constructed in two stages. In the first stage, we introduce an on-line pruning algorithm to reduce the computation cost by using class labels in learning. In the second stage, we optimize the structure of the SGNT in multiple classifier systems to improve the generalization capability by pruning the redundant leaves after

learning. In the optimization stage, we introduce a threshold value as a pruning parameter to decide which subtree's leaves to prune and estimate with 10-fold cross-validation [13]. After the optimization, the SONG improves its classification accuracy as well as reducing the computation cost. We use bagging [2] and boosting [14] as a resampling technique for the SONG.

We compare the SONG with support vector machine (SVM) [15] to investigate the computational cost and the classification accuracy using ten problems in the UCI machine learning repository [16].

The rest of the paper is organized as follows. The next section shows how to construct the SONG. Section 3 shows the experimental results. Then section 4 is devoted to some experiments to investigate the incremental learning performance of SONG. Finally we present some conclusions, and outline plans for future work.

2. CONSTRUCTING SELF-ORGANIZING NEURAL GROVE

First, we mention the on-line pruning method in the learning of SGNT. Second, we show the optimization method in constructing the SONG. Finally, we show a simple example of the pruning method for a two dimensional classification problem.

2.1. ON-LINE PRUNING OF SELF-GENERATING NEURAL TREE

SGNN are based on SOM and are implemented as an SGNT architecture. The SGNT can be constructed directly from the given training data without any intervening human effort. The SGNT algorithm is defined as a tree construction problem of how to construct a tree structure from the given data which consists of multiple attributes under the condition that the final leaves correspond to the given data. First, we mention the on-line pruning method in the learning of SGNT. Second, we show the optimization method in constructing the SONG.

Before we describe the SGNT algorithm, we denote some notations.

- input data vector: $\boldsymbol{e}_i \in \mathbb{R}^m$.
- root, leaf, and node in the SGNT: n_j.
- weight vector of n_j: $\boldsymbol{w}_j \in \mathbb{R}^m$.
- the number of the leaves in n_j: c_j.
- distance measure: $d(\boldsymbol{e}_i, \boldsymbol{w}_j)$.
- winner leaf for $\boldsymbol{e}_i$ in the SGNT: n_{win}.

The SGNT algorithm is a hierarchical clustering algorithm. The pseudo C code of the SGNT algorithm is given as follows:

```
Algorithm (SGNT Generation)
Input:
   A set of training examples E = {e_i},
   i = 1, ... , N.
   A distance measure d(e_i,w_j).
 Program Code:
   copy(n_1,e_1);
   for (i = 2, j = 2; i <= N; i++) {
      n_win = choose(e_i, n_1);
      if (leaf(n_win)) {
         copy(n_j, w_win);
         connect(n_j, n_win);
          j++;
      }
      copy(n_j, e_i);
      connect(n_j, n_win);
      j++;
      prune(n_win);
   }
 Output:
   Constructed SGNT by E.
```

In the above algorithm, several sub procedures are used. Table 1 shows the sub procedures of the SGNT algorithm and their specifications.

Table 1. Sub procedures of SGNT algorithm.

Sub procedure	Specification
copy(n_j, $\boldsymbol{e}_i$/$\boldsymbol{w}_{win}$)	Create n_j, copy $\boldsymbol{e}_i$/$\boldsymbol{w}_{win}$ as $\boldsymbol{w}_j$ in n_j.
choose($\boldsymbol{e}_i$, n_1)	Decide n_{win} for $\boldsymbol{e}_i$.
leaf(n_{win})	Check n_{win} whether n_{win} is a leaf or not.
connect(n_j, n_{win})	Connect n_j as a child leaf of n_{win}.
prune(n_{win})	Prune leaves if the leaves have the same class.

In order to decide the winner leaf n_{win} in the sub procedure *choose*($\boldsymbol{e}_i$,n_1), competitive learning is used. This sub procedure is recursively used from the root to the leaves of the SGNT. If an n_j includes the n_{win} as its descendant in the SGNT, the weight w_{jk}(k = 1,2, … , m) of the $n_j$is updated as follows:

$$w_{jk} \leftarrow w_{jk} + \frac{1}{C_j} \cdot \left(e_{ik} - w_{jk}\right), 1 \le k \le m. \qquad (1)$$

In the SGNT, the input vector $\boldsymbol{x}_i$ corresponds to $\boldsymbol{e}_i$, and the desired output y_i corresponds to the network output o_i which is stored in one of the leaf neurons, for (x_i, y_i) ∈D. Here, D is the training data set which consists of data $\{\boldsymbol{x}_i, y_i | \ i=1,\dots, N\}$, $\boldsymbol{x}_i \in \mathbb{R}^m$ is the input and y_i is the desired output. After all training data are inserted into the SGNT as the leaves, the leaves each have a class label as the outputs and the weights of each node are the averages of the corresponding weights of all its leaves. The whole network of the SGNT reflects the given feature space by its topology.

We explain the SGNT generation algorithm using an simple example. In this example, m is one and the four training data (x_i, y_i) is (1,1), (2,2), (3,3), and (4,4). Hence, e_{11} =1, e_{21} = 2, e_{31} = 3, and e_{41} = 4. Fig. 1 shows an example of the SGNT generation. First, e_{11} is just copied to a neuron n_1 as the root, and e_{11} is substituted to w_{11} (Fig. 1 (a)). In Fig. 1, the circle is the neuron, the integer in the circle is the number of neuron j, the integer of left-upper of the circle is c_j, and the integer of under the circle is w_{j1}. Next, n_2 and n_3 are generated as the children of n_1 with w_{21}=1, w_{31}=2. w_{11} is updated by e_{21} to 1+1/2(2-1)=1.5 (Fig. 1 (b)). Next, the winner in $\{n_1, n_2, n_3\}$ is n_3 since $d(\boldsymbol{e}_3,\boldsymbol{w}_1)$ = 1.5, $d(\boldsymbol{e}_3,\boldsymbol{w}_2)$ = 2, and $d(\boldsymbol{e}_3,\boldsymbol{w}_3)$ = 1; and thus, n_4 and n_5 are generated as the children of n_3 with w_{41} = 2, w_{51} = 3. w_{31} is updated by e_{31} to 2+1/2(3-2)=2.5 and w_{11} is updated by e_{31} to 1.5+1/3(3-1.5)=2 (Fig. 1 (c)). Finally, n_6 and n_7 are generated as the children of n_5 with w_{61} = 3, w_{71} = 4. w_{51} is updated by e_{41} to 3+1/2(4-3)=3.5, w_{31} is updated by e_{41} to 2.5 + 1/3(4-2.5) = 3, and w_{11} is updated by e_{41} to 2 + 1/4(4-2) = 2.5 (Fig. 1 (d)).

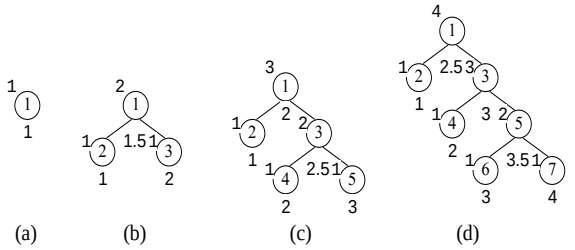


Fig. 1 – An example of the SGNT generation.

Note, to optimize the structure of the SGNT effectively, we remove the threshold value of the original SGNT algorithm in [8] to control the number of leaves based on the distance because of the trade-off between the memory capacity and the classification accuracy. In order to avoid the above problem, we introduce a new pruning method in the sub procedure *prune*(n_{win}). We use the class label to prune leaves. For leaves that have the n_{win}'s parent node, if all leaves belong to the same class, then these leaves are pruned and the parent node is given to the class.

2.2. OPTIMIZATION OF THE SONG

The SGNT has the capability of high speed processing. However, the accuracy of the SGNT is inferior to the conventional approaches, such as nearest neighbor, because the SGNT has no guarantee to reach the nearest leaf for unknown data. Hence, we construct the SONG by taking the majority of multiple SGNT's outputs to improve the accuracy (Fig. 2).

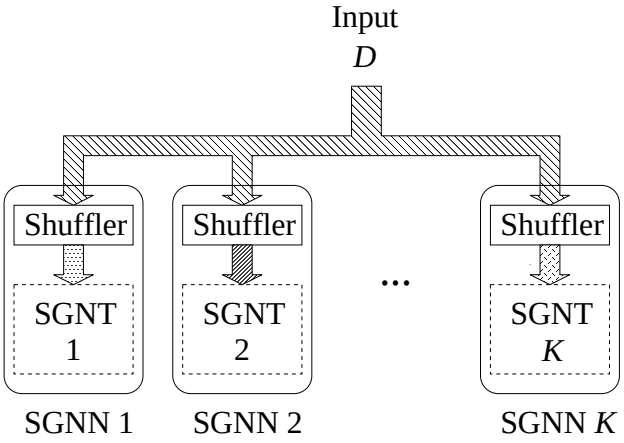


Fig. 2 – The SONG which is constructed from *K* SGNTs. The SONG's output is decided by voting outputs of *K* SGNTs.

Although the accuracy of the SONG is superior or comparable to the accuracy of conventional approaches, the computational cost increases in proportion to the increase in the number of SGNTs in the SONG. In particular, the huge memory requirement prevents the use of SONG for large datasets even with the latest computers.

In order to improve the classification accuracy, we propose an optimization method of the SONG for classification. This method has two parts, the merge phase and the evaluation phase. The merge phase is performed as a pruning algorithm to reduce dense leaves. The merge phase algorithm is given as follows:

Algorithm (Merge phase)
1. begin initialize j = the height of the SGNT
2. do for each subtree's leaves in the height j
3. if the ratio of the most class $\geq \alpha$,
4. then merge all leaves to parent node
5. if all subtrees are traversed in the height j,
6. then $j \leftarrow j-1$
7. until $j = 0$
8. end.

This phase uses the class information and a threshold value α to decide which subtree's leaves to prune or not. For leaves that have the same parent node, if the proportion of the most common class is greater than or equal to the threshold value α, then these leaves are pruned and the parent node is given the most common class.

The optimum threshold values α of the given problems are different from each other. The evaluation phase is performed to choose the best threshold value by introducing 10-fold cross validation. The evaluation phase algorithm is given as follows:

Algorithm (Evaluation phase)

1. begin initialize $\alpha = 0.5$
2. do for each α
3. evaluate the merge phase with 10-fold CV
4. if the best classification accuracy is obtained,
5. then record the α as the optimal value
6. $\alpha \leftarrow \alpha + 0.05$
7. until $\alpha = 1$

end.

2.3. SIMPLE EXAMPLE OF THE PRUNING METHOD

We show an example of the pruning algorithm in Fig. 3. This is a two-dimensional classification problem with two equal circular Gaussian distributions that have an overlap. The shaded plane is the decision region of class 0 and the other plane is the decision region of class 1 by the SGNT. The dotted line is the ideal decision boundary. The number of training samples is 200 (class0: 100, class1: 100) (Fig. 3 (a)).

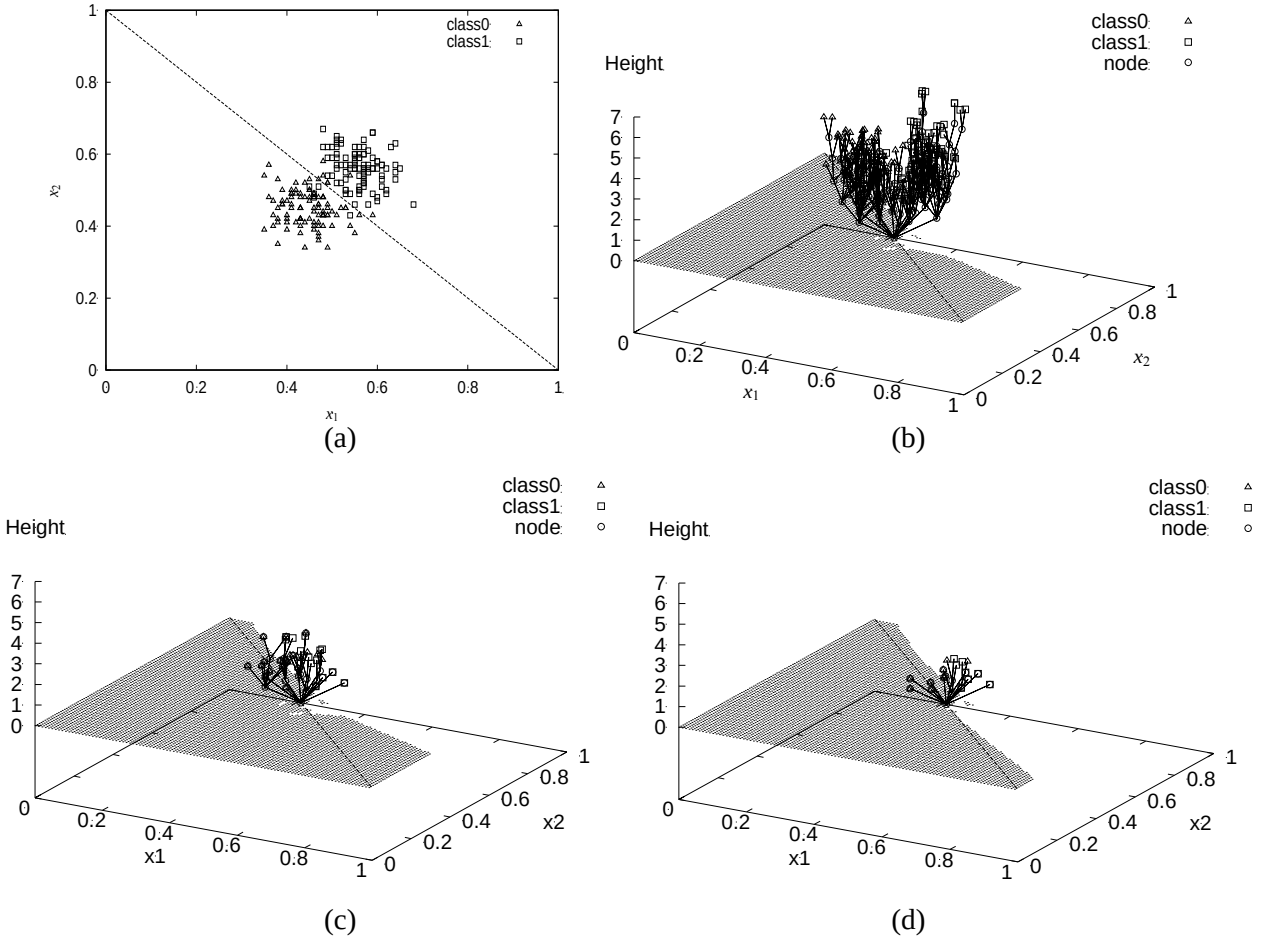


Fig. 3 – An example of the SGNT's pruning algorithm, (a) a two dimensional classification problem with two equal circular Gaussian distribution, (b) the structure of the unpruned SGNT, (c) the structure of the pruned SGNT (α= 1), and (d) the structure of the pruned SGNT (α = 0.6).The shaded plane is the decision region of class 0 by the SGNT and the doted line shows the ideal decision boundary

The unpruned SGNT is given in Fig. 3 (b). In this case, 200 leaves and 120 nodes are automatically generated by the SGNT algorithm. In this unpruned SGNT, the height is 7 and the number of units is 320. In this, we define the unit to count the sum of the root, nodes, and leaves of the SGNT. The root is the node which is of height 0.The unit is used as a measure of The decision boundary is the same as the unpruned SGNT. Fig. 3 (d) shows the pruned SGNT after the merge phase in α = 0.6. In this case, 182 leaves and 115 nodes are pruned away and only 23 units remain. Moreover, the decision boundary is improved more than the unpruned SGNT because this case can reduce the effect of the overlapping class by pruning the SGNT.

In the above example, we use all training data to construct the SGNT. The structure of the SGNT is changed by the order of the training data. Hence, we can construct the MCS from the same training data by changing the input order. We call this approach “shuffling”.

3. EXPERIMENTAL RESULTS

We investigate the computational cost (the memory capacity and the computation time) and the classification accuracy of the SONG with bagging for ten benchmark problems in the UCI machine learning repository [16]. Table 2 presents the abstract of the datasets.

Table 2. The brief summary of the datasets. *N* is the number of instances, *m* is the number of attributes.

Dataset	*N*	*m*	classes
balance-scale	625	4	3
Breast-cancer-w	699	9	2
glass	214	9	6
ionosphere	351	34	2
iris	150	4	3
letter	20000	16	26
liver-disorders	345	6	2
new-thyroid	215	5	3
pima-diabetes	768	8	2
wine	178	13	3

We evaluate how the SONG is pruned using 10-fold cross-validation for the ten benchmark problems. In this experiment, we use a modified Euclidean distance measure for the SONG as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{m} a_i \cdot (x_i - y_i)^2}, \quad (2)$$

$$a_i = \frac{1}{\max_j - \min_j}, (1 \le j \le N). \quad (3)$$

Since the performance of the SONG is not sensitive to the threshold value α, we set the different threshold values α to vary from 0.5 to 1; α = [0.5,0.55,0.6, … , 1]. We set the number of SGNT *K* in the SONG as 25 and execute 100 trials by changing the sampling order of each training set. All experiments in this section were performed on an UltraSPARC workstation with a 900 MHz CPU, 1 GB RAM, and Solaris 8.

The figures and tables must be numbered, have a self-contained caption. Figure captions should be below the figures; table captions should be above the tables. Also, avoid placing figures and tables before their first mention in the text.

Table 3 shows the average memory requirement and of 100 trials for the SONG. As the memory requirement, we count the number of units which is the sum of the root, nodes, and leaves of the SGNT. The average memory requirement is reduced from 65 % to 96.6 % and the classification accuracy is improved 0.1 % to 2.9 % by optimizing the SONG. Table 4 shows classification accuracy of 100 trials for the SONG. In Table 4, the standard deviation is given inside the bracket ($\times 10^{-3}$). The classification accuracy is improved 0.1 % to 2.9 % by optimizing the SONG. These results support that the SONG can be effectively used for all datasets with regard to both the computation cost and the classification accuracy.

Table 3. The average memory requirement of 100 trials for the bagged SGNT in the SONG.

Dataset	pruned	unpruned	ratio
balance-scale	107.68	861.18	12.5
breast-cancer-w	30.88	897.37	3.4
glass	104.33	297.75	35
ionosphere	50.75	472.39	10.7
iris	15.64	208.56	7.4
letter	6197.5	27028.56	22.9
liver-disorders	163.12	471.6	34.5
new-thyroid	49.45	298.21	16.5
pima-diabetes	204.4	1045.03	19.5
wine	15	238.95	6.2
Average	693.88	3181.96	16.9

Table 4. The classification accuracy of 100 trials for the bagged SGNT in the SONG. The standard deviation is given inside the bracket ($\times 10^{-3}$).

Dataset	pruned	unpruned	ratio
balance-scale	0.866 (6.36)	0.837 (7.83)	+2.9
breast-cancer-w	0.97 (2.41)	0.966 (2.71)	+0.4
glass	0.714 (13.01)	0.709 (14.86)	+0.5
ionosphere	0.891 (6.75)	0.862 (7.33)	+2.9
iris	0.962 (6.04)	0.955 (5.45)	+0.7
letter	0.956 (0.77)	0.955 (0.72)	+0.1
liver-disorders	0.648 (12.89)	0.636 (13.36)	+1.2
new-thyroid	0.958 (7.5)	0.957 (7.49)	+0.1
pima-diabetes	0.749 (7.05)	0.728 (7.83)	+2.1
wine	0.976 (4.41)	0.972 (5.57)	+0.4
Average	0.869	0.858	+1.1

Table 5 shows the average classification accuracy of 10 trials for the SONG with boosting. On boosting, we implement AdaBoost [14] to the SONG. Since original AdaBoost algorithm have been proposed for binary classification problems, we use four binary classification problems. In comparison with boosting, bagging is superior to boosting on all of the 4 datasets. In short,

bagging is better than boosting in terms of the classification accuracy.

Table 5. The average classification accuracy of 10 trials for the SONG with boosting. The standard deviation is given inside the bracket ($\times 10^{-3}$).

Dataset	SGNT	SONG	ratio
breast-cancer-w	0.96 (6.47)	0.957 (4.13)	-0.3
ionosphere	0.854 (18.26)	0.773 (17.4)	-8.1
liver-disorders	0.588 (17.9)	0.572 (24.3)	-1.6
pima-diabetes	0.696 (12.2)	0.722 (6.82)	+2.6
Average	0.775	0.756	-1.9

To show the advantages of the SONG, we compare it with SVM on the same problems. In the SONG, we choose the best classification accuracy of 100 trials with bagging. In SVM, we use *C*-SVM in libsvm [17] with radial basis function kernel. We select the parameters of SVM, the cost parameters C and the kernel parameters γ, from $15 \times 15 = 225$ combinations by 10-fold cross validation; $C = [2^{12},2^{11},2^{10}, \dots , 2^{-2}]$ and $\gamma = [2^{4},2^{3},2^{2}, \dots , 2^{-10}]$.We normalize the input data from 0 to 1 for all problems in SONG and SVM. All methods are compiled by using gcc with the optimization level -O2 on the same workstation.

Table 6, Table 7, and Table 8 show the classification accuracy, the memory requirement, and the computation time achieved by the SONG and SVM respectively. Next, we show the results for each category.

Table 6. The classification accuracy of 10 trials for the best pruned SONG and SVM.

Dataset	SONG	SVM
balance-scale	0.885	**0.992**
breast-cancer-w	**0.976**	0.973
glass	**0.758**	0.738
ionosphere	0.912	**0.954**
iris	**0.973**	0.96
letter	0.958	**0.977**
liver-disorders	0.685	**0.73**
new-thyroid	0.972	**0.977**
pima-diabetes	0.764	**0.766**
wine	0.983	**0.989**
Average	0.887	**0.904**

First, in view point of the classification accuracy, the SONG superior to SVM 3 of the 10 datasets and degrade 1.7 % in the average in Table 6.

Second, in terms of the memory requirement, even though the SONG includes the root and the nodes which are generated by the SGNT generation algorithm, this is less than SVM for 8 of the 10 datasets. Although the memory requirement of the SONG is totally used K times in Table 7, we release the memory of SGNT for each trial and reuse the memory for effective computation. Therefore, the memory requirement is suppressed by the size of the single SGNT.

Table 7. The memory requirement of 10 trials for the best pruned SONG and SVM.

Dataset	SONG	SVM
balance-scale	109.93	**60.6**
breast-cancer-w	**26.8**	79.6
glass	**91.33**	132.4
ionosphere	**51.38**	147.9
iris	**11.34**	51.3
letter	**6208.03**	7739.7
liver-disorders	**134.17**	214.5
new-thyroid	45.74	**44.1**
pima-diabetes	**183.57**	363.5
wine	**11.8**	62.2
Average	**687.41**	889.58

Table 8. The computation time (in sec.) of 10 trials for the best pruned SONG and SVM.

Dataset	SONG	SVM
balance-scale	**0.82**	4.77
breast-cancer-w	1.18	**0.64**
glass	**0.36**	0.61
ionosphere	1.93	**1.25**
iris	0.13	**0.06**
letter	**208.52**	2359.39
liver-disorders	**0.54**	2.07
new-thyroid	0.23	**0.22**
pima-diabetes	**1.72**	5.63
wine	0.31	**0.15**
Average	**21.57**	236.88

Finally, in view of the computation time, although the SONG consumes the cost of K times of the SGNT to construct the model and test for the unknown dataset, the average computation time is faster than SVM in Table 8. The SONG is slower than SVM for small datasets such as glass, ionosphere, and iris. However, the SONG is faster than SVM for large datasets such as balance-scale, letter, and pima-diabetes. Especially, in letter, the computation time of the SONG is faster than SVM about 11 times. We need to repeat 10-fold cross validation many times to select the optimum parameter for α, k, C, and γ. This evaluation consumes much computation time for large datasets such as letter. Therefore, the SONG based on the fast and compact SGNT is useful and practical for large datasets. Moreover, the SONG has the ability

of parallel computation because each classifier behaves independently. In conclusion, the SONG is a practical method for large-scale data mining compared with SVM.

4. CONCLUSION

In this paper, we proposed a new pruning method for the multiple classifier system based on self-generating neural trees, which is called the self-organizing neural grove, and evaluated the computation cost, that is the computation time and the memory capacity, and the classification accuracy. We introduced an on-line and off-line pruning methods and evaluated the self-organizing neural grove by 10-fold cross-validation. Experimental results showed that the memory requirement reduced remarkably, and the accuracy increased by using the pruned self-generating neural tree as the base classifier of the self-organizing neural grove. The self-organizing neural grove is a useful and practical multiple classifier system to classify large datasets. In future work, we will study a parallel and distributed processing of the self-organizing neural grove for large scale data mining.

6. REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2000.

[2] L. Breiman, Bagging predictors, *Machine Learning*, (24) (1996), pp. 123-140.

[3] R. E. Schapire, The strength of weak learnability, *Machine Learning*, (5) 2 (1990), pp. 197-227.

[4] J. R. Quinlan, Bagging, Boosting, and C4.5, *the Thirteenth National Conference on Artificial Intelligence*, Portland, OR, (August 4-8, 1996), pp. 725-730.

[5] G. Ratsch, T. Onoda, K. R. Muller, Soft margins for AdaBoost, *Machine Learning*, (42) 3 (2001), pp. 287-320.

[6] S. Haykin, *Neural Networks: A Comprehensive Foundation*, second ed., Prentice-Hall, Upper Saddle River, NJ, 1999.

[7] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification*, second ed., John Wiley & Sons Inc., New York, 2000.

[8] W. X. Wen, A. Jennings, H. Liu, Learning a neural tree, *the International Joint Conference on Neural Networks*, Beijing, China, (November 3-6, 1992), Vol. 2, pp. 751-756.

[9] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995.

[10] H. Inoue, H. Narihisa, *Improving generalization ability of self-generating neural networks through ensemble averaging*, in: T. Terano, H. Liu, A.L.P. Chen (Eds.), PAKDD 2000, Lecture Notes in Computer Science, Vol. 1805, Springer, Heidelberg, 2000, pp. 177-180.

[11] H. Inoue, H. Narihisa, *Optimizing a multiple classifier system*, In: M. Ishizuka, A. Sattar (Eds.), PRICAI 2002, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), Vol. 2417, Springer, Heidelberg, 2002, pp. 285-294.

[12] H. Inoue, K. Sugiyama, Self-organizing neural grove, *the 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Berlin, Germany, (September 12-14, 2013), pp. 319-323.

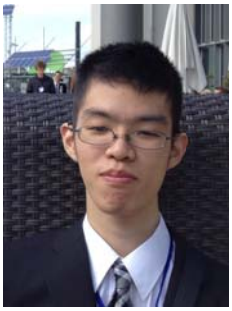
[13] M. Stone, Cross-validation: A review, *Math. Operations for sch. Statist. Ser. Statistics*, (9) 1 (1978), pp.127-139.

[14] Y. Freund and R. E. Schapire, *Boosting: Foundations and Algorithms*, MIT Press, Cambridge, MA, 2012.

[15] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, (2) 27 (2011), pp. 1–27, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm

[16] C. Blake, C. Merz, UCI repository of machine learning databases, 1998. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

***Hirotaka Inoue**, received his MSc degree in Okayama University of Science in 1999 and PhD degree in the same university in 2002, both in engineering. His research interests are in neural networks, pattern recognition, and combinational optimization. He is currently an associate professor at Kure National College of Technology in Japan.*



***Kyoshiro Sugiyama**, is a student at Advanced Engineering Faculty, Kure National College of Technology. He will be a graduate student at Nara Institute of Science and Technology from April 2014. His research interests are in neural networks and pattern recognition.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

COMPREHENSIVE MULTILEVEL SECURITY RISK ASSESSMENT OF DISTRIBUTED INFORMATION SYSTEMS

Igor Kotenko and Elena Doynikova

St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS)
39, 14 Liniya, St. Petersburg, Russia
ivkote@comsec.spb.ru, http://www.comsec.spb.ru/kotenko/
doynikova@comsec.spb.ru, http://www.comsec.spb.ru/doynikova/

Abstract: The paper suggests the multilevel approach to the risk assessment that is based on the system of security metrics and techniques for their calculation. Proposed techniques are based on attack graphs and service dependencies. They allow evaluating security of network topologies, malefactors and attack characteristics, and integral security properties and characteristics calculated on the basis of the cost-benefit and zero-day vulnerability analysis. Classification of these characteristics and separation of the security information on static, dynamic and historical allows defining different assessment levels. The paper considers the main issues and recommendations for using the risk assessment techniques based on the suggested approach.

Keywords: Cyber security, security metrics, risk assessment, attack graphs, service dependencies.

1. INTRODUCTION

Information risk management is an important aspect of operation of modern complex distributed information systems.

For effective security decisions it is necessary to get accurate and actual information about security state of the system.

Different security metrics can provide this information. Calculation and analysis of security metrics are involved in the risk assessment process as part of the risk management process.

There are a lot of security metrics that are related to the different aspects of the security of the distributed systems [1, 3, 4, 7, 11, 12, etc.]. This paper proposes the integrated system of security metrics and techniques for their calculation.

Main elements of the suggested approach include analysis of the static, dynamic and historical security information, calculation of security metrics on the base of attack graphs and service dependencies, usage of the standards of the Security Content Automation Protocol (SCAP) to represent data about system platforms, applications and configurations in unified form and to assess vulnerabilities.

The approach is based on taking into account the current research in the area of security metrics.

Proposed metrics allow evaluating a set of parameters: the security level of network topologies, malefactors and attack characteristics, and integral security properties calculated on the basis of the cost-benefit and zero-day vulnerability analysis.

We define several assessment levels according to the type of the analyzed characteristics and used data – a topological, an attack graph, a malefactor, events and the whole system. Each of the level contains risk assessment metrics that we propose to calculate. Such approach allows assessing risk on different stages of the system operation and getting time-dependent results. Also the paper discusses the most characteristic techniques that allow evaluating the proposed metrics with different accuracy: express risk assessment static technique, performance-based technique and technique based on historical data.

The express risk assessment technique allows simple and quick assessment of the security level of the system in the static operation mode. The performance-based technique permits assessment of the security level in the operation mode according to the detected security events. The technique based on the historical data allows detailed assessment with consideration of the data about previous security incidents.

One of the main problems in the risk assessment is the lack of data on the security incidents. We use Common Vulnerabilities Scoring System (CVSS) [29] scores as basis for calculations of the security

metrics. But it is still heuristic assessments. To manage uncertainty we use Bayes techniques and consider possibility of zero-day vulnerabilities that are not included to the Common Vulnerabilities and Exposures (CVE) [9] database.

This paper is an extended version of the paper presented on IDAACS'2013 [21]. It structured as follows. *Section 2* reviews related works in the area of security metrics. On the base of the review we outline the classes of known security metrics (topology, malefactor, and attack characteristics), consider integral metrics, and also metrics calculated on the basis of the cost-benefit and zero-day vulnerability analysis. The proposed hierarchical security assessment framework allows to evaluate the considered system from different aspects and to specify the risk value according to the available data. *Section 3* describes requirements to the suggested approach and proposes the system of security metrics. *Section 4* considers the techniques that are suggested for evaluation of the proposed security metrics. *Conclusion* outlines main results of the work and directions of future research.

2. RELATED WORK

There is the number of different *classifications of the security metrics* according to their goals, way of computation or value types.

For example, The Center for Internet Security (CIS) divides metrics according to six business functions [7]: incident management, vulnerability management, patch management, application security, configuration management, and financial metrics. In [2] eight categories of metrics are differentiated according to the value type: existence (indicator of whether something exists); ordinal (subjective qualitative measure); score (numeric values for qualitative measure); cardinal (number); percentage; holistic (based on external data sources); value (consider value loss); uncertainty (include stochastic or probabilistic aspect). In [14] metrics are divided according to the way of computation on primary (calculated on the base of the attack graphs or service dependencies graphs) and secondary (calculated on the base of the primary metrics).

In [18] metrics are distinguished according to the used model: logical dependencies graph (used to define, for example, attacker skill level, attack potentiality) and service dependencies graph (used to define, for instance, attack/response impact, response benefit).

Based on the current research in the area of security metrics we outline the following main groups of metrics: topology characteristics, malefactor characteristics, attack and response characteristics, system characteristics (integral metrics), cost characteristics and characteristics that are used in analysis of the zero-day vulnerabilities.

Topology characteristics are considered, for example, in [28] and [7].

In [28] the following metrics are outlined: *Host Criticality* (impact from loss of the host to the business), *Exposure* (it is defined by the reachability of the host and easiness to exploit the vulnerabilities on the host), *Business Value* (it is similar to *Criticality*, but expressed in monetary units), *Risk* (it is defined by *Exposure* and *Business Value*) and *Downstream Risk* (it is cumulative risk over the hosts attackable from the current host).

