Information Systems in Management
Primary version of the journal is an electronic version.

Editor
Department of Informatics, Warsaw University of Life Sciences – SGGW

Editorial Committee
Dr hab. inż. Arkadiusz Orłowski – Editor-in-Chief
Dr inż. Piotr Jałowiecki – Scientific Secretary / Executive Editor
Prof. nadzw. Kris Gaj – Linguistic Editor
Dr hab. Wiesław Szczesny – Statistical Editor
Dr Piotr Łukasiewicz – Technical Editor

Editorial Council
Dr hab. inż. Arkadiusz Orłowski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie – Chairman
Prof. dr hab. inż. Ryszard Budziński – Uniwersytet Szczeciński
Prof. dr hab. Witold Chmielewski – Uniwersytet Warszawski
Dr hab. inż. Leszek Chmielewski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr hab. Maciej Janowicz – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr inż. Waldemar Karowski – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Prof. Yuiry Kondratenko – Black Sea State University, Ukraina
Prof. dr hab. Marian Niedźwiecki – Uniwersytet Łódzki
Dr inż. Jerzy Pejaś – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
Dr hab. Marian Rusek – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Dr hab. inż. Ludwik Wicki – Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
Prof. dr hab. inż. Antoni Wiliński – Zachodniopomorski Uniwersytet Technologiczny w Szczecinie

Address of the Editor
Faculty of Applied Informatics and Mathematics, WULS – SGGW
ul. Nowoursynowska 166, 02-787 Warszawa, Poland
e-mail: isim@sggw.pl, www.isim.wzim.sggw.pl

ISSN: 2084-5537

Wydawnictwo SGGW
ul. Nowoursynowska 166, 02-787 Warszawa, Poland
e-mail: wydawnictwo@sggw.pl, www.wydawnictwosggw.pl

Table of contents

Janusz Baczyński, Michał Baczyński
TELEROBOTIC TECHNOLOGIES IN E-LEARNING ....................................................... 171

Krzysztof Cabaj, Marek Denis, Michał Buda
MANAGEMENT AND ANALYTICAL SOFTWARE FOR DATA GATHERED
FROM HONEYPOT SYSTEM ....................................................................................... 182

Helena Dudycz
HEURISTIC EVALUATION OF VISUALIZATION OF THE SEMANTIC
SEARCHING ECONOMIC INFORMATION. THE COMPARATIVE
ANALYSIS OF FOUR EXPERIMENTS ....................................................................... 194

Rafik Nafkha, Marcin Olejniczak
PRELIMINARY GUIDELINES FOR THE CONSTRUCTION OF AN EXPERT
SYSTEM FOR SUPPORT OF THE IMPLEMENTATION PROCESS
OF INFORMATION SYSTEMS .................................................................................... 207

Victor Shevchuk
MODELLING OF PRICE AND INCOME EFFECTS ON UKRAINE’S
AGRICULTURAL EXPORT GROWTH ........................................................................ 214

Zofia Stawska, Piotr Milczarski
ALGORITHMS AND METHODS USED IN SKIN AND FACE DETECTION
SUITABLE FOR MOBILE APPLICATIONS .................................................................. 227

Tomasz Ząbkowski, Krzysztof Gajowniczek
SMART METERING AND DATA PRIVACY ISSUES .............................................. 239
Telerobotic Technologies in E-Learning

Janusz Baczyński a), Michał Baczyński b)
a) Faculty of Physics and Applied Informatics, University of Lodz
b) Noe Enterprise Sp. z o.o. – The member of European Robotics Research Network

Today, e-learning methods and techniques are commonly used. In the Internet age they mainly employ different standard forms of the transfer of text and audio-video streams.

However, there are disciplines where the education process cannot be realized by means of the standard e-learning technologies, e.g. physics, chemistry or other practical educational courses. The education process requires on-site presence, e.g. in specialist labs.

The telerobotic technologies can allow e-learning for the courses including practical training. We have adapted a few types of the robotic manipulators to can use them for e-learning. Herein, we also present control systems and software developed by us for this idea. The presented works include the most sophisticated haptic equipment also.

Keywords: E-learning, Distance education, Telepresence, Robotics, TeleRobotics, Cartesian Robots, Haptic technology

1. Introduction

E-learning is the term that has a very broad meaning. In especially, it means the use of various multimedia techniques for distance education. The knowledge can be transferred by text, audio, images, animation, and streaming video. Currently, such processes of education are based on Information and Communication Technologies (ICT) [1]. For example, Web sites that contain a variety of didactic courses are the simplest and most common form of e-learning [2].
The advantages of e-Learning are evident. Knowledge can be transferred at any place, at any time, worldwide. Whoever is taking part in an e-learning course can learn at his own pace in the most appropriate and convenient location. Even the exams and other tests may be carried out via network – most often by Internet. Today, you can so easily imagine that a student can complete his studies and had never be at university.

However, there are learning areas that need a physical on-site presence - for example: physics, chemistry and engineering science. Students in these fields need to practice in different laboratories. Of course there are hybrid (blended) courses for such studies. In this cases, the teaching course is divided into two parts. One of them is the e-learning course and the second part that is hands-on training. The solution is not always optimal.

In recent years, it is observed drastically decrease of interest in learning physics, chemistry and so on. This applies to courses ranging from basic to advanced. In a few years, it may have not enough specialists in many fields of industry and science. To solve this problem, the special didactic projects should be implemented for those who want to study in these areas. From economic point of view, the maintenance of the specialist teaching staff and laboratories in most schools is not cost-effective for such small groups of students. The perfect solution is to support the teaching by e-learning methods. There are a lot of ways for such teaching. Of course, every student can complete an e-learning course before joining a hands-on training. The practical training can be done in specialist learning centers. Of course such training requires the presence in these centers - it is rather unthinkable for students of primary and secondary schools. Today, there are technical possibilities to eliminate the need for the physical presence in many laboratories for practical exercises. In recent years, many tools useful for telepresence [3] have been developed [4] – for example, webcams are widely available. They deliver audio video streams in real time. Many of them can be remotely controlled by a common web browsers. In short, there is ready hardware and software to watch something that is happening somewhere far away. So the existing audio-video tools for the passive telepresence allow the users to feel as if they were present in remote location. Importantly, these elements are commonly used. They are very popular and commercially available and low-cost also.

To achieve effect of active telepresence, the adequate telerobotic tools are needed.

They are needed to cause different events and manipulate various objects in remote location.

Below, there are described our works on telerobotic technologies dedicated to create tele-laboratories that can be used in e-learning. We have developed special software system to test different types of robots, including the most sophisticated haptic equipment also.
2. Analysis of type of robotic manipulators for e-learning

We have done analysis of types of robots in terms of their usefulness in e-learning applications. Generally, robots are electro mechanical devices (most often controlled by computer system) that can perform various physical tasks [5]. There are currently known software robots also. They can be used to replace teachers in performing some repetitive functions. Of course they may find very wide and different use in e-learning but it is not a subject of interest in this work.

Fundamentally, the physical robots consist of an electro-mechanical manipulators and computer systems to drive the manipulators. The computer system controls the manipulator and can do so in two ways: autonomous or in manual mode. Basically, for the e-learning tasks, the manual mode is useful. Of course, the system must be adapted to use the manual mode from any distance through network.

![Classical robotic arm](image1.png) ![Cartesian manipulator](image2.png)

Figure 1. The static robotic manipulators

There are a lot of various constructions of the robotic manipulators [6, 7] – e.g. mobile machines, humanoid robots, static robotic arms and Cartesian robots and many, many others. It seems that only the last two of these groups of manipulators (see the Figure 1) can be most useful in tele workshops.

3. Classical robotic arms

A robotic arm is a type of the manipulator with similar functions to a human arm. The human arm ends a hand with fingers. The end of the robotic arm is called the end effector. A gripper can be the end effector and then it is analogous to the human hand. Thus, it seems natural that the construction is optimal to replace student hand in remote location. There is a great choice of different constructions of the robotic arms ranging from industrial equipment and ending on the gadgets for robotic fans – see the Figure 2. We tested both robotic arms presented on the figure 2.
It is worth noting that the mentioned robotic gadgets are very cheap and they are often equipped with electronics to control them by computers. Some of them have a metal structure but despite all their capabilities are very limited - maximum load and coverage are very small. Moreover, many of them do not have position sensors but often enough to visually determine the position of the end effector. In summary, the mentioned disadvantages make that the equipment can be helpful only for building models of tele laboratories, in especially for testing software.

![Industrial robotic arm](image1.png) ![Small arm - toy for robotic fans](image2.png)

**Figure 2.** The radically different models of robot arms

The considered models of the robots are very easy to adapt for controlling their from a distance by network. The simplest way is the use of so called Remote Desktop Protocol (RDP) that is embedded in many computer operating systems, including the well-known 32-bit and 64-bit versions of WINDOWS – see the Figure 3.

![Remote machine](image3.png)

**Figure 3.** The scheme of remote controlling for simple small models of robots

The tested industrial arm has a lifting capacity up to 3 kg and a coverage of about 0.6 meters. The device is very effective and very efficient and long life equipment that can fully replace the human hand in a large number of activities. Of course the arm is equipped with position sensors with high accuracy better than 0.1 mm. The arm is controlled by a special autonomous controller which practically is a computer equipped with standard ports and a network card. However, the computer is not ready for direct implementation of a tele-control system. The spe-
cial computer software of the controller frees users from problem of solving the inverse kinematics equation \([x]\). Normally, it is needed for calculating all joint angle values of the manipulator to set its arm in desired position. In this case, only values of the linear position and the angular orientation of the end effector should be sent to the controller. It is done by serial port RS-232 – see the Figure 4. The data is sent from the proxy computer that has the TCP server installed (based on Transmission Control Protocol). The special TCP server has been designed by us. It ensures exchanging data between network and the computer serial port. The server software has been designed for WINDOWS platforms.

User computers (client machines) can be a full-fledged WINDOWS computers or any other devices with TCP Client software (e.g. mobile phones or tablets and so on).

![Figure 4. The scheme of remote controlling for the industrial arm](image)

Despite the many advantages, this is by far the most expensive equipment tested in this work.

There is another problem, that is however symptomatic for all family of robotic arms regardless whether they are industrial devices or robotic gadgets. There is a risk that users can perform operations that could damage something in the tele laboratory. Ideally, it would be protected the tele-manipulator system against such adverse activities. However, it is difficult to effectively limit the scope of operations performed by such manipulators. It is relatively easy to limit the workspace for the end effector but there are other operations that can be undesired (e.g. wrong angular orientation in a given point in space, and so on). Thus, the utility of such manipulators for e-learning is debatable and if possible it should be use the other robotic tools.

4. Haptic tools

In the systems presented above, their users can control the tele-manipulators by standard input devices as a keyboard, a touch screen or a computer mouse. The results of their actions may watch through audio-video streams obtained from
a remote location. It is good if a software interface is friendly and it is possible to instinctively control different operations taking place far away from us. Generally, this is not the most intuitive way for controlling performed operations.

It would be ideal if intuitive interfaces for the robotic arms could be based on natural possibilities of a human hand. An linear and angular position of the hand should clearly define appropriate position of the robot. Also it is desirable to sense touch. This can be useful to be able to feel the resistance when the hand is moved too fast in relation to the possibility of the manipulator. Also we should feel resistance when the end effector is on the border of the allowed area of operation. Moreover, the manipulator operates in 3D space so intuitive control devices have just such spatial properties.

To realize such intuitively interface, the commercially available input device has been used. It is called “Phantom Omni® Haptic Device” – see the Photo 5. This device is a practical example of the use of haptic technology [8-10] – it is a tactile feedback technology which would allow a user to feel stimuli from the remote environment.

![Figure 5. The PHANTOM Omni® haptic device](image_url)

The implemented tool is equipped with the special stylus that features 6-DOF positional sensing in the cuboid space (Width = 160 mm, High = 120 mm, Depth = 70 mm). The stylus tip is gimbaled and it is able to detect Yaw, Pitch, Roll and angles with about ± 5% linearity. Moreover the haptic provides the sense of touch (tactile) by generating force feedback up to maximum value about 3,3 N. It is enough to employ the device as the instrument for the operator to control the manipulator.

This haptic device has been implemented in our works not only for the intuitive controlling. It can be also useful for testing the hardness of materials or to give sense of touch of objects placed in distant locations. For example the softness of various materials can be tested if the manipulator is equipped with proper force sensors (standard components in robotics). In especially this system can be built from the pair of the presented haptic devices (one haptic is used as the manipula-
tor). All the tasks can be performed using the computer system as shown on the Figure 6.

In the simplest terms the system works as follow. The linear and angular coordinates are read out cyclically from the haptic controller by the user computer. They are transferred to a manipulator controller (computer) via network. There the data are analyzed also taking into account the dynamic of the operator’s gestures. Now, the manipulator is controlled accordingly to the results of the analysis. The computer controller returns the data determining the amount of force to be generated on the user’s hand.

![Figure 6. The architecture of the haptic system](image)

It should be noted that the haptic controllers may be used also to control virtual objects. Thus, it is possible to create virtual e-learning laboratories from which we receive not only audible and visual stimuli but also touch stimuli.

The main disadvantage of such methods for controlling a remote operation is that they require from users to possess the haptic devices and the haptic equipment is very expensive.

5. Cartesian manipulators

Cartesian manipulators are very often encountered machines in small business and large industry. The robots based on these manipulators are also called linear robots since their end effectors can move linearly along three mutually perpendicular axes [11]. One of the most popular applications for the manipulators are computer numerical control (CNC) machines. These machines are relatively very cheap. Also the required electronics to these machine is very simple. There are many standard commercially available components dedicated for the purpose. Usually, the step motors are used to move the end effector. The motors are controlled by electronic drivers interfaced with a PC computer. Through very many years the CNC machines were interfaced with computers by so called a parallel port – often also called LPT (Line Print Terminal). The typical control system for the CNC machine is shown on the Figure 7.
The step motors drivers must be clocked by pulses with frequencies up to several or dozen kHz. Therefore, the PC controllers (common PC computer) that generate the clocked pulses are often equipped with real time operating systems. It is the good solution when the use of the PC controller is limited only to control the manipulator movement.

![Figure 7. The control hardware of the CNC machine](image)

The screenshot

![The simplified flow chart of the timer](image)

**Figure 8.** The application “Cartesian Manipulator Driver “
In this case all other operations are performed on other computer machines that are linked with the PC controller. For the tele operation applications the computer with multitasking operating system is much more convenient. We have used the PC controller with the system WINDOWS XP and have developed special multi-threaded application to control the manipulator. It has named “Cartesian Manipulator Driver” – see the Figure 8. In this application we have employed the system mechanism called the high-resolution performance counter - in the PC used by us, the resolution of this counter is equal to 279 nsec. In this way we have achieved the possibility of clocking step motor drivers with frequency up to 10 kHz.

The Cartesian Manipulator Driver consists of two threads. One of them is a timer – Fig. 8. The main thread is used to display the current position on the manipulator end effector and for so called “arm solution” – solution of equations of motion of the manipulator [12]. The two thread structure allow the calculation of all parameters needed for the manipulator movement without slowing down clocking the stepper motor drivers.

