Capturing Expert Knowledge to Guide Data Flow and Structure Analysis of Large Corporate Databases

Gergő Balogh[†], Tamás Gergely[†], Árpád Beszédes[†], Attila Szarka[‡], Zoltán Fábián[‡]

[†] University of Szeged, Department of Software Engineering
Dugonics tér 13, 6720 Szeged, Hungary
E-mail: {geryxyz,gertom,beszedes}@inf.u-szeged.hu

[‡] Clarity Consulting Kft.
Erzsébet Királyné útja 29/b, 1145 Budapest, Hungary
E-mail: {Szarka.Attila,Fabian.Zoltan}@clarity.hu

Abstract: Maintaining and improving existing, large-scale systems, that are based on relational databases has proven to be a challenging task. Among many other aspects, it is crucial to develop actionable methods for estimating costs and durations in the process of assessing new feature requirements. This is a very frequent activity during the evolution of large database systems and data warehouses. This goal requires the analysis of program code, data structures and business level objectives at the same time, which is a daunting task if made manually by experts. Our industrial partner started to develop a static database analysis software package that would automate and ease this process in order to make more accurate estimations. The goal of this work was to create a quality assessment model that can effectively help developers to assess the data flow (lineage) quality and the database structure quality of data warehouse (DWH) and online transaction processing (OLTP) database systems. Based on the relevant literature, we created different models for these two interconnected topics, which were then evaluated by independent developers. The evaluation showed that the models are suitable for implementation, which are now included in a commercial product developed by our industrial partner, Clarity.

Keywords: database systems; data warehouses; cost estimation; software quality models; data flow; database structure; data lineage

1 Introduction and Motivation

Maintaining and improving existing large-scale systems that are based on relational databases has proven to be a challenging task. For example, from an IT operation manager's point of view, it is crucial to develop professional methods to estimate costs and durations when a new feature requirement needs to be assessed.

These estimations are usually performed by senior experts (e.g. senior database developers), who walk through main system components, data structures and program code to review everything that needs to be modified. These experts need to understand not only the nature of the change itself, but all of the affected systems as well. Understanding a large system such as an all-round corporate Data Warehouse (or DWH) system is never easy but estimating the impact of a medium sized change on the system's operation is even harder. Any method that can help experts to better understand what actually happens behind the lines of program code is a large step towards a more accurate and faster (cheaper) estimation of the above mentioned consequences of modifications [8] [18] [22].

Our goal is to help the experts of our industrial partner, Clarity Consulting Ltd. during the analysis of large industrial database systems (OLTP databases or DWHs). As dynamic or online analysis of these is rarely feasible due to compliance or IT security reasons, we established a static analysis methodology that provides an objective toolkit for data lineage (data flow) analysis for DWH systems and database structure quality assessment for OLTP systems. To assemble this framework, we made a manual assessment of the existing workflows Clarity uses to perform such analysis, and set up a measurement model that captures the experts' knowledge. We also sought practical ways to reduce the time needed to understand data flows and database structures in these large software systems.

Clarity Consulting Ltd. is a privately founded Hungarian consulting company, established in 2001, specialized in management and IT-related consultancy services, as well as the implementation of IT solutions for large companies. The company covers the full range of solution delivery to solve business problems (consulting, design, implementation, testing, deployment). The company also develops front-ends (e.g. CRMs, transaction systems, special applications), and database-driven systems (like campaign management databases, data warehouses). They also have products for data cleaning, migration, and DWH-supportive systems. Their clients are typically multinational companies and government-related agencies, e.g. MNB (the central bank of Hungary), Aegon (insurance company), MAK (Hungarian State Treasury). There are several huge systems managed or developed by Clarity. For example, one of their large-bank systems serving a local bank's customers presents 40 man years of development, in a 9 year life-cycle, that is currently used by 2000 active users and serves 8 business areas; another large DWH also developed by Clarity for 10 years incorporates tens of thousands of tables and more than a million columns.

In this paper, we present the work performed in order to assemble a methodology and quality model that can, through semi-automatic analysis, help cost estimations of Clarity's staff (including developers, project managers, quality maintenance staff). In particular, we present our experiences regarding the capture and encoding of the expert's knowledge in the resulting quality model, and thus, hopefully, help other organizations facing similar challenges.

The rest of the paper is organized as follows. We elaborate on the related literature in Section 2. In Section 3, we describe the whole process we followed during this research, then in Section 4 we introduce the steps related to the model construction in detail. We show how the results were validated in Section 5 and describe the resulting quality models in Section 6. In Section 7 the threats to validity are elaborated and we sum up in Section 0.

2 Related Literature

There are several studies that deal with the assessment of database quality. Chaudhuri et al. [1] provided a method to identify faulty program parts and bad programming practices. They used dynamic database logs to detect bad practices about the data flow of the system that prevent client and server-side query optimization. This method attempts to identify several databases related problems (e.g. setting the number of queues returned by queries, which reduces data traffic; identifying dynamic data within the queries and marking parameters that allow the server to perform more accurate optimizations; formulating suggestions and detecting potential indices based on successive queries). In another work the authors described how tools can support these kind of analyses [2]. Chen also combined static and dynamic techniques to help developers to improve the performance of the database-intensive systems by 3-88% [3]

Wassermann et al. [20] used static analysis techniques to detect type-related problems in dynamically generated queries. Their method is based on context-free language analysis and is able to detect the problems like type conflicts, incorrect variable types, or context-specific deviations. Haraty et al. [8] presented a method to prepare Control and Data Flow Graphs for database systems. They used column level entities and defined different connection types to represent data query and data manipulation instructions. Dasgupta et al. [4] examined embedded SQL queries, and combined the data flow in SQLs and non-SQL parts of the system. This way they could perform a more precise analysis of data flows.

Genero et al. [7] concentrated on the structure of the database and defined metrics derived from its static schema to describe its quality. Wedemeijer [21] and Papastefanatos et al. [11] [12] used schema-based metrics to describe the amount of changes made to the database schema. Wedemeijer used metrics designed for the different types of the database schema, while Papastefanatos et al. used graph based metrics. In our work, we derive separate but interconnected metrics from the data flow information and from the database structure.

Another way to assess the quality of a database system is to define different rules and check whether these rules are observed; or count how many times they are violated. Delplanque et al. [6] implemented a tool called *DBCritics* that analyzed DDL instructions and checked them against some rules. Their work focused on the

schema evolution problem, but many of their rules can be used on the schema itself to check the actual quality of the schema. Rodic et al. [17] dealt with the data quality processes of data warehouses and provided a method to implement data quality rules. Their rules can be used to mark or correct the defective records in the selected tables. The data quality process is integrated into the ETL process enabling automatic, quick and correct operation. Using their research results, a rule generator used in the industrial (banking) sector was prepared. They used several rules that checked whether components that ensure data integrity are present in the database schema. Nagy and Cleve also used rules to detect bad smells in (embedded) SQL queries, based on code, database schema and data analysis [10].

The most complete list of database quality rules we found are collected on the *red-gate community pages*[1]. The lists included several Microsoft SQL Server specific rules, but most of them were either general or could be used as a template for general or Oracle specific rules. We defined rules only for the database structure (i.e. for DDL instructions), and not for the data manipulation instructions[2].

In our research, we only found loose definitions of high-level database metrics. Although high-level definitions of the so called *QoX* (Quality of X) metrics [5] [13] are also known, and some of these metrics are used in other areas of software development, we are not aware of low-level (implementation-close) definitions, which would be generally accepted. In the studies, researchers generally interpret and clarify these definitions themselves, but we did not find an official or de facto standard.

Dayal et al. [5] defined several high-level quality metrics for database systems based on the regular high-level software metrics. These metrics capture the quality of the software from different human-understandable points of view. Simitsis et al. [19] and Pavlov [13] have examined these metrics and their relations to the classic software metrics. Herden [9] also published a methodology including several high-level quality criteria to assess the quality of the database system. Piattini et al. [14] [15] [16] conducted several studies where they measured low-level attributes of the schema to express high-level quality attributes of the database. In our methodology introduced below, we partially relied on the high-level metrics introduced by Simitsis et al.

1 https://www.red-gate.com/hub/

2 This was required by our industrial partner, as while there are several static analysis tools available for Oracle PL/SQL code, no suitable tool was found for database structures (DDLs). Data flow analysis is a different topic from this point of view because information could be extracted from DMLs to construct data lineage graphs.

3 Description of the Manual Assessment Process

Clarity Consulting relies on a manual method (supported by some automated analysis and measurement tools) to assess the resources and time required to perform modifications on a database system or a DWH. In practice, simple call graphs are created and affected database objects are identified. Simple metrics are also computed by supporting tools to enable estimation of modification cost and duration. This process consumes a lot of resources (expert and computation time), because a manual walkthrough of the code and database structures are required. The current method is adequate but due to the high ratio of manual analysis, it is expensive and slow. During the process, only those parts of the system are examined that are considered important (making the analysis more subjective than an automatic analysis). The actual assessment consists of the following steps:

1. Read and understand the change request (e.g. rewriting the structure, optimizing a component for performance, or inserting a new one)
2. Identify the relevant system objects (programs, modules, interfaces)
3. Examine the affected table structure
4. Manually analyze the affected program code
5. Estimate the amount of resources required for the development (estimate the size and complexity of the code need to be constructed)
6. Estimate the resources required for testing, documentation and go-live
7. Cross-validate and confront different estimations

In this process, the examination of the table structure is supported by tools at low-level. In step 3, graphs that describe the structure of the database and the database-related data flow are constructed to help the experts in understanding how the workflows operate. Then, the experts examine the program code, ETL processes, their complexity, etc. and make estimations on the development.

Our goal was to support these steps using automated software tools. According to our analysis, several low-level attributes of the database structure and of the data flow could be automatically detected based on the graphs and the source code itself. These low-level values could then be used directly in the experts' estimations, but it is also possible to further help the experts by computing high-level metrics. The estimations would still be made by the experts, but the high-level metrics have a more direct connection to the experts' estimations than the low-level metrics. In other words, part of the experts' estimation knowledge could be captured by the model that states how to compute the high-level metrics from the low-level ones.

4 Capturing Expert Knowledge

As elaborated earlier, our main goal was to help the experts with a (semi-)automatic system. The developers will use this system to accomplish repetitive audit tasks and it will allow users to retrieve objective data on the quality through reproducible measurements. To achieve this, the underlying models should contain the aggregated knowledge of the experts. Our industrial partner, Clarity included our findings in features of its database analysis software package. This software has two modules: DALIA (Database Lineage and Impact Analyzer) and DEXTER (Database Structure Analyzer).

DALIA is a database lineage tool capable of parsing Oracle PL/SQL code while identifying data connections implemented in DML statements. The data flow graph is constructed using static analysis, i.e. the extraction process does not require access to the working database instance (and its actual data); it works offline on uploaded structure and PL/SQL code extracts. DALIA can display a data flow graph thus enabling the evaluation of data dependencies at the database schema, table or column level (or a mix of these). The edges and vertices of this graph are labelled with calculated values of high-level concepts like maintainability or complexity. It also supports impact analysis by estimating the efforts and costs of planned modifications. The edges and vertices of this graph will be labelled with estimated values of high-level concepts.

DEXTER is a database structure analysis tool that is able to measure the quality of database models or structures. Its operation is similar to static source code analysis tools, but it does not work on programming languages, rather on the implemented database structure itself. It evaluates most of the database objects (tables, indexes, triggers, etc.) against simple or complex rules to gain an understanding of the quality of the database model. It helps to understand which actions could be made to increase the performance, maintainability or scalability of the database model.

DALIA and DEXTER use the above-mentioned models to compute certain metrics and rules to help our goal of supporting the comprehension and development of database-intensive systems.

We performed a multi-phase expert's knowledge capturing process to construct these models. Figure 1 shows an overview of the whole process. It started with the evaluation of the related literature to retrieve suggestions and best practices. Based on the findings, we conducted a series of informal interviews to collect general information and opinion from the developers. Then, based on the literature and the experts' knowledge, we defined the metrics and rules that would help the experts' work of assessing the data lineage and structural quality of a system.

As the analysis and the models should be implementable in DALIA and DEXTER, the project management of Clarity were also involved in the rule and metrics definition phase. We defined low-level (directly measurable) and high-level (conceptual) metrics. Low-level metrics represent some objective,

quantitative attributes of the system that can be measured directly from the database model and/or the source code (including number of rule violations of each rule). There were several rules that required some parameters; we asked the experts to set them based on their experience. High level metrics cover some quality-related conceptual properties and their values are computed from other (usually low-level) metrics. These computational processes (which, in our case, are weighted linear aggregations) are called the *quality model*. We performed a survey to collect the experts' knowledge regarding these models (i.e. to provide weights to the models).

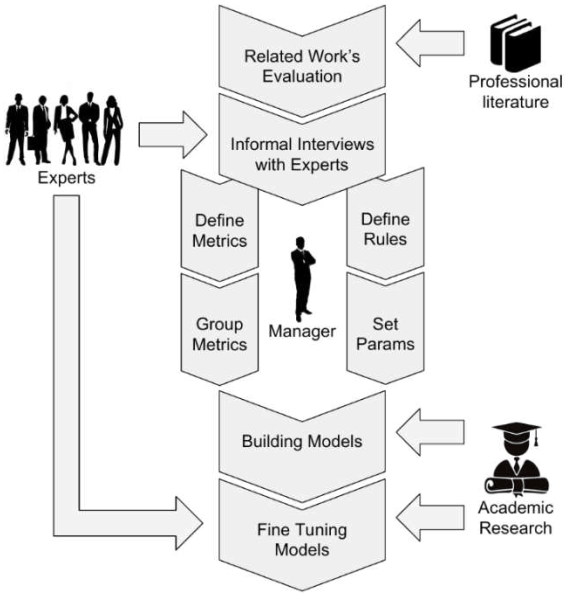


Figure 1
Overview of Expert's Knowledge Capturing Process

The following sections provide details about each step of our knowledge capturing process.

4.1 Evaluation of Related Literature

During this phase we evaluated 173 articles published by more than 400 different authors published at 41 distinct forums (conferences or journals). We grouped these papers into 89 topics according to their major contribution. These topics included: *conceptual model, data flow, data uncertainty, database as a service, database complexity, database generation, database migration, database summarization, database testing, ER model extraction, keyword search, metrics, object oriented database, online tuning, parallel join, performance evaluation,*

persistence, query comprehension, query optimization, query prediction, query validation, reverse engineering, schema analysis, schema expansion, schema filtering, schema summarization, standard, summarization, UML, workload estimation.

In this phase we sought answers to the following questions:

- Which methods and techniques are used to analyze the database structure and its data flow connections?
- Are there any widely accepted techniques, metrics, or rule sets for analyzing databases?

The first question was addressed in Section 2 of this article by listing the relevant related works. During the evaluation of these papers we could not find any widely accepted, general methods or techniques for database analysis. There are several key concepts and methods commonly used in various works, like data flow analysis and the use of the so-called *QoX* (Quality of X) metrics, but we could not find any (de facto) standards for this topic. These common methods are usually context sensitive and constructed to solve a specific problem, or they lack any strict definition. To conclude this phase, related work provided a good general basis for the research, but there were several details that had to be worked out in order to meet the industrial needs of Clarity.

4.2 Informal Interviews with Developers

To collect preliminary information about the experience of the developers, we performed informal interviews at the office of Clarity, with all participants present at the same time. Altogether 15 experts were involved: 5 juniors with less than 4 years of experience, 5 seniors with 4-10 years of experience, and 5 experts with more than 10 years of experience. We have also assigned different roles to the participants: there were 5 back-end, 2 front-end, and 3 lead developers, 2 testers, and 3 project managers.

The goal of these guided discussions was to collect the expert's professional viewpoints in various topics related to the evolution challenges of database constructs and technologies. Based on these data we were able to define the following topics Clarity was interested in:

- Generated PL/SQL statements
- EXECUTE IMMEDIATE statements and related code chunks
- Various graph topology descriptors, like count of cycles
- Error handling and dblink-connected database items
- Data flow connections of different database items

- Various weights based on the complexity of the implementation
- Property distribution among layers of the database

4.3 Metric Categorization and Rule Sets

The next steps of the knowledge capturing process were the definition of the metrics and rules. Rules were defined to check suspicious locations and constructs in the structure of the database. These rules can directly help developers to improve the quality of such systems, but cannot be directly used in the quality model. We have derived metrics from these rules by counting how many times were they violated, and these rule violation metrics were then used in the quality model. Data connection metrics were defined to quantitatively capture different properties of data, flowing in the system from column to column. Note, that we have not defined metrics or rules that used both database structure and data flow information; this was a technical decision made by Clarity to enable the standalone operation of the two modules, DALIA and DEXTER.

In the following, we briefly describe the different metric and rule categories.

4.3.1 Data Structure Rules

There can be several constructs in a database and its structure that can cause the loss of some quality attributes, e.g. lack of indexes can hinder performance or very similar names can hinder understandability. These constructs or properties can be detected by analyzing the database structure whether it satisfies some predefined rules or not. We have collected a large set of general and database system specific rules that were used in the practice, selected and modified the most relevant and appropriate ones to fit the needs of Clarity (relying on the opinion of its experts). There were some rule violations that required some parameters to be set. We've done this together with Clarity's team. Developers and managers were also asked to define the importance of the different rules on a scale from 1 (least important) to 5 (most important). Finally, we defined 53 rules in five categories.

Table rules define rules about the tables and the relationships between them. This rule set includes rules like a check for isolated tables (without foreign key and referencing tables), a check for using proper column types (e.g. use DATE, not CHAR or INTEGER), a check for using too many or too few indexes (which may reduce performance). The above mentioned rules are considered to be important (level 4) by the experts; the average importance of the rules in this group is 3.2.

The *key rules* capture primary and foreign key related issues and, if kept, they help in maintaining data integrity. Rules like checking for the existence of primary keys, using monolithic primary keys, or checking whether a foreign key refers to a key are included. The mentioned rules have high-level of importance (at least 4), and the average importance of the group is 3.4.

The *type rules* provide help for safe and efficient use of data types. This group includes rules for checking deprecated types, checking whether fixed or variable length types are more appropriate at a certain place, or proposing Unicode types where those seem to be more appropriate. Although some of these rules have the highest importance, the average importance of the group is 3.1.

The *syntactic rules* help to improve readability which is necessary for understanding the code during a manual analysis, and to avoid bad coding practices that would otherwise make maintenance activities more error prone. This set includes rules like checking the use of reserved words as identifiers, whether indexes have descriptive names, if there are very similar identifiers in use, or whether the same name is used for several elements (in different contexts). As following these rules can severely reduce the time required for understanding the system, many of them have high importance (with an average of 3.3).

Other rules include various rules that do not belong to any of the above categories but are still important for the quality assessment of a database. These rules suggest, for example, to use static database models (do not change the structure during operation), to avoid using the Entity-Attribute-Value model, or to restrict column values with additional tables and foreign keys instead of constraints. These rules have lower importance in general, as they can be reasonably ignored in certain systems (however, Clarity experts feel them appropriate).

4.3.2 Data Structure Metrics

Beside the rules, we defined metrics that capture some quantitative properties of the database structure. We have two groups of such metrics: one considering the different elements (like tables, views, columns, indexes, procedures, keys), and the other one considering the relations (like connected components, foreign keys). For both groups, we defined metrics to count the number of given elements or relations, and, if available, average and total number of them. For example, the number of indexes can be counted for each table but can also be summarized for the database, and an average index count per table can also be computed.

4.3.3 Data Connection Metrics

Beside the structure, data connections and data flow of the database also affects its quality. There are several metrics that quantitatively express the data flow related attributes of the system. In this work, we have defined 83 data connection metrics to be measured. Some of the metrics are local, meaning that they can be computed for smaller structures (e.g. for a stored procedure) and then be easily aggregated for larger structures (e.g. for all of the program code that exist in the database system). Other metrics are global, meaning that they have to be computed directly for larger structures (and cannot be aggregated from the values of smaller ones).

We can also make difference between *source code level* and *low-level* metrics; we defined 23 and 60 of these, respectively. Source code level metrics are directly

measurable numeric characteristics that express certain attributes of the source code. These metrics are based on the relationships between elements, size, and complexity. Low-level metrics express the number of elements, the number of relationship between them, or the proportion of these. Each such metric has a domain, which specifies what type of items of the data flow graph the metric is computed for (column, table, schema, database). These metrics are based on "EXECUTE IMMEDIATE commands", generated code, complexity, graph description, usage, and grouping.

4.4 High Level Concepts

On one hand, rules and low-level metrics express some well-defined properties of the database. On the other hand, high-level metrics are proposed in the literature to express some concepts like maintainability or reliability of the system. In this work, we have used 8 high-level: MAINTAINABILITY, RELIABILITY, ROBUSTNESS, TECHNICAL COST, SCALABILITY, LOG RATE, FLEXIBILITY and INTEGRITY. These high-level *QoX* metrics can be used by the experts to assess the overall quality of a database system and estimate the cost and duration of a modification.

4.5 Fine Tuning Quality Models

We use low-level metrics and rule violation counts to estimate the value of high-level concepts. There are many publications that elaborate on what features can be used (and how) to calculate high-level metrics of a database. This computational process is called the *quality model*. However, papers rarely provide specific models, instead, they examine how the automatically computed values of low-level metrics and the manually assessed values of high-level metrics are correlated in real database systems. As the goal of the quality model is to compute high-level (abstract) concepts using low-level (measurable) metrics, each model determines a kind of aggregation of lower-level metrics to the high-level ones.

As noted earlier, in our case, source code metrics describe the PL/SQL code itself, low-level and high-level metrics are interpreted on data flow graphs and on the database model, while rule violation checking is also applied on the latter. Several source-code based metrics are used to weigh the edges of data flow graphs, hence an abstraction level shift can be observed between the code level and the low-level data flow metrics since the data flow graph could be interpreted as an abstraction over the source code. The formal definition of higher-level metrics in our model is provided by the aggregation of lower-level metrics. These principles define a three-level model in the case of the data connection (source, low-level metrics and high-level concepts) and a two-level model in the case of the database structure.

For each lower-level metric to be aggregated on a higher level, we defined three higher-level metrics: the median, average and standard deviation values of the corresponding lower-level metric values. This model makes it possible to connect

any source-, low- and high-level metrics. The user is allowed to set the weights between any metric pairs to fine-tune for an exact situation or problem. These weights enabled us to capture the connection between various concepts commonly known to database experts. To define initial values for these weights we asked the developers and experts to weigh every connection between two consecutive levels. Results are shown on Figures 3, 4 and 5.

5 Evaluation

Our evaluation process consists of several phases (see Figure 2). All of our results were checked by Clarity's experts to correct any misinterpretation. Finally, we asked several independent experts from another company working in a similar domain to express their opinion about our model and methodology.

Clarity's first impressions about the methodology were that it is well thought out and allows developers to easily understand the concepts. The data flow metrics and database structure rules made the implementation of client inquiries smoother.

To collect the opinion of independent experts we used an electronic survey, which contained six sections and 20 questions. The questions covered all of the relevant steps of our validation process. These steps are highlighted with green background in Figure 2. The survey took 30-50 minutes to complete by an expert. We used open-ended questions to collect personal ideas without any bias from our side. The closed-ended questions targeted rankings and often meant single-choice questions.

We collected seven responses altogether. Two of these were given by database users and five were filled by developers. Interviewed experts' solutions for system evaluation and cost estimation varied from person to person, and although they have mentioned the usage of various (semi-)automatic analysis methods, everyone emphasized the importance of the connection with the original developers (face-to-face discussions, documentation). As turned out, the most useful techniques to solve these problems are the well-known static analysis of source code and inspection of the structure of database. The interviewed persons assigned similar scores to data connection analysis either based on data (1.29 of 0-3) or source code (1.33 of 0-3).

It suggests that although the data-based connection analysis plays an important role, the source code based heuristics are not discarded by experts. Two of the key entities in our model, namely directly measurable metrics and high-level concepts also got higher appreciation when solving the above-mentioned problems. We think that the lower scores of the intermediate or derived metrics (like ratios and other compound measures) are explained by their low interpretability.

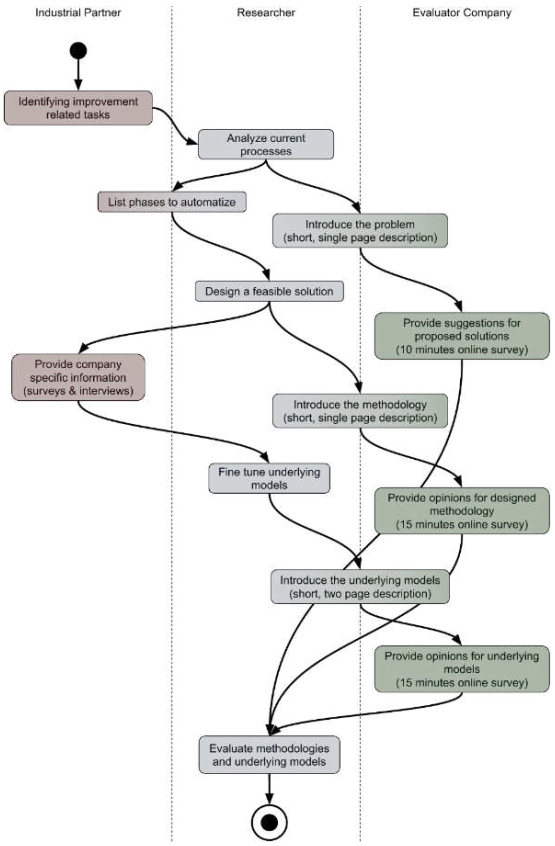


Figure 2
Overview of the Evaluation Process

We also asked experts about the usefulness of various information collecting techniques for different roles. They agreed that in most cases the opinion of the developers and experts could be useful, regardless of the actually used information collection method (e.g. interviews, discussions or surveys). Opinions of users have medium scores, while the value of project managers' concept shows a more divert picture. There are some areas where they placed the importance of developers and experts higher (for example using formal interviews or surveys to capture data about previous assessment processes). In general, we could conclude that our subjects prefer informal methods and techniques over strict or formal options.

Almost all of the asked experts preferred social or personal oriented methods (like interviews and discussions) to collect and fine tune the information required for the automatic assessment model. They also mentioned various documentation and other auxiliary artifacts as main information sources. Evaluation by independent experts or already validated frameworks were also favored. These concepts coincide our previously advised and executed knowledge capturing processes.

6 Quality Models

In this section, we describe our final quality models. As mentioned earlier, it was a strategic decision by Clarity to build separate models for the database structure and the data flow of database systems. Clarity plans to utilize the incorporated knowledge during its quality assessment and cost estimation processes implemented in their database analysis software package (DALIA and DEXTER). Some properties of the possible target systems (on which the analysis will be performed) are shown in Table 2.

Table 2
Target Systems

	count of			time	
Systems	tables	fields	LOC	Age	ver.
DWH 1 (Large Bank)	40.00K	1200K	3500K	15	3rd
DWH 1 (Large Bank)	9.00K	360K	700K	10	2nd
CRM 1 (Large Bank)	0.35K	10K	30K	10	5th
CRM 1 (Large Insurance Company)	0.90K	8K	100K	10	3rd

6.1 Measuring Database Structure

We have defined rules and metrics for the database structures. Rules provide direct feedback on potential problems but cannot be directly used in the model. However, rule violation counts can be, as described in Section 4.3.1. In our model, the values of these metrics directly affect the high-level metrics.