In [7] the topology metrics from the applications point of view are suggested – *Number of Applications*, *Percentage of Critical Applications*, and topology characteristics that consider information about vulnerabilities – *Percent of Systems Without Known Severe Vulnerabilities*, *Mean-Time to Mitigate Vulnerabilities*, *Number of Known Vulnerability Instances*, and topology characteristics that consider information about attacks – *Severity* of the vulnerability and its *Access Complexity*, that can be considered when the attack likelihood is calculated.

One of the most important *malefactor characteristics* is *Attacker Skill Level*, which is considered in [17] and [32]. It is valuable indicator about the attacker ability to carry-on an attack scenario and thus to attain his objectives.

In [10] calculation of the risk level on the base of the malefactor behavior is proposed. In some other papers, for example, in [5], attack attribution metrics are considered (the concrete actors involved, equipment and tools used, geographic position, motives, etc.).

Attack characteristics include *Attack Potentiality* (dynamic metric that is computed with respect to the attacker position in the attack graph).

In [35] *Attack Potentiality* is defined on the base of the *Confidence Level* metric – a confidence level in the fact that attack is in progress.

In [38] *Compromised Confidence Index* is suggested.

In [1, 16, 19, 36] another attack characteristic is proposed – *Attack Impact*. In [16, 19, 38] such *response metrics* as *Response Efficiency*, *Benefit of the Response* and *Response Collateral Damage* are considered.

Two main metrics should be outlined to consider the total *risk level* of the system (i.e. *integral metrics*): *Attack Surface* (it is defined on the base of the *Damage Potential-Effort Ratio* metric) [27] and *Security Level* [10, 15, 20, 22, 23, 35].

For the *cost-benefit analysis* the following metrics can be used: *Net Benefit* (total benefit in case of implementation of safeguard), *Annual Loss*

Expectancy (product of an incident's annual frequency times its total losses) [13], and *Return on Response Investment* [19].

Metrics that are used in the *analysis of zero-day vulnerabilities* involve *Probabilistic Vulnerability Measure* [1], which defines how likely the vulnerability will be released for a service over some period and the vulnerability's expected severity, and *k-zero day safety*, which defines network resistance to zero-day vulnerabilities [37].

3. THE SYSTEM OF SECURITY METRICS

We define the classification of security metrics on the base of the review above. Classification of security metrics is demonstrated in Fig. 1 and includes the following classes of security metrics:

(1) Host/topology characteristics;

(2) Malefactor characteristics (define malefactor skills, position etc.);

(3) Attack characteristics (define attack potential and possible impact);

(4) System characteristics (define integral security characteristics);

(5) Zero-days characteristics (define possibility of zero-day attacks);

(6) Cost-benefit characteristics (define cost of the attack and response).

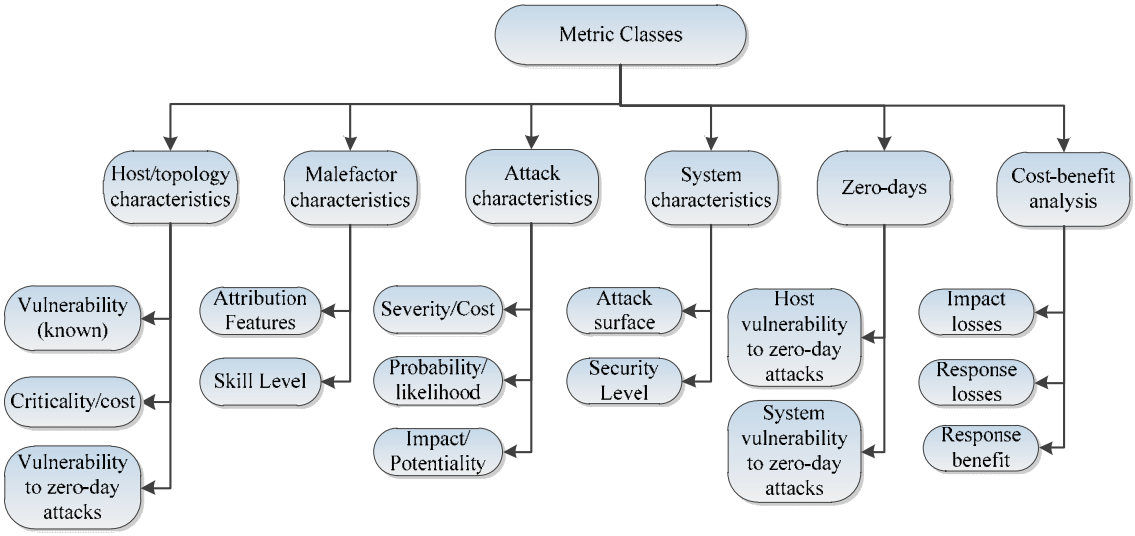


Fig. 1 – Classes of metrics

The system of security metrics as proposed in this paper is designed for the risk assessment approach that involves the next stages:

(1) generation of the attack and service dependency graphs on the base of the data about the network topology;

(2) consideration of the malefactor skills and position and generation of the profile attack graphs;

(3) analysis of the system events to monitor current security situation;

(4) calculation of security metrics on the base of this data.

This approach is implemented in the *Security Evaluator* [22-25].

We propose to use the SCAP protocol and particular standards included to this protocol [34] to present input security data for the *Security Evaluator*. SCAP, produced by the National Institute of Standards and Technologies (NIST) [30], includes a collection of specifications intended to standardize the way the security software solutions communicate software security flaw and configuration information.

SCAP contains the following standards: Common Configuration Enumeration (CCE) specifies features of the configurations that negatively influence the system security [6], Common Platform Enumeration (CPE) allows to create the list of the used platforms and applications [8], CVE allows to specify the list of the vulnerabilities, CVSS assesses negative influence of the configurations and vulnerabilities that allows the most critical vulnerabilities to be defined. CPE, CVE and CVSS are used for the attack graph generation.

Besides the configuration information, the list of used platforms and applications and the list of vulnerabilities, the input data includes software weaknesses represented by the Common Weakness Enumeration (CWE) standard, attack patterns specified using the Common Attack Pattern Enumeration and Classification (CAPEC) standard, security remediations in the format of the Common Remediation Enumeration (CRE) standard, security events in the Common Event Expression (CEE) format, the service dependencies (allow to define

impact propagation), security policies, malefactor models, etc.

Results of the processing of *Security Evaluator* include: attack graphs; calculated security and exposure metrics (some of them are calculated on the base of the attack graph); countermeasures (are defined on the base of the calculated security and exposure metrics).

Input and output information for the suggested *Security Evaluator* is outlined in Fig. 2.

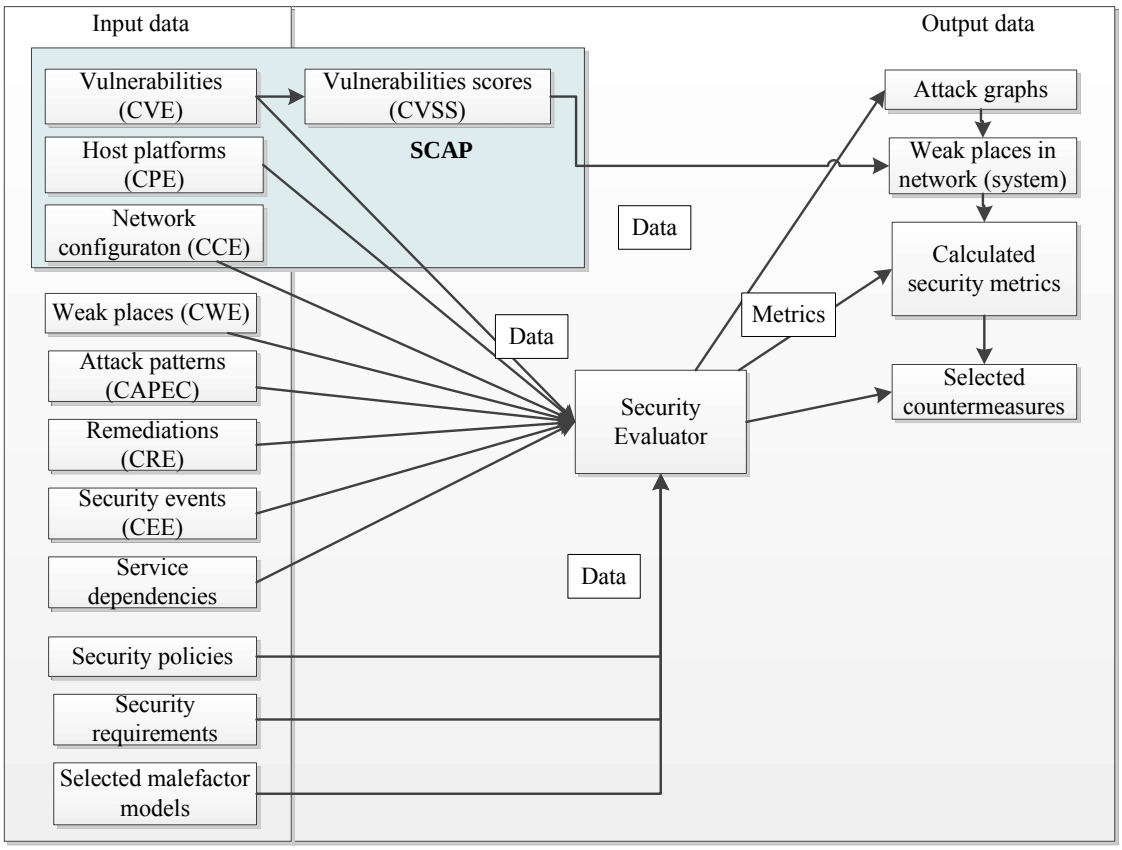


Fig. 2 – Data transformations in Security Evaluator

In process of security metrics development we considered the state-of-the-art in security metrics research as well as the architecture of the *Security Evaluator*.

Offered techniques cover two operation modes of the security evaluation system: online and offline.

The online mode stipulates limitations on the calculation time; however, it takes into account the current security situation (events, system configuration, etc.). So we can monitor position and skills of malefactor, and we can define the direction of the attack more carefully.

Off-line mode has no hard time limitations. In this case, historical data are used, and attack graphs are analyzed. This mode allows making a more profound and detailed risk assessment, as it applies more meaningful metrics that characterize the malefactor and attacks. Besides, computations of this level can be used as the base for the on-line mode.

On the base of the considered aspects we can arrange security metrics in our framework by levels, as it is shown in Fig. 3.

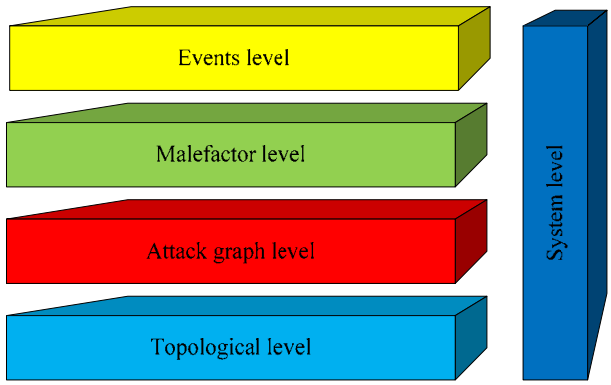


Fig. 3 – Levels of the suggested framework

Metrics of the higher levels are defined on the base of metrics of the lower levels, except for the system level metrics that are specified on each level via metrics of the appropriate level:

- *Topological level* – metrics of this level can be calculated by the administrator on the base of the system topology. We consider the following examples of metrics of this level: vulnerability

level of host, criticality level of host and vulnerability of host to the zero-day attacks.

- *Attack graph level* – on this level we consider information from the attack graph for the metrics generation. The metrics of this level are *Attack Likelihood* and *Attack Impact* (in this case the impact is defined only by the target criticality and the attack severity). When presenting the attack graph to the user we can highlight the most critical attack paths (from the risk level point of view, i.e. combination of the attack probability and attack impact).
- *Malefactor level* – on the base of the metrics of this level the dependency from the malefactor profile is introduced (including his position and skills), this allows presenting the profile attack graph [10], which includes only attacks that can be implemented by the appropriate malefactor.
- *Events level* – this level is actual, when the Security Evaluator works in online mode (in real time), because on this level the security events are considered. It allows monitoring attack deployment and malefactor profile according to incoming events. When new events come, we can represent the current position of the malefactor (host and access rights) on the attack graph and possible attack paths (all possible paths and the most probable).
- *System level* – *Common Security Level* of the system and *Attack Surface* are defined on this level. One more common metric that can be used on this level is the resistance to zero day attacks. Approach to computation of these metrics depends on taken into consideration parameters, thus it differs for all upper levels.

The values of metrics calculated on the lower levels are specified with new data, for example, the probability of the attack becomes higher if we get a security event that confirms appropriate attack.

4. THE RISK ASSESSMENT TECHNIQUES

Fig. 4 depicts all levels described above and their relation to the appropriate security metrics.

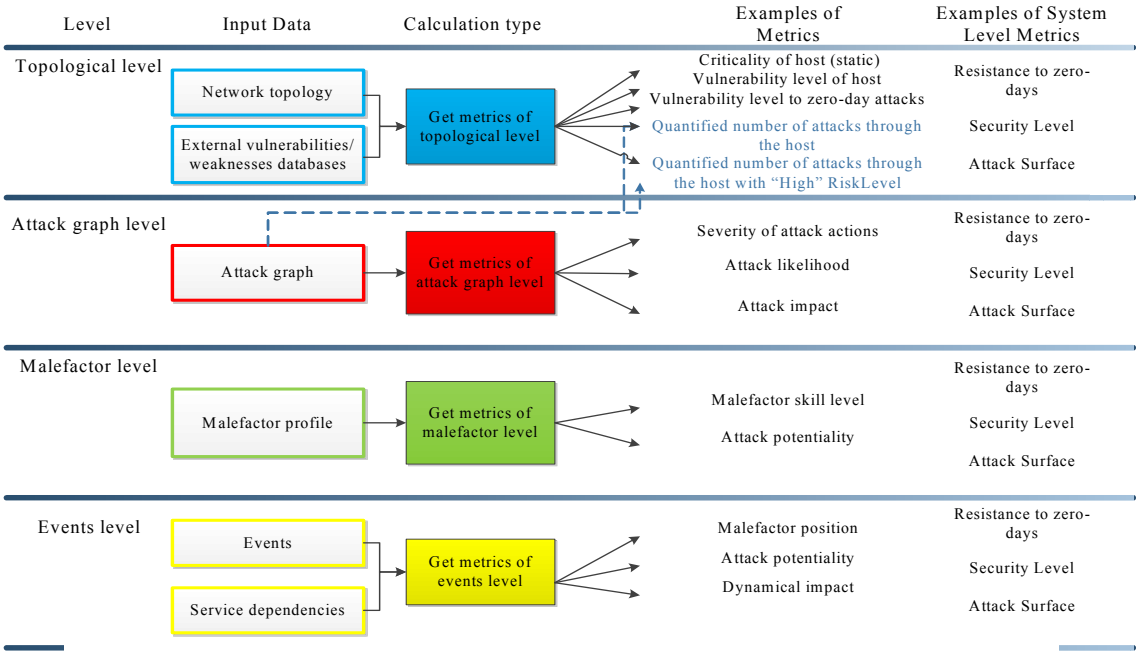


Fig. 4 – Security metrics and assessment levels

For the *topological level* we outline several host/system metrics that can be useful for the administrator and can be applied as base metrics (metrics that are used in calculations of another metrics) on the other levels. The examples of these metrics are as follows:

- *Criticality of Host* (or *Host Criticality*) gives information about the host criticality, can be defined for example on the base of CVSS (High, Medium or Low);
- *Vulnerability Level of Host* (or *Host Vulnerability*) (this metric can consist of the several values, for example, number of known vulnerabilities for the host, number of vulnerabilities with "High" CVSS *BaseScore*, etc. These values can be determined on the base of the CIS metric *Number of Known Vulnerability Instances* (*NKVI*), CVE and NVD);
- *Vulnerability Level to Zero-day Attacks* (for example, number of weaknesses, number of weaknesses with "High" CWSS *BaseScore*, etc.

These values can be determined on the base of CWE). For this metric we can use two calculation techniques: (1) considering vulnerability to zero-days only for single host (*Host Weakness*), and (2) detecting integrated system resistance to zero-days (*Vulnerability to Zero-day Attacks*), as, for example, in [37];

- *Percent of Systems (hosts) Without Known Severe Vulnerabilities (PSWKSV)* or number of hosts with "High" criticality in the system, etc.

The following two metrics can be calculated after adding information about the attack graph:

- *Quantified Number of Attacks through the Host*, which is equal to (Number of attacks that go through the host)/(Total number of attacks in the graph);
- *Quantified Number of Attacks through the Host with "High" RiskLevel*, which is equal to (Number of attacks with "High" *RiskLevel* through the host)/(Total number of attacks with "High" *RiskLevel* in the graph).

For techniques of the *attack graph level* we outline the metrics that consider paths in the attack graph. Such type of techniques allows us to define the security level of the system rather simply without considering attacker skills and likelihood of ongoing attack. The following metrics can be used here:

- *Severity of Attack Actions* ("base" metric);
- *Attack Severity*;
- *Attack Complexity*;
- *Attack Likelihood* (here we consider static and statistical approaches);
- *Attack Impact* (as static impact).

Here we should note that on the base of these metrics the express risk assessment technique is suggested. This technique does not consider attacker skills and attack changes in time and dynamic impact, but has low computational complexity. It can be improved by using the metrics of malefactor and events levels.

Techniques of the *malefactor level* deal with attacker skills (static/ statistical/ historical) and attack probability.

Here we suggest using the following *techniques and their modifications*:

- *Assessment where a malefactor profile is set statically by an administrator*. On the base of the malefactor skill level it is possible to create the profile attack graphs and evaluate the system security for each group of malefactors;
- *Probability assessment that uses the attack graphs for attacks representation and the Bayesian inference for risk analysis* (to get probability of attacks on the base of the probability of malefactor skills) [10];
- *Historical data assessment* that considers historical data to get the posterior probability of the attacks [32].

At the *events level* the risk assessment techniques take into account the dynamic aspect of risk assessment, using information (from sensors, correlation or intrusion detection components) on new events, probability of false and missed alarms, etc. We suppose the techniques of this level should be based on the following metrics:

- Malefactor position and compromised hosts;
- *Attacker Skill Level* (dynamic metric) [16];
- *Attack Potentiality* (here the event-based [35] and historical data-based [38] techniques can be used);
- *Dynamical Attack Impact* on the base of service dependencies [19, 37] (here we also can add such metric for the host as availability);
- The risk of the event can be evaluated, for example, as in [26].

On the base of the considerations above we outlined *three risk assessment techniques*:

- Express risk assessment technique;
- Performance-based (dynamic) technique;
- Technique based on historical data.

The *express risk assessment technique* is used for the express risk evaluation. It is a static technique that incorporates qualitative and quantitative approaches to the risk assessment and allows defining the common security level of the system.

This approach considers traditional understanding of the risk as the result of probability of the threat and its consequences for the system.

To assess risk we use CVSS (to define criticality of the attack action) and procedures of the FRAP (Facilitated Risk Analysis Process) technique [31].

Technique includes definition of the severity levels for hosts and attack actions, calculation of the impact from the attack actions, calculation of the impact from threats realization and complexity of threats realization, then on the base of this data the risk level for all threats is determined, and common security level of the system is calculated.

Main difference of the *performance-based technique* is that fact that it is oriented on the real time situation, when we can monitor the current attacker position and his (her) path in the network, but have hard time limitations for calculations.

Main modules of the *Security Evaluator* that are responsible for this technique, input data and links between them are represented in Fig. 5.

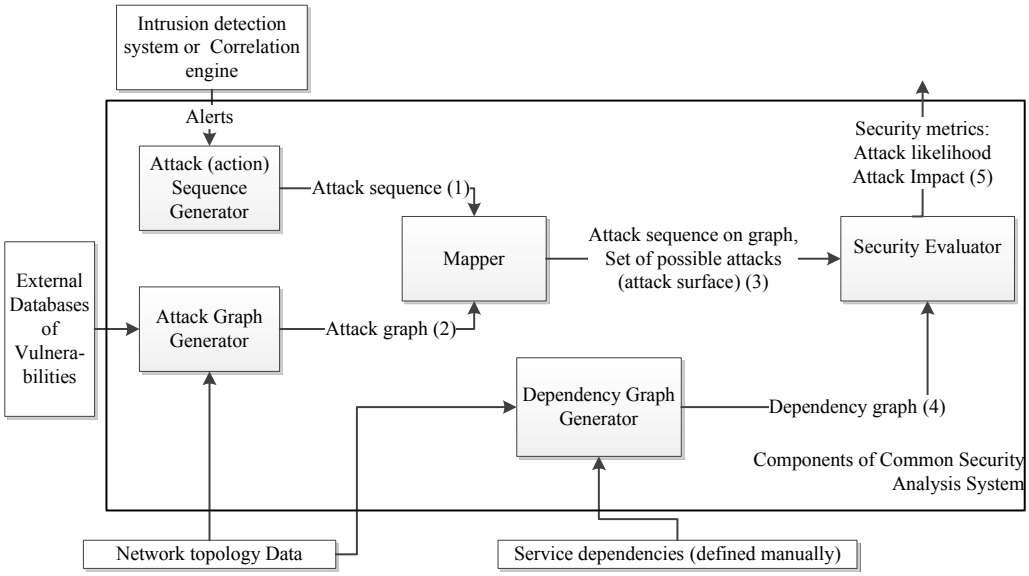


Fig. 5 – Input data and main modules of Security Evaluator for performance-based risk assessment

Main stages of the technique are as follows:

(1) Attack (action) Sequence Generator builds the attack (action) sequence on the base of the alerts from correlation engine (CEP).

(2) Attack Graph Generator forms the graph of possible attacks on the base of the data about known vulnerabilities and the network topology.

(3) Mapper reflects the attack sequence (generated on stage 1) on the attack graph (generated on stage 2) to define realized attack sequence and the malefactor position on the attack graph.

(4) Dependency Graph Generator builds the service dependency graph on the base of the data about service dependencies in the analyzed network.

(5) Security Evaluator calculates the set of security metrics on the base of the attack graph, the defined malefactor position, the set of realized steps, and the service dependency graph.

This approach also defines the risk as the result of probability of the threat and its consequences for the system.

Probability of the threat is defined with the *Likelihood of Successful Attack* metric. In case of the ongoing attack it can be defined on the base of the following elements: *Severity* of the attack action (calculated on the base of the CVSS), *Attacker Skill Level* (calculated on the base of the *AccessComplexity* of the realized steps), *Attack Potentiality* (it is equal *number of realized attack steps*/*total number of attack steps*), *Reliability* coefficient (it is evaluated on the base of the *False Positive Rate*).

Consequences of the threat for the system are defined with the *Impact of Successful Attack* metric. It can be divided on the impact of the attack action (a *Native Impact*) and the *Propagated Attack Impact*. *Native Impact* is defined as vector {*Confidentiality Impact*, *Integrity Impact*, and *Availability Impact*} on the base of the CVSS indexes. *Propagated Impact* is defined via the service dependencies.

The *technique based on historical data* uses the same approach as performance-based technique, but when attack likelihood is defined we add additional weights that are related to historical data. It concerns such metrics as *Attack Potentiality* and *Attacker Skill Level*.

For the *Attack Potentiality* weight is defined by the relation of the number of cases when detected attack sequence led to the assessed attack to the total number of the occurrences of the detected attack sequence.

For the *Attacker Skill Level* statistical data is used to define probability that attacker with appropriate skills initialize assessed attack. The data is stored in the historical database.

5. CONCLUSION

The paper proposed the comprehensive multilevel system of security metrics and techniques for their calculation. We analyzed the state-of-the-art in the area of risk assessment.

On the base of this research we outlined different classes for known risk assessment metrics. The system of risk assessment metrics that we suggest to calculate in the *Security Evaluator* is proposed. The most characteristic techniques that allow evaluating the proposed metrics are discussed. Brief description of the security metrics for each level and techniques of their calculation is given. We implemented

suggested techniques in the scope of the Common Security Analysis System [22- 25].

The suggested framework proposes structural organization of security metrics for the security analysis on the base of attack graphs. The metrics are defined on the base of the main elements of such analysis (system model, attacker model, attack model, events level). The framework allows to assess each element separately and to use it in the common risk assessment on the base of the suggested techniques. The techniques are selected according to the last research in the risk analysis area. In dynamic case information from the security events is considered. More information leads to the more accurate assessments.

On this moment we do not consider reliability of the data from the security events but we suppose to consider it in the future research. Techniques and metrics that are described in the paper will be considerably extended and detailed. Also we plan to test the *Security Evaluator* on real examples and analyze effectiveness of security assessment.

6. ACKNOWLEDGMENT

This research is being supported by grant of the Russian Foundation of Basic Research (13-01-00843, 13-07-13159, 14-07-00697, 14-07-00417), and Program of fundamental research of the Department for Nanotechnologies and Informational Technologies of the Russian Academy of Sciences (#2.2).

7. REFERENCES

[1] M. S. Ahmed, E. Al-Shaer, L. Khan, A novel quantitative approach for measuring network security, *Proceedings of the 27th Conference on Computer Communications* (*INFOCOM'08*), Phoenix, AZ, USA (April 13-18, 2008), pp. 1957-1965.

[2] C. W. Axelrod, Accounting for value and uncertainty in security metrics, *Information Systems Control Journal*, (6) (2008), pp. 1-6.

[3] R. Barabanov, S. Kowalski, L. Yngstrom, *Information Security Metrics. State of the Art*, DSV Report series, No. 11-007 (March 2011).

[4] N. Bartol, *Practical measurement framework for software assurance and information security (Version 1.0)*, Software Assurance Measurement Working Group (2008). Available at https://buildsecurityin.us-cert.gov/swa/downloads/SwA_Measurement.pdf.

[5] B. A. Blakely, *Cyberprints Identifying cyber attackers by feature analysis*, Doctoral Dissertation, Iowa State University, 2012.

[6] Common Configuration Enumeration (CCE) [Electronic resource]. Available at http://cce.mitre.org/.

[7] *The CIS Security Metrics*, The Center for Internet Security, 2009.

[8] Common Platform Enumeration (CPE) [Electronic resource]. Available at http://cpe.mitre.org/.

[9] Common Vulnerabilities and Exposures (CVE). [Electronic resource]. Available at http://cve.mitre.org/.

[10] R. Dantu, P. Kolan, J. Cangussu, Network risk management using attacker profiling, *Security and Communication Networks*, (1) (2009), pp. 83-96.

[11] L. Hayden, *IT Security Metrics: A Practical Framework for Measuring Security & Protecting Data*, McGraw-Hill, 2010, 396 p.

[12] D. S. Herrmann, *Complete Guide to Security and Privacy Metrics*, Auerbach Publications, 2007, 848 p.

[13] K. J. S. Hoo, *How much is enough? A risk-management approach to computer security*, PhD thesis, Stanford University, CA, 2000.

[14] N. C. Idika, *Characterizing and Aggregating Attack Graph-based Security Metrics*, CERIAS Tech Report 2010-23, Center for Education and Research Information Assurance and Security, Purdue University, August 2010.

[15] ISO/IEC 27005:2008, Information technology – Security techniques – Information security risk management, 2008.

[16] M. Jahnke, C. Thul and P. Martini, Graph-based metrics for intrusion response measures in computer networks, *Proceedings of the 3rd IEEE Workshop on Network Security, held in conjunction with 32nd IEEE Conference on Local Computer Networks*, Dublin (2007).

[17] W. Kanoun, N. Cuppens-Boulahia, F. Cuppens, J. Araujo, Automated reaction based on risk analysis and attackers skills in intrusion detection systems, *Proceedings of the third International Conference on Risks and Security of Internet and Systems* (*CRiSIS'08*), Toezer, Tunisia (2008), pp. 117-124.

[18] N. Kheir, *Response policies & counter-measures: Management of service dependencies and intrusion and reaction impacts*, PhD Thesis, Telecom Bretagne, 2010.

[19] N. Kheir, N. Cuppens-Boulahia, F. Cuppens, H. Debar, A service dependency model for cost-sensitive intrusion response, *Proceedings of the 15th European Symposium on Research in Computer Security* (*ESORICS'10*), Athens, Greece (2010), pp. 626-642.

[20] I. Kotenko, M. Stepashkin, Attack graph based evaluation of network security, *Proceedings of the 10th IFIP Conference on Communications*

and Multimedia Security (*CMS'2006*), Heraklion, Greece (2006), pp. 216-227.

[21] I. Kotenko, E. Doynikova, Security metrics for risk assessment of distributed information systems, *Proceedings of the IEEE 7th International Conference on "Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications" (IDAACS'2013)*, Berlin, Germany, pp. 646-650.

[22] I. Kotenko, A. Chechulin, and E. Novikova, Attack Modelling and Security Evaluation for Security Information and Event Management, *Proceedings of the International Conference on Security and Cryptography (SECRYPT 2012)*, Rome, Italy, pp. 391-394.

[23] I. Kotenko, A. Chechulin, Common Framework for Attack Modeling and Security Evaluation in SIEM Systems, *Proceedings of the 2012 IEEE International Conference on Green Computing and Communications, Conference on Internet of Things, and Conference on Cyber, Physical and Social Computing*, Besançon, France, *IEEE Computer Society*, Los Alamitos, California (2012), pp. 94-101.

[24] I. Kotenko and A. Chechulin, Attack Modeling and Security Evaluation in SIEM Systems, *International Transactions on Systems Science and Applications*, (8) 2012, pp. 129-147.

[25] I. Kotenko and A. Chechulin, A Cyber Attack Modeling and Impact Assessment Framework, *5th International Conference on Cyber Conflict 2013 (CyCon 2013)*, *Proceedings. IEEE and NATO COE Publications*, Tallinn, Estonia, pp. 119-142.

[26] J. M. Lorenzo, *AlienVault Users Manual. Version 1.0*, AlienVault, 2010-2011.

[27] P. K. Manadhata, J. M. Wing, An attack surface metric, *IEEE Transactions on Software Engineering*, (37) 3 (2011), pp. 371-386.

[28] A. Mayer, *Operational Security Risk Metrics: Definitions, Calculations, Visualizations*, Metricon 2.0. CTO RedSeal Systems, 2007.

[29] P. Mell, K. Scarfone, S. Romanosky, *A Complete Guide to the Common Vulnerability Scoring System Version 2.0*, June 2007.

[30] National Institute of Standard and Technologies. Available at http://www.nist.gov/.

[31] T. R. Peltier, *How to complete a risk assessment in 5 days or less*, Auerbach publications, 2008, 55 p.

[32] T. Olsson, Assessing security risk to a network using a statistical model of attacker community competence, *Proceedings of the 11th International Conference on Information and Communications Security (ICICS'2009)*, Beijing, China, pp. 308-324.

[33] N. Poolsappasit, R. Dewri, I. Ray, Dynamic security risk management using Bayesian attack graphs, *IEEE Transactions on Dependable and Security Computing*, (9) 1 (2012), pp. 61-74.

[34] S. Quinn, D. Waltermire, C. Johnson, K. Scarfone, J. Banghart, *The technical specification for the security content automation protocol (Version 1.0)*, Gaithersburg, MD: National Institute of Standards and Technology, 2009. Available at http://csrc.nist.gov/publications/nistpubs/800-126/sp800-126.pdf.

[35] N. Stakhanova, S. Basu, and J. Wong, A cost-sensitive model for preemptive intrusion response systems, *Proceedings of the 21st International Conference on Advanced Networking and Applications*, Washington, DC, USA, IEEE Computer Society (2007), pp. 428-435.

[36] T. Toth and C. Kruegel, Evaluating the impact of automated intrusion response mechanisms, *18th Annual Computer Security Applications Conference (ACSAC)*, 2002, pp.301-310.

[37] L. Wang, A. Singhal, S. Jajodia, and S. Noel, k-zero day safety: measuring the security risk of networks against unknown attacks, *Proceedings of the 15th European conference on Research in computer security*, Springer-Verlag Berlin, Heidelberg (2010), pp. 573-587.

[38] Y.-S. Wu, B. Foo, Y.-C. Mao, S. Bagchi, and E. H. Spafford, Automated adaptive intrusion containment in systems of interacting services, *Computer Networks: The International Journal of Computer and Telecommunications Networking*, (51) (2007), pp. 1334-1360.