The following architecture of the tele control system for e-learning applications of the Cartesian manipulators has been devised – see the Figure 9. The prototypes of all software components has been realized and practical tested. The described system has been tested on the local network (LAN) and the Internet. All tests were very successful and showed that the Cartesian manipulators can be easily controlled from any distance. Moreover we have designed the special mechanism to limit the range of movement of the manipulators. It is useful to protect the telelaboratories against irresponsible users.

![Figure 9. The architecture of the tele control system for the Cartesian manipulators](image-url)
6. Conclusions

The aim of this work was to analyze the different types of the robotic manipulators in terms of their suitability for e-learning. The two selected types have been tested in practice by special developed software. Addition, the haptic controllers was tested and for this purpose the special software has been developed also. The system has been tried successfully on a small group of computer science students at Faculty of Physics and Applied Informatics of University of Lodz.

The classical “robotic arms” are most efficient and effective. However the equipment has two very essential disadvantages:
- very high price of these manipulators,
- it is difficult to protect the remote environment against the undesired operations.

The haptic controllers greatly improve the intuitive control from a distance. However, they are expensive and each user of the tele laboratory would have to have the equipment.

The Cartesian manipulators are most optimal devices for remote operations. They are relatively cheap – even about ten times cheaper than industrial robotic arms. They are easy to protect them against unwanted operations performed by the users of the tele laboratories.

REFERENCES


180


MANAGEMENT AND ANALYTICAL SOFTWARE FOR DATA GATHERED FROM HONEYPOT SYSTEM

KRZYSZTOF CABAJ, MAREK DENIS, MICHAŁ BUDA

Institute of Computer Science, Warsaw University of Technology

The paper describes details concerning systems used for analysis and the result of data gathered from two various HoneyPot systems, implemented at Institute of Computer Science. The first system uses data mining techniques for the automatic discovery of interesting patterns in connections directed to the HoneyPot. The second one is responsible for the collection and the initial analysis of attacks dedicated to the Web applications, which nowadays is becoming the most interesting target for cybercriminals. The paper presents results from almost a year of usage, with implemented prototypes, which prove it's practical usefulness. The person performing analysis improves effectiveness by using potentially useful data, which is initially filtered from noise, and automatically generated reports. The usage of data mining techniques allows not only detection of important patterns in rapid manner, but also prevents from overlooking interesting patterns in vast amounts of other irrelevant data.

Keywords: HoneyPot systems, data-mining, monitoring

1. Introduction

Security of computer systems directly connected to the Internet, especially Web applications, becomes more and more important each day. The usage of thousands compromised computers for continuous searching for vulnerabilities in computer systems, inevitably leads to next successful attacks. In order to learn motives, tactics and tools used nowadays by the attackers, HoneyPot systems can
be easily utilized. The HoneyPot is specially crafted and configured machine, or only a chosen service, which is connected to the Internet as a trap for attackers. However, those systems are not used for the production purposes, as its only role is associated with gathering as many information as possible while is being compromised. Software used for implementing various types of HoneyPot systems is easily available. Nonetheless, there is lack of software which could support analysis of gathered data. Using knowledge acquired during many years of HoneyPot system operation and analysis of collected data, the support software was developed and integrated with operational HoneyPot systems.

The paper describes details concerning novel systems used for analysis and the result of data gathered from two various HoneyPot systems, implemented at Institute of Computer Science, Warsaw University of Technology. The first system uses data mining techniques for the automatic discovery of interesting patterns in connections directed to the HoneyPot. The second one is responsible for the collection and the initial analysis of attacks dedicated to the Web applications, which nowadays is becoming the most interesting target for cybercriminals. The paper presents results from almost a year of usage, with implemented prototypes, which prove it's practical usefulness. The person performing analysis improves effectiveness by using potentially useful data, which is initially filtered from noise, and automatically generated reports. The usage of data mining techniques allows not only detection of important patterns in rapid manner but also prevents from overlooking interesting patterns in vast amounts of other irrelevant data.

The paper is organized as follows. The second section describes HoneyPot systems. The third section presents the Miner system, which uses data mining techniques for analysis data gathered from the HoneyPot system. The fourth section is devoted to WebHP system and its monitoring and management software. In section fifth results from initial deployment and operational use of both prototype system are presented. The final sixth, section concludes performed works and indicates future directions and possible improvements.

2. HoneyPot systems

The role of the HoneyPot can be performed by any resource that can be used for observing hostile or unexpected activity. The only common feature of this resource is that it is not used for production purposes. The HoneyPot is mostly specialized machine or software; however, this role can take a fake record in the data base or the account in the important computer system. Any access to the resource, for example, an attempt to read or login, is a sign of unexpected activity. Historically, specially configured computers were used as the HoneyPot system. The configuration enables various monitoring mechanism that during attack gather as many as possible data concerning the attacker activity. For this purpose can be
used logs from operating systems, logs from network devices placed between HoneyPot and Internet or even traces of all traffic directed to it. This solution was ideal for caching and tracking a human attacker but has many drawbacks. The first and the most important is associated with an additional risk. If the attacker detects and disables all monitoring mechanism, the HoneyPot can be used for other hostile activity. Additionally, the initial deployment or cleaning the HoneyPot after a successful attack is very labor intensive. This kind of systems are called high interaction HoneyPots. In the [1] details concerning one of the first well documented development of the HoneyPot and description of further monitoring and tracing real attacker can be found.

In the era of automatic threats, like worms, e-mail viruses or auto-rootsers, dedicated high interaction HoneyPots systems used for gathering copies of malicious code new samples are not efficient and very risky. After each infection the HoneyPot system must be cleaned. This process, even with the support of virtualization, is relatively slow. A better solution for gathering information related with malware is usage of low interaction HoneyPots. The low interaction HoneyPot is dedicated software that imitates vulnerable services. Depending on purpose, it can be very simple, for example, only listing for incoming connections and returning standard banners of simulated service. On the other hand, there are very complicated systems dedicated to downloading new samples of malware. This kind of low interaction HoneyPots simulates high level protocols in which vulnerabilities appears, emulates incoming shellcode used by worm during vulnerability exploitation and downloads next stages of the malware. The most important low interaction HoneyPots are HoneyD [2], Nepenthes [3] and its successor Dionaea [4]. During our research on automatic threats, conducted at Institute of Computer Science, only low interaction HoneyPots are used. Due to limitations of available systems, associated with very poor simulation of Web applications, a custom solution was introduced.

3. Miner

The Miner software was developed as a solution that can automatically detect interesting patterns in data gathered from HoneyPot system. It is integrated with low-interaction HoneyPot Dionaea [4], which provides data for later analysis. Using XMPP protocol information concerning all connections from the Internet that reach the HoneyPot are transferred to the separate analytical system and stored in data base. Later, following a cyclical pattern data from last hour, six hours and 24 hours are analyzed using data mining techniques. This process is implemented in Quechua and Quechua-jep modules. Results, detected interesting patterns, are stored in the same data base. Web interface is used for presenting all detected
patterns. For this purpose custom module called miner was developed and integrated with an open source monitoring system carniwwwhore [5].

Figure 1 presents all elements of the system deployed in network of the Institute of Computer Science, Warsaw University of Technology. Presented arrows shows direction of data transfers.

![Diagram of the system](image)

**Figure 1.** The Miner systems, its Web interface and integration with Dionaea HoneyPot

As previously mentioned, the Miner software uses data mining techniques for analysis. For this purpose two types of patterns are used – frequent sets and jumping emerging patterns. The first pattern was proposed in so called basket analysis, as solution for detection of product sets that are frequently bought together in the markets [6]. In the described system each connection recorded by the HoneyPot is treated as an itemset consisting of five items, associated respectively with source and destination IP address, source and destination port and used protocol. By the definition, frequent set is a subset which appears \( \text{minSup} \) or more times in the analyzed data set. Parameter \( \text{minSup} \) is called minimal support and is given by person who performs analysis. Table 1 presents a sample data set with the connections recorded by HoneyPot.

<table>
<thead>
<tr>
<th></th>
<th>Protocol</th>
<th>Source IP</th>
<th>Source Port</th>
<th>Destination IP</th>
<th>Destination Port</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tcp</td>
<td>10.1.XX.XX</td>
<td>54333</td>
<td>192.168.YY.YY</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>tcp</td>
<td>10.1.XX.XX</td>
<td>54333</td>
<td>192.168.YY.YY</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>tcp</td>
<td>10.1.XX.XX</td>
<td>54333</td>
<td>192.168.YY.YY</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>tcp</td>
<td>172.16.ZZ.ZZ</td>
<td>42356</td>
<td>192.168.YY.YY</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>tcp</td>
<td>172.16.ZZ.ZZ</td>
<td>42456</td>
<td>192.168.YY.YY</td>
<td>8080</td>
</tr>
<tr>
<td>6</td>
<td>tcp</td>
<td>172.16.ZZ.ZZ</td>
<td>44895</td>
<td>192.168.YY.YY</td>
<td>1080</td>
</tr>
</tbody>
</table>
Assumed that we set parameter \( \text{minSup} \) to the value three, various frequent sets can be detected, for example, \(<\text{tcp}, *, *, *, *, *>, <\text{tcp}, *, *, *, 80>, <\text{tcp}, *, *, 192.168.YY.YY, 80>, <\text{tcp}, 10.1.XX.XX, 54333, 192.168.YY.YY, 80>\) or \(<\text{tcp}, 172.16.ZZ.ZZ, *, 192.168.YY.YY, *>\). Asterisk sign presented in the example frequent sets respectively supports initial item sets in ranges, 1-6, 1-4, 1-4, 1-3 and 4-6. The most interesting are the last two which are called maximal, due to the fact, that there is no other detected frequent sets in this data set that are over-sets of them. For further analysis only maximal frequent sets are considered. They are searched in all patterns which are discovered by Miner software using Apriori algorithm.

The second pattern, used in the developed system is called jumping emerging pattern (JEP) [7]. This kind of pattern could be defined between two data sets in which frequent sets are detected. The JEP is a frequent set that is detected in one data set and is not present in the second one. In the Miner system frequent sets are detected in the cyclic pattern in various length intervals: one hour, six hours and 24 hours. JEPs are detected between two adjacent intervals, that have the same duration. In case that some repeated activity interacts with the HoneyPot for longer period, frequent set associated with this events due to usage of JEP is presented only once, in the first interval. The usage of JEPs highlights changes in detected frequent sets, reduces number of patterns that should be inspected by human operator and in the effect reduce possibility of important pattern omission. Figure 2 presents Web interface of the Miner with list of performed detections of frequent sets in variable length intervals.

<table>
<thead>
<tr>
<th>No</th>
<th>From</th>
<th>To</th>
<th>Interval</th>
<th>No. of all frequent items</th>
<th>No. of one generating frequent items</th>
<th>No. of JEPs</th>
<th>No. of JEPs with remote values</th>
<th>Top</th>
</tr>
</thead>
<tbody>
<tr>
<td>30001</td>
<td>2010-01-01 00:00</td>
<td>2010-01-01 01:00</td>
<td>1 hour</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30002</td>
<td>2010-01-01 01:00</td>
<td>2010-01-01 02:00</td>
<td>1 hour</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30003</td>
<td>2010-01-01 02:00</td>
<td>2010-01-01 03:00</td>
<td>1 hour</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30004</td>
<td>2010-01-01 03:00</td>
<td>2010-01-01 04:00</td>
<td>1 hour</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30005</td>
<td>2010-01-01 04:00</td>
<td>2010-01-01 05:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30006</td>
<td>2010-01-01 05:00</td>
<td>2010-01-01 06:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30007</td>
<td>2010-01-01 06:00</td>
<td>2010-01-01 07:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30008</td>
<td>2010-01-01 07:00</td>
<td>2010-01-01 08:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30009</td>
<td>2010-01-01 08:00</td>
<td>2010-01-01 09:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30010</td>
<td>2010-01-01 09:00</td>
<td>2010-01-01 10:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30011</td>
<td>2010-01-01 10:00</td>
<td>2010-01-01 11:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30012</td>
<td>2010-01-01 11:00</td>
<td>2010-01-01 12:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30013</td>
<td>2010-01-01 12:00</td>
<td>2010-01-01 13:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30014</td>
<td>2010-01-01 13:00</td>
<td>2010-01-01 14:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30015</td>
<td>2010-01-01 14:00</td>
<td>2010-01-01 15:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30016</td>
<td>2010-01-01 15:00</td>
<td>2010-01-01 16:00</td>
<td>1 hour</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Appearance of the Miner system Web interface, with list of performed pattern discovery in variable length intervals.
In the table various detailed information concerning detection of frequent sets and JEPs are presented. In subsequent columns id of given calculations, start and stop time of the interval, interval length, number of detected frequent sets, number of maximal frequent sets, number of detected JEPs and number of interesting patterns are presented. Figure 3 presents details concerning detected by the Miner system frequent sets and JEPs.

<table>
<thead>
<tr>
<th>id</th>
<th>counter</th>
<th>proto</th>
<th>remote host</th>
<th>remote port</th>
<th>local host</th>
<th>local port</th>
<th>JEP</th>
<th>Interesting</th>
</tr>
</thead>
<tbody>
<tr>
<td>553600</td>
<td>5</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>3009</td>
<td></td>
<td></td>
<td>YES</td>
<td>no</td>
</tr>
<tr>
<td>553681</td>
<td>9</td>
<td>tcp</td>
<td>111.253.250.137</td>
<td>127.0.0.1</td>
<td></td>
<td></td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

**Figure 3.** Appearance of the Miner system Web interface with details concerning detected frequent sets in given interval

More results concerning information that can be detected using the Miner system are discussed with details in the section number five.

### 4. WebHP and HPMS software

WebHP and HPMS (HoneyPot Management System) software was developed due to limited capabilities associated with gathering details connected with data exchange in application layer between attacker and a low interaction HoneyPot. WebHP was developed as specialized data capture script implemented in PHP language. It must be placed in each monitored page of prepared Web HoneyPot static pages or an application. It is responsible for logging all request send from attacker to database used by HPMS management system. Additionally, in the implemented Web HoneyPot custom error page was prepared, which included data logging script, too. This allows the capture of any request, even if requested page is not present in the Web HoneyPot. The HPMS system was implemented in Python language using Django framework. It allows easy access to all data captured by Web HoneyPot, for example, searching for interesting requests and plotting activity in given time range. Moreover, the user can define rules, which automatically tag all requests matching certain conditions. Figure 4 presents elements of WebHP with HPMS system deployed in network of Institute of Computer Science, Warsaw University of Technology. Figure 5 shows sample screen shot of the HPMS Web interface.
Figure 4. Integration of WebHP and HPMS systems

Figure 5. Appearance of sample HPMS Web interface
5. Results

Both systems described in the previous sections were deployed at the end of the year 2012 in the network of Institute of Computer Science, Warsaw University of Technology. The HoneyPot sensors are placed in the same network using IPv4 addresses, few addresses in distance one from another. Both sensors are freely available from the Internet. The access to the management interfaces were secured only for users working internally or those who have valid access to the internal network via VPN.

Even though HoneyPot sensors are not used for any other activity, and its addresses were not specially announced, during this period of time vast amounts of data have been captured. The Dionaea HoneyPot, which was integrated with the Miner system, received from September 2012 to the end of September 2013, more than 827,5 thousands connections. The WebHP, which was analyzing only connections directed to the WWW services, received from the beginning of November 2012 to the end of September 2013 more than 22,7 thousands connections. This numbers proves that analysis of gathered data manually without specialized software is almost impossible. In the following part of this section the most interesting findings, discovered using implemented management systems, are presented.

