Information technology applications

ITA

2 2023

PANEURÓPSKA VYSOKÁ ŠKOLA
FAKULTA INFORMATIKY

PAN-EUROPEAN UNIVERSITY
Faculty of Informatics

VZDELÁVANIE VEDA VÝSKUM
EDUCATION SCIENCE RESEARCH

SSKI
SLOVENSKÁ SPOLOČNOSŤ PRE KYBERNETIKU A INFORMATIKU PRI SAV
SLOVAK SOCIETY FOR CYBERNETICS AND INFORMATICS

International Journal of Information Technology Applications (ITA)

Volume 12, Number 2, December 2023

AIMS AND SCOPE OF ITA

The primary aim of the International Journal of Information Technology Applications (ITA) is to publish high-quality papers of new development and trends, novel techniques, approaches and innovative methodologies of information technology applications in the broad areas. The International Journal of ITA is published twice a year. Each paper is refereed by two international reviewers. Accepted papers will be available online with no publication fee for authors. The International Journal of ITA is being prepared for the bibliographic scientific databases.

International Journal of Information Technology Applications (ITA)

Instructions for authors

The International Journal of Information Technology Applications is welcoming contributions related with the journal´s scope. Scientific articles in the range approximately 10 standard pages are reviewed by two international reviewers. Reports up to 5 standard pages and information notices in range approximately 1 standard page are accepted after the decision of editorial board. Contributions should be submitted via e-mail to the editorial office. The language of contributions is English. Text design should preserve the layout of the template file, which may be downloaded from the webpage of journal. Papers in this journal are provided under diamond acess policy (no fee, open website) and authors declare their own Creative Commons license.

Deadlines of two standard issues per year (special issues are possible)

paper submission deadline	– end of May/end of November
review deadline	– continuous process
release date	– June/December

Editorial office address

Faculty of Informatics, Pan-European University, Tematínska 10, 851 05 Bratislava, Slovakia
juraj.stefanovic@paneurouni.com

Published by

Pan-European University, Slovakia, https://www.paneurouni.com/
Paneurópska vysoká škola, n.o., Tomášikova 20, 821 02 Bratislava, IČO 36 077 429
Civil Association EDUCATION-SCIENCE-RESEARCH, Slovakia, https://v-v-v.eu/
OZ VZDELÁVANIE -VEDA-VÝSKUM, Andrusovova 5, 851 01 Bratislava,
IČO 42 255 180
Slovak Society for Cybernetics and Informatics (SSKI)
at the Slovak Academy of Sciences, **Slovakia**
Slovenská spoločnosť pre kybernetiku a informatiku pri SAV (SSKI),
Ústav automobilovej mechatroniky, Fakulta elektrotechniky a informatiky STU,
Ilkovičova 3, 812 19 Bratislava 1, IČO 00 178 730, https://www.sski.sk/

Electronic online version of journal

https://www.itajournal.com/
(Open Journal System)

visit Archive and Instructions for authors:

Print version

Multigrafika s.r.o., Rajecká 13, 821 07 Bratislava

Print issues

Contact the editorial office,
print issues are available until they are in stock.

ISSN: 2453-7497 (online)
ISSN: 1338-6468 (print version)

Registration No.: EV 4528/12

Contents

Editorial

Dear readers,

this issue contains three articles giving interests to information technology applications in logistics and related activities. Another two articles are devoted to misconducts they emerge within the use of information technology.

Several seasons before, we have modified design and style of our journal, its cover, articles and imprint information. Language was set to English only. Third publisher joined to this journal (SSKI). All articles were free of charge for authors and online readers since the very beginning in the year 2012 and we are going to keep this *diamond access* style in the future. Now we ask authors to declare their Creative Commons license by each paper. In autumn 2023, the whole journal was transferred to own webpage under Open Journal System (OJS) - all its issues since beginning in 2012. In 13th season 2024, editorial board will be reconfigured, to get fresh and more involved contacts. This journal is not registered in known databases yet, the interest of prospective authors is certainly small and the next step is to finish all needed adjustments to be incorporated into databases.

Information Technology Applications - the name of our journal is covering a wide range of IT research and practice topics. In usual circumstances, our priority topics shold be narrowed to domain of *Industry 4.0* and *augmented reality* and their interconnection eventually. Special issues and extraordinary issues of ITA journal are possible every time.

Thank you for your contributions, we welcome your latest research results and solutions and we look forward to further collaborations.

Ing. Juraj Štefanovič, PhD.
ITA Executive editor

International Journal
Information Technology Applications (ITA)

Volume 12, Number 2, December 2023

ADVANCED INFORMATION AND COMMUNICATION NETWORKS FOR VEHICLE CONTROL

Florian Dietze, Štefan Kozák

Abstract:

The paper deals with the design and application of advanced communication technologies for vehicle process control. We propose the modification of existing bus systems in order to create a unique system which grants benefits for a radical approach. The sum of all modifications will lead to a revolutionary new bus access method which is suitable for existing control units and fit for future demands by eliminating the restrictions of existing bus technologies and simultaneous implying the benefits of the different access methods.

Keywords:

Applied informatics, communication networks, Bus technology, CAN bus, FlexRay Bus, Most Bus.

Introduction

From the very beginning of automobile creation in the late 19th century, the benefits of individual travelling caused more and more people to use cars. The demands of the owners and legal regulations to increase safety have increased the features enormously and force the automotive industry to continuously develop better cars. The electronics have a severe influence to the convenience and safety of the customer. Just 50 years back, the light was the only electronic device on board. Now, electronic is a major part of every single car concept. While some electronics are hidden (like electronic fuel injection), others bring new experiences to the customer, like global positioning system (GPS) based navigation.

For all vehicle concepts, which are considered in this document, there are requirements for the cross-linking, which may be identical. This is because all vehicles have the commonness to meet the requirements of traffic safety. So, in any small cars as well as the most expensive luxury vehicle blinkers, we can find wipers and brake lights, to list just a few common examples. While the design of the systems often differs fundamentally, the connection to a vehicle networking concept does not necessarily. Whether chic LED or old-fashioned light bulb taillights: in both cases, a signal from the brake pedal to the brake light in the rear is routed. There are also components that despite the same tasks have very different demands on the network. The headlight of a small car has light bulbs for low beam, high beam, parking lights and turn signals. To realize a luxury car add some motors and positioning controllers to adaptive headlights, automatic headlamp levelling, bi-xenon function and different brightness of the LED strip to DRL or sidelight. The fog lights in addition unilaterally light up when cornering. Hence, there is need to create individual concepts for different classes of vehicles.

However, this can then be used for several variants, as another body style (e.g. station wagon instead Saloon) has no dramatic effect on the functions. In modern vehicles almost all electrical components are interconnected [2]. The degree of crosslinking is dependent on the number of electronics, the field of use of the vehicle and in particular the class.

In order to control all the functions required of the vehicle, the car manufacturer could use a central computer, which would have to be very fast and expensive and had to connect each actuator, sensor or switch with at least one line (safety components additional monitoring lines).

This would result in a thick and heavy harness and in addition lead to higher fuel consumption, significantly higher manufacturing cost, since copper cables are not cheap at all. Alternatively, you can split the functions of the car. For similar tasks create functional blocks, which can then be laced into a closed area to avoid long connections. Another advantage of this concept is that it is characterized by short connection injects less conducted and radiated interference into the line. In the automotive world these function blocks are called electronic control units (ECUs) or components.

The components are in most cases not able to perform its task without information from other components. But not all of the data of each control device is important for the function. The stability management certainly needs the wheel speed sensor data, but no information on whether the rear wiper is currently active. The components must indeed exchange data, but they can be summed up in one another (more / less) independent groups. The demands on the data rate are quite different. A power window switch is usually operated less often than a wheel sensor passes its speed value. The faster the data transfer needed, the more expensive is the essential bus. The current bus systems are currently used in the following chapter explained again in detail. Certain information is important for all components, such as whether the ignition is turned on, or the car is parked and every single component should change to a low-power mode in order to save battery power. The essential information needed by several groups, are transported via a gateway control device into the respective group. Individual groups of controllers can be combined in bus systems that are significantly different.

The choice of the appropriate bus system depends on many factors: the required data rate and quantity, safety requirements, portability to other vehicle concepts, expandability and more. But certainly a big factor is the affordability of the concept. Especially in the compact car class, cost pressure assumes enormous influence on the decision of the applicable options. In general, the price of a bus system increases with its transmission speed. Universal and widely used systems are also cheaper than exotic custom solutions. Such special buses are found hardly. Systems from other areas (eg, Ethernet), which, although cheap and widely spread and would therefore be seen gladly for financial reasons. But these systems often do not meet the stringent technical requirements for automotive electronics. Thus, for example, Ethernet devices designed for room temperature and plugs of patch cables do not provide sufficient protection against dust and vibration in the engine compartment. Systems for military applications provide certainly technically perfect conditions, but considering the cost would cause tears to the watchful eyes of the controlling department. For derivatives that are designed for entry markets, rather simple networking concepts are in the specifications. This allows cost savings in the double sense:

The number of networked control units is reduced and also the wiring harness of the vehicle is less expensive and lighter. Since common networks, especially in the low-price segment are based on copper interconnects, this means cost reduce due to currently high raw material prices.

High-priced luxury cars have networked features that would not be possible without networking: camera-based environmental sensors, ultrasonic based parking systems, distance RADAR and LIDAR can be merged so that the sensor cloud creates a digital image of the vehicle environment. The mutual plausibility of data of different measurement methods can be life saving in certain circumstances, important decisions, such as pre-crash detection, multi-collission brake assists, safety isolation of high-voltage systems.

Here classic CAN-based networking solutions are reaching their limits. Firstly, the bus load increases with the number of networked control units and the associated bus load. Too high bus load makes the crosslinking unstable, so in these cases the control units, after their installation or area separated into different sub-buses (which means that, in some currently produced SUVs up to seven CAN bus via a gateway, communicate with each other). Secondly, just safety-critical applications need to check the plausibility of data not only the readings but also a temporal mapping in order to derive the change and the rate of change from it.

1 Automotive Communication Systems

This part of paper provides information about bus systems which are used in nearly every single car produced nowadays.

A bus is a subsystem that transfers data between components inside a device or between devices inside of a system. Unlike a point-to-point connection, a bus can logically connect several electronic control units over the same set of wires.

Each bus defines its set of physical connectors, timing and access methods together. In the automotive industry, several bus system have been established, tailored to the requirements that are expected in today's vehicles. CAN and LIN are the most established, MOST has found its place with the implementation of multimedia interfaces in 21st century. The latest bus system is currently FlexRay.

2 Advanced Automotive Bus Systems

FlexRay marks the latest development of bus systems at the moment (Fig.1). It was created by a consortium of car manufacturers, suppliers and semiconductor producers in the years 2000 to 2010. It is found in high-class premium cars like Audi Q7, BMW X5 and 7series, Mercedes S-class.

It is a serial deterministic and fault tolerant bus system, which should be up to the task on the vehicle network in the near future. Advantage over the CAN is the real-time capability, higher data transfer rates and reliability. Several channels (typically two) that are synchronized with each other, can interconnect the participants. In case of short circuit /interruption of a single channel communication still takes place [1]. Similarly to CAN, data is packed into frames, but the access to the bus is managed differently. There is a static segment, in which participants are allowed to send in a defined sequence data packets of fixed length (a ECU may also several packets send by it is assigned multiple IDs) and a dynamic segment, messages in the subscriber in a fixed oder flexible length. The fixed order is necessary since (unlike CAN) no arbitration is possible due a missing recessive bus level. FlexRay has three states: "0", "1" and "idle", and all three are dominant. The sender of a message that reads back the same time and therefore cannot see any bus errors due to collisions. It is not specified to send error frames by other participants on the bus during transmission. Fault is detected at the end of the frame for several error sums (CRC header and data CRC).

FlexRay charakteristics and parameters:

- max. data transmission: 10 Mbit/s per channel
- max. length of wiring: 24m (48m in an active star topology)
- based on twisted pair copper harness
- guaranteed latency
- flexible topology

The FlexRay communications bus is a deterministic, fault-tolerant and high-speed bus system developed in conjunction with automobile manufacturers and leading suppliers. FlexRay delivers the error tolerance and time-determinism performance requirements for x-by-wire applications (i.e. drive-by-wire, steer-by-wire, brake-by-wire, etc.). This appendix covers the basics of FlexRay. Increasing Communications Demands For automobiles to continue to improve safety, increase performance, reduce environmental impact, and enhance comfort, the speed, quantity and reliability of data communicated between a car's electronic control units (ECU) must increase. Advanced control and safety systems--combining multiple sensors, actuators and electronic control units--are beginning to require synchronization and performance past what the existing standard, Controller Area Network (CAN), can provide.

Coupled with growing bandwidth requirements with today's advanced vehicles utilize over five separate CAN busses, automotive engineers are demanding a next-generation, embedded network. After years of partnership with OEMs, tool suppliers, and end users, the FlexRay standard has emerged as the in-vehicle communications bus to meet these new challenges in the next generation of vehicles. Adoption of a new networking standard in complex embedded designs like automobiles takes time. While FlexRay will be solving current high-end and future mainstream in-vehicle network challenges, it will not displace the other two dominant in-vehicle standards, CAN, and LIN. In order to optimize cost and reduce transition challenges, the next generation of automobiles will contain FlexRay for high-end applications, CAN for mainstream powertrain communications and LIN for low-cost body electronics.

Understanding how FlexRay works is important to engineers across all aspects of the vehicle design and production process. This article will explain the core concepts.

Many aspects of FlexRay are designed to keep costs down while delivering top performance in a rugged environment.

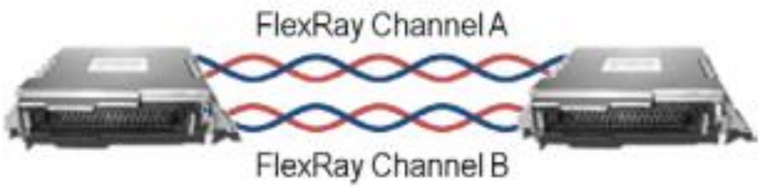


Fig.1. FlexRay system basic communication structure.

FlexRay uses unshielded twisted pair cabling to connect nodes together. FlexRay supports single- and dual-channel onfigurations which consist of one or two pairs of wires respectively. Differential signaling on each pair of wires reduces the effects of external noise on the network without expensive shielding. Most FlexRay nodes typically also have power and ground wires available to power transceivers and microprocessors.

Dual-channel (Fig.1) configurations offer enhanced fault-tolerance and/or increased bandwidth. Most first-generation FlexRay networks only utilize one channel to keep wiring costs down, but as applications increase in complexity and safety requirements, future networks will use both channels.

FlexRay buses require termination at the ends, in the form of a resistor connected between the pair of signal wires. Only the end nodes on a multi-drop bus need termination. Too much or too little termination can break a FlexRay network. While specific network implementations vary, typical FlexRay networks have a cabling impedance between 80 and 110 Ohms, and the end nodes are terminated to match this impedance. Termination is one of the most frequent causes of frustration when connecting a FlexRay node to a test setup. Modern PC-based FlexRay interfaces may contain on-board termination resistors to simplify wiring.

2.1 FlexRay Topology and Layout

One of the things that distinguishes FlexRay, CAN and LIN from more traditional networks such as ethernet is its topology, or network layout. FlexRay supports simple multi-drop passive connections as well as active star connections for more complex networks. Depending a vehicle's layout and level of FlexRay usage, selecting the right topology helps designers optimize cost, performance, and reliability for a given design.