Igor Kotenko *graduated with honors from St. Petersburg Academy of Space Engineering and St. Petersburg Signal Academy. He obtained the Ph.D. degree in 1990 and the National degree of Doctor of Engineering Science in 1999. He is Professor of computer science and Head of the Laboratory of Computer Security Problems of St. Petersburg Institute for Informatics and Automation.*

Elena Doynikova *graduated with honors from St. Petersburg Electrotechnical University "LETI". She is researcher of the Laboratory of Computer Security Problems of St. Petersburg Institute for Informatics and Automation.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

THE LEARNING ANALYTICS APPLICATION LEMO – RATIONALS AND FIRST RESULTS

Margarita Elkina [1], Albrecht Fortenbacher [2], Agathe Merceron [3]

[1] Hochschule für Wirtschaft und Recht Berlin, Germany
margarita.elkina@hwr-berlin.del, http://www.hwr-berlin.de
[2] Hochschule für Technik und Wirtschaft Berlin, Germany,
albrecht.fortenbacher@htw-berlin.de, http://www.htw-berlin.de
[3] Beuth Hochschule für Technik Berlin, Germany,
merceron@beuth-hochschule.de, http://www.beuth-hochschule.de

Abstract: LeMo is an open source application for learning analytics, which collects data about learners' activities from different platforms. This article describes design principles of LeMo in the context of creating an efficient tool for learning analytics. Focus is on the LeMo system architecture, user path analysis employing algorithms of sequential pattern mining, and visualization of learners' activities implemented in the current version. A case study shows first results.

Keywords: learning analytics, visual analytics, explorative visualization, sequential pattern mining, user path analysis.

1. INTRODUCTION

The Horizon Report 2011 [1] identifies learning analytics as an emerging technology, which most likely will have a significant impact on higher education within the next years [2]. On the first LAK conference 2011 [3], learning analytics was defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs".

In 2011, the research project "LeMo: monitoring of the learning process on personalizing and non-personalizing learning platforms" started at three universities of applied sciences in Berlin [4]. It aims at developing an analytics and visualization tool, to facilitate learning analytics for teachers, E-learning providers and researchers. Basic requirements were connectivity to various platforms for online learning, powerful analysis and mining of data obtained from the platforms, and an intuitive visualization of analytic results, providing insight into the learning process.

The following section describes the methodical approach of LeMo, which guided requirements analysis and refinement throughout the project. Then, basic design principles of the LeMo application are explained. The following sections focus on user path analysis and visualization. In a case study at HWR Berlin, LeMo was used in a regular class, showing applicability of LeMo and its usability for teachers. The last section concludes the findings of the LeMo project, and refers to future work.

2. METHODOLOGY

Requirements analysis in LeMo is based on a survey amongst stakeholders (project partners), representing the three target groups teachers, E-learning providers and researchers. This generic qualitative method was targeted at supporting teaching with a variety of didactic methods and different degrees of technology support, e.g. E-learning as a supplement to face-to-face teaching, or online learning as a basis for distance learning.

The result of the survey was a catalogue of about 80 questions and assumptions regarding students' learning behavior and the use of online media. Some of the questions were beyond the scope of an analytics tool which relies on learners' activities obtained from a learning (management) platform.

The catalogue of questions was segmented into six categories representing different areas of interest. The first two categories contain questions about users and groups of users, as well as learning objects and type of learning objects. Interactions with the learning environment are modeled through the categories "usage analysis", "user path

analysis", "communication and collaboration", and "performance".

Based on a catalog of questions provided by the LeMo project partners, 27 indicators were derived, that provide data analysis to help answering the referred questions. The indicators "activity / time" and "activity / learning object" are associated with the category "usage analyses", visualizing how students access learning material. The indicators "frequent paths" and "activity graph" belong to the category "user path analysis", where visualization shows user activities ordered by time.

The survey showed a considerable interest in analyzing the sequence of user interactions. This interest can be contextualized and supported by the theoretical approach of connectivism. According to G. Siemens, learning is organized in a network – internally as a neural network and externally "as a network of nodes, representing a specialization and aggregates as actions concerning the environment" [5]. In a first step, this network is visualized as an "activity graph" in LeMo.

Consequently, the four indicators mentioned above were the first indicators realized in LeMo. Whereas "usage analysis" is quite common in comparable tools, the indicators "frequent paths" and "activity graph", together with intuitive visualization, are unique characteristics of the LeMo tool.

3. THE LEMO APPLICATION

3.1. SYSTEM ARCHITECTURE

LeMo is designed as a 3-tier application, which facilitated distributed development among the three universities [6]. The first tier (data) contains a data model as well as functionality for data management, which includes connectivity to different learning platforms. Data analysis takes place in the second tier (data). Results of data analysis are passed to the third tier (presentation), where the data are visualized.

Filtering data, e.g. selecting male or female students within a course, is presently done in the data tier, but eventually should be performed in the third tier, thus reducing the number of requests to the data tier.

The LeMo application consists of a *data management server*, which basically comprises the first and second tier, and an *application server*. For each user interaction, a request is generated by the application server, which is passed via a web service interface to the data-management server. Both servers are realized as web apps, which could be run by a single Tomcat server, but also, for performance and security reasons, be distributed within a network.

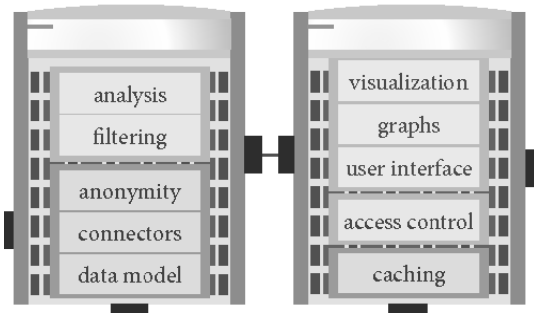


Fig. 1 – Two-server architecture.

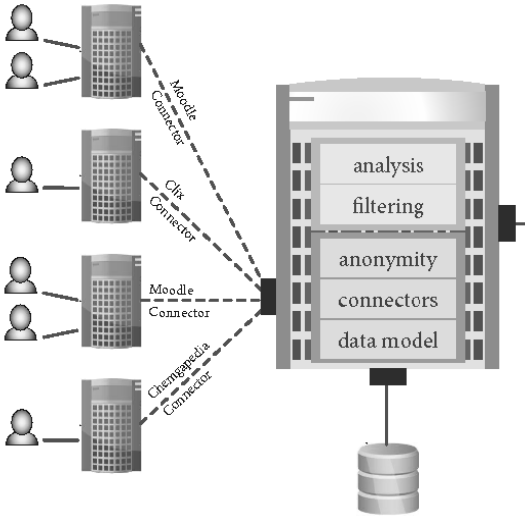


Fig. 2 – Data management server.

The data management server may connect to more than one LMS (learning management system). Presently, connectors for different versions of moodle, clix and Chemgapedia are implemented.

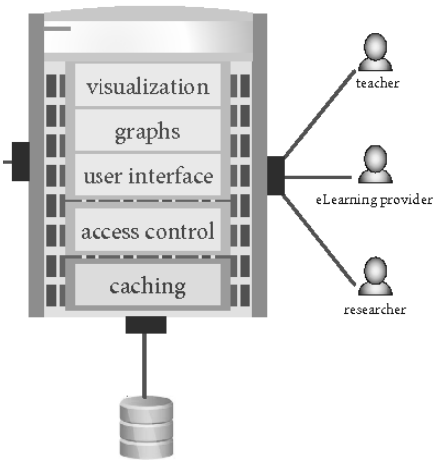


Fig. 3 – Application server.

The application server provides for visualization and user interaction. Using a JavaScript framework, this is done locally on the client, reducing server load and network traffic.

3.2. DATA MANAGEMENT

In [7], three factors for the development of learning analytics are listed: online learning, big data, and political concerns. This list is reflected by the design principles of LeMo.

Data analyzed by the LeMo tool are basically user activities obtained from platforms for online learning. LeMo connects to various platforms, including the learning management systems (LMS) Moodle[1] and Clix[2], as well as the online encyclopedia Chemgapedia[3]. The LeMo connectors for Moodle and Clix are implemented to access the underlying databases directly. In case of Chemgapedia, which is a web application, data stem from the server log file.

During the ETL (Extraction – Translation – Load) phase, data are imported into a database with a unified data model. While LMS provide detailed information on content and on users, server log files just contain page visits. Here, session information is needed to identify anonymous users (students). The LeMo data model contains entities for learning objects, e.g. courses, resources, wikis, or tests, and associations between learning objects, e.g. resource X belongs to course A. There exist students and teachers related to courses. Degree programs can group courses, a department is responsible for different degree programs, and an institution consists of several departments. The data model also includes "user activities", which basically are represented as tuples *[u,l,t,a]* (user, learning object, time stamp, action). An "action" may be one of "view", "download", "modification", "creation", "attempt", or "submit". And finally, platforms are represented in the data model. Given an element *e*, which was imported from platform *x*, a unique primary key for this element can be constructed from the pair *[e,x]*. Thus, it is possible to aggregate user data from different platforms into a single data model.

In the case of Chemgapedia, some data are not available, examples being department/faculty, or a student's enrollment in courses. This leads to sparse data, and eventually to a reduced number of available analyses.

Big data denote huge sets of, generally unstructured, data, which typically originate from social networks, and which are processed using analytic methods ("social media analytics"). Recording activity data on platforms for online learning, over a long period of time, a huge amount of data is produced. For interactive analytics applications, it is critical to handle big data efficiently, which can be achieved via data compression, efficient data retrieval and efficient mining algorithms.

Reviewing usage statistics on the Chemgapedia web server, it became apparent that log data have to be preprocessed in order to exclude activities not corresponding to learners (e.g. traffic generated by web crawlers). To solve this problem, an optional functionality was added which excludes all log entries with one of the following characteristics: multiple accesses per second, frequent repetition of time intervals between accesses, and many accesses to the same page. The use of this filter reduced the number of log entries by about 47 %, while excluding just 5 % of all users [8].

With the actual LeMo prototype, data are stored in a relational data base. First studies show that this is sufficiently efficient for institutions with about 10,000 students. Non-standard ways of storing data, e.g. in a NoSQL data base, will be investigated in the future.

3.3. DATA PRIVACY

One political concern of learning analytics is data privacy. Traditionally, personalized interactions and user modeling have significant implications on data privacy. Personal information about a user is collected and analyzed, which might not be in the interest of the user. Recently, collection of user data in social networks, and possible violation of data-privacy legislation, have been published frequently. As a result, even more strict data-privacy regulations at universities are established, which in turn limits the use of learning analytics.

The LeMo approach towards data privacy aims at achieving a high level of anonymity. The activity sequence, i.e. the "learning path" of a particular student may be analyzed, but personal data, which could serve to identify the student, must not be exposed. This is done by omitting all personal data in the ETL process, the only exception being gender information which is valuable for learning analytics. Furthermore, small data samples must not be used for a specific analysis, if this leads to the identification of a particular student (k-anonymity [9]). An example is a course with just one female (male) participant; this student could be identified easily using a gender filter.

For most of the analyses, courses must be identified by their name. Also, the title of learning objects like learning material (e.g. powerpoint presentations), the name of threads in a course-related forum, or results of a test provided in the course, are essential for efficient analysis, but must not allow for identification of learners (students).

1 https://moodle.org/
2 http://www.im-c.de/
3 www.chemgapedia.de/

To implement a role "teacher", courses taught by a given teacher must be identified. This cannot be done using some unique attribute, like name, personal ID or login name (user ID). All these attributes are omitted during the ETL process. Instead, a mapping from teachers to courses can be realized by a pseudonym (hash value) derived from a unique attribute during ETL. If the LeMo application uses the same authentication scheme as the connected platform (e.g. authentication via LDAP, the light way directory access protocol), the pseudonym is derived from the login name. When a teacher logs in, she or he sees the list of "his" or "her" courses.

4. USER PATH ANALYSIS

Computing frequent paths in LeMo is based on two algorithms, namely BIDE [10] and Fournier-Viger [11]. Implementations for both algorithms are provided by the SPMF (sequential pattern mining framework) library[4].

Given a course with a set of learning objects L, and a set of users (students) U, we define a path as a sequence of learning objects, allowing multiple occurrences of the same learning object, and L^* as the set of all possible paths (in terms of formal languages, each path is a word over the alphabet L). By $P(u)$ we denote the sequence of all learning objects accessed by user u, in the correct order implied by the time stamp of the corresponding activity.

$$p_1 = s_1 q_1 s_1 q_1 f_1 s_2 q_1 \quad (1)$$

As an example consider a course containing the following learning objects: two sets of slides s_1 and s_2, two quizzes q_1 and q_2, and one forum f_1. Let p_1 be a path and u be a User with $P(u) = p_1$. This corresponds to the following behavior: The user u first accesses the slides s_1, then quiz q_1, then again s_1 and the same quiz q_1. Then she visits the forum f_1, accesses slides s_2 and finally again quiz q_1 (1).

A path $p = l_1 \dots l_m$ is *contained* in another path $p' = l'_1 \dots l'_n$, if it can be embedded into p', preserving the order of the learning objects:

$$p \sqsubseteq p' \Leftrightarrow \exists 1 \le i_1 < \dots < i_m \le n : l_j = l'_{i_j} \forall 1 \le j \le m \quad (2)$$

A path $p = l_1 \dots l_m$ is a *subsequence* of another path $p' = l'_1 \dots l'_n$, if it can be embedded into p' directly, preserving the original sequence of p:

$$p \subseteq p' \Leftrightarrow \exists 0 \le k \le n - m : l_j = l'_j + k \forall 1 \le j \le m \quad (3)$$

4 http://www.philippe-fournier-viger.com/spmf/

Let's look at two paths

$$p_2 = s_1 s_2 q_1 \quad (4)$$

$$p_3 = s_1 q_1 f_1 s_2 q_1 \quad (5)$$

The path p_2 could reflect the teacher's intended "learning path", i.e. first accessing slides s_1, then s_2, and then trying to pass quiz q_1. p_2 is contained in p_1 (at positions 1, 6, 7 or 3, 6, 7), but p_2 is not a subsequence of p_1. The path p_3 could reflect the following learning activities: read slides s_1, go to quiz q_1 (and fail), go to the forum f_1 and ask for help, read slides s_2 and pass the quiz q_1. p_3 is a subsequence of path p_1, starting at position 3.

For both relations "is contained" and "is a subsequence", we can define the set of frequent user paths:

$$FP_{\sqsubseteq}(s) \Leftrightarrow \{p \in L^* : |\{u \in U : p \sqsubseteq P(u)\}| \div |U| \ge s\} \quad (6)$$

$$FP_{\subseteq}(s) \Leftrightarrow \{p \in L^* : |\{u \in U : p \subseteq P(u)\}| \div |U| \ge s\} \quad (7)$$

With a *support* value of *0.5*, the set of frequent paths consists of all learning object sequences, which are followed by at least 50 % of all users.

Given four users with learning paths

$$P(u_1) = s_1 s_2 q_2 q_1 \quad (8)$$

$$P(u_2) = q_1 q_2 f_1 s_1 s_2 \quad (9)$$

$$P(u_3) = f_1 \quad (10)$$

$$P(u_4) = s_1 q_1 s_1 q_1 f_1 s_2 q_1 \quad (11)$$

we get the following frequent paths:

$$FP_{\subseteq}(0.7) = \{s_1, s_2, q_1, f_1\} \quad (12)$$

$$FP_{\sqsubseteq}(0.7) = \{s_1, s_2, q_1, f_1, s_1 s_2\} \quad (13)$$

In this example, all paths are followed by at least 3 users.

The BIDE algorithm computes frequent paths with respect to the relation "is contained in" (6). Given a constant support value, the complexity of BIDE is $O(n_1 n_2 n_3)$, where $n_1 n_2 n_3$ is the product of the number of users $|U|$, the average and the maximal length of user paths $P(u)$. Since BIDE uses an apriori approach to determine reoccurring sub-paths, the processing time increases dramatically when the number of candidate sequences is large [10].

The Fournier-Viger algorithm computes frequent paths with respect to the relation "is a subsequence of" (7). It was developed for a specific application domain, namely learning with tutorial systems for

robotics, and extends the well-known PrefixSpan algorithm [12]. Fournier-Viger obtains significantly less paths than the BIDE algorithm. Since the number of candidate sequences is low (compared to BIDE), the algorithm can be easily applied to low support values, which is not the case for BIDE.

Fig. 4 and Fig. 5 illustrate the difference between these two approaches, both in the results they deliver and in their performance. Both algorithms have been run on exactly the same data set with the same minimal support of 50 %. The result of the BIDE approach is shown in Fig. 4: 20 frequent paths of length 3 have been found. Different learning objects are shown with different colors, making it easy to identify the same learning object within one or several paths. Clicking on a path (Fig. 4: the left path) shows the names of the learning objects. At least 50 % of the students followed the path: *resource1.1.1.6 resource1.1.1.6 resource1.1.1.5*.

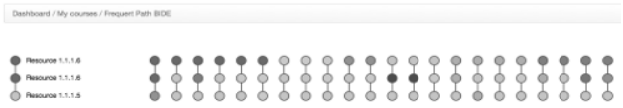


Fig. 4 – Frequent paths with BIDE.

Fig. 5 – Frequent paths with Fournier-Viger.

The result of the Fournier-Viger approach is shown in Fig. 5: For a support value of 50 %, 10 paths have been found, among them the paths (with one learning object) *resource1.1.1.5* and *resource 1.1.1.6*. The maximal length is 1.

Combining the two results, we conclude that all students who accessed *resource1.1.1.6*, later again *resource1.1.1.6* and later *resource1.1.1.5* have all accessed different learning objects in between. Lowering the support to 10 % gives the results shown in Fig. 6 obtained by the Fournier-Viger algorithm, while after 10 minutes the BIDE algorithm is still calculating.

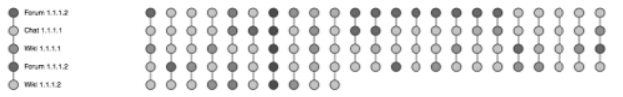


Fig. 6 – Fournier-Viger with lower support.

The drastically increased processing time of BIDE for low support values poses a serious problem to the usability of the LeMo application. As a solution, a "default value" for minimal support is determined, depending on parameters "number of user paths", "average length of user paths", and "number of learning objects" within a given course.

A typical users' behaviour is to try out low support values (which might cause BIDE to run "for ever"), and then increase the values until a result is achieved within a reasonable response time. This might start a couple of BIDE algorithms running as parallel threads, and eventually bring down the system. To avoid this, the algorithm must be terminated when a user requires the next analysis. Thus, in a multi-user environment, the number of running BIDE algorithms is limited to the number of active users. Furthermore, after a predefined lapse of time, the algorithm is terminated, which does not leave hung threads in the system.

5. VISUALIZATION

The design of the user interface in LeMo follows the explorative metaphor "overview first, zoom and filter, then details-on-demand", which assists analysis of learning data in an explorative way [13]. Especially when displaying large amounts of data, this approach supports a highly intuitive reception of the results of analyses.

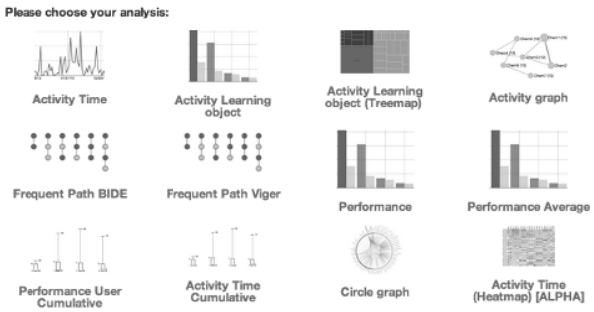


Fig. 7 – Different analyses / visualizations.

LeMo provides a variety of different visualizations, which are suitable for analysis of specific learning situations and processes. For example, the performance of students based on self-tests can be visualized. Alternatively, students' performance could be used as a filter for other analyses, exploring correlations between performance and navigation [14].

Visualization was designed according to the following guidelines [13]:

- **pre-attentive perception**: use of visual attributes, e.g. shape, size, color and position
- **provide details on demand**: detailed information on a specific learning object is available (context sensitive), without overloading the initial visualization

- **facilitate exploration**: the interface provides the possibility to customize the visualization, to best fit the personal expectations, through features like translation, rotation, filtering, and zooming

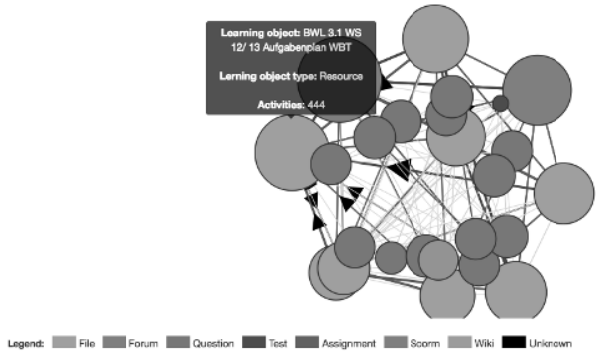


Fig. 8 – Navigation graph.

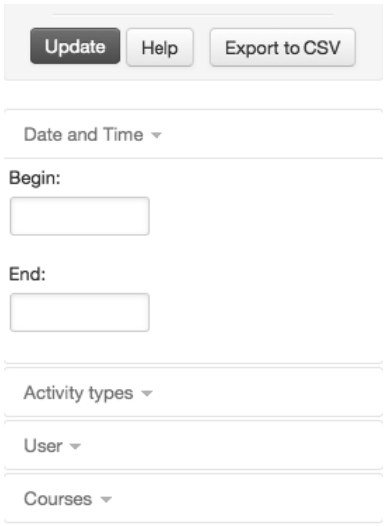


Fig. 9 – Filters.

Zooming and filtering can be applied to various visualizations, e.g. “activity / time”, “activity / learning object”, “treemap”, “frequent paths” or “activity graph”. Also, results of a self-test (quiz) can be used to filter user activities. The following example shows the application of the filters “learning object type” and “students”, as well as zooming into areas of interest.

Activities over time show a pattern, which corresponds with the concept of the course (blended learning).

Zooming provides more insight into a peak of activities around 2013-11-26.

Selecting arbitrarily (anonymous) students gives hints on different learning behaviour.

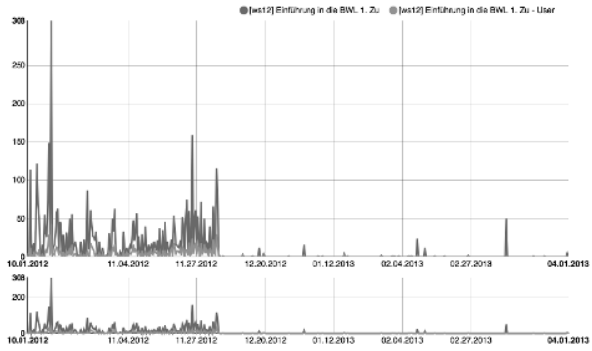


Fig. 10 – Course “Betriebswirtschaft”, 2012/13.

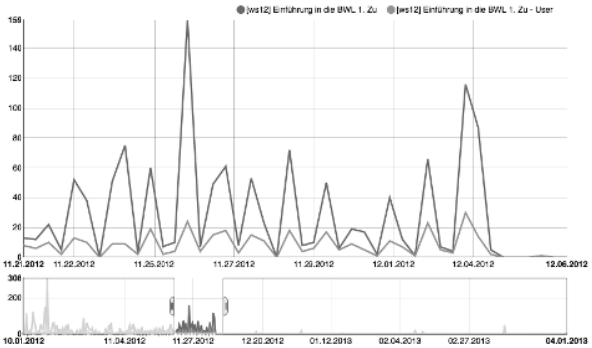


Fig. 11 – Zooming in.

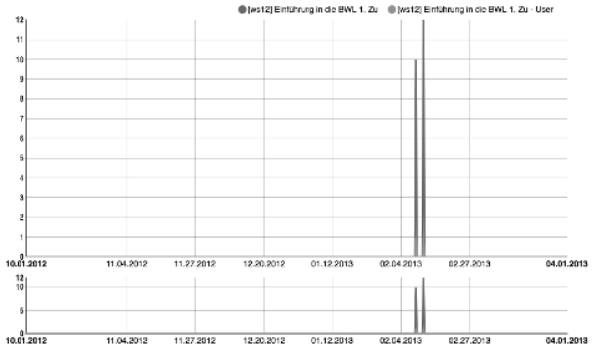


Fig. 12 – Selecting a single student.

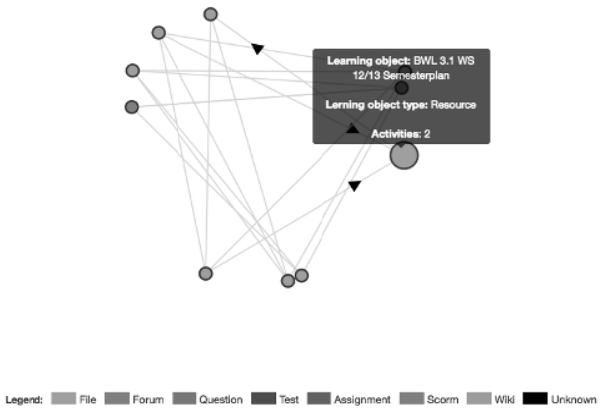


Fig. 13 – Single students' navigation.

The analysis "activity graph" provides a lot of information about the accumulated navigation behavior of all students, which in cases proves difficult to analyze and interpret. Selecting the same student as above, a selective usage of learning objects is displayed.

The results of an analysis "activity graph" are visualized by a navigation network. This analysis can be used to obtain information on the degree of cross-linking between learning objects. The activity graph is visualized by a graph, with learning objects as nodes and navigation steps as edges. Nodes are color-coded to support the visual perception of content type transitions. Learning objects are adjusted in size to encode the absolute number of user requests. Edges are weighted and color-coded to encode the amount of navigational steps.

Detailed information on specific elements of the graph can be made visible by interacting with the visualization, including tool tips with information on learning objects, and rearranging the graph, focusing on nodes of interest.

For the visualization of frequent paths (results of the BIDE algorithm, or Fournier-Viger algorithm), the following premises are important:

- A path is visualized by a sequence of nodes (learning objects) and edges (navigation steps). Learning objects are color-coded to allow a visually easy correlation of navigational patterns across different paths.
- Detailed information on specific elements of the path is visible through user interaction.

In some cases, LeMo provides even different "visual styles" for the same analysis, an example being "activity graph". The first version features a volatile visualization of the navigation graph, which facilitates user interaction (Fig. 8). The second version is a more static visualization, which, as an overview, just displays learning objects, where the size of the circles corresponds to the number of related user activities. Edges are omitted - selected edges appear, when a specific learning object is selected. The two visualizations support different ways of perceiving information, allowing users to develop their personal style of doing learning analytics [15].

6. A CASE STUDY

The case study is based on a bachelor degree course in informatics, which used Moodle as a learning platform. In addition to the standard file upload (slides, documents), the Moodle course contained quizzes and a wiki for individual work. The questions for the quizzes, and the assignments in the wiki were formulated by the students: each student provided one question or one assignment. The forum in Moodle was not used frequently, since this course assumed the presence of students in the lectures. The goal of this case study was an analysis of the usage of quizzes and wiki, which extends the regular possibilities provided by Moodle.

6.1. ACTIVITY / TIME

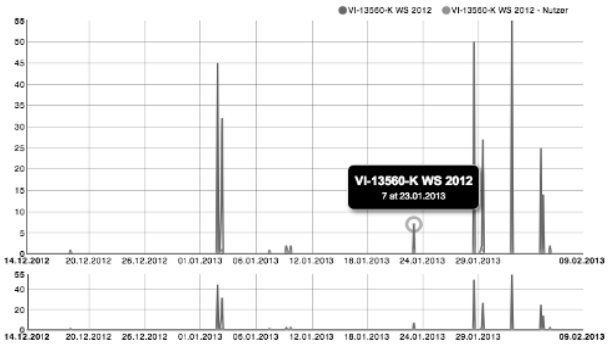


Fig. 14 – Actitivity / time (quizzes selected).

The upper part (blue) displays the number of activities, the lower part (orange) indicates the number of users who started these activities. Tool tips show the exact numbers of activities and users, and the exact time. The scale of the axis is adapted to the current value range.

Fig. 16 shows activites on quizzes, which can be achieved by the filter "learning object type". The analysis of activities indicates, that students' activities are rather high at the end of the semester. There was no examination at this time, the final examination is at the end of the second semester. It appears that the students were motivated to test their knowledge in order to evaluate themselves.

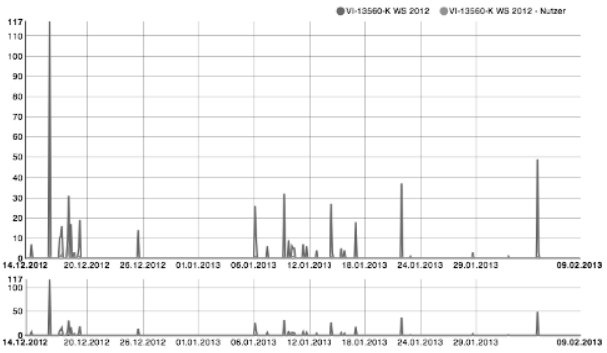


Fig. 15 – Activity / time (wiki selected).

Part of the didactic concept was to use a wiki for student collaboration. After the deadline for submission of the assignments, which had to be provided via the wiki, continuous (optional) activities in the wiki showed that students accepted this concept.

6.2. ACTIVITIES DURING THE WEEK

LeMo offers the visualization "activity time cumulative", which displays all activities during the week as box plots. Optionally, each day can be split into four slots, with six hours each. The box plots display the span of activities during the weekdays, as well as quartiles and median of the distribution.

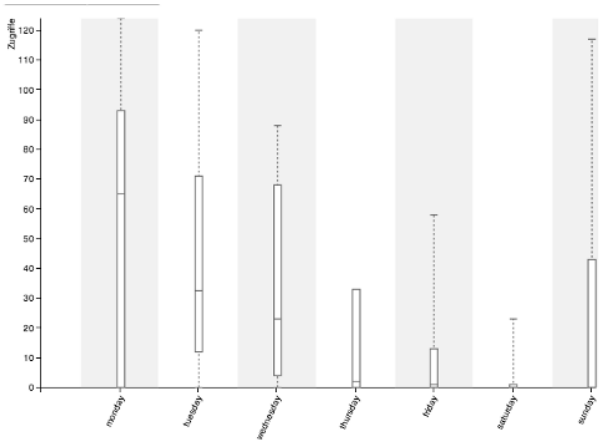


Fig. 16 – Cumulative activites.

From the visualization, it is obvious that students started to learn on Sunday, became most active on Monday and stayed rather active on Tuesday. Saturday is the day with the lowest activity. At first sight, Sunday activities were unexpected, although it can be explained logically: on Monday, students attend other courses as well. The slides for the next lecture were always made accessible on Monday morning; didactically, it would make more sense to provide access already on Saturday.

6.3. RESULTS OF THE CASE STUDY

The analysis offered by LeMo could be adapted to the particular course, so that the expected results of the learning scenarios could be tested. Interesting aspects of the learning behavior of the students have been disclosed, providing hints on how to improve of the courses. These conclusions were drawn mostly from a combination of the various visual analyses in LeMo.

7. CONCLUSION AND FURTHER WORK

The actual LeMo prototype proved to be a useful tool for teachers, giving hints on the learning behaviour of their students, and helping with the evaluation of their didactic concept. The combination of user path analysis combined with a powerful interactive visualization distinguishes LeMo from other tools for learning analytics.

Focus of future work will be on improving the LeMo application with respect to efficiency and usability, as well as on enhancing data analytics and visual analytics. System improvements will include a standardized ETL interface, connectivity with further platforms, a (non-standard) data model for efficient data retrieval, and additional pattern-mining and data-mining algorithms.

Functionality will be enhanced by realizing more indicators derived from the stakeholders' original list of requirements (questions). Of special interest are indicators referring to "group of users", "students' performance", and "communication and collaboration".

The actual prototype realizes the role "teacher", where a teacher can analyze his/her own courses. In the future, the roles "E-learning provider" and "researcher" should be defined and implemented.

For all improvements on the functionality and usability of the LeMo tool, an evaluation study is essential and will be conducted in the immediate future.

8. REFERENCES

[1] L. Johnson et al., *The 2011 Horizon Report*, The New Media Consortium, Austin, 2011, http://www.nmc.org/pdf/2011-Horizon-Report.pdf.

[2] A. Fortenbacher, Learning analytics for higher education – perspectives and challenges, in *Pratsi: Scientific, Science and Technology Collected Articles*, Issue 1(40), Odessa National Polytechnic University, 2013, pp. 184–187.

[3] *1st International Conference on Learning Analytics and Knowledge*, ACM Digital Library, 2011.

[4] L. Beuster et al., Learning analytics und visualisierung mit dem LeMo-Tool, in *Proceedings of Interaktive Vielfalt – DeLFI*, Bremen, 2013.

[5] G. Siemens, Connectivism: Learning Theory of Pasttime of the Selfamused, http://www.elearnspace.org/Articles/connectivism_self-amused.html. 2006.

[6] A. Fortenbacher, L. Beuster, M. Elkina, L. Kappe, A. Merceron, A. Pursian, S. Schwarzrock, B. Wenzlaff, LeMo: a Learning Analytics Application Focussing on User Path Analysis and Interactive Visualization, *The 7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS'2013)*, 12-14 September 2013, Berlin, Germany, Vol. 2, pp. 748-753.

