ITHEA

International Journal INFORMATION THEORIES & APPLICATIONS

2020 Volume 27 Number 3

International Journal

INFORMATION THEORIES & APPLICATIONS
Volume 27 / 2020, Number 3

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application. IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA are given on *www.ithea.org*.

Responsibility for papers *published in* IJ ITA belongs to authors.

International Journal "INFORMATION THEORIES & APPLICATIONS"
Vol. 27, Number 3, 2020

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with: University of Telecommunications and Posts, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politécnica de Madrid, Spain,
Hasselt University, Belgium, University of Perugia, Italy,
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia
St. Petersburg Institute of Informatics, RAS, Russia,

Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editor: **Ina Markova**

ISSN 1310-0513 (printed) **ISSN 1313-0463 (online)**

USING APRIORI ASSOCIATION RULES AND DECISION TREE ANALYSIS FOR DETECTING CUSTOMS FRAUD

Hamzah Qabbaah, George Sammour, Koen Vanhoof

***Abstract**:*

All over the world some people and companies try to avoid taxes whenever possible. Customs are not an exception to this. In this paper we investigate how customs fraud can be detected using data mining on logistics transaction data. We used both the Apriori algorithm for association rules and decision tree analysis to do so. We first transformed the continuous variables using both k-means clustering and CHAID decision tree analysis for the continuous variables in the data set. Analysis of the rules detected by our analysis indicates that it is possible via this methodology to find indicators of customs fraud cases. Moreover, it was possible to describe in detail the situation in which the fraud occurred (product type, country of origin, e-shipper, the weight and the price of the products shipped). The results of the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, and the evaluation using the classification measures (Accuracy, Precession, Recall and F1) based on the confusion matrix shows high percentages. This confirms the value of our conclusions.

***Keywords**: Customs fraud, CHAID decision tree, Apriori association rules, Data mining, Logistics.*

1. Introduction

Any act by which a person deceives, or attempts to deceive, the customs and thus evades, or attempts to evade, wholly or partly, the payment of import or export duties and taxes constitutes customs fraud (ACFE 2018) . Clearly both companies and governments would benefit from a system that could detect this as soon as possible in order to avoid hassle.

The literature on fraud detection focuses mainly on internet fraud by owners of websites to increase their income, the so-called ‘Click-fraud’. This fraud can have several origins. Some web-owners are simply dishonest and use automation to generate traffic to defraud advertisers (Metwally, Agrawal et al. 2007). Several fraud detection algorithms have been proposed for this problem. A new architecture for web fraud detection using an Apriori algorithm for association rule mining in a web advertising network was introduced by Tripathi et al. (Tripathi, Nigam et al. 2017). Zhang et al. presented a technique to detect duplicate clicks in jumping windows and sliding windows. They did so using two innovative algorithms that make only one pass over click streams (Zhang and Guan 2008). Haddadi defined bluff ads, which are a group of ads that join forces with the intention of increasing the effort level for click-fraud spammers (Haddadi 2010). Mittal et al. investigated a possible way to find fraud by falsely manipulating forbidden information on customers (Mittal, Gupta et al. 2006).

Finally, Triepels et al. investigated whether intelligent fraud detection systems can improve the detection of document fraud by miscoding and smuggling by analysing large sets of historical shipment data. They developed a Bayesian network that predicts the presence of goods on the cargo list of shipments. The predictions of the Bayesian network were compared with the accompanying documentation of the shipment to determine whether document fraud was perpetrated (Triepels, Daniels et al. 2018).

None of the above literature talks about customs fraud. To the best of our knowledge, our field of application is therefore novel. We used two algorithms to find indicators of customs fraud (detection being a novel dimension for the phenomenon as well). They are based on the Apriori association rules and decision tree analyses.

2. Problem statement and Research question

All over the world some people and companies try to avoid customs whenever possible. Therefore, detecting customs fraud is important part to logistics companies, their customers and governments in a world of growing e-commerce.

Consequently, the following research question has to be answered: How can customs fraud be detected using data mining on logistics transaction data? Developing a methodology capable to find some indicators of customs fraud in such logistics transactions on the basis of data mining technology is the objective of this research.

3. Methodology

To answer our research question we start from a real life dataset, we will create a proxy for customs fraud as we will explain in the data section later. Three models will be applied on the data. The Apriori algorithm for association analysis and decision tree analysis will be used for this fraud detection research. We will use Apriori analysis as it is one of the most frequently used methods for generating association rules and searching patterns in large databases (John and Shaiba 2019, Silva, Varela et al. 2019). It is very fast and memory friendly when generating rules comparing to other association rule algorithms. Moreover, it can be used as a supervised method to discover interesting patterns and rules for selected variable (Prithiviraj and Dr.R 2015). Decision tree analysis will be used to verify whether its results could establish a connection between the association rules resulting from the apriori analysis. The confusion matrix will be used in this study for the evaluation by determining the classification measures: Accuracy, precision, recall and F1. In this paper, we will first explain the details of the association rules and the Apriori algorithm and the CHAID decision tree analysis. Second, we will follow the research process from data collection via preprocessing of the data to explanation of the analyses and results.

3.1 Apriori Association Rules

Association Rules are a data mining technique. Basic association analysis deals with the simultaneous occurrence of several items with one another whereas deeper association analysis can take into consideration the quantity of the joint occurrence and sequence of occurrence, etc. The method for finding association rules through data mining involves several steps (Linoff and Berry 2011, Kotu and Deshpande 2014). The sequence is the following:

- Prepare the data: an association algorithm needs input data to be formatted in a particular format.
- Short-list frequently occurring item sets. Item sets are combination of items. An association algorithm limits the analysis to the most frequently occurring items. The final and meaningful rules are extracted in the next step.
- Finally, the algorithm generates and filters the rules based on the interest measure (Kotu and Deshpande 2014).

All association rules algorithms try to find frequently occurring item sets within a base of possible item sets. **The Apriori algorithm** is the most frequently used. It leverages some simple logical principles on the item sets to reduce the number to be tested for a certain support measure (John and Shaiba 2019). The algorithm is based on the principle that ‘If an item set is frequent, then all its subset items will be frequent’ (Tan, Steinbach et al. 2005). The name of the algorithm is based on the fact that the algorithm uses prior knowledge of properties of frequent items (Kotu and Deshpande 2014).

Apriori association rules employ an iterative approach known as a level-wise search, where k- item sets are used to explore (k +1) item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy a minimum support. The resulting set is denoted by L1. Next, L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database (Kotu and Deshpande 2014).

3.2 The Chi-Squared Automatic Interaction Detection (CHAID) algorithm

Decision tree analysis is a data mining technique that encompasses several algorithms to predict a dependent variable (Qabbaah, Sharawi et al. 2017). These predictions are determined by the influence of independent predictor variables (Qabbaah., Sammour. et al. 2019). It is "a structure that can be used to divide a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. Decision trees "use a set of rules for dividing a large heterogeneous dataset into smaller, more homogeneous groups with respect to a particular target variable" (Linoff and Berry 2011). With each successive division, the members of the resulting sets become more and more similar to one another"(Linoff and Berry 2011).

Each procedure goes as follows. "After the Dataset is split into n nodes, the process is repeated on each node until the data are completely partitioned or until a stopping rule is met. This means that there are no longer enough cases to partition, or only one case remains for the last node. The entire classification of data is then visualized through a graphical tree, which explains the interaction of x (x1, x2, …, xm) and y" (L. and Jr 2012). The classification procedures will attempt to provide as much separation as possible with regards to classifying data into correct cases (Long, Griffith et al. 1993). Graphically, decision trees will produce a tree T which is composed of a root node and child nodes (De'ath 2000) (Ripley and Hjort 1995). The tree is essentially a connected graphic that is inverted. However, the tree display is user specific (Loh and Shih 1997, L. and Jr 2012).

There are several decision tree classification procedures. The algorithm used in this research is **Chi-Squared Automatic Interaction Detector (CHAID)**. CHAID algorithm is one of the most popular classification procedures (Loh and Shih 1997) because it can manage both continuous and categorical variables and handle a very large dataset with high speed performance (Díaz-Pérez and Bethencourt-Cejas 2016). CHAID uses Chi-square test for splitting

With regards to splitting, CHAID is not bound to binary splits and allows for multiple level splits at each node. The algorithm searches for the best possible split for the different values of the independent variable then determines if the data should be merged to form a node or split in the data. Essentially, the best predictor is selected to begin splitting the data. During this process, CHAID selects the best predictor based on comparisons of the adjusted p-value for each predictor (L. and Jr 2012). This is a three step process for determining splitting. The splitting and merging procedures for CHAID are independent of one another.

The algorithm uses as sequential process where splitting and merging occur at the same time. The basis behind this aspect of the algorithm is rooted in the computation time of T. CHAID has difficulty in determining when and where to stop (Loh and Shih 1997). Stopping is also a three steps process determined by the chi-square statistic. As such, stopping occurs when the critical level of the chi-square test fails to meet the test.

4 Data and Data preprocessing

The research framework of this paper is: data collection, preprocessing of the data, analysing and obtaining the results.

The data used in this research were obtained from an international logistics services company. Cleaning and merging tables of the data has been applied in order to obtain the final data set. The total number of transactions in the final dataset is equal to the size of the sample (n=8,243). Some important elements have to be mentioned that guided us in selecting the variables used in the analysis models.

The product type, weight, quantity, destination and country of origin are logically used. This is amongst others based on what the customs regulation in Gulf Cooperation Council states (GCCStates 2018). We will use COD Value USD and Customs Values USD to define the fraud proxy. Table 1 shows the description of the variables used in this research.

Table 1. The description of the data variables.

Variable	Data type	Data description
Variables in the original dataset		
ID	Integer	The ID of the order
Weight In KG	Double	The weight of shipment in KG
CODValueUSD	Double	The amount of cash on delivery in USD
Product Group name	String	Product category group name of the shipment.
Product group ID	Integer	Product category group ID
Customs Value USD	Double	The declared value of the goods for customs department in USD
Destination	String	Destination city of shipment
Origin Country	String	Country of origin of the shipment
DestCountry	String	Country of destination of the shipment
ShipperID	Integer	The ID of the E-commerce companies
Created variables		
COD Customs amount USD	Double	The customs value of the goods measured using COD Value USD
Customs amount USD	Double	The customs value of the goods measured using the declared customs value USD
Weight K-mean class	Integer	Transformation classes of the weight using k-means
Weight CHAID class	Integer	Transformation classes of the weight using CHAID decision tree
Value USD K-mean class	Integer	Transformation classes of the COD value of using k-means
Value USD CHAID class	Integer	Transformation classes of the COD value using CHAID decision tree
Customs Fraud class	Boolean	Indicators of the fraud. Class:Yes(1),No(0)

The variables 'COD Customs amount USD' and 'Customs amount USD' have been calculated using the import duties of the products multiplied by the goods value (COD or the declared value to customs department) taking into consideration whether the shipment was sent within the same country or not. The variable 'Customs Fraud class' has been created by dividing the value of customs paid using the declared value of the goods to customs department ('Customs amount USD') over the new calculated customs value using the COD value which is the real amount of the goods that the customer should pay ('COD Customs amount USD'). If the answer is less than 0.8 (indication of fraud) or greater than 1.25 (overpayment, not matched in our data) then we identified the Fraud class as 1 (there is fraud), otherwise the Fraud class was identified as 0 (no fraud).

All the variables should be nominal or categorical as an input for the Apriori association rules. Therefore, converting the continuous variables such as weight and price to nominal is essential. This transformation process will be performed using clustering and decision tree analysis as follows:

1. **Transformation using k-means clustering:** The main objective of K-means is to partition the dataset into k clusters in which each instance belongs to the cluster with the nearest mean (Qabbaah, Sammour et al. 2019).

   To transform the continuous variables taking into consideration 'Fraud Class' which is our goal in this research, we created a k-means cluster model with three attributes, 'Fraud class', 'weight in KG' and 'COD Value USD'. Table 2 shows the results. The variables 'Weight K-mean class' and 'Value USD K-mean class' have been created for each transaction.

Table 2. Transformation results using K-means clustering.

Clusters	Avg.Weight In Kg	Avg.COD Value USD	Fraud class	Number of Items
Cluster 1	0.79383	72.166	1	551
Cluster 2	2.0341	131.96	1	778
Cluster 3	6.891	158.02	1	258

2.Transformation using decision tree: To transform the continuous variables taking into consideration 'Fraud Class' which is our goal in this research, we created two CHAID decision tree models. The dependent variable of the first model is 'Fraud class' and the independent variable is 'Weight in KG', whereas the dependent variable of the second model is 'Fraud class' and the independent variable is 'COD Value USD'. Table 3 shows the categories of 'Weight in kg' variable and Table 4 shows the categories of 'COD Value USD' variable.

Table 3. 'Weight in kg' categories using CHAID decision tree analysis.

Class	Value	Class	Value
1	<= 0.230	6	(1.640 , 2.450]
2	(0.230 , 0.490]	7	(2.450 , 5.380]
3	(0.490 , 0.750]	8	(5.380 , 8.130]
4	(0.750 , 1.120]	9	(8.130 , 19.780]
5	(1.120 , 1.640]	10	>19.780

Table 4. 'COD Value USD' categories using CHAID decision tree analysis.

Class	Value	Class	Value
1	<= 39.5790	5	(134.702 , 168.061]
2	(39.579 , 98.397]	6	(168.061 , 263.188]
3	(98.397 , 117.657]	7	>263.188
4	(117.657 , 134.702]		

5 Analyses and Results

After applying the pre-processing phase, we are ready to use the data for further analyses. In this section, we will first present the results of the Apriori association rules analysis. Next we will show the results of the CHAID Decision tree analysis.

5.1 The analyses and the results of customs fraud detection apriori models (models 1 and 2)

Two models will be created for this fraud detection. Model-1 uses K-means transformation and model-2 uses CHAID transformation.

1. The variables used in **model-1** are (weight class k-mean, COD value class k-mean, product group name, destination, origin country, shipper-id, fraud class). Table 5 shows the Apriori run information for the first model using K-means transformation.

Table 5. Apriori run information for the first model using K-means transformation.

Instances	Minimum support	Minimum metric <confidence>
8234	0.01	0.1

Table 6 shows the summary of the important rules found as a result of the Apriori association rules algorithm using K-means transformation for the continuous variables.

Table 6. Summary of the important rules found when detecting customs fraud with model-1 using k-means transformation for the continuous variables.

N	The Rules
1	Weight class k-mean=2, COD class k-means=3, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 224 ==> _Fraud class=1. 224 cases, confidence:(1)
2	Weight class k-mean=2, COD class k-means=2, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 178 ==> _Fraud class=1. 178 cases, confidence:(1)
3	Weight class k-mean=2, COD class k-means=3, Destination=JED, Origin Country=HK, Shipper ID=599818. 168 ==> _Fraud class=1. 168 cases, confidence:(1)

N	The Rules
4	COD class k-means=3, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 137 ==> _Fraud class=1. 137 cases, confidence:(1)
5	Weight class k-mean=2, COD class k-means=2, Product Group Name=Apparel, Destination=JED, Origin Country=HK Shipper ID=599818. 113 ==> _Fraud class=1. 113 cases, confidence:(1)
6	Weight class k-mean=2, Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 115 ==> _Fraud class=1. 104 cases, confidence:(0.9)
7	Weight class k-mean=2, Destination=DHA, Origin Country=HK, Shipper ID=599818. 113 ==> _Fraud class=1. 101 cases, confidence:(0.89)
8	Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 342 ==> _Fraud class=1. 126 cases, confidence:(0.37)
9	Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 1274 ==> _Fraud class=1. 607 cases, confidence:(0.48)
10	Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 312 ==> _Fraud class=1. 133 cases, confidence:(0.43)
11	Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 1094 ==> _Fraud class=1. 453 cases, confidence:(0.41)
12	Weight class k-mean=2, Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 129 ==> _Fraud class=1. 108 cases, confidence:(0.84)
13	Product Group Name=Apparel, Destination=DHA, Origin Country=HK, Shipper ID=599818. 272 ==> _Fraud class=1. 118 cases, confidence:(0.43)
14	Weight class k-mean=1, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 779 ==> _Fraud class=1. 169 cases, confidence:(0.22)

N	The Rules
15	Weight class k-mean=1, COD class k-means=1, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 751 ==> _Fraud class=1. 151 cases, confidence:(0.2)
16	Weight class k-mean=1, COD class k-means=1, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 741 ==> _Fraud class=1. 131 cases, confidence:(0.18)
17	Weight class k-mean=1, Product Group Name=Beauty Supplies, Destination=JED. 106 ==> _Fraud class=1. 15 confidence:(0.14)

As an example, rule number one reads as follows: If the shipment sent from Hong Kong to Riyadh and bought via e-commerce-ID=599818 and the product category was 'Apparel' with price class= 3, and weight class=2 as shown in Table 6.15, then 224 transactions were indicated as a customs fraud case. All rules have to be interpreted in this way. So, rule number 8 reads as: If the origin country of the shipment is 'Hong Kong' and bought via e-commerce-ID=599818 and the product category was 'Bag/case', then 126 transactions are indicated as customs fraud cases.

2. The variables used in **model-2** are (weight class CHAID, COD value class CHAID, product group name, destination, origin country, shipper-id, fraud class). Table 7 shows the Apriori run information for the second model using CHAID transformation.

Table 7. Apriori run information for the second model using CHAID transformation.

Instances	Minimum support	Minimum metric <confidence>
8234	0.01	0.1

Table 8 shows the summary of the important rules found in model-2 as a result of the Apriori association rules algorithm using CHAID transformation for the continuous variables.

Table 8. Summary of the important rules found when detecting customs fraud with model-2 using CHAID transformation for the continuous variables.

N	The Rules
1	COD value class Chaid=5, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 135 ==> _Fraud class=1. 135 cases, confidence:(1)
2	COD value class Chaid=6, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 94 ==> _Fraud class=1. 94 cases, confidence:(1)
3	Weight class Chaid=6 COD, value class Chaid=5, Origin Country=HK, Shipper ID=599818. 93 ==> _Fraud class=1. 91 cases, confidence:(0.98)
4	Weight class Chaid=7, COD value class Chaid=5, Origin Country=HK, Shipper ID=599818, 99 ==> _Fraud class=1. 91 cases, confidence:(0.92)
5	COD value class Chaid=5, Destination=JED, Origin Country=HK, Shipper ID=599818. 112 ==> _Fraud class=1. 102 cases, confidence:(0.91)
6	Weight class Chaid=8, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 105 ==> _Fraud class=1. 88 cases, confidence:(0.84)
7	Weight class Chaid=7, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 166 ==> _Fraud class=1. 120 cases, confidence:(0.72)

N	The Rules
8	COD value class Chaid=4, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 229 ==> _Fraud class=1. 151 cases, confidence:(0.66)
9	COD value class Chaid=4, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 159 ==> _Fraud class=1. 103 cases, confidence:(0.65)
10	Weight class Chaid=6 COD, value class Chaid=4, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 193 ==> _Fraud class=1. 121 cases, confidence:(0.63)
11	Weight class Chaid=6, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 201 ==> _Fraud class=1. 114 cases, confidence:(0.57)
12	Weight class Chaid=6, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 264 ==> _Fraud class=1. 146 cases, confidence:(0.55)
13	Weight class Chaid=5, COD value class Chaid=3, Origin Country=HK, Shipper ID=599818. 171 ==> _Fraud class=1. 84 cases, confidence:(0.49)
14	Weight class Chaid=5, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 244 ==> _Fraud class=1. 119 cases, confidence:(0.49)

N	The Rules
15	Weight class chaid=6, COD value class chaid=4, Origin Country=HK, Shipper ID=599818. 245 ==> _Fraud class=1. 157 cases, confidence:(0.64)
16	COD value class Chaid=3, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 208 ==> _Fraud class=1. 95 cases, confidence:(0.46)
17	Product Group Name=Apparel, Destination=DHA, Origin Country=HK, Shipper ID=599818. 272 ==> _Fraud class=1. 118 cases, confidence:(0.43)
18	Product Group Name=Shoes, Origin Country=HK, Shipper ID=599818. 312 ==> _Fraud class=1. 133 cases, confidence:(0.43)
19	COD value class Chaid=3, Destination=JED, Origin Country=HK ,Shipper ID=599818. 214 ==> _Fraud class=1. 91 cases, confidence:(0.43)
20	Weight class Chaid=6, COD value class Chaid=3, Origin Country=HK, Shipper ID=599818. 208 ==> _Fraud class=1. 85 cases, confidence:(0.41)
21	Product Group Name=Bag/Case, Origin Country=HK, Shipper ID=599818. 342 ==> _Fraud class=1. 126 cases, confidence:(0.37)
22	Weight class Chaid=5, Destination=JED, Origin Country=HK, Shipper ID=599818. 290 ==> _Fraud class=1. 101 cases, confidence:(0.35)

N	The Rules
23	Weight class Chaid=2 COD, value class Chaid=2, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 314 ==> _Fraud class=1. 104 cases, confidence:(0.33)
24	Weight class Chaid=4, Product Group Name=Apparel, Origin Country=HK, Shipper ID=599818. 442 ==> _Fraud class=1. 127 cases, confidence:(0.29)
25	COD value class Chaid=2, Product Group Name=Apparel, Destination=JED, Origin Country=HK, Shipper ID=599818. 643 ==> _Fraud class=1. 169 cases, confidence:(0.26)
26	COD value class Chaid=2, Product Group Name=Apparel, Destination=RUH, Origin Country=HK, Shipper ID=599818. 629 ==> _Fraud class=1. 161 cases, confidence:(0.26)
27	Weight class Chaid=5, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 427 ==> _Fraud class=1. 93 cases, confidence:(0.22)
28	Weight class Chaid=4, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 514 ==> _Fraud class=1. 109 cases, confidence:(0.21)
29	Weight class Chaid=3, COD value class Chaid=2, Origin Country=HK, Shipper ID=599818. 407 ==> _Fraud class=1. 85 cases, confidence:(0.21)

Using this transformation, rule number 15 looks for instance as follows: If the origin country of the shipment is Hong Kong and bought via e-commerce-ID=599818 with a price value is between (117.657, 134.702] and the weight value between (1.640 , 2.450] then 157 transactions are indicated as customs fraud cases

Rule number 27 can be explained as follows: If the origin country of the shipment is Hong Kong and bought via e-commerce-ID=599818 and the weight value being between (1.120 , 1.640] and the price value being between (39.579 , 98.397] then 93 transactions are indicated as a customs fraud case

When we look at all the decision rules mentioned in both models, we can notice that most of the cases of customs fraud have some common characteristics. The origin of the shipments is Hong Kong, the customers were mostly customers of the same e-commerce-ID company (ID=599818) and the cases involved different product categories, but mostly ‘Apparel’ of different values. A practical conclusion for the forwarding logistics company might consequently be to look more carefully at who is operating the e-commerce site and maybe train the people there to understand the customs rules better and to apply them more carefully. Moreover improvement can be monitored.