Multi-drop Bus FlexRay is commonly used in a simple multi-drop bus topology that features a single network cable run that connects multiple ECUs together. This is the same topology used by CAN and LIN and is familiar to OEMs, making it a popular topology in first-generation FlexRay vehicles.

Each ECU can "branch" up to a small distance from the core "trunk" of the bus. The ends of the network have termination resistors installed that eliminate problems with signal reflections. Because FlexRay operates at high frequencies, up to 10 Mbit/s compared to CAN's 1 Mbit, FlexRay designers much take care to correctly terminate and lay out networks to avoid signal integrity problems.

The multi-drop format also fits nicely with vehicle harnesses that commonly share a similar type of layout, simplifying installation and reducing wiring throughout the vehicle.

2.2 Star Network

The FlexRay standard supports "Star" configurations which consist of individual links that connect to a central active node. This node is functionally similar to a hub found in PC ethernet networks. The active star configuration makes it possible to run FlexRay networks over longer distances or to segment the network in such a way that makes it more reliable should a portion of the network fail. If one of the branches of the star is cut or shorted, the other legs continuing functioning. Since long runs of wires tend to conduct more environmental noise such as electromagnetic emissions from large electric motors, using multiple legs reduces the amount of exposed wire for a segment and can help increase noise immunity.

2.3 Hybrid Network

The bus and star topologies can be combined to form a hybrid topology. Future FlexRay networks will likely consist of hybrid networks to take advantage of the ease-of-use and cost advantages of the bus topology while applying the performance and reliability of star networks where needed in a vehicle.

The FlexRay Protocol protocol is a unique time-triggered protocol that provides options for deterministic data that arrives in a predictable time frame (down to the microsecond) as well as CAN-like dynamic event-driven data to handle a large variety of frames. FlexRay accomplishes this hybrid of core static frames and dynamic frames with a pre-set communication cycle that provides a pre-defined space for static and dynamic data. This space is configured with the network by the network designer. While CAN nodes only needed to know the correct baud rate to communicate, nodes on a FlexRay network must know how all the pieces of the network are configured in order to communicate. As with any multi-drop bus, only one node can electrically write data to the bus at a time. If two nodes were to write at the same time, you end up with contention on the bus and data becomes corrupt.

There are a variety of schemes used to prevent contention on a bus. CAN, for example, used an arbitration scheme where nodes will yield to other nodes if they see a message with higher priority being sent on a bus. While flexible and easy to expand, this technique does not allow for very high data rates and cannot guarantee timely delivery of data. FlexRay manages multiple nodes with a Time Division Multiple Access or TDMA scheme. Every FlexRay node is synchronized to the same clock, and each nodes waits for its turn to write on the bus.

Because the timing is consistent in a TDMA scheme, FlexRay is able to guarantee determinism or the consistency of data deliver to nodes on the network. This provides many advantages for systems that depend on up-to-date data between nodes.

Embedded networks are different from PC-based networks in that they have a closed configuration and do not change once they are assembled in the production product. This eliminates the need for additional mechanisms to automatically discover and configure devices at run-time, much like a PC does when joining a new wired or wireless network. By designing network configurations ahead of time, network designers save significant cost and increase reliability of the network. For a TDMA network such as FlexRay to work correctly, all nodes must be configured correctly. The FlexRay standard is adaptable to many different types of networks and allows network designers to make tradeoffs between network update speeds, deterministic data volume, and dynamic data volume among other parameters. Every FlexRay network may be different, so each node must be programmed with correct network parameters before it can participate on the bus.

To facilitate maintaining network configurations between nodes, FlexRay committee standardized a format for the storage and transfer of these parameters in the engineering process. The Field Bus Exchange Format, or FIBEX file is an ASAM-defined standard that allows network designers, prototypers, validaters, and testers to easily share network parameters and quickly configure ECUs, test tools, hardware-in-the-loop simulation systems, and so on for easy access to the bus.

2.4 The Communication Cycle

The FlexRay (Fig.2) communication cycle is the fundamental element of the media-access scheme within FlexRay. The duration of a cycle is fixed when the network is designed, but is typically around 1-5 ms. There are four main parts to a communication cycle:

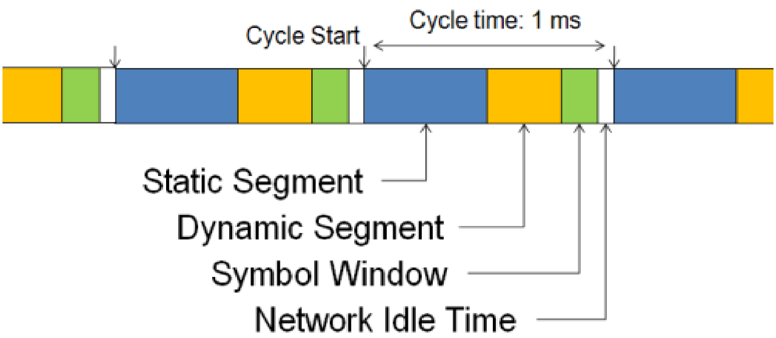


Fig.2. Communication cycle in FlexRay.

2.4.1 Static Segment

Actual FlexRay networks may contain up to several dozen static slots - reserved slots for deterministic data that arrives at a fixed period (Fig.3).

The static segment, represented as the blue portion of the frame, is the space in the cycle dedicated to scheduling a number of time-triggered frames. The segment is broken up into slots, each slot containing a reserved frame of data. When each slot occurs in time, the reserved ECU has the opportunity to transmit its data into that slot. Once that time passes, the ECU must wait until the next cycle to transmit its data in that slot.

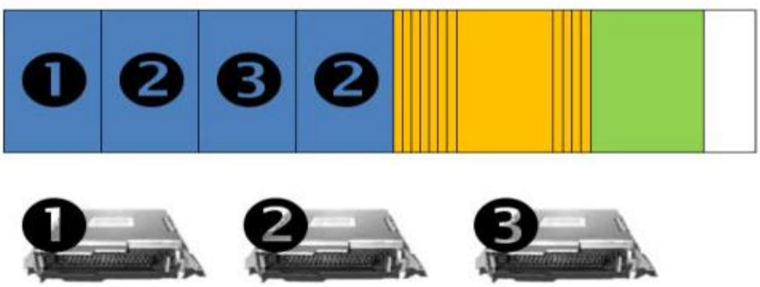


Fig.3. Illustration of a static segment with 3 ECUs transmitting data to 4 reserved slots.

2.4.2 Dynamic Segment

The dynamic segment behaves in a fashion similar to CAN and is used for a wider variety of event-based data that does not require determinism.

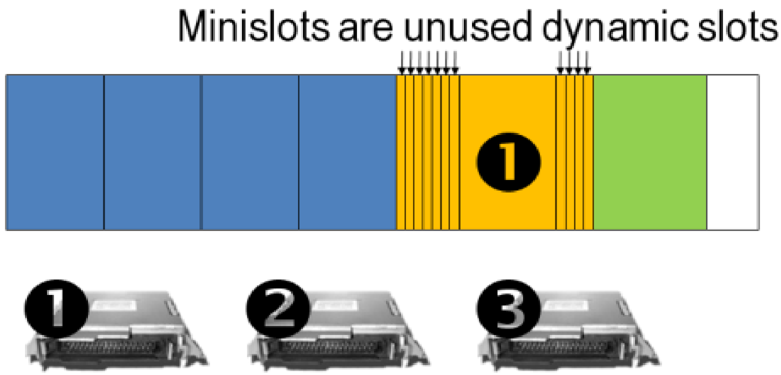


Fig.4. Illustration of FlexRay dynamic slots with one ECU broadcasting data.

Most embedded networks have a small number of high-speed messages and a large number of lower-speed, less-critical networks. To accommodate a wide variety of data without slowing down the FlexRay cycle with an excessive number of static slots, the dynamic segment allows occasionally transmitted data. The segment is a fixed length, so there is a limit of the fixed amount of data that can be placed in the dynamic segment per cycle. To prioritize the data, minislots are pre-assigned to each frame of data that is eligible for transmission in the dynamic segment.

2.4.3 Symbol Window

The Symbol window is primarily used for maintenance and identification of special cycles such as cold-start cycles. Most high-level applications do not interact with the symbol window. Typically used for network maintenance and signalling for starting the network.

2.4.4 Network Idle Time

A known "quiet" time used to maintain synchronization between node clocks.

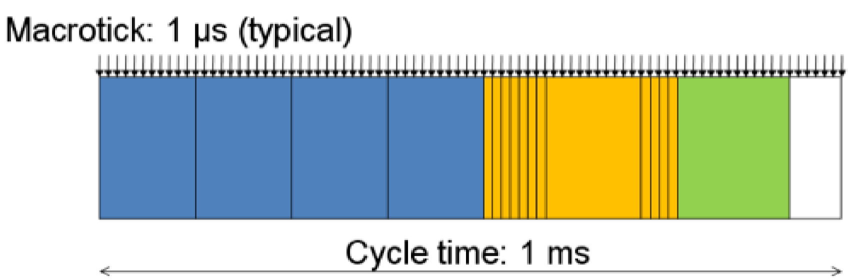


Fig 5. Detail of the FlexRay macrotick.

The smallest practical unit of time on a FlexRay network is a macrotick. FlexRay controllers actively synchronize themselves and adjust their local clocks so that the macrotick occurs at the same point in time on every node across the network. While configurable for a particular network, macroticks are often 1 µs long. Because the macrotick is synchronized, data that relies on it is also synchronized.

3 Modification and Improving Existing Bus Technology

The proposed modifications will lead to a revolutionary new bus access method which is suitable for control units and fit for future demands by eliminating the restrictions of existing bus technologies and simultaneous implying the benefits of the different access methods:

Modification of the physical layer

The aim is to slightly modify the physical properties. It would be perfect to find a way of using the complete physical properties of an existing bus technology, since this would mean to be able to keep all established communication modules, dirvers, wiring, impedance networks, architecture and monitoring. The scientific scope of this paper is to describe a modification, which causes minimal modification effort on established technology. This is the base for building of the bus infrastructure needed to develop the communication systems itself. It is necessary to introduce three logical levels to realize differential signals with two dominant levels and a recessive level for arbitration algorithms. The choice of levels themselves is not really easy: on the one hand a low differential voltage offer advantages in current consumption, when establishing a system with quite low impedance (around 120 Ohms). On the other hand the logical levels have to differ quite enough during arbitration in order to be assigned correctly, even in critical situations (more than one ECU is trying to manipulate the logical level at slightly different).

Time management

The time management is essential for a TDMA by topology. Therefore most of the scientific work will be in this domain. The demand is to create a way to literally avoid collisions on the data highway. It is comparable to trying to find a way of merge railway and motorway. The challenge is to put a rail on the same lane a car uses without delaying the time schedule. Scientific contribution is to provide ideas to future bus access methods. The ultimate goal is to keep all existing priority tables of arbitration systems and time tables of deterministic systems.

Applied algorithm with software modification

Modification of both physical layer and network timing properties logically leads to modifications of the data containers themselves. The aim is to adapt the data security, error and fault management to the changed environment. The scientific work is to restructure the frame format of each communication package: The header is responsible for both arbitration and determinism. The payload needs to be flexible and predictable. The trailer contains error and fault confinement. But in order to minimalize the influence of a new bus communication, the changes should not result in dramatic restructuring of the used software. It is the clear aim to carry over as much as possible of given software models. This leads to acceptance. There would be no benefit of introducing new features at the cost of complete remodelling of software.

Conclusion

The paper deals with design and modification of existing BUS technology for modern vehicles. The proposed control methodology for existing FlexRay BUS guaranteeing network stability and high performance. The scope of this paper is to presents a new methods and modification of physical layer, time management and control software in order to implement arbitration into a system. This means a technical adaption of new access methods on existing bus technology. The benefit is to provide hard- and software engineers with ideas for designing new bus structures, which are able to introduce.

Acknowledgement

The paper was supported by the Slovak Research and Development Agency under the project No. APVV-0772-12.

References

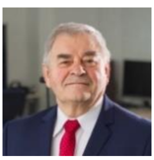
[1] Voss, W.: “Comprehensible Guide to Controller Area Network” Copperhill Media Corporation, 2005. ISBN 978-0976511601.

[2] Kurose, J. and Ross, K: “Computer Networking: A Top-Down Approach” 4the Edition, Addison Wesley, 2007. ISBN: 0321497708.

Authors

Florian Dietze
Porsche AG, Stuttgart - Zuffenhausen, Germany
Institute of Automotive Mechatronics
florian.dietze@outlook.de

M.Sc. Florian Dietze, PhD. is a young researcher working on advanced communication systems for automotive vehicles. He completed his PhD studies at the Institute of Automotive Technology, FEI STU in Bratislava in 2015 in Study Program Automation and Control. He is currently a researcher at Porsche AG, Stuttgart - Zuffenhausen, Germany.



prof. Ing. Štefan Kozák, PhD.
Faculty of Informatics,
Pan-European University in Bratislava, Slovakia
stefan.kozak@paneurouni.com

His research interests include system theory, linear and nonlinear control methods, numerical methods and software for modeling, control, signal processing, IoT, IIoT and embedded intelligent systems for digital factory in industry and medicine.

International Journal
Information Technology Applications (ITA)
Volume 12, Number 2, December 2023

STATE OF THE ART IN APPLICATION OF AI METHODS FOR TRANSPORT AND LOGISTICS

Štefan Kozák, Eugen Ružický, Juraj Štefanovič, Alena Kozáková

Abstract:

Research, development and application of intelligent methods and techniques represents one of the most dynamic factors influencing the quality and safety of processes in transport and logistics. Intelligent information, communication systems represent such complex systems that, using intelligent elements and systems, intelligent control methods and modern information, communication and control technologies using the Internet of Things (IoT), Industrial Internet of Things IIoT, virtual and mixed reality, can ensure and significantly improve the current functionality of transport and logistic systems and processes. This paper addresses the current issue of modeling and intelligent control of complex transportation systems and processes with the support of advanced integrated HW and SW technologies, artificial intelligence methods, adaptive and self-learning techniques capable effectively improve and optimise the conventional transport and logistic processes.

Keywords:

Artifical intelligence, fuzzy logic, transport system, logistics systems, neural networks, IoT - Internet of Things, mathematical model.

Introduction

Intelligent Transport Systems (ITS) are a group of technologies and methods that can nowadays significantly influence and optimise the control of complex transport systems and processes, public transport, as well as individual decisions regarding many aspects of goods transport and travel. Current ITS technologies include state-of-the-art wireless, electronic and automated technologies to improve safety, efficiency and passenger comfort. An Intelligent Transportation System (ITS) [5] is a complex integrated system composed of SW and HW subsystems aimed at transport to improve its safety, mobility and reliability. ITS encompasses a wide range of applications that process and share information in order to minimize, e.g., congestion, improve traffic management, minimize environmental impact, and enhance the benefits and advantages of transportation for both commercial users and the public [4]. ITS today uses advanced artificial intelligence methods in the field of information communication and control technologies such as Internet of Things (IoT) for interaction vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I), infrastructure-to-infrastructure (I2I) technologies implemented by wireless and wired communication, information and electronic technologies (Fig.1).

Intelligent Transportation System (ITS) is complex system which applies advanced technologies of electronics, informatics, communications, computers, artificial intelligence, control, sensing and detecting in all kinds of transportation system in order to improve safety, efficiency and service, traffic situation through transmitting real-time information.

IoT technology is mainly used to process vehicle information and transmit it to other vehicles, other modes of transport users (Fig.2), (e.g. pedestrians or cyclists), local and remote infrastructure. ITS methodology is continuously evolving and has a significant impact on transportation in a wide range of applications such as electronic toll collection, ramps, traffic light cameras, traffic signal coordination, transit signal priority, and traveler information systems. The quality and implementation of ITS is expected to increase especially in applications, which will use artificial intelligence methods for monitoring (image processing - machine learning, convolutional neural networks, deep learning), toll management, ticket management, transit pricing, telematics and traffic monitoring (Fig.2).