[7] R. Ferguson, The state of learning analytics in 2012: a review and future challenges, Tech. rep. KMI-12-01, Knowledge Media Institute, The Open University, UK, 2012.

[8] S. Schwarzrock, Analysis of learner navigation on web-based platforms using algorithms for sequential pattern mining, in *Pratsi: Scientific, Science and Technology Collected Articles*, Issue 1(40), Odessa National Polytechnic University, 2013, pp. 18–21.

[9] V. Ciriani et al., K-Anonymity, in *Secure Data Management in Decentralized Systems*, Ed. by T. Yu and S. Jajodia, Vol. 33, Advances in Information Security, Springer, Berlin Heidelberg, 2007, pp. 323–353.

[10] J. Wang and J. Han, Efficient mining of frequent closed sequences, in *Proceedings of the 20th International Conference on Data Engineering*, Boston, MA, USA. 2004.

[11] P. Fournier-Viger, R. Nkambou, and E. Mephu Nguifo, A knowledge discovery framework for learning task models from user interactions in intelligent tutoring systems, in *Proceedings of MICAI*, Ed. by A. Gelbukh and E. F. Morales, *Lecture Notes in Artificial Intelligence*, Vol. 5317. Springer-Verlag Berlin Heidelberg, 2008, pp. 765–778.

[12] Jian Pei et al., Mining sequential patterns by pattern-growth: the prefixspan approach, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, Issue 11, 2004, pp. 1424-1440.

[13] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in *Proceedings of the IEEE Symposium on Visual Languages*, 1996, pp. 336–343.

[14] A. Merceron et al., Visual exploration of interactions and performance with LeMo, in *Proceedings of the 6th International Conference on Educational Data Mining*, Memphis, USA, 2013.

[15] L. Beuster et al., Prototyp einer plattformunabhängigen Learning Analytics Applikation – fokussiert auf Nutzungsanalyse und Pfadanalyse, in *E-Learning Symposium 2012 Aktuelle Anwendungen, innovative Prozesse und neueste Ergebnisse aus der E-Learning-Praxis*, Ed. by U. Lucke, Universitätsverlag Potsdam, 2012, pp. 69–72.

Prof. Dr. Margarita Elkina *is teaching the courses on informatics, program and software development at the Berlin School of Economics and Law, Berlin, Germany. After obtaining her PhD degree in applied mathematics at the Lomonosov Moscow State University, she was teaching at the university, then turned to industry and for over 10 years performed research in software development. She managed many German and international projects, designed graphical user interfaces and implemented different software solutions. Starting from 2011, she also belongs to the management of the project "Monitoring of learning processes in personalizing and non-personalizing learning platforms". Currently her main research interests are Learning Analytics, Data Mining and Visualization.*

Prof. Dr. Albrecht Fortenbacher *graduated in computer science from the University of Karlsruhe, Germany. After one year at IBM Thomas J. Watson Research Center, Yorktown Hights, USA, where he contributed to the IBM computer algebra system Scratchpad/AXIOM, he received in 1988 his Ph.D. from University of Karlsruhe. Then, Albrecht Fortenbacher joined the IBM Scientific Center in Heidelberg, where he worked, and did research, on supercomputing and parallel computation.*

In 1994, Albrecht Fortenbacher was appointed professor for distributed and secure systems at the HTW Berlin, University of Applied Sciences. His research focus is on Teaching Learning Technology, recent projects range from E-learning infrastructure (LMS), ePortfolios, virtual classroom and ad-hoc meetings, to learning analytics. Albrecht Fortenbacher is active in the German learning analytics community, with a focus on introducing learning analytics to universities.

Prof. Dr. Agathe Merceron *holds a Ph.D in Computer Science of the University Paris VII. She joined Beuth University of Applied Sciences in 2006 and is the head of the online degrees Computer Science and Media, Bachelor and Master as well as head of the Online-Learning Laboratory. She teaches courses such as programming, foundation of computer science, algorithms or data mining. Her research interests presently focus on Information Systems and Knowledge Management – application to E-Learning, Technology Enhanced Learning, Educational Data Mining, Learning Analytics. She is a board member of the International Educational Data Mining Society and an associate editor of the International Journal of Educational data Mining.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

E-LEARNING ENVIRONMENT FOR THE REMOTE STUDY IN MATERIAL PROPERTIES COURSES

Peter Arras [1], Yelizaveta Kolot [2], Galyna Tabunshchyk [2], Tomas Kozík [3]

[1] Faculty of Engineering Technology, KULeuven-campus De Nayer,
Sint Katelijne Waver, J. P. De Nayerlaan 5, Belgium,
E-mail: peter.arras@kuleuven.be, http://iiw.kuleuven.be
[2] Software Tools Department of Zaporizhzhya National Technical University,
Zhukovskogo, 64, Zaporizhzhya, 69063, Ukraine,
E-mail: galina.tabunshchik@gmail.com, http://www.zntu.edu.ua
[3] Department of Technology and Information Technologies, Faculty of Education,
Constantine the Philosopher University in Nitra,
Dražovská 4, 949 01 Nitra,
E-mail: tkozik@ukf.sk, http://www.ukf.sk

Abstract: In this paper a newly developed e-learning environment for the support of the study on Material Properties is presented. It was developed to support the blended learning of the material and shape stiffness. Course structure is organized in HTML content, and virtual and remote laboratories are integrated in the computer aided learning module (CALM), which support both teachers during the hand-on teaching and students during self-study.

Keywords: CALM, remote laboratory, virtual laboratory, material stiffness, shape stiffness.

1. INTRODUCTION

The field of material science is defined as a key knowledge area for engineers. Many – if not all – innovations in modern technologies are related to the use of new materials, or new technological methods of using existing materials. However, study time in existing curricula is limited and new findings in material sciences cannot be fitted in. E_learning modules could help make up for this lack of time. This paper is devoted to the Computer Aided Learning Module (CALM) [1], developed for the supporting e-learning courses in Material Science Properties.

2. STRUCTURE OF THE COMPUTER AIDED LEARNING MODULE

A complete driving learning module for material sciences is constructed. The structure of it can found in Fig. 1. It is the framework in which fits the theoretical courses, and the different laboratories described.

The present remote lab in the Computer Aided Learning Module (CALM) is intended to be used to study the phenomenon of material versus shape stiffness. The CALM contains all the learning contents students are supposed to study on the subject. The CALM will be used in classroom teaching and demonstrated to instruct students on how to use the remote lab.

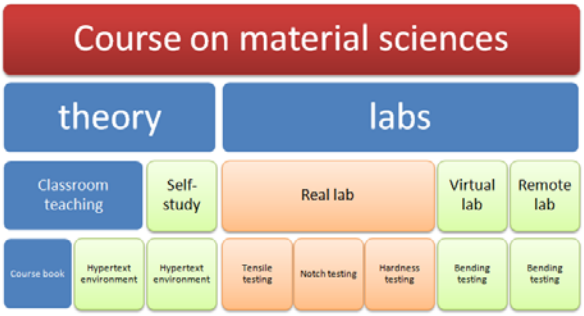


Fig. 1 – Structure of the course on material science, integrated in the CALM.

Students will study theory afterwards at their own pace, using the hypertext environment, and next simulate different combinations of materials and shapes in the virtual laboratory. Finally students can experiment on real material and shapes in the remote laboratory.

Different techniques are combined: hypertext linked contents, virtual lab (using Shockwave Flash)

and a remote lab (controls for it using Javascript) (Fig. 2). Slideshows are provided for documentation purposes and to include the classroom presentations (Fig. 3).

The structure of the theory on the CALM reflects the materials in the student's physical course books. As such the look and feel of the e_learning environment is similar to what students experience in the classroom sessions [2].

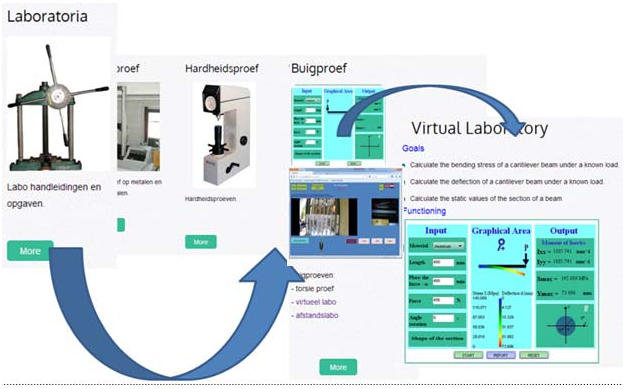


Fig. 2 – Navigation in the hypertext.

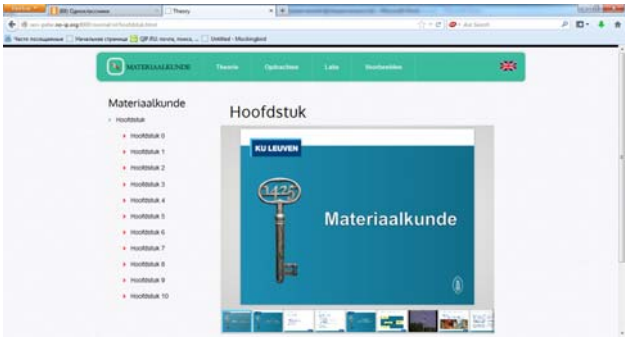


Fig. 3 – Lectures presentation.

3. VIRTUAL LABORATORY

The virtual laboratory is a simulation of a test in the physical laboratory for material properties study [2]. It was constructed using the basic formulas for the bending of a cantilever beam. In the virtual lab a number of specimens with standard sections can be tested. Applied force, shape is changeable to offer a wide variety of possibilities. The only limitation with the basic formulas is that asymmetrical test specimens will yield erroneous values, because bending outside the plane of the specimen is not considered in the basic formulas (Fig. 4).

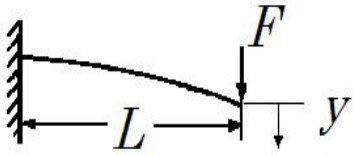


Fig. 4 – Deflection of a cantilever beam.

Used formulas for bending of a cantilever beam [3]: Maximum deflection at the free end:

$$y = \frac{F \times L^3}{3 \times E \times I} \tag{1}$$

Maximum bending stress in the cantilever beam (at the supported end):

$$\sigma = \frac{F \times L}{I / v} \tag{2}$$

Moment of inertia (I): e.g. of a rectangle:

$$I = \frac{b \times h^3}{12 \times I} \tag{3}$$

Table 1. Used symbols.

Used symbols		
Symbol	***Name***	***Unit***
E	Young's modulus	MPa
I	Moment of inertia	mm^4
F	Load	N
s	Maximum bending stress in the cross-section	MPa
y	Maximum deflection	mm
L	Length of beam	mm
v	Distance to neutral fiber	mm

3.1. FUNCTIONAL REQUIREMENTS FOR THE VIRTUAL LAB

1. Calculation and visualization of the deflection of a cantilever beam, with measurement of deflection at the end and indication of maximum stress and force.

2. The values for different Young's modulus for a variety of materials should be chosen according to reported values [3].

3. For a variety of shapes (rectangular section, H-beams, U-beam, hollow shapes…) the shape stiffness can be used.

4. For different orientations of the same shape (increment angles 1°) about the principal axes.

5. For a variety of dimensions in the same shape. (cross section dimensions and length). The different values should be entered by the student experimenter.

3.2. SUPPLEMENTARY REQUIREMENTS FOR THE VIRTUAL LAB

Experimenting students should be able to select all different possibilities with the aid of pull down

menus and selection panels, and dialogue boxes for dimensions.

Readings of deflection, force and stresses with the aid of look a like analogue meters.

Force should be applicable with the aid of a slider bar, and real time adjustment of readings should be taken care of.

Readings should be exportable: available in a window from which to copy, together with material and cross section data. This feature is necessary for the students to make their reports. As such they can copy the data without retyping, but also without any special formatting in the virtual lab, so it can be fit in any report.

3.3. VIRTUAL LABORATORY REALISATION

For the virtual lab realisation were used HTML5, ActionScript and JavaScript. The component diagram of developed application is presented on the Fig. 5.

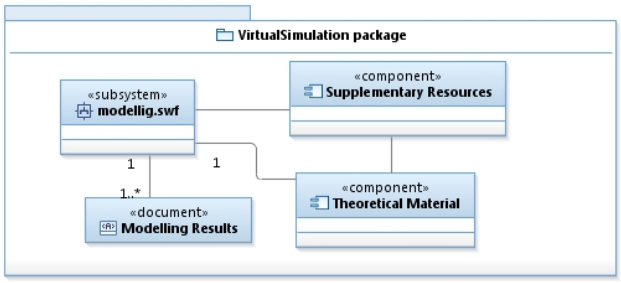


Fig. 5 – Component Diagram for Virtual Lab.

Subsystem "Modelling" is an interactive animation, based on Flash and ActionScript. Its consists of 3 components: a visual component, a component for calculations and the component showing different shapes. Such realizations make it possible to easily add new shapes. The realization of the calculation for different shapes is realized by classes put in package beam, which is part of subsystem "modelling" shown on the Fig. 6.

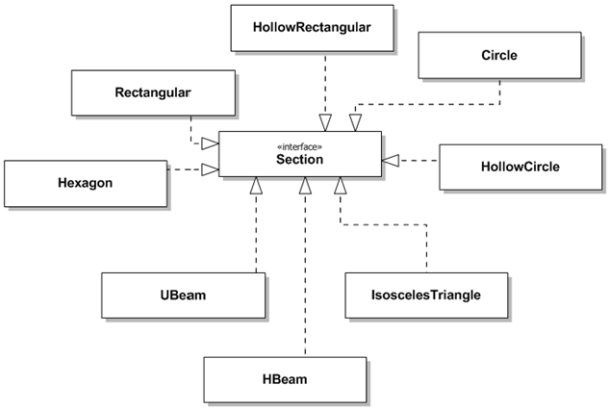


Fig. 6 – Class Diagram for the cantilever beam calculation depending from the shape.

3.4. VIRTUAL LABORATORY FUNCTIONALITY

To run the experiment on the client side user should use any web-browser with an installed Flash Player. The main screen of the virtual lab is in Fig. 7.

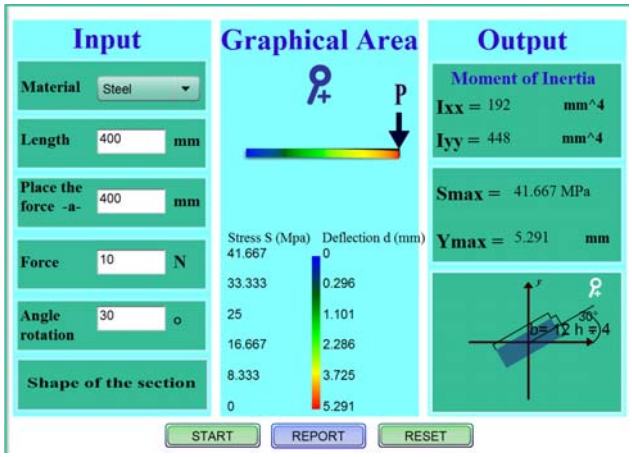


Fig. 7 – Screen shot of the actual virtual lab.

All results can be viewed in an additional window and printed (Fig. 8). The report window is a simple text window which allows the student/experimenter to copy paste these results in his lab reports, without having to type all texts.

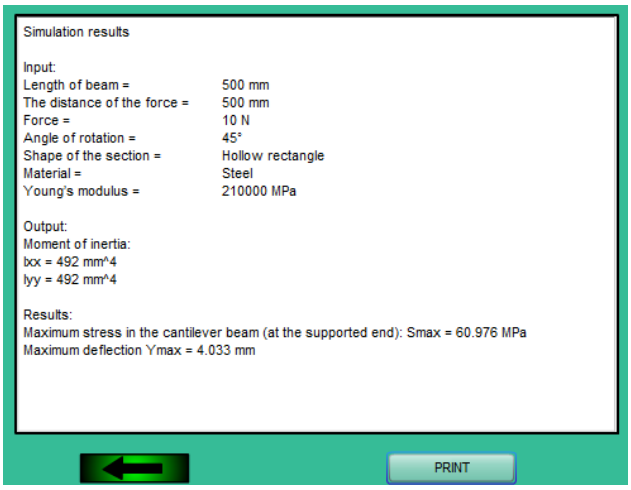


Fig. 8 – Screen shot of the report screen.

4. REMOTE LABORATORY

The objective of the experiment in the remote lab is similar to the objectives of the real physical lab on bending. First objective is to see and to measure the deflection of the cantilever beams. These values can be compared to the theoretical results from the virtual laboratory. Second goal of the physical lab which is substituted by the remote lab is to let students understand the influence of errors in

measuring values on the uncertainty of their results. Students need to calculate the (possible) error on their values by means of the theory of errors, considering all sources of errors in the experiment (accuracy of tools, uncertainty of loads, distances...)

4.1. HARDWARE CONSTRUCTION

The construction of the experiment consist of sets of beams. The beams can be any basic shape (round, square, rectangular, hollow). The beams as deformed by pulling them with pneumatic cylinders (Fig. 9). A reading scale is attached at the front end to measure the deflection by means of a camera. The reading scale resolution is 1 mm. The improve the read-out of the deflection, at the end of each beam, a blue marker was attached, to contrast as much as possible with the reading scale. Electronic control of the lab is with a simple relay board, on the USB-port of the computer.

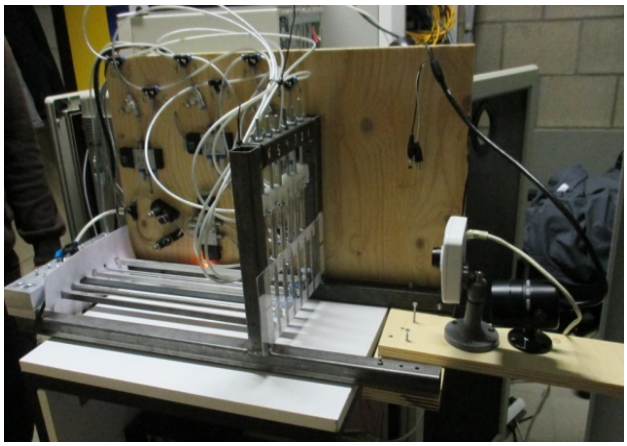


Fig. 9 – The hardware construction.

In the lab uses an IP-cam for the measurement (reading the scale) and a web-cam for board control and monitoring the deflection.

4.2. REMOTE LABORATORY REALIZATION

The control software driving the remote lab and integrating it in the CALM-website was developed with the collaboration with ZNTU, leading to a graphical user interface containing camera pictures and controls in one screen.

In general the software for the remote lab consists of three main subsystems:

- for control of the cameras;
- for board control;
- for lab control;

The subsystem for lab control is a controller which allows to control the process: initiate the session, safety control, process the logic of the experiment, close the session (Fig. 10).

The subsystem for board control task is for switching the relays on the relay board.

In Fig. 11 is shown the class diagram for these two subsystems.

The subsystem for camera control is an independent system, which allows to output video-streams using the rtsp-protocol and to save screenshots of the device with high resolution for further measurements. Using the rtsp-protocol allows to display the video in browser after installing the vlc-plugin.

For the realization was chosen the Spring MVC Framework. The basic logic was realized with java, for visualization were used HTML, CSS, Javascript, JQuery, JSP.

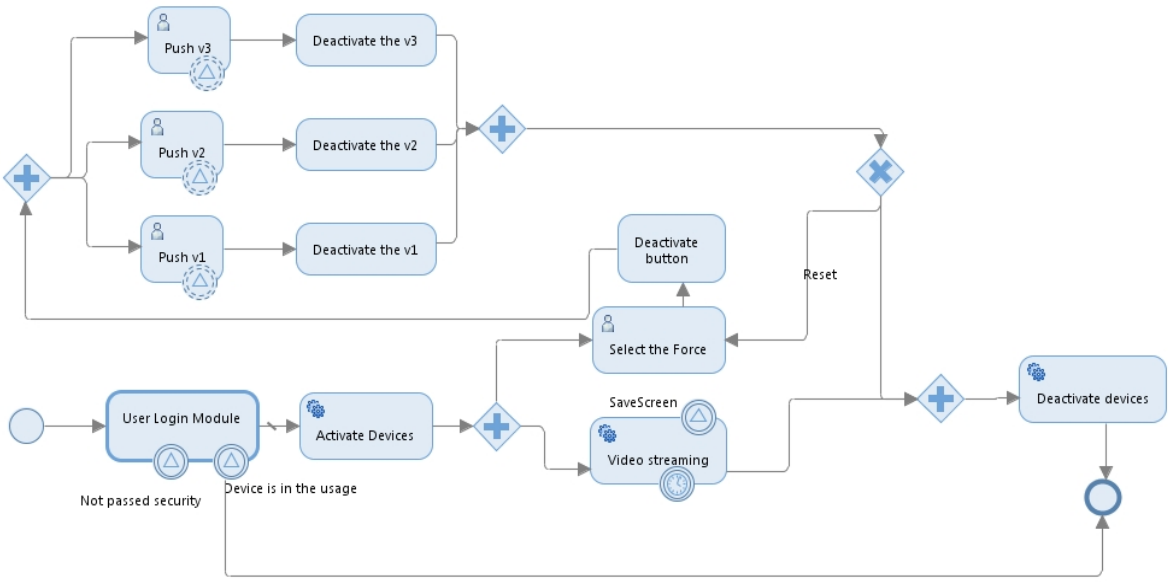


Fig. 10 – Remote Laboratory functionality.

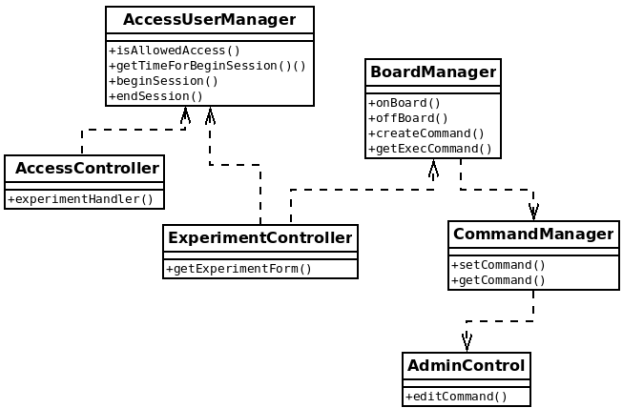


Fig. 11 – Class diagram.

4.3. FUNCTIONALITY OF LAB-CONTROL MODULE

After getting access to the remote lab student can choose the force applied to the beams and activate the forces on the different test specimen (beams) by pressing the different valve buttons.

To cope with possible problems on images and video, a section for saving screenshots was integrated for offline working with the experimental results. The different screenshots can be reviewed to measure the deflection at the end of the beams. As such experimenting time could be reduced to 10 minutes per student. Security measures are installed to avoid automated/hacked control of the lab.

The view of remote laboratory control through web interface can be seen in Fig. 12.

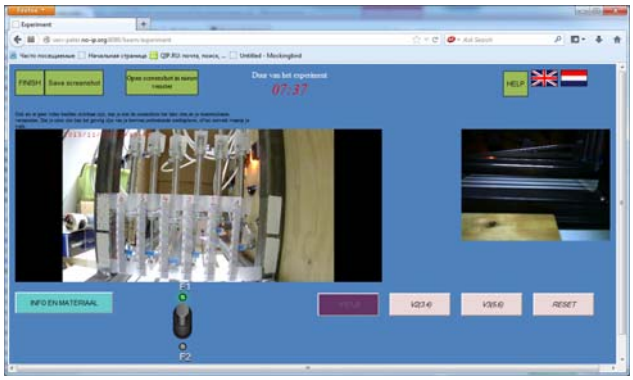


Fig. 12 – Remote Laboratory: left IP-cam for measuring, right webcam for viewing the deflection.

First evaluations of the lab show that overall functionality of the lab is appreciated by the students. Most of the problems arise with the video (speed of connection for life-streams) and the interference or no connection with plug-ins. Also the use of the lab in different browsers (Internet Explorer, Google Chrome, Firefox) show causes of ineffectiveness as all 3 browsers treat videos differently. This makes it somewhat difficult for the students, as they (sometimes) need to tune their browser first.

5. CONCLUSION

The blended learning approach shows the most efficient results among different approaches with the usage of e-learning systems. The developed e-learning environment CALM is useful both for students needs (self-study needs) as for teachers (central source of all course related materials). It consists of different e-learning tools – html content of learning material as theoretical background and recommendations for practical tasks, tools for lecture presentation, elements of experiments with material properties simulation and remote laboratory, which give possibilities for a more practical approach to learning.

In future, new tasks on expanding the CALM are to modify the architecture to incorporate new infrastructure for other remote laboratories and to enhance the security model, as there are plans to increase number of students using the remote labs. Security is necessary to ensure the safety of the equipment and to avoid early wear or destruction by unwanted use of the remote labs.

6. REFERENCES

[1] P. Arras, Computer aided learning approach for the study of the Properties of Materials, *Proceedings of the Conference Vzájomná informovanost' – cesta k efektívnemu rozvoju vedecko-pedagogickej činnosti*, 13 June 2013, Pedagogická fakulta UKF v Nitre, pp. 5-11.

[2] P. Arras, G. Tabunshchyk, T. Kozik, E-learning concept for the properties of materials remote study, *Proceedings of the 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Berlin, Germany, 12-14 September 2013, Vol. 2, pp. 742-747.

[3] R. C. Hibbeler, *Statics and Mechanics of Materials*, vol. Bending, New York: Pearson/Prentice Hall, 2004, pp. 511-544.

***Peter Arras**, drs, graduated as engineer in 1985 at De Nayer institute in Belgium. Got a pedagogical degree in 1986. Teaching design methods and numerical simulation methods in the faculty of engineering technology of KULeuven Since 2006 international studies co-ordinator at the department.*

Areas of interest in his scientific work are design related methods, CAD/CAM/CAE-systems, e_learning methods.

Yelizaveta Kolot – *PhD student of Zaporizhzhya National Technical University, received and Engineering Diploma in Informational Technologies in Design in 2013 with qualification analytic of Informational Systems. Scientific Interests – software development, reliability of informational systems.*

Galyna Tabunshchyk *PhD, Assoc. Prof. of Software Tools Department of Zaporizhzhya National Technical University, finished Zaporizhzhya State Technical University in 1998, made her PhD work in 2004 in Control Systems and Processes. The main topics of interests are Software Engineering, System Analysis, Testing and Reliability of Computer Systems.*

Tomaš Kozík, *Prof, Ing, DrSc, graduated in physics of solid states from the Slovak University of Technology in Bratislava, Slovakia in 1968. After defending his Doctoral thesis in condensed matter physics he was granted title PhD.*

In 1989 he defended his DrSc in technology of ceramic materials processing and was appointed associate professor in branch of physics of condensed compounds and acoustics at the Slovak University of Technology in Bratislava and later in 1994 full professor in branch of electrical engineering materials at the Faculty of Electrical Engineering, STU in Bratislava. Since 1990 he has been working at the Department of Technology and information Technologies, University of Constantine the Philosopher in Nitra and as an external professor at the Slovak University of Technology in Bratislava, at the Faculty of Materials Science and Technology in Trnava. His main research activities are focused on physical properties and technology of the classical and progressive ceramic materials, special glasses and plastics treatment, as well as didactics of technical vocational subjects.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

COMPARISON OF CONVENTIONAL AND SPREAD SPECTRUM SIGNALS TIME OF FLIGHT ERROR CAUSED BY NEIGHBORING REFLECTIONS

Linas Svilainis, Arturas Aleksandrovas, Kristina Lukoseviciute, Valdas Eidukynas

Signal Processing Department, Kaunas University of Technology,
Studentu str. 50-340, LT-51368, Kaunas, Lithuania
linas.svilainis@ktu.lt

Abstract: Performance comparison of conventional and spread spectrum signals in Time of Flight estimation is given. Ultrasonic measurement under multiple reflections condition is analyzed. It was indicated that if two reflections are in close proximity the neighboring signal induces energy leak into its opponent signal. Due to such situation ToF estimate is obtained with bias error. Narrow signals like single rectangular pulse, should suffer less from the aforementioned phenomena. But use of spread spectrum signals is preferred thanks to their compressibility. It was hypothesized that such long signals will have worse bias error due to neighbor reflection. Goal of the investigation was to compare the performance in multiple reflections environment in Time of Flight estimation for classical signals and spread spectrum signals. Investigation revealed that spread spectrum signals have better performance in a sense of bias error caused by neighboring reflection.

Keywords: time of flight; time delay estimation; ultrasonic measurements; multiple reflections.

1. INTRODUCTION

Ultrasonic imaging and measurement offers direct interaction of the probing signal with the mechanical properties of the test object. Many measurement systems explore the ultrasound delay time: temperature [7], load measurement [9], food product quality monitoring [3]; thickness [4], non-destructive imaging [6] or biomedical applications [8]; flow rate measurement [5]. Essential procedure carried out in such measurements is the estimation of signal delay or time-of-flight (ToF) [1, 2].

Usually, pulse signals are used because easy to generate [10]. Yet these signals are not able to deliver sufficient energy if measurement precision is needed. Toneburst signals are used when high signal energy is needed [4]. But if pulse signals have good temporal resolution, tonebursts do not have this property. Spread spectrum (SS) signals [11-14] offer both: high energy and high resolution. Most widely used signals are chirp, linear frequency modulation and coded sequences [12]. New class of SS signals was suggested recently [15]: trains of pulses of arbitrary pulse width and position (APWP). For single reflection the SS signals have clear advantage: both sharp main correlation lobe and high energy increase the estimation accuracy. Estimation of signal arrival time gets complicated when multiple reflections are in close proximity: in [26] it was shown that additional error occurs.

Investigation presented was aimed to compare the ToF estimation performance of conventional and SS signals in multiple reflections case.

2. MULTIPLE REFLECTIONS PROBLEM

Thin plate thickness measurement creates a challenge: front face and back wall reflections are close in time. If short pulse signals are used, close reflections can be resolved (Fig. 1).

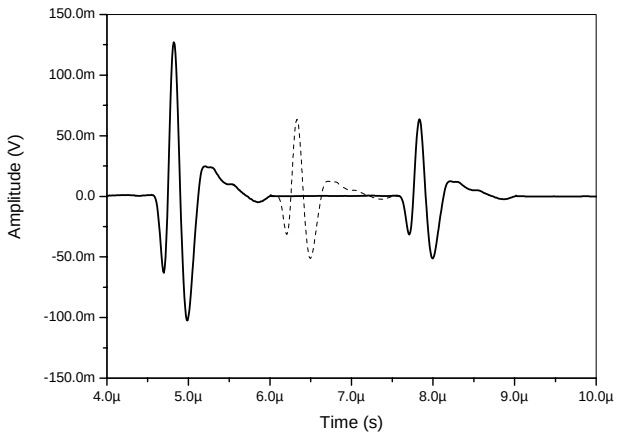


Fig. 1 – Multilayer reflections example when short pulse is used for probing.

But the shorter is the pulse the less energy it has and the lower is the attainable accuracy. In such case SS signals can be considered (Fig. 2).

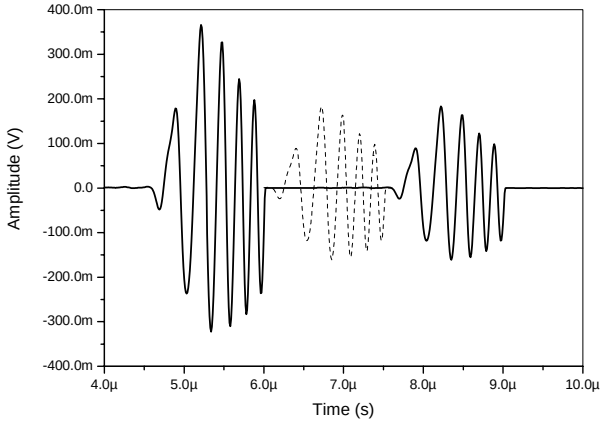


Fig. 2 – Multilayer reflections example when spread spectrum (chirp) signal is used for probing.

Application of SS signals is offering high energy yet it seems that resolution could be worse since operation in closer proximity (Fig. 2 dashed curves) could be complicated. Yet application of correlation processing compresses the signal and temporal resolution should not be jeopardized (Fig. 3).

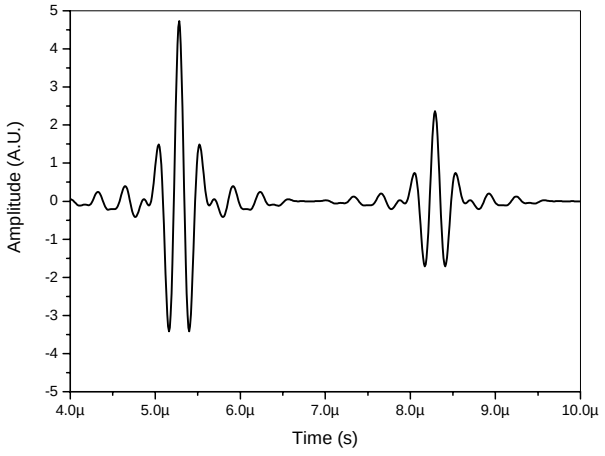


Fig. 3 – Signal compression result for chirp.