5.2 The results of customs fraud decision tree model (model-3)

The goal of this model (**model-3**) is to try to detect customs fraud using CHAID decision tree analysis and verify whether the results of the two previous models using association rules actually are similar to this model and in how far the results support each other.

The dependent variable in this model is Fraud class. The independent variables are (Weight In KG, Product Group Name, COD Value USD, Origin Country, Shipper ID). Table 9 presents the summary of Model-3.

The total number of transactions present in each node is based on Fraud class (category ‘Yes-indication of Fraud’ or category ‘No-Not Fraud). The total number of nodes in the model is 24, while the number of terminal nodes is 17. Figure 1 shows the tree map of the Fraud model created.

We can conclude for instance that if the shipper ID is ‘599148’ and weight in Kg <=0.230, there are 14 cases indicated as a fraud (node 5 explained). If the shipper ID is ‘599818’ and the values of ‘COD Value USD’ are between (117.6-134.7) and the product group name are ‘Beauty supplies’ or ‘Apparel’, then there are 288 cases indicated as a fraud (node 19 explained).

We can draw some conclusions about the similarity between the rules conducted from association rules models and the decision tree model.

The rules of model-3 are in general somewhat similar to the rules conducted by the Apriori analysis. Some examples of this similarity are the following. Most of the cases were for instance attributed to Shipper ID= 599818 and occurred with the product groups ‘Apparel’, ‘Shoes’ and ‘Bag/Case’. Moreover, ‘weight in kg’ and ‘COD’ variables show some similar values in the nodes of the CHAID model and the association rules. For instance, we can see the occurrence of the values 0.231, 0.49, 1.12 for the ‘weight in kg’ as variables in model-2 and model-3, and the values 98.39 and 134.7 for the ‘COD’ variable in both models as well. The transformation in model-1 was performed using k-means as shown in table One, the splitting of ‘weigh in kg’ and ‘COD value USD’ variables are different (in total 3 classes), but at the same time we can also see values of the ‘COD’ variable close to one another: the average of the first class in model-1 was 72.1 and the average of the second class was 131.9, values relatively close to the values found by model-3 (72.2 and 134.7). The rules performed by Apriori models (one and two) are more comprehensive than the results of the CHAID decision tree model. The decision tree model (model-3) splits the independent variables using the variable that has the most significant effect on the dependent variable one by one and sequentially.

Therefore, the results of Model-3 confirm our results in model-1 and model-2, but the rules extracted of the Apriori models are more comprehensive and are therefore capable of detecting fraud in deeper detail.

Table 9. Fraud class decision tree model summary.

Model Summary		
Specifications	**Growing Method**	CHAID
	Dependent Variable	Fraud class
	Independent Variables	Weight In KG, COD Value USD, Product Group Name, Origin Country, Destination, Shipper ID
	Validation	Cross Validation, n=25
Results	**Independent Variables Included**	Shipper ID, COD Value USD, Weight In KG, Product Group Name, Destination
	Number of Nodes	24
	Number of Terminal Nodes	17

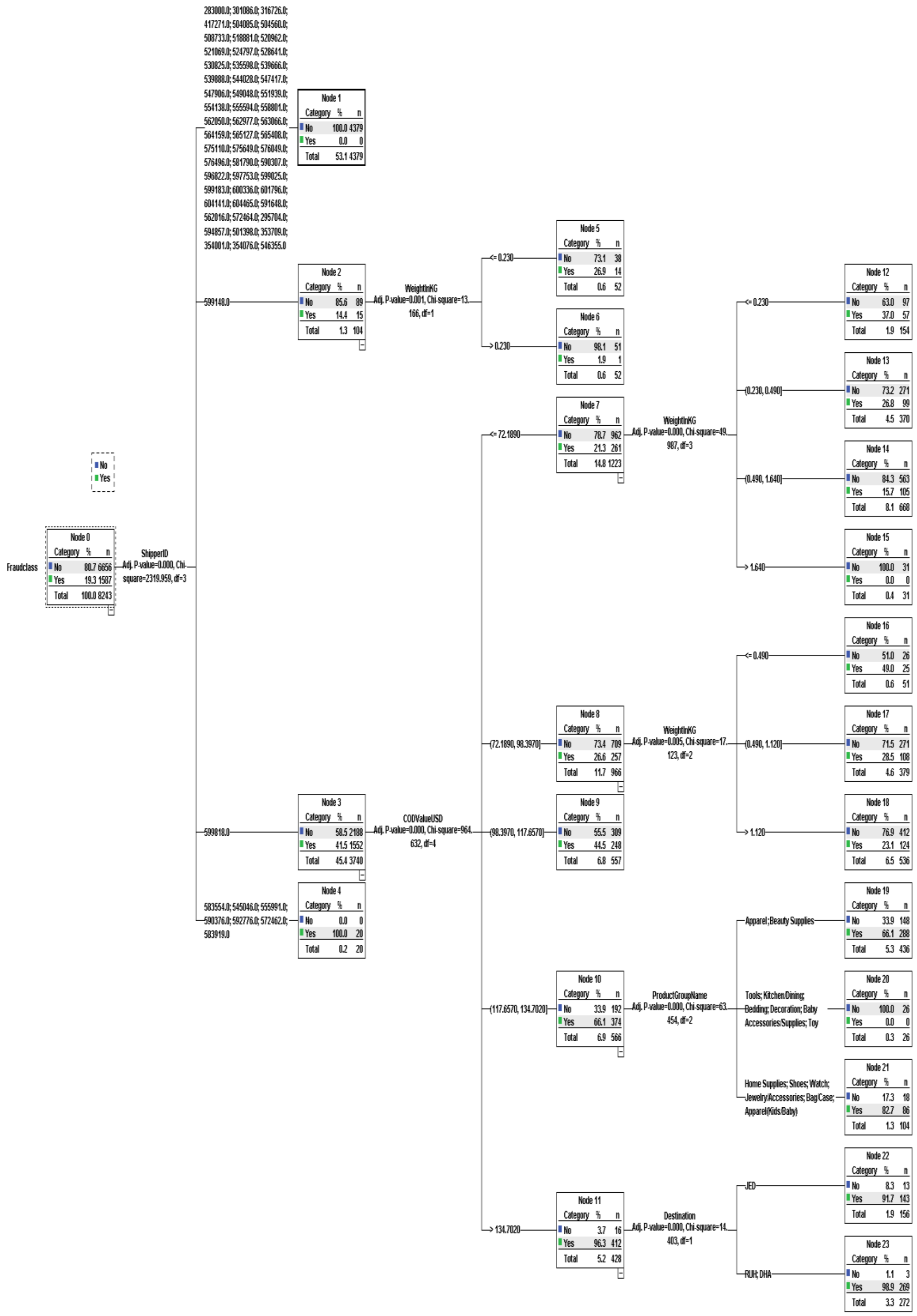


Figure 1. CHAID decision tree for Model 3

- **Evaluation the Customs fraud CHAID model using classification measures tests**

To evaluate the quality of the CHAID customs Fraud model, the classification measures (accuracy, recall, precision and f-score) have been used on the test data based on the confusion matrix.

Figure 2 shows the classification results on the test data of the CHAID customs Fraud model.

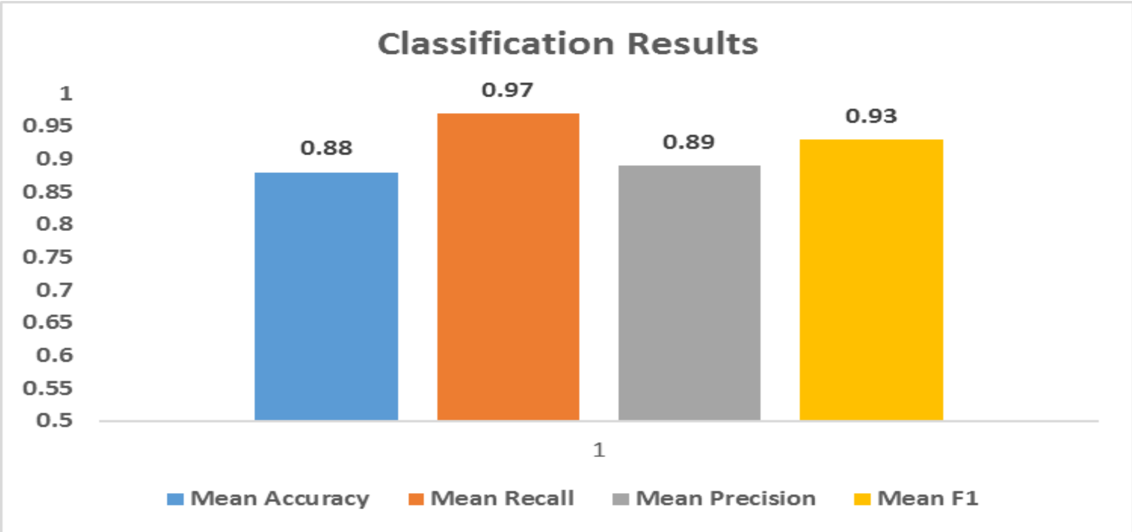


Figure 2: Classification results of the CHAID customs fraud model.

We can conclude that our CHAID model has a high average percentage of accuracy with 88%. The average of the recall, precision and f-score also have a high percentage with 97%, 89% and 93% respectively.

5.3 Discussion of the results of the custom fraud detection models

We have used the Apriori association rules and decision tree analysis on a partial section of our data base to answer the research question whether and how we can find indications of customs fraud using the data base only. After transformation of some of the continuous data using both k-means clustering and CHAID decision tree analysis, we have indeed succeeded in finding rules that allow to check whether there are cases of customs fraud in the data base. Some of these rules have been explained in detail.

This is highly significant for the users of the data base. Detecting customs fraud is indeed an important element in getting online business shipment systems within the realm of completely honest and trustworthy business on a global scale. This is important to all parties involved: forwarders, customers and governments. Forwarders like the company providing our database can more reliably offer fraud-proof systems to their customers, customers can faithfully rely on the forwarders that they will not be confronted with government claims afterwards and given the actual rules of customs application, and governments can trust forwarders more in not avoiding customs as fraud cases are detected before.

Our detection method is capable of doing so. To the best of our knowledge, it proves to be the first research method that detects fraud and proves to perform well. Moreover, the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, thus corroborating our research conclusions. Consequently this research constitutes an important contribution to both the scientific literature on data mining as a technique to find indication of customs fraud and to the practice of improving e-commerce behaviour.

6. Conclusion

In this section the major results of our different analyses are related to the research question will be presented.

Our research question read as follows: **‘How can customs fraud be detected using data mining on logistics transaction data?’**

The objective was to find evidence of potential customs fraud on different logistics trajectories based on the variables in the data base such as product categories and logistics routes.

This objective was achieved using both the Apriori algorithm for association rules and decision tree analysis. First we had to transform the continuous variables using both k-means clustering and CHAID decision tree analysis. This was done for the weight and price variables. Analysis of the rules detected in our analysis indicates that it is possible via this methodology to find indicators of customs fraud cases. Moreover, it was possible to describe in detail the situation in which the fraud occurs, indicating e-commerce customers (ID), types of goods (mostly apparel in our data base) and country of origin (mostly Hong Kong in our database).

Moreover, the results of the decision tree analysis proved to verify the connection between both Apriori models used and showed similar results, which confirm the value of our research conclusions.

This should allow logistics companies to offer a customs fraud service to their customers-partners, causing these customers and also government instances less hassle. As it is to our knowledge the first data mining technique capable of finding indications of customs fraud accurately, our research has a significant contribution to both data mining research and the logistics world.

Bibliography

ACFE (2018). REPORT TO THE NATIONS, 2018. GLOBAL STUDY ON OCCUPATIONAL FRAUD AND ABUSE.

De'ath, G. (2000). "Classification and regression trees: a powerful yet simple technique for ecological data analysis." Ecology v. 81(no. 11): pp. 3178-3192-2000 v.3181 no.3111.

Díaz-Pérez, F. M. and M. Bethencourt-Cejas (2016). "CHAID algorithm as an appropriate analytical method for tourism market segmentation." Journal of Destination Marketing & Management 5(3): 275-282.

GCCStates. (2018). "Common Customs Law for GCC States ", from https://www.fca.gov.ae/en/homerightmenu/ pages/gccstates.aspx?SelectedTab=14.

Haddadi, H. (2010). "Fighting online click-fraud using bluff ads." SIGCOMM Comput. Commun. Rev. 40(2): 21-25.

John, M. and H. Shaiba (2019). "Apriori-Based Algorithm for Dubai Road Accident Analysis." Procedia Computer Science 163: 218-227.

Kotu, V. and B. Deshpande (2014). Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner, Morgan Kaufmann Publishers Inc.

L., S. and B. Jr (2012). AN EMPIRICAL COMPARISON OF LOGISTIC REGRESSION TO DECISION TREE INDUCTION IN THE PREDICTION OF INTIMATE PARTNER VIOLENCE REASSAULT. Doctor of Philosophy, Indiana University of Pennsylvania.

Linoff, G. S. and M. J. A. Berry (2011). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley.

Loh, W.-Y. and Y.-S. Shih (1997). "SPLIT SELECTION METHODS FOR CLASSIFICATION TREES." Statistica Sinica 7(4): 815-840.

Long, W. J., J. L. Griffith, H. P. Selker and R. B. D'Agostino (1993). "A Comparison of Logistic Regression to Decision-Tree Induction in a Medical Domain." Computers and Biomedical Research 26(1): 74-97.

Metwally, A., D. Agrawal, A. E. Abbad and Q. Zheng (2007). On Hit Inflation Techniques and Detection in Streams of Web Advertising Networks. 27th International Conference on Distributed Computing Systems (ICDCS '07).

Mittal, S., R. Gupta, M. Mohania, S. K. Gupta, M. Iwaihara and T. Dillon (2006). Detecting Frauds in Online Advertising Systems, Berlin, Heidelberg, Springer Berlin Heidelberg.

Prithiviraj, P. and P. Dr.R (2015). "A Comparative Analysis of Association Rule Mining Algorithms in Data Mining: A Study."

Qabbaah, H., G. Sammour and K. Vanhoof. (2019). Using K-Means Clustering and Data Visualization for Monetizing logistics Data. 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE.

Qabbaah, H., L. Sharawi, G. Sammour and K. Vanhoof (2017). Development of Customs Prediction Model for Online Ordering. 2017 International Conference on New Trends in Computing Sciences (ICTCS), IEEE.

Qabbaah., H., G. Sammour. and K. Vanhoof. (2019). "DECISION TREE ANALYSIS TO IMPROVE E-MAIL MARKETING CAMPAIGNS." International Journal “Information Theories and Applications” 26(1): 303-330.

Ripley, B. D. and N. L. Hjort (1995). Pattern Recognition and Neural Networks, Cambridge University Press.

Silva, J., N. Varela, L. A. Borrero López and R. H. Rojas Millán (2019). "Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm." Procedia Computer Science 151: 1207-1212.

Tan, P. N., M. Steinbach and V. Kumar (2005). Association analysis: Basic concepts and algorithms.

Triepels, R., H. Daniels and A. Feelders (2018). "Data-driven fraud detection in international shipping." Expert Systems with Applications 99: 193-202.

Tripathi, D., B. Nigam and D. R. Edla (2017). "A Novel Web Fraud Detection Technique using Association Rule Mining." Procedia Computer Science 115: 274-281.

Zhang, L. and Y. Guan (2008). Detecting Click Fraud in Pay-Per-Click Streams of Online Advertising Networks. 2008 The 28th International Conference on Distributed Computing Systems.

Authors' Information

Hamzah Qabbaah *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B6b, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium*

e-mail: hamzah.qabbaah@uhasselt.be

Major Fields of Scientific Research: Data mining, Digital marketing, Business Intelligence, Machine Learning, Knowledge Management

George Sammour *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan.*

e-mail: George.Sammour@psut.edu.jo

Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development

Koen Vanhoof *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium*

e-mail:koen.vanhoof@uhasselt.be

Major Fields of Scientific Research: data mining, knowledge retrieval

SUBSTANTIVE FORMULATION OF UAV GROUPS ROUTES OPTIMIZATION PROBLEM

Maksym Ogurtsov, Vyacheslav Korolyov

Abstract: *Submissions of several specific problems of optimizing UAV routes for individual and group use are provided. The object of the study is the problems of combinatorial optimization of routes for UAVs and their groups that collect and transmit information to control points. General formulation of the target search problem in the area was given. An analysis of how to evaluate the quality of the search and what is the criteria for its effectiveness was conducted. The formal statement and mathematical model of the UAV routing problem are presented, which deals with the stages of the operation planning and the visiting the set targets by UAVs at the same time. Algorithms, based on this mathematical model can become the basis for solving several complex technical problems: increasing the tactical and technical (minimizing the territory exploring time by the UAV group) plus technical and economic indicators (minimizing the UAVs amount) while using UAVs groups. In this case, the UAVs would act as a team, which is tasked with examining or servicing several objects. As a result of the performed works, the routing problem meaningful statement for the UAV radio network with and without retransmission of control signals from the ground control center and the possibility of the UAV autonomous movement on the part of the route has been further developed.*

Keywords: *UAV routes, combinatorial optimization, routing optimization.*

Introduction

In recent years, we have a clear trend for the unmanned vehicles usage expansion, in particular, UAVs. The worldwide experience of the UAV development, especially in such high-technology countries as the USA, Germany and Israel, shows that UAVs in the next 10-15 years will be able to perform most of the problems that are now being tackled by manned means. Already, UAVs are used to control the technical condition, safety and operation of various objects and systems, in particular in ecology, agriculture and forestry, in railway transport, and the organization of marine search and rescue operations [Ponda et al, 2015], [Austin, 2011].

The rapid expansion of UAV applications in recent years has shaped the need for new classes of problems.

Formulation of the Problem

Mathematical models and combinatorial optimization methods development for real-time problems is an actual problem, as evidenced by a large number of specialized monographs, journals and conferences, which is conditioned both by practical needs and the complexity of theoretical problems that arise. At the same time, the problem is not only finding solutions, but also saving limited resources – fuel, computing resources, time, etc. This determines the need to develop software-algorithmic tools for the creation of appropriate information technology to save UAV resources and ensure optimal routing.

This article is aimed at the development of new mathematical models and algorithms oriented to use in the construction and operation of specialized networks, the nodes of which will be UAVs, stationary and mobile ground control points of different levels.

Analysis of Recent Research and Publications

To construct criteria for optimizing UAVs operations the estimates of the search efficiency used in the theory of operations were proposed in monographs [Jaiswal, 2012], [Golden et al, 2008].

The problem of finding objects during air surveillance and research began in the 1940s based on US business performance measurement approaches, which later became part of the theory of operations research [Ventzel, 1964]. Accumulated data on combat actions allowed to highlight critical parameters for assessing the outcome of the operation, formulate performance criteria, and perform a meaningful statement of optimization problems. The solution of the problem allowed to increase the effectiveness of their own sentinel devices in time, reduce their number, identify problems in their distribution in the areas of responsibility, etc.