Our model is the following: we compute a weighted average of the low-level metrics as

$$H(I) = \frac{\sum_{L \in LLM} w_{H,L} L(I)}{|LLM|}$$

where H(I) is the high-level, L(I) is the low-level, normalized metric value for the I item, LLM is the set of low-level metrics, and $w_{H,L}$ is the weight of metric L in the model of metric H. The weights were set by the experts. They were asked to fill questionnaires about how strongly the number of rule violations and the attributes captured by the metrics affect the high-level concepts (strongly, weakly, not at all) and in what direction (positively, negatively). The answers were summarized and a weight between -1 and 1 were assigned to each metric-concept pairs.

In Figures 3 and 4 two examples for the answers given by the experts for metrics of items and relations are shown. As can be seen, while the larger number of entities in a database negatively affects the high-level concepts in most of the cases (red lines in Figure 3), the number of relations usually aid them (green lines in Figure 4).

Before the rule violation counts or the low-level metric values are used, normalization is done. Normalization can be different for the different metrics, but for all of them it is done in a way to eliminate the size bias of the system (which practically means a division by some size related metrics).

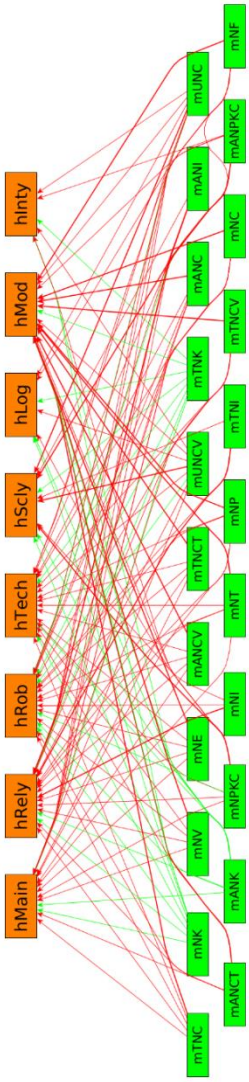


Figure 3
Example Database Structure Metric Model (low-level entity metrics). Red connections depict a negative, and green is a positive influence.

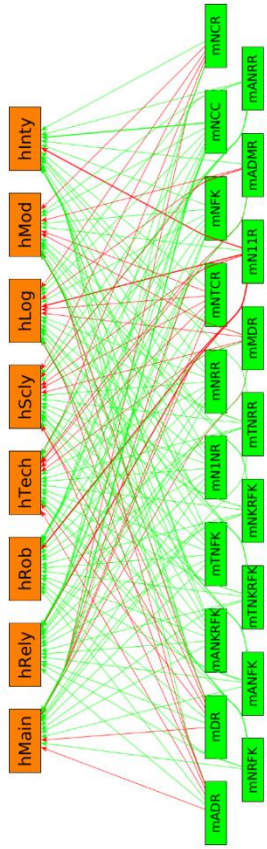


Figure 4
Example Database Structure Metric Model (low-level relation metrics). Red connections depict a negative, and green is a positive influence.

6.2 Measuring Data Connection

As the result of the above elaborated process, we constructed several models that help assess the quality of databases. Because we were not able to dynamically access the stored data, we decided to rely on the data connection graph (DCG) for the analysis. There are several types of components in relational databases; our models use four of these: columns, tables, schemas and databases (instances). In the constructed DCGs these components are represented by nodes. There are connections between these if at least one code chunk exists, which reads data from the source and presumably writes it into the target component. We created one global and 8 local models. The essence of the global model is to accumulate the *low-level global metrics* using a simple aggregation function. The result is a numeric descriptor of the whole system.

Local metrics can be divided into two groups according to their domains: metrics for the edges or nodes. They can be further divided by levels: columns, tables, schemas, and databases (instances). This grouping enables the construction of 8 independent local models to be built. These probabilistic models are based on the deviation of the values at the given level, and the aggregate values of these deviations are propagated towards the higher-level metrics.

As an example, the metric model for *connection between tables* is shown in Figure 5. The basic element of the models is a directed, non-circular graph (DAG) that describes the dependencies between each low-level metric and high-level characteristic. This graph forms the base of upward aggregation, where values of low-level metrics are determined first from the directly computable values, and then propagated along the edges up to the higher levels.

7 Threats to Validity

Although our methods and resulted models were accepted and approved by our industrial partner, there are some threats to validity of this work.

7.1 External Validity

Because of our goal was to provide a context specific system (and methodology to construct it), we do not have any data about the degree of generalization. The current phase of development and integration with the above-mentioned DALIA and DEXTER software systems made it difficult to produce any measurement on real life systems using the new model. Clarity plans to conduct such kind of empirical evaluation after the launch of the first version of the software tools based on our model.

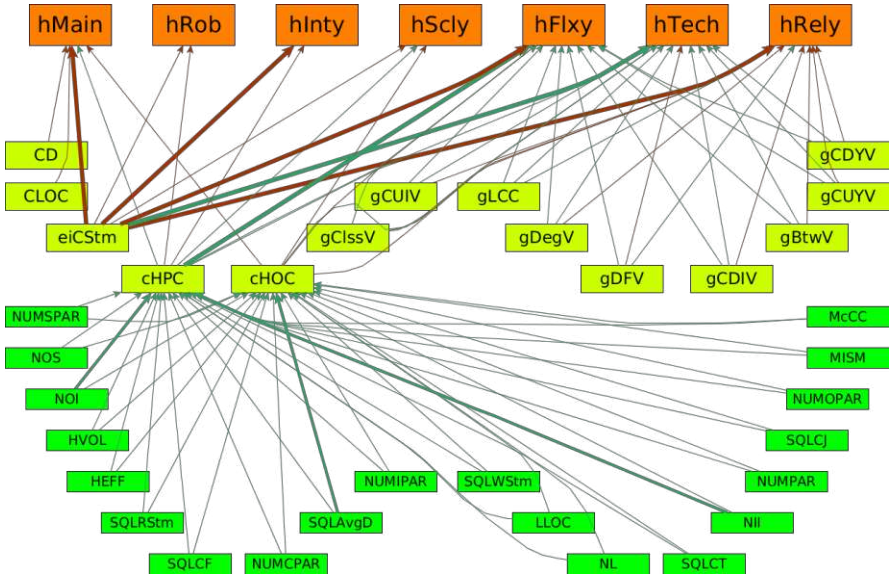


Figure 5
Metric Model for Connections between Tables

We used the previously mentioned survey to assess the validity of our methodology and our findings. We constructed the survey to minimize unintentional bias of opinions, but we could not eliminate this factor completely. The retrieved number of answers are quite low, which increases the chance of drawing insignificant or over-fitted conclusions. Note, that due to the ongoing development connected to this project and some other privacy considerations the number of potential subject audience were limited.

There were several open-ended questions in the survey. While our opinion is that these types of questions are useful to collect personal impressions and opinions, it could also lead to misinterpretation or subjective evaluation.

7.2 Internal Validity

Building a model where directly measurable metrics are used to estimate high-level, abstract concepts are prone to errors. These errors may emerge because the selected metrics could be unable to capture relevant information. We addressed this problem by carefully selecting and defining our metric sets based on the related works. These metrics were validated using informal interviews and guided discussions with developers and experts. We tried to minimize unintentional bias, but we could not eliminate this factor completely.

There are several parameters of the individual metrics and rules as well for the whole model. These properties are double-edged swords. They provide high-level

customizations increasing the scalability of the measurement system, but it is always possible to set these parameters to a sub-optimal value. We reduced this risk by asking the developers and experts to estimate these parameters using their experiences.

The choice of the aggregation method is another threat to validity. There are several ways to propagate the information to a higher level of abstraction. During the design of the model, there were two main factors needed to be taken into account. First, as the knowledge capturing was a direct process (i.e. we did not ``train'' the model, but asked the experts to assess its parameters), the parameters must have had a clear interpretation. The second constraint was the implementability and expected performance of the algorithms. This lead to our final decision to use the probability and distribution-based methods in compound multi-level parts of the model, and a weighted average-based method in the case of simpler parts.

Conclusions

In this work, our goal was to create a methodology that helps the Clarity Company in the cost estimation of database system development and maintenance. We initially collected related works and existing solutions, and assessed the actual process and the needs of Clarity. Then, we defined different low-level metrics to be automatically measured on the database systems, together with high-level metrics and models. These help developers and managers to assess the quality of a database system and incorporate this knowledge in cost estimation tasks. The models were parametrized and fine-tuned based on discussions and interviews with Clarity experts. The resulting models were checked by independent developers, and finally accepted by Clarity, who will build them in into their static analysis systems.

Although the methodology and the models have been evaluated by independent developers, they have not yet been actually used: the process and the models were not applied on real systems, they were merely "statically" validated. To make a more thorough evaluation, we plan to use the models for a certain period in real projects, during which we will record different aspects of them (e.g. accuracy of estimations) and then compare them to projects estimated without this support.

Nevertheless, we believe that our experiences, reported in this paper, will help other organizations and teams working on similar initiatives and facing similar challenges.

Clarity has already included a part of the knowledge and results of this work in two modules of their database analysis software. These modules are DALIA (Database Lineage and Impact Analyzer) which utilizes the dataflow analysis results, and DEXTER (Database Structure Analyzer), which takes advantage of the results related to the database structure.

Although the methodology and the models have been evaluated by independent developers, this did not include the use of all of them: every process and the models were not thoroughly applied on real systems, but some relevant parts were applied on large DWH and CRM systems (see Table 5 above). To make a more thorough evaluation, we plan to use the models more exhaustively in real projects, while we record different aspects of them (e.g. accuracy of estimations) and then compare them to projects estimated without this support.

For further information on Clarity Consulting, please visit http://clarity.hu/en or the product page http://daliaflow.com about DALIA (Database Lineage and Impact Analyzer).

Acknowledgement

This work has been supported by GINOP-2.1.1-15 (Economic Development and Innovation Operational Programme) funded by the European Union.

References

[1] S. Chaudhuri, V. Narasayya, and M. Syamala: Bridging the application and DBMS profiling divide for database application developers. 33rd International Conference on Very Large Databases, 2007, pp. 1252-1262

[2] S. Chaudhuri, V. R. Narasayya, and M. Syamala: Database Application Developer Tools Using Static Analysis and Dynamic Profiling. IEEE Data Eng. Bull., 37(1), 2014, pp. 38-47

[3] T. H. Chen: Improving the quality of large-scale database-centric software systems by analyzing database access code. 31st IEEE International Conference on Data Engineering Workshops, 2015, pp. 245-249

[4] A. Dasgupta, V. Narasayya, and M. Syamala: A static analysis framework for database applications. IEEE 25th International Conference on Data Engineering, 2009, pp. 1403-1414

[5] U. Dayal, M. Castellanos, A. Simitsis, and K. Wilkinson: Data integration flows for business intelligence. 12th International Conference on Extending Database Technology Advances in Database Technology, 2009

[6] J. Delplanque, A. Etien, O. Auverlot, T. Mens, N. Anquetil, and S. Ducasse: CodeCritics applied to database schema: Challenges and first results. 24th IEEE International Conference on Software Analysis, Evolution, and Reengineering, 2017, pp. 432-436

[7] M. Genero, G. Poels, and M. Piattini: Defining and validating metrics for assessing the understandability of entity-relationship diagrams. Data and Knowledge Engineering, 64(3), 2008, pp. 534-557

[8] R. A. Haraty, N. Mansour, and B. A. Daou: Regression test selection for database applications. Advanced Topics in Database Research, Vol. 3, 2004, pp. 141-165

[9] O. Herden: Measuring quality of database schemas by reviewing–concept, criteria and tool. Oldenburg Research and Development Institute for Computer Science Tools and Systems, Escherweg, 2:26121, 2001

[10] Cs. Nagy and A. Cleve: A Static Code Smell Detector for SQL Queries Embedded in Java Code. IEEE 17th International Working Conference on Source Code Analysis and Manipulation, Shanghai, 2017, pp. 147-152

[11] G. Papastefanatos, P. Vassiliadis, A. Simitsis, and Y. Vassiliou: Design Metrics for Data Warehouse Evolution. 27th International Conference on Conceptual Modeling, 2008, pp. 440-454

[12] G. Papastefanatos, P. Vassiliadis, A. Simitsis, and Y. Vassiliou: Metrics for the Prediction of Evolution Impact in ETL Ecosystems: A Case Study. Journal on Data Semantics, 2(1), 2012, pp. 75-97

[13] I. Pavlov: A QoX model for ETL subsystems: Theoretical and industry perspectives. In ACM International Conference Proceeding Series, Vol. 767, 2013, pp. 15-21

[14] M. Piattini, C. Calero, and M. Genero: Table oriented metrics for relational databases. Software Quality Journal, 9(2), 2001, pp. 79-97

[15] M. Piattini, C. Calero, H. A. Sahraoui, and H. Lounis: Object-relational database metrics. L'Objet, 7(4), 2001, pp. 477-496

[16] M. Piattini, M. Genero, and L. Jiménez: A metric-based approach for predicting conceptual data models maintainability. International Journal of Software Engineering and Knowledge Engineering, 11(6) 2001, pp. 703-729

[17] J. Rodic and M. Baranovic: Generating data quality rules and integration into ETL process. ACM Twelfth International Workshop on Data Warehousing and OLAP, 2009, pp. 65-72

[18] R. Saint-Paul, G. Raschia, and N. Mouaddib: General purpose database summarization. 31st International Conference on Very Large Databases, 2005, pp. 733-744

[19] A. Simitsis, K. Wilkinson, M. Castellanos, and U. Dayal: QoX-driven ETL design: Reducing the cost of ETL consulting engagements. 35th SIGMOD international conference on Management of data, 2009, pp. 953-960

[20] G. Wassermann, C. Gould, Z. Su, and P. Devanbu: Static checking of dynamically generated queries in database applications, ACM Trans. Softw. Eng. Methodol. 16(4), 2007, p. 14

[21] L. Wedemeijer: Defining metrics for conceptual schema evolution, Lecture Notes in Computer Science, Vol. 2065, 2001, pp. 220-244

[22] X. Yang, C. Procopiuc, and D. Srivastava: Summary graphs for relational database schemas, Very Large Data Bases 4(11), 2011, pp. 899-910

A Fuzzy Brain Emotional Learning Classifier Design and Application in Medical Diagnosis

Yuan Sun, Chih-Min Lin*

Department of Electrical Engineering, Yuan Ze University, Tao-Yuan 320, Taiwan, sungirl609@126.com
*Corresponding author, cml@saturn.yzu.edu.tw

Abstract: This paper aims to propose a fuzzy brain emotional learning classifier and applies it to medical diagnosis. To improve the generalization and learning ability, this classifier is combined with a fuzzy inference system and a brain emotional learning model. Meanwhile, different from a brain emotional learning controller, a novel definition of the reward signal is developed, which is more suitable for classification. In addition, a stable convergence is guaranteed by utilizing the Lyapunov stability theorem. Finally, the proposed method is applied for the leukemia classification and the diagnosis of heart disease. A comparison between the proposed method with other algorithms shows that this proposed classifier can be viewed as an efficient way to implement medical decision and diagnosis.

Keywords: brain emotional learning classifier; neural network; medical diagnosis

1 Introduction

Computational intelligent models have been widely used over the past decade, some popular approaches are fuzzy inference systems and fuzzy control systems []1-7], evolutionary computing techniques [8, 9] and cerebellar model articulation controllers (CMACs) [10-15]. Although, emotion was traditionally considered playing an insignificant role in intelligence for a long time, there has been a discovery of important value of emotion in human mind and behavior announced by J. E. LeDoux in 1990s [16, 17]. On the basis of this theory, a brain emotional learning (BEL) model based on neurophysiology was created by J. Moren and C. Balkenius in the early 21st century [18-20]. They have implemented a computational model of the amygdala and the orbitofrontal cortex, and tested that in simulation. Thus, there has been an increasing interest in constructing the model of emotional learning process in recent years. One of the successful implementations of this model, the Brain Emotional Learning Based Intelligent Controller (BELBIC), was introduced by C. Lucas et al. [21, 22], which became an effective method for control systems. This model is composed of two parts, mimics to the Amygdala and the Orbitofrontal cortex, respectively. The sensory

signals and an emotional cue signal are combined to generate the proper action regarding the emotional situation of the system. As an adaptive controller with fast self-learning, simple implementation and good robustness, the BELBIC has been developed to numerous applications, such as power systems [23-25], nonlinear systems [26, 27], and motor drives [28, 29].

In addition to engineering techniques, intelligent algorithms applied to medical diagnoses are also experiencing continuous attention in recent years. Early prediction with computer-aided diagnosis (CAD) acts a significant role in the clinical medicine, which can greatly increase the cure rate. Several approaches have been utilized for disease prediction or classification. For example, the logistic regression could be applied to medical diagnosis [30]. Moreover, a support vector machine [31, 32] and some other neural networks [33, 34] are also designed in this field. Most of the medical diagnosis problems are nonlinear and complex, which may be difficult to obtain definitely results even for a medical expert. Thus, it has motivated the design of more accurate and robust medical diagnosis CAD algorithms.

In this paper, we propose a classifier by incorporating fuzzy inference system with a BEL model, called FBELC model, which could not only offer a unique and flexible framework for knowledge representation, but also processes the quick learning ability of a BEL. Moreover, the online parameter adaptation laws are derived and the stable convergence is analysis for the proposed FBELC classifier. And the suggested model is designed for classifying different diseases and distinguishing presence or absence of heart disease using the posed conditions.

The specific contribution of this work involves the learning model with the definition of the emotional signal in the learning rules for classification problems. The desired goal could be obtained by appropriately choosing the system's emotional condition, which means, with the suitable definition of the reinforcing signal, the generalization quality and accuracy of prediction could be improved. The simulation results and analyses are performed to illustrate the effectiveness of the proposed model.

The remainder of the article is organized as follows: following this introductory section, Section 2 introduces the fuzzy brain emotional learning classifier, including the structure of the network, the learning algorithm and the convergence Analyses. The classifying process of diagnosis of diseases are described and the simulation analysis are shown in Section 3. Conclusion is given at last.

2 Fuzzy Brain Emotional Learning Classifier Design

This section reviews the structure of a fuzzy brain emotional learning classifier and the corresponding learning algorithm. As a whole, the general BEL model could be divided into two parts, the amygdala and the orbitofrontal cortex.

The former part receives inputs from the thalamus and from cortical areas, while the latter part receives inputs from the cortical areas and the amygdala. The output of the whole model is the output from the amygdala subtracting the inhibitory output from the orbitofrontal cortex. Usually, for control systems, the sensory input must be a function of plant output and controller output. Moreover, another reinforcing signal should be considered as a function of other signals, which is supposed as a cost function validation i.e. award and punishment are applied based on pervious defined cost function [35]. However, as a classifier, the sensory input and the reinforcing signal need to be reconsidered due to the characteristics of the input features and the cost function of the model. Detailed descriptions are presented in the following.

2.1 Structure of a FBELC

Fig. 1 shows a FBELC model with six spaces: the sensory input, sensory cortex, thalamus, orbitofrontal cortex, amygdala, and the output space. The fuzzy inference rules are defined as

$$If\ I_1 \text{ is } S_{1j} \text{ and } I_2 \text{ is } S_{2j}, \cdots, \text{and } I_{n_i} \text{ is } S_{n_i j}, \text{Then A} = V_{ij} \quad (1)$$

$$If\ I_1 \text{ is } S_{1j} \text{ and } I_2 \text{ is } S_{2j}, \cdots, \text{and } I_{n_i} \text{ is } S_{n_i j}, \text{Then O} = W_{ij} \quad (2)$$

$$\text{for } j = 1,\ 2,\ \cdots,\ n_j$$

where n_i is the number of input dimension, S_{ij} is the fuzzy set for the i-th input and j-th layer, and A is the output of amygdala and O is the output of orbitofrontal cortex; V_{ij} is the output weight of the amygdala and W_{ij} is the output weight of orbitofrontal cortex.

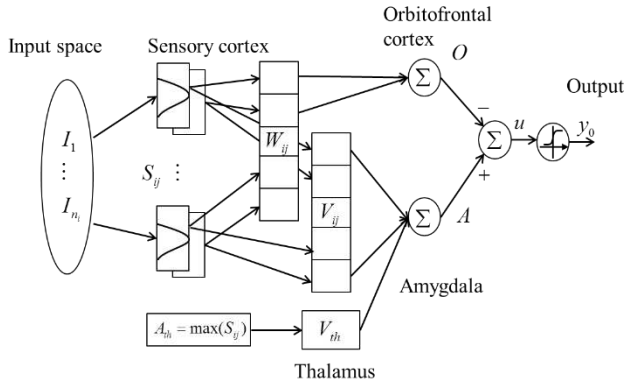


Figure 1
Architecture of a FBELC

The main functions of these six spaces are as follows:

a) Input space: For the input space $I = [I_1, \cdots, I_i, \cdots, I_{n_i}]^T \in R^{n_i}$, each input state variable, I_i can be quantized into discrete regions (called elements or

neurons). n_i is the number of input state variables. According to a given classification problem, it can also be considered as the feature dimension.

b) Sensory cortex: In this space, the sensory input should be defined and transmitted then to the orbitofrontal cortex and amygdala space. Different from the definition of the sensory input of a normal BEL model, this fuzzy brain emotional learning classifier only confirms a few activated neurons entering into the subsequent space. This is due to the consideration of improving the generalization ability and operating speed. Each block performs a fuzzy set excitation of the sensory input. Gaussian function is adopted here as the membership function, which can be represented as

$$S_{ij} = exp\left[\frac{-(I_i - m_{ij})^2}{\sigma_{ij}^2}\right], \quad \text{for} \quad i = 1,2,\cdots,n_i, j = 1,2,\cdots,n_j \tag{3}$$

where S_{ij} represents the *j*-th block and the *i*-th sensory with the mean m_{ij} and variance σ_{ij}.

c) Amygdala space: Each sensory fired value $S_{ij}\ is$ multiplied by a weight V_{ij}, then the *Amygdala space* output will be

$$A = \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij}\, V_{ij} \tag{4}$$

d) Orbitofrontal cortex: Similarly, the orbitofrontal cortex will be

$$O = \sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij}\, W_{ij} \tag{5}$$

Equations (4) and (5) correspond to the weighted sum outputs of the fuzzy rules in (1) and (2), respectively.

Both the weights V_{ij} and W_{ij} could be adjusted by certain learning rules. This will be addressed in Section *2.2*.

e) Output space: A single-output fuzzy brain emotional learning classifier is designed as

$$y_0 = 1/[1 + exp(-u)] \tag{6}$$

where u sums all the output from amygdala (including $A_{th}V_{th}$ term), and then subtracts the inhibitory outputs from the orbitofrontal cortex, as

$$u = A - O \tag{7}$$

2.2 The Learning Algorithm for FBELC

The learning rule of amygdala is given as follows [35]:

$$\Delta V_{ij} = \lambda_V(S_{ij} max(0, REW - A)) \tag{8}$$

where λ_V is the learning rate in amygdala. It needs to be emphasized that, the reward signal REW, as the emotional signal, is flexible and should be determined by the system according to the biological knowledge. Many terms contribute to the definition of the reward signal for control system. Exploring the prior researches, there are different formulas used for developing the reward signals. If the network is applied to control system, the reward signal is always connected with control error, frequency deviation [36], or some other control signals, such as sliding-mode control signal [37]. Likewise, for the classification problem, it is important to define this reward signal properly. And, the formula should be determined based on the knowledge of the problem. Meanwhile, the definition of the emotional signal and tunning of the gains are not complicated. In this work, the reward signal REW can be arbitrary function of error signal and output of the model, which is selected as:

$$REW = k_1 e + k_2 u \tag{9}$$

where e is the error of the model. k_1 and k_2 are both weighting factors, which are tuned for the expectation of error reduction and output separately. Usually, the value of k_1 is biggier than that of k_2 since the error of the model is always increasingly smaller in the process of learning.

$$\Delta W_{ij} = \lambda_W (S_{ij}(A - A_{th}V_{th} - O - REW)) \tag{10}$$

From (8), obviously, ΔV_{ij} has the same plus or minus with S_{ij}, which means, in the amygdala space, once an appropriate emotional reaction is learned, it should be permanent. However, in the orbitofrontal cortex, in order to inhibit or correct signals in the amygdala space and speed up the learning process toward to the expected value, ΔW_{ij} can be increased or decreased, shown from (10).

Define the output error as

$$e = T_0 - y_0 \tag{11}$$

where T_0 is the known target of samples and y_0 is the actual output.

In most situations of a BEL controller, the sensory input is computed in the sensory cortex and is directly sent to the orbitofrontal cortex and the amygdala space. There is not any learning process in the sensory cortex. In this architecture, from (2), updating of the mean and variance of the Gaussian function should be considered, which means the sensory cortex has the learning rules. Here, the gradient descent is applied to adjust the parameters and the adaptive law of mean and variance of Gaussian function are represented as

$$\Delta m_{ij} = -\lambda_m \frac{\partial E_0}{\partial m_{ij}} = -\lambda_m \frac{\partial E_0}{\partial e}\frac{\partial e}{\partial y_o}\frac{\partial y_o}{\partial u}\frac{\partial u}{\partial S_{ij}}\frac{\partial S_{ij}}{\partial m_{ij}}$$
$$= \lambda_m e \cdot y_o(1 - y_o) \cdot (v_{ij} - w_{ij}) \cdot S_{ij} \cdot \frac{2(I_i - m_{ij})}{\sigma_{ij}^2} \equiv \lambda_m e y_m \tag{12}$$

$$\Delta\sigma_{ij} = -\lambda_\sigma \frac{\partial E_0}{\partial \sigma_{ij}} = -\lambda_\sigma \frac{\partial E_0}{\partial e}\frac{\partial e}{\partial y_o}\frac{\partial y_o}{\partial u}\frac{\partial u}{\partial S_{ij}}\frac{\partial S_{ij}}{\partial \sigma_{ij}}$$
$$= \lambda_\sigma e \cdot y_o(1-y_o) \cdot (v_{ij} - w_{ij}) \cdot\cdot S_{ij} \cdot \frac{2(I_i - m_{ij})^2}{\sigma_{ij}^3} \equiv \lambda_\sigma e y_\sigma \quad (13)$$

where $E_o = \frac{1}{2}e^2$, $y_m = y_o(1-y_o) \cdot (v_{ij} - w_{ij}) \cdot\cdot S_{ij} \cdot \frac{2(I_i - m_{ij})}{\sigma_{ij}^2}$ and $y_\sigma = y_o(1-y_o) \cdot (v_{ij} - w_{ij}) \cdot\cdot S_{ij} \cdot \frac{2(I_i - m_{ij})^2}{\sigma_{ij}^3}$.