But in case of SS signal correlation sidelobes are higher (Fig. 3) than for pulse signal (Fig. 4).

Research presented in [26] was aimed to investigate the neighboring reflection influence on ToF estimation bias errors. Real signals were recorded using wideband transducer TF5C6N-E with delay line attached. Signals were averaged to produce noise-free reference signal ref_k. Here k=1…K represents the sample number. This reference signal later was used to construct a signal representing two reflections: front face, ff_k, with unity amplitude and backwall, bw_k with 0.7 amplitude. One signal was placed at original position which was not altered. Another signal was artificially shifted by introducing the Δt_n (n=1..N) delay via the phase alteration in frequency domain:

$$bw_{1..k}\left(t-\Delta t_n\right)=IFFT\left(FFT\left(ref_{1..k}\right)\cdot e^{j\omega_{1..k}\Delta t_n}\right) \quad (1)$$

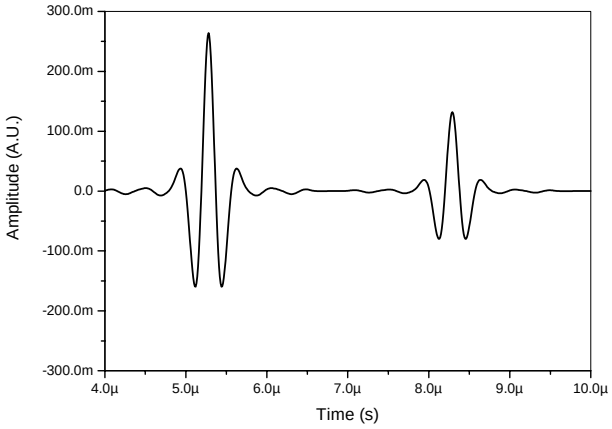


Fig. 4 – Signal compression result for pulse.

Then received signal is the sum of ff_k and bw_k:

$$sig_{1..k}=ff_{1..k}+bw_{1..k} \quad (2)$$

Several TOF estimation techniques exist [18, 27, 28]. If cross-correlation function x of reference and received signal is:

$$x(\tau)=\int_{-\infty}^{\infty} ref(t)\cdot bw(t-\tau)dt \quad (3)$$

Then the peak position of the cross-correlation function is the ToF estimate:

$$ToF_{DC}=\arg\left[\max\left(x(\tau)\right)\right] \quad (4)$$

But it is more practical to produce the cross-correlation function in digital space. Therefore signals analyzed are discrete.

$$x_m=\sum_{k=1}^{K} ref_{k-m}\cdot bw_k \quad (5)$$

If ToF errors below the sampling period are expected then interpolation is used to estimate the ToF between the samples. In ideal case, *sinc* function should be used for this purpose. But usually speed is required, so several truncated interpolation techniques exist [21-23]. Cosine interpolation was suggested as the best candidate in [26]. If CCF peak for a relatively narrowband signal was assumed to be harmonic function, then cosine interpolation can use samples x_{m-1}, x_m and x_{m+1} around the peak:

$$\Delta ToF_{\cos} = -\frac{\beta}{f_s \alpha}, \quad (6)$$

where

$$\alpha = \arccos\left(\frac{x_{m-1} + x_{m+1}}{2x_m}\right) \beta = \arctan\left(\frac{x_{m-1} - x_{m+1}}{2x_m \sin\alpha}\right) \quad (7)$$

Signals synthesized by (1) and (2) were used to obtain two ToF values, ToF_{ff} and ToF_{bw}: for ff_{k} and bw_{k} correspondingly. Peaks were located by gating the area of first and second reflection. Estimates were obtained using (4), (5), (6) and (7). Since there was N values for spacing Δt_{n} between the signal, every delay value was used to obtain the bias error $\varepsilon(ToF_e)$, by subtracting the introduced delay Δt_{n} (0 ns to 4000 ns for second reflection and 0 ns for first reflection):

$$\varepsilon(ToF_n) = ToF_n - \Delta t_n \quad (8)$$

ToF bias error is clearly seen in Fig. 5.

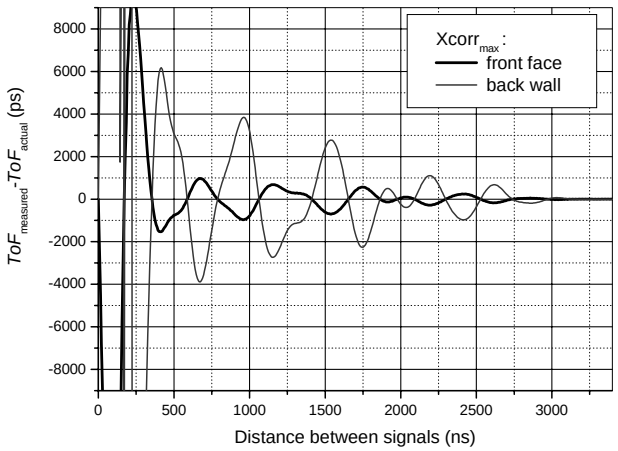


Fig. 5 – ToF bias error induced by neighboring signal.

It can be seen that presence of the neighboring reflection creates bias error. The influence is of opposite sign and depends on opposing signal amplitude: first reflection is larger so has less influence from second reflection and second reflection is smaller so has stronger influence from first reflection. The closer are the signals, the larger is the error. Situation can be expected from Fourier analysis point: pulse signal can be disassembled into several frequency components which exist beyond signal peak (Fig. 6).

But which type of signals will have lower errors? It was hypothesized that long SS signals which correlation lobes are larger should have worse bias error due to neighboring reflection.

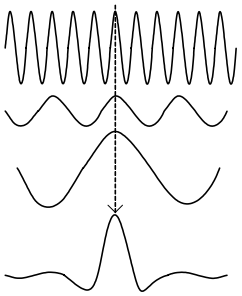


Fig. 6 – Pulse signal components exist beyond the peak.

3. EXPERIMENTAL RESULTS

In this investigation we aimed to compare the neighbor reflection influence on bias error for short single pulse and SS signals. Same transducer TF5C6N-E was used for real signals collection. Signals were collected from attached delay line, averaged to produce reference signal *ref*. Chirp signal with frequency 1-10 MHz and 2 μs duration was used as SS signal representative (Fig. 2). Conventional signal was represented by 100 ns rectangular pulse (Fig. 1).

Reference signals were artificially shifted as per equation (1). ToF values obtained using (4), (5), (6) and (7). Bias error due to neighbor reflection was obtained using (8).

At large spacing (Fig. 7, more than 3 μs) all signals had same performance bias error due to neighboring reflection is below few ps (ToF interpolation error).

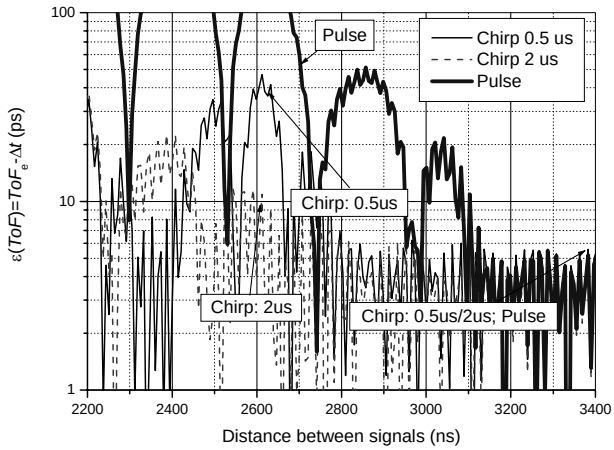


Fig. 7 – Bias error due to neighboring reflection vs. signal type at large spacing.

Performance degrades significantly if distance between reflections is reduced. Error for pulse signal is larger even in case of small spacing (Fig. 8). Such behavior of pulse signal is unexpected: being more concentrated in time it should be less sensitive to neighboring reflections.

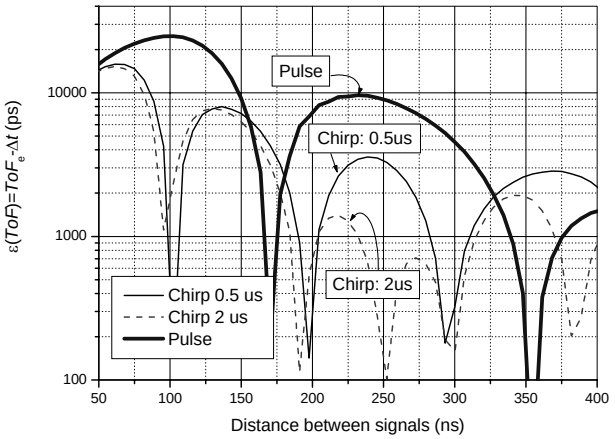


Fig. 8 – Bias error due to neighboring reflection vs. signal type at small spacing.

Obtained ToF error was normalised by the period of the transducer center frequency, f_0:

$$\delta(ToF)_{est} = \frac{\varepsilon(ToF)}{T_0} \cdot 100\% \quad (9)$$

Results for relative ToF error obtained for 100 ns rectangular pulse (matched for for 5 MHz transducer center frequency) are presented in Fig. 9.

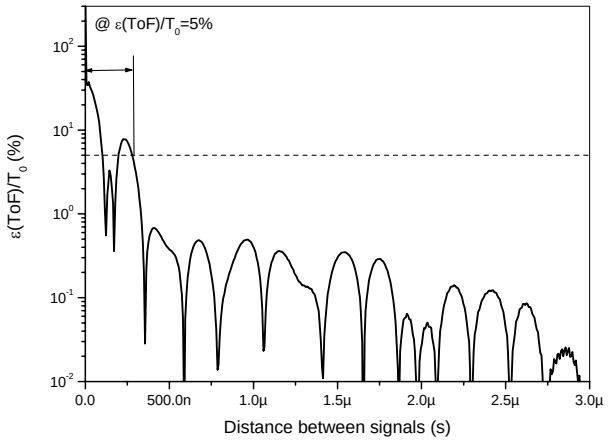


Fig. 9 – Relative time of flight estimation error for pulse signal.

Comparison of Fig. 9 results with errors for spread spectrum (2 μs chirp, Fig. 10) signal indicate the advantage of the last: errors are in the order of magnitude lower for spread spectrum excitation.

For instance, relative ToF estimation error is 5 % at 200 ns for spread spectrum signals and at 280 ns for pulse signal. For 0.1 % relative error signal spacing should be 1.25 μs for spread spectrum signals and 2.5 μs for pulse signal. Relative error is virtually zero if spacing is more than 1.3 μs for spread spectrum.

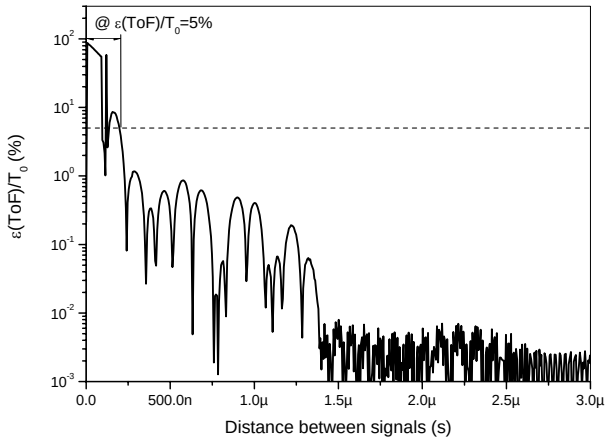


Fig. 10 – Relative time of flight estimation error for spread spectrum signal.

Amplitude reconstruction performance was studied. Once the position *ToF* is found then estimation A_{est} of the reflection amplitude in received signal s_k can be expressed via the reference signal ref_k:

$$A_{est} = \frac{\sum_{k=1}^{K} ref_k \cdot s_k}{K\sum_{k=1}^{K}(ref_k)^2 \cdot \sum_{k=1}^{K}(s_k)^2} \quad (10)$$

Then, this estimated amplitude can be compared to actual amplitude A_{act} of used in experiment to obtain the amplitude estimation error:

$$\varepsilon_{est} = \frac{A_{est} - A_{act}}{A_{act}} \cdot 100\% \quad (11)$$

Results for larger (first reflection in Fig. 1) pulse amplitude reconstruction are presented in Fig. 11.

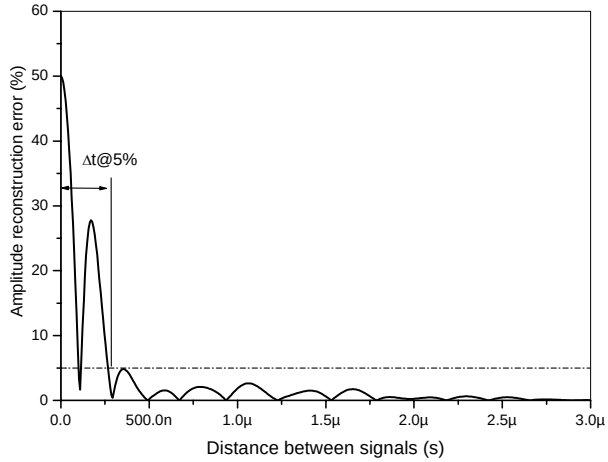


Fig. 11 – Amplitude estimation error due to neighboring reflection for pulse signal.

It can be seen that estimation error is reduced with distance and at 270 ns spacing is below 5 %. Amplitude estimation performance is slightly worse for spread spectrum (chirp) signal (Fig. 12): amplitude estimation error for chirp signal drops below 5 % at 300 ns distance.

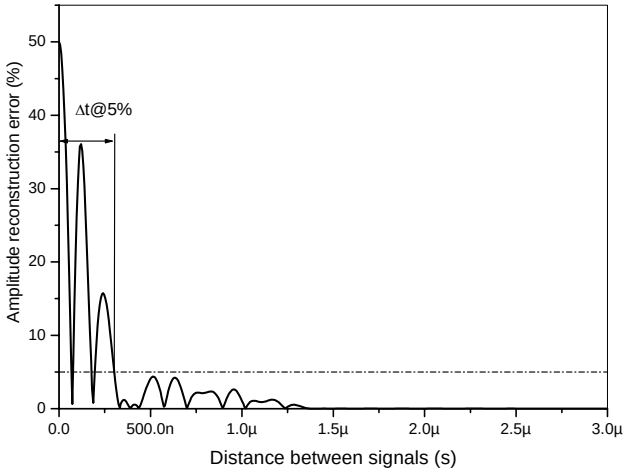


Fig. 12 – Amplitude estimation error due to neighboring reflection for chirp signal.

But it can be seen that amplitude estimation error for chirp signal is much better at large spacing: it is virtually zero at spacing above 1.3 μs. While for pulse signal such reduction is achieved only at more than twice the distance (3 μs).

5. CONCLUSIONS

It was demonstrated that presence of the neighboring reflection introduces time of flight estimation error. For separation larger than the duration of the reference signal, only interpolation errors prevail. In close proximity, additional time of flight estimation bias error is introduced which is of opposite sign for counteracting signals and depends on the amplitude of opposing signal. Situation can be explained from the signal theory: the neighboring signal produces the energy leak into its opponent. Narrow signals like single rectangular pulse, should suffer less from the aforementioned phenomena. But use of spread spectrum signals is preferred thanks to their energy and compressibility property. It was hypothesized that spread spectrum signals, being long signals will have worse bias error due to neighbor reflection. Comparison time of flight estimation error bias caused by neighbor reflection of pulse and chirp signal revealed that spread spectrum signals have better performance compared to short pulse signals.

6. ACKNOWLEDGMENT

This research was funded by a grant (No. MIP-058/2012) from the Research Council of Lithuania.

7. REFERENCES

[1] M. C. Amann, T. Bosch, M. Lescure, et al., Laser ranging: a critical review of usual techniques for distance measurement, *Opt. Eng*., (40) (2001), pp. 10-19.

[2] R. Kažys, L. Svilainis, L. Mažeika, Application of orthogonal ultrasonic signals and binaural processing for imaging of the environment, *Ultrasonics*, (38) (2000), pp 171-175.

[3] P. Pallav, D. A. Hutchins, T. H. Gan, Air-coupled ultrasonic evaluation of food materials, *Ultrasonics*, (49) (2009), pp. 244-253.

[4] N. B. Luciano, S. C. J. Alberto, et al., Development of an ultrasonic thickness measurement equipment prototype, *2010 Conference on Electronics, Communications and Computer*, Cholula Puebla, Mexico (February 22-24, 2010), pp. 124-129.

[5] V. Magori, Ultrasonic measuring arrangement for differential flow measurement, particularly for measurement of fuel consumptions in motor vehicles with a fuel return line, US patent No. 4,409,847, 1983.

[6] R. Kazys, R. Raisutis, E. Zukauskas, et al., Air-coupled ultrasonic testing of CFRP rods by means of guided waves, *2010 International Congress on Ultrasonics,* Physics Procedia, (3), 2010, pp. 185-192.

[7] W. Y. Tsai, et al., New implementation of high-precision and instant-response air thermometer by ultrasonic sensors, *Sensors and Actuators,* (117) (2005), pp. 88-94.

[8] P. G. M. Jong, T. Arts, A. P. G. Hoeks et al., Experimental evaluation of the correlation interpolation technique to measure regional tissue velocity, *Ultrasonic Imaging*, (13) (1991), pp. 145-161.

[9] S. Chaki, G. Corneloup, I. Lillamand, et al., Combination of longitudinal and transverse ultrasonic waves for in situ control of the tightening of bolts, *Journal of Pressure Vessel Technology*, (129) (2007), pp. 383-390.

[10] G. Athanasopoulos, C. Stephen, J. Hatfield, A high voltage pulser ASIC for driving high frequency ultrasonic arrays, *2004 IEEE Ultrasonics Symposium*, Montreal, Canada (August 23-27, 2004), pp. 1398-1400.

[11] F. Honarvar, H. Sheikhzadeh, M. Moles, A. N. Sinclair, Improving the time-resolution and signal-to-noise ratio of ultrasonic NDE signals, *Ultrasonics*, (41) (2004), pp. 755-763.

[12] M. Pollakowski, H. Ermert, Chirp signal matching and signal power optimization in pulse-echo mode ultrasonic nondestructive testing, *IEEE Transactions on Ultrason., Ferroelectr., Freq. Control,* (41) (1994), pp. 655-659.
[13] M. M. Saad, C. J. Bleakley, T. Ballal, S. Dobson, High-accuracy reference-free ultrasonic location estimation, *IEEE Transactions on Instrum. Meas.*, (61) (2012), pp. 1561-1570.
[14] Q. Licun, R. L. Wang, Performance improvement of ultrasonic Doppler flowmeter using spread spectrum technique, *2006 IEEE International Conference on Information Acquisition*, (1) (2), Shandong, China (August 20-23, 2006), pp. 122-126.
[15] L. Svilainis, K. Lukoseviciute, V. Dumbrava, et al., Application of arbitrary position and width pulse trains signals in ultrasonic imaging: correlation performance study, *Electronics and Electrical Engineering*, (19) (2013), pp. 57-60.
[16] L. Draudvilienė, L. Mažeika, Analysis of the zero-crossing technique in relation to measurements of phase velocities of the S0 mode of the Lamb waves, *Ultragarsas*, (65) (2010), pp. 11-14.
[17] V. G. Ivchenko, A. N. Kalashnikov, et al., High-speed digitizing of repetitive waveforms using accurate interleaved sampling, *IEEE Transactions on Instrum. Meas.*, (4) (2007), pp. 1322-1328.
[18] G. Jacovitti, G. Scarano, Discrete time techniques for time delay estimation, *IEEE Transactions on Signal Processing*, (41) (1993), pp. 525-533.
[19] H. Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.
[20] C. Rao, Information and the accuracy attainable in the estimation of statistical parameters, *Bulletin of Calcutta Mathematics Society*, (37) (1945), pp. 81-89.
[21] L. Svilainis, V. Dumbrava, S. Kitov, A. Chaziachmetovas, The influence of digital domain on time of flight estimation performance, *2012 International Congress on Ultrasonics*, Gdansk, Poland (September 5-8, 2011), pp. 479-482.
[22] X. Lai and H. Torp, Interpolation Methods for time-delay estimation using cross-correlation method for blood velocity measurement, *IEEE Transactions on Ultrason., Ferroelectr., Freq. Control*, (46) (1999), pp. 277-290.
[23] R. Kazys, Delay time estimation using the Hilbert transform, *Matavimai*,(3) (1996), pp. 42-46.
[24] L. Svilainis, A. Aleksandrovas, Application of arbitrary pulse width and position trains for the correlation sidelobes reduction for narrowband transducers, *Ultrasonics*, (53) (2013), pp. 1344-1348.
[25] R. Kažys, L. Svilainis, Ultrasonic detection and characterization of delaminations in thin composite plates using signal processing techniques, *Ultrasonics*, (35) (1997), pp. 367-383.
[26] L. Svilainis, A. Aleksandrovas, K. Lukoseviciute, V. Eidukynas, Investigation of the time of flight estimation errors induced by neighboring signals, *Proceedings of the 2013 IEEE 7th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, Berlin, Germany, 12-14 September 2013, pp. 413-418.
[27] L. Svilainis, Review of high resolution time of flight estimation techniques for ultrasonic signals, *2013 International Conference NDT*, Telford, UK, (Sep. 8-12, 2013), pp.1-12.
[28] L. Svilainis, K. Lukoseviciute, V. Dumbrava, A. Chaziachmetovas, Subsample interpolation bias error in time of flight estimation by direct correlation in digital domain, *Measurement*, (46) 10 (2013), pp. 3950–3958.

***Linas Svilainis**, acquired PhD degree in 1996 from the Kaunas University of Technology (Lithuania). Since 2009 he is Full time Professor at the Department of Signal Processing, Kaunas University of Technology, Lithuania. His current areas of research are: ultrasound electronics and signal processing; LED video displays research and development; Electromagnetic compatibility and EMI protection of electronic systems.*

***Arturas Aleksandrovas**, acquired MSc degree in 2013 from the Kaunas University of Technology (Lithuania). Since 2013 he is a doctoral student at the Department of Signal Processing, Kaunas University of Technology, Lithuania. His current areas of research are spread spectrum ultrasonics and signal electronics.*

***Kristina Lukoseviciute**, received the Ph.D. degree in informatics from Kaunas University of Technology, Lithuania, in 2012. Since 2012 she is a lecturer at the Department of Mathematical Research in Systems, Kaunas University of Technology.*

Her current research interests are evolutionary computations; time series analysis; time of flight estimation.

***Valdas Eidukynas**, acquired PhD degree in 1996 from the Kaunas University of Technology (Lithuania). Since 1997 he is Associate Professor at the Department of Engineering Mechanics, Kaunas University of Technology, Lithuania.*

His current areas of research are: biomechanics; structural safety; mechanical processes simulation.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ENHANCING THE COVERAGE OF INDOOR RADIO LOCALIZATION BY DISTRIBUTED COMPUTATIONS

Andreas Fink, Helmut Beikirch

Department of Computer Science and Electrical Engineering, Rostock University
A.-Einstein-Str. 2, D-18059 Rostock, Germany
{andreas.fink | helmut.beikirch}@uni-rostock.de
www.igs.uni-rostock.de

Abstract: The prevalent evaluation criterion for indoor local positioning systems (ILPS) is the achievable accuracy in terms of Euclidean distance between estimated and true position. Systems relying on received signal strength (RSS) ranging often use a distributed collection of RSS sensor data at reference nodes and a centralized position estimation. For this direct remote positioning, the accuracy is dependent on the reference node density and thus, is indirect proportional to the achievable coverage. To split up the dependency between these two criteria, we propose a distributed weighted centroid localization (dWCL) strategy with a hierarchical sensor data field bus. Accuracy and coverage of centralized and distributed WCL algorithms are compared for a one-dimensional tracking simulation and 196 reference nodes, arranged in up to 28 gateway segments. Using distributed computations, the localization system's coverage is increased by factor ten while the location estimation error increases only slightly.

Keywords: Centroid Localization, Distributed Computing, Human Tracking, RSS Ranging.

1. INTRODUCTION

The establishment of an active protection system for the monitoring of personnel in underground coal mines is crucial to save human lives and therefore, is an essential part of the development strategy of leading mining companies. The objective for an indoor local positioning system (ILPS) is to cover the whole area with the same quality of service, determined by the localization accuracy.

In underground coal mines, three different mining methods are used which define the structure of the mining area [1]:

- *Cut-and-fill mining* is used for irregular coal zones with steep dips. The sectional mining area consists of several types of tunnels.
- *Room and pillar mining* is used for relatively flat-lying coal deposits, e.g. following a particular stratum. The mining area looks like a large room with randomly shaped pillars.
- *Longwall mining* is used if the coal seam has a large and thin shape. This specialized type of mining process is characterized by minimal manual handling and highly productive automated control systems.

For longwall mining, the complexity of the self-advancing mining equipment, e.g. the electro-hydraulic shields for ceiling, is a mortal danger for maintenance staff. A direct remote positioning [2] for the longwall application using the received signal strength (RSS) of radio signals is shown in Fig. 1.

A set of fixed reference nodes receive the RF packets of mobile blind nodes and transmit the corresponding RSS values to a central data concentrator [3]. For the data backbone, the well-known and failsafe CAN bus is used which offers line-topologies up to the kilometers range and data rates up to 1 Mbit/s [4]. At the data concentrator, the user positions are calculated and used to intervene in the control system of the mining process and stop regarding shields.

The number of shields which have to be blocked depends on the accuracy and availability of the localization system. RF diversity for an improved availability of radio localization has been analyzed and successfully implemented in [5]. The localization accuracy is dependent on the reference node density and thus, is indirect proportional to the achievable coverage. Since a certain accuracy is required for a productive mining, the maximum distance between the reference nodes is limited, e.g. to a few meters for sub-meter accuracy [3]. For the used CAN bus, the maximum number of nodes on a bus segment is limited to 32 due to bus load

constraints of typical transceiver circuits [4]. Assuming a reference node distance of 3 m, this means an overall system coverage of only 96 m which is not sufficient for the 300 m longwall application, shown in Fig. 1.

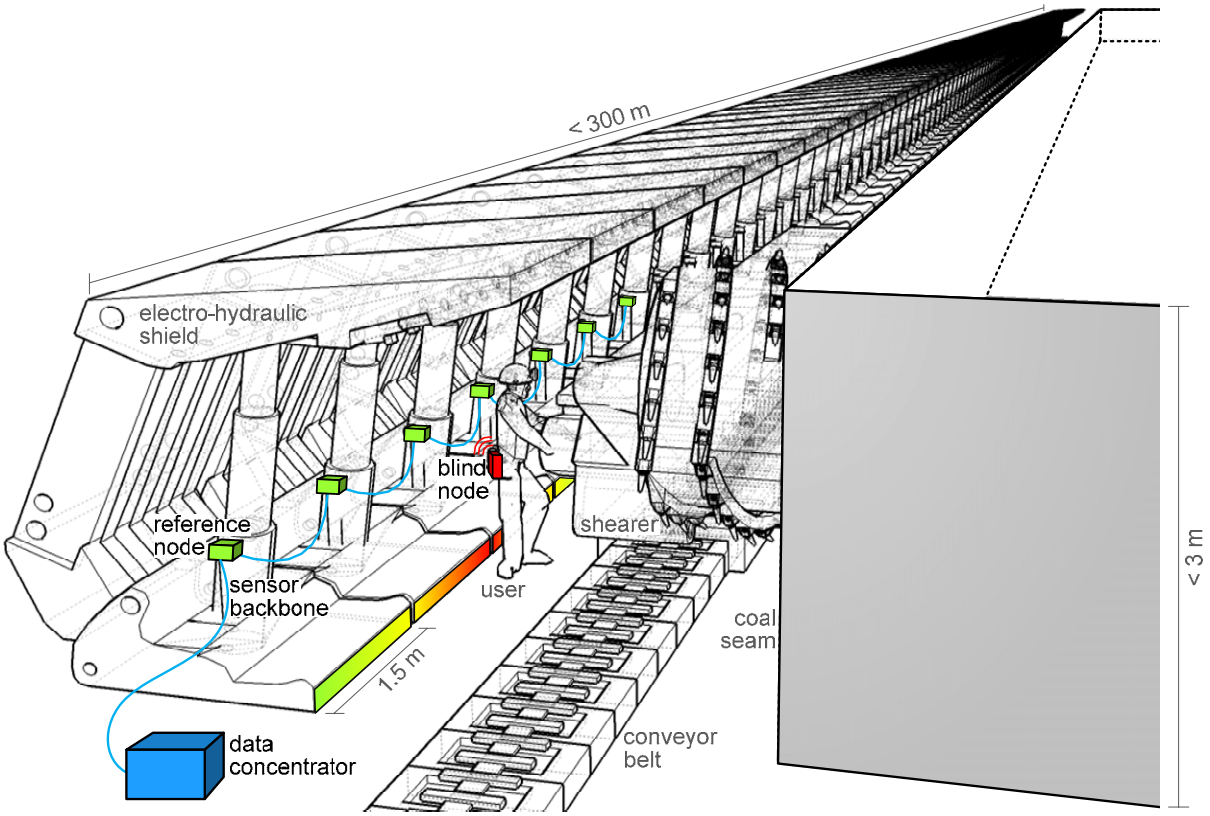


Fig. 1 – Indoor radio localization for personnel tracking in underground longwall coal mining using direct remote positioning (20 users along 300 m).

To break up the dependency between the two criteria accuracy and coverage, we propose a distributed weighted centroid localization (dWCL) strategy with a hierarchical sensor data field bus in [6]. In the current paper, we enhance the underlying RSS simulation model and investigate the accuracy limits of RSS-based localization.

The remainder of the paper is organized as follows. In section 2, the general classification of indoor localization and a taxonomy of RSS-based positioning are presented. The radio propagation model and consequential accuracy issues are analyzed in section 3, in particular for the propagation in tunnels. In section 4, the localization system's coverage is deduced from the CAN specifications, especially addressing practical CAN cable lengths and data rates for centralized and distributed location estimation strategies. The distributed weighted centroid localization (dWCL) approach is presented in section 5. In section 6, we compare the centralized and distributed localization performance using a one-dimensional tracking simulation. In the last section 7, the results are discussed and investigated in terms of an outlook for further system developments.

2. INDOOR RSS LOCALIZATION

Modern indoor local positioning systems for the tracking and navigation of people and objects show a variety of applied sensor technologies [7]. The main evaluation criteria are the accessible coverage and accuracy of the position estimations. An overview of available technologies with a taxonomy according to these two criteria is deduced from [8] and shown in Fig. 2. Of course, also other criteria like costs for installation and maintenance, the scalability, the user acceptance and specific system limitations should be taken into account. A further taxonomy of localization techniques and system implementations according to these criteria are proposed in [9] and [10].

In Fig. 2, the desired coverage and accuracy for the personnel tracking in the longwall mining is highlighted. Furthermore, the dashed graphs indicate systems with manual sampling like geodetic surveying systems. The dotted graphs indicate systems for typical outdoor use, while their use in indoor scenarios is limited due to intensive signal attenuations. The light solid graphs indicate systems which are limited to line-of-sight (LOS) scenarios.

All shaded graphs indicate systems which are more or less applicable for non-line-of-sight (NLOS) conditions.

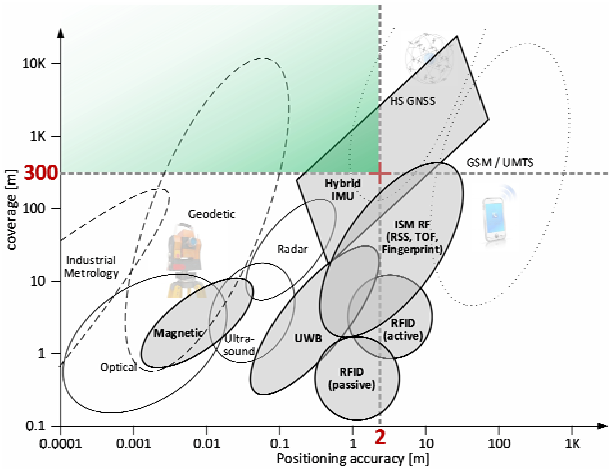


Fig. 2 – Overview of common localization sensor techniques with a classification according to coverage and accuracy, cf. [8].

For harsh indoor environments like a longwall coal mine, the received signal strength (RSS) of RF waves, e.g. using the worldwide harmonized 2.4 GHz ISM band, is a low-cost but valuable range-based localization sensor [11]. The performance of RSS-based localization is given by the spider chart in Fig. 3. As also stated in Fig. 2, accuracy and coverage are at medium level. The low-cost hardware and low algorithmic complexity together with low communication overhead enable an overall low cost solution.