The listed approaches to the search organization mostly did not include the use of a group of UAVs integrated into the network to perform search operations [Kumari et al, 2020], [Ye et al, 2020], [Корольов et al, 2017], [Корольов & Огурцов, 2017], [Огурцов & Ходзінський, 2016], [Поліновський et al, 2017], [Гуляницький & Огурцов, 2020], so this remain an actual scientific problem.

The tasks of route optimization of both single UAVs and their groups ([Coutinho at al, 2018], [Horbulin at al 2020]) belong to NP-complete problems, therefore exact algorithms cannot be used in practice. This applies to even the most simple and researched tasks, when you want to find the optimal route of the single UAV between two points: start and finish (see, for example, [Zhao at al, 2018] or [Otto at al, 2018]). Researchers have proposed a number of metaheuristic methods, including ant colony optimization algorithms ([Cekmez at al, 2017], [Perez-Carabaza at al, 2018] and [Гуляницький & Рибальченко, 2018]), genetic ([Chiang at al, 2019], [Binol at al, 2018]), and artificial bee colony algorithms ([Xu at al, 2010], [Tian at al, 2018]) and particle flow optimization [Shakhatreh at al, 2017].

Moreover, additional restrictive conditions should often be taken into account (for example, flight range without recharging or refueling, load capacity, minimum turning radius, weather conditions, etc.), as it was described in [Ponda et al, 2015], [Austin, 2011], [Tsourdos at al, 2010], [Shima and Rasmussen, 2009].

In some cases, the formulation of such problems is similar to the known class of vehicle routing problems (VRP) [Toth and Vigo, 2014] [Braekers at al, 2016], in particular, VRP with several depots [Karakatič and Podgorelec, 2015], [Soto at al, 2017].

The Goal of the Work

The goal of the work – automation of UAV groups flight routes building and relaying signals between them, which allows us to increase the radius of effective action. For example, if the effective range of one UAV is 8-10 km, then using a network of several UAVs in the relay mode, you can increase the area of reconnaissance and exploration at least three times.

Research methodology – combinatorial optimization theory, combinatorial analysis.

UAV Group Operations

The authors in previous studies [Корольов et al, 2017], [Корольов & Огурцов, 2017], [Огурцов & Ходзінський, 2016], [Поліновський et al, 2017], [Гуляницький & Огурцов, 2020] considered some routes optimization problems for ground and aviation robotized systems. Below is a more detailed discussion of the management of a UAV group that is associated with the functioning of a group as a team that performs tasks with the ability to apply relay control signals.

The coordination of UAV groups flights, which execute a territory survey, the creation of images (in particular, stereo images) will allow us to control the situation. When planning a territory survey for UAVs, two types of problems can be distinguished, each of which has certain peculiarities that are important for a solution.

The general formulation of the target search problem in the area

The search problem is to scan a specific territory for an optimal way to identify objects [Корольов et al, 2017], [Поліновський et al, 2017]. Let’s assume that the region is divided into lesser areas or cells (not necessarily identical), but the probability that the target is known in the i-th area, where (i= 1, ..., N). According to the problem formulation, it is necessary to find such a division of the total amount of time t at time intervals Δt for each district in such a way as to maximize the probability of identifying the target. The problem has an analytical solution: in each region the probability of detecting a target has a uniform distribution [Jaiswal, 2012]. This is based on the assumption that the size of the area is large compared with the sector maximally explored by the UAV sensor in a stable position. The probability of detecting a target, given that it is present in area *i*, is:

$$P_i=1-\exp(-\Delta t_i/T_i),$$

where T_i is the average time to detect object in the explored area.

Then the discrete optimization problem for finding the object in the area is formulated as follows [Jaiswal, 2012]:

$$\max P=\Sigma p_i(1-\exp(-\Delta t_i/T_i))$$
$$\Sigma\Delta t_i=t,\ \Delta t_i\geq 0$$

Similar statements do not take into account the possibility of using a UAV group that will simultaneously search for an object. Also, to solve this problem, a search may be considered as optimal, and then look for spatial distribution parameters inside the group [Jaiswal, 2012].

Analysis of search quality evaluation

The problem of objects finding from the air began to be resolved during the 1940s based on the approaches to determining the process effectiveness in the United States, which later became part of the theory of operations research [Ventzel, 1964]. The accumulated data allowed us to allocate critical parameters to assess the outcome of the operation, to formulate the effectiveness criteria and to perform a meaningful statement of optimization problems. The problem solution allowed to increase the effectiveness of their sentinel agents in time, reduce their number, identify problems in their distribution in the areas of responsibility, etc.

Search efficiency criteria. Such criteria can be used to calculate the input data of the routing problem for the UAV group.

Relative search speed. Relative search speed considering the estimation of search efficiency used in the theory of operations to construct criteria for optimizing UAV search and patrolling operations [Ventzel, 1964]. The efficiency measure is introduced as the ratio of the real to theoretical achievable search speeds:

$$Z_E=V_p/V_m$$

This is, in essence, the ratio of the average density of objects we are searching for in the search area to the amount of their detection by UAVs per unit time.

Real search speed in the area. Real search speed in the area may be calculated according to the next equation:

$$V_p=(K/T)/(O/S)$$

where K is the number of detected objects, T is the period of time for which these objects were found, O – the potential number of objects we are searching for in the area, S is the cell size of the search territory.

The theoretical speed of UAV searches:

$$V_m=L\cdot V$$

where L is the effective range of the object detection by UAVs, V – the UAV movement speed.

The ratio determines the productivity of using UAV tools. A comparison of real and theoretical search efficiency is a way of tracking the performance of search operations. This measure of effectiveness is not prize-winning for immediate use during district intelligence by the UAV group, since it uses a lot of a posteriori information; in the examples of its application, the indicator is measured during the month [Ventzel, 1964].

The relative area of the surveyed territory. As the criterion of the object search effectiveness in a given territory, the average value of the explored area with the given confidence until the specified period of time can also be used [Ventzel, 1964].

The effectiveness of UAV objects search is determined as follows:

$$Z_{OS}=(\Delta S/s)\cdot p$$

where ΔS – area, explored by one UAV, s – the total area of the UAV search at a single moment, p – probability of the successful task completion. The purpose of the UAV's search is based on the analysis of independent events: the correct recognition of the target object, obstacles (bad weather etc.) absence; therefore, the probability of fulfilling the task will be their product: the probability of correct recognition of the target object p_{or} and the likelihood of avoiding the influence of the obstacle p_{oi}. So, the stability of task completion:

$$p = p_{or}\, p_{oi}$$

For a group of N UAV, which facing, for example, M bad weather condition, interfering with successful control by the operator, the Bernoulli experiments repeat theorem to determine the probability of operation success will work [Корольов & Огурцов, 2017]:

$$P_{BW\ M}=(C_N)^M W^M \cdot (1-W)^{N-M}$$

where W – the probability of losing control over UAV due to bad weather conditions. Since the UAVs mostly have a short range of activities, they will fall into the zone of the same bad weather condition, respectively, the probability of stalling control will be one-sided.

Since events such as the UAV stability control and losing this control are mutually exclusive, then:

$$p_{oi}=1-P_{BW\ M}$$

If N UAVs are involved in the search territory problem, then if UAVs detect objects independently of each other, the trajectories of the survey intersect, duplication of data is not eliminated, the effectiveness of objects search process for UAVs will be:

$$Z_{OS} =(1/S)\cdot\Sigma\Delta S\cdot(1-(1-p)^k)$$

where ΔS – the area surveyed by one UAV at a single moment of time. So, the formulas above can be used to plan the object search.

When searching for an object according to the optimal plan, the distance between the UAVs during the search process should be at least half of the target detection range [Корольов & Огурцов, 2017]:

$$B=\Sigma d_i+2L$$

where B is the width of the observation area, d_i – the distance between the UAVs during the search process, L – the maximum target range. The sum is counting according to the number of UAVs.

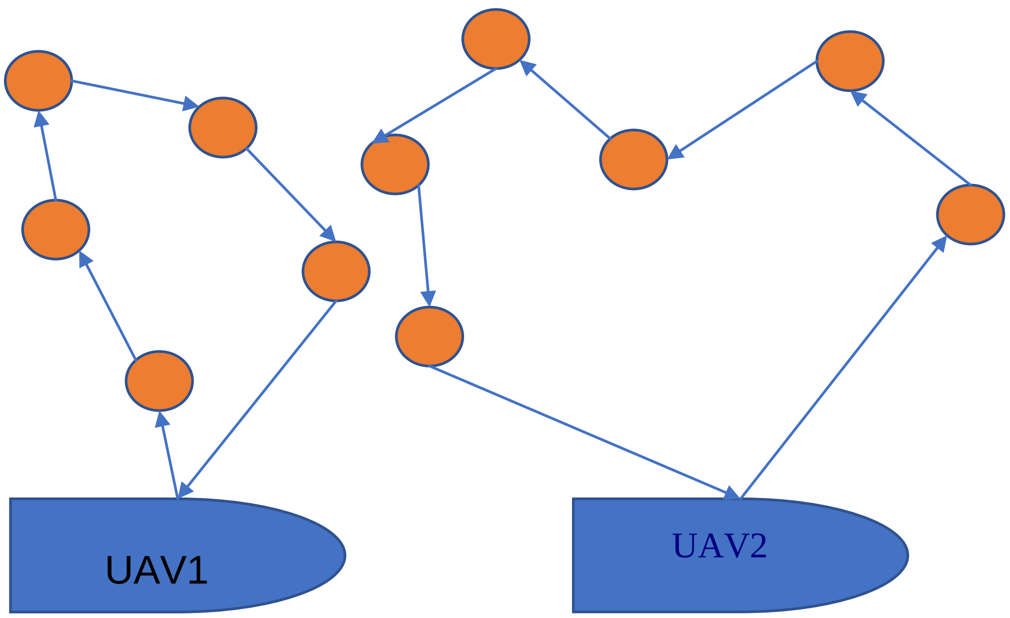
Observance of the distance value between the UAV group members by the formula above prevents the object from being skipped without re-registering the image of the surface. The given dependencies allow us to obtain initial conditions and parameters for the meaningful statement of the combinatorial optimization problem. The listed approaches to the search organization did not include the use of a group of UAVs integrated into the network to perform search operations [Корольов et al, 2017], [Поліновський et al, 2017].

The problem of a given territory general study

According to the problem terms, we have a certain territory, the area of which must be fully investigated to identify possible objects of interest. In this case, the territory may include predefined points of special interest that must necessarily be visited. Another priority when performing this problem is to cover given territory with existing UAVs as much as possible during a given time (or in the case of enough UAVs – fully research the entire territory with minimal fuel/time consumption).

For the successful performance of the tasks UAVs facing, it is necessary to solve the optimization problem of constructing routes with and without autonomous movement. During the territory study, it can be divided into smaller areas, taking into account the technical capabilities of the UAV equipment, which, for example, will correspond to the viewing angle of video recording equipment. In this case, some such problems can be reduced to building routes along with the points of centers of these smaller areas.

The problem of optimizing routes for UAV groups (without autonomous movement). According to the problem conditions, it is necessary to construct routes to each of the objects to be visited during the task execution, with a minimum travel time to the destinations and the minimum number of UAVs needed to complete the task (see image 1).



Img. 1 Example of route construction for two UAVs without autonomous movement

At the same time, each UAV is constantly under the direct control of the operator, and the UAV flew out and land at a place that we will call the depot.

Let's formalize the problem.

Incoming data.

- Maximum transmission range of the control signal;
- UAVs maximum number;

- coordinates of objects to visit;
- edge coordinates of territory to explore;
- depot coordinates;
- maximum range of UAV flight.

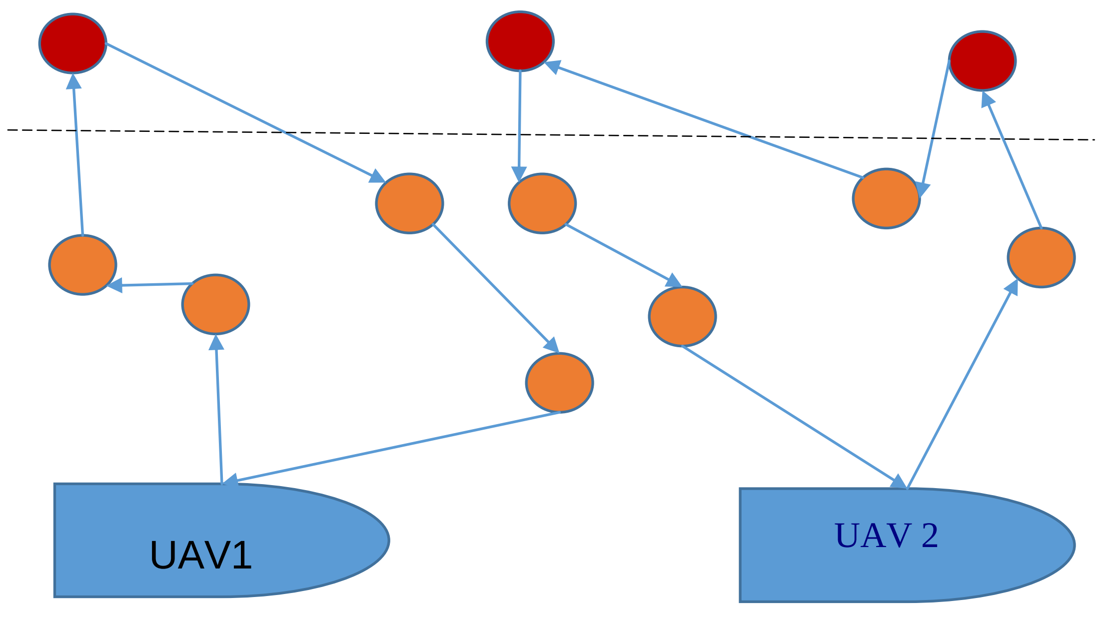
Output.

The route of each UAV with topographical binding, the route of the UAV is laid out in real coordinates and can be displayed on the area map.

Optimization Criteria.

- Minimization of the UAV routes total length;
- Minimization of the UAVs total number.

Routes optimization problem for UAV groups (with autonomous movement). Under the conditions of this problem, it is necessary to build routes to visit the objects (from the problem conditions) with the total minimum length of these routes, the minimum number of UAVs necessary for the task completion, and to determine the minimal time of autonomous movement (see image 2).



Img. 2 Example of route construction for two UAVs with autonomous movement

The autonomous movement refers to the UAV movement without the direct remote real-time control by the operator (movement on autopilot), following the predetermined flight plan.

The test area can be expanded by relaying a control signal using the UAV as a repeater.

Points of special interest, highlighted in image 2 with red color, are outside the UAV control radius and should be visited on autopilot.

Input data. Input data for the route optimization problem for UAV groups with autonomous movement:

• UAVs number;

• coordinates of objects to visit;

• depot coordinates;

• coordinates of the ground control points (operators);

• maximum range of UAV flight.

• UAV launch intervals and the number of UAVs, able to be in the air simultaneously (as supported by the control station);

• time of the UAV preparation for flight;

• bandwidth of the UAV relay equipment;

• minimum required channel bandwidth;

• maximum allowable delay for the control signal;

• the reach radius of the control signal from the ground control points.

The coordinates of the ground control points are needed to determine the maximum possible range of the UAV direct control (without signal retranslation). The UAV maximum flight range is determined by its technical characteristics (without taking into account the ability to control the UAV at such a distance).

Problem parameters.

- the maximum number of allowed retranslations;
- Frequency of UAV polls.

Criteria.

- minimizing sum length of UAV routes;
- minimizing the number of UAVs needed to complete the task;
- minimization of the number of control signals retranslations;
- minimizing the time spent outside the direct control zone;
- maximizing the amount of information being transmitted.

Output.

The route of each UAV with topographical binding, the route of the UAV is laid out in real coordinates and can be displayed on the area map.

When solving these problems three-dimensional (or, for calculations simplification, two-dimensional) coordinate grid is used.

Mathematical model

This routing problem relates to combinatorial type optimization problems that can be presented using a special graph [Корольов & Огурцов, 2017]. There are a set of points V_e to visit to avoid the refounding objects and reaching the target areas, there is a set of (depot) points V_d from which UAVs can be launched, as well as a set of points V_r where they can return.

Thus, based on these sets, a planar graph may be placed onto the map.

Let's introduce the following legend.

$G(V,E)$ – graph of the problem, where, $V=V_e \cup V_d \cup V_r$ and E is the set of graph edges;

C – matrix of integral distances (path values) C_{ij} between points v_i, v_j belong to V;

m – number of UAVs;

R_i – the route of the i-th UAV (i=1, …, m);

R – a set of routes R_i (i=1…m);

$C(R_i)$ – the cost of the route R_i (i=1, …, m);

$K(v)$ – the weight of each vertex v that belongs to V_e, which determines the importance of its attendance;

q_i – the movement resource of the UAV i (i=1, …, m).

Each point (graph vertex of the problem G) vi is given by coordinates (x_i, y_i). The routing problem consists in determining a set of m routes with a minimum total cost so that each vertex of the V_e subset is visited by only one UAV at least once. Also, the routes must begin at the vertices of the subset V_d and end at any point in the subset of V_r.

The target function in the general case is the cost of the solution of the problem:

$$F_{VRP}=(\Sigma K(v_i))/\Sigma C(R_i)$$

where $K(v_i)$ is the weight of the vertices v_i – part of the route R_i, i = 1, ..., m; $C(R_i)$ – the sum of the lengths of the route R_i paths, i = 1, ..., m.

In other words, it is necessary to find a valid solution with the maximum value of the target function on the formula above.

Conclusions and Prospects for Further Research

As a result of the performed works, further development of the routing problem meaningful statement for the UAV radio network with and without retransmission of control signals from the ground control center and the possibility of the UAV autonomous movement on the part of the route has been completed.

The general formulation of the target search problem in the area was given. An analysis of how to evaluate the quality of the search and what is the criteria for its effectiveness was conducted.

The formal statement of the UAV routing problem is presented, which deals with the stages of the operation planning and visiting the set targets by UAVs at the same time.

A mathematical model of a given problem has been built. Combinatorial optimization algorithms for this prblem can become the basis for solving a number of complex technical problems: increasing the technical (minimizing the territory exploring time) plus economic indicators (minimizing the UAVs amount) while using UAVs groups.

Implementation and practical use of the results obtained will allow us to increase the effectiveness of individual and group use of UAVs while performing different operations types and increase the maximum possible distance of UAV control by using relays.

Main directions of possible further research:

- development of combinatorial optimization algorithms for this problem;
- problems of the tasks formalization and routes optimization of a UAVs group, combined within a secure network, using the aircraft carrier;
- integral solving of the routing problem with minimization of necessary relays as an additional parameter.

Bibliography

[Ponda et al, 2015] Ponda, S. S., Johnson, L. B., Geramifard, A., & How, J. P. (2015). Cooperative mission planning for multi-UAV teams. *Handbook of unmanned aerial vehicles, 2*, 1447-1490.

[Austin, 2011] Austin, R. (2011). Unmanned aircraft systems: UAVS design, development and deployment (Vol. 54). *John Wiley & Sons*.

[Jaiswal, 2012] Jaiswal, N. K. (2012). Military operations research: quantitative decision making (Vol. 5). *Springer Science & Business Media*.

[Golden et al, 2008] Golden, B. L., Raghavan, S., & Wasil, E. A. (Eds.). (2008). The vehicle routing problem: latest advances and new challenges (Vol. 43). *Springer Science & Business Media*.

[Ventzel, 1964] Вентцель, Е. С. (1964). Введение в исследование операций. *Сов. радио*.

[Kumari et al, 2020] Kumari, A., Gupta, R., Tanwar, S., & Kumar, N. (2020). A taxonomy of blockchain-enabled softwarization for secure UAV network. *Computer Communications, 161*, 304-323.

[Ye et al, 2020] Ye, L., Zhang, Y., Li, Y., & Han, S. (2020, June). A Dynamic Cluster Head Selecting Algorithm for UAV Ad Hoc Networks. In 2020 *International Wireless Communications and Mobile Computing (IWCMC)* (pp. 225-228). IEEE.

[Корольов et al, 2017] Корольов В.Ю., Огурцов М.І., Ходзінський О.М. (2017). Про задачу побудови маршрутів руху для групи рухомих дистанційно керованих систем. *Обчислювальний інтелект (результати, проблеми, перспективи): праці міжнар. наук.-практ. конф., 16-18 травня 2017р.: Київ-Черкаси: наук. ред. В.Є. Снитюк. – К. ВПЦ «Київський університет»*, с. 247-248.

[Корольов & Огурцов, 2017] Корольов В.Ю., Огурцов М.І. (2017) Транспортно-комунікаційна задача для груп безпілотних апаратів. Математичні машини і системи, 1, 82-89.

[Огурцов & Ходзінський, 2016] Огурцов, М. І., & Ходзінський, О. М. (2016). Розробка алгоритмів розв'язання задачі маршрутизації транспортних засобів з часовими вікнами. *Компьютерная математика*, с. 134-142.