The learning objective could be divided into two parts. First, the parameters of the fuzzy part, m_{ij} and σ_{ij}, are adjusted by the gradient descent algorithm, shown in equations (12) and (13), respectively, which could minimize the training error, theoretically. And the other weights ΔV_{ij} and ΔW_{ij} are updated by equations (8) and (10), respectively, which are adjusted according to the structure of the brain emotional learning model. These special weights adaptation laws are the major feature that the brain emotional learning model distinguishes from the other intelligent algorithms.

A summary of this proposed FBELC model is given as below: first, original input signals are received from the features of samples. After introducing sensory input and reward signal by (2) and (9) respectively, the sensory input S_{ij} is used to form the thalamus input. From (3), a maximum term of the sensory input is selected as A_{th}. Then, signals are entered to the orbitofrontal cortex and the amygdala space by (4) and (5), respectively. By this process, the S_{ij} signal is used for both the orbitofrontal cortex and the amygdala space; however, A_{th} signal is only used for amygdala space. Before giving the total output, a comparison between the orbitofrontal cortex and the amygdala space is generated from (7). Moreover, in order to obtain a probability value between 0 to 1 for the binary classification problem, the sigmoid function is selected, as shown from (6). At last, learning processes exist in both the orbitofrontal cortex and the amygdala space by (8) and (10). They are related to the sensory inputs, reward signal, and the outputs. Besides, learning process is also done in the sensory cortex, and the adaptive law of mean and variance of Gaussian function are generated by (12) and (13).

2.3 Convergence Analyses

In the previous discussion, the learning laws in (12) and (13) call for a proper choice of the learning rates. Large values of learning rates could speed up the convergence; however, it may lead to more unstable. Therefore, we introduce the convergence theorem for selecting proper learning rates for FBELC to guarantee the stable convergence of the system.

Theorem 1: Let λ_m and λ_σ be the learning rates for the parameter of FBELC m_{ij} and σ_{ij}, respectively. Then, the stable convergence is guaranteed if λ_m and λ_σ are chosen as

$$0 < \lambda_m < \frac{2}{y_m^2} \tag{14}$$

and

$$0 < \lambda_\sigma < \frac{2}{y_\sigma^2} \tag{15}$$

Proof: A Lyapunov function is selected as

$$L(k) = \frac{1}{2}e^2(k) \tag{16}$$

The change of the Lyapunov function is

$$\Delta L(k) = L(k+1) - L(k) = \frac{1}{2}[e^2(k+1) - e^2(k)] \tag{17}$$

The predicted error can be represented by

$$e(k+1) = e(k) + \Delta e(k) \cong e(k) + [\frac{\partial e(k)}{\partial m_{ij}}]\Delta m_{ij} \tag{18}$$

where Δm_{ij} denotes the change of m_{ij}.

Using (12), it is obtained that

$$\frac{\partial e(k)}{\partial m_{ij}} = -y_o(1-y_o)\cdot(A-O)\cdot S_{ij}\cdot\frac{2(I_i-m_{ij})}{\sigma_{ij}^2} \tag{19}$$

Substituting (12) and (19) into (18), yields

$$e(k+1) = e(k) - \lambda_m e(k)[y_o(1-y_o)\cdot(A-O)\cdot S_{ij}\cdot\frac{2(I_i-m_{ij})}{\sigma_{ij}^2}]^2 = e(k) - \lambda_m e(k) y_m^2 \tag{20}$$

Thus,

$$\Delta L(k) = \frac{1}{2}[e^2(k+1) - e^2(k)] = \frac{1}{2}e^2(k)[(1-\lambda_m y_m^2)^2 - 1] \tag{21}$$

If λ_m is chosen as (14), $\Delta L(k)$ in (21) is less than 0. Therefore, the Lyapunov stability of $L(k) > 0$ and $\Delta L(k) < 0$ is guaranteed. The proof for λ_σ can be derived in similar method, which should be chosen as (15). This completes the proof.

3 Simulation Results

The proposed FBELC model is evaluated for two illustrative examples, including classification of cancer using gene microarray data and prediction of heart disease.

These two research data sets used for the medical classification in this analysis respectively are called the Leukemia ALL/AML dataset and Statlog Heart Dataset.

The design method of the classification system consists of the following steps:

- Step 1. Obtain the information of the dataset, including the the attributes of samples and their labels. Partial attributes or entire attributes are selected as the features of samples, which are also the inputs of the model.
- Step 2. Divide the data into training set and testing set. A training set could be selected from the dataset in a certain proportion, and the rest as the test set. Sometimes, k-fold cross-validation or other cross validation methods are also used for the partition.
- Step 3. Set the initial conditions and inputs. In general, the initial values of parameters of the learning algorithm, such as m_{ij}, σ_{ij}, V_{ij}, W_{ij}, V_{th} are chosen as random. Stop criterion is set when the mean square error equals to a certain defined small value, or the iteration value reaches an upper set limit value. Besides, the number of blocks and the values of learning-rates could be determined by trial-and-error.
- Step 4. The selected features are put as the input to the model, and the output are obtained by formula (2)-(7). Then, the differences between the actual output and the given label of the training samples are adopted to modify the parameters in the network, by formula (8)-(13).
- Step 5. After training for a certain number of times, the test set are applied to the trained model and the performance are available.

3.1 Leukemia ALL/AML Dataset

a) Data description

In this experiment, we target the prediction of leukemia disease using the standard Leukemia ALL/AML data [38], which is available at website http://www.molbio.princeton.edu/colondata. This dataset contains 72 samples taken from 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Each sample has its class label, 1 and 0, which is either ALL or AML. That means, this prediction problem could be modelled as a binary classification problem. Also, 7129 gene expression values corresponding to each patient are provided.

b) Experiment Methods

Previously, various types of gene selection methods are applied for classification on the Leukemia Datasets. These studies mostly consider how to automatically select appropriate genes, which could be associated with medical knowledge, and then obtain good results. Indeed, it is confirmed that no more than 10% of these 7129 genes are relevant to leukemia classification [38]. And without gene

selection, it is unnecessary or even harmful to classify such a few samples in such a high dimensional space. That means, it is impossible to use all the 7129 genes as features to classify this problem. Different from previous research, this work focuses on the comparison of the performance with other approaches using the same genes, and also the analysis of stability of the performance using different genes, simply because we mainly intend to illustrate the classification performance of the FBELC model. Our experiments are carried out using different selected genes as the inputs of the classifier, and the results are compared with other models.

For the classification system, the number of inputs is exactly the number of genes which are selected. If five genes are selected as features, that means, the input consists of five dimensions. The number of blocks could be adjusted from 10 to 50, determined by trial-and-error. It mainly influences the training time in the learning process. The training data are used to train the proposed FBELC offline. Other parameters, including m_{ij}, σ_{ij}, V_{ij}, W_{ij}, V_{th} are randomly initialized and the weighting factors k_1 and k_2 are tuned to be 100 and 1, respectively. Stop criterion is set to the limit of 200 training epochs with the learning rate $\lambda_V = \lambda_W = \lambda_m = \lambda_\sigma = 0.0001$.

c) Result Analysis

In our experiment, two evaluation methods are considered to divide the available data into a training dataset and a test dataset. Firstly, as provided by the standard Leukemia ALL/AML data, the training dataset consists of 38 samples (27 ALL and 11 AML) from bone marrow specimens, while the testing dataset has 34 samples (20 ALL and 14 AML), which are prepared under different biological experimental conditions.

Table 1 shows the comparison of performance of different methods on Leukemia dataset by training on 38 samples and testing on 34 samples. We adopt the accuracy to measure the performance of all the approaches. Shujaat Khan [39] proposes an RBF network with a novel kernel and selects the top five genes [40] for the experiment and provides 97.07% of accuracy. Meanwhile, using the same five genes as the features of our model, the same result is obtained. It is evident that the 97.07% of accuracy means there are only one of the 34 samples has been misclassified. Considering the small sample size for testing, this result is acceptable. The other two approaches, those of Tang [41], and Krishna Kanth [42], respectively select 15 genes and 2 genes as inputs of each classifier by achieving both 100% accuracy, that means the 34 testing samples are all classified correctly. It is shown in Table 1 that, using the FBELC model, whether choosing 15 genes or 2 genes, the accuracy is both 97.07%. This result illustrates that the benefit of the proposed model seems to be its stability and high-accuracy performance.

Moreover, another evaluation method, Leave-One-Out Cross Validation (LOOCV) is also used to verify the proposed classifier. The LOOCV method is usually used to select a model with good generalization and to evaluate predictive performance.

At each LOOCV step, this method holds out one sample for testing while the remining samples are used for training. The overall test accuracy is calculated based on each testing samples. The LOOCV method has the advantages of producing model estimates with less bias and more ease in smaller samples.

Table 1
Comparison of performance of different methods on Leukemia dataset by training on 38 samples and testing on 34 samples

Method	**No.of genes/features**	**Gene accession number/Gene index**	**Accuracy(%)**	
			References	Our work
RBF with a novel kernel [39]	5	Top-5-ranked	97.06	97.06
FCM-SVM-RFE [41]	15	M83652, X85116, D49950, U50136, M24400, Y12670, L20321, M23197, M20902, X95735, M19507, L08246, M96326, X05409, M29610	100	97.06
MFHSNN [42]	2	X95735, M27891	100	97.06

Hence, for this Leukemia ALL/AML dataset, LOOCV method consists of splitting the dataset randomly into 72 samples. At each of the 72-th iteration, 71 samples will be used as training sample and the left-out sample will be used as the test sample. For each step we obtain a test accuracy and the final accuracy equals to the average value of 72 iterations. If we get the 98.61% accuracy, it means in each test of LOOCV, there is only one error of 72 times test.

Table 2
Comparison of performance of different methods on Leukemia dataset by LOOCV

Method	**No.of genes/features**	**Gene accession number/Gene index**	**Accuracy(%)**	
			References	Our work
Wrapper method +SVM [43]	5	Top-5-ranked	98.61	98.61
NB [43]	4	Top-4-ranked	94.44	98.61
Wrapper method + NB [43]	3	Top-3-ranked	98.61	98.61
SVM [44]	3	X95735, M31523, M23197	97.22	98.61

Table 2 compares the performance of different methods on Leukemia dataset by LOOCV. The first three approaches, those of Peng [43], report 98.61% of accuracy with the top-5-ranked genes by wrapper method and SVM, 94.44% of accuracy with the top-4-ranked genes by NB and RBF and 98.61% of accuracy

with the top-3-ranked genes by NB method. The last approach is that of Wang [44], involves the model of MFHSNN and get the 97.22% of accuracy. However, according to the same selected genes in previous work, respectively, from the results in Table 2, the proposed FBELC model, gives 98.61% accuracy consistently. In other word, we can confirm in spite of that the attained performance being similar to other methods, the proposed FBELC is superior to other approaches for its good generalization.

3.2 Statlog Heart Dataset

a) Sample Datasets

The Statlog heart disease dataset used in our work is published and shared in the UCI machine learning database [45]. It contains 270 observations, which belongs two classes: the presence and absence of heart disease. Every sample includes 13 different features, including some conditions and symptoms of the patients. Thus, some of the attributes are real value and some are binary or nominal type.

b) Performance Evaluation

This Statlog dataset is commonly used among researchers for classification. Some studies used all the 13 features as inputs of the classifiers. Others proposed some of features are irrelevant to the learning process and feature selection was used to improve learning accuracy and decrease training and testing time. To illustrate the effectiveness of the proposed algorithm, both of these scenarios are considered and the results are compared with other studies.

Table 3
Comparison of performance of different methods on Statlog dataset with all features

Author(Year)	Method	Training-test partitions	Accuracy(%)
Subbulakshmi [46] (2012)	ELM	70%-30%	87.5
Lee [47] (2015)	NEWFM	5-fold CV	81.12
Hu [48] (2013)	RSRC	5-fold CV	84.0
	Our work	70%-30%	89.41
	Our work	5-fold CV	85.93
	Our work	75%-25%	92.54

Table 3 compares the accuracies of our algorithm with other approaches using all 13 features. The first method, that of Subbulakshmi [46], involves an extreme learning machine (ELM) for two category data classification problems and evaluated on the Statlog datasets. And the accuracy of 87.5% is the mean value for 50 runs. Those of Lee [47] and Hu [48], use 5-fold cross validation (CV) method to make results more credible and the accuracies of 81.12% and 84.0% are obtained, respectively. The results in our study, separately using the same training-test partitions as other approaches, shows better performance evidently.

Besides, using 75%-25% training-test partitions, the highest classification performance is also provided in Table 3.

On the other hand, there are some other feature selection and classification approaches based on this Statlog dataset. They use not all the 13 features as the inputs of models. The comparison of the classification accuracies of our study and previous methods, not using all features, are summarized in Table 4. The first approach [49], based on the LSTSVM model with 11 features, has achieved the accuracies of 85.19%, 87.65% and 83.93% using 50-50%, 70-30% and 80-20% partitions, respectively. The other method, that of Lee [47], selects 10 features and obtains the accuracy of 82.22%, using 5-fold CV. Using the same selected features and training-test partitions, the results shows our proposed method obtains superior and consistent performance. Indeed, the last two approaches select fewer features and obtain higher accuracies than other researches. For example, Liu [50] uses four classifiers with the same 7 features, and achieves the accuracy of 83.33%, 85.19%, 87.03% and 92.59%, respectively. Ertugrul [51] selects only 3 features, which are all nominal type and obtains the accuracy of 85.93% by extreme learning machine. All these feature selection methods are combined to certain algorithms. The accuracy displayed in Table 4 remains relatively high value, which demonstrates good robustness of the proposed model.

The results of our proposed algorithm are based on the fixed structure of the FBELC model and the parameters setting for the simulations using different selected features are presented in Table 5. It exhibits that the weighting factors, number of blocks and learning rate does not change much when using different selected features.

From the simulation results of two examples, with different feature selection or different training-testing partitions, the fuzzy emotional learning classifier can always perform well. From theoretical analysis, it is proved that satisfactory performance could be obtained by choosing appropriate emotional signals, according to the characteristics of the classification problems.

Table 4

Comparison of performance of different methods on Statlog dataset with certain features

Author(Year)	**Method**	**No. of genes/features**	**Training-test partitions**	**Accuracy(%)**	
Tomar [49] (2014)	LSTSVM	11	50%-50%	85.19	87.41
			70%-30%	87.65	88.89
			80%-20%	83.93	90.74
Lee [47] (2015)	NEWFM	10	5-fold CV	82.22	85.19
Liu [50] (2017)	RFRS	7	70%-30%	92.59	85.19
Ertugrul [51] (2016)	ELM	3	9-fold CV	85.93	83.33

Table 5
List of classification parameters using different features

Parameters	Value			
	11 features	10 features	7 features	3 features
k_1	50	50	30	30
k_2	1	1	1	1
No. of blocks	50	50	30	30
Learning rate	0.00001	0.00001	0.00001	0.00001

Conclusions

This study has successfully proposed a FBELC for classification. The novelty of this paper lies in the proposed approach: the incorporation of the fuzzy inference system and a BEL model, and the application to medical diagnosis. The classification efficiency can be improved specifically because of the fuzzy set and the novel setting of reward signal in the model, which can cause better generalization and faster learning. Meanwhile, two medical datasets are applied to test the developed FBELC model. From the simulation results, it is shown that the proposed algorithm can perform high generalization and good accuracy, while being simple and easily implementable. Therefore, the results indicate that the proposed classifier can be used as a promising alternative tool in medical decision and diagnosis. The data used for simulations all come from public medical experimental datasets; in the future, we could cooperate with some medical institutions and apply our algorithm in practical experiments.

References

[1] T. Islam, P. K. Srivastava, M. A. Rico-Ramirez, Q. Dai, D. Han and M. Gupta, "An exploratory investigation of an adaptive neuro fuzzy inference system (ANFIS) for estimating hydrometeors from TRMM/TMI in synergy with TRMM/PR," Atmospheric Research, Vol. 145-146, pp. 57-68, 2014

[2] K. Subramanian, S. Suresh and N. Sundararajan, "A metacognitive neuro-fuzzy inference system (McFIS) for sequential classification problems," IEEE Transactions on Fuzzy Systems, Vol. 21, No. 6, pp. 1080-1095, 2013

[3] M. Shanbedi, A. Amiri, S. Rashidi, SZ. Heris and M. Baniadam, "Thermal performance prediction of two-phase closed thermosyphon using adaptive neuro-fuzzy inference system," Heat Transfer Engineering, Vol. 36, No. 3, pp. 315-324, 2015

[4] S. Babu Devasenapati and K. I. Ramachandran, "Hybrid fuzzy model based expert system for misfire detection in automobile engines," International Journal of Artificial Intelligence, Vol. 7, No. A11, pp. 47-62, 2011

[5] T. Haidegger, L. Kovács, R. E. Precup, B. Benyo, Z. Benyo, S. Preitl, "Simulation and control for telerobots in space medicine," Acta Astronautica, Vol. 81, No. 1, pp. 390-402, 2012

[6] M. Jocic, E. Pap, A. Szakál, D. Obradovic, Z. Konjovic, "Managing big data by directed graph node similarity," Acta Polytechnica Hungarica, Vol, 14, No. 2, pp. 183-200, 2017

[7] S. Vrkalovic, E. C. Lunca and I. D. Borlea, "Model-free sliding mode and fuzzy controllers for reverse osmosis desalination plants," International Journal of Artificial Intelligence, Vol. 16, No. 2, pp. 208-222, 2018

[8] F. Jiménez, G. Sánchez and JM. Juárez, "Multi-objective evolutionary algorithms for fuzzy classification in survival prediction," Artificial Intelligence in Medicine, Vol. 60, No. 3, pp. 197-219, 2014

[9] A. Gotmare, SS. Bhattacharjee, R. Patidar and NV. George, "Swarm and evolutionary computing algorithms for system identification and filter design: A comprehensive review," Swarm and Evolutionary Computation, Vol. 32, pp. 68-84, 2017

[10] J. S. Guan, G. L. Ji, L. Y. Lin, C. M. Lin, T. L. Le and I. J. Rudas, "Breast tumor computer-aided diagnosis using self-validation cerebellar model neural networks," Acta Polytechnica Hungarica, Vol. 13, No. 4, pp. 39-52, 2016

[11] C. M. Lin and Y. F. Peng, "Adaptive CMAC-based supervisory control for uncertain nonlinear systems," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 34, No. 2, pp. 1248-1260, 2004

[12] C. M. Lin, and H. Y. Li, "Adaptive dynamic sliding-mode fuzzy CMAC for voice coil motor using asymmetric Gaussian membership function," IEEE Transactions on Industrial Electronics, Vol. 61, No. 10, pp. 5662-5671, 2014

[13] C. H. Lee, F. Y. Chang and C. M. Lin, "An efficient interval type-2 fuzzy CMAC for chaos time-series prediction and synchronization," IEEE Transactions on Cybernetics, Vol. 44, No. 3, pp. 329-341, 2014

[14] C. M. Lin and C. H. Chen, "Robust fault-tolerant control for a biped robot using a recurrent cerebellar model articulation controller," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 37, No. 1, pp. 110-123, 2007

[15] C. M. Lin and H. Y. Li, "Intelligent control using the wavelet fuzzy CMAC backstepping control system for two-axis linear piezoelectric ceramic motor drive systems," IEEE Transactions on Fuzzy Systems, Vol. 22, No. 4, pp. 791-802, 2014

[16] J. E. LeDoux, The Amygdala: Neurobiological Aspects of Emotion, Wiley-Liss, New York, 1992, pp. 339-351

[17] J. E. LeDoux, "Emotion: clues from the brain," Annual Review of Psychology, Vol. 46, pp. 209-235, 1995

[18] J. Morén and C. Balkenius, "A computational model of emotional learning in the amygdala," From Animals to Animats 6: Proceedings of the 6th International Conference on the Simulation of Adaptive Behavior, Cambridge, Mass. The MIT Press. 2000

[19] C. Balkenius and J. Moren, "Emotional learning: A computational model of the amygdala," Cybernetics and Systems, Vol. 32, No. 6, pp. 611-636, 2001

[20] J. Moren, Emotion and Learning-A Computational Model of the Amygdala, PhD Dissertation, Lund University, 2002

[21] C. Lucas, D. Shahmirzadi and N. Sheikholeslami, "Introducing BELBIC: Brain emotional learning based intelligent controller," International Journal of Intelligent Automation and Soft Computing, Vol. 10, No. 1, pp. 11-21, 2004

[22] M. Roshanaei, E. Vahedi and C. Lucas, "Adaptive antenna applications by brain emotional learning based on intelligent controller," IET Microwaves, Antennas & Propagation, Vol. 4, No. 12, pp. 2247-2255, 2010

[23] M. Hosseinzadeh Soreshjani, G. Arab Markadeh, E. Daryabeigi, N. R. Abjadi and A. Kargar, "Application of brain emotional learning-based intelligent controller to power flow control with thyristor-controlled series capacitance," IET Generation, Transmission & Distribution, Vol. 9, No. 14, pp. 1964-1976, 2015

[24] S. A. Aghaee, C. Lucas, and K. Amiri Zadeh, "Applying brain emotional learning based intelligent controller (BELBIC) to multiple-area power systems," Asian Journal of Control, Vol. 14, No. 6, pp. 1580-1588, 2012

[25] S. A. N. Niaki, R. Iravani and M. Noroozian, "Power-flow model and steady-state analysis of the hybrid flow controller," IEEE Transactions on Power Delivery, Vol. 23, No. 4, pp. 2330-2338, 2008

[26] G. Huang, Z. Zhen and D. Wang, "Brain emotional learning based intelligent controller for nonlinear system," 2008 Second International Symposium on Intelligent Information Technology Application, Shanghai, 2008, pp. 660-663

[27] M. M. Polycarpou, "Fault accommodation of a class of multivariable nonlinear dynamical systems using a learning approach," IEEE Transactions Automatic Control, Vol. 46, pp. 736-742, 2001

[28] H. A. Zarchi, E. Daryabeigi, G. R. A. Markadeh and J. Soltani, "Emotional controller (BELBIC) based DTC for encoderless synchronous reluctance motor drives," 2011 2nd Power Electronics, Drive Systems and Technologies Conference, Tehran, 2011, pp. 478-483

[29] M. A. Rahman, R. M. Milasi, C. Lucas, B. N. Araabi and T. S. Radwan, "Implementation of emotional controller for interior permanent-magnet synchronous motor drive," IEEE Transactions on Industry Applications, Vol. 44, No. 5, pp. 1466-1476, 2008

[30] S. K. Agarwal and R. Kumar, "Explication of a logistic regression driven hypothesis to strengthen derivative approach driven classification for medical diagnosis," 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, Bhopal, 2014, pp. 1-6

[31] H. Azzawi, J. Hou, Y. Xiang and R. Alanni, "Lung cancer prediction from microarray data by gene expression programming," IET Systems Biology, Vol. 10, No. 5, pp. 168-178, 2016

[32] M. Tan, B. Zheng, J. K. Leader and D. Gur, "Association between changes in mammographic image features and risk for near-term breast cancer development," IEEE Transactions on Medical Imaging, Vol. 35, No. 7, pp. 1719-1728, 2016

[33] J, Zhao, L. Y Lin, and C. M. Lin, "A general fuzzy cerebellar model neural network multidimensional classifier using intuitionistic fuzzy sets for medical identification," Computational Intelligence and Neuroscience, 2016

[34] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," IEEE Transactions on Medical Imaging, Vol. 35, No. 5, pp. 1207-1216, 2016

[35] AR Mehrabian, C. Lucas, "Emotional learning based intelligent robust adaptive controller for stable uncertain nonlinear systems," World Academy of Science, Engineering and Technology Vol. 19, pp. 1027-1033, 2008

[36] E. Bijami, R. Abshari, S. M. Saghaiannejad and J. Askari, "Load frequency control of interconnected power system using brain emotional learning based intelligent controller," 2011 19th Iranian Conference on Electrical Engineering, Tehran, 2011, pp. 1-6

[37] C. F. Hsu, C. T. Su and T. T. Lee, "Chaos synchronization using brain-emotional-learning-based fuzzy control," IEEE Joint, International Conference on Soft Computing and Intelligent Systems, 2016, pp. 811-816

[38] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, Vol. 286, No. 5439, pp. 531-537, 1999

[39] S. Khan, I. Naseem, R. Togneri and M. Bennamoun, "A novel adaptive kernel for the RBF neural networks," Circuits Systems and Signal Processing, Vol. 36, No. 4, pp. 1639-1653, 2017

[40] H. C. Peng, F. H. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine

Intelligence, Vol. 27, No. 8, pp. 1226-1238, 2005

[41] Y Tang, Y. Q. Zhang and Z Huang, "FCM-SVM-RFE gene feature selection algorithm for leukemia classification from microarray gene expression data," The 14th IEEE International Conference on Fuzzy Systems, Reno, 2005, pp. 97-101

[42] B. B. M. Krishna Kanth, U. V. Kulkarni and B. G. V. Giridhar, "Gene expression based acute leukemia cancer classification: a neuro-fuzzy approach," International Journal of Biometric & Bioinformatics, Vol. 4, No. 4, pp. 136-146, 2010

[43] Y Peng, W. Li, and Y. Liu, "A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification," Cancer Information, Vol. 2, No. 1, pp. 301-311, 2006

[44] S Wang, H Chen, R Li and D. Zhang, "Gene selection with rough sets for the molecular diagnosing of tumor based on support vector machines," International Computer Symposium, Taiwan, 2006, pp. 1368-1373

[45] UCI Repository of Machine Learning Databases, http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

[46] C. V. Subbulakshmi, S. N. Deepa and N. Malathi, "Extreme learning machine for two category data classification," 2012 IEEE International Conference on Advanced Communication Control and Computing Technologies, Ramanathapuram, 2012, pp. 458-461

[47] S. H. Lee, "Feature selection based on the center of gravity of BSWFMs using NEWFM," Engineering Applications of Artificial Intelligence, Vol. 45, pp. 482-487, 2015

[48] Y. C. Hu, "Rough sets for pattern classification using pairwise-comparison-based tables," Applied Mathematical Modelling, Vol. 37, No. 12-13, pp. 7330-7337, 2013

[49] D. Tomar and S. Agarwal, "Feature selection based least square twin support vector machine for diagnosis of heart disease," International Journal of Bio-Science and Bio-Technology, Vol. 6, No. 2, pp. 69-82, 2014

[50] X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the rfrs method,"Computational and Mathematical Methods in Medicine, Vol. 2017, No. 3, pp. 1-10, 2017

[51] F. Ertugrul, "Determining the order of risk factors in diagnosing heart disease by extreme learning machine," International Conference on Natural Science and Engineering, Kilis, 2016, pp. 10-19

An Integrated MCDM Approach to PLM Software Selection

Sanja Puzovic, Jasmina Vesic Vasovic, Miroslav Radojicic, Vladan Paunovic

University of Kragujevac, Faculty of Technical Sciences Cacak, Svetog Save 65, 32000 Cacak, Serbia

E-mail: sanja.puzovic@ftn.kg.ac.rs; jasmina.vesic@ftn.kg.ac.rs; miroslav.radojicic@ftn.kg.ac.rs; vladan.paunovic@ftn.kg.ac.rs

Abstract: This paper presents the development of a hybrid multi-criteria decision-making (MCDM) approach for Product Lifecycle Management (PLM) software selection, as an essential part of the PLM concept implementation. The approach is based on the hybrid MCDM process that integrates the Fuzzy Analytic Hierarchy Process (FAHP) and the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE). The Fuzzy AHP has been applied in order to overcome the problem of the vagueness of decision-makers' judgments in the process of the criteria relative significance assessment, whereas, the PROMETHEE method has been applied in order to evaluate the pieces of software. This paper's findings should indicate the broad possibilities of the proposed model for an objective evaluation of PLM software, on the basis of their total suitability against the global goal according to the established criteria, and capability for efficiently overcoming the problem of data vagueness that decision-makers are facing during the process.