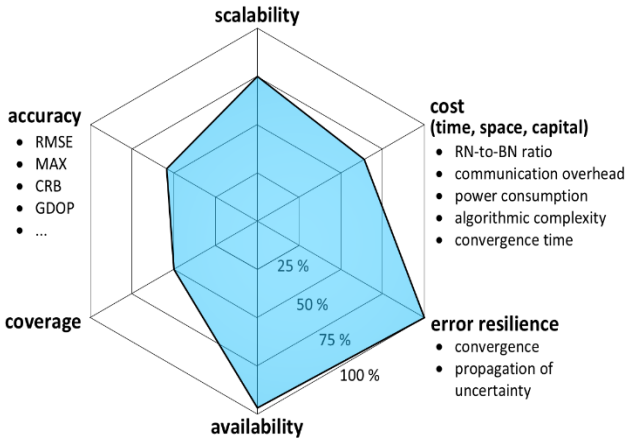


Fig. 3 – Taxonomy of RSS positioning.

3. RSS MODEL AND ACCURACY LIMITS

To find a relationship for the distance-dependent RSS and to compare the accuracy of centralized and distributed position estimation strategies in multipath fading environments, it is necessary to have a look at the indoor radio signal propagation. In this context, the limitations of RSS-based localization will be discussed using latest research results from literature and own path loss experiments.

3.1. LARGE-SCALE PATH LOSS

For both, indoor and outdoor radio channels, the distance-depending average path loss shows a logarithmic dropping of power with a linear increasing distance according to the LOS graph, shown in Fig. 4. The *log-distance* path loss model [12] is used to describe the average power at the receiver. With (1) the average path loss $PL(d)$ (in dB) over a LOS distance d between transmitter (TX) and receiver (RX) is given by the reference path loss $PL(d_0)$ over a reference distance d_0 and the environment-specific propagation coefficient n.

$$PL(d) = PL(d_0) + 10n\log\left(\frac{d}{d_0}\right) \qquad (1)$$

The value of $PL(d_0)$ is influenced by the effective radiated power (ERP) of the RF transmitter and the gain of the transmitting and receiving antenna [6]. The slope of the path loss is given by n and is influenced by the specific environmental propagation conditions and the used frequency as given by experiment results in Table 1.

Table 1.Path loss coefficients for typical indoor environments.

Frequency	Reference	PL coeff. *n*
900…1900 MHz	Office [12]	2.4…2.6
2000 MHz	Office [13]	2.5…4.0
2400 MHz	Office [14]	1.9…3.8
5500 MHz	Office [14]	1.7…5.0
868 MHz	Factory hall [5]	2.1…2.9
1300 MHz	Factory hall [15]	1.9…2.4
2440 MHz	Factory hall [5]	1.9…3.0
868 MHz	Corridor	2.4…2.8
2440 MHz	Corridor	1.8…2.3

In [15] and [12], values for n between 1.9 and 2.6 are given for frequencies between 900 MHz and 1900 MHz in obstructed indoor environments where not only NLOS conditions but also multipath signal fading affects the RF signal propagation. For 2000 MHz radio signals in an obstructed office environment, values between 2.5 and 4.0 are given in [13]. In the experimental investigations in [5], the path loss coefficient was determined with n=2.31 for 868 MHz and n=2.62 for 2400 MHz. All these values for the propagation coefficient were found by experiment and are typical values for spacious

indoor fading environments. They all show the general behavior of an increasing path loss and larger uncertainties for larger frequencies.

The mayor implication of the average path loss for the RSS-based ranging is an imbalanced resolution of the RSS over the distance. With usual receiver hardware, the RSS is sampled over a certain part of the received packet and stored as an 8 bit RSS indicator (RSSI) value which directly scales with the signal strength with 1 dBm resolution [16].

For RF communication systems with the purpose of data transmission like WLAN, the slope should be very flat to guarantee the same quality of operation over a large area. For RSS-based ranging and localization the opposite behavior is preferable [17].

Depending on the distance to the transmitter, two neighboring RSSI will translate into completely different distances. Near the transmitter, where the slope of the path loss graph is very steep, the 1 dBm resolution corresponds to a small distance in the centimeters range. Thus, the limited resolution will not lead to a downgrading of localization accuracy. Looking at the 2440 MHz path loss in Fig. 4, the steep slope from 1 m to 5 m is well suitable for RSS ranging. Far away from the transmitter, the slope is very flat and 1 dBm deviation corresponds to several meters RX-TX separation. The 2440 MHz path loss in Fig. 4 shows this flat behavior without larger signal dropouts for distances between 16 m and 24 m. Here, the resolution of the range measurements will lead to a coarse grained location estimation.

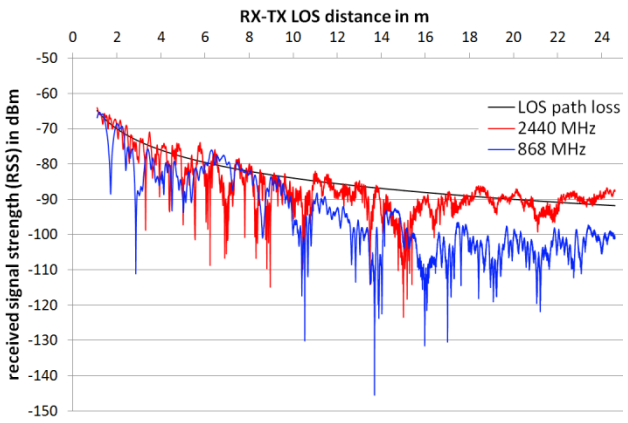


Fig. 4 – Path loss for 868 MHz and 2440 MHz in a narrow corridor.

3.2. SMALL-SCALE FADING

The distance dependent path loss from (1) is an average value and therefore not suitable to entirely describe a real channel. In general, a Rayleigh distribution of the RSS is assumed for NLOS conditions. When there is also a dominating LOS component, the fading is modeled with a Rician distribution. The probability density function of the Rician distribution is given by

$$f(x,v,\theta)=\frac{x}{\theta^2}\exp\left(\frac{-(x^2+v^2)}{2\theta^2}\right)I_0\left(\frac{xv}{\theta^2}\right), \qquad (2)$$

where θ is the amplitude of the superimposed RSS from all multipath components and has a scaling functionality. The parameter v is the peak amplitude of the LOS component. For $v=0$, the Rician distribution reduces to a Rayleigh distribution. Thus, the Rician distribution is a generalization of the Rayleigh distribution. For $v\to\infty$, the distribution reduces to a Gaussian distribution. The ratio of the LOS component compared to the superimposed signal can be expressed by the Rician K-factor according to

$$K(dB)=10\log\frac{v^2}{2\theta^2}dB, \qquad (3)$$

The environment-specific values of v and θ can be evaluated with an empirical analysis of multiple path loss observations using a two moment-based estimator as described in [18].

Several implications for RSS-based ranging and localization arise from the small-scale signal behavior. It is obvious, that the signal distribution in random environments also is random and the relation between distance and RSS could not be expressed by an equation [19]. Even with a training of the path loss coefficient n, RSS localization is known to be error-prone in multipath environments. The expected performance bound for a wide variety of algorithms shows a median location estimation error of three to four meters [20]. For changing environments, an automatically adjustment of the path loss coefficient should be used [21].

For RSS, the computational complexity of the location estimation is not proportional to the gained accuracy. Especially fine grained algorithms like least squares or maximum likelihood approaches show poorer accuracies than approximative ones [22]. Weighted centroid localization, for example, only needs a few computations and is especially suitable for erroneous range-sensors [3].

3.3. PROPAGATION IN TUNNELS

For the application of personnel tracking in longwall mining the environment can be seen as a tunnel with a uniform circular cross-section. When the diameter of the tunnel is small, it can be regarded as an over-sized imperfect waveguide. The propagation will exhibit the guided wave

characteristics and the propagation loss can be even smaller than that in free space [23].

Some additional assumptions for the small-scale behavior have to be considered. As for other indoor environments with a dominating LOS component, the fast fluctuations are modeled with a Rician distribution [24]. In [1], experiments with 450 MHz and 900 MHz support the theoretical waveguide model. The fast fluctuations are only apparent in the field near the transmitting antenna. As the frequency increases, the fast fluctuating region is prolonged.

Since the waveguide effect is larger for smaller wavelengths, the general large scale behavior of an increasing path loss for larger frequencies does not hold in tunnel environments. The waveguide acts as a high pass filter, so signals with higher frequency attenuate slower. This effect is examined in [1] and also by our experiment results in a narrow corridor. In Fig. 4, we compare the resulting path loss for 2440 MHz and 868 MHz. For comparison, both graphs are matched to the same reference path loss $PL(1\,m) = -67\,dBm$. For 2440 MHz, the path loss shows a slighter dropping than for 868 MHz. For distances above 20 m, the path loss is even lower than the LOS path loss for free-space propagation.

To sum up the influence of indoor radio propagation on the accuracy of RSS-based ranging and localization, in particular for one-dimensional environments like mining tunnels, we propose the following guidelines:

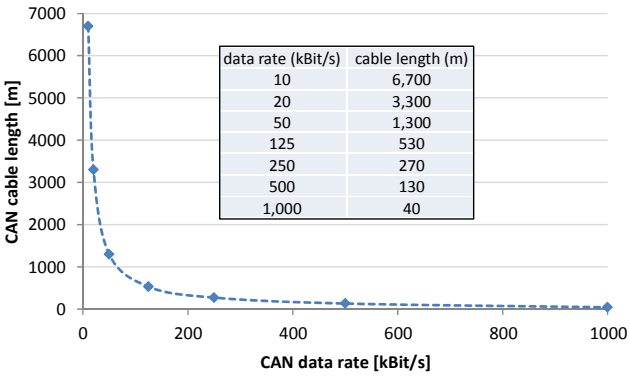
- For tunnels with waveguide characteristics, the larger frequency has a better communication range which is non-intuitive when looking at outdoor and open space indoor radio propagation.
- The channel bandwidth should be chosen as large as possible to suppress small-scale fading.
- For a proper RSS resolution, the distance between the mobile BN and the reference nodes shall be limited. This also limits the distance between the RNs and the overall coverage of the system.

4. SENSOR BACKBONE COVERAGE

To define the region which can be covered by the reference infrastructure, it is necessary to make some assumptions which are given by the ISO CAN specification [4] and the surrounding conditions of the application:

- The achievable location estimation accuracy is correlated to the distance between the reference nodes. Real life experiments have shown that useful system accuracies can be realized with node distances up to 3 m [3, 5].
- The CAN cable length is three times the node distance. When the distance between the reference nodes is 3 m, for a suitable mounting each CAN cable between the nodes has to be 9 m in length.
- A deterministic communication without large latencies is feasible with 50 % bus load. Thus, the practical CAN data rate is two times the CAN gross data rate.
- For a more generalized applicability of the investigations, the maximum number connected nodes on a single CAN bus is limited to 32. This criterion fits the demands of the electrical loads using typical low-speed and high-speed CAN transceivers.

With these predefined assumptions and the relationship between CAN cable length and CAN data rate (cf. Fig. 5) it is possible to define the maximum coverage of the one-dimensional localization system.



data rate (kBit/s)	cable length (m)
10	6,700
20	3,300
50	1,300
125	530
250	270
500	130
1,000	40

Fig. 5 – Maximum specified data rate and cable length for CAN bus lines [4].

The number of reference nodes (influencing the accuracy) and the number of simultaneously covered blind nodes affect the required data rate. The net data rate to collect the sensor information from n blind nodes using m reference nodes is given with

$$DR_{net} = m\left\lceil \frac{nx}{8} \right\rceil f_{pos}\, 64\, Bit\,, \tag{4}$$

where x is the number of bytes for each sensor and f_{pos} is the position update rate. For the CAN gross data rate, no generally valid value exists. The CAN header and inter frame gap have deterministic values. Since the number of stuff bits is dependent on the actual bit stream, the overall number of bits per CAN message is variable. A worst case scenario derived from the bit timing considerations in [25] leads to

$$DR_{gross} = m\left\lceil \frac{nx}{8} \right\rceil f_{pos}\left(\left\lceil \frac{34+64}{4} \right\rceil + 47 + 64\right) Bit \tag{5}$$

The practical data rate for a nearly deterministic communication rate depends on the specific application with allocation of more or less high prioritized message identifiers. The maximum tolerated bus load for the given application is 50 % Thus, using (6) the practical data rate is two times the gross data rate.

$$DR_{\text{det}} \geq \frac{DR_{gross}}{0.5} \qquad (6)$$

4.1. CENTRALIZED LOCATION ESTIMATION

The infrastructure components for the centralized location estimation are shown in Fig. 6. Up to 32 reference nodes (RNs) receive radio packets from up to 20 blind nodes (BNs) twice a second. The according RSS values are stored with one Byte and are transmitted to the data concentrator PC using a single CAN bus. In Fig. 6, an example distribution of the RSS values from all RNs is shown for a single BN located at RN 14.

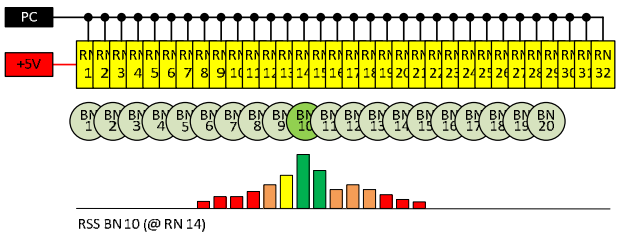


Fig. 6 – Infrastructure components for centralized location estimation of 20 blind nodes (BNs) using RSS values from 32 reference nodes (RNs).

A net data rate of $DR_{net} = 12.3\,kBit/s$ has to be realized by the CAN backbone using (4). The gross data rate with additional CAN header, bit stuffing and inter frame gap is $DR_{gross} = 26.2\,kBit/s$. As stated above, for an error-free communication without the suppression of less prioritized CAN packets, the data rate should be twice as large as the theoretical gross bus data rate. With $m = 32$ RNs and $n = 20$ mobile BNs, a practical CAN data rate of $DR_{\text{det}} = 52.4\,kBit/s$ leads to a specified CAN data rate of $DR_{CAN} = 125\,kBit/s$.

Thus, the theoretical maximum CAN cable length is 530 m with a corresponding maximum system coverage of approximately 177 m. Furthermore, the maximum number of 32 connected nodes on a single CAN bus together with the given node distance of 3 m limit the overall system coverage to 96 m.

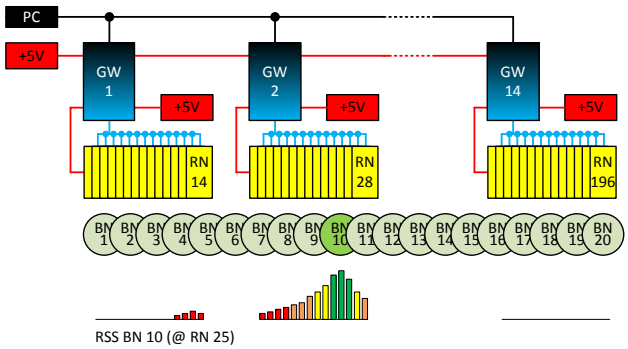


Fig. 7 – Hierarchical infrastructure for decentralized location estimation of 20 blind nodes (BNs) using 196 reference nodes (RNs) which are connected over 14 gateways (GWs).

4.2. DISTRIBUTED LOCATION ESTIMATION

To increase to coverage of the localization system, the distributed RSS localization uses additional infrastructure components to realize a hierarchical backbone bus. Up to 14 gateways with two CAN interfaces are used to connect up to 196 reference nodes to the data concentrator (cf. Fig. 7).

The embedded gateway architecture is shown in Fig. 8. To meet the requirements of ignition protection for the use in explosion-risk areas, the two CAN interfaces are electrically isolated. The CAN 1 transceiver has a separate 5 V power supply and is connected to the Cortex M4 system MCU via opto-couplers. Beside a large set of peripheral components, the MCU contains two dedicated CAN controllers.

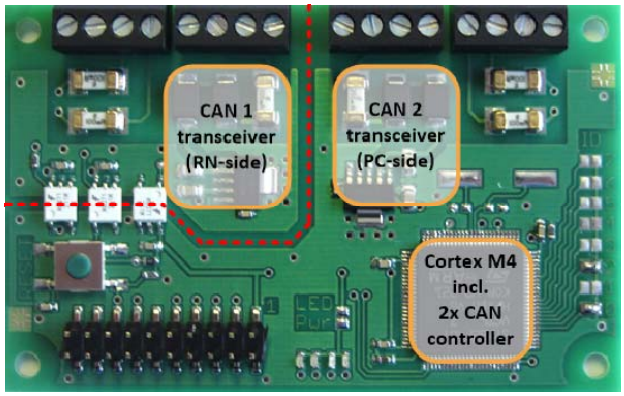


Fig. 8 – Embedded gateway architecture, top side of PCB showing CAN transceivers and Cortex M4 MCU.

For the centralized localization computation without data compression in the gateways, the PC-backbone CAN has to offer a minimum of $DR_{net} = 75.3\,kBit/s$. The theoretical gross bus data rate with additional CAN header and inter frame gaps has to be larger than $DR_{gross} = 159.4\,kBit/s$.

With 196 RNs, organized into 14 gateway subnets and 20 mobile BNs, this results in a practical CAN data rate of $DR_{\text{det}} = 318.8\,kBit/s$. For the given scenario we have to set $DR_{CAN} = 500\,kBit/s$, which is defined in the CAN specification. Looking to Fig. 5, this data rate limits the maximum cable length to 130 m. With this hierarchical bus concept, the maximum number of reference nodes can easily be scaled according to the application needs. Even though, the maximum cable length of 130 m for the top hierarchical bus-segment limits the system's coverage to less than 43 m. For the given application example of localization in the underground mining, at least 300 m coverage along a line of reference nodes is required.

To obtain a larger coverage, the backbone cable length has to be increased and thus, the backbone data rate has to be decreased by the same order of magnitude. The required data compression with a distributed location computation on the gateways is described in the following. For each BN, the gateways process the 14 byte RSS data (from all connected RNs) to compute a 1 byte subnet position information. Only the subnet position information are transmitted to the data concentrator PC. With this data compression by factor 14, the required data rate of the top hierarchical bus-segment is reduced significantly.

For the given example, the data rate of all 15 CAN segments (14x RN-side, 1x PC-backbone) can be set to $DR_{CAN} = 50\,kBit/s$. Thus, the specified maximum cable length for each segment is 1,300 m, resulting in an overall system coverage of 433 m. Compared to the centralized localization computation the coverage of the localization system is increased by factor ten. The detailed location computation steps on the gateways and the PC will be described in detail in the following section.

5. DISTRIBUTED WCL ALGORITHM

In the previous chapter, we introduced a segmentation of the sensor data backbone with data compression in the gateways. The data compression requires a distributed computation of a subnet BN positions and an additional weighting factor, denoting the influence of each subnet on the final centralized computation.

At all gateways, the subnet BN position is calculated according to the SAWCL algorithm given in [26]. First, the distances between all receiving RNs and the BN are calculated using

$$d_{ij} = 10\left(\frac{RSS_{ij} - A}{10n}\right), \tag{7}$$

where the reference signal strength A and the path loss coefficient n are taken from the corridor path loss experiment (cf. Fig. 4). The distances are used to compute a weight according to

$$w_{ij} = \frac{1}{(d_{ij})^g} \tag{8}$$

A weighting factor of $g = 3.0$ was found to be a good value for obstructed indoor environments [27].

The final calculation of the local position $d_i(x)$ of BN i is given by the weighted centroid of all RN positions $B_j(x)$ using

$$p_i(x) = \frac{\sum_{j=1}^{m}\left(w_{ij}B_j(x)\right)}{\sum_{j=1}^{m} w_{ij}}, \left\{p_i \in \mathbb{N} \mid 0 < p_i \le m\right\} \tag{9}$$

This one-dimensional position is rounded to an integer value, representing the ID of the nearest RN. Thus, using $m = 14$ RNs, the rough position fits in a 4 bit representation when it is transmitted to the data concentrator PC. Additionally, the gateways process the 1 byte RSS values from all m connected RNs to compute a specific weighting factor v_i with 4 bit representation. The weighting factor v_i is calculated using the average RSS of all m RNs according to

$$v_i = \frac{s_{\min} - \frac{\sum_{j=1}^{m} RSS_{ij}}{m}}{s_{\min} - A} \cdot 15, \left\{v_i \in \mathbb{N} \mid 0 < v_i \le 15\right\}, \tag{10}$$

where A is the reference RSS in a distance of 1 m and $S_{\min}$ is the receiver's sensitivity level. The value of v_i is normalized to a range of 0..15 to fit into the 4 bit representation.

The gateways forward the compressed data $p_i(x)$ and v_i to the data concentrator PC over the backbone CAN. On the PC the information from all gateways are processed to calculate the final position $P_i(x)$ of the mobile BN with ID i using

$$P_i(x) = \frac{\sum_{j=1}^{q}(v_i' p_i(x))}{\sum_{j=1}^{q} v_i'} \tag{11}$$

The impact of each segment position $p_i(x)$ can be tuned by modifying the weight v_i according to

$$v_i^{'} = v_i^{k}\,, \tag{12}$$

where k is an additonal weighting factor. Thus, it is possible to adapt the position estimation algorithm to different scenarios with varying infrastructure node setups. Beside the node distance, the ratio between the overall number of nodes and the nodes per subnet influence the optimum value of k.

6. TRACKING SIMULATION

For detailed investigations of different bus topologies, a simulation of the distance depending RSS is better suited than experiments, since the setup for hundreds of reference nodes would take a lot of time and space. Furthermore, real world experiments also limit the modification of process parameters (e.g. number of gateways), since the raw values are already processed on the gateways and only position information are available at the data concentrator PC. Thus, one set of measurement values cannot be analyzed with different gateway utilizations. For a parameterized computation using simulations of RSS distributions this is not an issue.

For the distance depending RSS we take the path loss model shown in Fig. 4. The average path loss is given with the log-distance model according to (1). For the varying part of the model we assume Rician fading using (2) and a Rician K-factor of $17.7\ dB$. As shown in the application's environment in Fig. 1, there exist a dominating LOS component of the superimposed RSS and only a few NLOS conditions might occur (e.g. shadowing due to human body). Thus, Rician fading is more suitable than Raleigh fading to model the multipath fading channel [18].

The infrastructure for the tracking simulation contains a line of 196 reference nodes 1 m beside the track and with 1 m node-to-node distance. A mobile blind node moves with a speed of $0.5\,m/s$ along the track and periodically sends out 2.4 GHz RF packets every 2 Hz. The reference RSS A in a distance of 1 m is set to $-67\ dBm$ and the receiver's sensitivity $S_{\min}$ is $-110\ dBm$.

In Fig. 9, the RSS values of all 196 reference nodes are plotted for a time duration of 392 s (784 iterations) and a blind node's linear motion from RN 1 to RN 196 (195 m overall track length). The color of the scatter plot indicates the RSS, where red corresponds to a high value near the reference path loss and dark blue to a relatively low RSS. Note that the z-axis starts at $S_{\min} = -110\ dBm$ and all values below this threshold are clipped off and neglected for the simulation model.

To figure out the influence of the distributed location estimation, two different utilizations are compared in the following. A centralized WCL computation (cWCL), corresponding to m=1 gateway and a distributed computation (dWCL) with m=14 gateways.

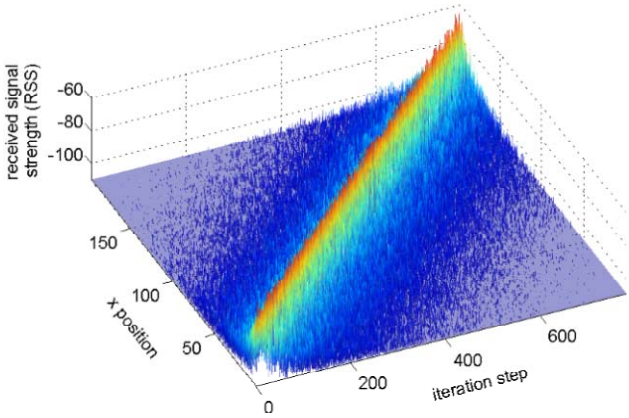


Fig. 9 – Received signal strength distribution for one-dimensional tracking simulation of a mobile blind node (BN) along 196 reference nodes (RNs).

The location estimation error (LEE) is given by the Euclidean distance between the estimated position and the true position, which is the absolute difference of x-positions in the one-dimensional case as given with

$$LEE(x) = \left\| \hat{x} - x \right\|_2 = \left| \hat{x} - x \right| \tag{13}$$

In Fig. 10, the accuracy of both configurations is compared for the partial track between RN 37 and RN 87, covering three complete subnets of 14 reference nodes in the distributed localization scenario.

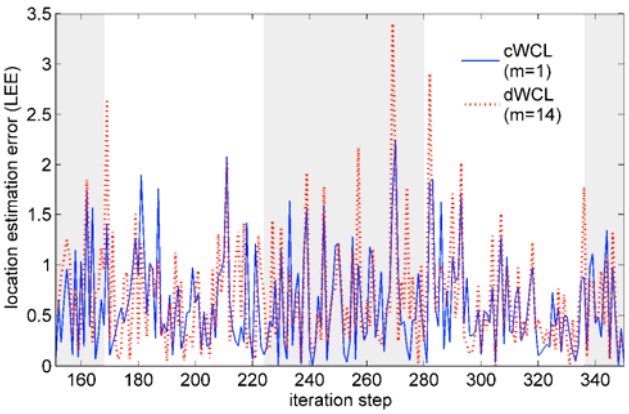


Fig. 10 – Location estimation error (LEE) for one-dimensional tracking simulation of a mobile blind node (BN). Comparison of centralized computation (cWCL) and distributed computation (dWCL) with m=14 gateways (extract showing 200 iteration steps for BN movement from RN 37 to RN 87, passing four gateway segment borders).

For most of the position estimations in the middle of a subnet there are only small differences between the two configurations. At the segment borders, the distributed computation has higher errors than the centralized computation.

A detailed comparison of both configurations is given with the error cumulative distribution functions in Fig. 11 and detailed error statistics in Table 2.

Table 2. Comparison of 3-sigma location estimation error (in m) for centralized and distributed weighted centroid localization.

error	cWCL m=1	dWCL m=1	dWCL m=4	**dWCL m=14**
median	0.48	0.50	0.50	**0.54**
sigma	0.22	0.26	0.39	**0.33**
3-sigma	2.78	3.00	3.70	**3.14**
maximum	3.56	4.00	5.54	**4.10**

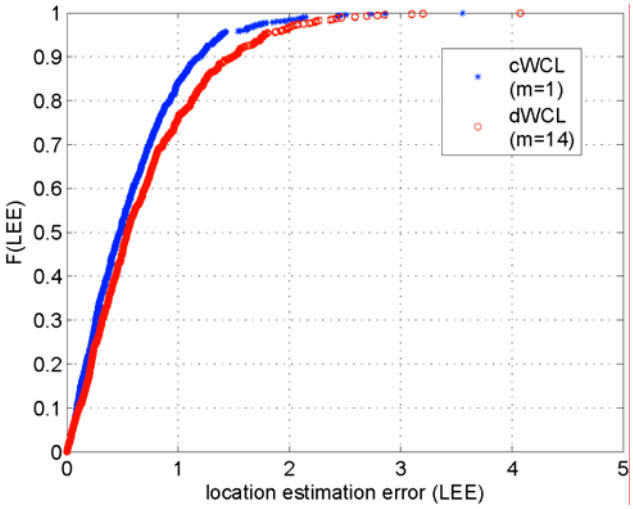


Fig. 11 –Cumulative location estimation error for 196 m linear track (empirical value), comparing centralized computation (cWCL) and distributed computation (dWCL) with m=14 gateways.

The influence of the number of gateways is analyzed in Fig. 12, where the 3-sigma error and the achievable coverage are shown for up to 28 gateways. The optimum accuracy for dWCL is reached with m=14 gateways. With the considerations in section 4, a CAN data rate of 50 *kBit/s* is sufficient to transmit all necessary sensor information. Looking at Fig. 5, a localization coverage along a line of 433 m can be realized with this setup. With a centralized WCL computation a specified CAN data rate of 500 *kBit/s* would be required which limits the system coverage to 43 m. This corresponds to a coverage enhancement by factor 10, when a distributed computation is used instead of a centralized one. At the same time, the maximum (3-sigma) error for the distributed computation with m=14 gateways compared to the conventional centralized one increases by 15.17 % (12.95 %).

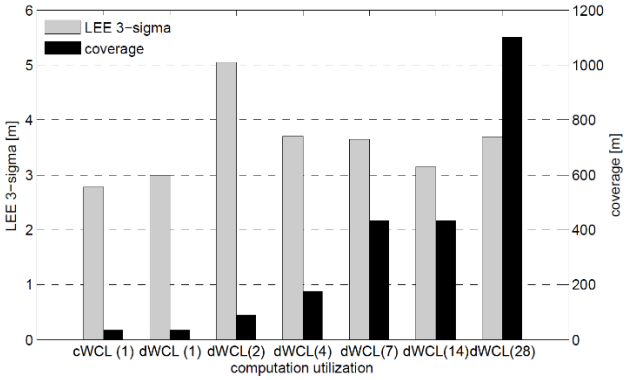


Fig. 12 –Comparison of the centralized (cWCL) and distributed (dWCL) weighted centroid location estimation, using 3-sigma location estimation error and one-dimensional localization coverage.

The influence of the weighting factor k on the location estimation error is analyzed in Fig. 13, where the cumulative error is shown for $1 \le k \le 50$. Additionally, the 3-sigma location estimation error is compared in Table 3 for selected values.

Table 3. 3-sigma location estimation error (in m) for different weighting factor k using distributed weighted centroid localization with m=14 gateways.

Error	k=2	k=4	**k=6**	k=8	k=10	k=12
3-sigma	7.25	3.16	**3.14**	3.31	3.69	3.95

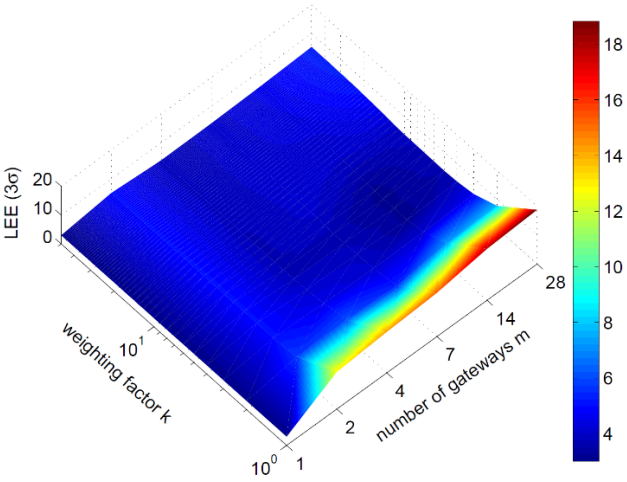


Fig. 13 – Combined influence of the number of gateways m and the weighting factor k on the 3-sigma location estimation error.

In the case a BN is located near the middle of a gateway segment, the position information from neighboring segments negatively influence the location estimation. This effect can be reduced by selecting an appropriate weighting factor k. For $k \le 2$ there exist large position errors, showing the

negative influence from neighboring gateway segments. The lowest error is reached with $k = 6$, where the influence from neighboring segments is shifted more to the segment borders. Larger values of k slightly decrease the estimation accuracy, since small systematic errors are introduced at the gateway segment borders.

7. CONCLUSION AND FUTURE WORK

The distributed WCL localization based on RSS readings is proposed to increase the system's coverage, limited by bus load constraints for collecting all sensor information. The results from a one-dimensional tracking simulation show, that it is possible to split up the location estimation into distributed parts of computation. Compared to the cWCL estimation, the increase in location estimation error for dWCL is relatively low and acceptable for most applications.

The main contribution of the decentralized location computation is a considerably increase of the tracking system's coverage up to the kilometers range. The system also can easily be scaled to the application's needs. An example scenario for personnel tracking in the underground longwall mining with 196 reference nodes shows a good localization performance for a utilization with 14 gateway segments. The corresponding system coverage of more than 400 m is sufficient for real world underground longwall mining setups.

Further system developments are focusing on enhancements of the dWCL algorithm, especially to encounter some systematic computation errors at the edge of the regarding track. The second issue for future work is a real life tracking measurement in the underground to compare the propagation effects with the corridor experiment and the corresponding assumptions for the simulated RSS distribution.

REFERENCES

[1] Z. Sun and I. F. Akyildiz, Channel modeling and analysis for wireless networks in underground mines and road tunnels, *IEEE Transactions on Communication*, (58) 6 (2010), pp. 1758–1768.

[2] J. Figueiras and S. Frattasi, *Mobile Positioning and Tracking - From Conventional to Cooperative Techniques*, 2010.