[Поліновський at al, 2017] Корольов, В. Ю., Поліновський, В. В., & Огурцов, М. І. (2017). Моделювання мереж зв'язку рухомих дистанційно керованих систем на базі HLA. *Вісник Хмельницького національного університету. Технічні науки*, 1, 160-165.

[Гуляницький & Огурцов, 2020] Гуляницький Л.Ф. Огурцов М.І. (2020). Визначення оптимального складу групи БПЛА для виконання поставленої задачі на основі теорії нечітких множин. *"Інформаційні технології та комп'ютерне моделювання"; матеріали статей Міжнародної науково-практичної конференції, м. Івано-Франківськ, 18-22 травня 2020 року. – Івано-Франківськ: п. Голіней О.М.*, с. 36-39.

[Coutinho at al, 2018] Coutinho, W. P., Battarra, M., & Fliege, J. (2018). The unmanned aerial vehicle routing and trajectory optimisation problem, a taxonomic review. *Computers & Industrial Engineering*, *120*, 116-128.

[Horbulin at al 2020] Horbulin, V.P., Hulianytskyi, L.F. & Sergienko, I.V.(2020). Optimization of UAV Team Routes in the Presence of Alternative and Dynamic Depots. *Cybern Syst Anal., 56, 2*, 195–203. https://doi.org/10.1007/s10559-020-00235-8.

[Otto at al, 2018] Otto, A., Agatz, N., Campbell, J., Golden, B., & Pesch, E. (2018). Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: A survey. *Networks*, 72(4), 411-458.

[Zhao at al, 2018] Zhao, Y., Zheng, Z., & Liu, Y. (2018). Survey on computational-intelligence-based UAV path planning. *Knowledge-Based Systems*, 158, 54-64.

[Гуляницький & Рибальченко, 2018] Гуляницький, Л. Ф., & Рибальченко, О. В. (2018). Формалізація та розв'язування одного типу задач маршрутизації БПЛА. *Теорія оптимальних рішень, 17*, с. 107-114.

[Perez-Carabaza at al, 2018] Perez-Carabaza, S., Besada-Portas, E., Lopez-Orozco, J. A., & Jesus, M. (2018). Ant colony optimization for multi-UAV minimum time search in uncertain domains. *Applied Soft Computing, 62*, 789-806.

[Cekmez at al, 2017] Cekmez, U., Ozsiginan, M., & Sahingoz, O. K. (2017, December). Multi-UAV path planning with multi colony ant optimization. In *International Conference on Intelligent Systems Design and Applications* (pp. 407-417). Springer, Cham.

[Chiang at al, 2019] Chiang, W. C., Li, Y., Shang, J., & Urban, T. L. (2019). Impact of drone delivery on sustainability and cost: Realizing the UAV potential through vehicle routing optimization. *Applied energy, 242*, 1164-1175.

[Binol at al, 2018] Binol, H., Bulut, E., Akkaya, K., & Guvenc, I. (2018, August). Time optimal multi-uav path planning for gathering its data from roadside units. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)* (pp. 1-5). IEEE.

[Xu at al, 2010] Xu, C., Duan, H., & Liu, F. (2010). Chaotic artificial bee colony approach to Uninhabited Combat Air Vehicle (UCAV) path planning. *Aerospace Science and Technology, 14(8)*, 535-541.

[Tian at al, 2018] Tian, G., Zhang, L., Bai, X., & Wang, B. (2018, July). Real-time dynamic track planning of multi-UAV formation based on improved artificial bee colony algorithm. In *2018 37th Chinese Control Conference (CCC)* (pp. 10055-10060). IEEE.

[Shakhatreh at al, 2017] Shakhatreh, H., Khreishah, A., Alsarhan, A., Khalil, I., Sawalmeh, A., & Othman, N. S. (2017, April). Efficient 3D placement of a UAV using particle swarm optimization. In *2017 8th International Conference on Information and Communication Systems (ICICS)* (pp. 258-263). IEEE.

[Tsourdos at al, 2010] Tsourdos, A., White, B., & Shanmugavel, M. (2010). Cooperative path planning of unmanned aerial vehicles (Vol. 32). *John Wiley & Sons*.

[Shima and Rasmussen, 2009] Shima, T., & Rasmussen, S. (Eds.). (2009). UAV cooperative decision and control: challenges and practical approaches. *Society for Industrial and Applied Mathematics*.

[Toth and Vigo, 2014] Toth, P., & Vigo, D. (Eds.). (2014). Vehicle routing: problems, methods, and applications. *Society for Industrial and Applied Mathematics*.

[Braekers at al, 2016] Braekers, K., Ramaekers, K., & Van Nieuwenhuyse, I. (2016). The vehicle routing problem: State of the art classification and review. *Computers & Industrial Engineering*, *99*, 300-313.

[Karakatič and Podgorelec, 2015] Karakatič, S., & Podgorelec, V. (2015). A survey of genetic algorithms for solving multi depot vehicle routing problem. *Applied Soft Computing, 27*, 519-532.

[Soto at al, 2017] Soto, M., Sevaux, M., Rossi, A., & Reinholz, A. (2017). Multiple neighborhood search, tabu search and ejection chains for the multi-depot open vehicle routing problem. *Computers & Industrial Engineering, 107*, 211-222.

Authors Information

Maksym Ogurtsov *– research assistant in V.M Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, 40 Glushkova ave., Kyiv, Ukraine; e-mail: ogurtsov.maksym@incyb.kiev.ua*

Major fields of scientific research: information security, cryptography, combinatorial optimization.

Vyacheslav Korolyov *– senior scientist in V.M Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, 40 Glushkova ave., Kyiv, Ukraine; e-mail: korolev@i.ua*

Major fields of scientific research: combinatorial optimization.

ISOMORPHISM OF PREDICATE FORMULAS AS THE BASE OF LOGIC ONTOLOGY CONSTRUCTION

Tatiana Kosovskaya

Abstract: *The paper describes an approach to the formation of logic ontology based on descriptions of objects in terms of the predicate calculus language. The main tool for creating such a description is an extraction of an elementary conjunction of predicate formula literals that is isomorphic to subformulas of some given formulas. The definition of an isomorphism of elementary conjunctions of atomic predicate formulas is given. The method of ontology construction is formulated. An illustrative example of such a construction and expansion of ontology is presented.*

Keywords: *Logic ontology, isomorphism of predicate formulas.*

ITHEA Keywords: *G.2. Mathematics of Computing, Discrete mathematics.*

Introduction

In philosophy, ontology is understood as abstraction, taking into account some properties of objects (or relations of parts of these objects) and ignoring others.

For mathematicians, an ontology is a oriented graph in which every node contains a set of objects having the same properties (or the same relations between parts of these objects). Moreover, all objects located at the end of the edge have the properties inherent in the objects at the beginning of the edge.

Construction of an ontology of an object domain is one of the directions for representing knowledge for its use in Artificial Intelligence systems.

If an object is represented as a set of its elements that have given properties and are connected with given relations, then the description of the object can be given by a predicates calculus formula.

To extract from a set of objects a subset with elements having common properties, a construction of a level description of the initial set of objects was proposed in Kosovskaya, [2008]. The main tool for creating such a description is an extraction of such an elementary conjunction of literals of predicate formulas that is isomorphic to subformulas of some initial formulas Kosovskaya, [2014]. In the present work, it is proposed to use the ideas of such a construction for the formation of an ontology.

Isomorphism of predicate formulas

The basis of further presentation is the concept of isomorphism of predicate formulas. Here is a definition of this concept.

Definition. *Two elementary conjunctions of atomic predicate formulas* $P(a_1, \ldots, a_m)$ *and* $Q(b_1, \ldots, b_m)$ *are called isomorphic if there exists such an elementary conjunction* $R(x_1, \ldots, x_m)$ *and substitutions of arguments* $a_{i_1}, \ldots, a_{i_m}$ *and* $b_{j_1}, \ldots, b_{j_m}$ *of formulas* $P(a_1, \ldots, a_m)$ *and* $Q(b_1, \ldots, b_m)$*, respectively, instead of all occurrences of the variables* $x_1, \ldots, x_m$ *of the formula* $R(x_1, \ldots, x_m)$*, that the results of these substitutions* $R(a_{i_1}, \ldots, a_{i_m})$ *and* $R(b_{j_1}, \ldots, b_{j_m})$ *coincide with the formulas* $P(a_1, \ldots, a_m)$ *and* $Q(b_1, \ldots, b_m)$*, respectively, up to the order of literals.*

The recieved substitutions $(a_1 \to x_{i_1}, \ldots, a_m \to x_{i_m})$ *and* $(b_1 \to x_{j_1}, \ldots, b_m \to x_{j_m})$ *are called unifiers of the formulas* $P(a_1, \ldots, a_m)$ *and* $Q(b_1, \ldots, b_m)$ *with the formula* $R(x_1, \ldots, x_m)$.

Note that the arguments of elementary conjunctions $P(a_1, \dots, a_m)$ and $Q(b_1, \dots, b_m)$ may be either object variables or object constants. In addition, the notion of isomorphism of elementary conjunctions of atomic predicate formulas differs from the notion of these formulas equivalence, since they may have significantly different arguments. In fact, for isomorphic formulas there are such permutations of their arguments, that they define the same relation between their arguments.

It was proved in Kosovskaya, Kosovskii [2019] that the predicate formula isomorphism problem is polynomially equivalent to the graph isomorphism problem, which is a so called "open" problem, about which neither its belonging to the class **P** nor its NP-completeness has been proved. So, the predicate formula isomorphism problem is GI-complete.

In the further presentation, to construct an ontology, it will be necessary to extract a maximal subformula isomorphic two given elementary conjunctions.

Definition. *An elementary conjunction is called a maximal formula isomorphic to subformulas of two given elementary conjunctions if it is isomorphic to some subformulas of these formulas, but after adding any literal to it, it is not isomorphic to any subformula of at least one of the given formulas.*

Example of a maximal formula isomorphic to subformulas of two given elementary conjunctions.

Let

$$P(a,b,c) = p_1(a,b) \;\&\; \neg p_1(a,c) \;\&\; p_2(b,c,a),$$

$$Q(a,b,c) = p_1(b,a) \;\&\; \neg p_1(a,c) \;\&\; p_2(a,c,b).$$

These elementary conjunctions have a common subformula with one literal

$$R_1(a,c) = \neg p_1(a,c)$$

and a subformula with two literals

$$R_2(x,y) = p_1(x,y) \;\&\; p_2(y,c,x),$$

which is isimorphic to subformulas of formulas $P(a,b,c)$ and $Q(a,b,c)$ because

$$P(a,b,c) = R(a,b) \;\&\; p_1(a,c),$$

$$Q(a,b,c) = R(b,a) \;\&\; p_1(a,c).$$

Adding the literal $\neg p_1(x,z)$ to the formula $R(x,y)$ turns it into a formula which is isomorphic to $P(a,b,c)$, but not isomorphic to the formula $Q(a,b,c)$. Adding the literal $\neg p_1(y,z)$ to it turns it into a formula which is isomorphic to $Q(a,b,c)$, but not isomorphic to the formula $P(a,b,c)$.

Thus, both the literal $\neg p_1(x,z)$ and the elementary conjunction $R(x,y) = p_1(x,y) \;\&\; p_2(y,c,x)$ are maximal formulas isomorphic to the subformulas of the formulas $P(a,b,c)$ and $Q(a,b,c)$.

An algorithm for checking the isomorphism of formulas, which has an exponential computational complexity, as well as an approximate polynomial algorithm for solving this problem is in Kosovskaya, Petrov [2017]. Note that, in contrast to the problem of checking formulas for isomorphism, the problem of extracting of maximal subformulas isomorphic to each other from two given elementary conjunctions is NP-hard.

Problem of constructing a logical ontology setting

Let a set of objects Ω be given. Every object ω from Ω is represented as a set of its elements $\omega = \{\omega_1, \dots, \omega_t\}$. A set of predicates $p_1, \dots, p_n$, which characterize properties of elements of ω and relations between them, is defined on ω.

If all predicates $p_1, \dots, p_n$ are binary then the logical description $S(\omega)$ of an object ω is a conjunction of all literals true on ω.

Note that the predicates $p_1, \ldots, p_n$ may not necessarily be binary. They can be finite-valued (defining the property of a numerical feature of an object to be in a given interval or, for example, the color of an object's element) or even infinite-valued (defining, for example, length, weight, etc.). A multi-valued n-ary predicate q^n with values from the set D_q can be interpreted as a $n+1$-ary predicate q^{n+1}, defined as $q^{n+1}(\overline{x}; y) \Leftrightarrow q^n(\overline{x}) = y$[1] and where variables from the string $\overline{x}$ are variables for the names of elements of the object itself, and the variable y is a variable for numbers from D_q. In fact, this means that among the arguments of the predicate there are not only variables for the names of the elements of the object ω or these names themselves, but also numerical arguments. We will call such predicates and numerical arguments quantitative.

To distinguish the variables for the names of the elements of the object ω or the names themselves from the names of numeric arguments, when writing the notation of the formula, we will write $P(\overline{x}; \overline{a}_1; \ldots; \overline{a}_m)$, where $\overline{x}$ is the argument list for the names of the elements of the object, $\overline{a}_1; \ldots; \overline{a}_m$ are the lists of quantitative arguments.

A logical description $S(\Omega)$ of the set Ω is the set of all descriptions $S(\omega)$ for $\omega \in \Omega$

$$S(\Omega) = \{S(\omega) : \omega \in \Omega\}.$$

Note that the set of formulas defining the description $S(\Omega)$ is divided into classes of equivalence with respect to the isomorphism of formulas. Therefore, we can assume that only nonisomorphic formulas, in which only variables are used as arguments, are presented in $S(\Omega)$[2].

Subsets $\Omega_1, \ldots, \Omega_m$ (possibly intersecting) of the set Ω are pointed out.[3].

Definition.

The greatest common property of elements from the set Ω is the formula with the largest number of literals for which each description of an object from Ω has an isomorphic subformula.

It is required to construct an oriented graph whose indegree node is labeled by the set Ω and corresponds to the set of non-isomorphic formulas, each of which is isomorphic to some descriptions of objects from Ω.

If there are edges from the node labeled by the set Ω_k to the nodes labeled by the sets $\Omega_{k_1}, \ldots, \Omega_{k_r}$, then

1. $\Omega_k = \cup_{i=1}^{r} \Omega_{k_i}$;

2. for each $i = 1, \ldots, r$, each formula in the description of objects from Ω_{k_i} is isomorphic to one of the formulas from the descriptio of Ω_k, i.e. objects from Ω_{k_i} have all the properties common to some objects from Ω_k;

3. if $i \neq j$, then the greatest common properties of elements from Ω_{k_i} and Ω_{k_j} are different.

Logic ontology construction

If descriptions of $S(\Omega)$ contain binary predicates, then each literal with such a predicate divides Ω into two disjoint subsets. The greatest common properties of these subsets (excluding literals with quantitative predicates) can be the same or different.

If the number of binary predicates is r , then Ω is divided into 2^r subsets. To avoid such a large number of subsets, the following procedure can be applied.

We single out the largest general properties of objects, taking into account only binary predicates. An algorithm for such an extraction is presented in Kosovskaya, [2014]. If some of the greatest common property is satisfied by elements from only one selected set, then we combine these properties.

We proceed to the analysis of the quantitative properties of objects. In the description of each of the selected subsets $\Omega_1, \ldots, \Omega_m$ of the set Ω we find the greatest common property of its elements $P_1(\overline{x}_1), \ldots, P_m(\overline{x}_m)$ and sort them by the number of arguments (numerical variables) in decreasing order $P_{k_1}(\overline{x}_{k_1}), \ldots, P_{k_m}(\overline{x}_{k_m})$.

[1]Hereinafter, $\overline{x}$ denotes a list of elements of the finite set x.

[2]Sometimes it is useful to highlight some values of quantitative constants and leave them as quantitative arguments.

[3]These subsets correspond to some well-known properties of objects that will be present in the ontology. The problem in which such subsets are to be extracted is not considered in this paper.

For $i = 1, \ldots, m$ we perform the following actions.

1. Among formulas $P_{k_{i+1}}(\overline{x}_{k_{i+1}}), \ldots, P_{k_m}(\overline{x}_{k_m})$ we find the formulas with the least number of arguments obtained from $P_{k_i}(\overline{x}_{k_i})$ by identifying the numerical arguments.

2. The node marked Ω_{k_i} we connect by oriented edges with the nodes corresponding to the found formulas.

When adding another subset of $\tilde{\Omega}$ to the selected subsets, we find the greatest common property of its elements $\tilde{P}(\overline{x}_{\tilde{}}))$. We determine its place among $P_{k_1}(\overline{x}_{k_1}), \ldots, P_{k_m}(\overline{x}_{k_m})$, assign it the corresponding number i_0 and increase per unit indices of formula numbers with $i \geq i_0$. We get $P_{k_1}(\overline{x}_{k_1}), \ldots, P_{k_{i_0}}(\overline{x}_{k_{i_0}}), \ldots, P_{k_{m+1}}(\overline{x}_{k_{m+1}})$. Then we perform 1, 2 for $i = i_0$.

Illustrative example: ontology of quadrangles

Consider the set of quadrangles. Each object (quadrangle) is defined by a sequence of four different names of its tops connected to each other in a cycle. So, for example, the same quadrangle can be given as (A, B, C, D), or (B, C, D, A), or (C, D, A, B) , or (D, A, B, C).

Each object is characterized by the lengths of its sides and the values of its internal angles. In this regard, we introduce predicates

$$length(x, y; l) \Leftrightarrow \text{“the length of the side } xy \text{ is } l\text{”},$$
$$angle(x; a) \Leftrightarrow \text{“the internal angle at the top } x \text{ is } a\text{”}.$$

For the convenience of further exposition, we will use the usual notations: $|xy| = l$ and $\angle x = a$. Here x and y are object arguments, l and a are quantitative arguments.

In addition, each object satisfies two conditions

– all angles are not equal to π (which excludes triangles with a dot on one side and segments with two points on it);

– the sum of all internal angles is 2π (which excludes the intersection of sides that do not have common ends).

Let the sample contains quadrangles of the type shown in Figure 1.

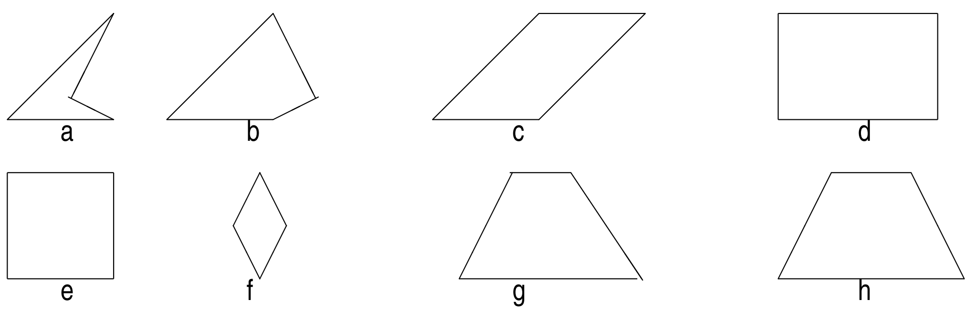


Figure 1. Quadrangles in the sample.

Since in our sample there are not only convex quadrangles, then one more predicate should be introduced

$$W(x, y, z, w) \Leftrightarrow \text{“one of the corners of the quadrangle } (x, y, z, w) \text{ is straight.”}$$

This predicate allows, at the first step, to divide the set of quadrangles into convex and not convex ones, since the former have a literal of the form $\neg W(x, y, z, w)$, and the latter have a literal of the form $W(x, y, z, w)$.

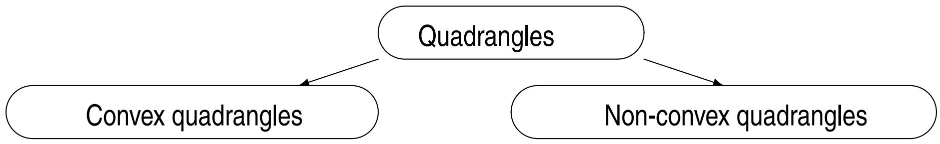


Figure 2. The beginning of ontology construction.

In this example, we will not further consider the set of non-convex quadrangles. In the set of convex quadrangles, all descriptions contain the literal $\neg W(x,y,z,w)$, so we will not take it into account for further ontology construction. In general, the description of a convex quadrangle in the selected signature has the form of not more than 2^{12} formulas, each of which is a description of a certain set of quadrangles

$$\begin{gathered} Q(x,y,z,w;l_1,l_2,l_3,l_4;a,b,c,d) = \\ |xy| = l_1 \,\&\, |yz| = l_2 \,\&\, |zw| = l_3 \,\&\, |wx| = l_4 \,\&\, \\ \angle x = a \,\&\, \angle y = b \,\&\, \angle z = c \,\&\, \angle w = d \end{gathered}$$

with possible restrictions (equal or not equal) on the lengths l_1,l_2,l_3,l_4 and values of angles a,b,c,d.