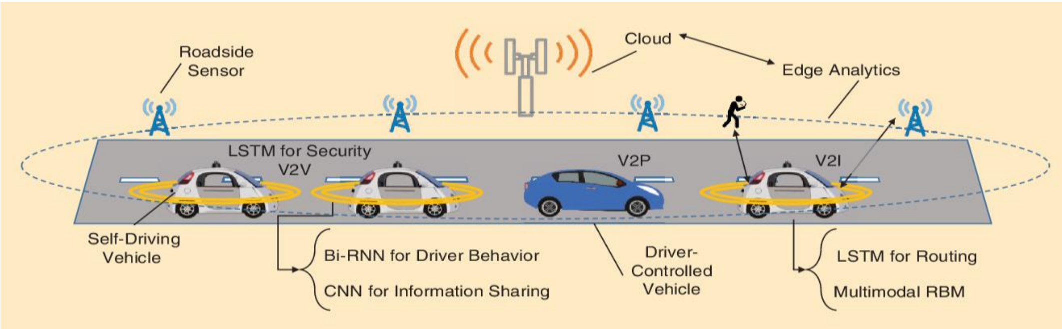


Fig.1. Intelligent Transport Systems with advanced technologies applications.

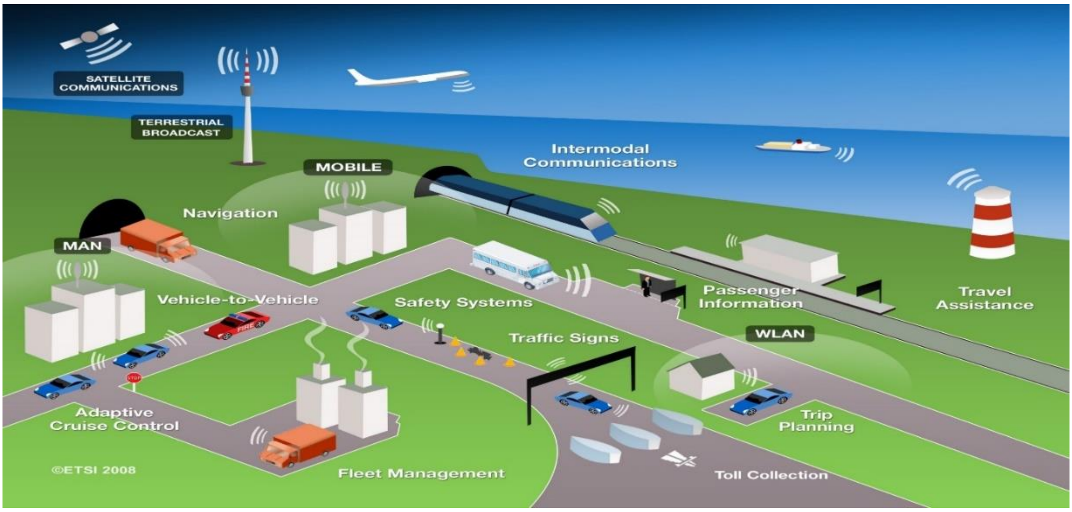


Fig.2. Application of Advanced Intelligent Information
and Communication Technologies to the transport.

The key beneficiaries of ITS use are passengers, businesses and transport agencies. The effective deployment of ITS is affected by the slow growth of global infrastructure and the high cost of implementation. Other issues that make ITS attractive to users are ensuring the security of highly advanced technologies and resistance to hacking and data protection.

Intelligent Transportation System (ITS) is aimed at achieving transport efficiency by minimising traffic problems and optimising the transport of people and goods. It provides users with the necessary comprehensive traffic information, linked to other local information in real time. The deployment of ITS can reduce commuters' travel time and increase their safety and comfort. Further development of ITS cannot be done nowadays without the deployment of IoT, Big Data, Clouds and Artificial Intelligence methods. The development and use of the new complex methodology for transport an logistics systems is closely related to the development and integration of the so-called exponential technologies (Fig.3).

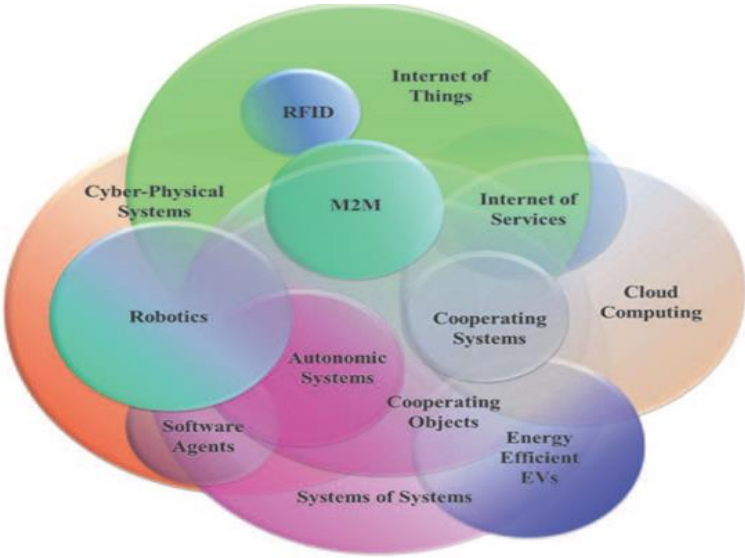


Fig.3. Convergence of Adavance Technologies for Transport and Logistics processes.

1 Artificial Intelligence in Transport and Logistics

Artificial Intelligence (AI) is a modern scientific field that deals with the research, development and application of methods and algorithms that enable intelligent behavior not only of humans but also of machines. It is a multidisciplinary science using the results of many other disciplines, especially computer science, mathematics, physics, automation, statistics, logic and neuroscience [2]. Research and applications of artificial intelligence (AI) methods have seen tremendous growth in the last decades in all areas of human activity. The use of AI methods in transportation [5] is largely based on pioneering research results from the 1960s and 1970s, in the areas of fuzzy logic, artificial neural networks, and genetic algorithms. In particular, biological systems, the behaviour of living organisms, the communication between them and, above all, the ability to learn, to react quickly and make optimal decisions in normal and critical situations have been the inspiration for the research, development and application of artificial intelligence methodologies [3].

One of the reasons for the use of artificial intelligence methods is their versatility and high ability to model, approximate and predict complex processes and processes in both technical and non-technical domains. The use of AI in transportation and logistics is significant. The use of conventional mathematical methods for modelling processes in transport is challenging and complex due to the non-linear descriptions and relationships of dynamic models.

Using mathematical models through AI methods and artificial neural networks, we can achieve much higher quality and accuracy of these "intelligent" models. The application of AI methods, which have the properties and capabilities of self-learning, adaptation and self-organization, allows us to efficiently design and apply AI models for the optimization of control and decision-making tasks in transportation and logistics (Fig.4). Artificial neural networks are nowadays mainly implemented through software modules or special hardware modules and special fast parallel computing systems and processors. These modules are the dominant representatives of the efficient use of AI algorithms and technical solutions related to direct deployment in practice.

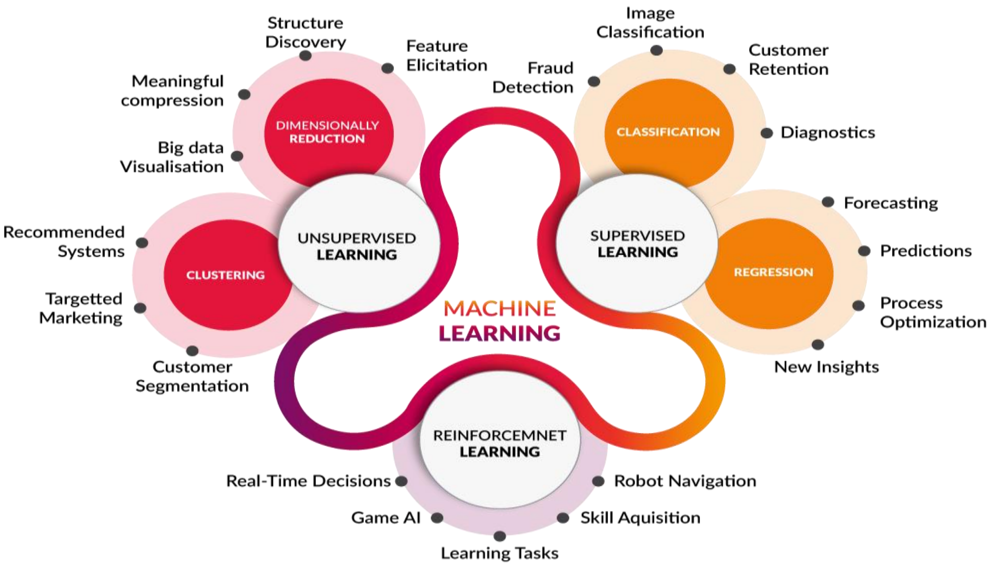


Fig.4. Development and applications of AI methods based on machine learning methods and algorithms.

Artificial neural networks are the basis for modeling of complex processes, representing a mathematical computational model constructed by abstracting the properties of biological neural systems. There are several types and possibilities of representing the behaviour of processes using artificial neural network models (multilayer perceptron, RBF networks, ART networks, Hopfield network, Kohonen network, etc.). The basic part of an artificial neural network is a model of a neuron with inputs and outputs. One of the most important properties of neural networks is the ability to abstract the rules of behavior between input and output values presented in a suitable form and then apply the obtained rules to any input values. The abstraction process represents the learning process. During the learning process, the values of the connection weights in the network are updated. The learning (training) process represents the modification of synaptic weights and thresholds according to the selected learning algorithm. The essence of learning is to select features and experiences from the input signals and adjust the network parameters so that the deviation (in a given metric) between the desired (reference) and actual output in response to the training patterns is minimal. Once learning is complete, the values of the weights no longer change and the network produces outputs according to the above rule applied to the input values. The neural network itself is composed of multiple layers with different numbers of neurons that are interconnected. There may still be hidden layers between the input and output layers, which is why these structures are called multilayer neural networks (Fig.5).

Another important feature of artificial neural networks is the principle of self-organization, which is used precisely when we do not know the target solution to the problem. Even in such a case, the artificial neural network is able to recognize the same or close features based on the input patterns to cluster them and sort them through clusters (clusters).

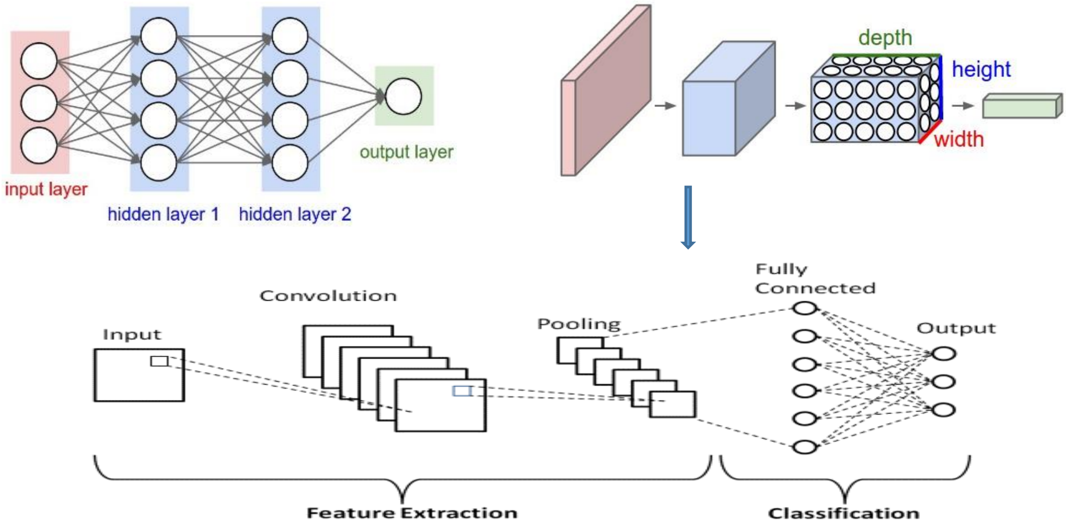


Fig.5. Conventional neural network (above left) and deep neural network structure.

Conventional machine-learning techniques were limited in their ability to process natural data in their raw form. discover the representation seeded for diagnosis, recognition and classification. Deep-learning methods are representation-learning methods with multiple levels of representation (Fig.5). Deep learning is making today many advances in solving complex problems modelling, control, visualization and prediction for large-scale tranports and logistics processes. The use of AI methods for a broad class of applications is shown in (Fig.6).

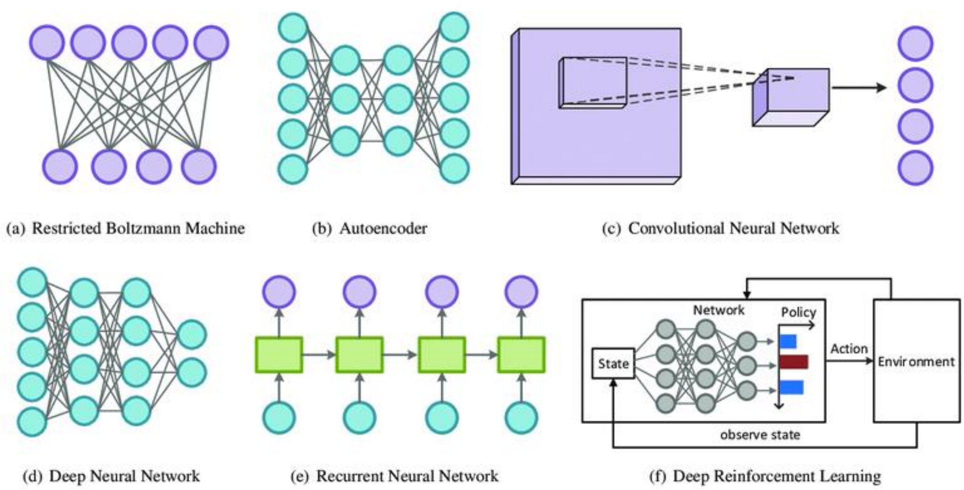


Fig.6. AI methods for a broad class of applications for transport and logistics processes.

The greatest today development of the application of artificial neural networks in transport and logistic is based on convolutional networks. They are particularly suitable and usable in the processing of image, audio and text information and directly applicable in transport. For image information processing, deep learning (DL) methods and algorithms used convolutional neural networks are the most powerful. Convolutional networks (Fig.7) are software networks that consist of layers of small computational units, called neurons, that hierarchically send each other the information to be processed. The system is conceptually similar to the biological system of neurons and synapses of the brain, hence its name. However, convolutional neurons are robust and are primarily used for tasks that involve processing image information. Other types of networks are, for example, recurrent (feedback)) neural networks, in which information does not pass from simpler to more advanced layers, and instead the information cycles dynamically between them. When convolutional neural networks are trained as part of a recognition process, they produce a representation whose complexity increases as part of a hierarchical process. Convolutional neural networks are a special kind of multilayer neural networks that are used for recognizing visual patterns directly from images using minimal preprocessing.

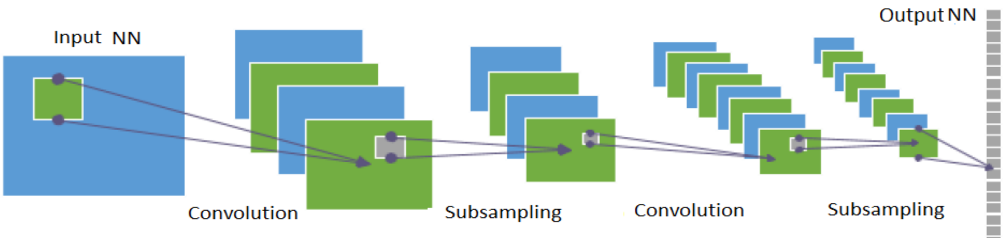


Fig.7. General convolution neural network architecture.