The first observation concerning data gathered by both systems shows that automatic scanning is performed for many addresses in given network, one by another. In most cases when some activity from the suspected address was observed in WebHP system, even broader data ware captured by Dionaea integrated with Miner software. Figure 6 shows exemplary request logged by WebHP that checks if it can be used as open proxy. Attacker uses IP address 115.24.164.179 and connects to the HoneyPot at 26 September 23:57. In the similar time, Miner software in six hour interval from 18:00 to 0:00 at 26th September detects frequent set which have item corresponding to this same IP address. Figure 7 shows detected pattern in the user Web interface of the Miner system.

![Figure 6. Searching for proxy logged by WebHP presented in HPMS Web interface. Marked line from IP address 115.24.164.179](image-url)
Detected frequent set has support equal to 16. This is caused by the fact that this scanner searches proxy in various ports, not only at the standard port 80. In this case, they are checked, for example, port number 8888, 8080, 3128, 8118 and 1080.

![Miner - Operation](image)

Figure 7. Detected by the Miner software pattern, which represents searching for proxy performed from IP address 115.24.164.179

The main advantage of the Miner system is associated with patterns discovery. In initial assumptions, each detected pattern represents logged activity, which should be manually inspected by the system operator. As the expected number of detected patterns should be smaller than the number of logged events. During the initial deployment phase, when real data gathered by Honeypot were analyzed, some additional constraints are introduced. In effect, patterns that do not carry interesting knowledge are omitted. For this purpose, in subsequent steps of system development, maximal frequent sets, jumping emerging patterns and interesting patterns are proposed. Maximal patterns cover from the operator all detected by subsets. When the maximal pattern, for example, \(<\text{tcp}, 10.0.XX.XX, *, 192.168.YY.YY, 80>\) is discovered in analyzed data, additionally its subsets are detected, too. In effect, an operator has to search useful data in many other frequent sets, for example, \(<\text{tcp}, *, *, *, >, <\text{tcp}, *, *, *, 80>, <*, *, *, 192.168.YY,80>\). The second improvement reduces additional data when hostile activity is performed for longer periods. If data from HoneyPot is analyzed only using discovery of frequent sets, than longer hostile activity produces many very similar or even identical patterns. The usage of pattern called JEJ reduces the number of generated patterns only to situations in which something changes in the analyzed data. In the effect, the first pattern will be generated, when hostile activity starts and the second when it stops. The last improvement is associated with frequent sets discovery behavior, that produces events which carried little new knowledge. The used algorithm tried to generate any frequent sets. In the effect, when first version of the Miner system was used in intervals with little activity, completely useless patterns were detected, for example, \(<\text{tcp}, *, *, 192.168.YY.YY, *>\) which represents connections using tcp protocol to our HoneyPot with any source address or port. Due to this fact, the definition of interesting patterns was introduced. The interesting pattern is such
frequent set, that is JEP and contains items associated with source port or address. All described in this section improvements reduce number of patterns, that the operator must check. In the analyzed period for more than 827,5 thousands events almost 67 thousands of frequent sets are discovered. In this number there are about 11 thousands of maximal frequent sets, about 5 thousands of JEPs and about 2 thousands interesting frequent sets. These numbers show a reduction of events that the operator must analyze. Moreover, when the operator does not have to find interesting events in vast amount of useless data some interesting data, firstly omitted can be observed. Figure 8 shows a sample analysis concerning one week time frame prepared by the Miner software.

<table>
<thead>
<tr>
<th>id</th>
<th>counter</th>
<th>proto</th>
<th>remote host</th>
<th>remote port</th>
<th>local host</th>
<th>local port</th>
<th>JEP</th>
<th>Interesting</th>
</tr>
</thead>
<tbody>
<tr>
<td>554156</td>
<td>40</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>23</td>
<td>no</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554157</td>
<td>24</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>25</td>
<td>no</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554170</td>
<td>23</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>4899</td>
<td>no</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554171</td>
<td>11</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>5900</td>
<td>no</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554179</td>
<td>12</td>
<td>tcp</td>
<td>123.151.42.61</td>
<td>127.0.0.1</td>
<td>no</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554225</td>
<td>16</td>
<td>tcp</td>
<td>6000</td>
<td>127.0.0.1</td>
<td>3128</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554227</td>
<td>8</td>
<td>tcp</td>
<td>4935</td>
<td>127.0.0.1</td>
<td>3389</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554230</td>
<td>11</td>
<td>udp</td>
<td>12051</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554231</td>
<td>12</td>
<td>udp</td>
<td>12052</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554232</td>
<td>7</td>
<td>udp</td>
<td>12053</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554234</td>
<td>8</td>
<td>udp</td>
<td>188.138.33.215</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554221</td>
<td>6</td>
<td>tcp</td>
<td>5004</td>
<td>127.0.0.1</td>
<td>1433</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>554223</td>
<td>6</td>
<td>tcp</td>
<td>219.235.8.245</td>
<td>127.0.0.1</td>
<td>1433</td>
<td>YES</td>
<td>YES</td>
<td></td>
</tr>
<tr>
<td>554229</td>
<td>5</td>
<td>tcp</td>
<td>188.116.21.86</td>
<td>42343</td>
<td>127.0.0.1</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554236</td>
<td>11</td>
<td>tcp</td>
<td>6000</td>
<td>127.0.0.1</td>
<td>3306</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554237</td>
<td>22</td>
<td>tcp</td>
<td>127.0.0.1</td>
<td>80</td>
<td>no</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>554233</td>
<td>7</td>
<td>udp</td>
<td>12054</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554238</td>
<td>8</td>
<td>udp</td>
<td>50.30.37.9</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554233</td>
<td>6</td>
<td>udp</td>
<td>188.138.41.34</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554229</td>
<td>5</td>
<td>tcp</td>
<td>6000</td>
<td>127.0.0.1</td>
<td>8080</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554222</td>
<td>6</td>
<td>tcp</td>
<td>122.0.66.98</td>
<td>127.0.0.1</td>
<td>1433</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554235</td>
<td>5</td>
<td>udp</td>
<td>173.242.117.164</td>
<td>127.0.0.1</td>
<td>5060</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554226</td>
<td>6</td>
<td>tcp</td>
<td>119.100.21.141</td>
<td>127.0.0.1</td>
<td>1433</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>554220</td>
<td>5</td>
<td>tcp</td>
<td>218.7.37.194</td>
<td>127.0.0.1</td>
<td>22</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8. Patterns detected by the Miner system in week interval. Very interesting patterns representing scanning activity from fixed source port (6000, 4935, 12051, 12052, etc.)
It is interesting that there are some frequent sets representing scanning activity performed from various IP addresses, which use the same source port. The activity of vulnerability scanners that use source port 6000 is well known in security field [8]. However, the detection of scanners that use source port 4935, 12051, 12052, 12053 and 12054 was astonishing. Moreover, without an automatic detection of patterns and implemented filtering function those facts cannot be revealed.

The main advantage of the WebHP is associated with gathering of application data, which can give better insight into attackers intentions. The automatic tagging feature can save an operator time and give opportunity to analyze only unknown activity. During search with the prototype of the system almost twenty distinct tags were discovered and configured in the HPMS systems. Some of them (Proxy and MS BingBot) can be observed in the Figure 6. Even more interesting results can be acquired when the ability to interact with the attacker is used. During conducted experiments, in the WebHP guest book without any “human detection” mechanism was deployed. After few months of inactivity, well organized process of posting hostile links began. During only one week more than 10 thousands of links were added. In this attack 480 distinct IP addresses were used. Further analysis shows that there were two kinds of Bots. The first, which was observed in 388 machines, sequentially placed new posts to the guest book. The second, smaller group which contains 92 machines only checks if guest book is still available, and posts are successfully added. Figure 9 presents plot from HPMS system showing an hourly number of distinct access to guest book pages. Before attack the average of 3 to 5 events occurred in one hour. However, during the attack more than one hundred request are sent to the HoneyPot. The attack has been stopped administratively by disabling the guest book.

![Figure 9. Plot from HPMS showing activity during SPAM attack at the HoneyPot guest book](image-url)
6. Conclusions

Both implemented systems, the Miner and WebHP with HMPS software, were developed using experience from operation use of various HoneyPots in Institute of Computer Science network. The first described system uses data mining techniques for analysis of data gathered from Dionaea HoneyPot. The usage of JEPs and frequent sets indicates the most important data for analysis by human operator. Additionally, filtering achieved by usage of those patterns allows discovery previously unknown patterns, for example, the behavior of some scanning programs, that use hardcoded source ports.

The second system gives insight into data transmitted during attacks on Web sites and Web applications with the level of details that previously was not able to be achieved. Automatic application of tags saves an operator time and allows to analyze only new, previously unseen activities. The ability of the better interaction with attackers gives additional data for further analysis, for example, potentially hostile links placed at the guest book.

Almost year of the operational usage of both systems proves that both systems increase knowledge about attacks directed to the implemented HoneyPots. Functionality build into the systems that does some tedious work, automatically increases productivity of operators and reduces the possibility of interesting events omission.

REFERENCES

The main goal of this paper is to discuss the research on heuristic evaluation of visualization in the semantic search of economic information. It is already the fourth experiment with participants. This time in the research we used two applications built in Protégé 4.1: for analysis of Return on Investment (ROI) indicator according to Du Pont model and for multidimensional early warning system. In the article we briefly described semantic networks as visual interface and premises of conducted study. Then we analysed and compared results of these experiments. Finally, we presented conclusions.

Keywords: interface, visual interface, visualization of semantic network, evaluation of usability of visualizing in searching information, ontology

1. Introduction

Issues of information search based on semantic network technologies is a subject of many studies and concerns various fields (see inter alia [1]; [2]; [16], [18], [19]). In this approach special attention is paid on the role of the visualization of a semantic network which is not only a tool for presenting data, but also provides an interface allowing interactive visual searching information (see inter alia [2], [8], [16]). The combination of data visualization in the form of semantic web and personal navigation can become an effective and efficient tool to perform various analyses, including economic data. The interface is described as good because
it has the proper presentation and efficient navigation allowing users to quickly access the information they need. (see: [9]). Used presentation of data has a major impact on the way in which for example decision–makers interpret the data and assess the usefulness of the system. For users the presentation layer is the most critical element of information and analytical system because it largely shapes the understanding of the basic data on the computer screen. (see: [17]).

Our research concentrates on the usage of visualization methods in searching information basing on semantic network. In this article we discuss four experiments from research on evaluation of visualization in the semantic search of economic information. The paper is structured as follows: in the next section we present shortly semantic networks as a visual interface. In section 3, assumptions of the experiments and the analysis of the results of the research and conclusions are described. Finally, in the last section we give a summary of this work and indicate future research prospects.

2. Semantic networks as a visual interface

One of ideas of collecting and searching data is semantic network, which besides the data themselves contains also information on relations between them, which are encoded in text format (as a metadata). There are many open formats of metadata. These are inter alia computer languages, so called web ontology languages, such as XML, RDF, OWL i SPARQ basing on natural language (see [7]; [16]). The basis of creating semantic network is worked out ontology that defines objects from some field of knowledge and relations between them. In this approach the special attention is paid on the role of the visualization of a semantic network which is not only a tool for presenting data, but also provides an interface allowing interactive visual searching information (see inter alia [8], [16]).

In semantic search visualization is essential, as it allows users to easier notice and understand various both semantic and structural dependences between topics. Basing on displayed semantic structure of topics it is possible to interactively choose analyzed topics or relations, changing area of presented details and obtaining source data. As empiric research carried out by S. Falconer indicates, visualization enhances understanding ontology, making users faster realize conceptual tasks requiring understanding and describing semantic of particular topic [6].

Interactive visualisation allows to actively include user in process of finding information, enabling him to build more accurate queries for specific set (see [13, p. 316]) and facilitating noticing relations between analyzed data. Presentation of data with the use of graphic method supports innovative look at them by user (i.e. decision –makers), allowing him to formulate new hypotheses and their validation. Such approach to visualization of information search is promising solution, because graphical methods and techniques can increase effectiveness of used au-
tomated exploration data methods by using perception and user’s general knowledge [10, p. 1767]). Visual information search consists in using graphic methods, allowing interactive browsing, analyzing and obtaining needed data with user’s active participation.

Using ontologies and semantic networks for visual interface supporting information search in information-analytic tools may solve following solutions [2, p. 216]:

- lack of support in defining business rules for getting proactive information and consulting in the process of decision making;
- lack of semantic layer describing relations between different economic topics;
- lack of support in presenting information on account of different users (employees) and their individual needs;
- difficulty in fast modification of existing databases and data factories in the enterprise in case of new analytic requirements.

Information search based on semantic network requires use of advances graphic interfaces, in which visual navigation in order to obtain needed information is essential.

3. Usability of visualization in the semantic searching economic information – research design

3.1. Assumptions of the research

The aim of the research is inter alia to verify the usability of visualization in the semantic searching economic information in the analysis of economic ratios.

In this article we concentrated on discussing four experiments, which used two applications built in Protégé 4.1: for analysis of Return on Investment (ROI) indicator according to Du Pont model and for multidimensional early warning system (MEWS). The applications created for built ontologies differ in scale of solution, which is important in verifying the usage of TM as a visual tool in searching information on account of semantic connections. In case of the ontology for ROI indicator 44 topics, 6 taxonomic classes with relation of Subclass-Of type and 13 binary relations, whereas in the ontology for an early warning system 142 topics, 23 classes with relation of Subclass-Of type and 20 binary relations were defined.

Study on evaluation of semantic network visualization in information search on account of contextual dependences was conducted using OntoGraf module in program Protégé 4.1 beta. The aim of module OntoGraf, which turned out to be sufficient to carry out the initial research, was to verify the usefulness of visual
semantic network in searching economical information that is contextually con-
nected.

In literature many methods of research and evaluation of human-computer in-
teraction are described (see inter alia [14]; [15]). The research of a prototype can be
conducted with the experts’ participation (e.g. heuristic evaluation of user inter-
face) and/or users (e.g. user testing, usability testing, eye tracking). It was decided
to carry out a research with the participation of users. In the all of four experiments
of the heuristic evaluation of visualization in searching economic information we
applied a combination of two methods of evaluating interface enabling human-
computer interaction, that is heuristic evaluation and usability tests. In literature
there are described many procedures using these methods. The research with the
use of these two methods is realized according to the following plan (see also
[2, pp. 177-178]):

1. Creating test task for the usability testing and questionnaire of heuristic evalua-
tion.

2. Study with participation of users:
   2.1. Selection of research participants.
   2.2. Carrying out study.

3. Data analysis on account of the following criterions:
   — correctness of performing tasks,
   — evaluation of easiness of finding information,
   — evaluation of interface usability,
   — identification of potential difficulties connected with used human-
     computer interaction.

4. Discussion of results and conclusions.

Presented procedure contains both tasks to be performed by research participants
and heuristic evaluation of visual searching information. All four experiments were
conducted according to this plan, but they differed in:
— application, that was used by participants during study,
— the tasks to be performed (without changing the context of implementation),
— time and content of training provided prior to the realization of commands.