The number of numerical variables included in this description is $N_Q = 8$.

All parallelograms have descriptions of the form

$$\begin{gathered} P(x,y,z,w;l_1,l_2;a) = Q(x,y,z,w;l_1,l_2,l_1,l_2;a,\pi-a,a,\pi-a) = \\ |xy| = l_1 \,\&\, |yz| = l_2 \,\&\, |zw| = l_1 \,\&\, |wx| = l_2 \,\&\, \angle x = a \,\&\, \angle y = \pi - a \,\&\, \angle z = a \,\&\, \angle w = \pi - a \end{gathered}$$

with $l_2 = l_1$ or $l_2 \neq l_1$.

The number of numerical variables included in this description is $N_P = 3$.

For rectangles we have

$$\begin{gathered} R(x,y,z,w;l_1,l_2;) = Q(x,y,z,w;l_1,l_2,l_1,l_2;\tfrac{\pi}{2},\tfrac{\pi}{2},\tfrac{\pi}{2},\tfrac{\pi}{2}) = \\ |xy| = l_1 \,\&\, |yz| = l_2 \,\&\, |zw| = l_1 \,\&\, |wx| = l_2 \,\&\, \angle x = \tfrac{\pi}{2} \,\&\, \angle y = \tfrac{\pi}{2} \,\&\, \angle z = \tfrac{\pi}{2} \,\&\, \angle w = \tfrac{\pi}{2} \end{gathered}$$

with $l_2 = l_1$ or $l_2 \neq l_1$.

The number of numerical variables included in this description is $N_R = 2$.

For squares we have

$$\begin{gathered} S(x,y,z,w;l;) = Q(x,y,z,w;l,l,l,l;\tfrac{\pi}{2},\tfrac{\pi}{2},\tfrac{\pi}{2},\tfrac{\pi}{2}) = \\ |xy| = l \,\&\, |yz| = l \,\&\, |zw| = l \,\&|wx| = l \,\&\, \angle x = \tfrac{\pi}{2} \,\&\, \angle y = \tfrac{\pi}{2} \,\&\, \angle z = \tfrac{\pi}{2} \,\&\, \angle w = \tfrac{\pi}{2}. \end{gathered}$$

The number of numerical variables included in this description is $N_S = 1$.

For rhombus we have

$$\begin{gathered} Rh(x,y,z,w;l;a) = Q(x,y,z,w;l,l,l,l;a,\pi-a,a,\pi-a) = \\ |xy| = l \,\&\, |yz| = l \,\&\, |zw| = l \,\&\, |wx| = l \,\&\, \angle x = a \,\&\, \angle y = \pi - a \,\&\, \angle z = a \,\&\, \angle w = \pi - a. \end{gathered}$$

with $a = b$ or $a \neq b$.

The number of numerical variables included in this description is $N_{Rh} = 2$.

For trapezoid we have

$$\begin{gathered} T(x,y,z,w;l_1,l_2,l_3,l_4;a,b) = Q(x,y,z,w;l_1,l_2,l_3,l_4;a,\pi-a,b,\pi-b) = \\ |xy| = l_1 \,\&\, |yz| = l_2 \,\&\, |zw| = l_3 \,\&\, |wx| = l_4 \,\&\, \angle x = a \,\&\, \angle y = \pi - a \,\&\, \angle z = b \,\&\, \angle w = \pi - b \end{gathered}$$

with $a \neq b$ and $l_2 \neq l_4$ and (all the rest pairs of l_i, l_j may be equal or distinct).
The number of numerical variables included in this description is $N_T = 6$.
For isosceles trapezoid we have

$$TI(x, y, z, w; l_1, l_2, l_3; a) = Q(x, y, z, w; l_1, l_2, l_1, l_3; a, \pi - a, \pi - a, a) =$$
$$|xy| = l_1 \ \& \ |yz| = l_2 \ \& \ |zw| = l_1 \ \& \ |wx| = l_3 \ \& \ \angle x = a \ \& \ \angle y = \pi - a \ \& \ \angle z = \pi - a \ \& \ \angle w = a$$

with $l_2 = l_1 \neq l_3$ or $l_3 = l_1 \neq l_2$ or $l_2 \neq l_1$ and $l_3 \neq l_1$.
The number of numerical variables included in this description is $N_{TI} = 4$.
Order the formulas defining the specified classes of objects, in decreasing order of the number of quantitative variables.

Object	Number of quantitative variables in description
Trapezoid	6
Isosceles trapezoid	4
Parallelogram	3
Rhombus	2
Rectangle	2
Square	1

The largest number of quantitative variables (6) is in the description of the trapezoid. Among the formulas that define a trapezoid, there are formulas that define an isosceles trapezoid. Among the formulas defining an isosceles trapezoid, there are no formulas defining any other objects. Therefore, we have a part of the ontology presented in Figure 3a.

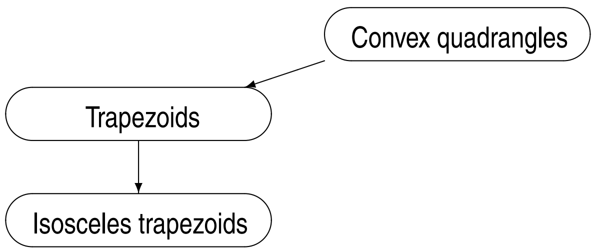


Figure 3a. Construction of ontology.

The largest number of quantitative variables (3) among the remaining descriptions of objects is in the parallelogram description. Among the formulas defining a parallelogram, there are formulas defining a rhombus and a rectangle (with the number of quantitative variables 2). Therefore, we have a part of the ontology presented in Figure 3b.

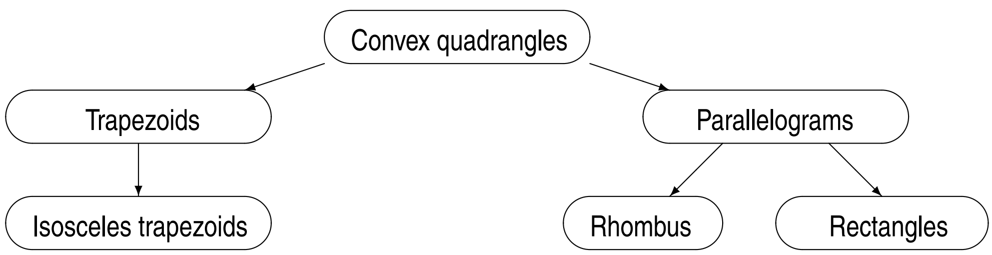


Figure 3b. Construction of ontology.

The sets of formulas defining the Rhombus and Rectangle have a common subset defining the Square (with the number of quantitative variables 1). Therefore, we have the ontology presented in Figure 3c.

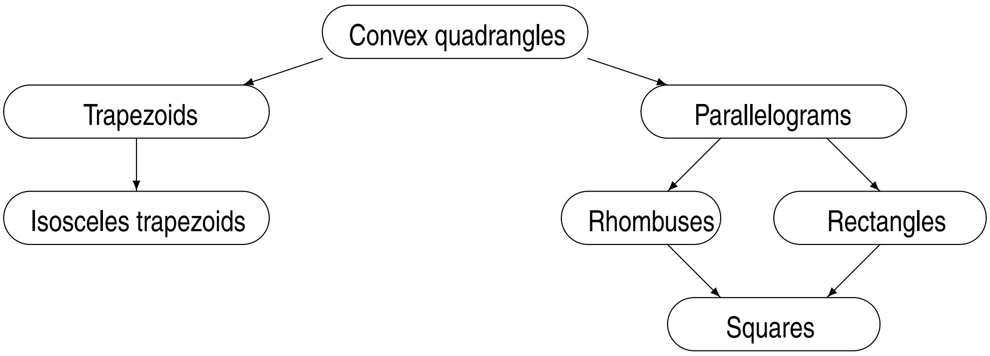


Figure 3c. Construction of ontology.

It should be noted that there is another interesting set of quadrangles, the descriptions of which form a set of formulas that are isomorphic to each other and do not coincide with any of the above. Typical representative of deltoids is shown in Figure 4.

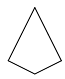


Figure 4. Deltoid.

Descriptions of these objects are isomorphic to formulas

$$F(x, y, z, w; l_1, l_2; a, b, c) = Q(x, y, z, w; l_1, l_2, l_2, l_1; a, b, c, a) =$$
$$|xy| = l_1 \,\&\, |yz| = l_2 \,\&\, |zw| = l_2 \,\&\, |wx| = l_1 \,\&\, \angle x = a \,\&\, \angle y = b \,\&\, \angle z = c \,\&\, \angle w = a$$

with $l_1 = l_2$ and $l_1 \neq l_2$ (and as a consequence with $a = c$ and $a \neq c$, respectively).

The number of numerical variables included in this description is $N_F = 5$.

In the deltoid, the number of quantitative variables is 4. Therefore, in ontology, this set of quadrangles is formed after the set of trapezoids. In addition, if $l_1 = l_2$, then not only $a = c$, but also $a = c = \ pi - b$. Therefore, a rhombus is a special case of a deltoid. Therefore, we have the ontology presented in Fig. 5.

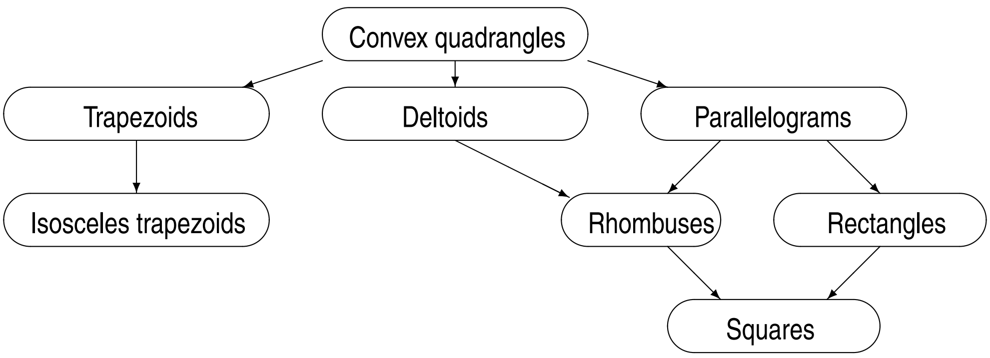


Figure 5. Extension of ontology.

Conclusion

The most interesting case of ontology construction is that some classes of objects which may be included in the ontology are not known in advance. In this case, it is required to extract such sets. If all predicates of the chosen signature are binary, then we can use the algorithm for constructing a level description of many objects Kosovskaya, [2014, 2018]. In fact, such a description is an ontology. However, in the presence of quantitative features of the object, the direct application of the algorithm for constructing a level description of many objects causes difficulties.

One of the areas of further research may be the development of algorithms for constructing a logical ontology for a set of objects, described, in particular, by quantitative features, for which subsets are not known that can serve as the basis for a “reasonable” classification of objects.

Bibliography

Kossovskaya T. M. “Level descriptions of classes for decreasing step number of pattern recognition problem solving described by predicate calculus formulas” // Vestn. St. Petersburg Univ.: Seria 10, Issue. 1. P. 64 – 72 (2008). (In Russian)

Kosovskaya T. “Construction of Class Level Description for Efficient Recognition of a Complex Object” // International Journal "Information Content and Processing", Vol. 1,No 1. 2014. P. 92 – 99.

Kosovskaya T. M., Petrov D. A. “Extraction of a maximal common sub-formula of predicate formulas for the solving of some Artificial Intelligence problems” // Vestnik of Saint-Petersburg University. Series 10. Applied mathematics. Computer science. Control processes. 2017, issue 3, pp. 250 – 263. (In Russian)

Kosovskaya T.M. “Predicate Calculus as a Tool for AI Problems Solution: Algorithms and Their Complexity”, Chapter 3 in: Intelligent System. Open access peer-reviewed Edited volume. Edited by Chatchawal Wongchoosuk Kasetsart University (2018) 1 – 20. https://www.intechopen.com/books/intelligent-system/predicate-calculus-as-a-tool-for-ai-problems-solution-algorithms-and-their-complexity

T. M. Kosovskaya, N. N. Kosovskii. “Polynomial Equivalence of the Problems Predicate Formulas Isomorphism and Graph Isomorphism” // VESTNIK ST. PETERSBURG UNIVERSITY, MATHEMATICS Vol. 52 No. 3, 2019. pp. 286 – 292.

Andrii Mykhailiuk, Mykola Petrenko. “Machine learning and ontologies as two approaches for building intellectual information systems” // International Journal “Information Technologies & Knowledge” Volume 13, Number 1, ï£¡ 2019, pp. 55 - 75.

Authors’ Information

Tatiana Kosovskaya *- Dr. of Sci., Professor, Head of Informatics Chair, St.Petersburg State University, University emb., 7 – 9, St.Petersburg, 199034, Russia; e-mail: kosovtm@gmail.com*
Major Fields of Scientific Research: Logical approach to artificial intelligence problems, Theory of Computational Complexity of Algorithms.

DIGITAL ANALYSIS OF GENETIC PROXIMITY OF THE BALKAN POPULATION

Georgi Gachev, Jordan Tabov

Abstract: *According to EUPEDIA the 12 basic haplogroups specific to the European population are:*

I1	**I2*/I2a**	**I2b**	**R1a**	**R1b**	**G**	**J2**	**J*/J1**	**E1b1b**	**T**	**Q**	**N**

Among the Bulgarians their distribution in percentage is:

4	**20**	**2**	**17**	**11**	**5**	**11**	**3**	**23.5**	**1.5**	**0.5**	**0.5**

Thus, it may be said that "the average Bulgarian" has "compound haplogroup" which consists of weighted partial inclusion of all basic haplogroups. Most common haplogroups for Bulgarian population are: E1b1b, I2/I2a, R1a, R1b and J2.*

Tabov and Koleva proposed a method for comparing "genetic closeness" between the nations: Each nation is presented as a point in 12-dimensional Euclidean space

Analysis of the "genetic" distances between the regional population groups of Europe makes it possible to associate these groups in "genetically close" clusters. One such cluster is the "Balkan".

From our analysis we conclude that the countries in the Balkan cluster are genetically closer to each other than all European countries among themselves. In fact the average distance in the Balkan cluster is 23 which is twice less than the average e-distance of 46 among European countries; and e.g. the e-genetic distance between North and South Germany is 21. Our results suggest also a comparison of the "genetic proximity" of the Balkan nations among themselves.

Keywords: *haplogroups, European population, genetic distance, cluster analysis, Balkan population, genetic center*

This paper is partially supported by the Task 1.2.5 of the Bulgarian National Scientific Program "ICT in Science, Education and Security", funded by the Ministry of Education and Science (Contract MES DOI-205/23.11.2018).

Introduction

The DNA analysis is a well-established scientific method that is increasingly used in practice. One of its popular applications is to resolve complicated family cases where the paternity is uncertain. Recently, the attention of both the general public and professionals has been focused on its use to study the genetic relations and origins of people and to classify people according to their genes and origins.

Most often the results of such a comparative analysis are presented in the form of a "tree diagram" similar to the family trees (**Error! Reference source not found.**), or as planar points, grouped "by adjacency" (**Error! Reference source not found.**), or in tabular form with percentage data of the predominant haplogroups (**Error! Reference source not found.**).

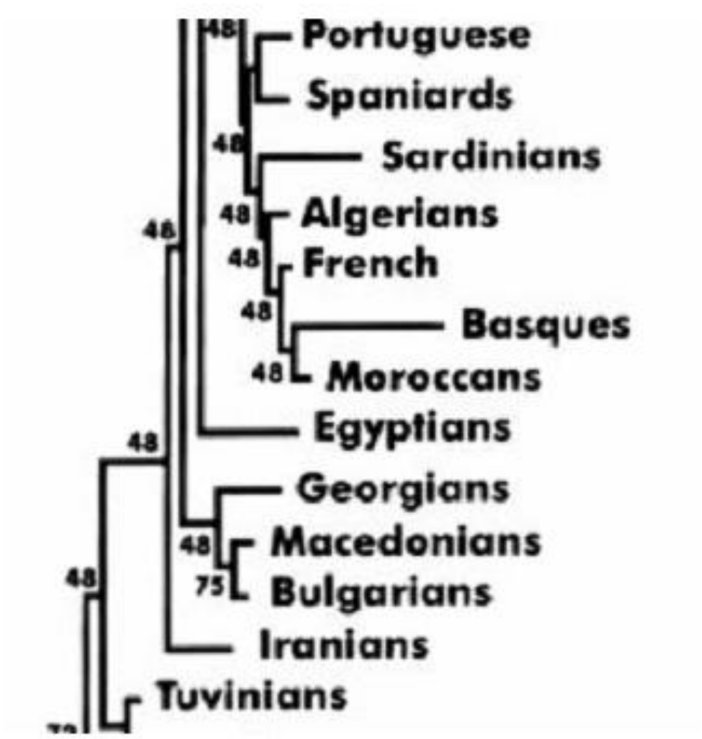


Figure 1. Part of the "Genetic Family Tree" of nations as result of a genetic research [Arnaiz-Villena et al., 2003].

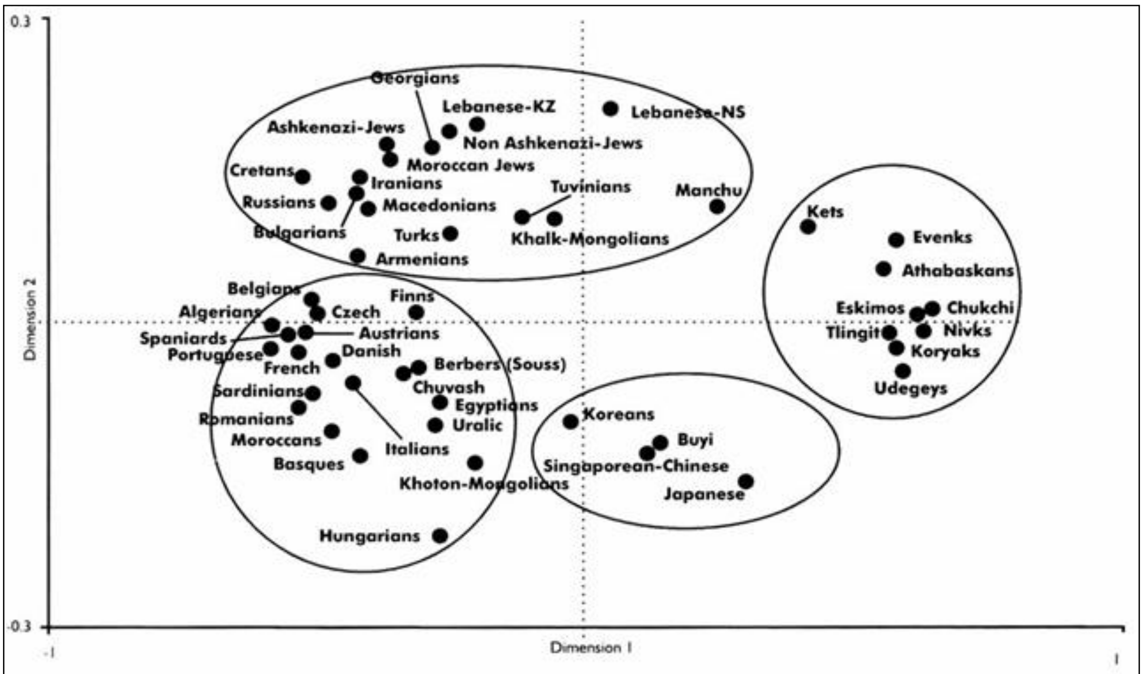


Figure 5. Correspondence analysis showing a global view of the relationship between Chuvash and other populations according to HLA allele frequencies in *n* dimensions (bidimensional representation). *HLA-DR* and *-DQ* generic typing allele frequencies data.

Figure 2. Grouping of people represented as points in the plane, according to their genetic adjacency [Arnaiz-Villena et al., 2003].

From the numerous publications in this field let us mention [Nazarova, 2009] [Balanovsky, 2012], [E., 2018], [Tabov & Sabeva-Koleva, 2018] [Chiaroni, Underhill, & Cavalli-Sforza, 2012] [Delev, 2018], [Eupedia, 2018], [Wikipedia, 2018a], [Chiaroni, Underhill, & Cavalli-Sforza, 2012], [FamilyTreeDNA, 2018], [Karmin, 2015], [Karlsson, Wallerström, Götherström, & Holmlund, 2006], [Wikipedia, 2018b] , [Trombetta B, 2011].