Keywords: PLM software; MCDM; software selection; Fuzzy AHP; PROMETHEE

1 Introduction

The rise of global processes leads to strengthening competition, strengthening of consumer's awareness and increasingly, more complex and stricter regulations that manufacturers have to adhere to. There is a constant increase of product complexity while the lifecycle shortens, also the market fluctuations and economic uncertainty create pressure on prices and expressed need for a quick response to the consumer's growing and interchangeable requirements. All these are the challenges brought by contemporary global flows, which on their part radically change market conditions and competitive relationships, to overcome those challenges requires adoption of the new knowledge and modifying problem

reaction and solution approach [1]. To face these challenges, manufacturers must adopt and incorporate the concept of products management process into their business flows throughout their lifecycle, from the idea of "end-of-life", i.e. PLM concept. The PLM refers to the strategic approach to creation, management and use of intellectual capital and products related information throughout their lifecycle, from the initial concept to their withdrawal. Addressing in depth the benefits of the PLM concept implementation is a major issue of many research studies [2, 3, 4, 5, 6]. According to Lämmer and Theiss [2] the PLM is considered as the basic concept for satisfying a series of business requirements with respect to the completeness, visibility and high product data transparency, financial requirements related to costs reduction, revenue growth, product related requirements with respect to innovations and its faster market placement, higher quality and a series of regulatory requirements. The PLM can be observed as an integrated information related approach consisting of people, processes and technologies [3], with purpose to create, store and seek data, information and knowledge about the products throughout their lifecycle [2].

The PLM strategy implementation, as a support to higher corporate goals, requires an appropriate software support, which will create a platform for spotting business possibilities, the PLM process standardization, increase in visibility of a product lifecycle phases and cost reduction. It also supports research-development efforts and product introduction to the market. However, this still remains an open issue, given the fact that there is no PLM software that will fully satisfy complex and specific users' requirements. Since the PLM software selection is a complex and challenging problem, therefore the solution lies in MCDM models for the multi-aspect assessment of the considered software, and their ranking based upon the overall suitability according to the global goal, also the concepts which will successfully deal with the problem of data vagueness accompanying this process.

2 Literature Review

Researching the problem of the software selection, is the field of interest for many authors, from different areas, resulting in various approaches, which are mainly hybrid, based on a combination of several MCDM methods, which makes it easier to handle the complexity of the problem.

Zaidan et al. [7] shows comparative analysis of the results obtained during the software selection for the Open Source of Electronic Medical Records (OS-EMR) by applying the AHP method integrated with different MCDM techniques, such as Weighted Product Method (WPM), Weighted Sum Model (WSM), Simple Additive Weighting (SAW), Hierarchical Additive Weighting (HAW) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). The combination of the AHP and TOPSIS methods is a model frequently used for

software selection, where the AHP is used in the criteria weight assessment, and TOPSIS for determining the rank of the alternatives. This approach has been used in the selection of the Extract, Transform and Load (ETL) software in the paper [8], whereas in the paper [9], the criteria assessment for the selection of collaborative software was performed by applying the fuzzified AHP method, while the TOPSIS method has been used to evaluate the alternatives. For software selection for the needs of an electronic company, Efe [10] proposes a hybrid multi criteria group decision-making approach, based on the integration of the Fuzzy AHP and Fuzzy TOPSIS. Yazgan et al. [11] attempted to address this issue using the Analytical Network Process (ANP) and the Artificial Neural Network (ANN), the approach implying that, in the creation of an ANP model, the opinions of a group of experts are reduced to the unique values with the help of the geometrical mean technique, after which obtained ANP results are used in creation of an ANN model in order to determine the priority of the considered ERP pieces of software.

Gürbüz et al. [12] propose a framework for ERP software evaluation, which integrates three methods: ANP, Choquet Integral (CI) and Measuring Attractiveness by a Categorical Based Evaluation Technique (MACBETH), where ANP defines the rank of alternatives, whereas the CI and MACBETH do the research in conjunctive or disjunctive criteria behavior. Lee et al. [13] propose AHP method application in solving the problem of the Open Source Customer Relationship Management (OS-CRM) software evaluation. Shukla et al. [14], developed a model for software selection, formed by combining the Stepwise Weight Assessment Ratio Analysis (SWARA) which is used for criteria prioritization and the PROMETHEE method, used for ranking alternatives.

Rouhani and Rouhani [15] have introduced a new approach in the selection of the Information Technology Service Management (ITSM) software, that implies software evaluation on the basis of functional and non-functional criteria by applying the Fuzzy Superiority and Inferiority Ranking (FSIR) method.

3 A Methodological Framework

The decision support model for the multi-criteria PLM software selection proposed in the paper represents a hybrid MCDM based approach, formed by integration of the PROMETHEE and the Fuzzy AHP methods. The problem structuring and criteria prioritization processes are realized by applying the Fuzzy AHP method, the results of this phase are further integrated within the process of evaluating alternatives by the PROMETHEE method.

Throughout the utilization of the PROMETHEE method, certain ambiguities related to designing the problem structure and the assessment of the criteria relative importance can be observed, limiting the rationality of decisions made; therefore, with the aim of overcoming these weaknesses, it is necessary to adopt

certain concepts from the other MCDM methods, and implement them within the PROMETHEE procedure. The first extension of the PROMETHEE method in this paper was performed by adoption of the problem structuring concept according to the AHP methodology, which allows problem solving by decomposing it down into hierarchy of decision-making elements of a different levels, which facilitates understanding of the importance of each element. On the other hand, the PROMETHEE method doesn't offer this structural possibility, which makes gaining an insight into the complexity of the problem more difficult and limits rationality of the final solution.

The PROMETHEE method does not provide clear guidelines to the assessment of criteria importance, either, but it is rather left to the decision-makers' rough estimate, and, as such, is imprecise and unreliable. The criteria used in the process of alternatives evaluation are expressed in different measure units with different requirements of minimization or maximization, they are changeable in time, heterogeneous, and frequently contradictory, as well [1, 16], whereas their relative significance is a variable category and depends on a concrete situation, as well as on the decision-maker's subjective perceptions. Besides, the significance attributed to criteria determines to a great extent the final selection, so, pursuant to the importance of this problem issue, it is necessary within the MCDM process, that much more complex approaches to the determination of the criteria relative weight than that offered by the PROMETHEE method should be included.

The criteria relative weights might be assigned in a normalized manner; however, a more suitable manner for their expressing is by using preference relation and linguistic expressions, given the fact that it is closer to the manner of human thinking. So far, several methods have been developed and used for criteria prioritization. In a study [17], the Linear Programming Technique for Multidimensional Analysis of Preference (LINMAP) is exposed, within which the relative weights of criteria are determined by means of the linear programming model. Hafezalkotob and Hafezalkotob [18] present the Fuzzy entropy-weighted MULTIMOORA method, within which the relative weights of criteria are determined on the basis of the entropy concept under the fuzzy environment, whereas in the paper [19] the geometric least square technique for the determination of the values of weight intervals from interval fuzzy preference relations is presented.

One of the most frequently applied principles for criteria prioritization – the Eigenvector Method (EM) – was introduced by Saaty [20], within the AHP method, within this principle the preference relations are described by a multiplied preferences relation, i.e. the pairwise matrix, which further translates into the problem of determining eigenvalues for the purpose of obtaining normalized weighted vectors. Wang and Chin [21] propose a modification of the EM by introducing a new approach, called the Linear Programming Approximation to the EM (LPAEM), which provides quite a consistent matrix of criteria comparison. There are several different prioritization techniques, presented in the literature,

that are used within the AHP procedure: Least Squares (LS) method [22], the Goal Programming (GP) method [23], the Fuzzy Preference Programming (FPP) method [24], the Linear Programming (LP) method [25], the Weighted Least Square method and Quadratic Programing method [26], the Logarithmic Least Squares (LLS) method [27] and the Geometric Mean (GM) method [28]. Also, herein, special attention is drawn to the approaches, based on describing a preference relationship, by means of the linguistic expressions, mathematically described by fuzzy numbers, which helps overcome the problem of the imprecision of the DM's assessments.

3.1 The Concept of the Fuzzy AHP Method

The AHP is an MCDM method developed by Saaty [20], it is frequently applied method in solving complex decision-making problems where it is necessary to include a series of different attributes that are difficult to formalize. Apart from its simplicity, the conventional AHP method shows certain weaknesses, which limit its application in situations when there is any indefiniteness whatsoever in the data about the problem that is being considered, and also, the AHP method is criticized that it doesn`t fully reflect a human way of thinking. Most often, authors point at the limitations that are related to the problem of the Eigenvector method, the unbalanced evaluation scale, and pairwise comparison. As the eigenvector solution is based on the description of the problem and an arbitrary order of the factors, this method shows shortcomings with respect to the adjustment of ratio measurements, so it is incapable of retaining the capacity of isomorphism between ratio and difference estimation problems [29].

The conventional AHP method offers an insufficiently precise ranking based on the unbalanced estimation scale because of the neglecting of the uncertainty that may appear when copying the decision-maker's imprecise perceptions onto the numerical estimation scale [30]. The final result in the AHP method is determined by the subjective estimation of the decision-maker while simultaneously the majority of them rely on subjective perceptions, their knowledge or prior experience. Decision-makers, however, are frequently unable to precisely express their preferences due to incomplete information, the complexity and indefiniteness of the decision-making problem, or yet a lack of an appropriate comparison scale. The conventional AHP method could be appropriate solution if the decision-maker's preferences might be express by means of static crisp values, which is not the case with unreliable and imprecise preferences; expressing them by linguistic descriptions is closer to human way of thinking; however, the uncertainty that can occur when copying imprecise perceptions onto the numerical estimation scale is neglected. This limitation is possible to overcome by using the fuzzy numbers that adequately represent fuzzy linguistic variables, that are broadly applied due to their capability to successful establish a compromise between the descriptive power and computational simplicity [31]. By using fuzzy numbers, it is possible to

quantitatively describe linguistic variables in an appropriate manner, by which the problem of their sharp classification on Saaty's Scale is successfully overcome.

The fuzzy AHP method was first introduced by Van Laarhoven and Pedrycz [32], who developed the fuzzy logarithmic least squares method. Within this process a system of normal equations for a fuzzy case with several degrees of freedom is obtained by minimizing logarithmic squares method. So far, several Fuzzy AHP models that mainly differ from one another in the manner of the fuzzification of the evaluation scale, (where the fuzzification of the linguistic expressions on Saaty's Scale by means of triangular and trapezoidal fuzzy numbers is the most present), have been developed: the geometrical mean method [33], the synthetic extend analysis method [34], the fuzzy method of the least square [35], the lambda-max method [36], the fuzzy preference programming method [37], the two-stage logarithmic programming method [38], the modification of logarithmic least squares method [39].

The criteria prioritization process within this paper was carried out by using the extent analysis method developed by Chang [34]. This Fuzzy AHP model is based on the fuzzification of imprecise preferences while performing a pairwise comparison by means of triangular fuzzy numbers, which is followed by implementation of the extent analysis method [34] into the process in order to determine relative weights by means of the synthetic extent value. Decision-makers express their subjective preferences when comparing criteria by means of linguistic expressions, each linguistic expression is assigned a numerical value that is, due to the indefiniteness of expressed preferences, it is given in the form of a triangular fuzzy number adopted from the fuzzified Saaty Scale (Table 1).

Table 1
Fuzzified Saaty Scale

Fuzzy number	Linguistic term	Scale of fuzzy numbers
'1	Equally important	(1, 1, 1)
'3	Weakly important	(2, 3, 4)
'5	Essentially important	(4, 5, 6)
'7	Very strongly important	(6, 7, 8)
'9	Absolutely important	(7, 8, 9)
'2,'4,'6,'8	Intermediate values ('*x*)	(*x*-1, *x*, *x*+1)
1/'x	Between two adjacent judgments	(1/*x* + 1, 1/*x*, 1/*x* – 1)

The fuzzified comparison matrix $M = \{M_{gi}^{j}\}$ with n goals and m attributes is given in (1). where $M_{gi}^{j} = (l_{ij,} m_{ij,} u_{ij})$ is the triangular fuzzy number that stands for a normalized and convex fuzzy set characterized by a closed confidence interval $[l_{ij}, u_{ij}]$ and the degree of uncertainty α (Figure 1).

$$M = \begin{vmatrix} M_{g1}^{1} & M_{g1}^{2} & \dots & M_{g1}^{m} \\ M_{g2}^{1} & M_{g2}^{2} & \cdots & M_{g2}^{m} \\ \cdots & \cdots & \dots & \cdots \\ M_{gn}^{1} & M_{gn}^{2} & \cdots & M_{gn}^{m} \end{vmatrix} \quad (1)$$

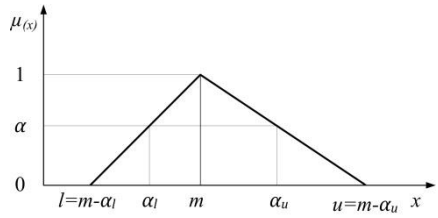


Figure 1
A triangular fuzzy number

Triangular fuzzy numbers are otherwise referred to as "linear" due to their linear membership function, defined as (2).

$$\mu_M(x) = \begin{cases} \frac{x-l}{m-l}, & x \in [l,m], \\ \frac{x-u}{m-u}, & x \in [m,u], \\ 0, & otherwise \end{cases} \tag{2}$$

The value of a linguistic expression, therefore, belongs to a closed interval $[l_{ij}, u_{ij}]$, where the borders of the interval represent the imprecision of the given expression, whereas m_{ij} represents the value of the linguistic expression in which the membership function has the highest value $m_{ij} = 1$.

In the case that the decision-making process includes n experts, the n fuzzy comparison matrices are obtained that are possible to aggregate by means of the fuzzy geometrical mean method [33], by which, the aggregated triangular fuzzy number of the assessment of the form $M_{gi}^{j} = (l_{ij,} m_{ij,} u_{ij})$ with the triangular membership function whose members are defined by (3), is obtained.

$$M_{gi}^{j} = \left(\prod_{k=1}^{n} M_{gik}^{j}\right)^{\frac{1}{n}} \tag{3}$$

where M_{gik}^{j} is the fuzzy relative importance according to the k^{th} expert's opinion, and n is the total number of the experts. Based on the obtained aggregated fuzzy comparison matrix M_{gi}^{j} it is possible to compute the fuzzy synthetic extent value S_i by respecting the algebraic rules for a triangular fuzzy number, by the (4).

$$S_i = \sum_{j=1}^{m} M_{gi}^{j} \otimes \left[\sum_{i=1}^{n}\sum_{j=1}^{m} M_{gi}^{j}\right]^{-1} \tag{4}$$

The S_i could be obtained from the previous relation by the (5).

$$S_i = (l_{i'}, m_{i'}, u_{i'}) \otimes \left(\frac{1}{\sum_{i=1}^{n} u_{i'}}, \frac{1}{\sum_{i=1}^{n} m_{i'}}, \frac{1}{\sum_{i=1}^{n} l_{i'}}\right) = \left(\frac{l_{i'}}{\sum_{i=1}^{n} u_{i'}}, \frac{m_{i'}}{\sum_{i=1}^{n} m_{i'}}, \frac{u_{i'}}{\sum_{i=1}^{n} l_{i'}}\right) = (l_i, m_i, u_i) \tag{5}$$

The further procedure requires the determination of the degree of the possibility that: $S_2 = (l_2, m_{2,} u_2) \geq S_1 = (l_1, m_{1,} u_1)$, according to the (6) which is based on the previously obtained fuzzy synthetic extent value:

$$V(S_2 \geq S_1) = sup\left[min(\mu_{S_1}(x), (\mu_{S_2}(y))\right] \tag{6}$$

This possibility can be expressed through the (7).

$$d = \begin{cases} 1, & m_2 \geq m_1 \\ 0, & l_1 \geq u_2 \\ \frac{l_1 - u_2}{(m_2 - u_2) - (m_1 - l_1)} & otherwise \end{cases} \tag{7}$$

where d represents the value of the ordinate on the abscissa that corresponds with the highest point of intersection between S_2 and S_1 (8), shown in Figure 2.

$$V(S_2 \geq S_1) = hgt(S_1 \cap S_2) = \mu_{S_1}(d) \tag{8}$$

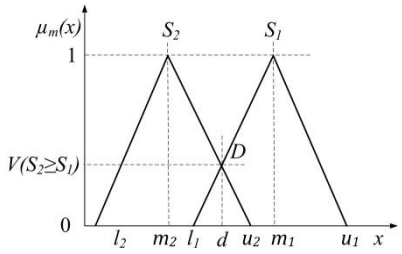


Figure 2

The intersection between S_1 and $S_2$and their degree of possibility [34]

By finding the preference of S_i and S_k , $(k = 1,2, \dots, n,$ where n is the number of the criteria) the degree of a possibility of obtaining a convex fuzzy number can be calculated by (9).

$$V\left(S \geq S_{1,}\, S_2, \dots, S_n\right) = [V(S \geq S_1)\ and\ V(S \geq S_2)\ and\ \dots\ V(S \geq S_n)] = min\, V(S \geq S_i)\ (i = 1,2, \dots, n) \tag{9}$$

Should $d'(A_i) = \min V(S_i \geq S_k)$, $(k = 1,2,3, \dots, n;\ k \neq i)$, then the obtained weight vector has the form as (10):

$$w' = \left(d'(A_1), d'(A_2), \dots, d'(A_n)\right) \tag{10}$$

By the normalization of the obtained weight vectors the weight of each individual criterion (11) is obtained in the form of a non-fuzzy number.

$$w = \left(d(A_1), d(A_2), \dots, d(A_n)\right) \tag{11}$$

3.2 The Concept of the PROMETHEE Method

The PROMETHEE method [40] belongs to the MCDM group of methods, and serves to rank the considered alternatives assessed in a multi-criteria system. This method enables the aggregation of the qualitative and quantitative criteria of different importance. So far, several approaches to MCDM-problem solving by

applying the PROMETHEE method have been developed: PROMETHEE I – for partial ranking; PROMETHEE II, i.e. the "net flow" method, which enables a complete order of all alternatives [40]; the PROMETHEE GAIA descriptive approach to the analysis of the results obtained [41]; PROMETHEE III, for the ranking based on intervals; PROMETHE IV for the multi-criteria analysis of an uninterrupted set of alternatives; PROMETHEE V, for optimization with segmentation limitations [42]; PROMETHEE VI, supportive of the manner which humans think in [43]; the PPROMETHEE CLUSTER, developed for normal classification [44], then Fuzzy PROMETHEE, based on the fuzzy outranking relation to overcome the problem of the indefiniteness, uncertainty and imprecision of data in decision-making [45], and the Modified PROMETHEE, based on the universal preference function [16].

The PROMETHEE procedure consists of the two steps: Construction of an outranking relation by means of the preference index and exploitation of the obtained relations for the purpose of solving the problem.

Let to define the MCDM problem as (12).

$$Max(f_1(a), f_2(a), \dots, f_n(a)) | a \in A \tag{12}$$

where A represents the final set of the alternatives subjected to the ranking, according to the defined criteria for the evaluation - f. For each alternative a from within set A, $f_i(a)$ represents the related value as per criterion f_i. The results of the comparison of the alternatives a and b ($a, b \in A$) can be expressed in the form of the preference function (13).

$$P(a,b) = P\big(f(a) - f(b)\big) = P(x) \tag{13}$$

The preference function $P(a,b)$ has the characteristics: $0 \le P(a,b) \le 1$, $P(a,b) \ne P(b,a)$, expresses the intensity of the preference of the alternative a in comparison with the alternative b and can be interpreted as: $P(a,b) = 0$ (indifference – $f(a) = f(b)$); $P(a,b) \sim 0$ (weak preference - $f(a) > f(b)$); $P(a,b) \sim 1$ (strong preference - $f(a) >> f(b)$); $P(a,b) = 1$ (strict preference - $f(a) >>> f(b)$). The preference function $P(a,b)$ is the non-falling function that acquires the zero value for the negative value $f(a) - f(b)$, and can graphically be interpreted as in Figure 3.

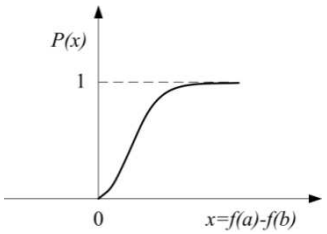


Figure 3
The general preference function

For each of the criteria, the type of the preference function is determined pursuant to the specificity of the criterion, as well as the related parameters. Brans and Vincke [40] propose six types of the general preference functions that enable the expression of a preference in the majority of real problems.

What follows is the calculation of the preference index according to the (14) that represents the total intensity of the preference of the alternative a over the alternative b, according to all of the acknowledged criteria.

$$IP(a,b) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} P_i(a,b) w_i \tag{14}$$

where w_i is the relative weight of the i^{th} criterion with the characteristics: $w_i \in [0,1]$ and $\sum w_i = 1$.

On the basis of the preference index, it is possible to determine the value of the leaving (15) and the entering (16) flows, and on the basis of these values a partial comparison of alternatives is performed (PROMETHEE I).

$$\emptyset^{+}(a_i) = \sum_{x \in A} IP(a,x) \tag{15}$$

$$\emptyset^{-}(a_i) = \sum_{x \in A} IP(x,a) \tag{16}$$

A complete order of the alternatives (PROMETHEE II) requires the balancing of the entering and the leaving flows, i.e. the consideration of the net flow (17).

$$\emptyset(a_i) = \emptyset^{+}(a_i) - \emptyset^{-}(a_i) \tag{17}$$

4 Problem Definition and Modeling

The decision-making problem considered in this paper relates to the selection of the adequate PLM software, it has been expressed a request for software that provides the necessary modules and functions for service provision and the integration of the basic PLM processes and the creation of the centralized data source for all the participants in the value chain, apart from which such software should satisfy certain requirements with respect to the technical characteristics. The software should also be learnable and efficient, with understandable models and concepts, adaptive to the specificities of production and business doing within different industries, as well as of acceptable costs.

The proposed model for PLM software selection is based on a hybrid multi-criteria approach. The basis of this process consists of the designing a MCDM base, which implies the generation of potential alternatives and the development of the system of criteria for their evaluation. Then, criteria prioritization, a multi-criteria evaluation of the alternatives, their ranking, the selection of the optimal solution and the results analysis by sensitivity analysis which facilitates the final

selection for the decision makers. The proposed model enables linking of all data and relations into a rational whole, which enables the analysis and understanding the problem with all of its logical connections, complexity, specificities and possible uncertainties, and then, the realization of a rational decision.

5 The Application of the Proposed Model for PLM Software Selection

The decision team formed of five experts from the industries such as: high-tech electronics, retail, medical devices, industrial manufacturing and automotive, has been involved through the entire process of PLM software selection. Those experts are with work experiences of 5-12 years and academic, engineering and IT background, also they are with experience in implementing the PML concept.

Step 1: The Construction of the Problem Hierarchical Structure - The expert team generated 12 alternatives – PLM pieces of software with required characteristics (Figure 4). The measuring and understanding of the total suitability of those alternatives against the global goal requires the recognition of different evaluation aspects, i.e. the translation of the general decision-making goals into the criteria on the basis of which alternatives will be evaluate, thus forming the criteria base that serves as the framework for the assessment of the alternatives.

Table 2
The criteria used for software MCDM selection

Criterion	Paper	Criterion	Paper
Usability	[7, 8, 13, 46, 47, 48]	Installation factors	[7, 50]
Functionality	[7, 8, 13, 48, 49, 50, 51, 52]	Perenniality	[9]
Technical specifications	[10, 53]	Implementability	[49]
User support and service	[7, 50, 52]	Learnability	[46]
Cost	[8, 10, 13, 47, 49, 50, 51, 52, 53]	Technology advance	[49, 51, 52]
Vendors factors	[8, 10, 13, 47, 49, 51, 52]	Weight of experts	[9]
Security	[7, 9, 47]	Machine-human interface	[9]
Flexibility	[13, 49, 51]	Technical architecture	[13]
Reliability	[8, 13, 47, 48, 50, 51]	Portability	[47, 48]
Developer support	[7]	Strategy-fit	[49]
Understandability	[46]	Operability	[46]
Ease of use	[50, 51, 53]	System overhead	[53]
Customizability	[7, 13, 50, 52, 53]	Attractiveness	[46]
Efficiency	[47, 48]	Maintainability	[48, 53]

In Table 2, the list of the criteria for the selection of software of different purpose, frequently used in the literature, is presented. As we perceive, there is no universality in criteria selection, but they depend on the type and purpose of software. Although the most frequently used are the criteria that consider functionality, usability and efficiency as the key aspects of software quality, the

majority of authors resort to expanding the base of the criteria for evaluation, including also the other aspects, such as technical specifications, flexibility, user support and service, ease of use, vendor factors… On the basis of this research, and the experts' and the PLM-software users' experience and assessments, a system of the seven basic criteria for the alternatives evaluation was identified, each of those criteria is explained in several sub-factors, which leads to a more rational evaluation (Figure 4). The selected criteria represent the main dimensions PLM software quality, and by their integration into the selection process all the aspects of the considered problem significant for finding the optimal solution are included.

As the result of the problem structuring carried out in compliance with the AHP methodology, the hierarchical structure of the problem was created (Figure 4).

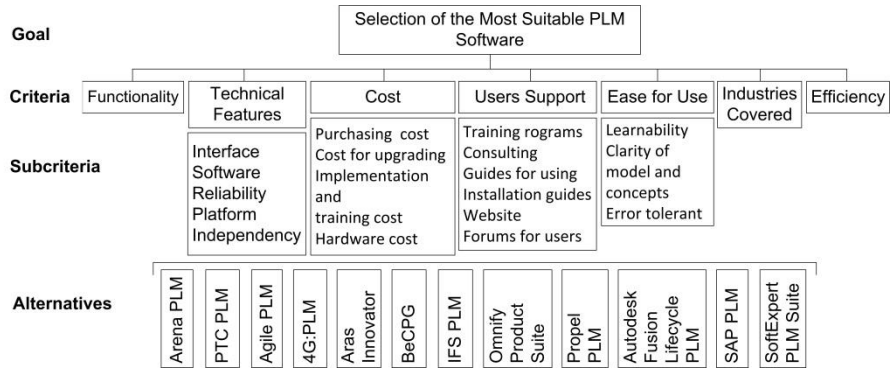


Figure 4

The hierarchical structure of the PLM software selection problem

In the context of PLM software, criterion Functionality considers whether the software provides the necessary functions, as well as the modules for providing integration of the basic PLM processes and offering a centralized data source for all participants in the value chain, such as: Bill of Material Analysis, Cost Tracking, Document Management, Product Data Management, Supplier Management, Product Analysis and other. Industries covered criterion relates to the level of the flexibility of the software, i.e. the extent to which the software is capable of responding to the specific production challenges faced by manufacturers from a wide range of industries.