In the first three experiments the duration of the introduction to performing tasks
by participants was similar (about 20-30 minutes), but the introduction differed in
content. The observation of users during the first test and analysis of the realization
of usability tests caused a modification of training and used vocabulary before the
next experiments. The second experiment was realized with lower number of
participants, because it was to be preparation for the experiment 3, in which
participants were using more complex application of the ontology for an early
warning system. This research was to answer the question whether the training was
substantially well prepared and whether modification of phrasing in tasks improved finding correct information.

Analysis of the data from the previous three experiments resulted (described in [2]) in the making the following assumptions for the experiment no. 4:

— each participant performs firstly tasks using the application for the ROI indicator, then the MEWS;
— knowledge of the participants differ in terms of both use of information systems and economics;
— introduction to the study takes about 10-15 minutes and it mainly discusses issues related to Protégé 4.1 beta (as in experiment no. 1);
— participants receive a prepared help on a paper (the identical as in experiment no. 2 and no. 3).

Such realization of four experiments resulted from the proposed research method (see [2]; [4]) and the model proposed by E. Brangier (see [5]). These studies enable to identify users’ needs precisely and may contribute to the development of innovations. The important element of the experiments are elaborated questionnaires.

3.2. Scope of questionnaires

According to the presented research plan, the first step was to develop questionnaires covering tasks to be performed using the application for the ontology of ROI indicator and for MEWS ontology as well as heuristic evaluation of the applied interface. To create them we used the conclusions from previous experiments. In the present study the structure of questionnaire is as follows:

• Part no. 1. User profile significantly expanded compared to experiment no. 1. In addition to questions related to personal data, there were also questions concerning used computer equipment.

• Part no. 2. Tasks to be performed in application for ROI indicator. That part of questionnaire consists of list of commands, where study participant records responses –found information. Furthermore in each task there is a table, in which after the execution of instruction the participant evaluates the difficulty of finding information. In the questionnaire five-grade scale was used: very easily (quickly), easily (quickly), average, hard (long), very hard (long). In case of study on application for ROI indicator, the six tasks were formulated, where several tasks are identical (as to the context and manner of performance) with instructions from experiment no. 1 and no. 2.

• Part no. 3. Tasks to be performed in application for MEWS. Same as in part 2, the commands were placed but in this case with use of application for ontology MEWS and assessment of ease of finding information. There are 7 tasks, most of them repeated questionnaire from experiment no. 3.
• Part no. 4. Criteria of an interface evaluation. This fragment of questionnaire is identical as part 2 in the experiments no. 1, no. 2 and no. 3. There were used four criteria for assessment. Each of these criterions is assessed by a user according to five-grade scale, i.e.: highly satisfactory, satisfactory, average, unsatisfactory, and very unsatisfactory.

• Part no. 5. List of potential problems. This part of the questionnaire concerns identification and evaluation of potential difficulties in using the system. This part of the questionnaire is identical as in part 3 in the experiment no 3, which differs by one additional position in relation to the experiment no. 1 and no. 2. Participants of the research choose one of the following answers: no problem, a small problem, an important problem.

The data obtained from the questionnaires used in the experiments can be divided in four main groups that concern:
— correctness of performing tasks,
— evaluation of easiness of finding information,
— evaluation of interface usability,
— identification of potential difficulties connected with used human-computer interaction.

Additionally in the experiment no. 4 the data might be analyze due to the profile of the participants involved in study e.g. their knowledge, gender, owned computer hardware.

3.3. Participants of the research

After the development of questionnaires the research was carried out. In the research (experiment no. 1, experiment no. 2 and experiment no. 3) potential users of topic maps participated. The selection of the participants cannot be random, as they are to fulfil a double role. The first one is to be a typical user, performing specific tasks in a topic map application for ontology indicators (research using the usability testing technique). The second role is to be an expert evaluating the usability of applied interface (research using heuristic evaluation of user interface). None of them either searched information basing on the visualization of ontology before or was familiar with the program Protégé.

In the first research 42 persons aged from 23 to 30 years, who had various experience and knowledge concerning economy and analysis of economic indicators as well as systems and information technology, that is with only computer education, computer science and econometrics education, economic education or non-computer education, took part.

In the second and the third research 14 and 46 persons, respectively, took part. In these two experiments the participants were 20-23 years old and had similar knowledge both of economic terms and computer systems. For the comparison and
verification of the results of this study we have decided to conduct two tests for
two different TM applications.

In the fourth research 41 persons aged from 23 to 27 years, who had various
experience and knowledge concerning economy and analysis of economic indi-
cators as well as systems and information technology (similarly as in the first re-
search). In addition, they indicated in the questionnaire whether their current inter-
est is more toward information technology issues, economic or both. Among those
with informatics secondary education three people indicated information technolo-
y and one - both. In the case of economic-informatics secondary education all
(100%) highlighted: information technology and economic issues. However, the
greatest diversity was among those included in the group of non-informatics educa-
tion (mainly economic education in this group), where economic issues indicated
four people, information technology issues chose three people, and the same num-
ber chose both, while one person answered "I do not know".

3.4. Analysis of the results

In the present paper we will focus on analyzing the results of the fourth
experiment in comparison with previous studies concerning:

— evaluation of interface usability,
— identification of potential difficulties connected with used human-computer
interaction.

Introduction to the experiment no. 4 took about 10-15 minutes, during which we
mainly discussed issues related to Protégé 4.1 beta (as in experiment no. 1). In this
study each participant firstly performed tasks using the application for the ROI
indicator, and then the MEWS. Although the application for the ROI indicator is
smaller and as such theoretically easier than the MEWS, success rate of the task
was significantly smaller. In the case of the application for the ROI indicator the
correct performance of 6 tasks (which consists of searching for proper information)
is shaped in the range of 32% to 95%. However, for the application for MEWS,
where participants performed seven tasks, is in the range of 90% to 100% (results
of this part of the experiment no.4 described in [3]). Analysis of these data allows
to tell that only minimal use of the Semantic Web visualization, without time-
consuming long training is sufficient to understand the idea of action as a semantic
network visualization interface.

In Appendix Table 1 there is data obtained from the research carried out so far
(experiment no. 1, experiment no. 2, experiment no. 3 and experiment no. 4). In the
columns percentages were calculated for the following number of research participants:

1) \( N = 42 \) – experiment no. 1;
2) \( N = 14 \) – experiment no. 2;
3) \( N = 46 \) – experiment no. 3,
4) \( N = 41 \) – experiment no. 4.
Data presented in the table shows that comparing to the experiment no. 1, in the other three experiments visual information search was much better evaluated. Setting about the second and the third experiments, we changed only the wording of the tasks (without changing their difficulty) and the content of training for participants that preceded the realization of the tasks. These changes were the consequence of both the results obtained from the first test and observing participants performing tasks. The number of participation in the experiment no. 2 was smaller than other experiments, because the main aim of this test was verification of preparation of the content of training.

Although in the experiment no. 4, there was the same introduction to the tasks as in the experiment no. 1, but the user ratings are much better (comparable to the experiment no 2 and no. 3). Providing a short but useful help on the handout could have significant importance for faster understanding of the functionality of the tool.

In conclusion, except for the first experiment, participants much better evaluated adopted solution according to the first three criteria. For these three criteria the dominant mark is satisfactory. Looking at this data on the Table 1, this is important information, that there is very small percentage of negative marks: unsatisfactory and very unsatisfactory. Three conclusions result from this data. Firstly, the proposed way of searching for information can be a useful solution for decision-makers carrying out an analysis of economic ratios. Secondly, we should focus more on preparing better content of training. Thirdly, user relatively quickly and easily learns the idea of information search based on visualization of semantic network.

In Appendix Table 2 the data concerning identification of potential difficulties connected to human-computer interaction is presented. Its initial analysis confirms conclusions formulated basing on the analysis of the data contained in Table 1. The modification of only the training and wording of tasks (they were clearer for research participants) significantly improved the evaluation of potential difficulties. In case of experiment no. 1 for four difficulties (i.e. no. 1, no. 2, no. 3 and no. 4) the dominant answer is: a small problem, whereas in case of experiment no. 2 in all seven difficulties considerably dominant answer is: no problem. In case of the third and fourth experiments only for two difficulties (no. 2 and no. 3) there is similar number of marks no problem and marks a small problem. In the other five difficulties (no. 1, no. 4, no. 5, no. 6 and no.7) the dominant answer is: no problem. In the experiment no. 3 and no. 4 an additional question in the questionnaire was introduced (no. 8). The dominant response in both experiments is: a small problem.

Summing up, we can make the same conclusion as from the previous table. Participants of the second, the third and the fourth studies evaluated interface much better than participants of the first study. Results obtained from the research are quite promising in the context of using visualization in the semantic searching economic information to present knowledge on economic indicators.
4. Conclusions and future work

In this article we discussed the results of the research on heuristic evaluation of the visualization in the semantic searching for economic information. We shortly described the proposed research method. We presented the research carried out and discussed the obtained results. An attention should be paid to the fact that in spite of the fact that in the experiment no. 4 we adopted a rule not to explain to participants of the study how to search for information using the Semantic Web visualization, in a relatively short period of time they "discovered" the idea of that interface. Also in the context of the assumptions, the results of the study: evaluation of interface usability and potential difficulties connected with used human-computer interaction are optimistic.

In the research we used heuristic methods of evaluation of human-computer interaction. Obtained data are presented in tables containing percentages of given event. In this research it is essential to get answers to the following questions:
— is it possible to use visualization in the semantic search of economic information as a useful interface in information systems for managers;
— how much time is needed to teach users of system to use visualization of semantic network as interface for searching needed economic information;
— what should be the scope of training to minimize time needed to teach users to use visualization of semantic network.

The essential factors of carried out experiments are inter alia: economic and computer knowledge and experience, duration of training and its content. Therefore analysis of obtained data from these experiments will be continued - multidimensional analysis with the use of statistical measures will be carried out.

The research will be continued in order to verify the creating of the ontology in formal and substantive respect, by testing created applications. At present research on evaluation of visualization in the semantic searching for economic information will be conducted within the project InKoM (described in: [11]; [12]).

REFERENCES


Internet and Information Visualization, V. Geroimenko, C. Chen (red.), Springer-Verlag London, pp. 137-153.


Appendix

Table 1. The heuristic evaluation of the visualization in searching for economic information in economic analysis indicators

<table>
<thead>
<tr>
<th>The criteria for assessment</th>
<th>Scale of usability interface evaluation</th>
<th>Breakdown of accomplishment of tasks (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test 1 (N = 42)</td>
<td>Test 2 (N = 14)</td>
</tr>
<tr>
<td>A. How would you rate the system in terms of visual clarity?</td>
<td>highly satisfactory</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>satisfactory</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>unsatisfactory</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>very unsatisfactory</td>
<td>2</td>
</tr>
<tr>
<td>B. How would you rate the system in terms of its functionality (in the context of searching information)?</td>
<td>highly satisfactory</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>satisfactory</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>unsatisfactory</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>very unsatisfactory</td>
<td>5</td>
</tr>
<tr>
<td>C. How would you rate the system in terms of flexibility of its structure and the presentation of information?</td>
<td>highly satisfactory</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>satisfactory</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>unsatisfactory</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>very unsatisfactory</td>
<td>0</td>
</tr>
<tr>
<td>D. How would you rate the way of searching information that bases on the visualization of semantic network?</td>
<td>highly satisfactory</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>satisfactory</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>unsatisfactory</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>very unsatisfactory</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2. The evaluation of the potential problems with the usage of the visualization of the semantic network in searching for economic information

<table>
<thead>
<tr>
<th>The list of the problems</th>
<th>Scale of the problem evaluation</th>
<th>Breakdown of accomplishment of tasks (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Test 1 N = 42</td>
</tr>
<tr>
<td>1. Understanding how to navigate the OntoGraf</td>
<td>no problem</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>10</td>
</tr>
<tr>
<td>2. Understanding how to execute tasks</td>
<td>no problem</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>7</td>
</tr>
<tr>
<td>3. Understanding the relation between information on the screen and the executed operation</td>
<td>no problem</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>17</td>
</tr>
<tr>
<td>4. Finding necessary information</td>
<td>no problem</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>14</td>
</tr>
<tr>
<td>5. The difficulty in reading information on the screen</td>
<td>no problem</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>12</td>
</tr>
<tr>
<td>6. Too many colors on the screen</td>
<td>no problem</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>17</td>
</tr>
<tr>
<td>7. The necessity to memorize too much information during execution of the task</td>
<td>no problem</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>10</td>
</tr>
<tr>
<td>8. Understanding names of relations between topics</td>
<td>no problem</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>a small problem</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>an important problem</td>
<td>---</td>
</tr>
</tbody>
</table>
PRELIMINARY GUIDELINES FOR THE CONSTRUCTION OF AN EXPERT SYSTEM FOR SUPPORT OF THE IMPLEMENTATION PROCESS OF INFORMATION SYSTEMS

RAFIK NAFKHA, MARCIN OLEJNICZAK

Department of Informatics, Warsaw University of Life Sciences (SGGW)

The process implementing of modern information systems in medium and large enterprises, more and more often is associated with a number of problems that can occur already at the planning stage. Often incorrect or incomplete separation of problem factors causes that the procedure of implementation of such a system can significantly lengthen, and in the worst case end in failure. Therefore, a crucial part of this process, it is possible to determine the potential problems and consequences thereof already at an early stage in the project. Modern technologies of artificial intelligence are increasingly becoming an indispensable tool supporting decision-making in different areas of economic activity. In the following work was analyzed the effectiveness of the use of an expert system in the implementation of information systems based on the skeletal system PC-SHELL. On the basis of research was developed scheme of building the knowledge base for future expert system to support the implementation of the information systems.

Keywords: artificial intelligence, the implementation of information systems, risk of failure, expert systems, domain knowledge

1. Introduction

The dynamic development of the economy has meant that the management of various areas of operation of enterprises have become extremely complex. Many companies comes to the conclusion that an important complement to the basic
strategy of the company, is the strategy of the development of information systems [7], which supporting management processes. Very often, companies choosing to implement the system, are not aware of the problems that can arise in various stages of implementation of the system, which in turn can lead to a situation that the implementation will fail [6].

The costs of failed implementation of information systems in the world each year consume 75 billion dollars and affects approximately 90% of the companies that decide to implement strategic informatics solutions [8]. Often incorrect or incomplete separation problem factors causes that the procedure of implementation of such a system can significantly lengthen, and in the worst case end in failure. Therefore, a crucial part of this process, it is possible to determine the potential problems and consequences thereof already at an early stage in the project. Modern technologies in the field of artificial intelligence can be an indispensable tool supports taking decisions at the level of failure probability estimation system implementation at every stage of its implementation. They are equipped with such a system in a knowledge base containing descriptions of identified good practices in the implementation of systems that can be valuable indicators for used to minimalize the risk of implementation informatics systems.

The following paper presents an analysis of the effectiveness of the use of an expert system in the implementation of information systems based on the skeletal system PC-SHELL. Based on the survey should have been developed a proposal for the construction of a knowledge base schema as the basis for the development of future expert system (SE) supporting the decision-making process concerning the procedures for implementation of information systems.

2. Advisory systems

Expert systems are widely used in many fields, especially as advisory systems for the tasks of identification, classification, control, simulation, diagnostics. They are often referred to as systems based on human knowledge to solve complex problems, generally narrow field that usually require human intelligence [1].