	У-хромозомни хаплогрупи											
Country/Region	I1	I2*/I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Bulgaria	4	20	2	17	11	5	11	3	23,5	1,5	0,5	0,5
Macedonia	3	23	1,5	13,5	12,5	4	14	2	21,5	1,5	0,5	0,5
Greece	3,5	9,5	1,5	11,5	15,5	6,5	23	3	21	4,5	0	0
Romania	4,5	26	2,5	17,5	12	5	13,5	1,5	15	0,5	0,5	0,5
Spain	1,5	4,5	1	2	69	3	8	1,5	7	2,5	0	0
Portugal	2	1,5	3	1,5	56	6,5	9,5	3	14	2,5	0,5	0
Belgium	12	3	4,5	4	61	4	4	1	5	1	0,5	0
France	8,5	3	3,5	3	58,5	5,5	6	1,5	7,5	1	0,5	0
Italy	4,5	3	2,5	4	39	9	15,5	3	13,5	2,5	0	0
Switzerland	14	1,5	8	3,5	50	7,5	3	0,5	7,5	0,5	1,5	1
South Germany	10,5	4,5	3	9,5	48,5	8	5,5	1	8	1,5	0,5	0,5
Austria	12	7	2,5	19	32	7,5	9	1	8	1	0,5	0,5
Czech Republic	11	9	4	34	22	5	6	0	6	1	1,5	0,5
Montenegro	6	29,5	1,5	7,5	9,5	2,5	9	0,5	27	0	2	1,5
Serbia	8,5	33	0,5	16	8	2	8	0,5	18	1	1,5	2
Hungary	8,5	16	2	29,5	18,5	3,5	6,5	3	8	0	0	0,5

Figure 3. Table presenting the content (in %) of twelve Y-DNA haplogoups typical for the respective European region [Tabov & Sabeva-Koleva, 2018] the data is from EUPEDIA project [Eupedia 2018].

Haplogroups of European nations

According to EUPEDIA the 12 basic haplogroups specific to the European population are:

I1 I2*/I2a I2b R1a R1b G J2 J*/J1 E1b1b T Q N

Among the Bulgarians their distribution in percentage is (Eupedia, 2018):

4 20 2 17 11 5 11 3 23.5 1.5 0.5 0.5

Thus, it may be said that “the average Bulgarian” has “compound haplogroup” which consists of weighted partial inclusion of all basic haplogroups. Most common haplogroups for Bulgarian population are: E1b1b, I2*/I2a, R1a, R1b and J2. For Serbia the corresponding percentages are:

8,5 33 0,5 16 8 2 8 0,5 18 1 1,5 2

These complex data show that the problem of genetic relation among nations is more complicated than the same problem but for individuals.

Genetic Distance

Tabov and Koleva (2018) propose a method for comparing “genetic closeness“ between the nations: Each nation is presented as a point in 12-dimensional Euclidean space. Percentage content of each haplogroup specific to the nation is one of the 12 coordinates of that point. The distance between the points is defined as **genetic distance**:

$$Distance(CountryA, CountryB) = \sqrt{\sum_{i=1}^{12}(CountryA_i - CountryB_i)^2} \quad (1)$$

This definition is **European specific**: for population in which the sum of these 12 Y-chromosome haplogroups is less than 95%, we obtain greater deviations compared to the values of most Y- chromosome haplogroups, therefore when we apply this definition for genetic distance we ignore them.

The genetic distances between European nations, calculated by formula (1), further will be called "e-genetic" (i.e. European genetic).

Two-dimensional visualization of genetic distances for four countries is not always appropriate. The reason is that the respective four points may not be in the same plane so the planar projection could be skewed. This may lead to speculations when such results are interpreted and discussed – see **Error! Reference source not found.** and **Error! Reference source not found.**.

Although the points of Bulgaria, Macedonia, Greece and Serbia are almost in the same plane so the planar view is acceptable approximation. The projection is shown in **Figure 4**.

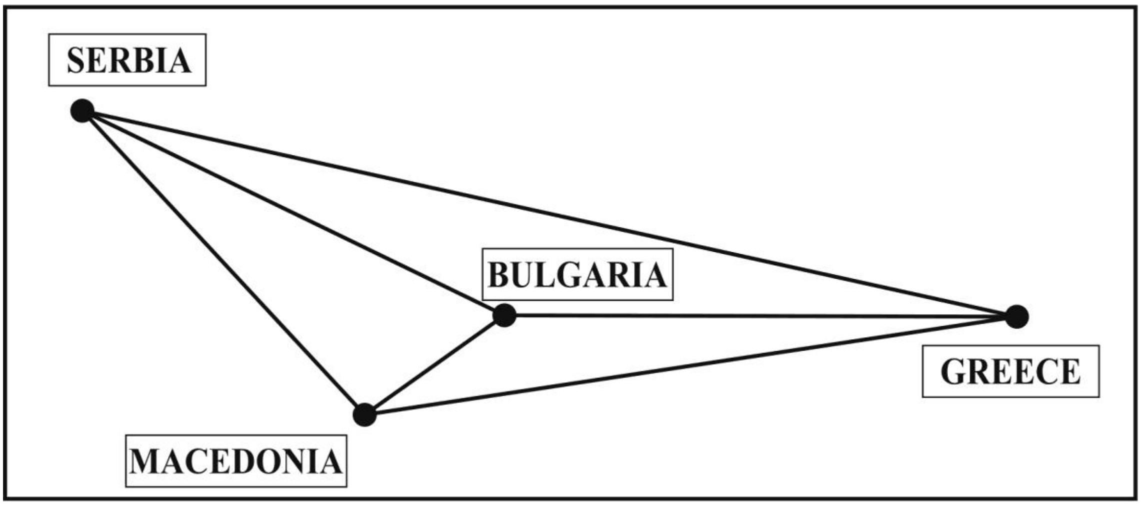


Figure 4. Visualization of genetic similarity of (Bulgaria, Macedonia, Serbia and Greece).

Diffuse distribution of the "European genes"

In the paper [Tabov & Sabeva-Koleva, 2018], the following hypothesis was offered:

Hypothesis: The change of ratio of haplogroups in Europe is a diffusion type process: it may be conjectured that the main part of this process has been played by “exchange of haplogroups” between neighboring nations. Most probably large scale migration has played less influence in this process. The results we are going to present below provide further arguments and clarification of this hypothesis.

In support of this hypothesis the following arguments are used in [Tabov et al., 2019]:

1- Genetic-geographical correlation: The first argument is connected to the correlation when “genetic distance” is compared to “geographical distance”. Since such a correlation exists this leads to the conclusion for significant diffusion type exchange of genes. Indeed, for all 38 European nations (we exclude small ones by territory and population) the correlation has coefficient of 0.61.

2- The second argument supporting the diffusion type distribution of male genes in Europe is the qualitative distribution (in percentages) of each of the 12 European haplogroups.

E.g. the percentages of haplogroup I2*/I2a are higher in Bosnia and Herzegovina; they are significant in neighboring countries – Serbia, Croatia, Montenegro, Rumania, Slovenia, Bulgaria and Moldova; they are to be found in most countries, neighbors of the above – Hungary, Slovakia, Ukraine, Belarus, Greece [Wikipedia 2018a].

Cluster analysis of e-genetic distances between the European people

Analysis of the "e-genetic" distances between the regional population groups of Europe makes it possible to associate these groups in "genetically close" clusters. One such cluster is the "Balkan", formed by the peoples of the following 8 Balkan countries: Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia; the table of the e-genetic distances between them is presented in **Figure 5**.

	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Bosnia-Herzegovina	39	22	53	37	32	33	25
Bulgaria		24	18	6	15	11	16
Croatia			38	23	25	16	13
Greece				17	27	22	31
Macedonia					13	9	15
Montenegro						17	13
Romania							12

Figure 5. Table of e-genetic distances between 8 Balkan countries: Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia.

Specific aspect of Slovenia and Albania

Analysis of the data for another two Balkan countries, Slovenia and Albania, shows that:

1 Slovenia is closer to another cluster of countries to which belong Slovakia and Ukraine. The e-genetic distances from Slovenia to Slovakia and Ukraine are respectively 8 and 14, which for example is less than the distances to Bulgaria – 31 and Greece - 40.

2 The result for Albania calculated by formula (1) is not reliable as the sum of percentages of all 12 European haplogroups is about 92%; therefore Albania is not included in the table in **Figure 4**. To have better approximation and improved view of Albania the "13th coordinate" is introduced. These are missed percentage values to sum of 100%. Thus for each country is added 13th "haplogroup" of unknown type just to fill the gap. Also this extends the formula (1) with another, 13th coordinate; the distance modified this way will be called b-genetic (i.e. Balkan genetic) distance.

Thus, the countries in the Balkan cluster become 9: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia; the table of the b-genetic distances between countries in the cluster is presented in **Figure 6**.

"Genetic Center" of the Balkan population

Virtual "genetic center" could be defined for the 9 countries from the "Balkan genetic cluster". This "genetic center" is geometric median of the cluster points, i.e the point which minimizes the sum of distances to all other 9 points representing corresponding countries. The B-genetic distance of a country to the "genetic center" characterizes its stronger or weaker belonging to the cluster.

The table with these b-genetic distances to the Balkan Genetic Center and b-genetic distances between the countries is presented in **Figure 7**.

	Bosnia-Herzegovina	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Albania	51	18	38	13	16	23	23	30
Bosnia-Herzegovina		39	22	53	37	32	33	25
Bulgaria			24	18	6	15	11	16
Croatia				38	23	25	16	13
Greece					17	27	22	31
Macedonia						13	9	15
Montenegro							17	14
Romania								12

Figure 6. Table of the b-genetic distances between countries in the Balkan cluster: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia

	Albania	Bosnia-Herzegovina	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Balkans genetic Center	21	31	9	18	22	6	11	6	9
Albania		51	18	38	13	16	23	23	30
Bosnia-Herzegovina			39	22	53	37	32	33	25
Bulgaria				24	18	6	15	11	16
Croatia					38	23	25	16	13
Greece						17	27	22	31
Macedonia							13	9	15
Montenegro								17	14
Romania									12

Figure 7. Table of the b-genetic distances between 9 Balkan countries: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia and their b-genetic distances to the "Balkan genetic center".

The table in **Figure 7** shows that closest countries to the Balkan Genetic Center are: Macedonia, Romania, Bulgaria and Serbia (with b-genetic distances to it, respectively 6, 6, 9, 9). Of these four, Bulgaria and Serbia are adjacent to each of the other two - Macedonia and Romania. It may be said that the population of these four countries together form the "closest b-genetic neighborhood" of the Balkan Genetic Center.

Conclusion

From the above analysis we may conclude that the countries in the Balkan cluster are genetically closer to each other than all European countries among themselves. In fact the average distance in the Balkan cluster is 23 which is twice less than the average e-distance of 46 among European countries; and e.g. the e-genetic distance between North and South Germany is 21.

The tight genetic relations are possible motivation for further improvement of the collaboration on the Balkans. The increasing of the number and the development of new genetic research methods will allow for more accurate comparisons and conclusions.

Appendix

Percentage of the quantity of haplogroups measured up to 20.11.2018 in population of 39 European nations/countries [Eupedia 2018]:

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Albania	2	12	1.5	9	16	1.5	19.5	2	27.5	1	0	0
Austria	12	7	2.5	19	32	7.5	9	1	8	1	0.5	0.5
Belarus	5.5	17.5	1	51	5.5	1.5	2.5	1	4	0	0	10
Belgium	12	3	4.5	4	61	4	4	1	5	1	0.5	0
Bosnia-Herzegovina	3	55.5	0	15	3.5	1.5	4	0.5	12	1	2	2
Bulgaria	4	20	2	17	11	5	11	3	23.5	1.5	0.5	0.5
Croatia	5.5	37	1	24	8.5	2.5	6	1	10	0.5	1	0.5
Czech Republic	11	9	4	34	22	5	6	0	6	1	1.5	0.5

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Denmark	34	2	5.5	15	33	2.5	3	0	2.5	0	1	1
England	14	2.5	4.5	4.5	67	1.5	3.5	0	2	0.5	0.5	0
Estonia	15	3	0.5	32	8	0	1	0	2.5	3.5	0.5	34
Finland	28	0	0.5	5	3.5	0	0	0	0.5	0	0	61.5
France	8.5	3	3.5	3	58.5	5.5	6	1.5	7.5	1	0.5	0
Germany	16	1.5	4.5	16	44.5	5	4.5	0	5.5	1	0.5	1
Greece	3.5	9.5	1.5	11.5	15.5	6.5	23	3	21	4.5	0	0
Hungary	8.5	16	2	29.5	18.5	3.5	6.5	3	8	0	0	0.5
Iceland	29	0	4	23	42	0	0	0	0	0	1	1
Ireland	6	1	5	2.5	81	1	1	0	2	0	0	0
Italy	4.5	3	2.5	4	39	9	15.5	3	13.5	2.5	0	0
Latvia	6	1	1	40	12	0	0.5	0	0.5	0.5	0.5	38

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Lithuania	6	6	1	38	5	0	0	0	1	0.5	0.5	42
Macedonia	3	23	1.5	13.5	12.5	4	14	2	21.5	1.5	0.5	0.5
Malta	1	10	1	3.5	32.5	6.5	21	8	9	4.5	1	0
Moldova	5	21	3	30.5	16	1	4	4	13	1	0	1.5
Montenegro	6	29.5	1.5	7.5	9.5	2.5	9	0.5	27	0	2	1.5
Netherlands	16.5	1	6.5	4	49	4.5	3.5	0.5	3.5	1	0	0
Norway	31.5	0	4.5	25.5	32	1	0.5	0	1	0	1	2.5
Poland	8.5	5.5	2	57.5	12.5	1.5	2.5	0	3.5	0.5	0.5	4
Portugal	2	1.5	3	1.5	56	6.5	9.5	3	14	2.5	0.5	0
Romania	4.5	26	2.5	17.5	12	5	13.5	1.5	15	0.5	0.5	0.5
Russia	5	10.5	0	46	6	1	3	0	2.5	1.5	1.5	23
Serbia	8.5	33	0.5	16	8	2	8	0.5	18	1	1.5	2

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Slovakia	6.5	16	1.5	41.5	14.5	4	2	1	6.5	0.5	0.5	3
Slovenia	9	20.5	1.5	38	18	1.5	2.5	0	5	1	0	0
Spain	1.5	4.5	1	2	69	3	8	1.5	7	2.5	0	0
Sweden	37	1.5	3.5	16	21.5	1	2.5	0	3	0	2.5	7
Switzerland	14	1.5	8	3.5	50	7.5	3	0.5	7.5	0.5	1.5	1

Bibliography

[Arnaiz-Villena et al., 2003] A. Arnaiz-Villena, J. Martinez-Laso, J. Moscoso, G. Livshits, J. Zamora, E. Gomez- Casado. HLA Genes in the Chuvashian Population from European Russia: Admixture of Central. European and Mediterranean Populations. *Human Biology, 75*(3), 375-392.

[Balanovsky, 2012] O. Balanovsky.Variability of the gene pool in space and time: the synthesis of data on the mitochondrial DNA genogeography and Y-chromosome. In B. O., *Abstract of dissertation for the degree of Doctor of Biological Sciences.* Moscow.

[Chiaroni et al., 2012] J. Chiaroni, P. Underhill, & L.Cavalli-Sforza. L Y chromosome diversity, human expansion, drift, and cultural evolution. *Proceedings of the National Academy of Sciences of the United States of America.*

[Delev, 2018] E. Delev. Population genetics of the Bulgarians, founders of the European civilization. *BOMI*.

[Eupedia, 2018] *Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by country in percent.* Retrieved 11 20, 2018, from EUPEDIA: https://www.eupedia.com/europe/european_y-dna_haplogroups.shtml

[FamilyTreeDNA, 2012] *Understanding Haplogroups (2012). Understanding Haplogroups: How are the haplogroups named? Family Tree DNA. Archived from the original on June 21, 2012*. Retrieved 11 11, 2018, from https://web.archive.org/web/20120621052330/ http:/www.familytreedna.com/understanding-haplogroups.aspx

[Karlsson et al., 2006] A. Karlsson, T. Wallerström, A. Götherström, & G. Holmlund. Y-chromosome diversity in Sweden - A long-time perspective. *European Journal of Human Genetics,* 963-970.

[Karmin et al. 2015] M. Karmin et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research*, 459-466.

[Nazarova, 2009] F. Nazarova. The closeness of the Finnish Ugric, Slav and Germanic populations according to anthropological and genetic data. *Journal of Evolutionary Biology Research 2009, 1*(1), 18-26.

[Tabov & Sabeva-Koleva, 2018] J. Tabov & N. Sabeva-Koleva. A method for assessing the genetic proximity of neighboring peoples. *Bulgarian Historical Readings 2018*.

[Tabov et al., 2019]. J. Tabov, G. Gachev, N. Sabeva-Koleva. On the genetic ties between European nations. *Khimiya, 28*, 615-627.

[Trombetta, 2011]. B. Trombetta. New Topology of the Human Y Chromosome Haplogroup E1b1 (E-P2) Revealed through the Use of Newly Characterized Binary Polymorphisms. *PLoS ONE*.

[Wikipedia, 2018a] *Haplogroup I2*/I2a (Y-DNA). From Wikipedia.* Retrieved 11 11, 2018, from https://ru.wikipedia.org/w/index.php?title=Гаплогруппа_E1b1b_(Y-ДНК)&oldid=97696115

[Wikipedia, 2018b]. *List of Y-chromosome in the population of the world.* Wikipedia. Retrieved 11 11, 2018, from https://en.wikipedia.org/wiki/List_of_Y-chromosome_haplogroups_in_populations_of_the_world

Authors' Information

Georgi Gachev *- Assistant researcher; Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria;*

Major Fields of Scientific Research: Mathematics Education, Modeling, Software dvelopment

Jordan Tabov *– Professor (retired); Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; tabov@math.bas.bg*

Major Fields of Scientific Research: Applications of mathrmatics and informatics in the Humanities, Didactics of mathematics and informatics

DIDACTICAL MAPPING IN TOPIC-ORIENTED DIDACTICAL UNITS

Borislav Yordanov Lazarov

***Abstract**: Didactic map is analogue to concept map when an educational process is under consideration. Two quantitative characteristics of a didactical map are introduced: weight and distance. These characteristics are used to evaluate the 'didactic price' of didactical map modifications. A complex structure of the map links is proposed aiming to perform a more coherent image of the connections between the entities of the map. The implementation of such complex structure in a topic-oriented didactical unit potentially facilitates the execution of a particular integrated didactic scenario. The theoretical frame is operationalized by a sample instruction for didactical map design. The scope of the study includes the innovative schools in Bulgaria, which are adopting integrated approach as an upgrade of the compulsory subject-oriented curricula.*

Keywords: didactic map weight, didactic map distance, integrated approach, processing matrix, topic-oriented education.

ITHEA Keywords: K.3 COMPUTERS AND EDUCATION

Introduction

The new Bulgarian legislation provides the innovative schools to arrange partially the school plan with introducing didactical approaches to ensure an effective teaching-learning process [MON, 2015]. Among them are project-oriented initiatives for upgrading the subject-oriented curricula through topic-oriented didactic scenarios in which the integrated approach is crucial [Lazarov, 2018]. The integrated approach supposes that any included subject in a topic-oriented didactical unit (TODU) comes with its own methodology and the subject teachers prepare the separate parts of the TODU scenario at their own discretion, having in mind some specific educational goals. Development of a concept map materially facilitates the analytical constituents of the ingredients of the topic under consideration [Giouvanakis et al., 2016]. Routine procedures are usually applied at this stage, though rethought in a new context. But at the synthesis level, in most cases, there are certain problems with the multifaceted presentation of the topic. Then it is necessary to concert the concepts and to merge analytical conceptual maps. Such a technical procedure has its "didactic price", which will be discussed below.

Theoretical Base

The concept map matrix shown in Figure 1 (based on [Kitano et al., 2008]) represents a modification of the concept map key introduced by Novak and Gowin [1984].

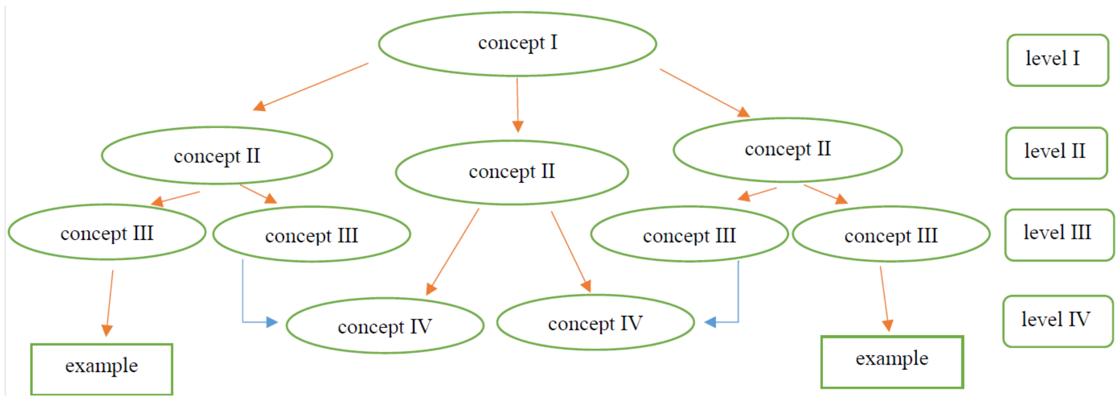


Figure 1. Concept map matrix

Here the concepts are accommodated in several hierarchic levels and a concept of higher level correlates with some concepts of lower levels (*higher* means smaller level number). The relationships between the concepts can be of two types: links (straight arrows in the figure) or crosslinks (broken arrows). The score of this particular map is calculated as

$$11 \cdot 1 + 4 \cdot 5 + 2 \cdot 10 + 2 \cdot 1 = 53.$$

This score corresponds to 11 links $(11 \cdot 1)$, 4 levels $(4 \cdot 5)$, 2 crosslinks $(2 \cdot 10)$ and 2 examples $(2 \cdot 1)$ [Kitano et al., 2008]. The values of the entities are commensurate with that of one link. In our opinion, the scoring formula is incomplete. It does not take into account the role of the concepts themselves. Simply expanding the formula with the addend (number-times-value) does not read the contribution of concepts from given level to the complexity of the map. The value 10 for a crosslink is also overestimated.