Step 2: The Assignation of the Criteria Relative Weights - According to the subjective preferences of the group of five experts, involved in the decision making process, about the importance of the established criteria, the individual fuzzy pairwise comparison matrices were formed by using the linguistic variables that were subsequently translated into the appropriate fuzzy numbers according to the fuzzified scale (Table 1). By calculating the fuzzy geometrical mean of the experts' individual matrices according to (3), the aggregated fuzzy pairwise

comparisons matrix (Table 3) is formed. The consistency of both all individual matrices and the aggregated matrix is less than 0.1, which is indicative of the consistency of the criteria evaluation and ensures a required level of the quality of the decision, so the procedure may continue.

Table 3

The fuzzy aggregated pairwise comparison matrix

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	(1,1,1)	(1.74,2.77,3.78)	(0.74,1.25,2.05)	(0.52,0.76,1.25)	(2.61,3.65,4.68)	(2.22,3.29,4.32)	(1.64,2.7,3.73)
C_2	(0.26,0.36,0.57)	(1,1,1)	(0.61,1.06,1.68)	(0.37,0.5,0.74)	(1.4,1.97,2.55)	(0.96,1.55,2.17)	(0.94,1.4,2.06)
C_3	(0.49,0.8,1.35)	(0.59,0.94,1.64)	(1,1,1)	(0.38,0.58,0.82)	(1.06,1.78,2.7)	(0.92,1.64,2.55)	(0.87,1.52,2.22)
C_4	(0.8,1.32,1.93)	(1.35,2,2.72)	(1.22,1.72,2.61)	(1,1,1)	(1.64,2.7,3.73)	(1.32,2.35,3.37)	(1.64,2.7,3.73)
C_5	(0.21,0.27,0.38)	(0.39,0.51,0.72)	(0.37,0.56,0.94)	(0.27,0.37,0.61)	(1,1,1)	(0.47,0.66,1.11)	(0.43,0.61,1)
C_6	(0.23,0.3,0.45)	(0.46,0.64,1,05)	(0.39,0.61,1.08)	(0.3,0.43,0.76)	(0.9,1.52,2.14)	(1,1,1)	(1,1.74,2.41)
C_7	(0.27,0.37,0.61)	(0.48,0.72,1.06)	(0.45,0.66,1.15)	(0.27,0.37,0.61)	(1,1.64,2.35)	(0.42,0.57,1)	(1,1,1)

On the basis of the aggregated fuzzy pairwise comparisons matrix the fuzzy synthetic extent values for each of the criteria (S_i) are computed according to (5), the value S_i for the remaining criteria is accounted for in Table 4.

Table 4

The fuzzy synthetic extent values, the possibilities matrix and the weight vectors

Criterion		S_i	The possibilities matrix							$w_{i'}$	w_i
			C_1	C_2	C_3	C_4	C_5	C_6	C_7		
C_1	Functionality	(0.12, 0.25, 0.5)	-	1	1	1	1	1	1	1	0.265
C_2	Technical Features	(0.06, 0.13, 0.26)	0.523	-	0.966	0.612	1	1	1	0.5234	0.138
C_3	Efficiency	(0.06, 0.14, 0.3)	0.596	1	-	0.677	1	1	1	0.5956	0.158
C_4	Cost	(0.11, 0.23, 0.46)	0.927	1	1	-	1	1	1	0.9268	0.245
C_5	Users Support	(0.04, 0.07, 0.14)	0.078	0.537	0.519	0.171	-	0.704	0.807	0.0776	0.021
C_6	Industries Covered	(0.05, 0.1, 0.21)	0.377	0.849	0.819	0.466	1	-	1	0.3765	0.1
C_7	Ease for Use	(0.05, 0.09, 0.19)	0.28	0.747	0.721	0.370	1	0.902	-	0.2797	0.074

The calculated fuzzy synthetic extent values represent the preference of a certain criterion over other criteria, whereas the possibility of the determined superiority of each individual criterion over other criteria, individually, $V(S_i \geq S_k)$ $(k = 1,2,3,\dots,n; k \neq i)$, can be determined by means of (7). The obtained values are shown in the possibilities matrix, shown in Table 4.

By minimizing the calculated possibilities, criterion weight vector $w_{i'}$ is obtained (according to (10)), by whose normalization the final weights of each individual criterion are obtained in the form of the non-fuzzy number w (11). The normalization is possible to perform by means of (18). The criteria weights vectors, as well as their normalized values, are presented in Table 4.

$$w = \frac{w_i'}{\sum_{i=1}^{n} w_i'} \tag{18}$$

The obtained results are indicative of the fact that the Functionality criterion has priority over the other criteria, with the weight coefficient of 0.265, so it will have

the greatest influence on the final decision. The Cost criterion with the weight of 0.245 will also have a significant influence on the choice. According to the assessments made by the experts involved in the decision-making process, the criteria with the least influence on the software selection are: Industries Covered, Ease for Use and Users Support.

Step 3: The Evaluation of the Alternatives - The ranking of the considered PLM pieces of software was performed by applying the PROMETHEE method. The basis of this procedure consists of the designing of the MCDM decision-making base (Table 5). Besides, for each criterion a type of the preference function that reflects the specificities of the given criterion in the most appropriate way was selected, after which the related parameters, as well as the requirements for extremization were determined (Table 5). The results obtained in the previous procedure of the criteria evaluation by means of the Fuzzy AHP method were used for the relative weights of the criteria.

Table 5

MCDM base

Criterion	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Generalized Criteria Type	III	VI	III	V	IV	V	VI
p	0.2	-	2.6	1.17	1	1.2	-
q	-		-	7.33	2.5	3	-
σ	-	0.49	-	-	-	-	0,76
Request	Max	Max	Max	Min	Max	Max	Max
Criteria Relative Weight	0.265	0.138	0.158	0.245	0.021	0.1	0.074

For the considered MDCM problem the evaluation matrix (Table 6) was constructed, which compliantly with the defined hierarchical structure of the problem (Figure 4) encompasses the 12 alternatives assessed in the system of 7 criteria. The evaluation matrix was constructed on the bases of the impressions that the experts, involved in the decision making process, have acquired during testing of the considered PLM pieces of software, also on the bases of the data about the software performance provided by the producers.

Table 6

The evaluation matrix

Criteria	Arena PLM	PTC PLM	Agile PLM	4G:PLM	Aras Innovator	BeCPG	IFS PLM	Omnify Product Suite	Propel PLM	Autodesk Fusion Lifecycle	SAP PLM	SoftExpert PLM Suite
Functionality	0.579	0.526	0.632	0.316	0.526	0.579	0.421	0.632	0.474	0.368	0.684	0.526
Technical Features	3.3	3.025	4.4	4.2	3.95	3.2	3.975	3.925	4.275	3.925	4.75	4.2
Efficiency	7.02	6.2	8.62	7.6	4.98	5.12	5.24	8.24	7.82	8.21	8.96	7.12
Cost	2.97	6	10	1.67	2.23	1.5	5.67	6.67	4	5	10	2.8
Users Support	8.23	7.27	10	6.72	6.19	8.77	9.62	7.84	8.96	6.95	9.33	7.56
Industries Covered	2.258	1.935	2.258	0.968	4.839	1.290	3.226	1.29	1.935	4.194	4.516	3.871
Ease for Use	8.18	7.3	7.77	8.3	6.7	8.1	6.45	8.68	7.77	7.89	7.17	9.2

It should be mentioned that, in the evaluation of the alternatives according to the Cost criterion, the costs of the using, implementation, upgrading and training for the software packages that only include some of the available modules which the user had expressed the need for are taken into consideration. Simultaneously, all of the considered software packages are of a roughly equivalent content when the modules included in them are concerned.

For each pair of the compared alternatives, the preference function $P(a,b)$ was determined, according to (13), then the index of the preference $IP(a,b)$ according to (14) that enables the construction of the outranking relation.

Step 4: The ranking of the alternatives - The alternatives rank is determined on the basis of the values of the entering, leaving and net flows (Table 7) calculated according to (15-17).

Table 7

The values of the positive, negative and net flows

Alternatives		$\emptyset^+(a_i)$	$\emptyset^-(a_i)$	$\emptyset(a_i)$	Rang
a_1	Arena PLM	0.228	0.179	0.049	5
a_2	PTC PLM	0.102	0.353	-0.251	11
a_3	Agile PLM	0.310	0.213	0.097	4
a_4	4G:PLM	0.214	0.309	-0.095	10
a_5	Aras Innovator	0.231	0.254	-0.023	7
a_6	BeCPG	0.221	0.277	-0.056	8
a_7	IFS PLM	0.095	0.379	-0.284	12
a_8	Omnify Product Suite	0.295	0.143	0.152	3
a_9	Propel PLM	0.210	0.174	0.036	6
a_{10}	Autodesk Fusion Lifecycle	0.195	0.266	-0.071	9
a_{11}	SAP PLM	0.449	0.203	0.246	1
a_{12}	SoftExpert PLM Suite	0.303	0.106	0.198	2

The obtained alternative rank (Table 7) based on the preference index indicates that to the greatest extent the SAP PLM software satisfies the established requirements, this alternative has the greatest value of the net flow of 0.246. Somewhat worse ranked alternatives are the pieces of software SoftExpert PLM Suite and Omnify Product Suite with the values of the net flow being 0.198 and 0.152, respectively. The obtained rank of the considered PLM pieces of software, provides a significant help in making a decision of the PLM software selection.

Step 5: The results analysis - Figure 5 presents the Geometrical Analysis for Interactive Aid (GAIA) plane for the PLM software selection problem. The GAIA plane represents a descriptive approach to the results analysis, understanding the specificities of a problem, the identification of a synergy or conflicts between criteria; it enables the highlighting of the alternatives clusters and the alternatives with exceptional preferences. The quality of the visual display is 80.9%, which indicates that very few information got lost by the results projection. The GAIA plane is indicative of the fact that, between the criteria of Functionality, Users support i Industries covered, there is a synergy; the other group of close criteria consists of Efficiency and Technical features, whereas the criteria Cost and Ease

for use are conflicting against the other criteria. The pieces of software SAP PLM, AGILE PLM and Omnify Product Suite demonstrate the best features with respect to the functionality and user support, but are of weaker characteristics with respect to the costs and the ease of use in comparison with the other pieces of software.

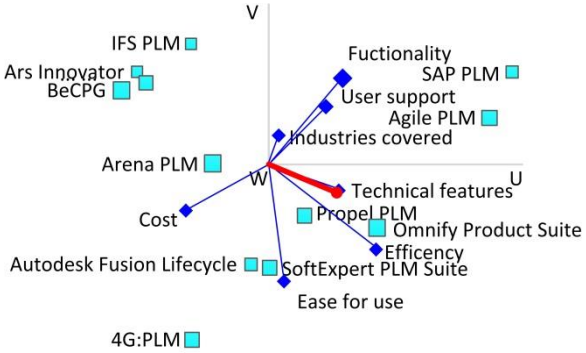


Figure 5
The GAIA plane for PLM software selection

Conclusion

The theoretical and practical applications of this work are indicative of the conveniences of the proposed model for PLM software selection. The model is formed by combining two techniques: Fuzzy AHP and PROMETHEE, which are integrated within one MCDM approach, by their inter-complementarity, the weaknesses that they show, which can represent a significant limitation in finding rational solutions, have been overcome. The Fuzzy AHP method enables for obtaining more consistent and more precise criteria weights, in comparison with those determined on the basis of the DM's intuition, which is the case, with the PROMETHEE method. Also, the incorporation of fuzzy set theory into the criteria prioritization process, makes it easier to handle the ambiguities that this process embodies. On the other hand, the PROMETHEE method enriches the proposed approach, by assigning appropriate preference functions to each one of the criteria, in which manner, it reflects their specificity and enables the rational ranking of the considered PLM pieces of software.

The proposed model reduces subjective outcomes and generates much more rational solutions, based on the reliable assessment of criteria weights, problem structuring and overcoming the problems and inconsistencies of human thinking. It offers a multi-aspect evaluation of the considered alternatives and the provision of adequate support to group decision-making and finally, visual analysis of the results obtained. Apart from the PLM software selection, this model is also applicable in solving other real problems that may include various conflicting criteria, in the fuzzy environment.

References

[1] J. Vesić Vasović, M. Radojičić, M. M. Klarin, V. K. Spasojević Brkić: Multi-Criteria Approach to Optimization of Enterprise Production Programme, Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, Vol. 225, No. 10, 2011, pp. 1951-1963

[2] L. Lämmer, M. Theiss: Product Lifecycle Management, In Concurrent Engineering in the 21st Century, Springer International, 2015, pp. 455-490

[3] H. Gmelin, S. Seuring: Achieving Sustainable New Product Development by Integrating Product Life-Cycle Management Capabilities, International Journal of Production Economics, Vol. 154, 2014, pp. 166-177

[4] P. S. G. D. Oliveira, D. D. Silva, L. F. D Silva, M. D. S. Lopes, A. Helleno: Factors that Influence Product Life Cycle Management to Develop Greener Products in the Mechanical Industry, International Journal of Production Research, Vol. 54, No.15, 2016, pp. 4547-4567

[5] L. Horváth, I. J. Rudas: Active Knowledge for the Situation-Driven Control of Product Definition, Acta Polytechnica Hungarica, Vol. 10, No. 2, 2013, pp. 217-234

[6] M. Cantamessa, F. Montagna, P. Neirotti: An Empirical Analysis of the PLM Implementation Effects in the Aerospace Industry, Computers in Industry, Vol. 63, No. 3, 2012, pp. 243-251

[7] A. A. Zaidan, B. B. Zaidan, M. Hussain, A. Haiqi, M. M. Kiah, M. Abdulnabi: Multi-Criteria Analysis for OS-EMR Software Selection Problem: A Comparative Study, Decision Support System, Vol. 78, 2015, pp. 15-27

[8] M. Hanine, O. Boutkhoum, A. Tikniouine, T. Agouti: Application of an Integrated Multi-Criteria Decision Making AHP-TOPSIS Methodology for ETL Software Selection, SpringerPlus, Vol. 5, No. 1, 2016, p. 263

[9] S. S. Kara, N. Cheikhrouhou: A Multi Criteria Group Decision Making Approach for Collaborative Software Selection Problem, Journal of Intelligent and Fuzzy Systems, Vol. 26, No. 1, 2014, pp. 37-47

[10] B. Efe: An Integrated Fuzzy Multi Criteria Group Decision Making Approach for ERP System Selection, Applied Soft Computing, Vol. 38, 2016, pp. 106-117

[11] H. R. Yazgan, S. Boran, K. Goztepe: An ERP Software Selection Process with Using Artificial Neural Network Based on Analytic Network Process Approach, Expert Systems with Applications, Vol. 36, No. 5, 2009, pp. 9214-9222

[12] T. Gürbüz, S. E. Alptekin, G. I. Alptekin: A Hybrid MCDM Methodology for ERP Selection Problem with Interacting Criteria, Decision Support Systems, Vol. 54, No. 1, 2012, pp. 206-214

[13] Y. C. Lee, N. H Tang, V. Sugumaran: Open Source CRM Software Selection Using the Analytic Hierarchy Process, Information Systems Management, Vol. 31, No. 1, 2014, pp. 2-20

[14] S. Shukla, P. K. Mishra, R. Jain, H. C. Yadav: An Integrated Decision Making Approach for ERP System Selection Using SWARA and PROMETHEE Method, International Journal of Intelligent Enterprise, Vol. 3, No. 2, 2016, pp. 120-147

[15] S. Rouhani, S. Rouhani: A Fuzzy Superiority and Inferiority Ranking Based Approach for IT Service Management Software Selection, Kybernetes, Vol. 46, No. 4, 2017, pp. 728-746

[16] M. Radojicic, M. Zizovic, Z. Nesic, J. Vesic Vasovic: Modified Approach to PROMETHEE for Multi-Criteria Decision-Making, Maejo International Journal of Science and Technology, Vol. 7, No. 3, 2013, pp. 408-421

[17] Y. Z. Liu, Z. P. Fan, G. X. Gao: An Extended LINMAP Method for MAGDM Under Linguistic Hesitant Fuzzy Environment, Journal of Intelligent and Fuzzy Systems, Vol. 30, No. 5, 2016, pp. 2689-2703

[18] A. Hafezalkotob, A. Hafezalkotob: Fuzzy Entropy-Weighted MULTIMOORA Method for Materials Selection, Journal of Intelligent and Fuzzy Systems, Vol. 31, No. 3, 2016, pp. 1211-1226

[19] X. Yang, Z. J. Wang: Geometric Least Square Models for Deriving-Valued Interval Weights from Interval Fuzzy Preference Relations Based on Multiplicative Transitivity, Mathematical Problems in Engineering, 2015

[20] T. L. Saaty: The Analytic Hierarchy Process, New York, NY: McGraw-Hill, 1980

[21] Y. M. Wang, K. S. Chin: A Linear Programming Approximation to the Eigenvector Method in the Analytic Hierarchy Process, Information Sciences, Vol. 181, No. 23, 2011, pp. 5240-5248

[22] T. L. Saaty, L. G. Vargas: Comparison of Eigenvalue, Logarithmic Least Squares and Least Squares Methods in Estimating Ratios, Mathematical Modelling, Vol. 5, No. 5, 1984, pp. 309-324

[23] N. Bryson: A Goal Programming Method for Generating Priority Vectors, Journal of the Operational Research Society, Vol. 46, No. 5, 1995, pp. 641-648

[24] L. Mikhailov: A Fuzzy Programming Method for Deriving Priorities in the Analytic Hierarchy Process, Journal of the Operational Research Society, Vol. 51, 2000, pp. 341-349

[25] B. Chandran, B. Golden, E. Wasil: Linear Programming Models for Estimating Weights in the Analytic Hierarchy Process, Computers and Operations Research, Vol. 32, 2005, pp. 2235-2254

[26] X. U. Ye-jun, D. A. Qing-li: Weighted Least-Square Method and Its Improvement for Priority of Incomplete Complementary Judgement Matrix, Systems Engineering and Electronics, Vol. 7, 2008, p. 021

[27] Y. Xu, R. Patnayakuni, H. Wang: Logarithmic Least Squares Method to Priority for Group Decision Making with Incomplete Fuzzy Preference Relations, Applied Mathematical Modelling, Vol. 37, No. 4, 2013, pp. 2139-2152

[28] B. T. Sivrikaya, A. Kaya, M. Dursun, F. Çebi: Fuzzy AHP–Goal Programming Approach for a Supplier Selection Problem, Research in Logistics and Production, Vol. 5, No. 3, 2015, pp. 271-285

[29] J. Barzilai: Deriving Weights from Pairwise Comparison Matrices, Journal of the Operational Research Society, Vol. 48, No. 12, 1997, pp. 1226-1232

[30] C. C. Sun: A Performance Evaluation Model by Integrating Fuzzy AHP and Fuzzy TOPSIS Methods, Expert Systems with Applications, Vol. 37, No. 12, 2010, pp. 7745-7754

[31] S. Nestic: M. Stefanovic, A. Djordjevic, S. Arsovski, D. Tadic: A Model of the Assessment and Optimisation of Production Process Quality Using the Fuzzy Sets and Genetic Algorithm Approach, European Journal of Industrial Engineering, Vol. 9, No. 1, 2015, pp. 77-99

[32] P. J. M. Van Laarhoven, W. Pedrycz: A Fuzzy Extension of Saaty's Priority Theory, Fuzzy Sets and Systems, Vol. 11, No. 1-3, 1983, pp. 229-241

[33] J. J. Buckley: Fuzzy Hierarchical Analysis, Fuzzy Sets and Systems, Vol. 17, No. 3, 1985, pp. 233-247

[34] D. Y. Chang: Applications of the Extent Analysis Method on Fuzzy AHP, European Journal of Operational Research, Vol. 95, No. 3, 1996, pp.649-655

[35] R. Xu: Fuzzy Least-Squares Priority Method in the Analytic Hierarchy Process, Fuzzy Sets and Systems, Vol. 112, No. 3, 2000, pp. 395-404

[36] R. Csutora, J. J. Buckley: Fuzzy Hierarchical Analysis: the Lambda-Max Method, Fuzzy Sets and Systems, Vol. 120, No. 2, 2001, pp. 181-195

[37] L. Mikhailov: Deriving Priorities from Fuzzy Pairwise Comparison Judgements, Fuzzy Sets and Systems, Vol. 134, No. 3, 2003, pp. 365-385

[38] Y. M. Wang, J. B. Yang, D. L. Xu: A Two-Stage Logarithmic Goal Programming Method for Generating Weights from Interval Comparison Matrices, Fuzzy Sets and Systems, Vol. 152, No. 3, 2005, pp. 475-498

[39] Y. M. Wang, T. M. Elhag, Z. Hua: A Modified Fuzzy Logarithmic Least Squares Method for Fuzzy Analytic Hierarchy Process, Fuzzy Sets and Systems, Vol. 157, No. 23, 2006, pp. 3055-3071

[40] J. P. Brans, P. Vincke: A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making), Management Science, Vol. 31, No. 6, 1985, pp. 647-656

[41] B. Mareschal, J. P. Brans: Geometrical Representations for MCDA, European Journal of Operational Research, Vol. 34, No. 1, 1988, pp. 69-77

[42] J. P. Brans, B. Mareschal: PROMETHEE V: MCDM Problems with Segmentation Constraints, INFOR: Information Systems and Operational Research, Vol. 30, No. 2, 1992, pp. 85-96

[43] J. P. Brans, and B. Mareschal, The PROMETHEE VI Procedure: How to Differentiate Hard from Soft Multicriteria Problems, Journal of Decision Systems, Vol. 4, No. 3, 1995, pp. 213-223

[44] J. Figueira, Y. De Smet, J. P. Brans: MCDA Methods for Sorting and Clustering Problems: PROMETHEE TRI and PROMETHEE CLUSTER, Université Libre de Bruxelles. Service deMathématiques de la Gestion (Working Paper 2) 2004

[45] M. Goumas, V. Lygerou: An Extension of the PROMETHEE Method for Decision Making in Fuzzy Environment: Ranking of Alternative Energy Exploitation Projects, European Journal of Operational Research, Vol. 123, No. 3, 2000, pp. 606-613

[46] A. Sanga, I. M. Venter: Algorithm for the Evaluation of Free and Open Source Software When the Evaluator is ”Uncertain”, International Journal of Management Science and Information Technology (IJMSIT) Vol. 17, 2015, pp. 36-55

[47] S. Rouhani, A. Z. Ravasan: A Fuzzy TOPSIS Based Approach for ITSM Software Selection, International Journal of IT/Business Alignment and Governance (IJITBAG), Vol. 5, No. 2, pp. 2014, 1-26

[48] G. Rincon, M. Alvarez, M. Perez, S. Hernandez: A Discrete-Event Simulation and Continuous Software Evaluation on a Systemic Quality Model: An Oil Industry Case, Information and Management, Vol. 42, No. 8, 2005, pp. 1051-1066

[49] Y. Kazancoglu, S. Burmaoglu: ERP Software Selection with MCDM: Application of TODIM Method, International Journal of Business Information Systems, Vol. 13, No. 4, 2013, pp. 435-452

[50] A. Benlian, T. Hess: Comparing the Relative Importance of Evaluation Criteria in Proprietary and Open Source Enterprise Application Software Selection–a Conjoint Study of ERP and Office Systems, Information Systems Journal, Vol. 21, No. 6, 2011, pp. 503-525

[51] Z. Ayağ, R. G. Özdemir: An Intelligent Approach to ERP Software Selection through Fuzzy ANP, International Journal of Production Research, Vol. 45, No. 10, 2007, pp. 2169-2194

[52] B. S. Sahay, A. K. Gupta: Development of Software Selection Criteria for Supply Chain Solutions, Industrial Management and Data Systems, Vol. 103, No. 2, 2003, pp. 97-110

[53] L. Kaur, D. H. Singh: Software Component Selection Techniques-A Review, International Journal of Computer Science and Information Technologies, Vol. 5, No. 3, 2014, p. 2

TopicAE: A Topic Modeling Autoencoder

Miroslav Smatana, Peter Butka

Faculty of Electrical Engineering and Informatics, Department of Cybernetics and Artificial Intelligence
TechnicalUniversity of Košice
Letná 9, 040 01 Košice, Slovakia
e-mails: {miroslav.smatana, peter.butka}@tuke.sk

Abstract: In this paper, we propose TopicAE, a simple autoencoder designed to perform topic modeling with input texts. Topic modeling has grown in popularity especially in recent years with a large number of digital documents and contributions from social media available. These texts usually contain useful information and methods in the area of topic modeling, show novel approaches to their automatic summarization, browsing and searching. The main idea of topic modeling is to uncover hidden semantic structures from the input collection of texts. There are several topic models to extract standard topics from with their evolution through time and hierarchical structure of the topics. In this paper, we propose techniques known from the area of neural networks. Our TopicAE model can be applied to solve all the tasks mentioned above. The performance of the proposed model was also tested and showed that TopicAE could solve the topic modeling problem and out performed standard methods (Latent Semantic Indexing and Latent Dirichlet Allocation) according to evaluation metrics.

Keywords: autoencoder; deep learning; neural networks; topic modeling

1 Introduction

In the last one and a half decade, the internet and especially social media have become one of the powerful communication tools providing new possibilities for data gathering and analysis. The Internet is a source of the enormous amount of data in various forms (audio, video, text, etc.). In this paper, we focus on textual data, which usually contain useful information such as opinions and attitudes related to different people, organizations, products, world events, etc. This information can be used in several ways, mostly by organizations to increase their profit, for example:

- Launching of new products – when a company introduces a new product to the market, topic modeling can be used to track topics in which the

product is discussed, see details of the opinions, problems with the product, or which products are most competitive according to users.

- Crisis analysis –in the time of a war conflict it is possible to monitor how users perceive the current situation; it is possible to track the evolution of reactions, make adequate actions.
- Targeted marketing – tracking what people usually discuss and predict which product they might buy.
- Analysis and protection of reputation – this represents an opportunity to monitor social media to catch different contributions with negative or positive opinions on a company or person.
- Media – in this case, topic modeling helps to analyze, summarize, and visualize the news, search for reactions of people on them, their evolvement and sharing.

Currently, digital textual data play a key role in many tasks. However, due to their significant amount, it is difficult to find helpful information, especially in social media sources. Such needs lead to new methods for the extraction and processing of textual data. For that reason, the so-called topic modeling became a popular and powerful tool for the automatic semantic analysis of large collections of texts. It shows new ways of browsing, searching, and summarizing such collections. The main idea of topic modeling is to uncover hidden semantic structures (topics) in texts, where a topic is represented as a probability distribution over the fixed vocabulary of words (terms).

An example of topic modeling output (for example of article from news in Associated Press) can be in this form:
"The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Julliard School. Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important...", where every color in the text represents a different topic (for example red color - arts, green color - budgets, etc.).