The main task of expert systems to support decision-making, using previously accumulated "knowledge" (based on empirical data), derived from human experts in the field. The result is a solution that offers the appropriate level of expertise, together with inference procedures [3]. Extensive use of expert systems is the result of the many advantages that determine their versatility [4]:

\(\alpha\) provide expert opinions, which are cheaper than the expertise of specialists;
\(\beta\) work faster than specialists;
\(\chi\) improve the quality of expertise, by consequence in drawing conclusions and fewer errors;

208
δ) reduce the length of breaks for systems with continuous operation;
ε) keep readily available expertise, which can be very significant when an insufficient number of experts;
φ) improve safety by replacing specialists in environments and situations hazardous to health or life-threatening;
γ) expertise is comprehensive (it is possible to obtain several alternative solutions);
η) allow a larger group of people to an expert role (a combination of knowledge of many people improve the quality of the knowledge base, and make the expert system will work better than a single expert);
ι) mentally resistant (expert system will operate without interruption, even under stressful conditions);
ϕ) knowledge base that can be easily expanded as experience is gained;
κ) have the ability to explain the found solutions to problems;

The process of creating expert systems can be based on a specialized programming language (e.g. Prolog language) or created on the basis of skeletal expert system. Using the skeletal system, the main task is to acquire and formalize the relevant expertise in the field. As a result, the knowledge accumulated in the system converts the skeletal system used in an appropriate expert system.

3. Scope of the research

The construction of an expert system that supports the processes of information systems implementation, will be based on the knowledge stored in the knowledge base, based on data obtained through a survey. The survey includes a set of threats that have found in the reports of The Standish Group [9].

This group since 1985 collects information about IT projects carried out in enterprises of all sizes with a number of industries including banking, industry, trade and services, health care and education. It provides reliable diagnoses uses of ICT systems and forecasts of future trends. Report prepared on the basis of research based on the conduct of interviews, which included 365 small, medium and large companies using 8380 of systems.

Table 1 contains a sample set of risks identified in the report, The Standish Group [9], which have been isolated over the years. The responses will be used to prepare the knowledge base on the basis of data obtained from respondents.
Table 1. Sample list of risks of failure that may occur during system implementation

<table>
<thead>
<tr>
<th>Lp.</th>
<th>Name of the risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lack or misspelled defined goals ERP system implementation, tailored for the requirements of &quot;strong&quot; department</td>
</tr>
<tr>
<td>2</td>
<td>Lack of support project by the Board</td>
</tr>
<tr>
<td>3</td>
<td>Wrong choice of system or suppliers</td>
</tr>
<tr>
<td>4</td>
<td>Lack of experience in designing contracts</td>
</tr>
<tr>
<td>5</td>
<td>Duty use the law of public procurement by an organizational unit</td>
</tr>
<tr>
<td>6</td>
<td>Lack of knowledge of product on customer side and lack of competence on the supplier side</td>
</tr>
<tr>
<td>7</td>
<td>Lack of knowledge if the system meets requirements for safety and formal regulations-legal (lack of knowledge of whether the system will be able to take into account the polish realities)</td>
</tr>
<tr>
<td>8</td>
<td>Lack of knowledge of current and final economic processes (absence of a centralized and structured organization of knowledge management on the progress of their business processes. Lack of documentation (descriptions) and mapping (flowcharts) business processes. No description of the current 'AS-IS' and the target (the implementation of the system) 'TO-BE'. difficulties in obtaining information / knowledge in the analysis of processes)</td>
</tr>
<tr>
<td>9</td>
<td>Inadequate (lack of performance, or performed as a minimum) rebuilding process BPR (Business Process Reengineering)</td>
</tr>
<tr>
<td>10</td>
<td>Lack of design methodology (lack of measurable business goals and measures of success evaluation of the project)</td>
</tr>
</tbody>
</table>

Source: own based on reports of The Standish Group

4. The concept of the system

As an expert system platform used skeletal system PC-SHELL, which is part of the package SPHINX [5]. All information collected in the process of acquiring knowledge, will be stored in the knowledge base. Stored in the database domain knowledge is declarative in nature (is in the form of rules and facts). The method used in the forward inference is based on the rule:

\[
\text{rules} > \text{facts} > \text{goal}
\]

Modeling of expert system will be carried out on the basis of the skeleton, whose architecture is shown in Figure 1.
The presented architecture of expert system comprises five main components [2]:

a) Knowledge base - contains information obtained from experts in the process of accumulation of knowledge (facts and rules of reasoning). Information is stored in a knowledge base in a declarative way in accordance with prescribed regimens;

b) requesting module - used in the process of inference all collected procedures;

c) explanatory module - allows step-by-step presentation of rules and facts (the reasoning), which enables to generate proposals;

d) module of collection of knowledge (knowledge base editor) - is used to collect domain knowledge, acquired from the experts;

e) User Interface - is used for communication between the user of the system, both in the process of knowledge acquisition as well as a presentation and explain how to fix the problem;

Interface module occupies a central role in the dialogue between the user and an expert system. On the one hand serves as a tool for knowledge acquisition, the other as a mechanism to present the results of the process of inference for users and provide for their understanding.

The proposed system supporting the implementation of information systems should consist of: an expert system, the knowledge base in the form of text files, and - in the case of system expansion with additional mechanisms - neural network
module (responsible for supporting the process of inference). Knowledge base will be supplemented by information obtained in the survey of companies with developed IT infrastructures.

The task of the expert system will support the team of implementation of information system. At each stage of the implementation will be possible to obtain a scenario of conduct to minimize the occurrence of the risk of failure of system implementation. As a result of the responses we get a conclusion on the likelihood of the correct completion of the project or the risk of its failure. After receiving the solutions it is possible to obtain information on the indications of the failure of the implementation of an IT project, and the statements to eliminate the cause of the failure.

5. Summary

Information Systems from year to year will be more and more complex, and the implementation process even more complicated. A key element of the implementation will be able to take on any stage right decisions that will allow to bring the whole process of implementation by the end of achieving the objectives posed at the beginning.

The proposed expert system can be used as a decision making tool for managers and project managers, supervising the implementation of information systems. The main advantages of the system, above all, the opportunity to acquire expertise in the dialogue process user - system. It is planned to make further development of modules statements based on neural networks, which will further improve the accuracy of the expert advice.

REFERENCES


[8] Siegel Wiktor (2012), Wdrożenia we mgle, Raport na temat zarządzania zmianą w Polsce, Door Group.

MODELLING OF PRICE AND INCOME EFFECTS ON UKRAINE’S AGRICULTURAL EXPORT GROWTH

VICTOR SHEVCHUK

Institute of Economics, Sociology and Philosophy, Cracow University of Technology

Using monthly data for the 2001—2013 period, this paper applies Error-Correction Model (ECM) to estimate export demand effects for Ukraine’s agricultural commodities and foodstuffs. According to our results, the long-run exchange rate sensitivity of export demand seems to be rather weak, whereas the domestic income effect is high enough across all four groups: (i) meat, fish and dairy products; (ii) wheat and vegetables; (iii) vegetable oil and (iv) foodstuffs. No evidence is found of the long-term relationship between agricultural exports and foreign trade-partner industrial output. However, both exchange rate and foreign output are established to strongly affect the demand for agricultural exports in the short-run. Also, there is evidence of a speedy short-run adjustment for all groups of agricultural exports to their long-run relationships.

Keywords: agricultural exports, error-correction model, exchange rate effects, price and income effects

1. Introduction

Ukraine’s agricultural sector can be characterized as being in a state of modernizing and steady growth of exports. Ukraine’s agricultural exports have been growing 22.6 percent in the period 2005 to 2012 to a total value of US$17.9 billion dollars, with a remarkable expansion in the exports of cereals, vegetable oil and foodstuffs over last few years (Fig. 1). Demand patterns are influenced by trade liberalization and greater openness. Similar to the Central and Eastern
European (CEE) countries [2; 4, pp. 1–10], it is expected that some key Ukrainian agricultural production (i.e., wheat, barley, maize, sunflower seeds, and sunflower oil) could potentially benefit from a preferential trade agreement with the European Union (EU), with potential gains estimated at US$200 million a year or 0.4 percent of GDP [7]. Among important domestic factors, a steep exchange rate depreciation of 50 percent in 2008–2009 combined with a steady growth in agricultural production over last decade are worth attention as potential determinants of stronger export growth of agricultural crops and foodstuffs.

While it is generally accepted that the nominal (real) exchange rate depreciation is an important determinant of agricultural exports [12, pp. 134–143; 14, pp. 134–143; 19; 22, pp. 160–170], this link either not always holds empirically [6; 15] or it is rather weak [3], especially for developing countries [9]. One of the most plausible explanations refers to the supply-side constraints of a ‘weak’ exchange rate [18, pp. 271–298]. To the same extent, it is not clear whether income and price elasticities are high enough in industrial and developing countries or across particular agricultural commodities. Existing evidences for CEE countries are rather mixed in this respect [2; 10; 21, pp. 463–466].

![Figure 1. Ukraine: agricultural exports (US$ million), 2001–2013](image)

**Source:** Ukraine’s State Statistical Committee

The primary purpose of this paper is to obtain empirical estimates of long-run and short-run price and income export demand effects across four groups of agricultural exports — meat, fish and dairy products (I), wheat and vegetables (II), vegetable oil (III), and foodstuffs (IV). It is important to establish how exchange rate, price and income effects vary by the groups of agricultural export.

Structure of this paper is as follows. In the next section, a brief survey of theoretical arguments and empirical results is presented. In Section 3, review of the data and empirical specification of statistical error-correction model (ECM) used in the empirical analysis of both short- and long-term relationships is provided. The
estimation results are reported in Section 4 along with the analytical interpretation and policy implications. Finally, Section 5 concludes.

2. Literature survey

As is common in empirical studies [4, pp. 1–10; 13, pp. 45–55; 15], the baseline equation for the level of agricultural exports is as follows:

\[ X_t = f(X_{t-1}, E_t, Y_t, Y^*_t, P^*_t, A), \quad X_E, X_{Y^*}, X_{P^*} > 0, \quad X_Y \ll 0 \] (1)

where \( X_t \) is the level of agricultural exports, \( E_t \) is the exchange rate, \( Y_t \) and \( Y^*_t \) are domestic and foreign income (output), respectively, \( P^*_t \) is the international commodity price level, \( A \) is the vector of other exogenous variables (for example, geographical distance, population density, dummies for common border or participation in free trade agreements etc.).

Derivable from theory, export demand is defined as a function of exchange rate, domestic and foreign output, international price level and country-specific exogenous variables. A priori it is assumed that demand for agricultural exports increases with the exchange rate depreciation, higher foreign output and better international prices. The impact of domestic output is not straightforward as this indicator can reflect either higher demand (in the case of a normal good), with a depressing effect on exports, or better supply, leading to stronger export activities. The association between the current and lagged values of exports is expected to be positive, reflecting the lack of constraints for an increase in export activities.

Though potentially export of agricultural commodities (to less extent of foodstuffs) is affected by the principles of competitive advantage, export demand could be enhanced by the exchange rate fluctuations, among other measures. As argued in an extended survey of exchange rate effects on agriculture by Kristinek and Anderson [17], strengthening of the exchange rate as a factor behind decreasing U.S. agricultural exports had attracted a lot of attention in 1970s and 1980s. Orden [19] demonstrates that an appreciation of the dollar leads to a decrease in the export value of agricultural exports, with much lower expansionary effect on imports. Although evidences in favor of a negative link between appreciation of the dollar and agricultural exports still prevail [22, pp. 160–175], to the same extent it is true that exchange rate changes are not the most important determinant of U.S. agricultural exports compared with foreign income [3]. Evidence in favor of positive exchange rate effects on agricultural exports are not lacking for developing countries, for example [12, pp. 70–82; 14, pp. 134–143].

Although it is standard to assume a positive link between exchange rate depreciation and exports at the aggregate level, it could be not a case on the sectoral level. Thus it is possible that exchange rate affects differently each group.
of agricultural exports. As obtained by Brunini et al. [6] for Uruguay, changes in exchange rate have no effect on beef and dairy exports, though positively affecting plastic exports. Idsardi [15] finds that the exchange rate is positively significant in the estimation of exports of South African hop cones only, not affecting other items (wheat, sunflower seeds, vegetables, pistachio nuts). Domestic output is positively correlated with most of agricultural exports. Contrary to expectations, the export of grain is negatively affected by foreign income. Based on the U.S. estimates over 12 commodity categories, Shane et al. [22, pp. 160–175] find that income and exchange rate effects are conditioned by differences between bulk and high value commodities and the type of foreign trade partner, with U.S. exports to low income countries being more exchange rate sensitive than exports to high income countries.

As established by Lamb [18, pp. 271–298] for panel data of 14 African countries, the exchange rate depreciation (in real terms) is negatively related to either export crop production or aggregate agricultural supply. It could be explained by shifts in production away from exports, even if changes in the exchange rate are fully passed into domestic prices, time lag or structural adjustment. At the same time it is confirmed that higher export prices contribute to an increase in agricultural exports. The supply-side factors can explain a recent finding by Colacelli [9] that the real exchange rate elasticities are close to unity for high-income countries and well below unity for developing countries.

Islam and Subramanian [16, pp. 221–231] report that estimates of price and income elasticities of demand for aggregate agricultural exports are low on average, with export price playing a relatively insignificant role in increasing export supply. Such findings are interpreted in favor of diversification of agricultural exports as a pro-growth tool. Similar results of low price and income elasticities for agricultural exports are obtained by Bond [5, pp. 191–227], which contrasts with results of rather high income elasticities for the U.S. exports of wheat and soybeans [13, pp. 45–55]. Abler [1] reports that income elasticities of demand for most agricultural products have been declining in the BRIC countries, with possible exceptions of meat and dairy products. Statistically significant positive link between agricultural exports and domestic agricultural production and international commodity price is found for Turkey [20, pp. 87–96].

Several studies on CEE countries utilize the gravity-type model of agricultural exports, which adds to determinants of bilateral trade flows such factors as geographical distance, population of the country or environment characteristics. Using panel data of 10 CEE countries, Bartošova et al. [2] find that foreign income is a significant positive determinant only for poultry, cheese, and sugar export, with a negative relationship obtained for milk powder. Relative prices contribute to higher exports across all groups. For total agricultural exports, the income elasticity is low but significant in average, while the price elasticities remain relatively large.
At the same time large and positive effects from membership in the EU are estimated on the majority of exports. Bojnec and Fertő [12, pp. 1–10] established that the EU enlargement had contributed to increases in exports of primary agricultural produce and intermediate food-processed goods from CEE countries, though less in higher value-added food-processed differentiated products.

As for individual country studies, Ševela [21, pp. 463–466] find for the Czech Republic that the agricultural exports is positively correlated with the level of income and negatively with geographical distance and GNP per capita (it is interpreted as indicator of low competitiveness at more developed economies). Similar results are obtained by Djurkovic [10] for export of Serbian corn, with negative correlation to income per capita explained by labor intensity in production and inferiority in tastes (it implies substitution for processed rice, fish and meat in line with the increase in income). Based on Poland’s regional data, Ciżkowicz et al. [8, pp. 206–224] established that exports of agricultural and food products is positively correlated with the labour productivity, practical skills of labour force, and negatively with population density and location in the country’s border region.