The first layer (Fig.7) is the so-called convolutional layer. The image is first fed to the input of the convolutional neural network, to which a filter is applied in order to extract the features. This filter is composed of neurons that share the same set of weight vectors between them, which reduces the number of network parameters and therefore the size of its capacity. Using these local receptive fields, the network can extract features such as edges or corners from images. Another advantage of convolutional neural networks is the sharing of weights, which speeds up the learning of the network by reducing the total number of network computations. This filter is referred to as a receptive field. This receptive field is successively applied to the entire input image, thus computing a single operation in different parts of the image. In most cases, several receptive fields are applied to a single input image, with the individual receptive fields differing in the values of the shared vector weights. The output of one receptive field that has been applied to the entire input image is called a flag map, and the number of flag maps corresponds to the number of receptive fields applied to the image.

2 Application of Artificial Intelligence in Transport and Logistics

The greatest development of the application of artificial neural networks is related to the research, development of the convolutional networks. They are particularly suitable and usable in the processing of image, audio and text information and directly applicable in transport. For image information processing, deep learning (DL) methods and algorithms based on convolutional neural networks are the most powerful. Convolutional networks are software networks that consist of layers of small computational units, called neurons, that hierarchically send each other the information to be processed. The system is conceptually similar to the biological system of neurons and synapses of the brain, hence its name.

However, convolutional neurons are robust and are primarily used for tasks that involve processing image information. Other types of networks are, for example, recurrent (feedback)) neural networks, in which information does not pass from simpler to more advanced layers, and instead the information cycles dynamically between them. When convolutional neural networks are trained as part of a recognition process, they produce a representation whose complexity increases as part of a hierarchical process. Convolutional neural networks are a special kind of multilayer neural networks that are used for recognizing visual patterns directly from images using minimal preprocessing.

Deep learning based AI systems are part of modern methods of so-called machine learning and AI routing and exploitation [1]. Deep neural network (DNN) architectures got their name because of the novel 3D artificial neural network architecture consisting of several levels. Meanwhile, each layer has its own information processing elements. The deep network methodology is characterized by the fact that it does not follow the set of the task rules, but the regularities and rules determine themselves based on the input data received, thus it does not determine exactly the procedure that the neural network must follow to achieve the correct result. Instead, the network is trained on a number of examples against which it evaluates its success rate. In this way, the neural network changes its internal parameters and tries to achieve the best possible results. It retains and further exploits the elements and rules that work for learning, and gradually discards those elements that lead to deteriorating results. The network is thus constantly evolving, self-organising and adapting. In case a symptom is identified using a receptive field, the location in which the symptom was identified becomes less relevant. Hence, there is an additional layer after the convolutional layer, which is referred to as the subsampling layer. The goal of this layer is to reduce the input size [3]. In practice, a function is applied to all the symptom maps that are the output of the convolutional layer to reduce these maps. There are several types of functions that can be applied in the subsampling layer such as averaging, maximum selection or linear combination of neurons in a given symptom map. (Fig.7) shows the application of convolutional networks for situation representation, traffic prediction and object identification in the traffic process. The tracked objects are continuously identified and transmitted to the computing center (Cloud) for optimization and prediction. The convolutional network and its software implementation algorithm allows to embed, for example, images of intersections, based on which the system determines what the traffic roughly looks like and then applies this knowledge to recognize people and cars even in new images that it has never seen before. It is not the case that someone has programmed the likeness of a person or vehicle into the neural network. In short, the system is given the information that there is a car in this particular image, not in the next, and yes in the third, and so on. The neural network tries to recognise certain specific elements and then verifies its methods itself. One of the current applications of artificial intelligence methods is autonomous vehicles.

Modern autonomous vehicles would not work today without the application of intelligent methods and the extensive use of machine learning methods. Today, only a small portion of conventional solutions are still used in autonomous vehicles to control the direction/movement of the vehicle. Autonomous vehicles use machine vision algorithms such as convolutional neural networks to recognize the road and obstacles and train cars to identify the road and react to hazards, (cars in other lanes and pedestrians). With training on the surroundings and these hazards, the cars can then move safely.

3 Applications of Artificial Intelligence Methods in Transport

Currently, there are several areas where neural networks have moved to a "superhuman" level of cognition. One important application area where artificial intelligence and its methods and principles can be effectively used in transportation and logistics systems. Artificial intelligence methods and algorithms in transportation can be used for a wide range of applications (Fig.8).

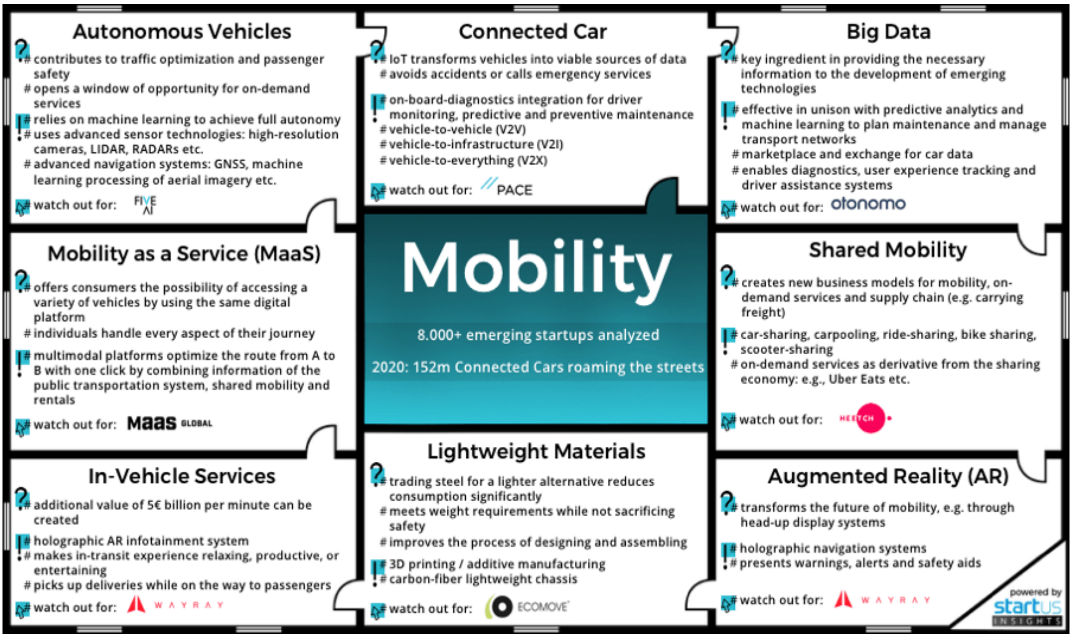


Fig.8. Applications of AI methods to the transport systems.

The advantages of artificial intelligence methods for process control in transport compared to conventional methods can be summarised as follows:

a. Conventional traffic control methods are largely based on mathematical models whose parameters and structures are difficult to identify based on known analytical and experimental methods.
b. Problems with mathematical traffic modelling arise especially when it is necessary to model system behaviour under conditions of uncertainty and contingency, which are often influenced by unexpected traffic phenomena and situations, human errors or accidents. In such cases, AI can help significantly.
c. AI, through its structures, uses sensed and observed data to correctly and optimally manage traffic processes and situations and even predict decisions based on learned and practiced situations.
d. Artificial neural networks and fuzzy-neuro systems are a versatile approximator for modeling complex traffic situations based on learning techniques allowing, together with efficient numerical algorithms based on genetic algorithms, to solve problems of unpredictability types. They are developed and implemented in different ways and their applications are conditioned by the use of powerful computing systems.

Recently, applications of AI are well known in the world, and in particular deep neural networks (DNN), which have demonstrated the ability to predict, e.g., traffic flow for traffic processes with large I/O datasets. While current existing DNN-based models provide much better results than conventional models, there is still the open question of fully exploiting the spatiotemporal characteristics of traffic flow to improve their performance. Significant achievements for HNS-based traffic flow prediction model are documented worldwide significantly affecting the improvement of prediction accuracy.

The HNS model fully exploits the weekly/daily periodicity and spatio-temporal characteristics of traffic flow. Inspired by recent work in the field of machine learning, a sensing and observation-based model was thus developed that automatically learns to determine the importance of past traffic flow.

A convolutional neural network was also used to extract spatial features and a recurrent neural network to catch up with the temporal features of the traffic flow. Through the visualization of traffic situations, it has been shown in practice how the HNS model accepts the traffic flow data and challenges conventional thinking about neural networks in the transportation domain, where neural networks are purely a "black box" model.

The scale of the data received in a traffic system and even the interaction of the different components of the system that generates the data have become a barrier to traditional data analytics solutions. Machine learning, on the other hand, is a form of artificial intelligence (AI) and a data-driven solution that can handle new system requirements. Machine learning learns latent patterns in historical data to model system behavior and to respond accordingly in order to automate analytical model building.

The use of AI methods has now gained significant use in the transportation industry and many applications have been shown to yield a higher return on investment compared to conventional approaches. The areas of application of these methods with effective implementation of AI methods are still open and require the training of new experts capable of configuring intelligent SW and HW modules. The main objectives of these solutions are to reduce congestion, improve safety and reduce human errors, mitigate adverse environmental impacts, optimize energy efficiency, and increase productivity and efficiency of surface transportation.

Potential areas of application of AI methods in transport include, but are not limited to:

- Monitoring, visualisation and performance management of the transport system.
- Management of autonomous vehicles and road traffic.
- Optimisation and operation of freight transport.
- Optimisation and management of road traffic.
- Predictive analysis and subsequent synthesis of intelligent public transport.
- Detection of critical situations and events from sensed image, audio and text information.
- Intelligent mobility services for transit planning, operations and reporting based on image and audiovisual data.
- Safety monitoring of vehicle movements.
- Passenger safety monitoring.
- Route safety monitoring and road damage analysis.
- Efficient journey planning and sharing of human-vehicle, vehicle-to-vehicle, vehicle-to-infrastructure communication.
- Traffic sign object recognition and detection.
- Passenger behaviour analysis.

4 Applications of Artificial Intelligence Methods in Logistics

Large logistics companies have long been using modern tools to organize and optimize transportation through intelligent neural networks complex architectures (Fig.9). As the volume of data grows, these companies are increasingly using and implementing AI (machine learning, deep learning and natural language audio and text processing) methods and algorithms. The use of AI brings efficiencies to the management and optimization of supply chains and logistics operations that already represent a significant competitive advantage.

The alignment of logistics with information technology is leading to the creation of 'cyber-physical systems'. Cyber-Physical Systems (CPS), which represent the intelligent interconnection of the real and virtual worlds.

Fig.9. Advanced intelligent architectures and AI methods for logistics.

Logistics is facing a fundamental change of approach with regard to the application of AI methods and algorithms, towards proactive, predictive, automated and individualised logistics operations in operational activities and customer contact. AI methods in logistics operations can be used for the application of image recognition methods (convolutional networks) in tracking the status of shipments and equipment, providing complete autonomous transportation, or predicting and minimizing fluctuations in total shipment quantities before they occur. The purpose of artificial intelligence methods and algorithms (Fig.9) in logistics applications is to enhance and develop human skills and also to eliminate routine work activities. Intelligent route planning systems represent a significant part of the use of AI in the transportation of goods and business transactions between customers.

Supply Chain applications for logistics can be built using machine learning algorithms with better predictive analytics tools. These prediction tools can accurately determine the average waiting time by just entering the location and day, it can also provide better predictions of appointment times and dynamic Estimated Time for Arrivals. Many Logistics & Supply Chain companies are utilizing **Chatbots** to carry out automated registration processes along with solving customer queries and feedback. AI-based chatbots can help in two essential ways:

- improve customer-based operations
- improve Supply Chain operations
- customers can interact with Chatbots for placing any sort of orders. The Chatbot will handle significant order details like pickup and delivery location, dates, and rates. It can even send receipts directly to the client’s email in an automated way.

Self-Driving Cars in transport and Logistics

- The idea of having self-driving car systems in logistics is going around for some time now. This technology is considered to be much more safer and reliable as it outperforms human driving capabilities. With the development of modern intelligent sensor systems and 3D Picture formulation visualization technology, self-driving cars are just 3-4 years ahead of us. Intelligent cars can recognize traffic signals and laws, avoid barriers, interpret road signs and many other things by applying advanced AI algorithms to sense and forecast changes in their surroundings.

Predictive analytics for logistics - with the implementation of Big Data, logistics companies are able to make accurate predictions and signicantly improve performance. AI-based nethods and algorithms can enhance predictive analytics and improve automation to achieve maximum efficiency.

Automated Robots for logistics - artificial intelligence in logistics can be effectively used for smart management of operations carried out in warehouses. Robots in process logistics (Fig.9) can be put in place, where they can identify, sort, move and track inventories, without the need for any human interference.

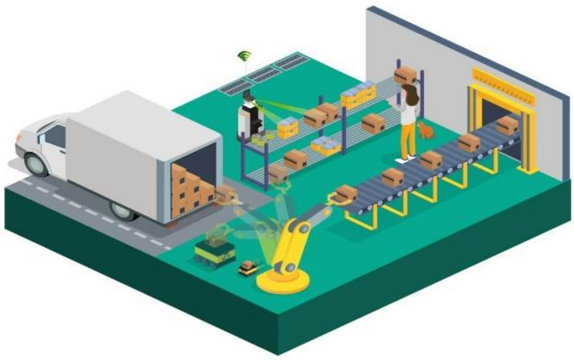


Fig.10. AI in logistics warehouses applications.

Logistics for 4.0/5.0 generation and advancing digitalization offer development opportunities through AI methods and algorithms in ever-growing and interconnected markets. Companies operating in the logistics industry and using AI, have a high probability of successfully advancing and creating modern logistics with autonomously managed multidimensional systems on a national and transnational scale. It shows that logistics companies with a proactive AI strategy in the transport and logistics sector have profit margins higher than conventional companies by more than 5%. On the other hand, competitors that do not use intelligent optimization methods see their margins decline. However, these key benefits have not led to widespread implementation of AI methods. It is estimated that about 21% of transportation and logistics firms have moved beyond the initial testing phase and have begun to deploy AI-based solutions. The introduction and acceptance of AI methodologies in supply chain and logistics requires large capital investments and organisational changes. Therefore, it can be expected that in the future, AI methods and algorithms will be implemented primarily by the "biggest players" in the logistics industry. And it is the speed of adaptation to new intelligent solutions that will determine a company's competitiveness in the future.

Conclusion

The use of artificial intelligence methods has now gained significant application in the transportation industry. In many applications, it has been shown to deliver higher quality and return on investment compared to conventional approaches. The areas of application of these AI methods with effective implementation of AI algorithms are still open and require the training of new experts capable of configuring and integration of intelligent SW and HW modules. The main objectives of deploying these intelligent solutions are to reduce congestion, improve safety and reduce human errors, mitigate adverse environmental impacts, optimize energy efficiency, and increase transport productivity and efficiency.

Acknowledgement

The paper was written within the research project MO SR SEMOD-80-7/2019 and project VEGA 1/0819/17.

References

[1] Hinton G.E., Tutorial on Deep Belief Networks, Machine Learning Summer School, Cambridge, 2009

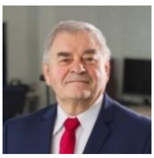
[2] Kozák, Š.: Advanced control engineering methods in modern technological applications, Published 2012 in IEEE Xplore Proceedings of the 13th International Carpathian Conference Process control, 2012.