There are already several approaches to topic modeling. Most of them focus on the analysis in the context of static corpora. Most of the standard methods are based on the recognition of topics as collections of terms and computation of their probabilities. Latent Semantic Indexing (LSI) [1] and Latent Dirichlet Allocation (LDA) [2] are well-known and often applied standard methods, which create a probabilistic model of topics from the input corpus of documents. However, in many problems of practical significance, a simple structure in the form of an extracted set of topics for the whole corpora is not enough. There are at least two interesting sub-tasks of topic modeling, which can be extracted within topics to make them more informative. The first is that changes in time or evolution of the topics are interesting, especially in the context of data streams such as social media. In this

case, it is essential to capture changes in the topic structure. The second is that in many cases it can be seen that some topic is more structured or complicated, and it has a hierarchical structure of subtopics.

Several methods, which extract the evolution or the hierarchical structure of topics, were also already introduced, and we describe some of them in the next section within related work. Standard methods share their main feature – all of them are generative probabilistic models. In this paper, we would like to introduce our model, which provides different non-probabilistic and non-generative approach based on neural networks.

In particular, the paper describes the proposed neural network layer, TopicAE (Topic AutoEncoder) which can be applied to solve the problem of building all three types of topic modeling tasks (basic topic model, the evolution of topics in time, the hierarchical structure of sub-topics). We also provide experiments with the selected datasets, where TopicAE is compared to standard topic models using several metrics.

The remainder of the present paper is structured as follows. In the next section, we present related work for each topic modeling problem with a more detailed description of the selected methods. In Section 3, we provide some necessary details on neural networks needed for the introduction of the proposed TopicAE approach in the following section. In Section 5, we present experiments with the selected dataset as well as comparison of TopicAE, standard topic modeling methods and selected neural network models.

2 Related Work

In this section, we focus on the selected methods applied to achieve the goals of particular topic modeling tasks. First, we start our discussion with a simple latent topic model which formalizes the basic ideas in the topic modeling. Next, we describe more advanced models which can capture time evolution of topics and hierarchical topic structures. We also mention some of the neural network approaches related to topic modeling tasks.

2.1. Latent Dirichlet Allocation

Latent Semantic Indexing (LSI) [1] can be considered as one of the first methods for topic analysis. Even though if it is not always perceived as a topic modeling method, it creates a base for probabilistic latent semantic analysis [3]. The basic ideas behind this approach lead to the most known topic modeling method - Latent DirichletAllocation (LDA) – first described by Blei et al. in [2]. LDA became the de-facto standard of topic modeling and is often used as a baseline method in comparisons with new approaches.

Before we give a brief detailed description of LDA, we could mention that there are also many LDA extensions such as Petterson et al. [4] or data stream extension [5]. Another type of topic modeling method is the HierarchicalDirichlet Process (HDP) [6], which became widely used as the core of present topic modeling methods.

Now we provide a short description of the LDA method (based on [2]) for mining of the standard topic model. LDA is a generative probabilistic model of corpus data. The main idea of LDA is that input documents are represented as random mixtures over latent topics, and each of these topics is characterized by a distribution over words. The basic terms are defined as:

- A word is a basic unit of discrete data and belongs to finite vocabulary indexed by $\{1, \ldots, V\}$. The v-the word in the vocabulary is represented by a V-vector of weights w so that $w^v = 1$, $w^u = 0$ and $u \neq v$.
- A document is a sequence of N words denoted by weights $w = (w_1, w_2, \ldots, w_N)$, where w_n is a weight for n-th word in a sequence.
- A corpus is a collection of M documents denoted by $D=\{d_1, d_2, \ldots, d_M\}$.

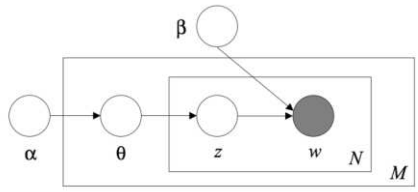


Figure 1

Probabilistic graphical representation of LDA model [2]. In this case, we have a collection of M documents represented by sequences of N words, which lead to topic models z. The whole process is controlled by parameters α and β.

LDA assumes the following generative process (graphical representation is shown in **Figure 1**) for each document in input corpus D:

1. Choose $N \sim Poisson$
2. Choose $\theta \sim Dir(\alpha)$
3. For every n-th word from N words:
    a. Choose a topic $z_n \sim Multinomial(\theta)$
    b. Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

This model has several assumptions. First, dimensionality k of the Dirichlet distribution (and thus the dimensionality of z) over $(k - 1)$ simplex is assumed known and fixed. Second, the word probabilities are parametrized by a $k \times V$ matrix β,where $\beta_{ij} = p(w^j = 1|z^i = 1)$. θ is k-dimensional Dirichlet random variable,

$\alpha$is a k-vector with components $\alpha_i > 0$. The α and β represent corpus level parameters and are sampled once in the process of processing a corpus. θ are document level variables and are sampled once per document. Variables z and w are word-levels and are sampled once for each word in each document.

2.2. Dynamic Topic Models

One can imagine that instead of simple topic modeling results, a more structured output from such approaches can be of interest for users. One of such extensions is related to time perspective in topic modeling and leads to the evolution of topics in time.

In this case, Blei and Lafferty [7] present a method called Dynamic Topic Models (DTM), which belongs to the family of probabilistic time series models and provides time evolution of topics in input collections of texts. This model works only with a discrete space model, so to resolve this problem of discretization Wang et al. present DTM extension called Continuous Time Dynamic Topic Models (cDTM) [8]. Moreover, Beykikhoshk et al. in [9] present a different approach to capturing topic time evolution based on the Hierarchical Dirichlet Process.

Differently to LDA, where documents are selected equivalently from the same set of topics, DTM approach supposes that input corpus is divided by time slice with K-component topic model and topics associated with slice t evolved from topics generated in slice $t-1$.

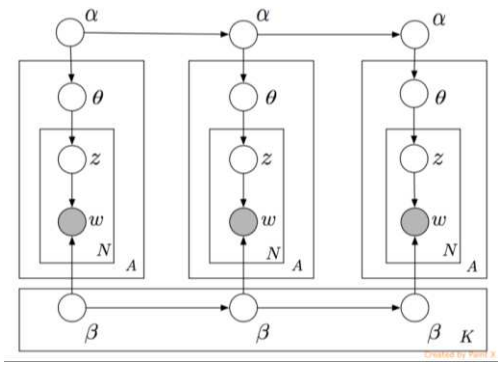


Figure 2

Graphical representation of probabilistic model known as DTM (Dynamic Topic Models) [7] used for modeling of topics evolution in time

Using the same notation as in LDA, DTM generative process for slice t is defined as follows (graphical model of DTM is given in [7]):

1. Draw topics $\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$
2. Draw $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$

3. For each document:
    a. Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
    b. For each word:
        i. Draw z ~ $Mult(\pi(\eta))$
        ii. Draw $w_{t,d,z} \sim Mult(\pi(\beta_{t,z}))$

Note that πmaps the multinomial natural parameters to the mean parameters, and $\mathcal{N}$ represents an extension of the logistic normal distribution to time-series simplex data [7].

2.3. Hierarchical Topic Models

A different family of topic modeling approaches is related to methods which can capture hierarchical topic structure. In this case, a more detailed analysis of particular topics from higher levels leads to their subtopics with the result in the form of topic hierarchy. Different methods have already been developed to achieve such a goal, e.g., Blei et al. presented a method based on a nested Chinese restaurant process in [10], Hoffman described a cluster-based method [11], Smith et al. [12] provided the hierarchical version of LDA. From other approaches we can mention Pachinko Allocation Model (PAM) [13], Hierarchical Latent Tree Analysis (HLTA) [14] or a method presented in [15], where authors developed a hierarchical model based on HDP.

Now we describe one of these approaches, a generative probabilistic model for learning the hierarchical structure of topics as an extension of LDA based on nested processes [10]. This hierarchy is an *L*-level tree, where each node is associated with a topic. In this approach, Bayesian perspective is applied to the problem of topic hierarchy extraction. Here, hierarchies are random variables and these random variables are specified procedurally. It is based on the Chinese restaurant process (CRP) [10] and is defined as follows (using notation as for LDA):

1. Let c_1 be the root restaurant
2. For each level $l \in \{2,...,L\}$:
    a. Draw a table from the restaurant c_{l-1}. Set c_l to be the restaurant referred to by table.
3. Draw an L-dimensional topic proportion vector θ from $Dir(\alpha)$
4. For each word $w \in \{1,...,N\}$:
    a. Draw $z \in \{1,...,L\}$ from $Mult(\theta)$
    b. Draw w_n from the topic associated with restaurant c_z.

Graphical representation of this hierarchical model is shown in Figure **3**. The node labeled *T* refers to a collection of an infinite number of *L*-level paths drawn from a nested CRP, γ, η are hyperparameters for T, β and distribution of *c* is defined by the nested Chinese restaurant process.

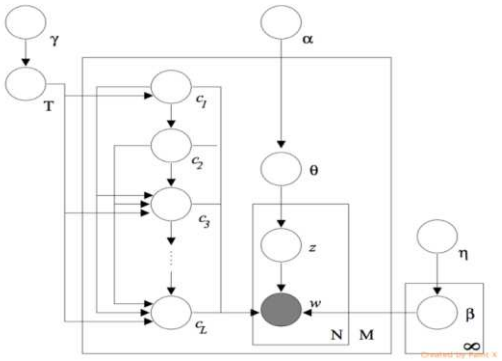


Figure 3
Graphical representation of hierarchical LDA topic model based on the nested Chinese restaurant process [10]

2.4. Neural Networks Approaches

Neural networks are techniques which are capable of automatically extracting a low dimensional representation of input data. Due to their growing popularity in recent years, these methods have become more popular in the field of natural language processing. Several works aimed to extract meaningful document representation (topics).

In their work, Salakhutdinov and Hinton [20] present a two-layer undirected graphical model called Replicated Softmax (RSM). Larochelle and Lauly [21] propose DocNADE, which is a neural autoregressive topic model that estimates the probability of observing a new word in the document by previously observed words in that document. As their results show, this approach outperforms the RSM model and solves RSM computational complexity on data with a large vocabulary. There are also other interesting models such as neural variational inference model NVDM [22], neural topic model NTM [23], k-competitive autoencoder KATE [24] or ProdLDA [26], which combine neural networks with the classic LDA model. In the field of topic evolution in time, Gupta et al. [25] present a model based on a recurrent neural network and replicated softmax to extract topical trends over time.

3 Preliminaries on Neural Networks

In this section, we shortly provide key properties of neural networks, which are necessary for the description of the proposed autoencoder TopicAE.Neural networks (NN) [16] is a computing system inspired by biological nervous systems. It is composed of a large number of highly interconnected elements called neurons, which cooperate to solve a specific problem. NN is also called a massive parallel processor model and like people can learn from examples and use gained knowledge in the future. To learn NN, we need training examples $(\boldsymbol{x}^{(i)},y^{(i)})$, where $i\in\{1,\ldots,n\}$ represents the index of i-th training example from our training set of n examples. Here, x represents a vector of input features and y represents output, which we want to predict (in this simplest case it is one value of prediction, but we can also have a vector of values in a more general case). For example, in the medical domain, every patient is a training example, where x is a vector of all measured values (symptoms), and y is the output value from $\{0,1\}$ if the patient has or has not a disease.

3.1 Single Neuron and Simple Neural Network

To describe the basics of neural networks, we will start with the description of the simplest neural network, which consists of a single neuron. The structure of neuron is shown in Figure **4**.

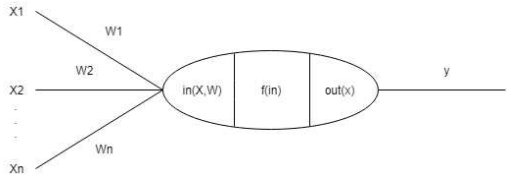


Figure 4

Structure of a single neuron in a neural network with inputs and weights represented by vectors x and w (producing aggregated input), activation function f and output y

A neuron is a computational unit, which has inputs - $\{x_1, x_2, \ldots, x_n\}$ and generates output y. A basic neuron consists of the followingparts:

- weights $\{w_1, w_2, \ldots, w_n\}$ - used to connect neurons within NN, bearers of information in NN,
- *in* - input to the neuron - is function of inputs $\{x_1, x_2, \ldots, x_n\}$. In most cases inputs are aggregated using sum function:

$$in = \sum_{j=1}^{n} w_j\, x_j \tag{1}$$

- *f(in)* - activation function - there are several types of activation functions, in this paper we use *f(in)* represented by sigmoid function:

$$f(in) = \frac{1}{1+exp(-in)} \tag{2}$$

- *out(x)* - output function, which is usually identical function - *out(x) = x*

A neural network is a structure, which connects many neurons, so that the output of one neuron is the input to another neuron. An example of a simple feed-forward NN is shown in Figure **5**. L1 represents input layer, L2 hidden layer (output values of neurons in this layer are not available in a training set), a L3 output layer and $a^{(i)}_j$ represent the output of a *j*-th neuron in the *i*-th layer.

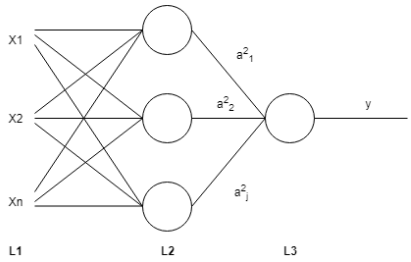


Figure 5

Structure of a single neuron in a neural network with inputs and weights represented by vectors x and w (producing aggregated input), activation function f and output y.

The number of layers gives the depth of the neural network model. Based on this, output *y* of the sample NN for inputs $\{x_1,x_{2,\ldots},x_n\}$is computed as the composition of application of functions of particular layers in feed forward way, e.g.:

$$y = f^{(3)}(f^2(f^{(1)}(x))) \tag{3}$$

3.2 Backpropagation Algorithm

Backpropagation algorithm (BP) is a gradient-based algorithm for learning neural networks using the training set $\{(x^{(1)}, y^{(1)}), \ldots, x^{(m)}, y^{(m)})\}$ of *m* examples. The goal of the BP procedure is to minimize cost (error) function, which represents the difference between the expected output for a training example and the real output of the network. While there are different types of cost function, for BP procedure explanation let's assume we applied the mean-square cost function:

$$J(t) = 0.5 \sum_{i=1}^{N_o} (eo_i(t) - y_i(t))^2 \tag{4}$$

where *t* is an index of *t*-th training example, N_o is a number of neurons in the output layer, and *eo* is a real (expected) output value from the training set. An optimization process is based on the modification of weights of the neuron for time *t*+1 from previous values in time *t* as follows:

$$w_{ij}(t+1) = w_{ij}(t) - \Delta w_{ij}(t), \Delta w_{ij}(t) = \alpha \frac{\partial}{\partial w_{ij}} J(t) \tag{5}$$

where α is the learning rate. The computation of Δw_{ij} can be written as:

$$\Delta w_{ij}(t) = \alpha \frac{\partial J(t)}{\partial in_i(t)} \frac{\partial in_i(t)}{\partial w_{ij}(t)}, \frac{\partial J(t)}{\partial in_i(t)} = \delta_i(t), \frac{\partial in_i(t)}{\partial w_{ij}(t)} = x_j(t) \tag{6}$$

then weight update can be written as:

$$\Delta w_{ij}(t) = \alpha \delta_i(t) x_j(t). \tag{7}$$

Now the problem of NN learning is to find $\delta_i(t)$ for every neuron in the network. This leads to a simple recursive equation for $\delta_i(t)$ computation (for *i*-th neuron in the layer), which represents backpropagation of error.

For output layer neurons equation for $\delta_i(t)$ is:

$$\delta_i(t) = \big(eo_i(t) - y_i(t)\big) f'(in_i(t)) \tag{8}$$

and for neurons in the hidden layer $\delta_i(t)$ is computed as (where N_o is number of neurons on the output layer to the current hidden layer neuron):

$$\delta_i(t) = f'(in_i(t)) \sum_{h=1}^{N_o} \delta_h(t) w_{hi}(t) \tag{9}$$

While backpropagation can be applied in the same way (or with some changes related to faster and more effective learning), the main difference of the expected output model of NN and its application can be achieved by the use of a different cost function.

4 Our Topic Modeling Autoencoder

An autoencoder is a type of neural network with a hidden layer, which is trained to provide the same output on its output layer as input on the input layer. Then, the hidden layer (with a smaller number of neurons) holds encoding information on training examples. Usually, autoencoder is used for dimensionality reduction or features learning. Simply, autoencoder network can be viewed as a two-part network with: the encoder function $h = f(x)$ and decoder function $r = g(h)$. The encoder is used to reduce dimensionality and produce a smaller number of input features, while the decoder is used to reconstruct original input from reduced representation in the hidden layer. The basic architecture of the autoencoder is shown in Figure **6**. In this case, we have layers $X \approx R$ and Z is a representation (coding) of inputs with reduced dimensionality.

4.1 Autoencoder for Topic Modeling

Our topic modeling autoencoder (TopicAE) is inspired by the sparse autoencoder presented in [17]. The architecture of the autoencoder is the same as in Figure **6**, where input layer *X*is used for documents represented by words (containing

variables x_i for every word in documents within a corpus) and hidden layer Z providing induced topics in its K neurons (with k-thtopic represented by z_k).

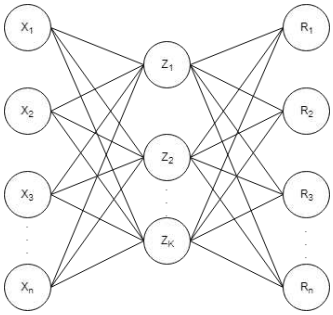


Figure 6

The architecture of autoencoder with n input/output neurons (with particular variables within input layer X and output layer R) and K neurons in a hidden layer. In the case of TopicAE, variables in layer X represent particular words in the vocabulary of documents in the corpus and Z contains K different topics.

For TopicAE, we propose a topic penalty on encoder (hidden) layer units, which is based on a sparse penalty. Moreover, we need to achieve generative properties of the method to provide an approach which imitates the behavior of generative topic models such as LDA. Such assumption leads to the approach where only several neurons in the hidden layer (representing topics) should be activated (neuron is activated when its value is near 1 and inactive when its value is close to 0) for each input (document), and also each topic should be activated only for several documents. To achieve this behavior, we added a topic penalty to cost function in the hidden neurons layer as described in the following equations:

$$J_{topic}(t) = J(t) + \Omega(t) \tag{10}$$

$$\Omega(t) = \alpha \sum_{i=1}^{m} KL(\rho||\rho_i') + \beta \sum_{i=1}^{m} KL(\varsigma||\varsigma_i') + \gamma \sum_{i=1}^{h} KL(\sigma||\sigma_i') \tag{11}$$

where α, β, γ controls the weight of the penalty terms in cost function (usually set to 1), m is a number of training examples, h is a number of neurons in the hidden layer. $KL(\rho||\rho_i')$ represents the penalty based on KL divergence (which describes the divergence of one probability distribution to another) chosen as follows:

$$KL(\rho||\rho_i') = \rho \, log \frac{\rho}{\rho_i'} + (1-\rho) log \frac{1-\rho}{1-\rho_i'} \tag{12}$$

where ρ_i' is average activation of the hidden units for i-th training example, ς_i' is median of activations of the hidden units for i-th training example, σ_i' is average activation of the hidden units over the training set and $\rho, \varsigma, \sigma$are constants (typically small values close to zero, e.g.,0.05 - to make average values of ρ_i',ς_i',σ_i' of each hidden neuron to be close to 0.05). To achieve similar distribution of topics for each training example (document) as in classical topic modeling methods, $\varsigma < \rho$ should be true.

In TopicAE, we represent each input text document as a log-normalized word count vector $\boldsymbol{x} \in R^d$, where each dimension is represented as:

$$x_i = \frac{log(1+n_i)}{max_{i \in V} log(1+n_i)}, \forall\, i \in V \tag{13}$$

where V is the vocabulary and n_i is the word count in the document for i-th word in the vocabulary.

For activation function in each layer we used sigmoid function, and as a cost function, we selected binary cross entropy:

$$C = -\frac{1}{n}\sum_{i=1}^{m}(eo_i\, ln\, y_i + (1 - eo_i)\, ln(1 - y_i)) \tag{14}$$

As mentioned above, topics for each input document can be obtained from the hidden layer, where each neuron represents one of the topics. To find specific words which describe the k-th topic we need to strongly activate particular k-th neuron (set its value to 1 and values of other neurons to 0), compute the output activations and obtain the words which correspond to output units.

In order to see the evolution of topics in time (from data streams or from the whole dataset with timestamps), TopicAE can be simply applied on a chronologically ordered input corpus with the usage of particular documents or their batches (smaller sets of documents from the defined period). The application of TopicAE in batches leads to the visualization of topics evolution in time, where it is possible to follow any topic and changes in its particular description (example of one topic evolution in time is shown in Table 3).

4.2 Extension for Hierarchical Topic Modeling

One of harder tasks in topic modeling is a possibility to extract a hierarchical structure of topics. To extract the hierarchical structure of topics we need to extend architecture and combine more autoencoders in a specific way. It means that for a hierarchical model of topics with depth h, we need to combine h TopicAE autoencoder layers. In practice, such architecture for $h=3$ is shown in Figure 7. Here we have three TopicAE hidden layers to learn three levels of hierarchy topics. Hence, our model is going to learn reconstruction function of input for every possible output $(X \approx R_1, X \approx R_2, X \approx R_3)$. Similarly as in TopicAE case,H_1, H_2, H_3 (where H_1 represents the most specific topics and $H_3$the most general topics) represent topics distribution for input and weights between these layers represent dependencies between topics in each layer.

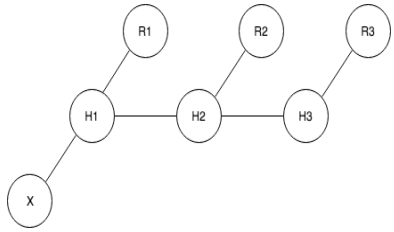


Figure 7
The architecture of the composition of three autoencoders (h=3) for extraction of the hierarchical topic model with three levels of subtopics

The main problem with learning such a composition of autoencoders is that using only J_{topic} penalty, we find it problematic to extract meaningful representation of dependencies between the topic layers. It is simply because TopicAE in the basic setup extracts topics dependencies across layers with similar weights. This behavior is expected if we do not need to have subtopics of some higher topics. On the other hand, in the hierarchical model we expect that whenever a topic is activated on a higher level, on a lower level only subtopics related to such activated parent topic are also activated. This is similar to J_{topic} penalty, but instead of relations topic-document (in simple TopicAE) we need relation topic-subtopic, which leads us to weights between autoencoders' hidden layers. In our composition of TopicAE autoencoders, this problem can be solved with a penalty added to weights between two topic layers, which will add such behavior to a composition. Therefore, in order to achieve this behavior we propose dependency penalty for weights between topics layers as follows:

$$J_{depencency}(t) = J(t) + \Omega(t) \tag{15}$$

and

$$\Omega(t) = \alpha \sum_{i=1}^{s1} KL(\rho||\rho_i') + \beta \sum_{i=1}^{s1} KL(\varsigma||\varsigma_i') + \gamma \sum_{i=1}^{s2} KL(\sigma||\sigma_i') \tag{16}$$

Where*s 1* is a number of weights at a more specific topic layer and *s2* a number of topics in a more general layer. The application of dependency penalty will learn weights between topic layers in the composition of TopicAEs, i.e., this network will activate only some topics in a more specific level (e.g., H_1) that belong to activated topics in a more general level (e.g., H_2). Then, it is only on our decision how many levels of subtopics we want to have and learn the composition of TopicAE autoencoders.

5 Experiments

In this section, we evaluate our proposed TopicAE with other topic modeling methods and its effectiveness in the extraction of topic structure in time as well as the hierarchical structure of topics.

The evaluation was performed on the Reuters Dataset[1], which contains 90 classes, 10788 documents and vocabulary consisting of 35247 unique words, and also on the 20Newsgroups dataset[2], which contains 18846 documents divided into 20 classes. For all of the evaluated methods we preprocessed datasets in the following way:

- Tokenization - split of texts into tokens (in our case words),
- Removing stopwords and words with a length smaller than three characters,
- Word lemmatization and normalization to lowercase form,
- Selection of words which occurred in at least ten documents, but in less than 50% of documents.

The preprocessing steps reduced the initial vocabulary of Reuters dataset from 35247 words to 4672 words. For 20Newsgroups dataset, we took 2000 most frequent words filtered by preprocessing.

For the evaluation of our experiments, we decided to select two standard evaluation metrics. First, we used UMass topic coherence [18] to evaluate the quality of the extracted topics. It represents the pairwise score of *n* top words of the topic and is defined as follows:

$$coherence = \sum_{i<j} log \frac{D(w_i,w_j)+1}{D(w_i)} \tag{17}$$

where $D(w_i)$ is defined as a number of documents containing the word w_i and $D(w_i, w_j)$ is a number of documents containing both words w_i and w_j.

As a second metric, we applied normalized mutual information (NMI) [19], which evaluates how diverse the topics are. For NMI we at first selected top N words (in our case N = 100) for each topic and divided them into 10 clusters $\{<w_1:w_{10}>,<w_{11}:w_{20}>,...,<w_{91}:w_{100}>\}$. Then, these clusters were evaluated on how similar two topics were according to N top words. The final value of the NMI evaluation metric was the average of NMI between every two topics. We also compared our proposed model with several neural network models (ProdLDA, NVDM) and LDA model variations described in paper [26]. For that purpose we also evaluated our model using normalized point wise mutual information (NPMI) topic coherence [26]:

$$NPMI(w_i) = \sum_{j}^{N-1} \frac{log\frac{P(w_i,w_j)}{P(w_i)P(,w_j)}}{-logP(w_i,w_j)} \tag{18}$$

[1]https://martin-thoma.com/nlp-reuters/
[2]http://qwone.com/~jason/20Newsgroups/

5.1 Standard Topic Modeling

Now, we describe the results from experiments in a standard topic modeling task, i.e., we evaluate the quality of extracted topics by TopicAE with other standard methods such as LDA and LSI. For these experiments, we ran TopicAE using autoencoder architecture with one hidden layer and the following parameters for the topic penalty: $\rho = 0.03, \varsigma = 0.01$, $\sigma = 0.03$. The learning phase consisted of 30 epochs. The values of parameters and number of epochs were selected by testing several settings of their values and we selected values which gave us better results.