3. Data and statistical model

The data includes the period 2001M1:2013M3, using monthly series of the four agricultural export groups and the set of independent variables, as it is implied by the equation (1). The exchange rate variable is proxied by a nominal effective exchange rate. As a measure of the international commodity price, indices of agricultural raw materials and food are used. Indices of industrial production in Ukraine, the Eurozone and Russia are used to approximate domestic and foreign output, respectively, as a more direct measure, gross domestic product, is not available at the monthly frequency. Agricultural export series in constant dollars, deflated by the U.S. Consumer Price Index, were taken from the Ukraine’s State Statistical Committee. All other data are obtained from the International Monetary Fund (IMF) International Financial Statistics online database. Since production and export variables reveal a marked seasonal pattern, the series are seasonally adjusted by the X11 procedure.

The stationarity of variables in the model (1) is tested using the ADF unit root test procedure (Table 1). Except foodstuffs (to some extent), the test results are not sensitive to lag length and this outcome stays intact even for higher lags. According to the MacKinnon critical values, for all series, the null of unit root cannot be rejected at 1 and 5 percent statistical significance level for their levels, while it is the case for first differences. As all variables are found to be integrated of order 1, it is necessary to investigate the cointegration relationship between them.
Table 1. Unit Root Test for agricultural exports

<table>
<thead>
<tr>
<th>Lags</th>
<th>Agricultural export groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meat, fish and dairy products (I)</td>
</tr>
<tr>
<td></td>
<td>L</td>
</tr>
<tr>
<td>3</td>
<td>−2.38</td>
</tr>
<tr>
<td>6</td>
<td>−2.20</td>
</tr>
<tr>
<td>9</td>
<td>−2.18</td>
</tr>
<tr>
<td>12</td>
<td>−2.23</td>
</tr>
<tr>
<td>15</td>
<td>−2.18</td>
</tr>
</tbody>
</table>

Note: a maximum lag length of 15 is chosen according to a Schwarz Information Criterion; the null hypothesis of a unit root can be rejected at 1 percent level of confidence (** at 5 percent level of confidence, *** at 10 percent level of confidence); L and FD stand for levels and first differences, respectively.

The Johansen cointegration test results for agricultural export groups, output, commodity prices and nominal effective exchange rate are presented at Table 2. According to the Trace statistics, the hypothesis that there is cointegration relationship between the variables is accepted for all groups of agricultural exports. Though for groups II, III and IV presence of two cointegrating equations is suggested at the 10 percent of statistical significance, this result is rather weak, so that existence of one cointegrating equation is a much more plausible outcome, thus enabling the use of a single equation ECM.

Table 2. Trace Test Statistics for Ukraine’s agricultural exports

<table>
<thead>
<tr>
<th>Number of cointegrating equations</th>
<th>Agricultural export groups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meat, fish and dairy products (I)</td>
</tr>
<tr>
<td>H₀: ( r = r₀ )</td>
<td>( r = 0 )</td>
</tr>
<tr>
<td></td>
<td>65.47*</td>
</tr>
<tr>
<td></td>
<td>61.45**</td>
</tr>
<tr>
<td></td>
<td>88.49***</td>
</tr>
<tr>
<td></td>
<td>74.04***</td>
</tr>
</tbody>
</table>

Note: * denotes rejection of the null hypothesis at the 1 percent level (** at the 5 percent level, *** at the 10 percent level)

As implied by the Engle–Granger two-step methodology, cointegration of the data containing unit roots in the individual time series allows to estimate the long-run relationship (in levels) by standard least-squares techniques and then use the lagged residuals to estimate a short-run dynamics (in first differences). ECMs are based on the assumption that a few time series exhibit an equilibrium relationship.
that determines both short- and long-run behaviour. Despite several potential restrictions, as endogeneity of export prices or infinitely elastic demand, the practice of estimating a single-equation export demand function is widespread, taking into account that empirical assessment of an export supply function is hampered by data constraints.

Following suggestions from other empirical studies, the long-run relationship between agricultural exports and its key determinants may be written:

\[ X_{i,t} = \alpha_0 + \alpha_1 X_{i,t-1} + \alpha_2 E_t + \alpha_3 Y_t + \alpha_4 Y_t^A + \alpha_5 P_t^* + \alpha_6 W_t + \alpha_7 Y_t' + z_t, \]  

(2)

where \( X_{i,t} \) is agricultural exports of group \( i \), \( E_t \) is the nominal effective exchange rate (the price of foreign currency), \( Y_t \) and \( Y_t' \) are domestic and foreign outputs, \( Y_t^A \) is the Ukraine’s agricultural production, \( P_t^* \) is the international commodity price, \( W_t \) is the nominal wage, \( z_t \) is the error term.

The short-run dynamics around the long-run relationship (2) is defined as:

\[ \Delta X_{i,t} = \beta_0 + \beta_1 \Delta X_{i,t-1} + \beta_2 \Delta E_t + \beta_3 \Delta Y_t + \beta_4 \Delta Y_t^A + \beta_5 \Delta P_t^* + \beta_6 \Delta W_t + \gamma \Delta Y_{t-1} + \epsilon_t, \]  

(3)

where \( \Delta \) is the operator of first differences, \( \epsilon_t \) is the error term.

The parameter \( \gamma \) on \( \Delta Y_{t-1} \) is the error-correction coefficient, which reflects the speed of short-run adjustment. According to the Engle-Granger specification, if the lagged error-correction term carries a negative and statistically significant coefficient, all variables are assumed to be converging towards their long-run equilibrium.

4. Empirical results

Our long-term coefficients are reported in Table 3 (the estimates were obtained with EViews 6.1 program). All specifications in levels are characterized by appropriate explanatory power, as measured by high values of the coefficient of determination \( R^2 \). Lack of autocorrelation of the residuals is confirmed by the Durbin-Watson (DW) test. According to the Augmented Dickey-Fuller (ADF) test, the null of hypothesis of a unit root for residuals could be rejected at no less than the 5 percent level of statistical significance. It implies that our regression models allow for correct interpretation of the results obtained.

Similar to other studies [2], all groups of agricultural exports are significantly influenced by the past export performance. Lagged coefficients are in the range from 0.57 to 0.68 and statistically significant at 1 percent. Such a relationship can
be interpreted either in favor of (i) investment activities stimulated by exports revenues or (ii) as an indication of non-saturation of the demand for Ukrainian agricultural commodities and foodstuffs at export markets.

Except meat, fish and dairy products, all other commodity groups are exchange rate inelastic in the long run. Domestic industrial output contributes to the export of meat and foodstuffs, but the inverse relationship is found for wheat and vegetables. Export of vegetable oil is neutral in respect to Ukraine’s industrial output. At the same time there is no long-run relationship to the level of foreign output, either in the Eurozone or in Russia, across all groups of the agricultural export. The highest long-run effect of export demand with respect to the international price level is found for vegetable oil, while the parameter estimates are much smaller for other groups of agricultural exports.

Table 3. Long-term estimates of agricultural export determinants

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Commodity groups</th>
<th>Meat, fish and dairy products (I)</th>
<th>Wheat and vegetables (II)</th>
<th>Vegetable oil (III)</th>
<th>Foodstuffs (IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagged</td>
<td></td>
<td>0.686 (12.73**)</td>
<td>0.576 (7.87*)</td>
<td>0.607 (8.70*)</td>
<td>0.590 (9.18*)</td>
</tr>
<tr>
<td>$E_t$</td>
<td></td>
<td>0.224 (1.63***</td>
<td>—</td>
<td>0.607 (8.70*)</td>
<td>0.590 (9.18*)</td>
</tr>
<tr>
<td>$P_t^*$</td>
<td></td>
<td>0.227 (1.62***</td>
<td>0.412 (1.82*)</td>
<td>0.742 (5.30*)</td>
<td>0.174 (1.67***)</td>
</tr>
<tr>
<td>$Y_t$</td>
<td></td>
<td>0.430 (3.25*)</td>
<td>−0.572 (−2.61*)</td>
<td>—</td>
<td>0.367 (4.19*)</td>
</tr>
<tr>
<td>$Y_{t-1}$</td>
<td></td>
<td>0.177 (1.59***</td>
<td>0.493 (1.88***)</td>
<td>—</td>
<td>0.230 (3.22*)</td>
</tr>
<tr>
<td>$I_t$</td>
<td>−0.123 (−3.48**)</td>
<td>0.347 (4.88')</td>
<td>—</td>
<td>0.132 (2.45**)</td>
<td>0.137 (3.21')</td>
</tr>
<tr>
<td>$W_t$</td>
<td>−</td>
<td>—</td>
<td>0.132 (2.45**)</td>
<td>0.137 (3.21')</td>
<td></td>
</tr>
</tbody>
</table>

Observations: 140 142 132 136

$R^2$: 0.74 0.88 0.93 0.96
DW: 2.12 2.26 2.12 2.28
ADF: −4.01* −4.72** −3.39** −3.79*

Note: *, **, *** denote statistical significance at the 1, 5 and 10 percent level, respectively

Ukraine’s industrial output ($Y_t$) contributes to export of meat and foodstuffs, but it reduces export of wheat and vegetables (no link to export of vegetable oil). As positive effect of domestic output on exports of groups I and IV can be explained by supply of finished goods, an opposite relationship with export of group II can be related to either intermediate character of wheat and vegetables as
production inputs or normality of these goods in private consumption. Not surprisingly, agricultural output \( Y^A_t \) is a factor behind agricultural exports (except vegetable oil). As measured by the nominal wage, domestic demand is likely not to hinder agricultural exports. Just the opposite, there is a positive link with export of vegetable oil and foodstuffs, which can result from the economy of scale in production and, consequently, better international competitiveness.

Import of meat, fish and dairy products ‘crowds out’ exports of the same group of agricultural exports, but it has a positive effect on exports of wheat and vegetables. While the former outcome is not so easy to be explained, substitution effects in private consumption induced by import of meat and fish can explain an increase of export of wheat and vegetables due to higher supply of export goods resulting from lower domestic demand for them.

### Table 4. Short-term estimates of agricultural export determinants

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Groups of agricultural exports</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Meat, fish and dairy products (I)</td>
</tr>
<tr>
<td>Lagged</td>
<td>0.480 (2.83(^*))</td>
</tr>
<tr>
<td>(\Delta Y_t)</td>
<td>0.813 (1.84(^**))</td>
</tr>
<tr>
<td>(\Delta L_Y)</td>
<td>0.628 (1.58(^***))</td>
</tr>
<tr>
<td>(\Delta I_t)</td>
<td>-0.699 (1.40(^*))</td>
</tr>
<tr>
<td>(\Delta I^A_t)</td>
<td>-</td>
</tr>
<tr>
<td>(\Delta I^{EURO}_t)</td>
<td>-</td>
</tr>
<tr>
<td>(\Delta I^{RUS}_t)</td>
<td>1.258 (1.67(^***))</td>
</tr>
<tr>
<td>(\Delta W_t)</td>
<td>-0.145 (1.91(^***))</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

| Error correction coefficients | \(\xi_{i-1}\) | \(-0.818\) (4.22\(^*\)) | -0.934 (5.13\(^*\)) | -0.889 (4.66\(^*\)) | -0.523 (6.68\(^*\)) |
| Observations            | 139                            | 139                        | 131               | 135               |
| R\(^2\)                 | 0.22                           | 0.36                       | 0.27               | 0.39               |
| DW                      | 2.06                           | 1.88                       | 1.95               | 2.06               |
| ADF                     | -4.30\(^*\)                   | -4.81\(^*\)               | -3.90\(^*\)       | -4.86\(^*\)       |
The short-term estimation results are given in Table 4. As it is quite natural for estimates in first differences, the value of the coefficient of determination $R^2$ is much lower if compared with the estimates in levels but it is still high enough. Similar to the long-term estimates (Table 3), statistical properties of our short-term regression models are relevant for correct interpretation.

The short-term estimates of agricultural export determinants are broadly consistent with those obtained above for the long-term relationships but there are several differences. For groups I–III, the coefficient on the nominal exchange rate is found to be positive, indicating an increase in the agricultural exports due to the depreciation of Ukrainian hryvna in the short-run. A one percent depreciation of the hryvna is found to increase exports by about 0.8 to 1.3 percent, which implies a very strong link between the exchange rate and export of agricultural commodities. However, a weakening of the hryvna is likely to cause a decrease in the export of foodstuffs, with the estimated short-term coefficient on the exchange rate of $-0.7$.

Significant relationship between exports and international prices is confirmed in the short-run for meat, fish and dairy products (I), wheat and vegetables (II) and vegetable oil (III), but it is lost in estimates for foodstuffs (IV). It means that export of processed food react to price incentives only in the long-run.

Similar to the long-run relationships, domestic industrial output is a factor behind an increase in export of foodstuffs combined with a decrease in export of wheat and vegetables. However, there is no more a positive income effect on export of meat, fish and dairy products. Similar weakening of the long-run relationship is obtained for agricultural output, though its impact on the export of wheat and vegetables and foodstuffs stay intact.

Short-term estimates restore a link between agricultural exports and trade partners’ income, but the effects are quite heterogeneous. Russia’s industrial output growth ($\Delta Y^{\text{RUS}}$) contributes to export demand for meat, fish and dairy products and foodstuffs, with the income effect of 1.3 and 0.8 percent, respectively. The impact of the Eurozone output growth ($\Delta Y^{\text{EURO}}$) is not that uniform across exports groups. Following a one percent increase in the Eurozone output growth, as measured by industrial production, it is found to increase export of vegetable oil by 2.1 percent, while contributing to a decrease in export of foodstuffs by 1.9 percent and wheat and vegetables by as much as 3.0 percent.

Short-term estimates confirm either expansionary effect of agricultural imports upon export of wheat and vegetables or restrictive effect on export of meat, fish and dairy products, as it is obtained by the long-term estimates. However, the same negative link emerges in the short-run between imports and export of vegetable oil. In the short-run, higher nominal wages re-emerge as a factor behind stronger export growth of foodstuffs, but the same positive link is reversed for export of vegetable oil.
All error correction (or adjustment) coefficients are negative and significant at one percent, implying that export of all agricultural groups move towards its equilibrium level, rather than diverging. Figures ranging from \(-0.5\) to \(-0.8\) indicate that it takes up to two months to reach equilibrium. In other words, there is a swift adjustment of the short-run relationships to the long-run links between agricultural exports across all groups and its determinants.

6. Conclusions

After estimating an error-correction model for agricultural exports across four groups (meat, fish and dairy products; wheat and vegetables; vegetable oil; foodstuffs), it is found that there is a uniform long- and short term relationship to the international prices and lagged export performance. No evidence has found of a long-term relationship between agricultural exports and foreign trade-partner industrial output. A comparison of other results across agricultural groups reveals that estimated coefficients on domestic and foreign output, agricultural production and exchange rate (to less extent) show great variation in either magnitude or coefficient signs. Especially, little evidence is found for a positive long-run effect of the nominal exchange rate, though a depreciation of the hryvna is useful for an increase in agricultural exports in the short-run. However, both exchange rate and foreign output are established to strongly affect the demand for agricultural exports in the short-run, though in a different way across specific groups of agricultural exports. As a depreciation of the hryvna contributes to export of such agricultural commodities, as wheat, vegetables or vegetable oil, it is likely to cause a decrease in the export of foodstuffs.