[3] Krizhevsky, A., Sutskever, I. & Hinton, G.: ImageNet classification with deep convolutional neural networks. In Proc. Advances in Neural Information Processing Systems 25, 1090 - 1098 (2012)

[4] Bezák, B., Balkó, G.: Traffic space as a sensitive public space in historic city centers. In: Polomová, B. Bardkontakt 2014: issues of urban heritage centres. Transport in historic centres. Proceedings of the scientific conference Bardejov 26-27th August 2014. 1st ed. Bardejov : City of Bardejov, 2014, ISBN 978-80-971733-8-8.

[5] Kozák, Š.: XX. Súčasné trendy využívania metód umelej inteligencie v doprave a logistike. (Current trends in the use of artificial intelligence methods in transport and logistics). Event: Dopravno-inžinierske dni (Traffic-engineer days), Czech republic, Mikulov 2019.

Authors



prof. Ing. Štefan Kozák, PhD.
Faculty of Informatics,
Pan-European University in Bratislava, Slovakia
stefan.kozak@paneurouni.com
His research interests include system theory, linear and nonlinear control methods, numerical methods and software for modeling, control, signal processing, IoT, IIoT and embedded intelligent systems for digital factory in industry and medicine.

Assoc. Prof. RNDr. Eugen Ružický, PhD.
Faculty of Informatics, Pan-European University, Bratislava, Slovakia
eugen.ruzicky@paneurouni.com
His research interests include applied informatics, virtual reality, modelling, visualisation and applications in medicine.

Ing. Juraj Štefanovič, PhD.
Faculty of Informatics, Pan-European University in Bratislava, Slovakia
juraj.stefanovic@paneurouni.com
His research interests include modeling and simulation of discrete systems in technology applications.

prof. Ing. Alena Kozáková, PhD.
Institute of Automotive Mechatronics,
Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology in Bratislava, Slovakia
alena.kozakova@stuba.sk
From 1985 to 2013 she was with the Department of Automatic Control Systems (later Institute of Control and Industrial Informatics). From 2018 she is a full professor at the Institute of Automotive Mechatronics, Faculty of Electrical Engineering and Information Technology, STU in Bratislava. Her research interests include robust, optimal and decentralized control.

International Journal
Information Technology Applications (ITA)

Volume 12, Number 2, December 2023

AUTOMATION – INFORMATICS – COMMUNICATION – ARTIFICIAL INTELLIGENCE, PAST, PRESENT AND FUTURE

Pavol Frešo, Štefan Kozák, Juraj Štefanovič

Abstract:

The paper presents state-of-the-art of the research, development of advanced control methods, control structures, information a communications technologies AI and their applications to the different types of industrial processes, health care, services. In this paper we analyse methodology and basic common principle of four close field scientific discipline automation, information, communications technologies, and mechatronics. Automation, IKT, AI and Mechatronics is becoming an increasingly important discipline in today's digital society. Automation technology is understood to be the use of such methods, control strategies, processes, and installations (hardware and software) which can fulfil defined objectives without the constant interference of man in a largely independent manner, i. e. automatically. Mechatronics integrates the fields of mechanical, electrical, control, and computer engineering. This concentration was created because knowledge across these disciplines is essential to improve and/or optimize the functionality of modern engineering systems. Motivated by the practical success of control engineering methods in consumer mechatronics products and industrial process control, there has been an increasing amount of work on development of new methods which are based on new robust, adaptive, effective numerical optimization techniques, soft computing strategies, and hardware realization of control algorithms using embedded controllers and FPGA circuits for fast dynamic processes.

Keywords:

Automation, information technology, communication technology, control engineering, PID controller, MPC controller, fuzzy sets, neural networks.

Introduction

Many industrial processes and consumer electronic products have incorporated advance control methods, microprocessors, programmable logic controllers and computers to enable and embed intelligence and more functionality in these systems. Mechatronics integrates the fields of mechanical, electrical, control, and computer engineering. This concentration was created because knowledge across these disciplines is essential to improve and/or optimize the functionality of modern engineering systems. The principal characteristics of the following dominant branches:

- Mechatronics is the engineering discipline of integrating technologies from mechanical engineering, electronics, and computing to create more intelligent devices and machines. (Fig.1a,b)
- Informatics is the science of processing data for storage and retrieval. The successful design of complex systems is highly dependent on how design information is represented, managed and retrieved.

- Automation - Control engineering concerns the design of process controllers, based on an understanding of dynamic characteristics, so that the process will behave in a desirable way.

Traditionally, mechatronics has been applied to manufacturing and other industrial automation: robotic automation found in car automated production lines, such as welding and assembly lines in computer-integrated manufacture. These mechatronic applications have been extended from industrial systems to domestic products. New products have been designed applying mechatronic principles and increasingly consumers and society have benefited tremendously from these new intelligent products, including the latest mobile phones with mechatronic features, intelligent robotic vacuum cleaners and intelligent wheelchairs. Mechatronics has contributed to progress in many industrial fields such as robotics, semiconductors, aerospace, automotive, consumer electronics, and medical. Well-known and well-established mechatronic systems include production systems, synergy drives, automated guided vehicles, automotive subsystems such as antilock braking systems, and commonly used spin-assist consumer products such as auto-focus cameras, hard disk drives, compact disc players, and washing machines. The main benefits that mechatronics has provided are an increased functionality and comfort level, energy savings, versatility, and flexibility.

Research in automation and mechatronics focuses on the fusion of mechanical and electrical disciplines in modern engineering processes, aimed at achieving a cost- effective, optimal balance between mechanical structure and their overall control (Fig.2). Research in the Automation and Mechatronics area varies from fundamental research in control engineering theory to the conception, design, and prototype evaluation of innovative mechatronic systems and applications to automation.

Research topics in this area include active and passive damping, adaptive learning and predictive, optimal and robust control of systems with uncertainty, automated manufacturing and remanufacturing, fuzzy and neural networks for control and identification, precision engineering and motion control, multimedia technology, intelligent sensors and acutators, vision-based motion control and teleoperation.

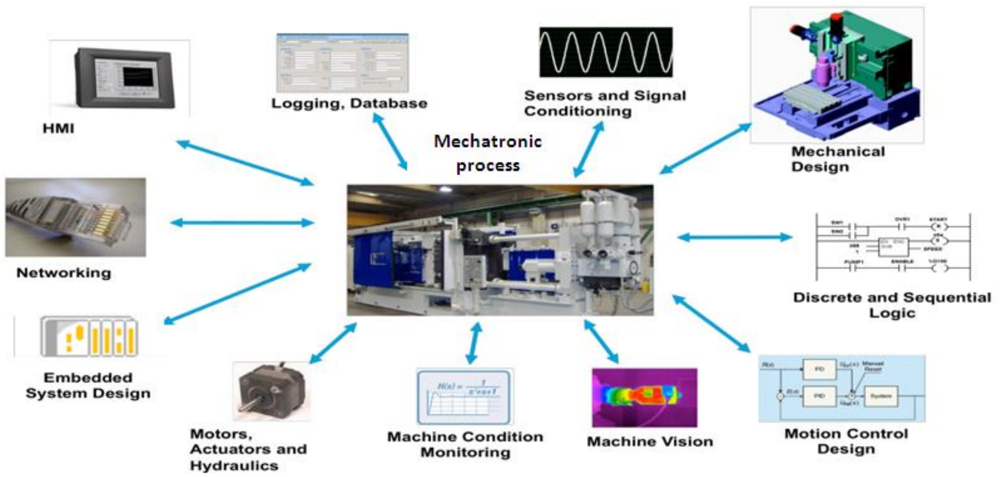


Fig.1a. Synergic connection of some important fields in effective complex mechatronics system design.

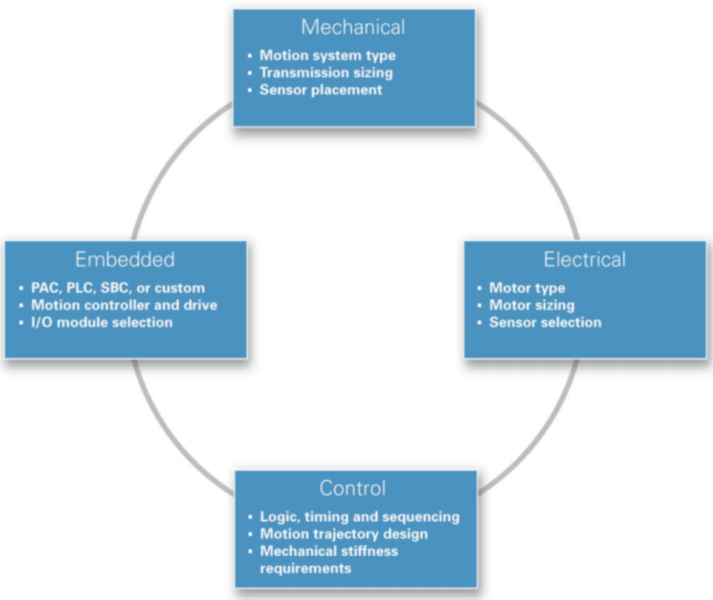


Fig.1b. Synergic connection of some important fields
in effective complex mechatronics system design.

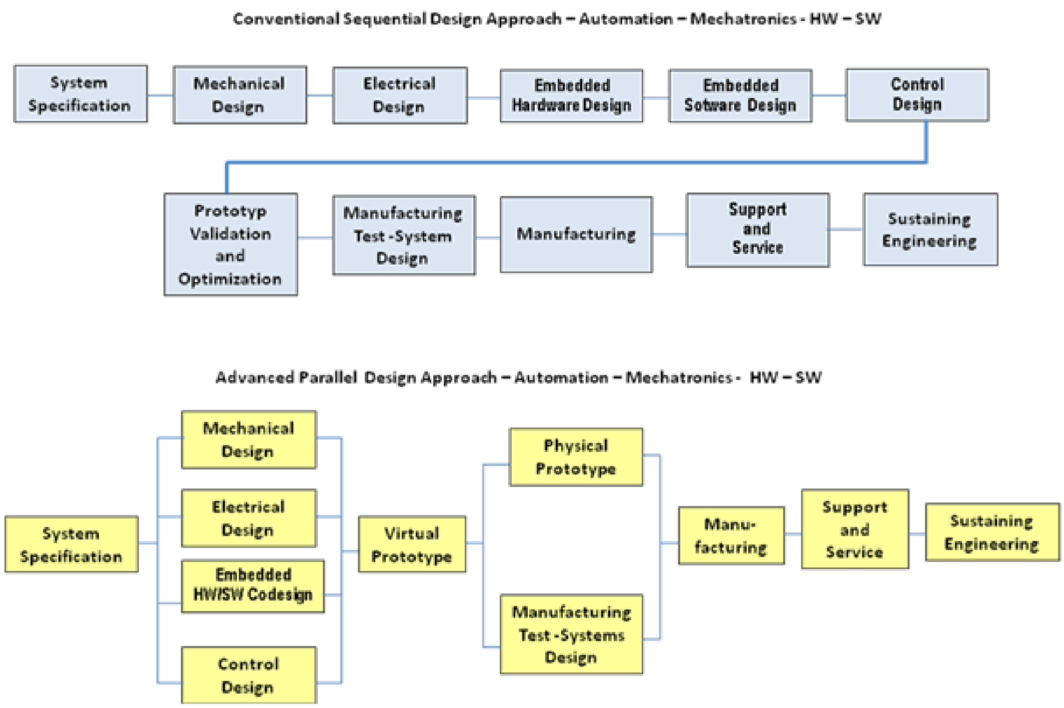


Fig.2. Conventional and advanced structures of mechatronic systems development.

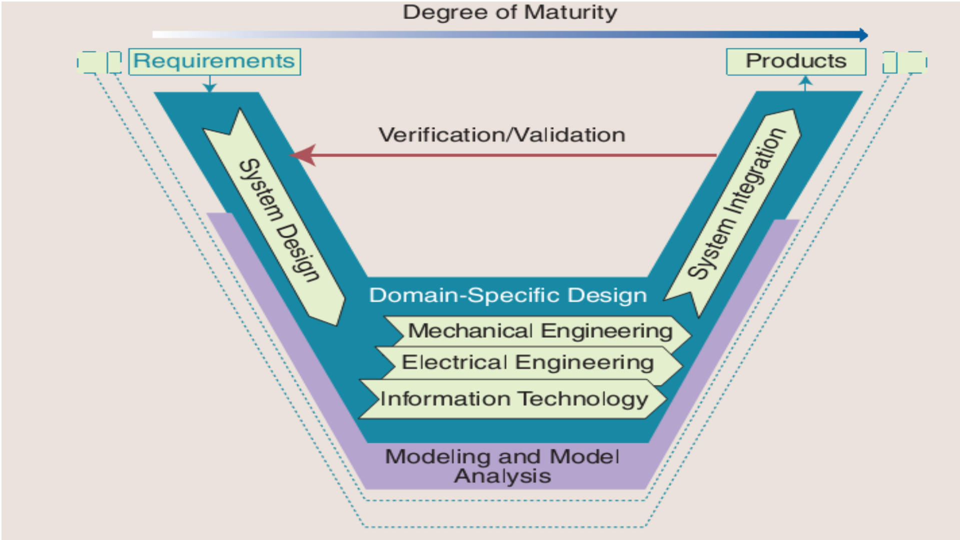


Fig.3. V-model scheme for development of mechatronic systems.

1 Advances in Automation and Mechatronics

Mechatronics systemic progress and several advanced software design tools are required during the design of mechatronic systems. The mechatronic design is an iterative and integrated process that includes different kinds of the domain-specific engineering (e.g., mechanical, electrical, electronic, information, automation, and multidisciplinary) for successful design, implementation and inspecting. The design step is the starting and most important procedure, and for the design aspects, system of objectives, applications, requirements, functions, active structure and shape and behavior should be considered. The implementation and inspecting step include the distribution of interdisciplinary task, the use of sensors and actuators, the electronic architecture, the software architecture, the different controller design (PID, LQ, MPC, etc) and system validation resulting in totally desired functions. The development scheme is represented in the form of a V-model, which distinguishes between the mechatronics system design and integration, as shown in (Fig.2) and (Fig.3). A control system and control engineering methods are at the heart of mechatronic systems where electronics are used to control mechanical systems. Control systems research has a long history of mathematical rigor, with application to diverse branches of science and engineering. The control methods, algorithms, and tools developed by control researchers have been widely used by generations of engineers to solve problems of practical importance with enormous impact on society. Control concepts have been crucial in the design and development of high-performance mechatronics systems (airplanes, fuel-efficient automobiles, industrial process plants, manufacturing enterprises, smart phones, planetary rovers, communication networks) and many other applications across various sectors of industry. In these and other complex engineering systems, control theory and its technological artifacts are also widely used to ensure reliable, efficient, and cost-effective operations.

As automation and intelligence are essential for mechatronic systems, the importance of sensors of mechatronic systems to meet the needs has grown steadily. An intelligent mechatronic system should be supported by various sensing devices. Various sensors (e.g., potentiometers, encoders, proximate switch, tachometers, acceleration sensors, and gyro sensors) have been used in mechatronic systems such as a robot system, manufacturing system, automotive vehicle system and aircraft vehicle system.