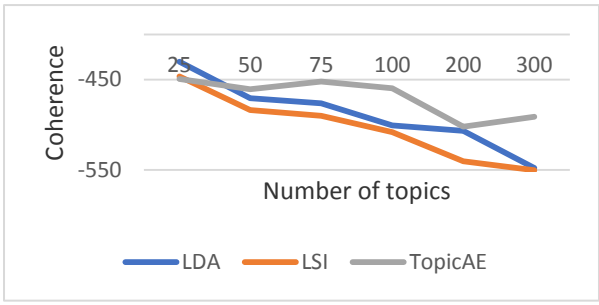


Figure 8
Average topic coherence for LDA, LSI, and TopicAE - Reuters dataset (higher value is better)

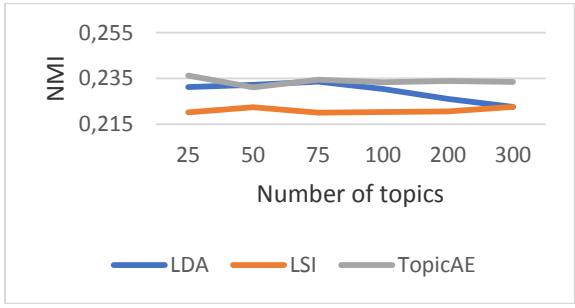


Figure 9
Comparison of average NMI for LDA, LSI, andTopicAE- Reuters dataset (lower value is better)

Figure 8 presents the results of average topic coherence for the compared methods using different numbers of topics (on Reuters dataset). As we can see from the graph, TopicAE out performs standard topic modeling methods according to coherence evaluation metric. The comparison of these methods using NMI (see **Figure 9**) shows that our proposed approach is more likely to generate topics with similar word collocations. Some of the extracted topics are illustrated in Table 1, where we show ad-hoc selected topics from 100 extracted topics for the whole dataset.

Table 1
Example of generated topics by TopicAE

Dollar	Cooper	Orange	Repurchase	Cable
Miyazawa	Ounce	Barley	Corp	Telecommunication
Dealer	Gold	Maize	Reserve	Wireless
Tokyo	Mine	Juice	Customer	Merge
Japan	Loss	Tonne	Temporary	Hold
Tonne	Mining	Argentine	Security	Share
Nakasone	Ton	Gallon	Tonne	Settlement
Stock	Quabec	Grain	Well	Company

Table 2 gives comparison of TopicAE with other topic models (described in paper [26]). From the results, it is obvious that our method out perform other methods by given metric, except ProdLDA.

Table 2
Average NPMI topic coherence on the 20 newsgroups dataset (higher value is better)

# topics	ProdLDA VAE	LDA VAE	LDA DMFVI	LDA Collapsed Gibbs	NVDM	TopicAE
50	0.24	0.11	0.11	0.17	0.08	0.18
200	0.19	0.11	0.06	0.14	0.06	0.17

5.2 Online Topic Evolution in Time

In this subsection, we show how it is also possible to use TopicAE to learn topics evolution in time and evaluate its quality against LDA and LSI. For the purpose of these experiments, we ran TopicAE using the same architecture as in the previous subsection with the same parameters of topic penalty. In this case we only did experiments with Reuters dataset because of time metadata for particular documents. To be able to simulate topic evolution we ordered documents chronologically and divided them into 10 batches $t=\{t_1,t_2,....,t_{10}\}$, each consisting of about 1050 documents.

TopicAE was first learned in the batch with the oldest documents (to be able to learn initial representation of topics we ran TopicAE using 50 iterations). Next, we used the learned TopicAE inthe next batch, and so on. After obtaining the initial topics representation from the first batch run, for all non-initial batches we only used 20 epochs for learning.

In Figure **10** we can see coherence for topics extracted by different models for each of the batches. TopicAE retains stable values of coherence during topics learning and outperforms LDA and LSI. Figure **11** shows that also NMI score has stable values and is still comparable to other topic modeling methods.

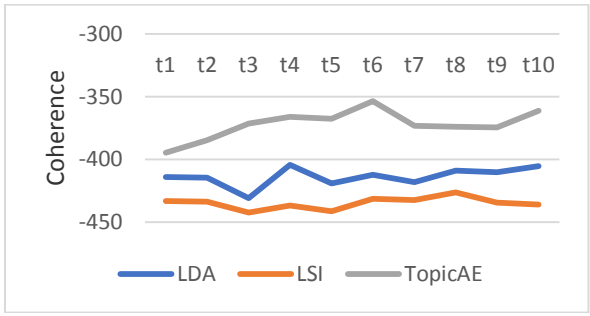


Figure 10
Comparison of coherence for learning topics evolution in time on Reuters datasest batches (higher value is better)

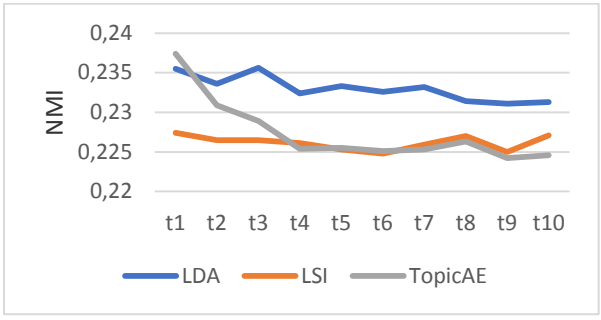


Figure 11
Comparison of NMI for learning topics evolution in time on Reuters datasest batches (lower value is better)

In Table 3 we illustrate an example of topic evolution over the time for one topic extracted using TopicAE. Here, we can see changes in the most characteristic words extracted for the topic during batches from T1 to T10 (words with higher probability are always top within the time batch).

Table 3
An illustrative example of the evolution of topic extracted by TopicAE

T1	T2	T3	T4	T5
Billion	Billion	Week	Week	Week
Week	Dlrs	Billion	Say	Say
Barrel	Week	Dlrs	Dlrs	Money
Rose	Rose	Rose	Barrel	March
Bank	Stock	Barrel	Money	Dlrs
Year	Barrel	Say	Supply	Supply
Stock	Fell	Fell	Rose	Ended
Crude	Supply	Supply	Bond	Growth
Fell	Year	Dollar	Stock	Barell
december	money	government	billion	Rose

T6	T7	T8	T9	T10
Week	Week	Week	Week	Week
March	Ended	Barrel	barrel	Say
Ended	Barrel	Distillate	Distillate	Barrel
Demand	March	Gasoline	Gasoline	Gasoline
Say	Distillate	Weekly	Weekly	Distillate
Crude	Gasoline	Demand	Demand	Stock
Economic	Crude	Stock	Stock	Weekly
Stock	Stock	Ended	Ended	Demand
Distillate	Say	Say	Say	Ended
supply	petroleum	crude	crude	crude

5.3 Hierarchical TopicAE

The composition of TopicAE autoencoders is also applicable for the extraction of the hierarchical structure of topics. For the purpose of these experiments we used TopicAE layered architecture with two hidden layers, one with 200 neurons (more specific layer) and second with 50 neurons in hidden layers (more general layer), and with the following parameters for each of the topics and dependency penalties: $\rho = 0.05, \varsigma = 0.03,\ \sigma = 0.05$. We used 30 epochs for learning on 20Newsgroups dataset, which was preffered for experiments with hierarchical structuring due to better separation of main classes. Similar to previous experiments for simple topic modeling, the parameters and number epochs were selected by previous testing of several settings.

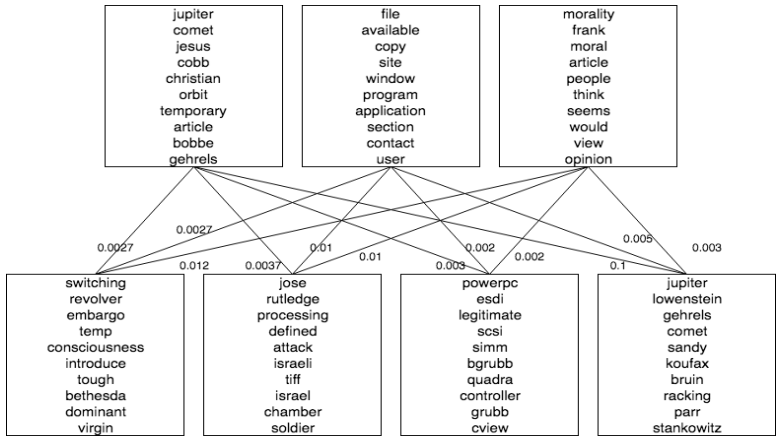


Figure 12

Part of the extracted hierarchical structure from the 20Newsgroup dataset (TopicAE with two levels)

In general, there is currently nohierarchy-based evaluation metrics available. Therefore, we applied evaluation metrics in a particular hierarchical level. Here, we achieved *coherence = -581.3* and *NMI = 0.25* for topics in a more specific TopicAE level (layer with 200 neurons). In a more general TopicAE level (with

50 neurons in the hidden layer), evaluation metrics had values of *coherence = -458.6.97* and *NMI = 0.235*. Figure **12** shows part of the extracted hierarchical structure. Here we can see that hierarchical TopicAE architecture was able to discover meaningful topics on each hierarchy level and also to find similar topics across the hierarchy.

Conclusion

In this paper, we proposeda novel neural network approach to solving the topic modeling problem usingTopicAE autoencoder. The main advantage of this solution is that it can be applied to classical topic modeling, topic modeling over time and to the extraction of the hierarchical structure of topics. Our experiments showed that TopicAE could extract topics with similar or better quality (measured by evaluation metrics) than other standardly applied models. Also, our model was likely to generate more topics with similar words.

For future work, we want to eliminate the problem of most topic models, which is that they are only able to work with a bag-of-words model for their input. We expect that it can be solved using additional layers in our network, i.e. it will be possible to represent the input using an embedding layer in our network.

Acknowledgment

The work presented in this paper was supported by the Slovak VEGA research grant 1/0493/16,and Slovak APVV research grants APVV-16-0213 and APVV-17-0267.

References

[1] T. K. Landauer, P. W. Foltz and D. Laham: An introduction to latent semantic analysis, *Discourse Process*, Vol. 25, pp. 259-284, 1998

[2] D. M. Blei, A. Y. Ng and M. I. Jordan: Latent dirichlet allocation, *JMLR*, Vol. 3, pp. 993-1022, 2003

[3] T. Hofmann: Probabilistic latent semantic indexing, *SIGIR99*, 1999, pp. 50-57

[4] J. Petterson *et al.*: Word features for latent dirichlet allocation, *NIPS*, 2010, pp. 1921-1929

[5] K. Zhai and J. Boyd-Graber: Online Latent Dirichlet Allocation with Infinite Vocabulary, *ICML*, 2013, pp. 561-569

[6] Y. W. Teh*et al.*: Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes, *NIPS*, 2004, pp. 1385-1392

[7] D. M. Blei and J. D. Lafferty: Dynamic topic models,*ICML*, 2006, pp. 113-120

[8] Ch. Wang, D. Blei and D. Heckerman: Continuous time dynamic topic models, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 579-586

[9] A. Beykikhoshk *et al.*: Discovering topic structures of a temporally evolving document corpus, *KAIS*, Vol. 55, pp. 599-632, 2018

[10] D. M. Blei, T. L. Griffiths and M. I. Jordan: The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies, *JACM*, Vol. 57, 2010

[11] T. Hoffman: The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data, *IJCAI*, Vol. 99, 1999

[12] A. Smith, T. Hawes and M. Myers: Hierarchie: Interactive Visualization for Hierarchical Topic Models, *ILLVI*, 2014, pp. 71-78

[13] W. Li and A. McCallum: Pachinko allocation: Scalable mixture models of topic correlations, *JMLR*, 2008, Submitted

[14] P. Chen *et al.*: Progressive EM for Latent Tree Models and Hierarchical Topic Detection, *AAAI*, 2016, pp. 1498-1504

[15] J. Paisley *et al.*: Nested hierarchical Dirichlet processes, *TPAMI*, Vol. 37, pp. 256-270, 2015

[16] *Deep Learning*, I. Godfellow, Y. Bengio and A. Courville, MIT Press, 2016 [Online] Available: http://www.deeplearningbook.org/

[17] A. Ng: Sparse autoencoder, CS294A Lecture Notes, Vol. 72, pp.1-19, 2011

[18] D. Mimno *et al.*: Optimizing semantic coherence in topic models, *Proceeding of EMNLP* 2011, pp. 262-272, 2011

[19] A. F. McDaid *et al.*: Normalized Mutual Informationtoevaluate overlapping community finding algorithms, 2011

[20] R. Salakhutdinov and G. Hinton: Replicated Replicated Softmax: an Undirected Topic Model, *NIPS*, 2009, pp. 1607-1614

[21] H. Larochelle and S. Lauly: A Neural Autoregressive Topic Model, *NIPS*, 2012, pp. 2708-2716

[22] Y. Miao, L. Yu and P. Blunsom: Neural Variational Inference for Text Processing, *ICML*, 2016

[23] Z. Cao*et al.*: A Novel Neural Topic Model and Its Supervised Extension, *AAAI*, 2015, pp. 2210-2216

[24] Y, Chen and M. Zaki: KATE: K-Competitive Autoencoder for Text, *KDD*, 2017, pp. 85-94

[25] P. Gupta *et al.*: Deep Temporal-Recurrent-Replicated-Softmax for Topic Trends over Time, *NACCL HLT*, 2018, pp. 1079-1089

[26] A. Srivastava and Ch. Sutton: Autoencoding Variational Inference For Topic Models, *ICLR*, 2017

Info-Chunk Objects as New Behavior Representation for System-based Model of Product

Yatish Bathla

Doctoral School of Applied Informatics and Applied Mathematics, Óbuda University, Bécsi út 96/b, H-1034 Budapest, Hungary
yatish.bathla@phd.uni-obuda.hu

Abstract: Requirement Functional Logical Physical (RFLP) structure has emerged as one of the prominent approaches for modeling the multidisciplinary products. Information Content (IC) provides effective interaction between the human and multidisciplinary product model. Though it controls the RFLP level by the Multilevel Abstraction based Self-Adaptive Definition (MAAD) structure, it needs to be further enhanced in terms of Human-Computer Interaction (HCI), multidisciplinary product behavior representation and structured processing of interrelated engineering objects to obtain coordinated decisions. Therefore, this paper introduces the Object-Oriented Principle (OOP) concepts in the IC for behaviors representation of the multidisciplinary product where Info-Chunk is considered as an object. Here, Behavior Info-Chunk (BiC) and Context Info-Chunk (CxiC) objects are proposed in the MAAD structure to model the behavior of the multidisciplinary product. Further, the concepts of Info-Chunk objects are extended to Intelligent Property (IP) that uses Initiative Behavior Context and Action (IBCA) structure to handle the RFLP structure. Based on the communication between the MAAD and RFLP structure, an API (Application Programming Interface) called "InfoChunkLib" is proposed. It can generate the graphs to represent the behaviors of a multidisciplinary product model. The API is handled by the information content to represent the behavior information and store the results in a database.

Keywords: Behaviors representation; Multidisciplinary product modeling; Info-Chunk based Information Content; RFLP structure; MAAD structure; IBCA structure

1 Introduction

Modeling of multidisciplinary products requires coordination of a significant amount of model information. The integrated definition is raised to the conceptual level of product design, which requires high-level abstraction. A four-leveled structure of the product model using Requirement Functional Logical Physical

(RFLP) structure [2] was introduced in the virtual environment. It is applied from system engineering and offers handling product and its model as a system. It accommodates product behavior definitions on its F and L levels. Product assembly is done in the specification tree (red square) of RFLP structure as shown in Fig. 1. Due to complex Human-Computer Interaction (HCI), Information Content (IC) was used to record and apply the content of information that is represented in the product model space [19]. In this content, an intent is defined by the human to control the definition of engineering objects [22]. IC [1] controls the RFLP level by the Multilevel Abstraction based Self-Adaptive Definition (MAAD) structure [2]. However, IC needs to be enhanced in terms of the practical feasibility of HCI, behavior representation and structured processing of interrelated engineering objects to obtain the coordinated decisions. To solve above-mentioned issues, this research work proposes the Info-Chunk objects and InfoChunkLib API (Application Programming Interface) is proposed in the IC.

Info-Chunk objects are based on the Object-Oriented Principle (OOP) concepts that are used in software programming. Previous research work deployed OOP concepts in the RFLP structure in the form of the Modelica language [6]. It is used for logical and physical modeling of a multidisciplinary product. Here, models and their components are defined by the object diagram. This research work uses OOP concepts in the Functional layer and Logical layer of the RFLP structure for behaviors representation of the multidisciplinary product modeling. Here, Info-Chunk entity is converted into the object first. Then, Behavior Info-Chunk (BiC) object and Context Info-Chunk (CxiC) object are introduced in the MAAD structure and the IBCA structure to store the behaviors of the multidisciplinary product. The proposed Info-Chunk objects are used to establish a link with the Layer Info-Chunk (LiC) objects of RFLP structure. InfoChunkLib API is coded based on the communication between the MAAD structure and RFLP structure. The Java language is used as a JavaFX application. It represents the behaviors of the components in the multidisciplinary product model. The generated output is shown using the graph between the components of engineering disciplines. IC imports the InfoChunkLib API and coded as a Web application. The multidisciplinary product model can be more efficiently handled through the IC instead of the Specification Tree. This paper begins with the conversion of Info-Chunk entity into the object. Further, behavior Info-Chunk (BiC) objects and context Info-Chunk (CxiC) objects are proposed in the MAAD structure and the IBCA structure. Then, behavior storing techniques for the multidisciplinary product are explained with the aforementioned concepts. Then, Info-Chunk objects based Information Content (IC) is emphasized. Here, rules for generating the BiC objects and CxiC objects in the MAAD structure are defined by using the pseudo-codes. Then, InfoChunkLib API is over viewed. Here, LiC objects of the RFLP structure, BiC and CxiC objects of the MAAD structure are demonstrated. Finally, InfoChunkLib API is imported in the IC to handles the multidisciplinary product model.

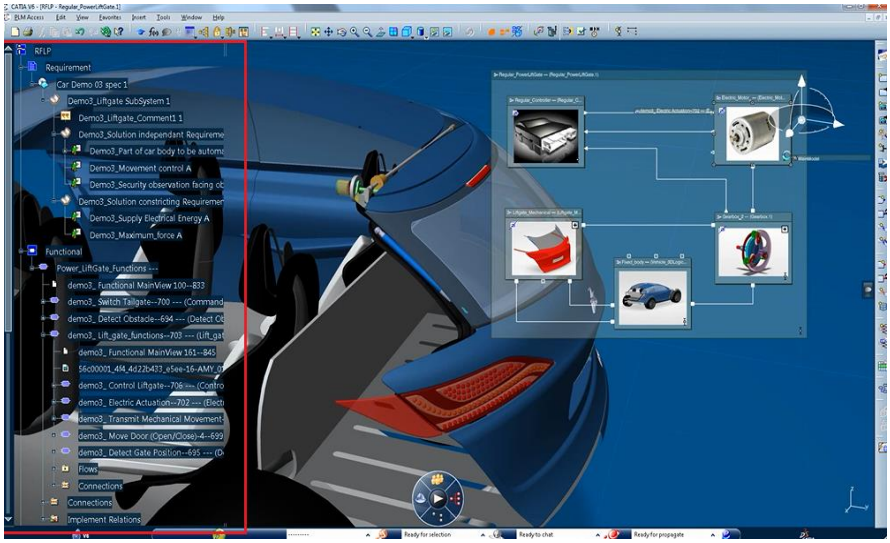


Figure 1
Specification tree of the CATIA V6 RFLP structure

2 Background

The Classical Product Model (CPM) [1] is limited to the physical level. The separated integrated mechanical engineering modeling increasingly demands multidisciplinary integration [5]. Modeling of a multidisciplinary product must have a means for the integration of discipline-specific models into a model with a unified structure. It makes the product model virtually executable. Higher abstraction is realized by using of RFLP structure product model [2]. It is compliant with the IEEE 1220 standard. Requirement against the product function to fulfill the requirement, product-wide logical connections, and representations of physically existing objects was organized in the highly contextual RFLP structure.

Human-Computer Interaction (HCI) during the multidisciplinary product modeling is a challenging task. Therefore, Information Content (IC) [1] assists in effective communication between engineers of different disciplines and information-oriented product modeling procedures. Community zones [17] are used in the IC to organize the product model entities and their relationship. Further, behaviors of the modeled entities are evaluated in the information content by the process plane [13]. IC requires MAAD structure to drive the levels of RFLP structure. This structure is used for self-adaptive modeling, where the *objectives and requests* level, *product behaviors* level, *contexts* level, *actions*

level, and feature objects are applied in order to connect engineers with RFLP implementations [4]. The MAAD modeling methods and model structures are introduced as a generalized means for the support of higher level abstraction-based generation of RFLP elements. The MAAD modeling was based on the knowledge representation, contextual change propagation, and extended feature definition capabilities for advanced modeling systems [4]. Further, active knowledge in a product model has become organized in the form of Intelligent Property (IP) of the company. Here, IP drives the RFLP level by the IBCA structure which represents active knowledge content [5].

To store the information of a multidisciplinary product, Info-Chunk entity is introduced in the logical level of the RFLP structure [9]. This entity is mapped with the information content to control the structure activities through the MAAD structure. Here, the Layer Info-Chunk (LiC) entity stores the information of the Logical layer and the Component Info-Chunk (CiC) entity stored the information of the logical component. Then, the Info-Chunk entity is defined in the Functional layer of the RFLP structure [19]. Here, the Layer Info-Chunk (LiC) entity stores the information of the main function of the Functional layer of the RFLP structure and Sub-function Info-Chunk (SFiC) entity stores the information of sub-function. Nowadays, OOP concepts are used in system engineering as object-oriented system engineering (OOSE) [3]. OOSE blends system engineering with software engineering.

3 Info-Chunk as an Object

In this research work, OOP concepts are used for multidisciplinary product modeling. Encapsulation, inheritance, and polymorphism are the three principles of OOP methodology. This work starts with the entities and their relationship. Info-Chunk [9] [19] is an entity defined in the RFLP structure. However, OOP concepts are not directly applicable to an entity. For InfoChunkLib API, Info-Chunk entity must be converted into the Info-Chunk object for communication between IC and RFLP structure. Based on the entity-object conversion process by Ou Y. [10] and Bernhard Thalheim [11]:

- The parameters of an Info-Chunk entity is equivalent to the attribute of an Info-Chunk object
- ER (Entity Relationship) between Info-Chunk is equivalent to the OR (Object Relationship). Here, the method of an Info-Chunk object is derived from the OR as per the requirement of a specific discipline

Then, behavior Info-Chunk (BiC) object and context Info-Chunk (CxiC) object are proposed in the MAAD structure and IBCA structure. According to the proposed concept of Info-Chunk objects:

- In the RFLP structure, logical layer Info-Chunk (LiCL) object consist of the attributes and methods of the Logical level and functional layer Info-Chunk (LiCF) objects consist of the attributes and methods of Functional level
- In the MAAD structure, behavior layer Info-chunk (BiC) object consist of the attributes and methods of Behaviors level and context layer Info-Chunk (CxiC) objects consist of attribute and method of Contexts level
- In the IBCA structure, behavior layer Info-chunk (BiC) objects consist of the attributes and methods of *Situation defining behaviors (SB)* level of Behavior substructures and context layer Info-Chunk (CxiC) objects consist of attribute and method of *Product definition Activity Contexts (AC)* level, *Adaptive Drive Contexts (DC)* level, *Product Feature Contexts (FC)* level of Contexts substructures

4 Behavior Storing Techniques using Info-Chunk Objects

Behavior is based on well-defined situations for sets of circumstances. It is represented in the Functional level and Logical level of the RFLP structure. BiC objects and CxiC objects represent dynamic behavior information. They are stored in the MAAD structure and IBCA structure to communicate with the LiC objects of the RFLP structure. Information Content operates the RFLP structure by the MAAD structure. Also, Intelligent Property (IP) operates the RFLP structure by IBCA structure. The behavior storing techniques are classified as the operation performed by the BiC objects and CxiC objects in the MAAD structure and IBCA structure.

4.1 Info-Chunk Objects-based MAAD Structure

The Behavior level of the MAAD structure drives the Functional level and Logical level of the RFLP structure. The relationship between the abstraction levels of the MAAD structure is described by using the Unified Modelling Language (UML) diagram as shown in Fig. 2. In the Object-Oriented Programming, a UML diagram is used to define the relationship and model the behavior of the product.

Here, Entities Relationship Modeling (ERM) of the MAAD structure is converted into object relationship modeling (ORM). As per the ORM concept, the relationship between objects is defined by the composition, aggregation, and association. Hence, there is a bi-directional association relationship between the *Objectives and Requests* level, *Behaviors* level, *Contexts* level, and *Actions* level.

Inside the Behaviors level, the behavior object has a composite relationship with the situation object, which further has an aggregation relationship with the circumstances object. Also, the behavior object has a bi-directional association with the Adaptive Drive object. In the MAAD structure, a behavior is represented at Behaviors and Contexts level. For behavior representation, communication between the RFLP structure and the MAAD structure is done by using the proposed BiC objects and CxiC objects.

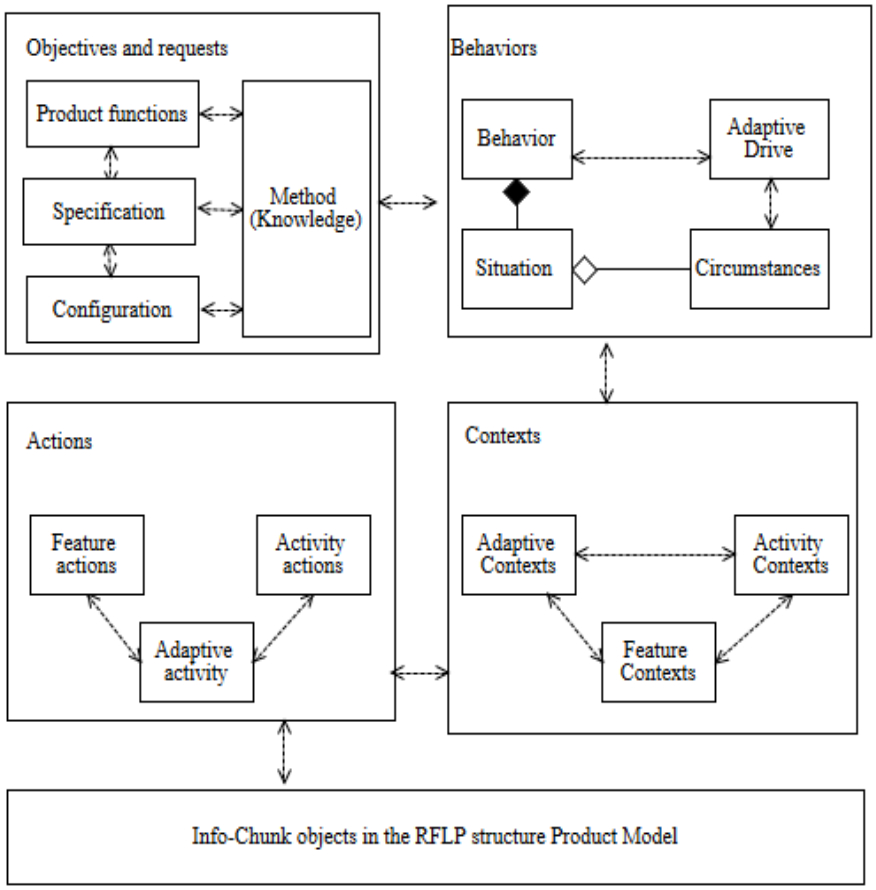


Figure 2
UML representation of MAAD structure with the Info-Chunk based RFLP Structure

The BiC objects communicate with the LiC objects as shown in Fig. 3, where the main contextual connections of the MAAD structure are organized as follow:

- The solid line is the inside contexts (C) of Behaviors levels for the MAAD structure. It is explained in the paper [18], where the contextual connection of model entities in the MAAD level is defined.