Our results also support previous findings that links between agricultural exports and income — both domestic and foreign — are quite heterogeneous. In the case of Ukraine, this feature also relates to such determinants of agricultural exports, as import of specific agricultural commodities or domestic wages. As it is obtained that the short-run adjustment is very fast, it implies the lack of any obstacles to convergence towards the equilibrium level of agricultural exports in Ukraine. Overall, our study indicates that Ukraine’s agricultural exports is not constrained by domestic demand or capacity of external markets, at least in the long run, which bodes well for the expansion of export-oriented activities in the agricultural sector.
REFERENCES


ALGORITHMS AND METHODS USED IN SKIN AND FACE DETECTION SUITABLE FOR MOBILE APPLICATIONS

ZOFIA STAWSKA, PIOTR MILCZARSKI

Department of Theoretical Physics and Computer Sciences,
University of Lodz, Poland

Face detection is one of the most important issues in the identification and authentication systems that use biometric features. In this paper we present algorithms for detecting skin colour. The selection and implementation of an algorithm for automated authentication system and face detection can significantly improve the effectiveness of such a system. In the paper we examine several algorithms and methods that can be used in mobile application for authentication purpose i.e. NFC payments.

Keywords: face identification, authentication, biometrics, detection of skin colour

1. Introduction

Authentication and identification systems are now being used increasingly. Nowadays more and more popular become NFC transactions like mobile payments. We want to protect our resources against unauthorized access to the compromised. The commonly used authentication using login and password appear to be insufficient and too easy, which are exposed to a high risk of attack. There is therefore a need to develop alternative methods which seem most promising biometric methods using individual human characteristics. Biometric methods can use a variety of physical characteristics such as fingerprints, hand geometry, ear and facial geometry, iris of the eye [6]. Among these features face as our individual identifier seems to be the most interesting and gives new features research [1, 2, 3].
Face area is now one of the most interesting elements of the image to the research on its location in the image, the appointment of facial symmetry, finding the significant points [15], searching for similarities between several images and compared the skin colour detection. Locating faces in the image using its search algorithms need to be able to work on the details and provides a basis for further research in the diagnosis of skin colour [6].

Searching for skin colour is much simpler than the other much more complex elements of the image. Skin colour is indicative - it is not a physical phenomenon from different angles as light colour will have a different degree of saturation, colour is determined by perception. Facial image captured by the camera in motion is changed due to weather conditions such as sunny or cloudy weather, the exposure of an object or its movement. Image captured by the camera may have different tolerances of colour, through which can recognize different skin colour for one object.

2. Skin colour classification

Study of image content and classification of skin colour in the image gained popularity through active research [11]. It is thanks to research it became possible to determine the distribution of points in the image and operation face. Locating faces and the use of colour information in the face is often the first step to locate her. Classification of colour is important. The research resulted in the creation of colour histograms.

2.1. Skin colour detection – histograms

Histograms show range of colours that identify the colour as your skin. In the first skin colour models used in the base images that have been taught by the elimination of noise in the picture, which property has a skin test and what it is not. It then examines the effectiveness of the photos. Working with photos to create database used in the applications give impressive results. However, histograms created from images did not give satisfactory results in the study of new unidentified photos. Therefore, new, created from the image recording systems based on RGB, YCbCr, HSV, HSI and TLS [7, 8, 9, 10].

2.2. Colour and its models in skin detection

RGB it is the most well-known and widely described method of storing digital images. Yet it is not very accurate because of the colour analysis. This model mixes colours in outdoor conditions such as exposure to cloudy weather.
Nevertheless, it is relatively simple and inexpensive to operate. In addition, recording of images taken for treatment of the following on each computer is in this system [7, 19].

HSV is describing the video recording using saturation and colour intensity. It determined the dominant colour in the image considered in relation to its brightness. This model is to represent the colours with numbers whose values describe what type and size of the image is hidden beneath it [16, 17, 18, 33].

YCrCb [8] is a non-linearly encoded RGB commonly used by television studios and used to work with image compression.

The simplicity of the transformation and a clear separation of components of images, such as exposure and saturation makes this type of recording is becoming very attractive for further studies [10, 17, 20, 21, 22].

TSL means the hue, saturation and exposure. In this case the transformation was made intuitively by using the following formula:

\[
S = \left[\frac{9}{5} (r'^2 + g'^2) \right]^{1/2}
\]

\[
T = \begin{cases} 
\arctan(r'/g')/2\pi + 1/4, & g' > 0 \\
\arctan(r'/g')/2\pi + 3/4, & g' < 0 \\
0, & g' = 0
\end{cases}
\]

\[
L = 0.299R + 0.587G + 0.114B
\]

where \(r' = r - 1/3\), \(g' = g - 1/3\) and the rig are calculated using the formula for the normalization of RGB

\[
r = \frac{R}{R + G + B}, \quad g = \frac{G}{R + G + B}, \quad b = \frac{B}{R + G + B}
\]

There are defined another possible delegated classifications of skin colour also: YES [24], YUV [25] and YIQ [26], [27]. CIE-xyz [14] it is very rarely used.

3. Skin colour detection algorithms

Recording an image in one of these models offers the possibility of further processing. Skin models are formed on the basis of qualifications of pixels contained in the test image, and eliminate those that belong to the skin, and those that do not belong there. Classification of pixels is followed by image storage systems.
Below there are the RGB values to classify areas as skin test pictures.

\[(R,G,B)\text{ is classified as skin if : }\]
\[R > 95 \text{ and } G > 40 \text{ and } B > 20 \text{ and} \]
\[\max\{R,G,B\} - \min\{R,G,B\} > 15 \text{ and} \]
\[|R - G| > 15 \text{ and } R > G \text{ and } R > B\]

Classification of the skin is followed by separating the pixel that was identified. Then pass the standardization process, which resulted in the imposition of a mask on the image and then the image has been subjected to binarization.

![Diagram](image)

**Figure 1.** Processing of the sample image for a particular system by normalization and binarization. The result finds skin component images in the test image


Normalization of pixel colour is based on saving images using the mask. Pixels create colours’ histograms we are interested in identifying the picture. The following example illustrates the process of normalization \(r + g + b = 1\) calculated using the following procedure in the image in RGB mode controlled entry.
It gives us the opportunity to study pictures, build algorithms by which we can classify the test area as the skin [17, 23, 32].

This method made it possible to locate the face in the test image using an algorithm and comparing the pixel colour histograms. The area in which the face becomes a field of research is between the coordinates (X1, Y1), (X2, Y2), (X3, Y3) and (X4, Y4) identified by points of the eyes and mouth.

The algorithm was formulated for the calculation of the frontal regions of the skin to picture the assumption of symmetry of the face varies from 90-110°

X1 and X4 are deployed at the same coordinate (Xi – 1/3 d (k));

X2 and X3 are deployed on the same coordinate (Xk +1 / 3D (k));

Y1 and Y2 are deployed at the same coordinate (Yi + 1 / 3D (k));

Y3 and Y4 deployed on the same coordinate (Yj – 1 / 3D (k));

The results achieved on the basis of the classification of HSI in conjunction with the proposed algorithm for finding faces in images yielded results of 90% of the classification of the skin in the test images. The database consists of 1100 images and the authors [4] obtained the following results for the different colour spaces.
Table 1. Performance of Singh algorithm

<table>
<thead>
<tr>
<th>Criterion</th>
<th>RGB</th>
<th>YCbCr</th>
<th>HIS</th>
<th>Singh algorithm [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photo frontal</td>
<td>56.46%</td>
<td>83.91%</td>
<td>82.27%</td>
<td>96.01%</td>
</tr>
<tr>
<td>Photo rotated</td>
<td>54.57%</td>
<td>80.14%</td>
<td>80.09%</td>
<td>92.42%</td>
</tr>
<tr>
<td>Profile photo</td>
<td>47.84%</td>
<td>80.11%</td>
<td>79.92%</td>
<td>91.29%</td>
</tr>
<tr>
<td>Image with complex</td>
<td>42.62%</td>
<td>73.72%</td>
<td>71.19%</td>
<td>95.18%</td>
</tr>
<tr>
<td>background</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time [s]</td>
<td>2.09</td>
<td>3.46</td>
<td>3.52</td>
<td>6.38</td>
</tr>
</tbody>
</table>

3.2. Tomaz algorithm [5]

Method using TSL (hue, saturation, luminance), distance Mahalanobis (search skin in images), based on the Gaussian model. It seeks to eliminate parts of the image containing the skin. Scaling and Calibration values are input to the algorithm taken from each test pixel after normalization in the range [0, 1.0] [14]. The Mahalanobis distance is applied:

\[ S = \frac{(S - \text{MinS})}{(\text{MaxS} - \text{MinS})} \]
\[ T = \frac{(T - \text{MinS})}{(\text{MaxT} - \text{MinS})} \]

Input filter, which was used in the calculation of the algorithm relates to the classification of pixels: RGB values are represented by 0 to 255.

\[
\begin{align*}
\text{if} & \quad ((B \leq 160 \land R < 180 \land G < 180)) \quad / / \text{too much blue} \\
\text{if} & \quad (G > 160 \land R < 180 \land B < 180)) \quad / / \text{too much green} \\
\text{if} & \quad (B < 100 \land R < 100 \land G < 100)) \quad / / \text{too dark} \\
\text{if} & \quad (G > 200) \quad / / \text{Green} \\
\text{if} & \quad ((R + G < 400)) \quad / / \text{a lot of red and green (yellow)} \\
\text{if} & \quad (G > 150 \land B < 90) \quad / / \text{yellow} \\
\text{if} & \quad (B \land (R + G > 40)) \quad / / \text{too much blue in contrast with other colours} \\
\text{if} & \quad (G \land (R + G > 40)) \quad / / \text{too much green in contrast with other colours} \\
\text{if} & \quad (R < 120 \land B < 170 \land G < 140 \land B < 160)) \quad / / \\
\end{align*}
\]

For skin segmentation Gaussian model was used:

\[ S = \frac{9}{\sqrt{5(r'^2 + g'^2)}} \quad T = \begin{cases} 
\tan^{-1}\left(\frac{r'}{g'}\right) + 0.5 & g' \neq 0 \\
0 & g' = 0 
\end{cases} \]

where

\[ r' = \frac{R}{R + G + B} - 1/3 \quad \text{and} \quad g' = \frac{G}{R + G + B} - 1/3 \]
The class definition describing the skin colour is determined by the formula:

$$C_s = \begin{bmatrix} \sigma^2 M_s & \sigma TM_s \\ \sigma TM_s & \sigma^2 M_s \end{bmatrix}$$

using the formula Mahalanobis of the distance

$$v_m = [M_i, M_j]^T.$$

Pixel distance in the vector $v_m$ calculated using the formula

$$|\lambda^2_{i,j}| = [X_{i,j} - v_m]^T C^{-1}_{j} [X_{i,j} - v_m]$$

It gives the results

$$\lambda^2_{i,j} = \left( \frac{T_{i,j} - M_i}{\sigma^2 M_i T_{i,j}^2 M_i} - \frac{S_{i,j} - M_i}{\sigma T_{i,j} (\sigma^2 M_i - T_{i,j}^2 M_i)} \right) (T_{i,j} - M_i) +$$

$$- \left( \frac{T_{i,j} - M_i}{\sigma T_{i,j} (\sigma^2 M_i - T_{i,j}^2 M_i)} \right) + \left( \frac{S_{i,j} - M_i}{M_i (\sigma^2 M_i - T_{i,j}^2 M_i)} \right) (S_{i,j} - M_i)$$

This method first calculates the pixel values using the formula above values are then subjected to normalization by $\lambda^2_{i,j}$ cons. If the values are greater than 0.7 are identified as skin and this method works quite well. This method detects skin in test images but as every method has errors and is prone to blur the image by clothes and other disturbances.

4. Mixture of gaussian classifiers

Another method based on the RGB system, and my e s a colour Strait Gaussian model is to compare the values of pixels in the histogram of the skin and the skin. Gaussian model is defined by the formula:

$$P(x) = \sum_{i=1}^{N} w_i \frac{1}{(2\pi)^{3/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where x is defined by a vector of RGB, and - the share of the Gauss-scalar value in the vector, u - diagonal diagonal matrix E.

This system gives the possibility of their qualifications. Characterized by the following formulas

$$P(\text{rgb}|\text{skin}) = \frac{s[\text{rgb}]}{T_s}, \quad P(\text{rgb}|-\text{skin}) = \frac{n[\text{rgb}]}{T_n}$$

where s [rgb] is the pixel value of the skin an [RGB] - not leather, T respectively, the amount of all pixels in the image test.
The skin is determined by formula (the most important factor in the equation is the value of the equation for the skin):

\[
\frac{P(\text{rgb}|\text{skin})}{P(\text{rgb}|-\text{skin})} \geq \Theta,
\]

where the detection threshold is set in

\[0 \leq \Theta \leq 1\]

and \(P\) is equivalent to the formula

\[P(\text{skin}) = \frac{T_s}{T_s + T_n}\]

The method gives good results provided clarification image recognition of the skin in the test images. The performance of the algorithm defined by calculation ROC curve [14].

![Figure 2. The effectiveness of the algorithm](image)

The method gives good results provided classification image recognition of the skin in the test images. The effectiveness of the algorithm described in the following Table 2, where the authors used the 4999 and 1909 test shots respondents.
Table 2. Performance of Jones’ method [26]

<table>
<thead>
<tr>
<th></th>
<th>Images correctly classified as human beings</th>
<th>Correctly classified images that are not human beings</th>
<th>The total number of correctly classified images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test images</td>
<td>83.0% (2488/2999)</td>
<td>70.6% (1412/2000)</td>
<td>78.0% (3900/4999)</td>
</tr>
<tr>
<td>Examined images</td>
<td>83.2% (835/1004)</td>
<td>71.3% (645/905)</td>
<td>77.5% (1480/1909)</td>
</tr>
</tbody>
</table>

5. The algorithm based on morphological filters

Another algorithm based on HSV system and the morphological filters [28, 29] used a picture after binarization. Image Filer used, which aims to eliminate noise in an image - such as too small faces, unidentified areas. Given the irregularity of the outline of the face to the location of the image used in the test structure of diamond. Then the hole that was present after filtration was filled with original picture to recover the nose, mouth or eyes. If you change the filter results were inspiring

\[ H = H \ast \text{Filter} \]
\[ S = S \ast \text{Filter} \]
\[ V = V \ast \text{Filter} \]

There was proposed algorithm based on a table that identifies the skin colour and created by Garcia [30]. For any value of \( X \ast Y \) pixels in the image below algorithm to be more efficient:

\[ O_p(XY) + O_m(XY) + O_f(XY) \]

where the \( O_p \) - evaluation of pixels, \( O_m \) - morphology filter, and \( O_f \) - filter pixels.

Figure 3. (a) The original photograph, (b) the result of the algorithm [31], (c) the result of the proposed algorithm
The method can be described as highly efficient method in the location of the face in the image. Due to the identification of the skin and comparison of pixels of the histogram one should use colourful pictures for the tests.

6. Conclusion

In this paper we presented the first part of the algorithms and methods for detecting sets pixels that can represent the human skin. Selected methods have high recognition performance areas of skin such as the face. And an extra characteristic feature is the simplicity and therefore do not require large computing power of the device, which will be applied. These algorithms can be used successfully in the face authentication system complains act on mobile devices on the client side.