Because of the advances of manufacturing, sensor technologies, and micro/nanotechnologies, more compact (micro/nanosize) and highly integrated mechatronic systems have been recently created. Since redundant sensory information is installed/designed in systems, doubt never arises about implementing multisensor fusion methods into smart mechatronic systems and hence results in more intelligent performances. The challenges and perspectives of advanced mechatronics are summarized as the key directions in research and applications in automation and mechatronics:

- intelligent mechatronics, vehicles, robotics, biomimetics, automation and control systems,
- opto-electronic elements and materials, laser technology and laser processing,
- elements, structures, mechanisms and applications of micro and nano technologies,
- teleoperation, telerobotics, haptics, and teleoperated semi-autonomous systems,
- sensor design, multi-sensor data fusion algorithms and wireless sensor networks,
- biomedical and rehabilitation engineering, prosthetics and artificial organs,
- control system modeling and simulation techniques and methodologies,
- AI, intelligent control, neuro-control, fuzzy control and their applications,
- industrial automation, process control, manufacturing process and automation.

The challenges and perspectives of mechatronics are summarized as shown in (Tab.1).

Conclusion

The need for increased performance in efficiency, productivity, and flexibility through automation is more than obvious. Automation, AI and Mechatronics has already improved existing products and developed new ones with better performance in other consumer and industrial areas and applications.

Automation, AI and robotics remain the key sciences of the 21st century because they bring forth all the comforts in life and solve many inconveniences. The challenges and future perspectives for the forthcoming applications and requests are described in this paper. Intelligent control engineering methods and structures have a wide spectrum and extend to other engineering sciences, such as the application of mechatronics, from the traditional to the high technical areas, such as car industry, medical areas, multisensor fusion, and micro/nano techniques application.

References

[1] R. Isermann: Mechatronic systems - Innovative products with embedded control, Control Eng. Practice, vol. 16, no. 1, pp. 14 - 29, Jan. 2008.

[2] R. C. Luo, C. C. Lai, and C. C. Hsiao: Enriched indoor environment map building using multisensor based fusion approach, inProc. IEEE Int. Conf. Intelligent Robots and Systems (IROS), 2010, pp. 2509 - 2064.

[3] S. Hutchinson, G. D. Hager, and P. I. Corke: A tutorial on visual servo control, IEEE Trans. Robot. Automat., vol. 12, no. 5, pp. 651 - 670, Oct. 1996.

[4] M. K. Habib: Mechatronics engineering the evolution, the needs and the challenges, in Proc. IECON 2006, pp. 4510 - 4515.

[5] R. C. Luo and C. C. Chang: Multisensor fusion and integration aspects to mechatronics, IEEE Ind. Electron. Mag., vol. 4, no. 2, pp. 20 - 27, June 2010.

[6] Š. Kozák: Advanced Control Engineering Methods in Modern Technological Applications. In: ICCC 2012 : Proceedings of 13th International Carpathian Control 2012, ISBN 978-1-4577-1866-3, pp. 392 - 397.

Authors

Ing. Pavol Frešo
pavol.freso@paneurouni.com
Ing. Pavol Frešo is one of the leading experienced workers dealing with information technologies and their applications in practice. His research activities are currently directed to research in the field of database systems and programming paradigms. In the field of pedagogy, he teaches two courses at FI PEVS in Bratislava.

prof. Ing. Štefan Kozák, PhD.
Faculty of Informatics, Pan-European University in Bratislava, Slovakia
stefan.kozak@paneurouni.com
His research interests include system theory, linear and nonlinear control methods, numerical methods and software for modeling, control, signal processing, IoT, IIoT and embedded intelligent systems for digital factory in industry and medicine.

Ing. Juraj Štefanovič, PhD.
Faculty of Informatics, Pan-European University in Bratislava, Slovakia
juraj.stefanovic@paneurouni.com
His research interests include modeling and simulation of discrete systems in technology applications.

International Journal
Information Technology Applications (ITA)
Volume 12, Number 2, December 2023

COMPARISON OF DOCUMENT SIMILARITIES USING FREQUENCY METHODS

Ján Cigánek, Filip Žemla

Abstract:

This paper analyses the current state of plagiarism. As a result of this work, the proposed tool is implemented and the accuracy of the solution is verified. The analysis and design address preprocessing techniques, file conversion to different formats, algorithms aimed at comparing document similarity and graphical techniques to represent these similarities. Article focuses in particular on frequency methods and graphical representation called Tag Cloud.

Keywords:

Plagiarism, detection, word frequency, tag cloud.

Introduction

Nowadays, plagiarism can be explained as the theft of thoughts and ideas, which the plagiarist later presents as his own. Plagiarism is divided into intentional and unintentional. However, both possibilities are unacceptable in an academic environment. The biggest threat that plagiarism poses to us is the fact that it is no longer just a last resort for students, but is also a common practice in colleges and universities. In this way, it creates a problem that serves to increase the disincentive to further study. It is almost impossible to accurately detect plagiarism when there are many students in particular subjects in schools. Therefore, it was decided to develop a plagiarism comparison tool to compare similarities in text documents. Existing methods will be compared and their suitability for use will be evaluated. Systems used by the public for similarity detection will be used and their effectiveness when used as a basis for the proposed implementation will be assessed. There are many anti-plagiarism tools, some of which are described in detail in Section 1, but they are mostly too difficult to be used in a standard classroom setting. Our target is to develop a simplified tool that would be challenging to fool. Thus, such a tool should not require too much time to refract, which would disqualify it against regular study of the topic under consideration. The tool should be used for the purpose of detecting plagiarism in subjects taught at Faculty of Electrical Engineering and Information Technology, Slovak University of Technology in Bratislava (FEI STU).

1 Analysis

Within our research we analyzed 2 existing applications used for document authentication at FEI STU. **PlaDeS** is a freely available application for detecting plagiarism in text documents created by students of FEI STU [1]. (Fig.1) shows the basic architecture of PlaDeS tool. It uses the 3-gram method for benchmarking. Acceleration is achieved by using parallelism to process pairs in separate threads and setting the similarity threshold above 0% [1].

The application supports .doc/docx, .pdf and .txt files. First, it analyses different files and preprocesses them by lemmatization and removing stop words. Subsequently, it uses the technique of n-grams, LSC, TF-IDF and metadata. The graphic output is a graph of dependencies. It filters similarity results and exports them to .pdf, .csv formats. This tool gives the possibility to save the project and display the statistics of document similarity.

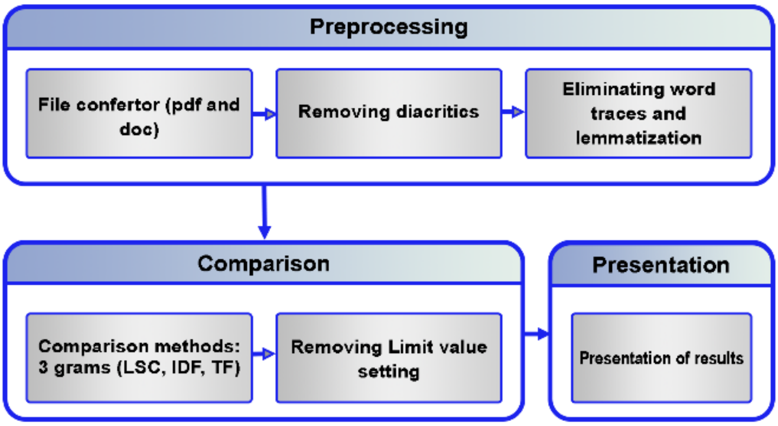


Fig.1. Architecture of PlaDeS.

SimPaD is based on the similarity of words in documents on the web. SimPaD takes into account not only substitution, addition and deletion, but also splitting and joining of sentences based on word similarity. It is used for English texts.

Short sentences and words are removed to achieve higher efficiency. Sentences that have less than 12 words before preprocessing are not taken into account. SimPaD also uses n-gram methods [2].

After a more detailed examination of various methods, metrics and methods for comparing similarities in text documents, we have worked out a possible suitable solution.

Since we want to compare texts in the Slovak language, and SimPaD detection is created for the primary purpose of finding similarities in another language, the basis for implementing the solution is the PlaDeS application.

The implemented method is word frequency. PlaDes application, which uses n-grams, will be used to compare the results. For visualization, the method of tag clouds and verbal notation of percentage similarity are tried [2, 3].

A. Methods for detecting plagiarism

The process by which we achieve the desired result, in our case finding the similarity of text documents, universally consists of three basic parts. These phases are not final and each uses different methods. Their sequence can be seen in Fig.2 [3].

The preprocessing phase includes the acquisition of text from different file formats, their conversion to the same form, and various operations with the given text documents for subsequent comparison purposes. There are several preprocessing factors, including the number of documents being processed simultaneously, the effort to understand the text, and the use of linguistic rules [4].

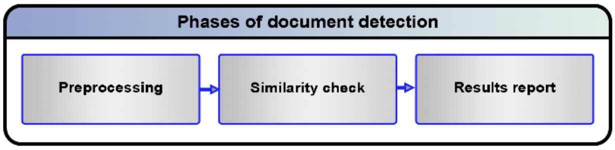


Fig.2. Phases of document similarity detection and their sequence.

The most common ways for preprocessing the text are the removal of stop words, tokenization, lemmatization, stemmatization, normalization, and synonyms replacement. Some of these methods require external resources, such as a dictionary of the relevant language for lemmatization.

The goal of lemmatization is to convert words into their basic form. We denote this shape by the term lemma, which means the "dictionary" shape of the token. It doesn't work for words that aren't in the dictionary, such as names and new colloquialisms. For lemmatization, we will use a database of words in a non-basic form, where a lemma is assigned to each word.

Stemmatization generalizes, shortens the text. Its basic task is the elimination of suffixes and prefixes. It converts words to their root, stem. In this modification, stemmatization does not consider the meaning of the word, nor the context of the text, and therefore it is not certain whether it will find the real root of the word.

By normalization, we check the replacement by hyperonyms and the replacement by synonyms, as the name suggests, we check the replacement by synonyms. Hyperonyms are words superior in meaning [4, 5].

The similarity check on already pre-processed texts uses the following techniques [4, 5]:

- n-gram technique,
- searching the longest common substring,
- word frequency technique,
- technique of inverse word frequencies,
- latent semantic analysis.

Searching the longest common subsequence (**Longest Common Subsequence**) - with this method, the longest common sequence of two documents is determined. Its advantage is a simple principle. However, it is slow, and for documents of more than ten pages, it starts to lose its functionality.

LSA (**Latent Semantic Analysis**) creates sets based on the contextual meaning of words in a set of documents. It assumes that words that are contextually close to each other will be found in the same parts of the text. It can use a matrix that records the occurrence of words in a document. The rows of the matrix represent terms and the columns documents.

Word frequency technique compares word histograms to produce hashmaps. The probability of similarity is calculated according to the intersection with the specified tolerance of the number of occurrences of words.

The term "**n-gram**" defines a sequence of n consecutive items from a given sequence. The model calculates the probability of occurrence of the last word of the n-gram from the previous n-grams. The method is devoted to the detection of overlapping parts of the document and the elimination of the problem of text displacement. The displacement of the internal part of the text is caused by sentence modification, which is the most common technique for masking the copied text and in most cases prevents the exact comparison of strings.

Part of the similarity check in this analysis is also the evaluation of the percentage similarity. That is, whether the given document is plagiarized or not. The best way is to visualize the results and leave the final decision to the individual user [4, 5, 6].

So, the last phase of similarity detection is the presentation of the results. The calculation of similarity proceeds according to the used comparison method: the number of common tokens, the most common n-grams, and a combination of several techniques.

Existing visualization methods are [3, 4]:

- graph of the connection of similar documents,
- pie chart of connection,
- color histogram,
- arc diagram,
- side-by-side,
- tag clouds,
- word spectrum diagram.

In our proposal, a decision is made between using side-by-side, which means color-coded similar blocks side by side, or tag clouds. We can see these two methods shown in (Fig.3) and (Fig.4). Tag clouds contain a list of words in a given document, and the words have a defined size based on their occurrence in the document. Words that appear in both documents are placed in the center or at the edge if they appear in only one document.

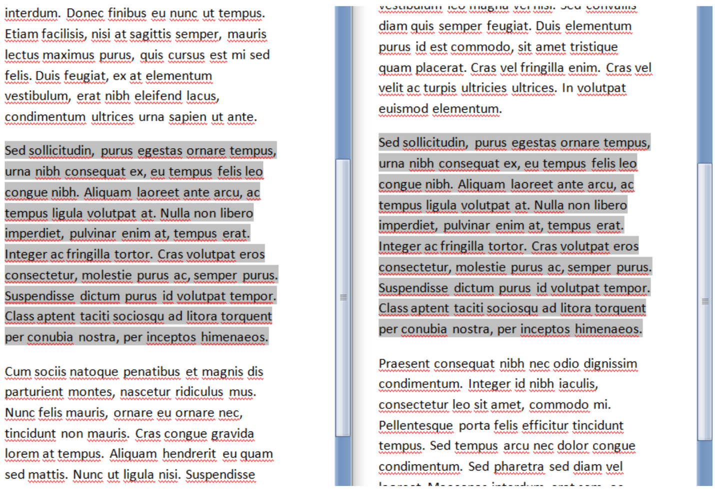


Fig.3. Visualization of similarity results using the side-by-side method.

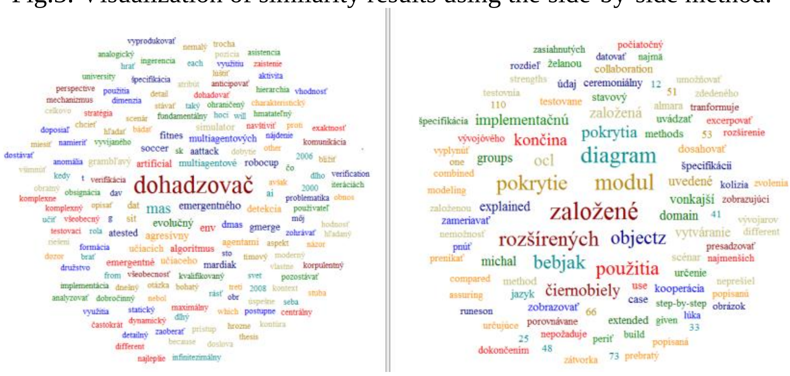


Fig.4. Visualization of similarity results using the tag clouds method.

For the design and subsequent implementation of the prototype, it is first necessary to specify the functional and non-functional requirements. Functionally, they speak directly about the functionality of the system, what it should do. Non-functional requirements are those that do not relate to what the system is supposed to do. They typically talk about system parameters.

Functional requirements for the system are the following.

- Retrieving text and loading it into the system is possible from different types of files.
- Running the loaded texts through tokenization and filtering.
- Individual words of the read text are lemmatized.
- Individual words of the loaded text are synonymized.
- Creation of a corpus of documents.
- The system will allow the user to compare documents.
- The user has the option to choose between one or more comparison methods.
- The similarity search algorithm does not have a long lifetime.
- Visualization of the resulting word vectors using the tag clouds or side-by-side method.

We added the use case of the system to the functional requirements. We can see it in (Fig.5).

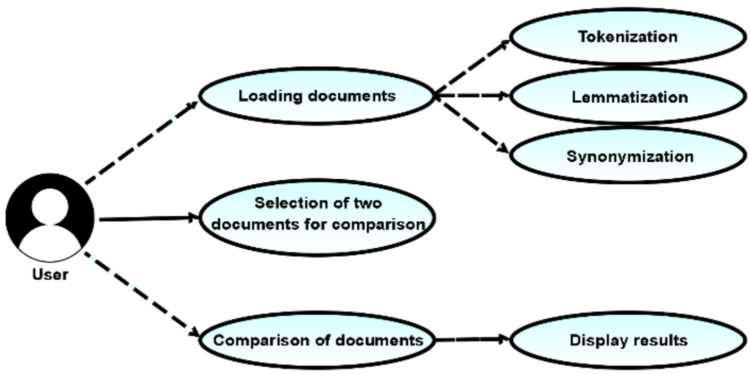


Fig.5. Use case of the system.