The relationships in Figure 1 constitutes an ontology and the classical Aristoteles' genus-species hierarchy is easily recognizable in it. An *similarity index* could be appropriated to any two concepts in such an ontology [Giouvanakis et al., 2016]. The authors give the definition via an example:

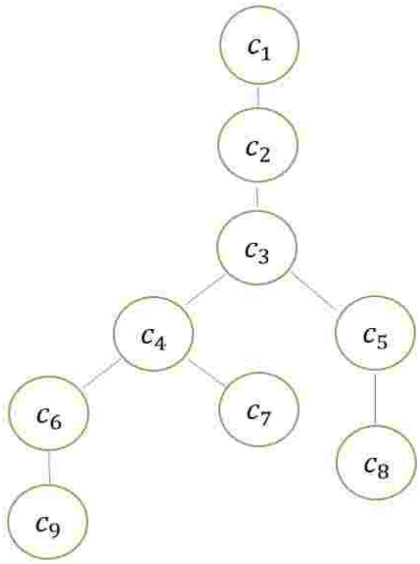


Figure 2. Ontology graph-tree

Following the notations in Figure 2, let:

– d_1 be the number of the links from the concepts c_6 to the common concept c_4;

– d_2 be the number of the links from c_7 to c_4;

– d_3 be the number of the links from c_4 to the genus concept c_2.

The similarity index for the concepts c_6 and c_7 is defined by the formula

$$\mathrm{sim}(c_6, c_7) = \frac{2d_3}{2d_3 + d_1 + d_2}.$$

About the Integrated Approach in Secondary School

Our standing point on the implementation of integrated approaches in secondary school is described more comprehensive in [Lazarov, 2018]. The fundamental idea is to amalgamate two (or more) methodical units from different subjects in one common methodical unit keeping the relative independence of any subject methodic. We speak for relatively independent inclusion of the methodic because the analytical knowledge developed by any methodic is expected to be synthesized in the frame of the common unit. The synthesis supposes transferability and multifunktionality of the complex *knowledge-skills-attitude* (KSA) into another context that differs from the ones, in which the analytical KSAs were developed. So, a kind of synchronization of the subject methodic should be orchestrated.

The *processing matrix* is a didactical tool that makes the implementation of the integrated approach via TODU more technological. It includes:

(1) definitions of the objectives, indicators of progress and benchmarks;

(2) selection of educational content to be examined;

(3) schedule and methodic for didactical support of the students;

(4) methodic for data collecting and processing the data collected;

(5) template for presenting information and frame for presentation design;

(6) assessment system;

(7) resources provided to the students and organizing a school conference;

(8) analysis of the outcomes of the initiative and drawing conclusions.

The processing matrix allows reconsidering the KSA developed in the traditional curriculum topics through the perspective of their potential application in a new context. Its proper design allows obtaining synergetic effect from the TODU, i.e. the resulting effect of the integrated teaching-learning process to exceed the effect of the expected one of any subject if it acts separately [ibid.]. But amalgamating methodic of different curriculum subjects is tricky business and some antinomies arise especially in points (2), (3) and (6) of the processing matrix. Resolving them could help a proper concept mapping.

Tuning the Concept Map Matrix

Basically, the structure of the concept map matrix in Figure 1 could be applied to a TODU for simple synchronization of concepts. However, a teaching methodic operates with more complex entities and relations than subject concepts and links between them. This is why we are going to conform the general concept map matrix to the purposes of the integrated approach. We call *didactical map* (DM) an analogue of the concept map for TODU with the modifications as follow.

First, we count the levels down up, i.e. the DM is supposed to be designed from the bottom to the top (bigger level number means higher level as in Figure 3). Next, let us define the nodes of the DM as *atoms*, which stands for ‘atomic didactical components of the subject methodic’. An atom could be a theorem, ethic category, historical fact etc. We call *beams* the edges of the concept map. This is to emphasize the multicomponent type of the relations between the atoms. The reformulation is not just a technical act but also a meaningful reconsideration of the didactical process of concept mapping. E.g. a beam that links the atoms (Hristo Botev) [TheFamousPeople, 2019] and (Bulgarian National Revival) in a DM could contain the knowledge in different levels about: the role of the hero in several stages of the revolutionary movement, some momentous of the historical period, his education etc. Any beam component reflects some specifics of the educational context in which the particular TODU takes place. In a DM, a beam can be either direct (d-beam), or crossing (c-beam).

In this connection, let us list several areas from a topic-oriented process of education in which the didactical mapping is potentially useful. It:

- illustrates the hierarchy and relationships between different elements of a topic;
- facilitates the perception and memorization of key elements (bases and beams) as fewer characters achieve greater information saturation;
- allows easy modification and restructuring in already constructed map fragments;

- allows for ideas to be shared, formalized quickly and accurately, and correctly displayed when preparing documentation;
- helps to clarify the logical consistency of presentation in the separate elements;
- allows evaluating more precisely the complexity of a topic, to define its complex structure, to decompose separate elements.

Developing the Theory

We are going to make more technological the advantages listed in section 4 by introducing two quantitative attributes related to the new atom-beam structure of the concept map matrix: distance and weight. We consider the educational objectives of a TODU in relation with some components of a synthetic competence that is developed via integrated approach [Lazarov, 2013]. Thus, the attributes of a beam should reflect the observed indicators for progress of these components of the synthetic competence. E.g. let the qualitative indicators I_k be the attributes of a d-beam, $k = 1,2,\dots,5$. Let us assign the benchmark i_k to I_k for the transition from the origin atom to the beam-connected one. We take the value $J = \sum_{k=1}^{5} i_k$ as a quantitative measure of the transition calling it beam-value. We take the beam-value instead of the corresponding link from the definition of the similarity index sim. When stating the indicators, it is important to consider the direction of the educational process. This means that any indicator reflects some elements of the progress in the transition from an atom to the succeeding one, thus the beams should be thought as vectors.

Distance in DM

We are going to invert the ontological similarity taking into account the structure of the concept map matrix in Figure 1 but with down up counting of the levels. Let C' and C'' be two atoms in the first (bottom) level both connected with an atom C^* (it could be also C' or C'') and having a genus root C (it could be C^*); let δ be the level number of C and let d' and d'' equal the sums of the beam-values J for the beams connecting C^* with C' and C'' respectively.

I. Suppose the DM is an ontology graph-tree, i.e. there are no c-beams. In this case, we define *ontological distance* as:

$$D_C(C', C'') = \frac{d' + d''}{\delta} \ \& \ \forall C: (D_C(C', C'') = 0 \ \Leftrightarrow C' \equiv C'') \,.$$

Indeed, the values of sim are fractions (decimals) between 0 and 1, which makes difficult to feel the similarity. On the contrary, the ontological distance shows quite clear the relation between atoms. Here it is a kind of motivation to introduce the concept of ontological distance:

- the root C (the topic under consideration) is in the area of knowledge for which the concept map is intended;

- a hierarchical chain of atoms, which are common for C' and C'', is formed up from C^* to C;

- down the atom C^* there is a split into two different subject branches, hierarchically arranged to any of the atoms C' and C'';

- the farther away C' and C'' are from C^*, the larger the distance;

- the higher the root C in the map, the shorter the distance.

The consideration is that a long distance reflects some differences in fundamental details of the atoms and vice versa – small distance means conceptual proximity. Another feature of D is its root sensitivity. This means that $D_C(C', C'')$ changes when taking the genus root C in different levels of the map. Figuratively, the distance decreases when upgrading the ontological tree "up". Let us clarify this by an example. Consider C' stands for ‘pekinese’, C'' stands for ‘bulldog’. Now taking $C \equiv C^*$ to be ‘dogs’ $D_C(C', C'')$ seems to be large, but taking C to be ‘pets’ (keeping C^* for ‘dogs’) the distance become smaller. Mathematically, larger level of the genus root increases the denominator of D, hence diminish the fraction. Didactically, it is crucial for the effectiveness of a teaching-learning the atoms of the educational process to be immersed in a deeper origin.

II. Now suppose that there are c-beams in the DM. The ontological distance in the case “there exists a c-beam between C' and C'' with the beam-value J” we define as

$$D_C(C', C'') = \frac{J}{\delta}.$$

III. In both cases I and II, if C' and C'' are in different levels, we consider δ as the relative level with respect to the lowest one among the levels of C' and C''.

The ontological distance could be extended on cases that are more sophisticated. For instance, when “C' and C'' are indirectly connected with c-beams” or “C' and C'' are indirectly connected with c-beams and d-beams”. In these cases, one should combine the ideas in stating $D_C(C', C'')$ in the two basic cases of DM presented in this section.

Weight of DM

We follow the Novak and Govin's ideas for the score of a concept map outlined above by modifying the algebraic sum. We will form the characteristic *weight of the DM* by considering levels, bases and beams with specific coefficients that depend on the structure of the respective component.

Let us evaluate the level structure of a graph-tree. In binary branching, each level doubles the number of atoms in the above level. If there is an atom in the top level then there are two atoms in the level below, four atoms in the level, which is two levels below and so on. The total number of atoms in the DM of a binary-graph-tree type with m levels equals $2^m - 1$. Thus the weight of the map receives $L = 2^m - 1$ points from the level-map-characteristic. We adopt this simplest binary structure as a medium model. Indeed, the graph-trees that are more sophisticated could have different type branching in any particular knot, e.g., single or triple, which in average give binary branching.

An atom t participates with an individual coefficient B_t. Suppose there are k_t elements of already studied material and n_t elements new material, which should be studied. The material of the atom-structure could be of rather different type: experimental work, learning a poem by heart, drawing a figure etc. We assume that the points, which this particular atom contributes to the map weight are $B_t = k_t + 2n_t$ in total. Motivation: In an atom, the usage of concepts requires an appropriate time resource and complicates the synthesis in the same time. Moreover, introduction of a new concept requires additional technological time for operationalization.

A beam t will contribute its beam-value J_t to the map weight. Here we make no formal difference between d-beams and c-beams. However, the complexity of the beam could be taken into account by stating the benchmarks i_k.

Thus, we define *weight W of a DM* having p atoms and q beams as

$$W = L + \sum_{t=1}^{p} B_t + \sum_{t=1}^{q} J_t.$$

Weight is a measure of the permissibility of a particular didactic scheme described by the DM. It is clear that when considering the values L, B, J, some interdependence and overlap are allowed. For example, concepts at one atom affect the indicators of both the incoming beam and the outgoing one. We do not consider this as a drawback, but rather as a collateral feature of the integrated approach.

Editing DM

A need for editing could appear when two independently designed DM are formally combined into one. E.g., if the total weight is too big or appear too long distances. The following operations are possible:

- adding an atom;
- removing an atom;
- replacement of an atom;
- adding a beam;
- removing a beam;
- adding an attribute to a beam;
- removing an attribute from a beam.

Any of these operations has a didactic price. It can reflect directly the corresponding action, but it can also take into account factors, which are determined by the educational context. The price can be in the direction of gain (usually associated with weight or distance reduction) or loss (increase in performance values).

Sample Didactical Mapping Technology for TODU

DM structurizes the didactic process: the hierarchy of concepts is clarified, and hence the order of their introduction; links are clarified, cross-domain links are rethought. The technology of DM design could be as follows:

1) a team of teachers (usually two subject teachers and a coordinator, preferably an ICT teacher) traces the topic for integrated consideration;
2) the subject teachers independently highlight the topic from their methodology, any one separately clarifies the educational goal and indicators of progress;
3) the subject teachers prepare a preliminary list of atoms, taking into account the concepts learned, the already developed students' knowledge and skills, as well as the need of introducing new concepts, building additional skills, forming the positive attitude;
4) atoms are ranked in levels, only d-beams are defined;
5) the subject teachers design concept maps independently of each other, where the topic is at the top;
6) the team jointly draws up a DM, uniting and editing the two DMs into one;
7) the weight of the DM is calculated and, if necessary, ways of reducing it are sought;
8) when drawing up the DM, the distance between the atoms is taken into account and, if appropriate, the DM is edited, possibly by adding c-beams.

Example

In this section, we present the didactical mapping in a TODU that could be titled ‘Hristo Botev – translator and propagandist’ (it was dedicated to the 140th anniversary of the Bulgarian hero Hristo Botev) in the experimental classes on mathematics at the 125th Secondary School – Sofia. Botev cannot be fully studied in neither history nor literature, since Hristo Botev's multifaceted appearances does not fall within the narrow scope of any particular school subject. Figure 3 gives us an idea of Botev's hypostases. The left branch of Botev’s subtree is subject of history, the right branch is studied in literature school course. (The examples in the rectangles on the bottom of this DM could be different and they do not form a separate level in the DM.)

The activities of the translator and the propagandist do not easily fit together, insofar as the translators are tasked with retelling some source, and the mission of the propagandists is to persuade through (their own) ideas. Thus, there is no place in the presented DM for a little known side in Botev's activity: the translation of the textbook "Lessons for the first four arithmetic rules" by Mikhailov (a detailed review of this translation was made in [Penkov and Chobanov, 1958]). Here, Botev's genius manages to synthesize a brilliant translation with propaganda, which gave us a reason to organize a corresponding TODU devoted to the celebration of his anniversary.

Figure 3. Ontology graph-tree of Botev's hypostases

Let us calculate the ontological distance between the atoms C' (publicist) and C'' (poet) taking $J = 5$ for all beams (a motivation for this value is given in the next paragraph). We have C^*(Botev), C (Bulgarian National Liberation Movement), $d' = d'' = 10,\ \delta = 4$. Thus $D_C(C', C'') = \frac{10+10}{4} = 5$. Let us note that if extract the branch 'Botev' from the Bulgarian-National-Liberation-Movement didactic map as an autonomous DM for which $C \equiv C^*$ (hence $\delta = 3$), then the distance increases: $D_{C^*}(C', C'') = \frac{20}{3} \approx 6.7$.

Another (third) hypostasis, namely ‘Hristo Botev – translator and propagandist’, cannot be find under the atom C^* (Botev) in Figure 3. It needs two new atoms to be added: (translator) in level III and (mathematician) in level IV. The transition (translator) → (mathematician) requires studying a fragment of Botev’s biography by Zahariy Stoyanov [1888]. The bottom row of this branch will be supplemented by the example [Lessons for the first four arithmetic rules]-rectangle. Let us consider the beam from the atom (translator) to the atom (mathematician) with attributes (indicators) I_k , $k = 1,2,\dots,5$, as follows: I_1 – knowledge, I_2 – skills, I_3 – attitude, I_4 – transferability of KSA, I_5 – multifunctionality of KSA. Let the benchmark of any indicator takes value from the set $\{0; 1; 2\}$. We state: $i_1 = 2$ (fundamental knowledge in Bulgarian history for that period); $i_2 = 1$ (basic math skills); $i_3 = 1$ (moderate positive attitude); $i_4 = 0$ (no requirements); $i_5 = 1$ (basic). Thus $J = 5$.

The introduction of the new atoms allows the atoms (revolutionary) and (enlightener) to be connected in a natural way through intersecting beams with beam-values $J_{1,2}$. Taking the genus root (Bulgarian National Liberation Movement), the distance in the chain

$$\{(\text{publicist}) \overset{J_1}{\leftrightarrow} (\text{mathematician}) \overset{J_2}{\leftrightarrow} (\text{poet})\}$$

calculated via c-beams equals $\frac{J_1}{\delta} + \frac{J_2}{\delta} = \frac{5}{4} + \frac{5}{4} = 2.5$. Comparing with the distance $D_C(C', C'') = 5$, which we obtained above, this is twice less. Such shortage of the distance could be done with a minimal didactic price: preparatory independent work for the students to study the text [Stoyanov, 1888], one additional lesson as well as monitoring the transferability and multifunctionality of the knowledge. In the Appendix, we give the test-item from the term test that checks these transferability and multifunctionality.

The shortage of the distance $D_C(C', C'')$ derives increasing the DM weight. Suppose the weight of the DM in Figure 3 is W (it could vary depending on the structure of the level III). Let W' be the weight of the new DM, which supplements the initial one with the branch

$$\{(\text{Botev}) \overset{J_4}{\leftrightarrow} (\text{translator}) \overset{J_3}{\leftrightarrow} (\text{mathematician})\}$$

and the c-beams that were defined above. Thus

$$W' = W + \sum_{t=1}^{2} B_t + \sum_{t=1}^{4} J_t = W + 26.$$

Skipping the motivation, here $B_1 = 2$ stands for the atom (translator), $B_2 = 4$ stands for the atom (mathematician). We take $J_3 = J_4 = 5$ analogously to the calculated above beam-values for $J_{1,2}$. Fortunately, the increase of the weight does not affect the didactic price in this particular case.

Conclusions

The didactical mapping we presented is still a theoretical model in an experimental phase. We tried to recruit some teachers for implementing the didactic map in their pedagogical practice but we met their tacit resistance. Such attitude is understandable on this stage of our research. However, our plans include elaborating map-shell that automatically calculates the weight and the ontological distance when stating the benchmark values. In our opinion, the technology and the model for didactical mapping could help in deeper reconsideration of didactic process of integrated approach in topic-oriented educational units.

The positive results reported in [Giouvanakis et al., 2016] encourage us to develop more TODU that integrate different school subjects, providing a priory calculated quantitative characteristics of the particular TODU. Further clarifications of the parameters are expected to be done by taking into account the educational context in which the teachers are going to implement the integrated approach.

Acknowledgement

The paper is published with a partial support by the Educational and Research Program *Chernorizec Hrabar* (www.math.bas.bg/ch).

Bibliography

[Giouvanakis et al., 2016] Giouvanakis, T., Evaggelos, K., Mpakavos, A. & Samaras, H. (2016) Creative thinking through concept mapping. International Journal for Cross-Disciplinary Subjects in Education, Volume 7, Issue 1. pp 2705-2711

[Kitano et al., 2008] Kitano, M., Montgomery, D., Vantassel-Baska, J. & Johnsen, S. (2008) Using the National Gifted Standards for PreK-12 Professional Development. NAGC&CEC-TAG. Corwin Press, Thousand Oaks, California.) p 124

[Lazarov, 2013] Lazarov, B. Developing Synthetic Competence along Individual Educational Trajectory. The Korean Society of Mathematical Education. Proceedings of the International Conf. on Math. Edu. on Creativity & Giftedness (August 9–10, 2013), pp 251-262.

[Lazarov, 2018] Lazarov B. Topic-oriented Upgrade of Subject-oriented Educational System. In Tarasenkova, N. (Eds). Current Status and Prospects of Mathematical Education: Monograph. Budapest, Hungary: SCASPEE. (2018) pp 37-53

[MON, 2015] Ministry of Education and Science. Regulation No 5 of 30.11.2015 for the general education. Darzhaven Vestnik, I. 95 of 08.12.2015. [in Bulgarian]

[Novak and Gowin, 1984] Novak,J. & Gowin,D. Learning how to learn. Cambridge, UK: Cambridge University Press, 1984 pp 15-54.

[Penkov and Chobanov, 1958] Penkov, B, & Chobanov, I. Botevite „Urotsi za parvite chetiri aritmeticheski pravila“. Fiziko-matematichesko spisanie., 1958, №1, pp 61-67. [in Bulgarian]

[Stoyanov, 1888] Stoyanov, Z. Hristo Botyov. Opit za biografia. [in Bulgarian] http://www.slovo.bg/showwork.php3?AuID=149&WorkID=3878&Level=3 (last visited Feb 2020)

[TheFamousPeople, 2019] https://www.thefamouspeople.com/profiles/hristo-botev-5649.php (last visited Nov 2019)

Appendix

Item 10 from the term test. Hristo Botev is the author of the following problem:

'Assen and Peter liberated Bulgaria from the yoke of the Greeks in 1190. After 206 years, Bulgaria fell under Turkish rule. In what year did our country lose its freedom?'

Write down the years in which the Bulgarian kingdoms fell under Turkish slavery, and the names of the respective Bulgarian rulers in which this occurred.