[3] A. Fink and H. Beikirch, Adding link quantity information to redundant RF signal strength estimates for improved indoor positioning, *3rd IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2012, pp. 1–6.

[4] ISO 11898-1, *Road Vehicles – Controller Area Network – Part 1: Data Link Layer and Physical Signaling*, ISO Std., 2003.

[5] A. Fink and H. Beikirch, Combining of redundant signal strength readings for an improved RF localization in multipath indoor environments, *15th International Conference on Information Fusion*, 2012, pp. 1–7.

[6] A. Fink and H. Beikirch, Radio-based human tracking for large indoor environments using distributed centroid location estimation, *7th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS)*, 2013, pp. 442-449.

[7] R. Mautz, *Indoor Positioning Technologies*, Habilitation Thesis, 2012.

[8] T. K. Kohoutek, R. Mautz, and A. Donaubauer, Real-time indoor positioning using range imaging sensors, *Proceedingsd of SPIE Photonics*, (7724) (2010).

[9] M. Allen, S. Baydere, E. Gaura, and G. Kucuk, Evaluation of localization algorithms, *Localization Algorithms and Strategies for Wireless Sensor Networks: Monitoring and Surveillance Techniques for Target Tracking*, 2009, pp. 348–379.

[10] P. Barsocchi, S. Chessa, and F. Furfari, Evaluating AAL solutions through competitive benchmarking: the localization competition. *IEEE Pervasive Computing*, (99) (2013).

[11] A. Fink, H. Beikirch, and M. Voss, Improved indoor localization with diversity and filtering based on received signal strength measurements, *International Journal of Computing*, (9) 1 (2010).

[12] T. S. Rappaport, *Wireless Communications – Principles and Practice*, Prentice Hall PTR, 2002.

[13] H. Benn and H. Thomas, GSM in the indoor business environment, in *GSM Evolution Towards 3rd Generation Systems*, Z. Zyonar, P. Jung, and K. Kammerlander, Eds. Kluwer Academic Publishers, 2002, pp. 211–234.

[14] D. Cheung and C. Prettie, *A path loss comparison between the 5 GHz UNII band (802.11a) and the 2.4 GHz ISM band (802.11b)*. Intel Labs, Technical Report, 2002.

[15] T. S. Rappaport and C. D. McGillem, UHF fading in factories, *IEEE Journal of Selected Areas Communications*, (7) 1 (1989), pp. 40–48.

[16] Texas Instruments, *CC2500: low-cost low-power 2.4 GHz RF transceiver*. rev. C, 2008.

[17] C. Beder, A. McGibney, and M. Klepal, Predicting the expected accuracy for fingerprinting based WiFi localisation systems, *2nd IEEE International Conference on Indoor*

Positioning and Indoor Navigation (IPIN), 2011, pp. 1–6.

[18] A. Abdi, C. Tepedelenlioglu, M. Kaveh, and G. Giannakis, On the estimation of the parameter for the rice fading distribution, *IEEE Communication Letteers*, (5) 3 (2001), pp. 92–94.

[19] A. T. Parameswaran, M. I. Husain, and S. Upadhyaya, Is RSSI a reliable parameter in sensor localization algorithms, an experimental study, *IEEE International Symposium on Reliable Distributed Systems*, 2009.

[20] E. Elnahrawy, X. Li, and R. P. Martin, The limits of localization using signal strength: A comparative study, *IEEE Sensor and Ad Hoc Communications and Networks (SECON)*, 2004, pp. 406-414.

[21] K. Whitehouse, C. Karlof, and D. Culler, A practical evaluation of radio signal strength for ranging-based localization, *SIGMOBILE Mobile Computing and Communications Review*, (11) (January 2007), pp. 41-52.

[22] G. Chandrasekaran, M. A. Ergin, J. Yang, S. Liu, Y. Chen, M. Gruteser, and R. P. Martin, Empirical evaluation of the limits on localization using signal strength, *IEEE Sensor and Ad Hoc Communications and Networks (SECON)*, 2009.

[23] Y. P. Zhang, G. X. Zheng, and J. H. Sheng, Radio propagation at 900 MHz in underground coal mines, *IEEE Transactions on Antennas Propagation*, (49) 5 (2001), pp. 757–762.

[24] M. Boutin, A. Benzakour, C. Despins, and S. Affes, Characterization and modeling of a wireless channel at 2.4 and 5.8 GHz in underground tunnels, *3rd International Symposium on Wireless Communication Systems (ISWCS)*, 2006, pp. 517-521.

[25] K. Tindell and A. Burns, Guaranteed message latencies for distributed safety-critical hard real-time control networks. Technical Report, 1994.

[26] A. Fink and H. Beikirch, RSSI-based indoor localization using antenna diversity and plausibility filter, *6th Workshop on Positioning Navigation and Communication*, 2009, pp. 159–165.

[27] A. Fink and H. Beikirch, Analysis of RSS-based location estimation techniques in fading environments, *2nd IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2011, pp. 1–6.

***Andreas Fink**, received his M.Sc. degree in Information Technology from Rostock University in 2008. Currently, he works as a research scientist at the Department of Computer Science and Electrical Engineering at Rostock University. His primary research interests focus on wireless communication sys-tems, radio-based indoor positioning and hybrid localization with inertial sensors.*

***Helmut Beikirch**, to hold the professorship „Electronic Devices and Circuit Design" at the Department of Computer Science and Electrical Engineering at Rostock University since 1998. He is the head of the research group in electronic circuit design and communication systems in industrial automation.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

BREAKING BLOCK AND PRODUCT CIPHERS

Albert H. Carlson [1], Robert E. Hiromoto [2], Richard B. Wells [3]

[1] Fontbonne University, St. Louis, MO 63105, USA, acarlson@fontbonne.edu,
[2] University of Idaho, Idaho Falls, ID 83402, USA, hiromoto@uidaho.edu,
[3] University of Idaho, Moscow, ID 83844-1010 USA, rwells@uidaho.edu

Abstract: The security of block and product ciphers is considered using a set theoretic estimation (STE) approach to decryption. Known-ciphertext attacks are studied using permutation (P) and substitution (S) keys. The blocks are formed from two (2) alphabetic characters (meta-characters) where the applications of P and S upon the ASCII byte representation of each of the two characters are allowed to cross byte boundaries. The application of STE forgoes the need to employ chosen-plaintext or known-plaintext attacks.

Keywords: Set theoretic estimation, block and product cipher, byte boundaries, known ciphertext attack, meta-characters.

1. INTRODUCTION

Cryptanalysis is the study of decryption techniques of encrypted information, which involves determining the secret key to the encryption. The key is a "secret" parameter that adds "noise" [1] to the original message and makes the message unreadable.

The process of decryption is a set of iterative applications of *a priori* knowledge applied to a noisy input in an effort to recover the message from the noise. The noise is intentionally injected into the message so that

$$C = E\,(M, k),$$

where C is the cipher text and $E\,(M, k)$ is the encryption function using a key, k, as a noise generator for the message M.

Noise generators take on numerous forms. The block and product ciphers apply multiple keys to the same data, one after the other. That is, the first cipher is applied to the plain text. The second cipher is applied to the ciphertext that results from applying the first cipher to the plaintext.

In this paper, we apply a rule-based set theoretic estimation (STE) [2] approach to capture the properties of *m*-grams that are derived from a *stylistic* use of the English language. We treat the resulting collection of *m*-grams as an *a priori* property set that is used to analyze the frequency distribution of *m*-grams symbols and as predictors for either *allowed* or *forbidden* letter combinations.

In the remainder of the paper, we provide background material on permutation, *P*, ciphers and their relation to the substitution, *S*, cipher. Results of STE applied to a block P and block S decryption algorithm is presented.

2. BACKGROUND

All modern block ciphers are product ciphers. The product cipher combines several encryption operations to provide both confusion and diffusion. Confusion substitutes one character for another while diffusion distributes information across the encrypted message [3]. Block ciphers of *PS* and *PSP* type are composed of a *P* cipher with a key representing a mapping. The key space for the *P* cipher is b!, where b is the number of bits in the block [4, 5, 6]. Similarly, the *S* cipher maps an alphabet A to another group of symbols, A′ in some manner, $A \mapsto A'$. For the *S* cipher, it is possible that A = A′. The key space is |A|!, where |A| is the number of symbols in the alphabet.

Another way to increase the key space is to increase the number of times a message is encrypted. Shannon concluded that compounding ciphers could increase the key space for a message, and as a result, increase its security. When the product ciphers use a good mixing transformation [5, 9], the number of keys increases by multiplying the key space for each

cipher in the product cipher. The key space for a product cipher with p ciphers is given by

$$\left|K_{productcipher}\right| = \prod_{i=1}^{p} K_i$$

Since Shannon introduced the concept of increasing security by compounding ciphers, it has generally been accepted that product ciphers of the form *PSP* [1, 3, 4] are more secure than a cipher employing only a permutation (See Fig. 1) or a substitution cipher. This is not true for block ciphers (See Fig. 2) whose encryption algorithms end at byte (character) boundaries and are encoded using ASCII.

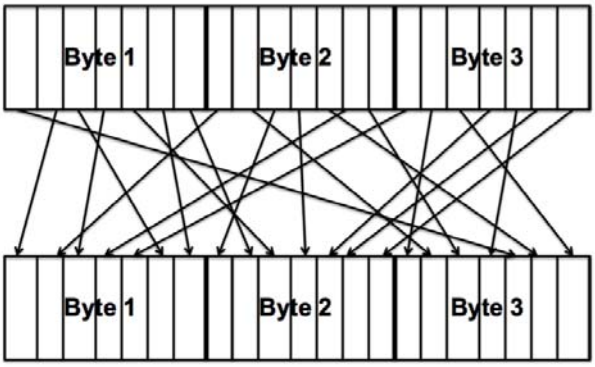


Fig. 1 – Permutation Cipher Mapping Key.

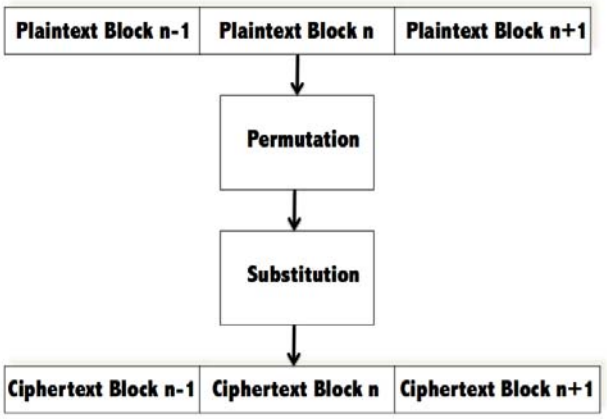


Fig. 2 – PS Type Cipher.

Blocks of integral character size suffer from a significant weakness; that is, information is confined within the block. As such, we suspect that as the block size of the *PSP* cipher increases above two symbols the additional security gained is insignificant as compared to a simple substitution cipher of the same block size. Combining both *P* and *S* ciphers into a block cipher using identical block boundaries results in a key space of size b!|A|!, where b is the number of bits in the block and |A| is the number of symbols in the alphabet A. Extending the block cipher to a *PSP* cipher, the size of the key space is b!|A|!b!. Let X be a compound symbol with an ordered n-tuple of characters $< x_0, x_1, \cdots, x_i >$ regarded as comprising a single (meta) symbol, where $< x_0, x_1, \cdots, x_i > x_i \in A_\lambda$ and A_λ is the alphabet of language λ. The language λ is a meta-language composed of meta-s-characters embedded in the same natural language, where *s* is the number of alphabetic symbols that make up a meta-character. A different meta-language exists for each meta-s-character size. For example, a meta-7-character 'flyinsk' is drawn from a meta-language with an alphabet using 'f', 'l', 'y', 'i', 'n', 's', and 'k'. In this representation, information is not restricted to the same byte of data in which it originated. For example, permutations on multi-byte blocks allow for any bit in the block to be permuted to any other location inside the block, even across byte boundaries. Ciphers that diffuse data are specifically chosen so that they allow diffusion across byte boundaries as depicted in Fig. 1. This is a much more difficult problem than byte-wise decryption. The problem is so difficult that cryptanalysts have chosen to create different attacks rather than deal with the diffusion [4, 12]. The algorithm described in this chapter deals directly with the diffusion across byte boundaries.

A *PSP* cipher is idempotent to an *S* cipher with identical block boundaries. This result follows by taking symbols in a block cipher as a compound symbol of the same size and having the same boundaries as the cipher block. Further, let the S block cipher be applied to the same compound symbol. Then *P* and *S* ciphers are equivalent within a symbol boundary. Therefore, *PSP* reduces to *SSS*. *S* ciphers are idempotent and associative with each other. Therefore *SSS* cipher reduces to a single *S* cipher. It can also be demonstrated that *P* reduces to *S*; however, *S* does not necessarily reduce to *P* [13, 14].

3. BCBB ALGORITHM

Defeating permutation across block boundaries requires treating the language as if the block of characters was actually a single character of a different language. The block comprises a character made up of characters, or a "meta-character," which is part of a "meta-language."

The Block Cross Byte Boundary (BCBB) algorithm is designed to decrypt block ciphers of *PS* and *PSP* type. The BCBB algorithm is based on the framework of STE as described earlier. The algorithm starts by reading in both the encrypted text and the data sets needed for decryption. The algorithm then sets up a solution matrix of possible mappings from ciphertext to plaintext meta-s-characters. The mapping is stored by means of a

hash-table that associates invalid key mappings for a particular ciphertext meta-s-character to a plaintext meta-s-character. Mappings that do not appear in the hash-table are still considered possible. Lists of meta-s-characters seen in the encrypted text and mappings found to be part of the key are also maintained.

When choosing property sets for use in decrypting multi-byte product ciphers, data sets are based on meta-s-characters of the block size being decrypted. In this case, only the meta-s-character frequency and allowed meta-s-gram (meta(s,m)) sets are used for decryption tests. For example, the text composed of 'theonl' is a meta(3,2) made up of two different meta-3-characters 'the' and 'onl'. Note that a meta(s,m) is equivalent to an m-gram of the size $s * m$.

The first property set applied to the BCBB algorithm for a product cipher is the global frequency property set. Global redundancy is applied only once. Global frequency is the frequency of characters in the meta-alphabet found in the message. Following the Law of Large Numbers [6], the larger the message, the more likely the meta-s-character frequency from the message will accurately reflect the meta-language alphabet frequency. The meta-2-character frequencies are studied by [13]. Both high and low frequency characters are of interest in this set. A large division in the data collected occurs between the first 10 and subsequent members of the meta-2-character set. This division is referred to as the "high frequency threshold." The top ten meta-2-characters on the frequency list are dubbed "high frequency" meta-2-characters. When interpreting data returned from the BCBB algorithm, higher frequency meta-2-character combinations are assumed to belong to the top ten-frequency set.

Similarly, the corpus identifies a set of low frequency characters. "Low frequency" meta-2-characters fall within the bottom 5% of the meta-2-character frequency list. Again, finding a frequency where it is possible to distinguish between sets of meta-2-characters sets the threshold. Any meta-2-character occurring more frequently than the threshold cannot be mapped to any member inside the low frequency set. Thus the mapping(s) can be eliminated. Together the initial application of the global frequency set results in a reduction of mappings in the solution matrix.

Other property sets selected for use are the frequency of meta-s-characters and the forbidden meta(s,m) set. The use of meta(s,m) sets subsumes the redundancy and multiple letter sets used for the algorithm, indicating that no additional property sets need to be applied.

Once the solution structure is set up, the algorithm began to eliminate mappings. The mapping elimination procedure follows the following steps: 1) the first property is applied to the global meta-s-character frequency data. The entire message is processed and then compared against a normalized global frequency list, 2) high frequency meta-s-characters passing the frequency threshold are mapped to a select set of plaintext characters that contained the only characters seen above the threshold, 3) the message is checked for redundant meta-s-characters in each meta(s, m) gram for all m selected for evaluation. Redundancy in a meta(s,m)-gram yields low entropy information. Therefore, processing the redundancies further eliminates mappings, 4) following frequency and redundancy checks, the main body of message analysis begins. A meta-s-character is read from the message and the meta(s,m)-gram set is applied to the portion of the message that is being analyzed. This process is iterated upon as new meta-s-characters are introduced from the message and reanalyzed until the message is decrypted or there are no more meta-s-characters left for evaluation. This entire process is called the *relaxation* stage. Additional details of the BCBB algorithm are found at [13].

4. EXAMPLE ATTACKS

Consider a message encrypted by a permutation cipher, followed by a substitution cipher (referred to as *PS*). Without loss of generality, let the plain text consist only of lower case alphabetic English characters with all spaces that delineate words within the message, and punctuations removed. The message uses standard ASCII encoding and the permutation employs a two-character (byte) block. Assume that the byte and block boundaries are known for the encrypted message. Blocks are restricted to coincide with byte boundaries. We treat the entire block as a single meta-character in a meta-language.

Each meta-symbol is analyzed for common language statistics based on their appearance in English. Language statistics constitute property sets that can be exploited using Set Theoretic Estimation (STE) and Set Methodology [2, 10]. The STE property sets used in the attack are listed in Table 1.

The ASCII representation used in this example requires the use of three special (constant) bits that indicate the case and alphabetic letters used. These bits are termed *static* bits, and are also permuted under the mapping of the key. The location of these bits can be determined by performing an XOR function between two blocks. Any bits that do not change are identified as *static* bits. A block diagram

illustrating the *static* bit algorithm is shown in Fig. 3.

The resulting information is used as one of the XOR operators and additional blocks are XOR'ed with the composite information. Since ASCII employs the same three most significant bits, nine bits are located (i.e., three bits per a two letter block), leaving 10 remaining bits to be mapped.

Table 1. STE Property Sets for Block Cipher.

Label	Property Sets
Φ_0	English Language
Φ_1	Static Bits
Φ_2	Byte Grouping
Φ_3	Bit Exclusivity
Φ_4	3-grams Forbidden

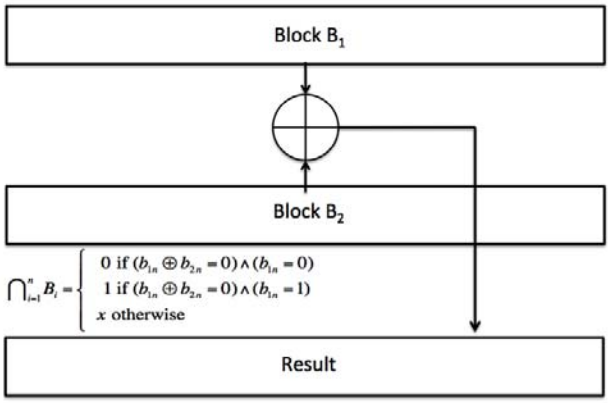


Fig. 3 – Static Bit Algorithm.

The remaining 10 bits are taken five at a time and permuted internally to identify all legitimate ASCII characters. The permutation ordering is defined relative to the corresponding block of the cipher text. Using the internal bit representation for ASCII letters, these letters are ordered in ascending order. If the group of five bits cannot be arranged to form a valid pattern for a letter, it is discarded. At the completion of this process, there should be at most two five-bit groupings, each containing one or several valid letters that arise through a unique permutation mapping. In addition, certain bit combinations are not allowed in the ASCII representation. Patterns of five "1"s or five "0"s, for example, are not valid and, therefore, disregarded. This is a property set that is referred to as *bit exclusivity*. This entire process is applied to *n* two-byte block (the order is not important) in the message.

At the end of processing *n* two-byte blocks, the resulting allowed permutation mappings in each block most likely contain the correct key and numerous spurious keys. At this stage, each bit within each byte of a block is assigned a unique index about which valid letter permutations are identified. In the next step, the *derived* information (allowed permutations) is analyzed by intersecting all possible permutations in each block across all block bytes. This step takes into account the fact that the same mapping key encrypts each block. With this in mind, the resulting intersections should determine the set of equivalent permutation mappings for the encryption key. At this point, the correct key that determines the correct ordering of letters is still unknown.

The final step of the process reduces to the application of the *Last-Man-Standing* technique described above. Each equivalent permutation contains a string of valid letters but only one sequence is the correct ordering sought. This problem now becomes identical to the substitution cipher problem that was solved by the application of *m*-grams. The nature of the two-byte block, therefore, suggests the application of 3-grams to determine correct ordering of letters.

Fig. 4 illustrates the intersections of property sets as applied in STE to decryption. The requirements for an STE application are: 1) the problem must mapped one-to-one and onto the solution domain; 2) the problem must have a bounded error for a given input; 3) each property set is composed of distinct elements; and 4) property sets differentiate inputs based on set membership.

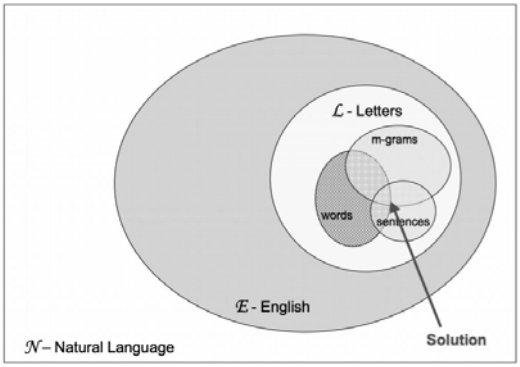


Fig. 4 – STE Application.

Any information (rules, constraints) known about the problem is encoded in sets in the solution space. Each rule or constraint is represented by its own unique set. Estimates may be a member of more than one set, but must be in at least one set to be considered. Information known about both the inputs and about the rules is treated similarly. Rules are expressed as assertions in STE. An assertion A takes the set of possible inputs and gives a set of resulting outputs, or solutions, for the operation, O, as specified in the rule.

The *m*-grams used in our STE approach are drawn from a survey of English prose styles from 1600 - 2000 AD [6] that are found in Project Gutenberg [11]. Each symbol in this meta-language represents an *m*-gram of English. This definition reduces the number of symbols in the meta-language alphabet to the number of unique *m*-grams allowed in English. All encryption occurs within a single meta-symbol. Results for block S and block P are shown in Table 2.

4.1. META-2-GRAM EXAMPLE

Consider the following encrypted input text:

fa_63_aa _fa_85_fa_63_aa_fa

where the plaintext is given by:

The_truth_is_the_truth

Table 2 gives the corresponding key for the encryption.

Table 2. Encryption Key.

Key		
63	→	et
85	→	is
aa	→	ru
fa	→	th

Applying the key to the encrypted input text results in the recovered plaintext:

th_et_ru_th_is_th_et_ru_th

The BCBB algorithm reads the encrypted meta-character input text from left to right as illustrated in Table 3.1.

Table 3.1. Encrypted Message Input String.

Read 1st	fa
Read 2nd	fa_63
Read 3rd	fa_63_aa
Read 4th	fa_63_aa _fa
Read 5th	fa_63_aa _fa_85
• • •	• • •

The decryption method takes the following form:

- Form sequences of two allowed 2-grams meta(2,2), e.g.,

  etru, isis, isth, ruth, thet, this

- Form a correspondence between *like* and *unlike* patterns using symbols such as A and B (Table 3.2) where A is different from B and vice versa.

Table 3.2. Meta(2,2) + Pattern Matching.

etru	isis	isth	ruth	thet	this
↕	↕	↕	↕	↕	↕
AB	AA	AB	AB	AB	AB

- Form sequences of three allowed 2-grams, meta(2,3) such as

  etruth, isthet, ruthis, thetru, thisth

- Form their associated patterns of meta(2,3) in Table 4.

Table 4. Meta(2,3) + Pattern Matching.

etruth	isthet	ruthis	thetru	thisth
↕	↕	↕	↕	↕
ABC	ABC	ABC	ABC	ABA

Finally for a maximum window size of four then

- Form sequences of four allowed 2-grams meta(2,4)

  etruthis, isthetru, ruthisth, thetruth, thisthet

- Form their associated matching patterns of (2,4) in Table 5.

Table 5. Meta(2,4) + Pattern Matching.

etruthis	isthetru	ruthisth	thetruth	thisthet
↕	↕	↕	↕	↕
ABCD	ABCD	ABCB	ABCA	ABAD

- Form sequence of encrypted input meta characters (window sizes two – four)
- Associate their patterns with patterns of the allowed sequence of m 2-grams (above), where m = {2, 3, 4}
- Use pattern matching to converge encrypted meta-character to plaintext characters

Table 6 illustrates the decryption matrix (d-matrix) before the start of the relaxation steps.

Table 6. D-Matrix.

	fa	63	aa	85
et	.	.	.	.
is	.	.	.	.
ru	.	.	.	.
th	.	.	.	.

where the decryption is built up from the meta-2-character inputs.

In this example, the input is read from left to right:

fa_63_aa _fa_85_fa_63_aa_fa

The relaxation steps are applied after each meta-character read of the inputs. In Read 1st, the meta-character fa is inserted into the d-matrix. Read 1st does not provide sufficient information to make an analysis. Thus relaxation waits on Read 2nd at which point the meta-character 63 is inserted into the d-matrix. The relaxation procedure can now form the combination fa63 and the associated pattern AB. Checking the allowed two 2-gram patterns: AB, AA, AB, AB, AB, AB from Table 3.2, we find that all patterns AB are possible so no change to the d-matrix is made. In the Read 3rd step, the meta-character aa is inserted into the d-matrix. At this point, the relaxation procedure forms the sequence fa63aa and assigns it the pattern ABC. This pattern is checked against the allowed three 2-gram, meta(2,3), in Table 4 with patterns: ABC, ABC, ABC, ABC, and ABA. Again since the pattern ABC occurs multiple times, no change is made to the d-matrix. After Read 4th, the relaxation step forms the sequence fa63aafa and its corresponding pattern ABCA. Checking allowed four 2-gram patterns, meta(2,4), in Table meta-4, we find the patterns: ABCD, ABCD, ABCB, ABCA, ABAD. Since there is only one occurrence of the pattern ABCA, this implies that A = 'fa' = 'th.' The mapping 'fa' to 'th' is termed a "*partial binding,*" that results in updating the d-matrix.

	fa	63	aa	85
et	0	.	.	.
is	0	.	.	.
ru	0	.	.	.
th	1	0	0	0

Next the meta-character 85 is read into the d-matrix. Since we have discovered that 'th' maps to 'fa', we form the following sequences: 63aa, 63aafa, 63aafa85 and their associated patterns AB, ABC, ABCD. Note that we have set the relaxation window size for analysis to be 4, which indicates that max (m) = 4. Using these patterns, we attempt to identify additional partial bindings: 63aa – AB is not bound to 'th', since 'fa' = 'th'; therefore, etru is the only mapping. Thus '63' = 'et' and 'aa' = 'ru' and we update the d-matrix to indicate the new partial bindings.

	fa	63	aa	85
et	0	1	0	0
is	0	0	0	.
ru	0	0	1	0
th	1	0	0	0

The d-matrix now has only one unique unresolved mapping; therefore, set '85' = 'is'

	fa	63	aa	85
et	0	1	0	0
is	0	0	0	1
ru	0	0	1	0
th	1	0	0	0

4.2 BCBB DECRYPTION RESULTS

Table 7, is a list of the books and authors that we used in testing the BCBB algorithm. Every text was correctly decrypted regardless of the cipher type employed. The time required for decryption was nearly identical for each cipher type (see Table 8). Variation in the time required for decryption appears to be dependent on several properties of the files. The properties identified are: 1) author style; 2) file size, non-standard English, such as name, place names, and imaginary words; 4) the ear in which the work was written; and 5) the original language in which the work was written.

Table 7. Files, Tiles, and Authors.

File	Title	Author
linc11.txt	The Writings of Abraham Lincoln	Abraham Lincoln
wr10.txt	On War	Carl von Clauswitz
alice30.txt	Alice in Wonderland	Lewis Carroll
anne11.txt	Anne of Green Gables	Lucy Maud Montgomery
hend10.txt	Howard's End	E. M. Forser
janc10.txt	Jefferson and His Colleagues	Allen Johnson
jmlta10.txt	The Jew of Malta	Christopher Marlowe
glass18.txt	Through the Looking Glass	Lewis Carroll
will10.txt	The Wind in the Willows	Kenneth Grahame
wrld10.txt	The Way of the World	William Congreve

Table 8. Two Byte Block Decryption Results (sec).

File	S	P	PSP	Mean
linc11.txt	675	669	676	670.4
wr10.txt	14507	14442	14273	14418
alice30.txt	44617	44690	44470	44527
anne11.txt	778	770	774	774
hend10.txt	861	847	848	841
janc10.txt	1387	1381	1388	1389
jmlta10.txt	12680	12723	12616	12626
glass18.txt	7851	7664	7603	7732.4
will10.txt	765	743	744	750.4
wrld10.txt	546	550	546	548.8

Authors have distinct styles of writing, including the use of same sentence structure and lexicon in all of their works. Reusing the same patterns in structure and words results in a set of m-grams trained with those patterns. As a consequence, authors that share similar patterns of styles should decrypt in similar times and number of ciphertext characters. For example, *Alice in Wonderland* and *Through the Looking Glass*, both by Lewis Carroll, show similar decryption results.

Of all the files tested, *Alice in Wonderland* had the greatest diversity of names. It took the longest time of all text files to decrypt. Correspondingly, *Through the Looking Glass* also took longer to decrypt than other test files, due to the imaginary words and names contained in the text. *The Jew of Malta*, a work that included a large number of foreign names and locations also had problems with the low frequency m-grams that result from those words. Patterns in those words, and consequently the m-grams, are not as likely to be represented in the m-gram sets.

During the time periods covered by the corpus, English usage evolved, changed, and has been re-characterized. Word and usage patterns regularly change with popularity. Changes in the lexicon and language habits can result in literary era dependent m-gram sets, and; therefore, give rise to different decryption performance. Customizing m-gram sets for a particular era, over which the language has remained relatively static, may increase future decryption accuracy and efficiency. Sets of data derived from the same time period as the message are more likely to consist of the same patterns of word usage and frequency as the message.

5. SUMMARY AND CONCLUSION

P is a subset of *S*. *P* has the proper that under *P* the number of 1's and 0's are preserved between the plaintext and the ciphertext. This proper is illustrated by the uniform decryption time seen in the results listed in Table 1. The application of the *S* cipher does not necessarily preserve the number of 1's and 0's. The *S* block cipher, therefore, requires greater time to converge to the correct cipher key. Since P is S, it is expected that *PSP* is also *S*. Therefore, under block *PSP* the decryption time should be proportional to a *S* cipher key. Therefore, under block *PSP* the decryption time should be proportional to *S*.

We have presented several results that indicate the application of the STE method on performing a known-ciphertext attack for block *S* and *P*. These results did not rely on either the use of the knowledge of the chosen-plaintext or the known-plaintext attacks. The statistic of the language is provided within the property sets of the STE in the form of m-grams. Letter frequencies are also used as a property set. In general, any block cipher method that employs products of *P* and *S* should be deciphered in *S* time, independent of the intermediate block product combinations.

As the size of the meta-s-character increases, the number of meta(s,m) grams in a language also increases. Successful decryption using the forbidden meta(s,m) sets necessitates having enough of the language represented in the sets to find valid language patterns for most messages. Variations in language style and lexicon affect the set size and membership. On the average, smaller allowed meta(s,m) gram sets are less likely to contain all of the meta(s,m) grams found in a message. The necessary size of the sets, compared to the meta-language, has not previously been studied and is unknown.

6. ACKNOWLEDGMENT

We are grateful to F. Mitchell and S. Mitchell for their extraordinary effort in assisting us with portions of the software development for which the results of this paper are based.

7. REFERENCES

[1] E. Fogal, and Y. F. Huang, On the value of information in system identification - bounded noise case, *Automatica*, (18) 2 (1982), pp. 229-238.

[2] A. Carlson, R. B. Wells, and R. E. Hiromoto, Using set theoretic estimation to implement Shannon secrecy theory, *The Proceedings of the Third IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS'2005*, Sofia, Bulgaria, September 5-7, 2005, pp. 435-438.

[3] C. E. Shannon, Communication theory of secrecy systems, *Bell Systems Technical Journal*, (28) (1949), pp. 656-715.

[4] B. Schneier, *Applied Cryptography: Protocols, Algorithms and Source Code in C*, 2nd ed., John Wiley and Sons Inc., New York, NY, 1996.

[5] R. B. Wells, *Applied Coding and Information Theory*, 1st ed., Prentice Hall, Upper Saddle River, NJ, 1999.

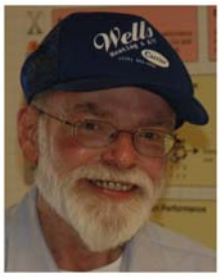
[6] S. Ross, *A First Course in Probability*, 1st ed., MacMillan Publishing, Inc., New York, NY, 1976.