Most of the currently available mobile devices such as smartphones have enough computing power to use one of the presented algorithms. These devices can perform the necessary calculations in a sufficiently short time. Therefore, it can be used in applications using face recognition and authentication of users [6, 15] i.e. contactless payment verification using NFC.

Based on the article table can be concluded that the percentage is about 80%. Because most faces were randomly distributed in the pictures, and facial image taken with the smartphone can be restricted to a certain area, their effectiveness in such defined conditions will increase significantly.

Of course, the implementation of these algorithms and methods will not provide transaction security. By forwarding photos of people in front of the camera can easily fool such applications. Therefore, such an application would be strengthened by comparing the series of images of 3D human being [12] or tracking the gaze [13].

REFERENCES


SMART METERING AND DATA PRIVACY ISSUES

TOMASZ ZĄBKOWSKI, KRZYSZTOF GAIOWNICZEK

Faculty of Applied Informatics and Mathematics
Warsaw University of Life Sciences – SGGW

Growing energy consumption enforces initiatives that look for alternatives aimed at better energy management and load balancing. Smart metering is a topic that meets these expectations and it seems to provide a value added for both, suppliers and customers. In this paper we focus on different issues of data and privacy protection for smart grids. In particular, we discuss security concerns related to system architecture, possible means of data protection and demonstrate the main research challenges in privacy assurance for smart grids.

Keywords: smart metering, data protection, privacy protection

1. Introduction

In general, smart metering concerns the usage of some intelligent metering devices at customer location and the regular process of reading, processing and giving the information about consumption to the customer. It needs to be stated that there is a clear distinction between a smart meter and smart metering. The first one is the individual device installed at customer house or facility, primarily measuring the energy consumption. The second one, is a general application of smart meters on a larger scale, connected in a grid. In particular, EU legal framework refers to “intelligent metering systems”. The European Commission's Interpretative Note on Directive 2009/72/EC [8] gives a description of the Commission's understanding of a metering system which is “the ability to provide bi-directional communication
between the consumer and the supplier/operator and to promote services that facilitate energy efficiency within the home.

Smart metering appears to be a remedy for rising prices of electricity and therefore, to encourage parties involved there are many benefits attributed to smart metering systems.

From the individual customer point of view (end user) the main benefits include [6, 7]:

- Access to detailed data to manage energy usage;
- More accurate and timely delivered billing;
- Possibility to benefit from demand flexibility;
- Possibility to introduce safety solutions of the household and equipment through better power quality and breakdown management;
- Other, such as home appliances failure detection, detection of waste, detection of unexpected activity or inactivity, what could be possible with smart home unit controllers.

For the energy supplier, the smart metering offers, among others [6, 7]:

- Possibility to introduce demand response approach what is especially important on electricity market dealing with peak loads;
- Reduced costs of metering readings compared to manual data gathering;
- Reduced back office rebilling process;
- Misuse and fraud detection;

On the other hand there are also costs associated with smart metering implementation. It is clear that the implementation of smart meters will entail number of costs, including the initial cost of the meters, communications costs and also possibly higher maintenance costs of electric devices.

Nevertheless, a serious costs related to smart metering systems concern customer data protection and privacy assurance. Therefore, while building the smart metering infrastructure and dedicated solutions a special attention should be paid to data protection and security issues in order to ensure secure data communication and protection of consumers private data against unauthorized access or hacking.

The purpose of the article is to systematize different issues of data and privacy protection for smart grids. In particular, we discuss different security concerns related to system architecture, possible means of data protection and demonstrate the main research challenges in privacy assurance for smart metering solutions.

2. Smart metering architecture and data flow

ICT (Information and Communication Technologies) systems including smart metering and grid automation possesses functionality, security and real-time
requirements that need to be fulfilled as whole and in a way that is technically and economically feasible. Security threats in smart meter solutions include data tampering in order to manipulate the billing, leakage of private data related to the lifestyle and financial situation of customers, and finally manipulation of grid control commands, which can threaten the whole network.

Particular challenges arise due the large scale of a smart grid and because system components are widely distributed in the field. For this reason the components need to be very stable and secure, particularly in the light of cyber security concept. This concept is defined to be aware of threads conveyed by computers and the protection of the assets from modification or damage from accidental or malicious misuse.

A typical smart metering architecture consists of the following elements [2]:

1) Metering device with associated devices on the customer's site, which can be optionally connected to a smart home controller to manage appliances usage (taking into account tariff information and energy costs);
2) Communication and data processing infrastructure between the customers devices and the transactional systems of the utility supplier;
3) Central data management system which is located on supplier’s site and it has possibility, inter alia, to start/shut down the utility supply, to process data for customer relationship purposes, to archive data according to legal requirements and and optionally to present data to the consumers, through the web page, for instance.

This system is formed by a collection of software, hardware, operators and information flow. However, for the sake of consistency, we will now briefly describe the parties involved in a smart metering showing that they are cross related not only due to electricity flow but also data flow. As shown on Fig. 1 we can distinguish:

- Customers (individual and business); They are end-users that receive the power supply. They generate usage patterns and specific individual consumption information. These are sensitive data that must be protected for preserving the consumers’ privacy. It is assumed that customers would have access to the data on various granularity levels in order to select an advantageous tariff and be able to manage their usage habits and electric appliances.
- Smart metering devices; They are installed at the customer location in order to record the consumed energy at different time windows and send the measurements to the customer and/or the aggregator. Each customer must be equipped with at least one meter, so they are typically small and cheap devices with limited computational power.
- Grid operator (Supplier); This is a company that controls the electricity distribution and transportation infrastructure. Data flow for the operators is
crucial since they may employ electricity usage data and distribution needs in order to manage their resources. With detailed consumption data a better load balancing is feasible.

- Communication network: It concerns communication among all the parties involved in the smart meter grid. Due to sensitive data transmission all communication channels must be secured.

- Electricity producer (power plant): This is a company that produces and then sells the electricity to the customers through the supplier’s infrastructure. The producer must take into account demand data (to adjust the produced electricity), and total consumption data for billing each consumer according the contracted tariff.

![Scheme of example smart meter architecture](image)

**Figure 1.** Scheme of example smart meter architecture

A fully centralized architecture gives the meters only the sensing function, with ability of sending the measurements to a central database. The database acts as a node and communicates with each smart meter. The data stored in
database is then used for consumption calculation, load balancing and billing. Each customer may access the stored data in order to get information about own consumptions. This approach seems to be considered as dominating scenario for smart metering implementation proposals.

A centralized management and data collection implies a trust on the grid operator as this party who would play the role of the chief data officer concentrating also the authentication and storage functionalities, and having access to all the fine-grained measurements, stored in a central database.

Undoubtedly, it possesses many technical and legal difficulties for the delivery of an actual and appropriate privacy preserving solutions.

3. Smart meter data and privacy concerns

There are two European directives that are relevant to data processing in smart meters:

1. The European Data Protection Directive which governs the processing of personal data by data controllers and grants rights to individuals.

2. The European Privacy and Electronic Communications Directive which aim to make it technology neutral.

Under these directives a number of requirements concerning data protection is specified. Firstly, personal data processing is allowed only if specific legal purposes apply. Secondly, personal data gathered for one purpose cannot be used for another purpose without permission. Thirdly, there are limitations on the personal data transfer to other countries. Finally, there is a strict obligation to ensure adequate security.

For this reasons, smart meters cannot transmit any sensitive data such as customer name or address, but to some extend it will involve transmitting personal data through the use of a smart meter ID number, which can be associated with a recipient. The information that smart meters will transfer from the customer to the supplier would include:

- smart meter ID number;
- meter readings on different granularity level;
- type of information transmitted (meter reading or unauthorised access alert);
- date and time;
- payment details for the customers using prepayment meter.

Readings from the smart meter will be gathered through remote access to prepare energy usage profile of the individual or the household. Supplier will be able to use those behavioural data (for instance, low energy usage when the customer is away from home) to prepare kind of energy profile. This can be basis for new services and new tariffs developments based on such energy profiles.
For this reasons, data collected from smart meters will be interesting both, for energy suppliers and consumers.

As the smart meter technologies will capture personal data, energy suppliers will be data controllers and responsible for acting with data protection laws, even if they outsource any data processing services. In fact, energy suppliers already have to comply with data protection laws, so the main requirement concerning existing obligations is adequate enforcement. If final arrangements concerning smart meter solutions take the form of supplier centric approach then data capture, use, storage and sharing will be domain of the energy supplier. Then a question arises: what would energy suppliers do with the data they get from customers? Although there are customers who are in favour of smart meters as a way of providing them more accurate and detailed usage information, some people are concerned about how suppliers will use the energy consumption data arguing this would be an invasion of privacy.

Therefore, there is a need for providing clear and understandable for the customers rules on:

- data capture including the clear statement of what data are allowed to be captured, stored in databases and for what purposes;
- data use (how the customer data are used);
- data storage (how the data are stored ad if they are secured);
- data sharing (what data and how can data be shared with other parties).

Some parties (customers lobby and governments) indicate a need for wide consultation [3] to tackle some key questions such as:

- Should the regulations limit data capture to only that which is necessary for the maintenance and proper functioning of the service?
- Should the regulations limit data usage in any way?
- Should the regulations provide limitations on data storage?
- Should the regulations provide further guidance on data sharing?

Some societies and movements which are opposed to the introduction of smart metering initiatives in United States compares it to the Big Brother’s and argue that it is against The Fourth Amendment (Amendment IV). It is the amendment to the US Constitution which is the part of the Bill of Rights that prohibits unreasonable searches and seizures and requires any warrant to be judicially sanctioned and supported by probable cause.

We have underlined that privacy is a crucial issue in smart metering. With a short example we can show possible privacy violation produced when collecting fine-grained readings from a household power consumption. These are real data gathered for the purpose of smart metering project in one of the households in Warsaw in September and October 2012. Such data reveals information about in-home activities that can be mined and combined with other available information to discover more about inhabitants’ behaviour.
Fig. 2 represents a set of hourly readings for a time span of 400 hours, where the non-consumption period of the household can be easily identified just by eye inspection. This so called “inactivity” period falls on a weekend and this information is very sensitive since it shows theoretical risk concerned with detailed data collection, as it could be used to find out when a home was empty in order to commit burglary, for instance.

![Figure 2. Household’s hourly smart meter readings](image)

Similar case is shown on Fig. 3, in which a set of minute readings for a time span of 24 hours is presented. There, we can distinguish two periods of household

![Figure 3. Household’s smart meter readings per minute](image)
inactivity. One is due to the rest during the night hours and the second one is evident during the afternoon hours. This example reflects of how important privacy protection will be when smart metering is widely deployed.

4. Challenges in data protection assurance

In this part we aim to identify and present several important challenges for considering the privacy issues in smart metering systems, taking into account their importance and influence on ease of adoption.

4.1. Challenges concerning the trust among the parties involved

Taking into account any privacy related scenario, there is an inherent dependency between the trust and privacy. According to [5] “(...) every entities, parties and infrastructure elements of a smart metering network that are trusted will need no privacy protection, and those elements in which privacy is enforced through a secure protocol will not need to be trusted”. Therefore, the definition of the trust model is extremely important for proper and effective customers privacy assurance. In case of smart metering scenario, the main trust relationships are established between the customers, the suppliers and grid operator. Customers’ trust is directly related with privacy of the metered data and it should be stated clearly which parties can access these data for a legitimate purpose. Conversely, the trust from the supplier or grid operator is focused on the data correctness, to provide the actual usage values without trying to forge these measurements and the corresponding bills. The traditional sealed meters accessible only at the customer’s location represented the mutual trust between the supplier and the customers, in a way that customers could not forge the measurements without manipulating the meter and the operator could only access approximate measurements. Therefore, adoption of smart metering changes the trust model and evolving it towards relation that is based on mutual trust, also taking into account the choice of system architecture.

4.2. Challenges concerning the smart meter hardware

For the smart metering growth to be economically feasible, from the grid operators point of view, smart meters devices must be cheap and easily replaceable. This corresponds to a scalability, what means that cost of meter devices deployment at users’ locations must be manageable and must be covered by the energy savings and consumption reduction as a result of smart use and the optimal load balancing. Therefore, smart meters cannot be well equipped and high performance computers but rather small devices with very limited computation resources and importantly characterized by small power consumption.
Due to these fundamental limitations, some of the proposals for privacy-preserving have addressed the use of simple homomorphic encryptions which is the conversion of data into cipher text that can be used or read as if it was still in its original form. Existing current meters don’t possess trusted elements capable of performing complex homomorphic encryption. Even if they have, they use symmetric cryptography [10] usually supporting rather light cryptographic functions like hashes and secret-key encryption/decryption and hash-based message authentication code signatures. Most of the proposed privacy preserving solutions require [9] application of tamper-proof cryptographic modules (similar to smart cards) to handle integrity, distributed authentication and heavy public key data encryption and signatures. Accordingly, if homomorphic processing is chosen then the smart meter devices must also cope with homomorphic operations that include large modular additions, multiplications and exponentiations [9]. This may be too difficult to achieve assuming the low manufacturing cost of these devices.

4.3. Challenges related to cryptographic protocols

In case of the system in which the grid operator company has the control and concentrates the need of trust from the customers, customers will assume that they will be billed correctly for their consumptions. On the other hand, if customer privacy is respected and guaranteed, then it is the utility company who must trust that the measurement and billing calculation are correctly performed, as it will not have access to particular individual measurements. That is why the grid operators are very sceptical to adopt a privacy preserving solution if it does not appear next to fraud detection mechanism and technical guarantees that cheating customers will not take place.

For private protocols based on homomorphic encryption [4], “(...) it is a common requirement that all the encrypted values are produced with the same key in order to be homomorphically combinable in such a way that the secret key is shared among several customers and even the utility company”. In a typical setting, key disclosure would imply losing the possibility of correct authentication. The other problems would include the risk of forgery by users familiar with the technology and decryption techniques. The solution to these problem is, for instance, unusual key distribution mechanisms, like the sub-key generation process proposed by [4] or the peer-to-peer key establishment by [1]. The last one concerns the case, in which each two coupled users share a uniquely computed key for each iteration of the private consumption calculation protocol.

Surely, for privacy protection, the on-going research should also invest in cryptography, also getting familiar with its benefits and certain limitations and, as a result to work out a good solution that is feasible to adopt in smart metering infrastructure.
5. Conclusions

In this paper, we identified a set of data protection and privacy problems in smart grid architecture and present an overview of existing research challenges for secure data processing what brings us towards better understanding of both, customer and system operator needs.

The mainstream of the smart metering systems critique is that it collects personal information. Customer data are collected and it gives a possibility for utility providers to monitor customers behaviours. In order to assure a customer privacy and data protection a certain initiatives should be undertaken, including (1) guidelines regulating access to data for customer service, (2) strong user control over information leaving the customer location, (3) protocols that can process most of the data at customer locations.

Undoubtedly, the topic is very complex and difficult since it tackles very sensitive issues. However, the consensus appears to be possible to achieve assuming that each of the parties involved would show a good will.

Acknowledgments

This research was financed by VEDIA S.A. leading a project partially supported by National Centre for Research and Development (NCBiR).

REFERENCES