On the contrary, the non-functional requirements are:

- Minimization of memory requirements.
- The application implements one or more comparison algorithms.
- The system's graphic interface is user-friendly and easy to operate.
- Similarity matching algorithm uses word frequency technique.
- The system is compatible with the Windows operating system.

2 Design of the Tool

During the implementation, we decided to use the C# programming language with the Visual Studio programming environment. For preprocessing, namely text extraction from the file, we decided to use AbiWord and iTextSharp.

AbiWord is a freely available program, similar to Microsoft Word. It is cross-platform, it has dictionaries for more than 30 languages. It requires to have at least Windows version 2000 and above and 16MB of RAM.

iTextSharp is a PDF library that allows to create, adapt and preserve PDF documents. iText is used in Java, .NET and Android. iTextSharp is portable to .NET [7, 8].

To better describe our application, we show the system architecture in (Fig.6).

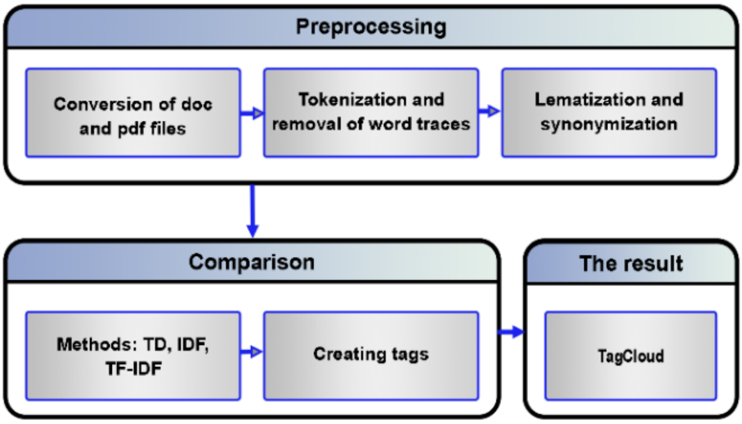


Fig.6. System architecture.

It is important that the environment of even the most demanding applications is simple and intuitive. We also adapted the user interface of our application as shown in (Fig.7).

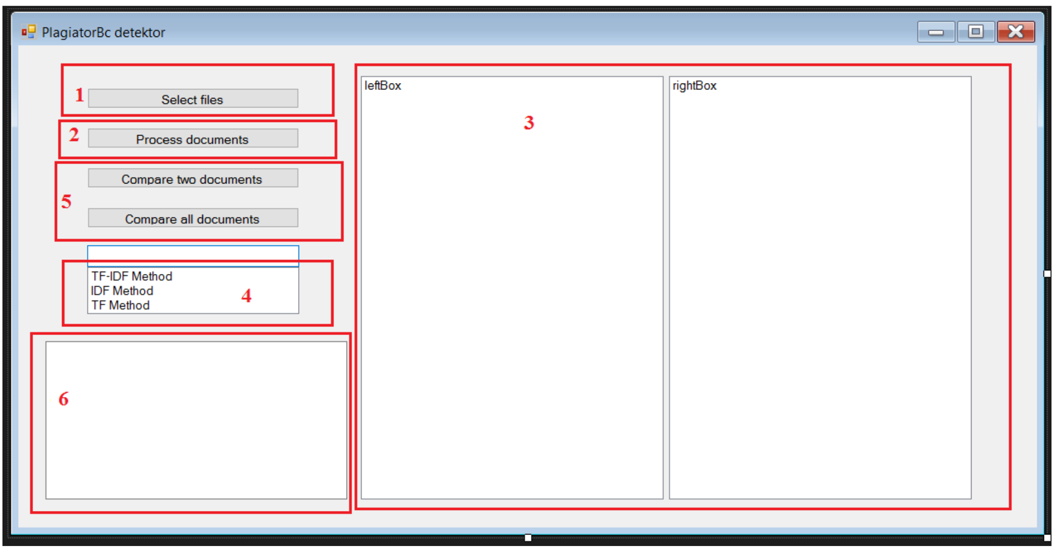


Fig.7. GUI of the system.

We have marked the image with numbers for better understanding. After starting the application, the user first selects the files (1) that he wants to compare. After adding them, he chooses document processing (2), after which they appear in two ListBoxes (3). For the calculation, it is possible to choose a method using ComboBox (4). For the subsequent calculation, he presses Button (5). The user has two Buttons to choose from. One compares two documents using Tag Clouds, and the other lists the percentage match of all documents in the corpus. We write all the information, that we want the user to have at hand, in the ListBox (6).

A. Document loading, conversion and preprocessing

When loading and converting documents, we implemented the IParser interface, which is followed by the Tokenizer class. IParser is used to convert *.doc and *.pdf format files. It is therefore used by two classes, namely ParserDoc and ParserPdf.

ParserDoc uses the external program AbiWord to convert a *.doc file into a plain text file with UTF-8 encoding. The text file is then read using the StreamReader class and processed into a string using the StringBuilder.

ParserPdf works with the external library ItextSharp.dll, which allows to retrieve strings on specific pages of a PDF file. We then merge the obtained strings into one single string using StringBuilder [8].

Tokenizer uses C# regular expressions for tokenization. Any string consisting of meaningful characters is considered a word. Among the meaningful signs we consider dot and comma. StopWords is a class representing the removal of words the listed in the stop list. The list of stop words is stored in a collection of type List<>. This list is read from a file where a line represents one stop word. The verified word must be recorded in lowercase letters.

When implementing lemmatization, we used the dictionary method. The dictionary was obtained from the work in CDB format. Using the procedure, we converted this dictionary to plain text with UTF-8 encoding. The method is implemented in the LematizerSK class. For faster lemmatization, the HashTable class is used. The dictionary is stored by lines. As with lemmatization, we also used the dictionary for synonymization in the Synonymizer class. An already existing dictionary was used, from which a text file was created. HashTable is also used here to speed up the synonymization process [1, 8, 10].

B. Determining the similarity of documents

To implement the visualization page of our application for the selected method - Tag Clouds, we used an external library. For integration into VS 2019, in which we programmed, we wrote an adapter for the given control.

The Tag Clouds method is displayed in the case when the user wants to compare only two specific documents. Each document has its own tag cloud created. The tag cloud is populated with information from the TagCloudData class, which uses TagWord. The TagWord class contains information about a specific document term: its weight, a list of other documents in the corpus where it is found, and the term itself. These TagWords form two sheets, one for each document. In addition, TagCloudData contains the names of the files from which the information was drawn and the percentage dependence of one document on another using the method implemented by us, described below. We also use this method if the user wants to compare all documents in the corpus at once [9].

For the determined similarities of the pre-processed documents, we used the Term Frequency - Inverse Dense Frequency methods. In addition, a simple method was added, which was used to calculate the percentage of documents matching and as a reference when testing other methods.

Our implemented method is based on the normalization of vectors that tell us how far the documents are from each other. After adding them up, we get a cumulative error that expresses the degree of dissimilarity. To obtain the overall similarity of the documents in percent, we subtract this absolute error from unity and multiply by one hundred. The resulting sum indicates the dependence of the given document on the pivot.

The normalization of a vector whose magnitude must be one is performed as follows. For each TagWord that expresses a term in the document, we count its frequency in the given document and divide by the number of all terms in the given document. In this way, we get the value of the normalized vector, which we subsequently use in formula (1). To reduce the complexity of the cycle, we remove all terms whose weight is equal to zero.

For each TagWord in the pivot document, we calculate the absolute value of the difference of the normalized vectors of two identical terms. The first of the vectors belongs to the term of the pivot document, and the second to the vector of the second document. We are looking for the document dependent on the pivot.

$$CumulativeError = \Sigma abs(PivotWordVector - DocumentWordVector) \tag{1}$$

$$Result = [1 - (CumulativeError)] * 100 \tag{2}$$

3 Verification of the Correctness of the Solution

We selected sixteen documents for the first corpus. In this sample, there are known pairs of plagiarisms on which we could verify the size of the estimated similarity. For the existing solution, we selected two comparison methods from an external application - PlaDes: Q-grams and TF-IDF. Q-grams are a modified 3-gram document comparison method.

We compared these two methods with our simple evaluation of two documents. The goal of this experiment is to find out whether the proposed method is close to the other two, and whether it is appropriate to continue using it to express the percentage similarity of two documents in our application.

Table 1. Experiment 1. average deviations

Method	**Average deviations**
Simple Method / Q-gram	8.16 %
Simple Method / TF-IDF	5.06 %

From (Tab.1) and (Fig.8), we can see that our chosen method has only a small deviation against the other two methods. To prove that we don't need a smaller deviation, we will point out that our program also has a visual side to compare two documents - tag clouds. These Tag clouds are created using TF, IDF or TF-IDF methods. As an example, we will show the real result of our prototype, where we will compare two documents that had the highest deviation when comparing Simple Method and TF-IDF.

The TF-IDF method, unlike the Simple Method, calculates the resulting value of term weights based on their frequency in several documents. We also took this fact into account in Tag Cloud. If the user wants to find out how many other documents from the corpus a given term is found in, it is enough to click on the given term in the Tag Cloud. A window with the required information will appear, as we see in (Fig.9).

For the second corpus, we selected the same documents, but compared them to each other, not only known pairs of plagiarists, but also among themselves. There will be comparatively more results than in the first experiment and thus the final result will be more objective. The goal, similar to the first experiment, is to find out credibility of the method.

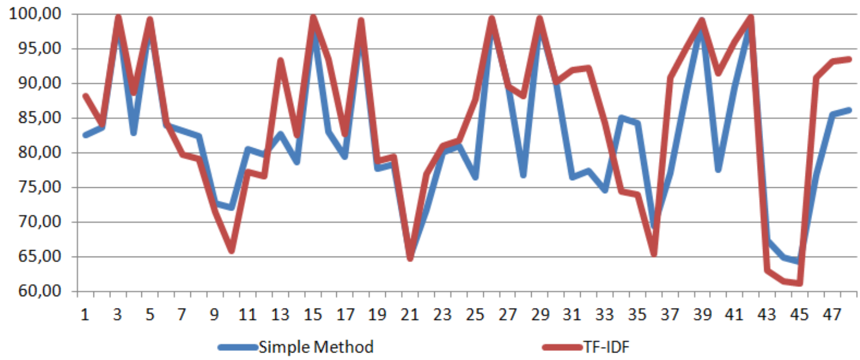


Fig.8. Similarity of documents (Simple Method – TF-IDF).

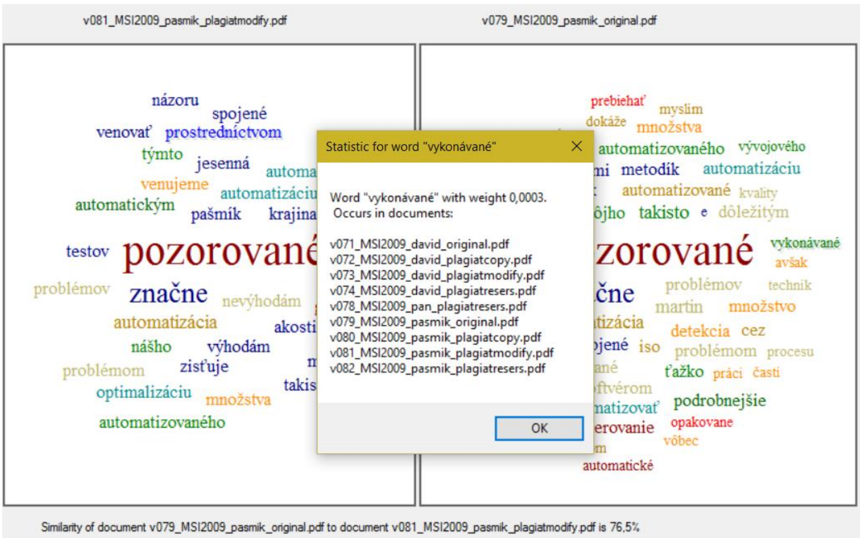


Fig.9. Graphical results in application.

Table 2. Experiment 2. table with statistics

Method	Statistics	Value
Simple Method / TF-IDF	Average deviation	18.07 %
	Number of comparisons with deviation 0-10 %	53
	Number of comparisons with deviation 10-20 %	87
	Number of comparisons with deviation 20-30 %	99
Simple Method / Q-gram	Average deviation	16.04 %
	Number of comparisons with deviation 0-10 %	40
	Number of comparisons with deviation 10-20 %	68
	Number of comparisons with deviation 20-30 %	132

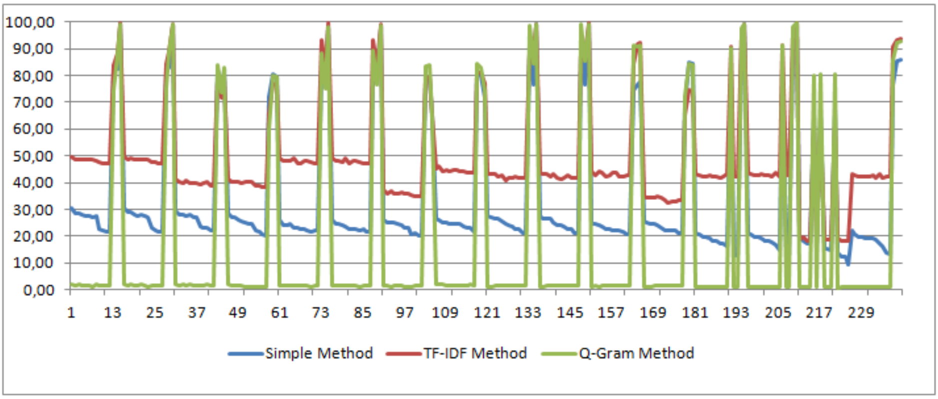


Fig.10. Similarity of documents for different methods.

In (Fig.10) we can see the similarity graph for the individual methods that were compared in this experiment. As you can see, our Simple Method has no problem comparing a high percentage of similarity and thus alerting to an attempted plagiarism. The difference seen for 50 percent or less agreement in methods can be attributed to the fact that these were papers from one problem area - Software Engineering Essays.

Conclusion

The main goal of the work was to supply an existing application for comparing texts and the corpus of submitted student works.

We compared the proposed preprocessing method and the implemented comparison methods with existing methods.

After verifying the correctness of the simple method, it is possible to claim that the used comparison methods are appropriately chosen and the presence layer was appropriately chosen, which will allow the user to display and manipulate the detected similarity results.

Identified problems of the prototype during the implementation and during experimentation, a few shortcomings of the prototype were revealed. Some were removed in time, but it will be continue to try to do so for the following. The visual page of the results of the comparison of all documents in the corpus is not acceptable for the simple and intuitive graphical interface. A possible solution is to find another display method where the results will be easy to read and the user could sort them according to his own will. The time complexity of the process of comparing all documents is too lengthy. 10 comparisons take place in one millisecond. To compare 93 documents with each other during experimentation, the elapsed time to obtain the result was 14 minutes. File formats allowed for comparison are pdf and doc. There is also a possible expansion for other formats.

In the future, the system could be modified to use multiple cores in the computer, or to implement the work of the algorithm in multiple threads. It would be advisable to sort the results or add the option of searching for a specific document in the results. At the same time, it would be good to add the possibility of exporting the results to a file.

Finally, there was implemented only one method of graphical representation of similarity in the paper. In future work, these methods could be expanded, and the user could be given the choice to display the result using, for example, the side-by-side method, or a graph of the connection of similar documents.

Acknowledgement

This paper was supported by the Slovak Grant Agency VEGA 1/0637/23 and KEGA 010STU-4/2023, and by the Scientific Grant APVV-21-0125.