[7] P. Garrett, *The Mathematics of Coding Theory*, 1st ed.. Pearson/Prentice Hall, Upper Saddle River, NJ, 2004.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2001.
[9] D. R. Stinson, *Cryptography: Theory and Practice*, 3rd ed., Chapman Hall/CRC Press, London, England, 2006.
[10] P. L. Combettes, The foundations of set theoretic estimation, *Proceedings of the IEEE*, (81) 2 (1993), pp. 182-208.
[11] The Gutenberg Project, Internet, http://www.gutenberg.net
[12] Fred Schweppe, Recursive state estimation: Unknown but bounded errors and system inputs, *IEEE Transactions on Automatic Control*, (13) 1 (1968), pp. 22-28.
[13] Albert Carlson, *Set Theoretic Estimation Applied to the Information Content of Ciphers and Decryption*, PhD Thesis, University of Idaho, 2012.
[14] Horst Feistel, Cryptography and computer privacy, *Scientific American*, (228) 5 (1973), pp. 15-20.

Albert H. Carlson *received his MS and Ph.D. degrees in Computer Science from the University of Idaho. Albert earned a BS in Computer Engineering from the University of Illinois, Urbana. With over 20 years experience as an Engineer and Engineering Manager, he has worked in safety critical circuits, consumer and industrial electronics, communications, control circuits, and embedded systems. Presently, he is an Assistant Professor at Fontbonne University, St. Louis, MO in the Department of Math and Computer Science.*

Robert E. Hiromoto*, received his Ph.D. degree in Physics from the University of Texas at Dallas. He is professor of computer science at the University of Idaho. His areas of research include ad hoc wireless mobile networks for unmanned autonomous vehicles, and information-based design of sequential and parallel algorithms, decryption techniques using set theoretic estimation, and parallel graphics rendering systems.*



Richard B. Wells *is Emeritus Professor and former Director of the UI Neuroscience Graduate Program, Professor of Electrical and Computer Engineering, and holds an Affiliate appointment with the Department of Physiology and Biophysics at the University of Washington School of Medicine. He also served as Associate Director of UI's MRC Institute and the Program Director for the UI's Neuroscience REU Site. He had 8 years of technical management experience at the Hewlett Packard Company prior to joining the UI in 1993. Dr. Wells is a Senior Member of IEEE.*

Hermann Heßling. The Challenge of Managing and Analyzing Big Data, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 204-209.

The amounts of data produced in science are growing exponentially. Traditional methods for storing and maintaining the enormous flood of data seem to be no longer sufficient anymore. The complexity of the data that will be distributed more and more worldwide, is going to constitute a considerable challenge for their analysis. According to Alex Scalay there soon will be produced so many data that they cannot even be stored and maintained anymore. The data have to be analyzed in real time in order to extract the relevant information. An outline of the project Large Scale Management and Analysis (LSDMA) is given. The status of our research group on distributed real-time computing is reviewed. Finally, a novel approach to time-dependent image processing based on local thermodynamical methods is presented.

Hirotaka Inoue, Kyoshiro Sugiyama. Self-Organizing Neural Grove: Efficient Neural Network Ensembles Using Pruned Self-Generationg Neural Trees, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 210-216.

Recently, multiple classifier systems have been used for practical applications to improve classification accuracy. Self-generating neural networks (SGNN) are one of the most suitable base-classifiers for multiple classifier systems because of their simple settings and fast learning ability. However, the computation cost ofthe multiple classifier system based on SGNN increases in proportion to the numbers of SGNN. In this paper, we propose a novel pruning method for efficient classification and we call this model a self-organizing neural grove (SONG). Experiments have been conducted to compare the SONG with bagging, the SONG with boosting, and support vector machine (SVM). The results show that the SONG can improve its classification accuracy as well as reducing the computation cost.

Igor Kotenko, Elena Doynikova. Comprehensive Multilevel Security Risk Assessment of Distributed Information Systems, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 117-225.

The paper suggests the multilevel approach to the risk assessment that is based on the system of security metrics and techniques for their calculation. Proposed techniques are based on attack graphs and service dependencies. They allow evaluating security of network topologies, malefactors and attack characteristics, and integral security properties and characteristics calculated on the basis of the cost-benefit and zero-day vulnerability analysis. Classification of these characteristics and separation of the security information on static, dynamic and historical allows defining different assessment levels. The paper considers the main issues and recommendations for using the risk assessment techniques based on the suggested approach.

Margarita Elkina, Albrecht Fortenbacher, Agathe Merceron. The Learning Analytics Application LeMo – Rationals and First Results, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 226-234.

LeMo is an open source application for learning analytics, which collects data about learners' activities from different platforms. This article describes design principles of LeMo in the context of creating an efficient tool for learning analytics. Focus is on the LeMo system architecture, user path analysis employing algorithms of sequential pattern mining, and visualization of learners' activities implemented in the current version. A case study shows first results.

Peter Arras, Yelizaveta Kolot, Galyna Tabunshchyk, Tomas Kozík. E-Learning Environment for the Remote Study in Material Properties Courses, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 235-240.

In this paper a newly developed e-learning environment for the support of the study on Material Properties is presented. It was developed to support the blended learning of the material and shape stiffness. Course structure is organized in HTML content, and virtual and remote laboratories are integrated in the computer aided learning module (CALM), which support both teachers during the hand-on teaching and students during self-study.

Linas Svilainis, Arturas Aleksandrovas, Kristina Lukoseviciute, Valdas Eidukynas. Comparison of Conventional and Spread Spectrum Signals Time of Flight Error Caused by Neighboring Reflections, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 241-247.

Performance comparison of conventional and spread spectrum signals in Time of Flight estimation is given. Ultrasonic measurement under multiple reflections condition is analyzed. It was indicated that if two reflections are in close proximity the neighboring signal induces energy leak into its opponent signal. Due to such situation ToF estimate is obtained with bias error. Narrow signals like single rectangular pulse, should suffer less from the aforementioned phenomena. But use of spread spectrum signals is preferred thanks to their compressibility. It was hypothesized that such long signals will have worse bias error due to neighbor reflection. Goal of the investigation was to compare the performance in multiple reflections environment in Time of Flight estimation for classical signals and spread spectrum signals. Investigation revealed that spread spectrum signals have better performance in a sense of bias error caused by neighboring reflection.

Andreas Fink, Helmut Beikirch. Enhancing the Coverage of Indoor Radio Localization by Distributed Computations, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 249-257.

The prevalent evaluation criterion for indoorlocal positioning systems (ILPS) is the achievable accuracy in terms of Euclidean distance between estimated and true position. Systems relying on received signal strength (RSS) rangingoften use a distributedcollection of RSS sensor data atreference nodes and a centralized position estimation.For this direct remote positioning, the accuracy isdependent on the reference node density and thus, is indirect proportional to the achievable coverage. To split up the dependency between these two criteria, we propose a distributed weighted centroid localization (dWCL) strategy with a hierarchical sensor data field bus. Accuracy and coverage of centralized and distributed WCL algorithms are compared for a one-dimensional tracking simulation and 196 reference nodes, arranged in up to 28 gateway segments. Using distributed computations, the localization system's coverage is increased by factor ten while the location estimation error increases only slightly.

Albert H. Carlson, Robert E. Hiromoto, Richard B. Wells. Breaking Block and Product Ciphers, *International Journal of Computing*, Vol. 12, Issue 3, 2013, pp. 259-266.

The security of block and product ciphers is considered using a set theoretic estimation (STE) approach to decryption. Known-ciphertext attacks are studied using permutation (P) and substitution (S) keys. The blocks are formed from two (2) alphabetic characters (meta-characters) where the applications of P and S upon the ASCII byte representation of each of the two characters are allowed to cross byte boundaries. The application of STE forgoes the need to employ chosen-plaintext or known-plaintext attacks.

Hermann Heßling. Завдання управління та аналізу великих обсягів даних, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 204-209.

Обсяг даних, якими оперують в наукових дослідженнях, зростає в геометричній прогресії. Традиційних методів зберігання та підтримки цього величезного потоку даних, схоже, більше не достатньо. Складність даних, які будуть розподілені все більше і більше в світовому масштабі, представляє собою значну проблему для їх аналізу. За словами Алекса Scalay, скоро буде вироблятися стільки даних, що вони більше навіть не зможуть зберігатися та обслуговуватися. Дані повинні аналізуватися в реальному часі, щоб отримувати з них відповідну інформацію. В статті представлено схему проекту «Великомасштабне управління та аналіз» та основні задачі нашої дослідницької групи з розподілених обчислень в реальному часі. Нарешті, представлено новий підхід до часово-залежної обробки зображень на основі локальних термодинамічних методів.

Hirotaka Inoue, Kyoshiro Sugiyama. Самоорганізована нейронна Grove: Ефективні ансамблі нейронної мережі з використанням відсічених само-генерованих нейронних дерев, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 210-216.

Останнім часом складні системи класифікаторів були використані для практичного застосування для підвищення точності класифікації. Самогенеруючі нейронні мережі (СГНМ) є одним з найбільш відповідних базових класифікаторів для складних систем класифікації через їх прості налаштування та можливість швидкого навчання. Проте, обчислювальні витрати складної системи класифікаторів на основі СГНМ зростає пропорційно чисельності СГНМ. У даній роботі ми пропонуємо новий метод відсічення для ефективної класифікації. Ми називаємо цю модель самоорганізованою нейронною grove (СОНГ). Були проведені експерименти для порівняння СОНГ з просіюванням, СОНГ з підсиленням та з машиною опорних векторів. Результати показують, що СОНГ можуть покращити точність класифікації, а також знизити обчислювальну складність.

Igor Kotenko, Elena Doynikova. Комплексна оцінка розподілених інформаційних систем багаторівневої оцінки ризику, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 117-225.

У статті пропонується багаторівневий підхід до оцінки ризиків, заснований на системі показників і методів безпеки для їх розрахунку. Пропоновані методи засновані на графіках залежностей атаки і сервісів. Вони дають змогу оцінити безпеку мережевих топологій, оцінити зловмисників та характеристики атаки, інтегральні властивості безпеки та характеристики, розраховані на основі затратно-прибуткового та середньо-нульового аналізу вразливості. Класифікація цих характеристик та поділ інформації про безпеку на статичні, динамічні та історичні дозволяє визначити різні рівні оцінки. У статті розглянуті основні питання та рекомендації щодо використання методів оцінки ризику, що засновані на запропонованому підході.

Margarita Elkina, Albrecht Fortenbacher, Agathe Merceron. Навчальний аналітичний додаток LEMO – цілі та перші результати, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 226-234.

LEMO є додатком з відкритим вихідним кодом для навчальної аналітики, яка збирає дані про діяльність осіб, що навчаються, з різних платформ. У цій статті описуються принципи проектування LEMO в контексті створення ефективного інструменту для навчальної аналітики. Акцент робиться на системній архітектурі LEMO, аналізі шляху користувача з використанням алгоритмів послідовного отримання образів та візуалізації діяльності осіб, що навчаються, реалізованих в поточній версії. Перші результати роботи показані на конкретному прикладі.

Peter Arras, Yelizaveta Kolot, Galyna Tabunshchyk, Tomas Kozík. Дистанційне навчання за допомогою електронного навчання у вивченні властивостей матеріалів, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 235-240.

У даній роботі представлене нещодавно розроблене середовище електронного навчання для підтримки дослідження з властивостей матеріалів. Це середовище було розроблене для підтримки змішаного навчання жорсткості матеріалів та форми. Структура курсу організовується в HTML зміст, і віртуальні та віддалені лабораторії інтегровані в автоматизовану систему навчального модуля (АСНМ), що підтримують викладачів під час викладання, і студентів під час самостійного вивчення.

Linas Svilainis, Arturas Aleksandrovas, Kristina Lukoseviciute, Valdas Eidukynas. Порівняння звичайних і широкоспектральних сигналів при визначенні помилки часу проходження електричного сигналу спричиненої сусідніми відображеннями, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 241-247.

В статті дається порівняння продуктивності звичайних і широкоспектральних сигналів при визначенні часу проходження електричного сигналу. Аналізується ультразвукове вимірювання при умовах багаторазових відображень. Було відзначено, що, якщо два відображення знаходяться в безпосередній близькості, то сусідній сигнал викликає витік енергії в протилежний сигнал. Через таку ситуацію оцінка часу проходження електричного сигналу виходить з помилкою зміщення. Вузькі сигнали подібні одиночним прямокутним імпульсам, повинні менше страждати від вищезазначених явищ. Але використання широкоспектральних сигналів є кращим завдяки їх стисненню. Було висловлено припущення, що такі широкі сигнали будуть мати гіршу помилку зміщення через сусіднє відображення. Мета дослідження полягала в порівнянні ефективності в середовищі з декількома відображеннями в момент визначення часу проходження електричного сигналу для класичних і широкоспектральних сигналів. Дослідження показали, що широкоспектральні сигнали мають меншу систематичну похибку, викликану сусідніми відображеннями.

Andreas Fink, Helmut Beikirch. Підвищення покриття внутрішньої радіо-локалізації за допомогою розподілених обчислень, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 248-258.

Поширеним критерієм оцінки для внутрішніх локальних систем позиціонування (ILPS) є досяжна точність з точки зору евклідової відстані між очікуваною та реальною позицією. Системи, спираючись на ранжування сили прийнятого сигналу (RSS), часто використовують розподілений набір даних датчиків RSS в опорних вузлах і централізовану оцінку позиції. Для цього прямого віддаленого позиціонування, точність залежить від щільності опорного вузла і, таким чином, є непрямо пропорційна до досяжного покриття. Щоб розділити цю залежність між цими двома критеріями, ми пропонуємо стратегію зваженого центру розподіленої локалізації (dWCL) з ієрархічною шиною даних датчика. Точність і покриття централізованих і розподілених алгоритмів WCL порівнюються для одновимірного відстеження модельованих і 196 опорних вузлів, розташованих в 28 сегментах шлюзів. Завдяки використанню розподілених обчислень, охоплення локалізаційними системами збільшується в десять разів в той час як похибка оцінки позиції збільшується незначно.

Albert H. Carlson, Robert E. Hiromoto, Richard B. Wells. Злам блочних та продукційних шифрів, *Міжнародний журнал Комп'ютинг*, том. 12, випуск 3, 2013, с. 259-266.

Розглядається безпека блочних та продукційних шифрів з використанням підходу теоретико-множинної оцінки (STE) до розшифрування. Відомі атаки на зашифрований текст вивчаються за допомогою ключів перестановки (P) і заміщення (S). Блоки формуються з двох (2) алфавітних символів (мета-символів), де застосування ключів P і S у виді ASCII байту кожного з двох символів, яким дозволено перетинати межі байтів. Застосування STE примушує до необхідності використання простого тексту або атак з використанням простого тексту.

Hermann Heßling. Задания управления и анализу больших объемов данных, *Международный журнал Компьютинг*, том. 12, выпуск 3, с. 204-209.

Объем данных, которыми оперируют в научных исследованиях, растет в геометрической прогрессии. Традиционных методов хранения и поддержки этого огромного потока данных, похоже, больше не достаточно. Сложность данных, которые будут распределены все больше и больше в мировом масштабе, представляет собой значительную проблему для их анализа. По словам Алекса Scalay, скоро будет производиться столько данных, что они больше даже не смогут храниться и обслуживаться. Данные должны анализироваться в реальном времени, чтобы получать из них соответствующую информацию. В статье представлена схема проекта «Крупномасштабное управления и анализ» и основные задачи нашей исследовательской группы по распределенных вычислений в реальном времени. Наконец, представлен новый подход к временно-зависимой обработке изображений на основе локальных термодинамических методов.

Hirotaka Inoue, Kyoshiro Sugiyama. Самоорганизированная нейронная Grove: Эффективные ансамбли нейронной сети с использованием отсеченных само-генерированных нейронных деревьев, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 210-216.

В последнее время сложные системы классификаторов были использованы для практического применения в повышении точности классификации. Самогенерирующие нейронные сети (СГНС) являются одним из наиболее подходящих базовых классификаторов для сложных систем классификации из-за их простых настройки и возможности быстрого обучения. Однако, вычислительные затраты сложной системы классификаторов на основе СГНС растут пропорционально численности СГНМ. В данной работе мы предлагаем новый метод отсечения для эффективной классификации. Мы назвали эту модель самоорганизованной нейронной Grove (СОНГ). Были проведены эксперименты для сравнения СОНГ с отсечением, СОНГ с усилением и с машиной опорных векторов. Результаты показывают, что СОНГ могут улучшить точность классификации, а также снизить вычислительную сложность.

Igor Kotenko, Elena Doynikova. Комплексная оценка распределенных информационных систем многоуровневой оценки риска, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 117-225.

В статье предлагается многоуровневый подход к оценке рисков, основанный на системе показателей и методов безопасности для их расчета. Предлагаемые методы основаны на графиках зависимостей атаки и сервисов. Они позволяют оценить безопасность сетевых топологий, оценить злоумышленников и характеристики атаки, интегральные свойства безопасности и характеристики, рассчитанные на основе затратно-прибыльного и средне-нулевого анализа уязвимости. Классификация этих характеристик и разделение информации о безопасности на статические, динамические и исторические позволяет определить различные уровни оценки. В статье рассмотрены основные вопросы и рекомендации по использованию методов оценки риска, основанные на предложенном подходе.

Margarita Elkina, Albrecht Fortenbacher, Agathe Merceron. Учебное аналитическое приложение LEMO – цели и первые результаты, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 226-234.

LEMO является приложением с открытым исходным кодом для учебной аналитики, которая собирает данные о деятельности обучающихся, с различных платформ. В этой статье описываются принципы проектирования LEMO в контексте создания эффективного инструмента для учебной аналитики. Акцент делается на системной архитектуре LEMO, анализе пути пользователя с использованием алгоритмов последовательного получения образов и визуализации деятельности обучающихся, реализованных в текущей версии. Первые результаты работы показаны на конкретном примере.

Peter Arras, Yelizaveta Kolot, Galyna Tabunshchyk, Tomas Kozík. Дистанционное обучение с помощью электронного обучения в изучении свойств материалов, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 235-240.

В данной работе представлено недавно разработанную среду электронного обучения для поддержки исследования свойств материалов. Эта среда была разработана для поддержки

смешанного обучения жесткости материалов и формы. Структура курса организуется в HTML содержание, и виртуальные и удаленные лаборатории интегрированы в автоматизированную систему учебного модуля (АСНМ), поддерживающих преподавателей во время преподавания, и студентов во время самостоятельного изучения.

Linas Svilainis, Arturas Aleksandrovas, Kristina Lukoseviciute, Valdas Eidukynas. Сравнение обычных и широкоспектральных сигналов при определении ошибки времени прохождения электрическоно сигналу вызванной соседними отображениями, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 241-247.

В статье дается сравнение производительности обычных и широкоспектральных сигналов при определении времени прохождения электрического сигнала. Анализируется ультразвуковое измерения при условиях многократных отражений. Было отмечено, что, если два отображения находятся в непосредственной близости, то соседний сигнал вызывает утечку энергии в противоположный сигнал. В такой ситуации оценка времени прохождения электрического сигнала получается с ошибкой смещения. Узкие сигналы подобные одиночным прямоугольным импульсам, должны меньше страдать от вышеупомянутых явлений. Но использование широкоспектральных сигналов является лучшим благодаря их сжатию. Было высказано предположение, что такие широкие сигналы будут иметь худшую ошибку смещения через соседнее отображение. Цель исследования состояла в сравнении эффективности в среде с несколькими отражениями в момент определения времени прохождения электрического сигнала для классических и широкоспектральных сигналов. Исследования показали, что широкоспектральные сигналы имеют меньшую систематическую погрешность, вызванную соседними отражениями.

Andreas Fink, Helmut Beikirch. Повышение покрытия внутренней радио-локализации с помощью распределенных вычислений, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 248-258.

Распространенным критерием оценки для внутренних локальных систем позиционирования (ILPS) является достижимая точность с точки зрения евклидового расстояния между ожидаемой и реальной позицией. Системы, опираясь на ранжирование силы принимаемого сигнала (RSS), часто используют распределенный набор данных датчиков RSS в опорных узлах и централизованную оценку позиции. Для этого прямого удаленного позиционирования, точность зависит от плотности опорного узла и, таким образом, является косвенно пропорциональной достижимому покрытию. Чтобы разделить эту зависимость между этими двумя критериями, мы предлагаем стратегию взвешенного центра распределенной локализации (dWCL) с иерархической шиной данных датчика. Точность и покрытие централизованных и распределенных алгоритмов WCL сравниваются для одномерного отслеживания моделируемых и 196 опорных узлов, расположенных в 28 сегментах шлюзов. Благодаря использованию распределенных вычислений, охват локализационными системами увеличивается в десять раз, в то время, как погрешность оценки позиции увеличивается незначительно.

Albert H. Carlson, Robert E. Hiromoto, Richard B. Wells. Взлом блочных и продукционных шифров, *Международный журнал Компьютинг*, том. 12, выпуск 3, 2013, с. 259-266.

Рассматривается безопасность блочных и продукционных шифров с использованием подхода теоретико-множественной оценки (STE) для расшифровки. Известны атаки на зашифрованный текст изучаются с помощью ключей перестановки (P) и замещения (S). Блоки формируются из двух (2) алфавитных символов (мета-символов), где применение ключей P и S в виде ASCII байта каждого из двух символов, которым разрешено пересекать границы байтов. Применение STE принуждает к необходимости использования простого текста или атак с использованием простого текста.

Prepare your paper according to the following guides:

Unconditional requirement - paper should not be published earlier.

(i) Suggested composition (frame) of paper:
- Issue formulation stressing its urgent solving; evaluation of recent publications in the explored issue
- short formulation of paper's purpose
- description of proposed method (algorithm)
- implementation and testing (verification)
- conclusion.

(ii) Use A4 (210 x 297 mm) paper. Size of paper has to be extended up to 6-8 pages.

(iii) Please use main text two column formatting;

(iv) A paper must have an abstract and some keywords;

(v) Place a full list of references at the end of the paper. Please place the references according to their order of appearance in the text.

(vi) An affiliation of each author is wanted.

(vii) The text should be single-spaced. Use Times New Roman (11 points, regular) typeface throughout the paper.

(viii) Equations should be placed in separate lines and numbered. The numbers should be within brackets and right aligned.

(ix) The figures and tables must be numbered, have a self-contained caption. Figure captions should be below the figures; table captions should be above the tables. Also, avoid placing figures and tables before their first mention in the text.

(x) As soon as you have the complete materials, the final versions should come electronically in MS Word'97 of MS Word 2000 format to the address computing@computingonline.net.

(xi) A hardcopy of your article is needed to be sent by regular mail for our publishing house.

(xii) Please send short CVs (up to 20 lines) and photos of every author.

(xiii) There is no other formatting required. The publishing department makes all rest formatting according to the publisher's rules.

Journal Topics:
- Algorithms and Data Structure, Software Tools and Environments
- Bio-Informatics
- Computational Intelligence
- Computer Modeling and Simulation
- Cyber and Homeland Security
- Data Communications and Networking
- Data Mining, Knowledge Bases and Ontology
- Digital Signal Processing
- Distributed Systems and Remote Control
- Education in Computing
- Embedded Systems
- High Performance Computing and GRIDS
- Image Processing and Pattern Recognition
- Intelligent Robotics Systems
- Internet of Things
- IT Project Management
- Wireless Systems

Основні вимоги до подання і оформлення публікацій наукового журналу "Комп'ютинг":

Безумовною вимогою є те, щоб стаття не була опублікована раніше!

(i) Наукові статті повинні мати такі необхідні елементи:
- постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями;
- аналіз останніх досліджень і публікацій, в яких започатковано розв'язання даної проблеми і на які спирається автор, виділення невирішених раніше частин загальної проблеми, котрим присвячується означена стаття;
- формулювання цілей статті (постановка завдання);
- виклад основного матеріалу дослідження з повним обгрунтуванням отриманих наукових результатів;
- висновки з даного дослідження і перспективи подальших розвідок у даному напрямку.

(ii) Використовуйте A4 (210 x 297 mm) формат сторінки. Загальний розмір статті має містити 6-8 сторінок.

(iii) Використовуйте двоколонкове форматування основного тексту;

(iv) Стаття повинна обов'язково містити основний текст українською мовою, анотацію (написану на Англійській і Українській мовах) і список ключових слів;

(v) В кінці статті розмістіть список літератури. Розміщуйте список літератури в порядку її цитування.

(vi) Необхідною є інформація про наукові звання, титули та посади авторів.

(vii) Текст повинен бути набраним одинарним інтервалом із використанням шрифту Times New Roman (11 points, regular).

(viii) Формули повинні відділятись від основного тексту пустими стрічками а також пронумеровані у круглих дужках та відцентровані по правому краю.

(ix) Таблиці і рисунки повинні бути пронумерованими. Заголовки рисунків розміщують під рисунком по центру. Заголовки таблиць розміщують по центру зверху таблиці.

(x) Завершені версії статей повинні бути надісланими в електронному MS Word'97 або MS Word 2000 форматі за адресою computing@computingonline.net.

(xi) Просимо надсилати поштою роздруковані копії статей.

(xii) В кінці кожної статті потрібно подати її назву, резюме (абстракт) і ключові слова англійською мовою.

(xiii) Просимо надсилати нам короткі біографічні дані (до 20 рядків) і скановані фотографії кожного із авторів.

(xiv) Видавництво здійснює остаточне форматування тексту згідно із вимогами друку.

(xv) У закордонних читачів можуть виникнути проблеми при ознайомленні з працями на російській та українській мовах. В зв'язку з цим редакційна колегія просить авторів додатково прислати розширений реферат (резюме), щоб б містило дві сторінки тексту англійською мовою, і супроводжувалось заголовком, прізвищами та адресами авторів. Авторам рекомендується використовувати у рисунках статті позначення переважно англійською мовою, або давати переклад у дужках. Тоді у розширеному резюме можна буде посилатися на рисунки у основному тексті.

Тематика журналу:

- Алгоритми та структури даних, програмні засоби та середовище
- Біо-інформатика
- Обчислювальний інтелект
- Комп'ютерне та імітаційне моделювання
- Кібернетична безпека та захист від тероризму
- Передача даних та комп'ютерні мережі
- Видобування даних, бази знань та онтології
- Цифрова обробка сигналів
- Розподілені системи та дистанційне управління
- Освіта в комп'ютингу
- Вбудовувані системи
- Високопродуктивні обчислення та ГРІД
- Обробка зображень та розпізнавання образів
- Інтелектуальні робототехнічні системи
- Інтернет речей
- Управління ІТ-проектами
- Безпровідні системи

Основные требования к подаче и оформлению публикаций научного журнала “Компьютинг”:

Безусловное требование – чтобы статья не была опубликована ранее!

(i) Научные статьи должны иметь такие необходимые элементы:
- постановка проблемы в общем виде и ее связь с важными научными или практическими задачами;
- анализ последних исследований и публикаций, в которых начаты решения данной проблемы и на которые опирается автор, выделение нерешенных прежде частей общей проблемы, которым посвящается обозначенная статья;
- формулирование целей статьи (постановка задача);
- изложение основного материала исследования с полным обоснованием полученных научных результатов;
- выводы из данного исследования и перспективы дальнейших изысканий в данном направлении.

(ii) Используйте A4 (210 x 297 mm) формат страницы. Общий размер статьи 6-8 страниц.

(iii) Используйте двухколоночное форматирование основного текста;

(iv) Статья должна обязательно содержать основной текст на Русском языке, аннотацию (написанную на Английском и Русском языках) и список ключевых слов;

(v) В конце статьи разместите список литературы. Размещайте список литературы в порядке ее цитирования.

(vi) Необходима информация о научных званиях, титулах и должностях авторов.

(vii) Текст должен быть набранным одинарным интервалом с использованием шрифта Times New Roman (11 points, regular).

(viii) Формулы должны отделяться от основного текста пустыми строками, а также пронумерованные в круглых скобках и отцентрованные по правому краю.

(ix) Таблицы и рисунки должны быть пронумерованными. Заголовки рисунков размещают под рисунком по центру. Заголовки таблиц размещают по центру сверху таблицы.

(x) Завершенные версии статей должны быть присланы в электронном MS Word'97 или MS Word 2000 формате по адресу computing@computingonline.net.

(xi) Просим присылать распечатанные копии статей по почте.

(xii) В конце каждой статьи необходимо предоставить ее название, резюме (абстракт) и ключевые слова на английском языке.

(xiii) Просим присылать нам короткие биографические данные (до 20 строчек) и сканированные фотографии каждого из авторов.

(xiv) Издательство осуществляет окончательное форматирование текста в соответствии с требованиями печати.

(xv) У зарубежных читателей могут возникнуть проблемы при ознакомлении с трудами на русском и украинском языках. В связи с этим редакционная коллегия просит авторов дополнительно прислать расширенный реферат (резюме), который содержал бы две страницы текста на английском языке, и сопровождался заголовком, фамилиями и адресами авторов. Авторам рекомендуется использовать в рисунках статьи обозначения преимущественно на английском языке, или давать перевод в скобках. Тогда в расширенном резюме можно будет посылаться на рисунки в основном тексте.

Тематика журнала:

- Алгоритмы и структуры данных, программные средства и среды
- Био-информатика
- Вычислительный интеллект
- Компьютерное и имитационное моделирование
- Кибернетическая безопасность и защита от терроризма
- Передача данных и компьютерные сети
- Добыча данных, базы знаний и онтологии
- Цифровая обработка сигналов
- Распределенные системы и дистанционное управление
- Образование в компьютинге
- Встраиваемые системы
- Высокопроизводительные вычисления и ГРИД
- Обработка изображений и распознавание образов
- Интеллектуальные робототехнические системы
- Интернет вещей
- Управление IT-проектами
- Беспроводные системы

TENTATIVE CALL FOR PAPERS

IDAACS'2015

The 8th IEEE International Conference on
Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications

September 24-26, 2015
WARSAW, POLAND

Organized by
Research Institute for Intelligent Computer Systems, Ternopil National Economic University and V.M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine
in cooperation with

Faculty of Electronics and Information Technologies, Faculty of Mathematics and Information Science, Warsaw University of Technology
Warsaw, Poland
www.idaacs.net

Conference Co-Chairmen
Anatoly Sachenko, Ukraine
Wiesław Winiecki, Germany
International Programme Committee Co-Chairmen
Robert E. Hiromoto, USA
Linas Svilainis, Lithuania
IDAACS International Advisory Board
Dominique Dallet, France
Richard Duro, Spain
Domenico Grimaldi, Italy
Vladimir Haasz, Czech Republic
Robert Hiromoto, USA
Theodore Laopoulos, Greece
George Markowsky, USA, Chair
Vladimir Oleshchuk, Norway
Fernando Lopez Pena, Spain
Peter Reusch, Germany
Anatoly Sachenko, Ukraine
Wiesław Winiecki, Poland

Important dates
Abstract submission: 1 February 2015
Notification of Acceptance: 31 March 2015
Camera ready paper: 15 May 2015
Early registration: 31 March – 15 June 2015

Correspondence
The correspondence should be directed to
IDAACS Organizing Committee:
IDAACS Organizing Committee
Research Institute for Intelligent Computer Systems
Ternopil National Economic University
3 Peremoga Square
Ternopil 46020 Ukraine
Phone: +380352-475050 ext.: 12234
Fax: +380352-475053 (24 hours)
E-mail: orgcom@idaacs.net

Tentative Streams and Topics

The conference scope includes, but it's not limited to:

1. Special Stream in Wireless Systems
2. Special Stream in Project Management
3. Special Stream in Cyber Security
4. Special Stream in High Performance Computing
5. Special Stream in eLearning Management
6. Special Stream in Advanced Information Technologies in Ecology
7. Special Stream in Intelligent Robotics and Components
8. Advanced Instrumentation and Data Acquisition Systems
9. Advanced Mathematical Methods for Data Acquisition and Computing Systems
10. Artificial Intelligence and Neural Networks for Advanced Data Acquisition and Computing Systems
11. Bio-Informatics
12. Software Tools and Environments
13. Data Analysis and Modeling
14. Embedded Systems
15. Information Computing Systems for Education and Commercial Applications
16. Intelligent Distributed Systems and Remote Control
17. Intelligent Information Systems, Data Mining and Ontology
18. Intelligent Testing and Diagnostics of Computing Systems
19. Internet of Things
20. Pattern Recognition and Digital Image and Signal Processing
21. Virtual Instrumentation Systems

It's our pleasure to invite the interested scientists to organize own Streams.

Підписано до друку 30. 08. 2013 р.
Формат 60х84 $^{1}/_{8}$. Гарнітура Times.
Папір офсетний. Друк на дублікаторі.
Умов. друк. арк. 11,5. Облік.-вид. арк. 13,1.
Зам. № 003-13П. Наклад 300 прим.

Тернопільський національний економічний університет
вул. Львівська, 11, м. Тернопіль, 46004

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців ДК № 3467 від 23.04.2009 р.

Віддруковано у Видавничо-поліграфічному центрі «Економічна думка ТНЕУ»
46004 м. Тернопіль, вул. Львівська, 11
тел. (0352) 47-58-72
E-mail: edition@tneu.edu.ua