The Kingdom of Tarnovo:

Year .... King ....

Kingdom of Vidin:

Year .... King ....

Which kingdom does Botev have in mind? ....

Answers. Tarnovo Kingdom: 1393 Tsar Ivan Shishman. Kingdom of Vidin: year 1396 Tsar Ivan Sratsimir. Botev refers to the kingdom of Vidin.

Authors' Information

Borislav Lazarov – Professor, Institute of Mathematics and Informatics – Bulgarian Academy of Sciences

e-mail: lazarov@math.bas.bg

Major Fields of Scientific Research: Didactical models, Integrated approach in education

ANALYSIS OF APPROACHES FOR EFFECTIVE APPLICATIONS DEVELOPMENT IN MONOLITH, SERVICE-ORIENTED AND MICROSERVICE ARCHITECTURE

Yurii Milovidov

Abstract: *The paper presents the features of the using the microservice architecture in the development process. The advantages of this approach are illustrated in comparison with the traditional monolithic approach. The connection between the use of the microservice architecture and the ability of the team to work within agile development methodologies is shown. Creating a new product is often a risk. And choosing the right architecture is an essential step toward success. If you're considering between a monolithic, service-oriented and microservice architecture, this paper will help you make the right choice.*

Key words: *Service Oriented Architecture, microservices, microservice architecture, efficient development, agile methodologies, Quality of Service.*

ITHEA keywords: *D.2 Software engineering, D.2.1 Requirement/Spesification, D.2.13 Reusable Software.*

Introduction

Software architecture modeling is an important part of the development process. The choice of a specific architectural solution depends on many factors, in particular, the purpose of the system being developed. Despite the variety of approaches to the architecture of applications, in the framework of modern projects that need easy scalability, lately, microservices have been preferred [Shadija D. Rezai M. Hill R.].

The term Microservice Architecture describes a software development technology in which an application is a collection of loosely coupled services. Services are built around business logic and are independently deployable. Each service works in its own process and interacts with other services using lightweight mechanisms, such as, for example, HTTP. Moreover, the development of each service can be carried out independently of the others. This allows you to organize the development process flexibly and distributed.

The main goal of this paper is to determine and describe the conditions necessary for the implementation of effective development based on microservice architecture.

Theoretical foundations

Service-Oriented Architecture (SOA) is a software architecture where functionality is grouped around business processes and packaged as interoperable services. SOA also describes IT infrastructure which allows different applications to exchange data with one another by passing data from one service to another as they participate in business processes. The aim is a loose coupling of services with operating systems, programming languages and other technologies which underlie applications. SOA separates functions into distinct units, or services, which are made accessible over a network in order that they can be combined and reused in the production of business applications.

The microservices approach has become a trend in recent years. Microservices have become the software architecture of choice for business applications. Initially originated at Netflix and Amazon, they result from the need to partition, both, software development teams and executing components, to, respectively, foster agile development and horizontal scalability.

The concept of monolithic software lies in different components of an application being combined into a single program on a single platform. Usually, a monolithic app consists of a database, client-side user interface, and server-side application. All the software’s parts are unified and all its functions are managed in one place. A monolithic architecture is comfortable for small teams. Components of monolithic software are interconnected and interdependent, which helps the software be self-contained. This architecture is a traditional solution for building applications.

Related works

Interoperability Constraints in Service Selection Algorithms [Kaczmarek P.]

Tasks: Existing methods and algorithms of service selection focus on the formal model of the complex service refraining from interoperability issues that may affect the integration process as services are deployed in heterogeneous runtime environments. Service integration in SOA requires resolution of interoperability issues.

Approaches for solving task: Alternative services realizing the same functionality differ in non-functional, Quality of Service (QoS) attributes such as performance, reliability and price. It is required that services with optimal QoS values are selected, while constraints are satisfied. Considering the problem, the author proposes a methodology that includes interoperability analysis in service selection algorithms. The data structures of service selection algorithms are extended with dedicated constraints that represent interoperability between services.

If QoS constraints are removed from the selection problem, a simplified model is created, in which services are selected considering solely the utility function. In this paper, interoperability is analyzed on the communication protocol level, which corresponds to Level 1 (Connected) in the LISI metric or Level 2 (Data/Object) in the LCI metric. Was assumed that if services are interoperable on that level, it is possible to exchange data between them in the context of CS construction.

The proposed extensions of existing models with interoperability constraints enable a possibly straightforward application of existing selection algorithms.

As a part of the research, was implemented the Work-flow Integrator system that enables selection of services during composite service development. The aim of the system is to support developers in the selection process depending on given requirements on utility function, QoS attributes, and interoperability constraints.

Were used workflow applications as a concrete implementation of the composite service concept, which enables us to leverage existing solutions in service description and invocation.

SOAQM: Quality model for SOA applications based on ISO 25010 [Franca J. Soares M.]

This paper addresses one gap in researches about SOA. There is a lack of quality models specific for SOA based on most recent ISO 25010. In this paper authors propose a quality model for SOA applications based on quality attributes proposed in ISO 25010. ISO 25010 is currently the most complete version of quality models of ISO. In order to make a proper evaluation of software, relevant quality characteristics of a software product have been proposed in many quality models, including ISO standards. The task of the paper is to analyze which important contributions were aggregated into the new ISO 25010 regarding SOA applications. In addition, this analysis aims to determine if limitations perceived in ISO 9126 were solved in the most recent ISO 25010. Another proposal of this paper is to investigate the applicability of ISO 25010 to SOA applications.

Approaches for solving task:

All characteristics and sub-characteristics proposed by ISO 25010 were analyzed regarding the real applicability to SOA. Accordingly, authors establish for each characteristics the degree of applicability to SOA. The next step of this analysis is to define how these characteristics can be adapted to the SOA context.

Was proposed a degree of importance that defines which characteristics are more relevant during the SOA application development process. Seven volunteers answered a questionnaire to define the degree of importance. The answer varies from 1 to 5 following the Likert Scale. Was created a Table that presents the results of the analysis of ISO 25010 characteristics and their applicability to SOA projects. First two columns show characteristics and sub-characteristics proposed by ISO 25010. Third column presents the degree of importance of each sub-characteristic with relation to the SOA context. Highlighted sub-characteristics in green are important or very important quality attributes for SOA applications. Sub-characteristics highlighted in yellow are not so important. Quality sub-characteristics considered less important or even irrelevant to SOA applications and therefore were highlighted in red. Finally, the last column presents the likely reasons that justifies the results.

All quality attributes proposed by ISO 25010were analyzed in this research from the SOA perspective using both experts opinion and literature review. As a result, most of quality attributes proposed by ISO 25010 may be applicable for SOA. On the other hand, ISO 25010 has a generic nature and therefore some quality attributes proposed do not apply to SOA domain.

Service Consumer Framework - Managing Service Evolution from a Consumer [Feuerlicht G. Tran H.]

In this paper, was described initial proposal for Service Consumer Framework that attempts to address this problem by providing resilience to changes in external services as these services are evolved or become temporarily unavailable. Service evolution has been the subject of recent research interest, however the focus of these efforts has been mainly on developing methodologies and tools that help service providers to manage service evolution. There is a pressing need to develop consumer-side methodologies and tools to address these issues from a service consumer perspective.

Approaches for solving task

Was described a proposal for a Service Consumer Framework (SCF) that attempts to address this problem by providing resilience for client applications to changes in external services as these are evolved or become temporarily unavailable. The SCF framework uses a combination of service adaptors and a service router to protect client applications from external changes. Evolution of services is supported by using service adaptors that transform service request and response messages between internal and external services. Service router determines which external services are evoked at runtime, based on their availability and pre-defined processing rules stored in the Service Repository.

The SFC Framework is illustrated in Figure 1 and consists of three layers: Process Layer, Adaptor Layer and Service Layer. The Service Layer incorporates both internal and external services and defines their native interfaces.

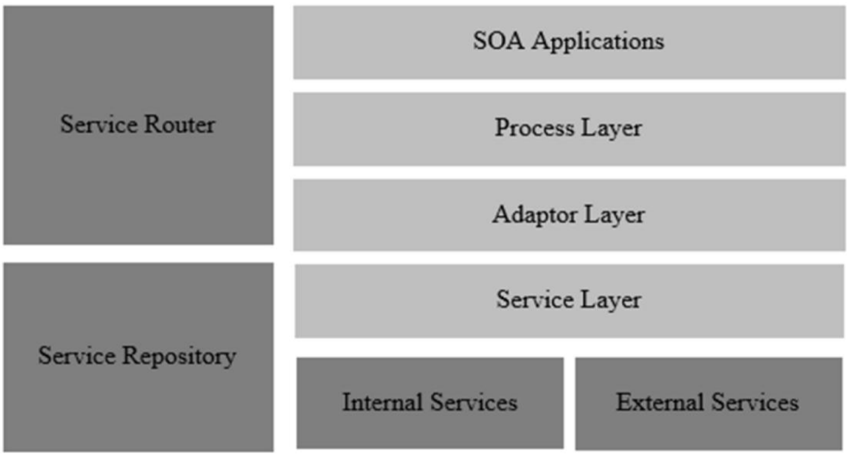


Figure 1. Service Consumer Framework

The Adaptor Layer contains adaptors that translate requests between the native services (e.g. PayPal payment service) and the corresponding internal services. The Process Layer defines processes that are implemented using the service router and processing rules stored in the service repository.Service router determines at runtime, which services are evoked based on their availability and pre-defined processing rules. Service repository maintains information about services allowing substitution of services with equivalent functionality to avoid service interruptions.

From a Monolith to a Microservices Architecture: An Approach Based on Transactional Contexts [Nunes L., Santos N., Rito Silva A.]

Currently, there is a large number of monolith applications that are being migrated to a microservices architecture. This article proposes the identification of business applications transactional contexts for the design of microservices. Therefore, the emphasis is to drive the aggregation of domain entities by the transactional contexts where they are executed, instead of by their domain structural inter-relationships. Additionally, authors propose a complete workflow for the identification of microservices together with a set of tools that assist the developers on this process.

Task

The purpose of this paper is to, based on a comparative analysis of different architectural approaches, determine and describe the conditions necessary for effective development on the basis of microservice architecture.

Research methods

This section describes the advantages of microservice architecture compared to monolithic architectures and describes the proposed principles for organizing effective development in the project team developing applications in microservice architecture.

Monolithic architecture

A monolithic architecture is comfortable for small teams. Components of monolithic software are interconnected and interdependent, which helps the software be self-contained. This architecture is a traditional solution for building applications, but some developers find it outdated.

Pros of a monolithic architecture:

- Simpler development and deployment;
- Fewer cross-cutting concerns;
- Better performance.

Cons of a monolithic architecture:

- Codebase gets cumbersome over time;
- Difficult to adopt new technologies;
- Limited agility.

SOA

A service-oriented architecture (SOA) is a software architecture style that refers to an application composed of discrete and loosely coupled software agents that perform a required function. SOA has two main roles: a service provider and a service consumer. Both of these roles can be played by a software agent.

Pros of SOA

- Reusability of services;
- Better maintainability;
- Higher reliability;
- Parallel development.

Cons of SOA

- Complex management;
- High investment costs;
- Extra overload.

Microservice architecture

Microservice is a type of service-oriented software architecture that focuses on building a series of autonomous components that make up an app. Unlike monolithic apps built as a single indivisible unit, microservice apps consist of multiple independent components that are glued together with APIs.

The microservice architectural style is an approach to developing a single application as a suite of small services, each running in its own process and communicating with lightweight mechanisms, often an HTTP resource API. These services are built around business capabilities and independently deployable by fully automated deployment machinery. There is a bare minimum of centralized management of these services, which may be written in different programming languages and use different data storage technologies.

To explai the microservice style it's useful to compare it to the monolithic style: a monolithic application built as a single unit. Enterprise Applications are often built in three main parts: a client-side user interface (consisting of HTML pages and javascript running in a browser on the user's machine) a database (consisting of many tables inserted into a common, and usually relational, database management system), and a server-side application. The server-side application will handle HTTP requests, execute domain logic, retrieve and update data from the database, and select and populate HTML views to be sent to the browser. This server-side application is a monolith - a single logical executable. Any changes to the system involve building and deploying a new version of the server-side application. The price of microservices pays off by reducing the costs of reducing team productivity while complicating the system. This is shown in Figure 2, which is presented in article [Fowler M.].

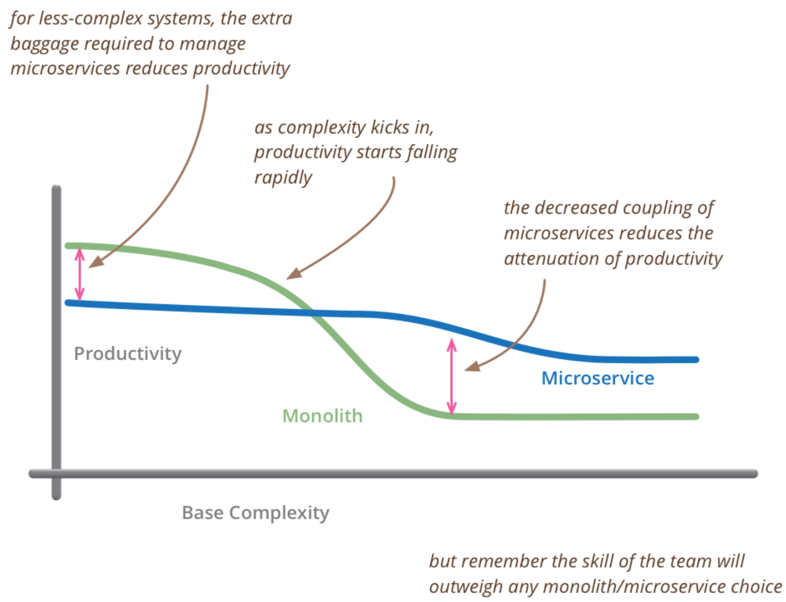


Figure 2. Dependence of team productivity from base complexity.

Pros of microservices

- Easy to develop, test, and deploy. The biggest advantage of microservices over other architectures is that small single services can be built, tested, and deployed independently. Since a deployment unit is small, it facilitates and speeds up development and release. Besides, the release of one unit isn’t limited by the release of another unit that isn’t finished. And the last plus here is that the risks of deployment are reduced as developers deploy parts of the software, not the whole app.
- Increased agility. With microservices, several teams can work on their services independently and quickly. Each individual part of an application can be built independently due to the decoupling of microservice

components. For example, you may have a team of 100 people working on the whole app (like in the monolithic approach), or you can have 10 teams of 10 people developing different services for the app. Increased agility allows developers to update system components without bringing down the application.

- Ability to scale horizontally. Vertical scaling (running the same software but on bigger machines) can be limited by the capacity of each service. But horizontal scaling (creating more services in the same pool) isn’t limited and can run dynamically with microservices. Furthermore, horizontal scaling can be completely automated.

Cons of microservices

- Complexity. The biggest disadvantage of microservices lies in their complexity. Splitting an application into independent microservices entails more artifacts to manage. This type of architecture requires careful planning, enormous effort, team resources, and skills.
- Security concerns. In a microservices application, each functionality that communicates externally via an API increases the chance of attacks. These attacks can happen only if proper security measurements aren’t implemented when building an app.
- Different programming languages. The ability to choose different programming languages is two sides of the same coin. Using different languages make deployment more difficult. In addition, it’s harder to switch programmers between development phases when each service is written in a different language.

When using microservices, the role of the project manager comes to the fore, since many different teams work on different parts of the application. In order to synchronize the work of microcommands and avoid problems, you should hire competent project managers.

Software architect Simon Brown suggested what a flexible software architecture would look like [14]. His description of flexible architecture fits quite well on microservice architecture. In his opinion, an architecture style built using small loosely coupled components (services) that interact with each other to respond to a user's request will help provide flexibility. Services can be modified and tested in isolation or even torn out and replaced depending on changing requirements, new components can be added and scaled, if necessary.

According to Agile Manifesto, successful and efficient teams make the necessary decisions together, and the recommended team size is at least 3 and no more than 9 people.

In this way, Agile provides support for effective collaborative development in the microservice architecture.

Leading technology organizations such as Amazon, Netflix, and Apple use the microservice architecture to provide the flexibility to quickly adapt and respond to their customers' requests.

The figure 3 shows the differences between the three different architectures that were discussed above.

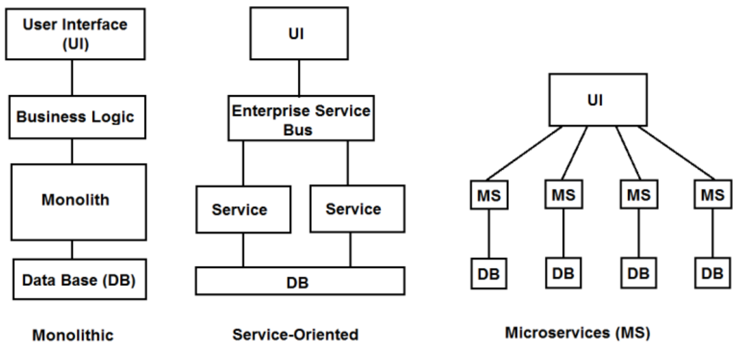


Figure 3. The differences between the three different architectures.

Monolithic applications consist of interdependent, indivisible units and feature very low development speed. SOA is broken into smaller, moderately coupled services, and features slow development. Microservices are very small, loosely coupled independent services and feature rapid continuous development.

Results

The essence of three different architectural approaches is shown below.

The monolithic model isn’t outdated, and it still works great in some cases. Some giant companies like Etsy stay monolithic despite today’s popularity of microservices. Monolithic software architecture can be beneficial if your team is at the founding stage, you’re building an unproven product, and you have no experience with microservices. Monolithic is perfect for startups that need to get a product up and running as soon as possible.

The SOA approach is best suited for complex enterprise systems such as those for banks. A banking system is extremely hard to break into microservices. But a monolithic approach also isn’t good for a banking system as one part could hurt the whole app. The best solution is to use the SOA approach and organize complex apps into isolated independent services.

Microservices are good, but not for all types of apps. This pattern works great for evolving applications and complex systems. Choose a microservices architecture when you have multiple experienced teams and when the app is complex enough to break it into services. When an application is large and needs to be flexible and scalable, microservices are beneficial.

Conclusions

By comparing three different architectural approaches, we can draw the following conclusion:

Choosing the right architecture is a daunting task.

When developing a software product, you need to seriously approach the choice of architectural solution. Microservices seem to be a suitable choice for the implementation of large systems involving long-term development. However, microservice architecture carries a lot of risks. In this case, the success of the project depends largely on the availability of an experienced system architect and project manager.

A team using microservice architecture should be formed on the basis of business opportunities. One of the prerequisites for the successful implementation of the microservice architecture project is the competence of the team.

Bibliography

[Kaczmarek P. 2012] Interoperability Constraints in Service Selection Algorithms, Proceedings of the 7th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE-2012), pp 23-32 ISBN: 978-989-8565-13-6 – Access mode: https://scitepress.org/papers/2012/39805/39805.pdf

[Feuerlicht G. Tran H. 2014] Service Consumer Framework - Managing Service Evolution from a Consumer Perspective, Proceedings of the 16th International Conference on Enterprise Information Systems (ICEIS-2014), pp 665-672 ISBN: 978-989-758-028-4 – Access mode: https://www.scitepress.org/papers/2014/49766/49766.pdf

[Franca J. Soares M. 2015] SOAQM: Quality model for SOA applications based on ISO 25010, Proceedings of the 17th International Conference on Enterprise Information Systems (ICEIS-2015), pp 60-70 ISBN: 978-989-758-097-0 – Access mode: https://www.researchgate.net/publication/276202563_SOAQM_Quality_model_for_SOA_applications_based_on_ISO_25010

[Nunes L., Santos N., Rito Silva A. 2019] From a Monolith to a Microservices Architecture: An Approach Based on Transactional Contexts.: Software Architecture ECSA 2019 pp 37-52 ISBN: 978-3-030-29982-8 – Access mode: https://link.springer.com/chapter/10.1007/978-3-030-29983-5_3

[Shadija D. Rezai M. Hill R 2017] Towards an understanding of microservices, 23rd International Conference on Automation and Computing (ICAC) ISBN: 978-1-5090-5040-6 – Access mode: https://ieeexplore.ieee.org/abstract/document/8082018

[Fowler M. 2015] Microservice Premium, – Access mode: https://martinfowler.com/bliki/MicroservicePremium.html

[Brown S. 2013] Coding the Architecture blog. , – Access mode: http://www.codingthearchitecture.com/2013/09/03/what_is_agile_software_architecture.html

Agile Manifesto – Access mode: http://agilemanifesto.org/

Authors' Information

Yurii Milovidov, *National University of Life and Enironmental Sciences of Ukraine, Kyiv, scientific Department of Programming Technologies. Senior Lecturer.*

E-mail: milovidov.email.ua

Programming technologies, web-design, Internet – technologies.

TABLE OF CONTENTS