References

[1] D. Chudá, T. Hlatký, M. Kompan, M. Ludvík, R. Švajdlenka. *PlaDes - Plagiarism Detection System*. Bratislava Faculty of Electrical Engineering and Information Technology of STU in Bratislava. [Online]. Available from: http://www2.fiit.stuba.sk/~chuda/plagiarism/PlaDeS.html

[2] Pera, Maria & Ng, Yiu-kai, SimPaD: *A word-similarity sentence-based plagiarism detection tool*, Web documents. Web Intelligence and Agent Systems 9. 27-41. 10.3233/WIA-2011-0203.

[3] D. Chudá. *Advances in database and knowledge technologies 1*. Petr Šaloun, Ján Genci. 1. VSB – Technical University of Ostrava. s. 99. ISBN: 978-80-248-2813-8.

[4] Z. Češka. *Use of modern approaches for plagiarism detection*. [Online] 2008. Available from: http://textmining.zcu.cz/publications/ModerniPristupyProPlagiaty-ITAT2008-Czech.pdf

[5] R. Garabík, L. Gianitsová, A. Horák, M. Šimková. *Tokenization, lemmatization and morphological annotation of the Slovak national corpus*. [Online] 2004. Available from: http://korpus.sk/attachments/publications/2004-garabikgianitsova-horak-simkova-tokenizacia.pdf

[6] M. Pataki. *Plagiarism detection and document chunking methods*. Proceedings of the twelfth international conference on World Wide Web, WWW2003. Budapest. NIIF, 2003.

[7] iText Group nv (HQ Belgium). *IText*. [Online]. Available from: https://itextpdf.com/

[8] the AbiSource community. *AbiWorld* [Online]. Available from: https://www.abisource.com/

[9] A. Shvets, G. Frey, and M. Pavlova. *Design patterns*. [Online]. Available from: https://sourcemaking.com/

[10] R. Garabik. *Slovak morphology analyzer based on levenshtein edit operations*, 2006.

Authors

Ing. Ján Cigánek, PhD.
Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology in Bratislava, Slovakia
jan.ciganek@stuba.sk
He was born in 1981 in Malacky, Slovakia. He received the diploma and PhD. degree in Automatic Control from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEI STU) in Bratislava, in 2005 and 2010, respectively. He is now Assistant Professors at Institute of Automotive Mechatronics FEI STU in Bratislava. His research interests include optimization, robust control design, computational tools, SCADA systems, big data, and hybrid systems.

Ing. Filip Žemla, PhD.
Faculty of Electrical Engineering and Information Technology,
Slovak University of Technology in Bratislava, Slovakia
filip.zemla@icloud.com
He received the diploma and PhD. degree in Automotive Mechatronics from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEI STU) in Bratislava, in 2020 and 2023, respectively. He is now External colleague at Institute of Automotive Mechatronics FEI STU in Bratislava. The research is oriented to virtualization and optimization of modern manufacture processes. His main skills are SCADA systems, database systems and front-end programming.

International Journal
Information Technology Applications (ITA)

Volume 12, Number 2, December 2023

RANSOMWARE ATTACKS AND DEFENSE AGAINST THEM

Miloš Vydarený

Abstract:

In this work, we focus on analyzing the current situation from the perspective of ransomware attacks. In the first part, we describe various types of ransomware and strive to provide a brief characterization of each. From the information gathered through the analysis of the issue, we present currently used software solutions and mention educational resources describing this type of attack, which can serve as a proper informational tool for those unfamiliar with the subject. With the knowledge acquired, we will create a set of preventive methods and procedures aimed at preventing the infection of information systems by ransomware. After implementing preventive solutions, we focus on detecting specific ransomware and on further obtaining information about their functionality and weaknesses based on the use of selected malware analysis methods. The knowledge obtained is then used to introduce measures to minimize their impacts.

Keywords:

Ransomware, Malware, encryption, security, Antivirus, attacks.

Introduction

Ransomware is a type of malicious software, or malware, that prevents the user from accessing computer files, systems, or networks and demands a ransom to restore access to the data. Ransomware attacks can cause costly disruptions to operations and loss of important information and data. An attack with this type of software often manifests only after infection. When infected, users typically do not know that their computer has been compromised by this software. It is usually discovered when they notice that their data is inaccessible or its content is corrupted, and they are then informed about the attack through a pop-up window or by finding a text document left by the attacker outlining what has happened on their computer, along with a specified form of payment method.

Types of Key Management For a successful attack, the ransomware must encrypt the user's data in such a way that the user cannot restore it themselves while ensuring a way for its recovery if the ransom is paid. Therefore, choosing the right method of key management is one of the primary elements for a successful ransomware attack. Symmetric Encryption In this type of encryption, the ransomware uses the same key for both encrypting and decrypting data. Currently, because of this property, symmetric systems are not often used as standalone encryption forms. They are most commonly found in less professional ransomware with a smaller number of infected devices, where protection can be established within a few days. One of the most used algorithms is AES (Fig.1). One of the main advantages of symmetric encryption is its speed. This property is highly sought after in this type of application when it is necessary to encrypt a large volume of data in the shortest possible time.

It is mainly because the ransomware needs to encrypt as much data as possible before an antivirus program might detect a certain pattern in file access and subsequent modification. It might also notice excessive use of cryptographic API. Thus, the more data we are able to encrypt before placing it into quarantine by the antivirus (Fig.2), the better.

One of the main disadvantages of symmetric encryption in this application is that poor hiding of the key can lead to its subsequent discovery. Immediately after encrypting the data, actions must be taken to ensure that the used key is moved out of the user's reach.

One method could be hiding the key on the user's disk. Subsequently, the attacker can send a decryption program capable of finding and using this key to decrypt. More often, however, the method used is such that the symmetric key is present on the victim's computer only during the encryption of files, after which the key is deleted from the computer and sent to the attacker's server. This method has a fundamental problem in that if the victim's PC does not have current internet access, the program can encrypt the data without sending the key to the attacker, meaning that no one can access it, resulting in the data being lost. Some ransomware using this method employs a strategy where they only encrypt if they can connect to the attacker's server, where key collection occurs, and if they do not connect to the server, the encryption process does not occur.

Asymmetric Encryption Asymmetric cryptography, also known as public-key cryptography, uses a mathematically related pair of keys, e.g., a public key for encryption and a mathematically paired private key for decryption (or vice versa). The RSA algorithm is an asymmetric cipher favored by ransomware. Currently, with the right choice of key length, it is not possible to break the encryption, i.e., to obtain the private key within a reasonable time frame, relying solely on the public key and knowledge of the algorithm. When properly implemented, this approach offers greater flexibility to attackers and prevents reversing the encryption without knowledge of the attacker's private key. This implementation usually requires a connection to the internet to function correctly due to sending the generated private key to the attacker's server. Sometimes, the method is also used where the attacker generates one private and one public key before distributing the ransomware, with the public key directly implemented into the ransomware code. The disadvantage of this method of implementation with the public key written directly into the code is associated with the existence of only one private key for decrypting data on all infected devices. Once one user obtains this key, they can make it public, thus making this ransomware irrelevant. However, the main disadvantage for attackers is the fact that encryption with asymmetric cryptography is slow and increases the size of the cryptogram compared to the corresponding plaintext. The encryption process is lengthy, and encrypted data require more storage space on the host.

When encrypting with RSA, we would need to chunk the data into blocks based on the length of the public key, e.g., when using a key with a length of 2048 bits, we can encrypt data with a size of 256 bytes. However, this comes with a problem if we wanted to encrypt data of size 230 bytes, as they would become 256 bytes when encrypted. This way, we could quite significantly occupy space on the disk when encrypting large volumes of data, which is not a desired phenomenon. For this reason, asymmetric encryption is mainly used to encrypt the symmetric key used to encrypt the data on the target PC. This procedure is called hybrid encryption.

1 Hybrid Encryption

Generally, it is known that recent ransomware has employed a hybrid approach combining the previous two models to utilize the best of both types. The user's data are encrypted with a symmetric cipher for speed, while the symmetric key used for encryption is then encrypted using the attacker's public key. This public key is most often part of the ransomware code.

Steps of the Attack Using Hybrid Form

A Ransomware accesses the target person's PC and launches. Most commonly, it gets onto the PC due to the user's actions (phishing, social engineering). Penetration into the PC without the user's fault happens very rarely, most often with the use of exploits.

B Ransomware uses the cryptographic API available on the host to generate the encryption key, very often using AES-256.

C Ransomware encrypts this newly generated symmetric key using the public key (e.g., RSA-2048) contained in the ransomware code and then sends this encrypted key to the attacker. If internet connectivity is unavailable, the key is stored on the disk (the person with the infected PC may have access to it, due to its worthlessness in encrypted form) and then sent when internet access is detected.

D The user's data are encrypted using the symmetric key.

E Subsequently, the ransomware securely deletes the unencrypted symmetric key located on the infected PC. This key is usually never written to the disk; it is stored only in the operational memory during the file encryption period.

F The user is presented with the ransomware's graphical interface with the expectation of paying the ransom.

Crypto Ransomware

This is one of the most widespread types of ransomware. Its primary method of attack involves penetrating the target person's computer through social engineering or phishing. The simplest way to access this type of software is through files and URLs sent via dubious emails or chat applications. Most often, it is a program that masquerades as another file or program. The main functionality of this "software" is to gain access to the user's data for the purpose of encrypting them. In most attacks of this type of ransomware, the user will still be able to use their PC. Thus, the attacker's intent is not to completely disable the target system, but only to encrypt the data and then inform the user of the ransom amount that must be paid within a certain time limit to restore the encrypted data. These types of ransomware achieve non-disabling of the infected system by selecting such components for encryption where no critical system files are located, or by filtering file extensions, e.g., .dll. If the user does not meet the ransom payment deadline, the data may be deleted or remain in encrypted form. The behavior in this scenario is purely based on the implementation chosen by the attacker in the code. What happens to the data after the time limit expires is not guaranteed just as what happens to them after paying the ransom. In many cases, even after paying the ransom, we may not receive the key to decrypt the data. The reason is that some circulating ransomware is currently available online without any attacker managing and caring for the decryption of data, or even collecting the ransom. This means that the money, which the victim sends, for example, to a Bitcoin wallet, may remain locked there forever, because the attacker may no longer have access to its private key.

WannaCry Ransomware

WannaCry (Fig.3) is the most famous representative of crypto-ransomware. The program reached users in the form of an EXE file through emails or downloading dubious files. However, its spread primarily involved the use of an exploit that worked on the Microsoft Windows platform. The exploit arose due to a faulty implementation of the Server Message Block (SMB) protocol, which helps various nodes in a network communicate, mainly when sharing files on the network. This unpatched version of the implementation allowed the running of programs on other computers on the network using specially crafted packets (EternalBlue exploit). Immediately after launching on the target PC, the malware begins to search the network for other PCs on port 445 with an unpatched version of the SMB protocol.

This network spread makes WannaCry ransomware one of the most widespread. It is an attack where a user can be infected even without their own fault, as the ransomware spreads automatically through the network and seeks out other vulnerable systems. Due to the speed and scope of this spread, WannaCry became globally known and extraordinarily costly ransomware that affected a large number of organizations and individuals worldwide. WannaCry uses asymmetric encryption to block access to files on the infected computer, while demanding a ransom be paid in bitcoins for the victims to obtain the decryption key and restore their data. The severity of the attack was emphasized by the fact that many hospitals, large companies, and government organizations were among those affected, which in some cases had serious consequences for providing healthcare and other critical services. The WannaCry case highlighted the importance of regular software updates and the need for robust security protocols in organizations.

Further steps included in the prevention and response to ransomware like WannaCry include:

- Software Patching: Immediate application of security patches to software and operating systems to prevent the exploitation of known vulnerabilities.
- Data Backup: Regular backup of important data on separate and secure media can limit the damage caused by a ransomware attack, allowing quick recovery without the need to pay.
- Employee Education: Strengthening security awareness among employees so they can recognize and avoid suspicious emails and links that may contain malicious content.
- Use of Security Tools: Implementation and maintenance of antivirus software and a firewall, as well as the use of intrusion detection and prevention tools (IPS/IDS).
- Network Segmentation: Dividing the network into smaller segments can limit the ability of ransomware to spread in case it enters the internal network.

Creating a comprehensive and layered security approach is key to protecting against ransomware attacks. Integrating preventative measures with response strategies ensures that organizations can effectively respond to threats and minimize potential damages caused by these attacks.

```java
private static String secretKey = "boooooooooom!!!!";
private static String salt = "ssshhhhhhhhhhh!!!!";

public static String encrypt(String strToEncrypt, String secret)
{
    try
    {
        byte[] iv = { 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 };
        IvParameterSpec ivspec = new IvParameterSpec(iv);

        SecretKeyFactory factory = SecretKeyFactory.getInstance("PBKDF2WithHmacSHA256");
        KeySpec spec = new PBEKeySpec(secretKey.toCharArray(), salt.getBytes(), 65536, 256);
        SecretKey tmp = factory.generateSecret(spec);
        SecretKeySpec secretKey = new SecretKeySpec(tmp.getEncoded(), "AES");

        Cipher cipher = Cipher.getInstance("AES/CBC/PKCS5Padding");
        cipher.init(Cipher.ENCRYPT_MODE, secretKey, ivspec);
        return Base64.getEncoder().encodeToString(cipher.doFinal(strToEncrypt.getBytes("UTF-8")));
    }
    catch (Exception e)
    {
        System.out.println("Error while encrypting: " + e.toString());
    }
    return null;
}
```

Fig.1. Example of encryption using an AES 256 key.

Fig.2. Modern antivirus applications.

Fig.3. Wannacry ransomware warning, ransom demand.

Conclusion

Addressing ransomware requires a comprehensive approach that includes legal, technical, and ethical aspects. Organizations should consider all the consequences of paying the ransom and focus on prevention, detection, and rapid response to cyber threats. At the same time, they should comply with applicable laws and regulations in their jurisdictions and maintain open communication with stakeholders about their security measures and procedures for addressing attacks.

Acknowledgement

This contribution was made with the support of Mr. Ralf Wunschov-Train Consult Korneuburg, Austria, and the training center Wifi Wien Währinger Gürtel, Vienna.

References

[1] Bratvogel, K., (2022). Network Fundamentals.
HERDT-Verlag. 227 p, ISBN 978-3-98569-047-3.

[2] Matoušek, D., (2016). C++ without prior knowledge.
p.19-30, ISBN 978-80-251-4640-8.

[3] martin.haller@patron-it.cz
Retrieved online, May 2024,
https://martinhaller.cz/ransomware/prednaska-zalohy-ktere-nepreziji-ransomware/

[4] martin.haller@patron-it.cz
Retrieved online, May 2024,
https://martinhaller.cz/ransomware/postrehy-ze-zasahu-pri-ransomware-utoku/

Authors

Miloš Vydarený
Faculty of Informatics, Pan-European University, Bratislava, Slovakia
xvydareny@paneurouni.com
His research interests include the field of cybersecurity, the topic of cryptography, ethical hacking and the complete area of computer networks from their design, implementation, deployment, and protection.

List of Reviewers

Issue 2/2023, in alphabetic order

doc. Ing. Petr Drahoš, PhD. Slovak University of Technology in Bratislava, Slovakia
doc. Ing. Oto Haffner, PhD. Slovak University of Technology in Bratislava, Slovakia
doc. Ing. Erik Kučera, PhD. Slovak University of Technology in Bratislava, Slovakia
prof. Danica Rosinová, PhD. Slovak University of Technology in Bratislava, Slovakia

ISSN: 2453-7497 (online)
ISSN 1338-6468