- The bold line is the driving contexts (D) of Behaviors levels for the MAAD structure. It drives the Functional level and Logical level of the RFLP structure. The dashed lines are the information retrieved by the BiC objects from the LiC objects of the Functional level and Logical level.

In the case of the Logical layer of RFLP structure, it retrieves the *situation* attribute of the LiCL object {$LiCL_1$, $LiCL_2$, .. $LiCL_o$} and corresponding *behavior* attribute of their CiC objects {CiC_1, CiC_2, .. CiC_n}. It is represented inside the oval shape in the diagram. The information retrieved by the driving contexts populates the BiC objects in the Behaviors level of MAAD structure. Here, n is the number of CiC objects in a LiCL object and o is the total number of LiCL objects in the logical layer. The information retrieved is the actual situation, circumstances for the situation and the adaptive drive to drive context definitions.

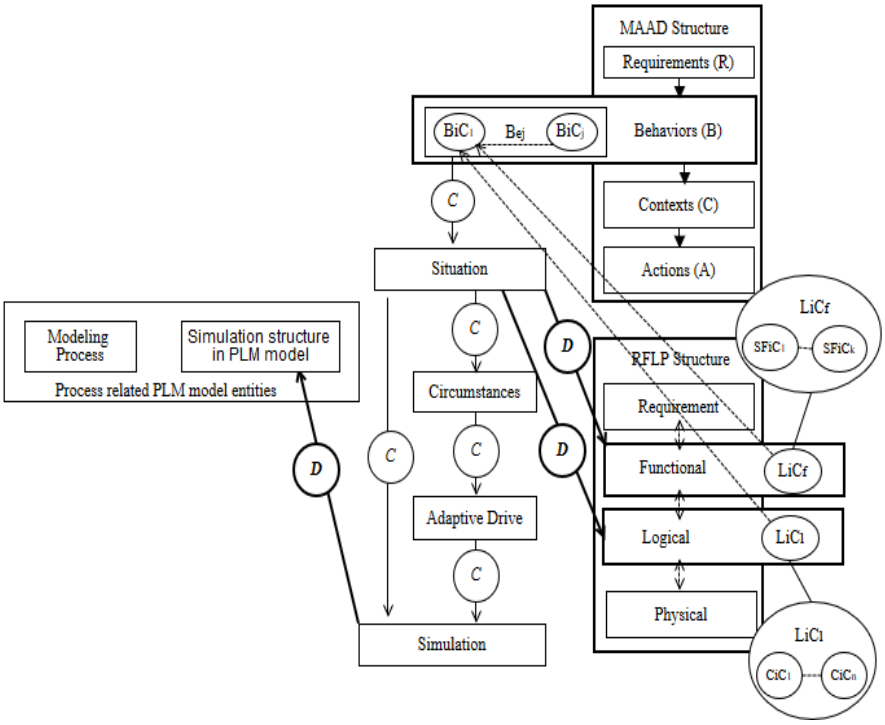


Figure 3
Communication between RFLP and MAAD structure at Behaviors level

In the case of Functional layer of RFLP structure, driving contexts (D) retrieves the *requirement class* attribute of the LiCF object {$LiCF_1$, $LiCF_2$, .. $LiCF_l$} and corresponding *elements description* attributes of the SFiC objects {$SFiC_1$, $SFiC_2$,.. $SFiC_k$}. It is represented inside the oval shape in the diagram. The information retrieved by the driving contexts populates the BiC objects in the Behaviors level

of MAAD structure. Here, k is the number of SFiC objects in a LiCF object and l is the number of LiCF objects in the functional layer of RFLP structure

The retrieved BiC objects are represented as {BiC_1, BiC_2,... BiC_j}. Here, j is the number of BiC objects in the Behaviors substructure. The CxiC objects communicate with the LiC objects is shown in Fig. 4, where the main contextual connections of the MAAD structure is organized as follows:

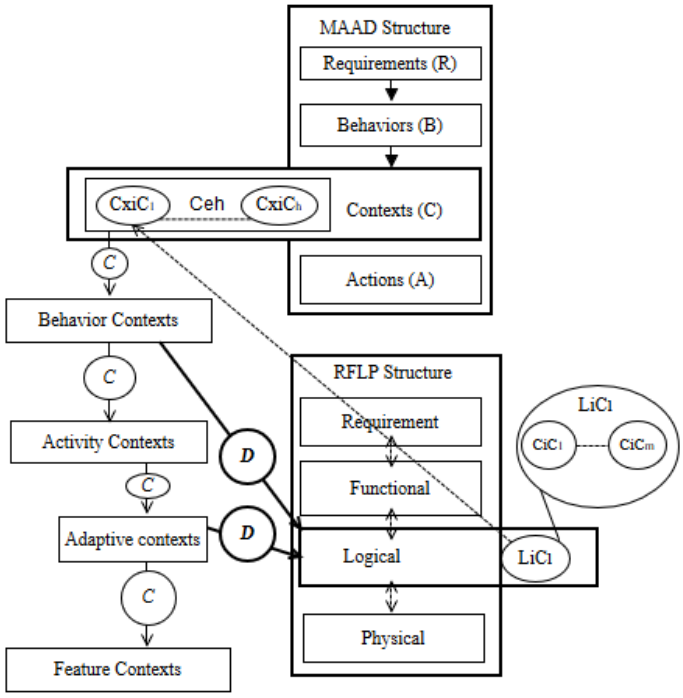


Figure 4
Communication between RFLP and MAAD structure at Contexts level

- The solid line is the inside contexts (C) of Contexts levels for the MAAD structure. It is explained in the paper [18], where the contextual connection of model entities in the MAAD level is defined.
- The bold line is the driving contexts (D) of Behaviors levels for the MAAD structure. It drives the Logical level of the RFLP structure. The dashed line is the information retrieved by the CxiC objects from the LiC objects of the Logical level.

In the case of the Logical layer of RFLP structure, it retrieves the *data model* attribute of the LiCL object {$LiCL_1$, $LiCL_2$, .. $LiCL_o$} and corresponding *data model* attribute of CiC objects {CiC_1, CiC_2, .. CiC_m}. Here, m is the number of CiC objects in a LiCL object and o is the total number of LiCL objects in the

logical layer. The information retrieved is the concept behavior, activity, adaptive and product feature contexts, connection behavior definitions, model definition activities, contexts for an adaptive drive, and context for physical level product and knowledge features. The retrieved CxiC objects are represented as {$CxiC_1$, $CxiC_2$,... $CxiC_h$}, where, h is the number of CxiC objects in the Contexts substructure.

4.2 Info-Chunk Objects-based IBCA Structure

The driving generation of the RFLP element is done by the Intelligent Property (IP). Human-initiated engineering activities with the company IP by using IBCA structure for the generation of RFLP elements. It leads to the analysis of self-adaptive product lifecycle management (PLM) modeling. The Info-Chunk objects based IBCA structure drives the RFLP structure as shown in Fig. 5. The solid lines are the interaction between the IBCA structure and RFLP structure. The dashed lines are the information retrieved by the BiC objects and CxiC objects from the LiC objects of the Functional level and Logical level. On the Behavior (B) level of the IBCA structure, situations defining behaviors (SB) substructure are configured to define behaviors by a set of BiC objects.

- In the Logical level of the RFLP structure, the *situation* attribute & *behavior* attribute of the LiCL object {$LiCL_1$, $LiCL_2$,... $LiCL_d$} and the corresponding *behavior* attribute of the CiC objects {CiC_1, CiC_2,... CiC_a} are stored in the BiC objects of the SB element. Here, a is the number of CiC objects in a LiCL object and d is the total number of LiCL objects.
- In the Functional level of the RFLP structure, *requirement* attribute of the LiC object {LiC_1, LiC_2,... LiC_c} and corresponding *elements description* attributes of the SFiC objects {$SFiC_1$, $SFiC_2$,.. $SFiC_b$} are stored in the BiC objects of the SB element. Here, b is the number of SFiC objects in a LiC object and c is the number of LiC objects in the functional level.

The stored information in the BiC objects is behavior definition (IEBD) and the related situation (IEBT) [12]. The total BiC objects obtained from the LiC objects of the functional and logical layer is represented as {BiC_1, BiC_2,... BiC_n}. Here, j is the number of BiC objects in the SB element. On the Contexts (C) level of the IBCA structure, *product definition activity contexts (AC)* level, *adaptive drive contexts (DC)* level, and *product feature contexts (FC)* level are configured to define behaviors by a set of CxiC objects. In the logical level of the RFLP structure, the *data model* attributes of LiC objects {LiC_1, LiC_2,... LiC_d} & CiC objects {CiC_1, CiC_2,.. CiC_a} are stored by the CxiC objects of AC, DC and FC elements. Here, a is the number of CiC objects in a LiC object and d is the total number of LiC objects. The stored information in the CxiC objects is the product behavior (IE_{CB}). The total CxiC objects obtained from the LiC objects of logical

layer is represented as {CxiC1, CxiC2, .. CxiCx}, {CxiC1, CxiC2, .. CxiCy}, {CxiC1, CxiC2, .. CxiCz}. Here, x, y, z are the number of CxiC objects stored in the AC, DC and FC elements.

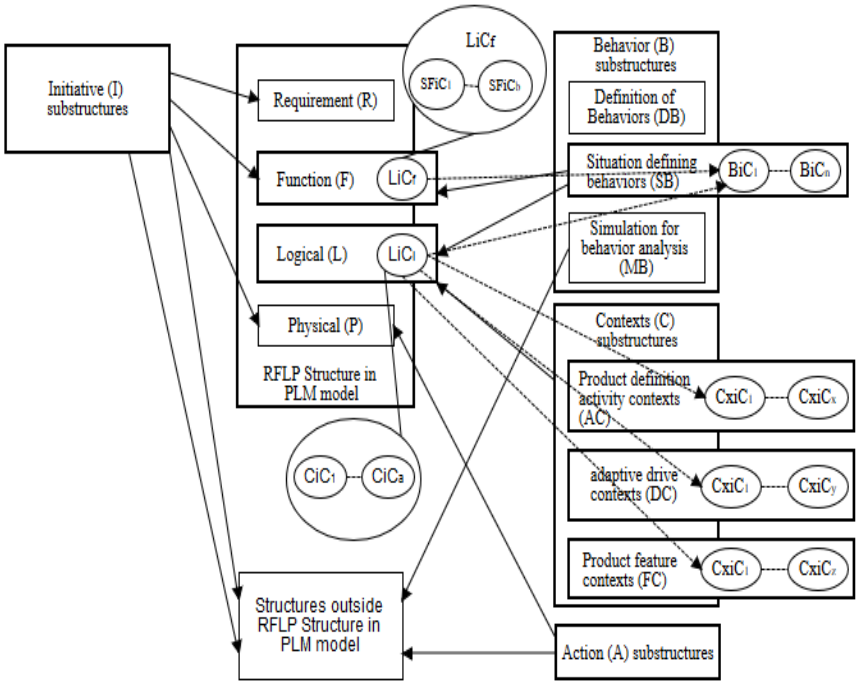


Figure 5
Communication between RFLP and IBCA structure at Behavior and Contexts substructure

5 Info-Chunk Objects-based Information Content

Behavior models with intelligent content involve specifications and knowledge for the design processes. The most appropriate forms of knowledge are formulas, rules, and checks. In the following sections, this work focuses on Info-Chunk object activities in the information content (IC). Here, the MAAD structure is the driving factor for representing the behavior of the RFLP structure.

5.1 Rules for the Generation of Info-Chunk Objects

Rules are the set of instructions that can be executed for generating and storing the Info-Chunk objects in the MAAD and IBCA structure. Rules are defined by using pseudo-codes.

- In the case of the MAAD structure, the behavior objects {BiC_1, BiC_2,... BiC_j} are stored in the behaviors level and the context objects {$CxiC_1$, $CxiC_2$,... $CxiC_h$} are stored in the contexts level. The Process plane of IC [13] can elaborate on the BiC and CxiC objects for the behavior representation of the multidisciplinary product. After the analysis process, the analyzed objects are stored with the nomenclature of BiC_{ab}. If a human wants to evaluate the context of one analyzed object on the other analyzed object, the context object undergoes the effect process. The resultant objects are stored as $CxiC_{ec}$. Further, If a human wants to optimize the contextual object, it is stored as BiC_{ob} after the optimization process. It is also possible to optimize the behavior of an object without analysis. Information content (IC) retrieve and store required objects at the *Engineering objectives* level to drive the behavior of RFLP structure.
- In the case of the IBCA structure, the behavior objects {BiC_1, BiC_2 , .. BiC_n} are stored in the behavior substructure and the context objects {$CxiC_1$, $CxiC_2$, .. $CxiC_x$}, {$CxiC_1$, $CxiC_2$, .. $CxiC_y$}, {$CxiC_1$, $CxiC_2$, .. $CxiC_z$} are stored in the contexts substructures. IP could retrieve and store these objects to drive the behavior of the RFLP structure. The IP level and process plane of IP are not defined yet. The behavior representation for IP is the topic of future work.

Pseudo Codes for BiC & CxiC objects

- BEGIN LOOP
- **Initialize** a Process
- IF 'Process' is 'Analysis'
    - BEGIN LOOP
    - Store 'BiC_{ab}' in 'Behaviors level' where $1 \leq ab \leq j$
    - IF 'Process' is 'Effect'
        - BEGIN LOOP
        - Store '$CxiC_{ac}$' in 'Contexts level' where $1 \leq ac \leq h$
        - IF 'Process' is 'Optimization'
            - BEGIN LOOP
            - Store 'BiC_{ob}' in 'Behaviors level' where $1 \leq ob \leq j$
            - END LOOP
        - END LOOP
    - END LOOP
- IF 'Process' is 'Optimization'

- BEGIN LOOP
- Store ‘BiCob’ in ‘Behaviors level’ where $1 \leq ob \leq j$
- END LOOP

6 Overview of the InfoChunkLib API

The *InfoChunkLib* API is coded in the JavaFX application as shown in Fig. 6. It consists of two Java packages. The *informationcontent* Package consists of all the classes related to the Information Content (IC) like *MAADStructure* class, *BiC* class, *CxiC* class, and *CommunityZone* class. The *rflp* Package consists of all the classes related to the RFLP structure like *LiCL* class, *LiCF* class, *CiC* class, and *SFiC* class.

6.1 Demonstration of Info-Chunk Objects in the RFLP Structure

To explain the proposed concepts in the system behavior, let us consider a car as an example. According to the community concepts [17], a car system is the combination of various communities where the *Electrical supply system* is one of the community. It consists of components like battery, starter, alternator, heater, fan, distributor, etc. Here, battery and alternator components are used for the Info-Chunk objects concept explanation. Then, the following scenario is considered as an example: The oil consumption of car depends on the engine behavior that must be modeled according to the situation such as the experience of the driver with the sets of circumstances like path traveled by car, the condition of the car, surrounding environment, etc.

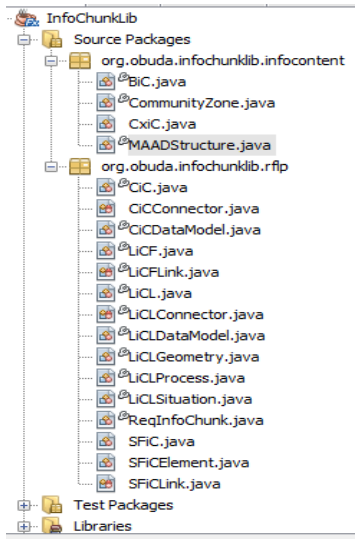


Figure 6
InfoChunkLib API

Here, the dynamic behavior of the engine strongly is influenced by the situation and weakly influenced by the circumstances. The parameters of the LiCL objects and CiC objects are described in the paper [9]. Also, the parameters of LiCF objects and SFiC objects are described in the paper [19]. The descriptions of LiCF

class and LiCL class are shown in the code below. The *LiCF* class is written to show the layer Info-Chunk object of the functional layer in the RFLP structure. Here, the array of SFiC objects, ReqInfoChunk object, LiCFLink object and String parameters are used as an argument in the constructor. The constructor arguments of the ReqInfoChunk object are populated from the Requirement layer of the RFLP structure. The *ReqInfoChunk* class is written to show the customer requirements in the requirement layer of the RFLP structure. The *SFiC* class is written to show the sub-functional Info-Chunk object in the functional layer of the RFLP structure. The LiCFLink is the enumeration class to store connector information. The concepts of constructor overloading are used so that LiCF can accept various sets of the argument depends on the initialization of the object.

```
// LiCF.java
/**This class is written to show the layer Info-Chunk
object of the functional layer in the RFLP structure
*@param func_name This parameter stores the name of a
function
*@param func_descrip This parameter stores the
description of a function
*@param comm_name This parameter stores the community
name of a function
*@param func_input This parameter stores the inputs to
a LiCF object
*@param func_output This parameter stores the outputs
from a LiCF object
*@param arrySFiC This parameter stores the array of the
SFiC(Sub-Function InfoChunk) objects
*@param req This parameter initializes the
RFiC(Requirement InfoChunk) object
*@param funct_link This parameter stores links between
the two function of LiCFLink type. It could be Data
flow or Control flow
*/
package org.obuda.infochunklib.rflp;
public class LiCF {
 String func_name, func_descrip, comm_name, func_input,
func_output; SFiC[] arrySFiC = null; ReqInfoChunk req =
null; LiCFLink func_link;
/**
*This constructor is used to initialize the LiCF
objects without information from the requirement layer
*@param name_func This parameter defines the name of a
function
*@param descrip_func This parameter defines the
description of a function
```

```
*@param name_comm This parameter stores the community
name of a function
*@param link This parameter stores links between the
two function of LiCFLink type. It could be Data flow or
Control flow
*@param input_func This parameter stores the inputs to
a LiCF object
*@param output_func This parameter stores the outputs
from a LiCF object
*@param subArry This parameter initialize the array of
the SFiC(Sub-Function InfoChunk) objects
*/
    public LiCF (String name_func, String descrip_func,
String name_comm, LiCFLink link, String input_func,
String output_func, SFiC[] subArry) {
func_name = name_func; func_descrip = descrip_func;
comm_name = name_comm; func_link = link; func_input =
input_func; func_output = output_func; arrySFiC =
subArry;}
/**
*This constructor is used to initialize the LiCF
objects with the information from the requirement layer
*@param spec_LiC This parameter stores the
specification of a LiCF object
*@param design_LiC This parameter stores the design of
a liCF object
*/
    public LiCF (String name_func, String descrip_func,
String name_comm, LiCFLink link, String input_func,
String output_func, SFiC[] subArry, String spec_LiC,
String design_LiC) {func_name = name_func; func_descrip
= descrip_func; comm_name = name_comm; func_link =
link; func_input = input_func; output_func =
func_output; arrySFiC = subArry; req = new ReqInfoChunk
(spec_LiC, design_LiC);
    }}
```

Further, the *LiCL* class is written to show the layer Info-Chunk object of the logical layer in the RFLP structure. Here, the array of CiC objects, LiCLGeometry object, LiCLSituation object, LiCLProcess object, LiCLDataModel object, LiCF object, LiCLConnector object, integer parameter, boolean parameter, string parameters, and the array of string parameters are used as an argument in the constructor. The constructor arguments of a LiCF object are populated from the Functional layer of the RFLP structure. The concepts of constructor overloading are used so that LiCL can accept various sets of the argument depends on the initialization of the object. The *CiC* class is written to show the component Info-Chunk object in the logical layer of the RFLP structure. The *LiCLGeometry* class

is written to show the geometry of the multidisciplinary product. Here, it could be possible for data retrieval of the product model and part model's STEP files in the LiCL class. In that case, LiCLGeometry constructor's arguments *part_info* and *assembly_info* are converted from string types into the STEP file format. Here, JSDAI API could be the possible approach to read and write the STEP file format. Then, LiCLGeometry object, affect zone and array of circumstances are used as a constructor argument for the LiCLSituation object. The *LiCLSituation* class is written to show the situation with a set of circumstances applicable to a LiCL object. The *LiCLProcess* class is written to show the process plane of the IC. It accepts String and Boolean values of processes as a constructor argument. The string values store the name of the processes whereas Boolean value stores the status of a process. The *LiCLConnector* is the enumeration class to store connector information. The get method returns the value of objects required to the main application. It is used in the next subsection.

```
// LiCL.java
/** This class is written to show the layer Info-Chunk
object of the logical layer in the RFLP structure*/
/**This class is written to show the layer Info-Chunk
object of the functional layer in the RFLP structure
*@param comp_name This parameter stores the name of a
component
*@param community_name This parameter stores the
community name of a component
*@param descrp_CiC This parameter stores the
description of a component
*@param contrib_product This parameter stores the
contribution of the component in the product modeling
*@param type_output This parameter stores the outputs
from the LiCL object
*@param type_input This parameter stores the inputs to
the LiCL object
*@param comp_connected This parameter stores the number
of connected components in a LiCL object
*@param connect This parameter stores information of
connector type. It could be Inner connector or Extended
connector
*@param components This parameter stores the array of
the CiC(Component InfoChunk) objects
*@param functionality This parameter stores the feature
of a LiCL object with LiCF type
*@param gmtry This parameter stores the geometry of the
components in a LiCL object
```

```
*@param situation This parameter stores the information
of influenced components and geometry of the components
in a LiCL object at the given situation
*@param process This parameter stores the process
involved in a LiCL object
*@param data_model This parameter stores the detail
description of a LICL object in the context of the
physical object such as process, geometry, and
situation
*/
package org.obuda.infochunklib.rflp;
public class LiCL {
private String comp_name, community_name, descrp_CiC,
contrib_product, type_input, type_output; int
comp_connected; LiCLConnector connect; CiC[] components
= null; LiCF functionality = null; LiCLGeometry gmtry =
null; LiCLSituation situation = null; LiCLProcess
process = null; LiCLDataModel data_model = null;
/**
*This constructor is used to initialize the LiCL object
without the information to the physical layer
*@param name_comp This parameter stores the name of a
component
*@param name_community This parameter stores the
community name of a component
*@param connected_comp This parameter stores the number
of connected components in a LiCL object
*@param product_contib This parameter stores the
contribution of the component in the product modeling
*@param input_type This parameter stores the inputs to
the LiCL object
*@param output_type This parameter stores the outputs
from the LiCL object
*@param affect_zone This parameter stores the
influenced components during the analysis in a LiCL
object
*@param part_info This parameter stores part
information in the geometry of a LiCL object
*@param assembly_info This parameter stores assembly
information in the geometry of a LiCL object
*@param form_features This parameter stores form
feature information in the geometry of a LiCL object
*@param circum This parameter stores the array of the
circumstance of a situation in a LiCL object
*@param arryCiC This parameter stores the array of the
CiC(Component InfoChunk) objects
```

```
*@param function This parameter stores the feature of a
LiCL object with LiCF type
*/
    public      LiCL(String      name_comp,      String
name_community,     int     connected_comp,     String
product_contib, String input_type, String output_type,
String   affect_zone,   String   part_info,   String
assembly_info, String form_features, String[] circum,
CiC[] arryCiC, LiCF function) {
comp_name = name_comp; comp_connected = connected_comp;
community_name  =  name_community;  contrib_product  =
product_contib; type_input = input_type; type_output =
output_type;   functionality   =   function;   connect
=connection;  components  =  arryCiC;  gmtry  =  new
LiCLGeometry(part_info, assembly_info, form_features);
situation  =  new  LiCLSituation(affect_zone,  circum,
gmtry);
    }
/**
*This constructor is used to initialize the LiCL object
with the information to the physical layer
*@param process_analysis This parameter stores the
status of the analysis process in a LiCL object
*@param process_effect This parameter stores the status
of the effect/contextual process in a LiCL object
*@param process_optimization This parameter stores the
status of the optimization process in a LiCL object
*@param value_analysis This parameter stores the array
of analysis process values in a LiCL object
*@param value_effect This parameter stores the array of
contextual process values in a LiCL object
*@param value_optimization This parameter stores the
array of optimization process values in a LiCL object
*@param connection This parameter stores information of
connector type. It could be Inner connector or Extended
connector
*@param contexual_PO This parameter stores knowledge of
contextual Physical object/s in a LiCL object
*@param connected_PO This parameter stores knowledge of
connected Physical object/s in a LiCL object
*/
    public      LiCL(String      name_comp,      String
name_community,     int     connected_comp,     String
product_contib, String input_type, String output_type,
String   affect_zone,   String   part_info,   String
assembly_info, String form_features, String[] circum,
Boolean   process_analysis,   Boolean   process_effect,
```

```
Boolean process_optimization, String[] value_analysis,
String[] value_effect, String[] value_optimization,
LiCLConnector connection, String contexual_PO, String
connected_PO, CiC[] arryCiC, LiCF function) {
comp_name=name_comp; community_name=name_community;
comp_connected = connected_comp; contrib_product =
product_contib; type_input = input_type; type_output =
output_type; functionality = function; connect
=connection; components = arryCiC;
gmtry = new LiCLGeometry(part_info, assembly_info,
form_features);
situation = new LiCLSituation(affect_zone, circum,
gmtry); process = new LiCLProcess(process_analysis,
process_effect, process_optimization, value_analysis,
value_effect, value_optimization); data_model = new
LiCLDataModel(contexual_PO,process,situation,connected_
PO, type_input, type_output);
    }

    private String getSituation() {
    return situation.affect_zone;}
    private String [] getCircumstances() {
    return situation.circumtnces ; }
    private CiC[] array_Components(){
    return components;}
    private LiCF getLiCF(){
    return functionality;}
    private LiCLProcess getProcessInfo(){
    return process;}}
```

6.2 Demonstration of Info-Chunk Objects in MAAD Structure

The *BiC* and *CxiC* class are the application classes for the behavior representation of the multidisciplinary product model as shown in the code below. The output is the graph between the components of various disciplines. It is the outcome of the process plane of the IC. The *BiC* class accepts the LiCL and LiCF objects as a constructor argument. Using the LiCL object, the LiCLProcess object can check the status of the analysis and optimization process. If the value is true, then it can generate the graph related to the process. The outcome of the Analysis process is shown in Fig. 7. The graph explains the displacement of the battery, starter and attenuator components w.r.t to time after the Thermal Analysis process. The outcome of the optimization process is shown in Fig. 8. The graph explains the voltage required w.r.t time for the optimized battery response.

```
//BiC Class
/**
This class is written to show the behavior Info-Chunk
object of the Behaviors layer in the MAAD structure
*@param bfunc This parameter initializes the LiCF
(Layer InfoChunk in the functional layer) object.
*@param blogic This parameter initializes the LiCL
(Layer InfoChunk in the logical layer) object.*/
public class BiC extends Application {
   LiCF bfunc = null;
   LiCL blogic = null;
/**
*This is the only constructor used to initialize the
BiC (Behavior InfoChunk) object with the information of
LiCF and LiCL object*/
   public BiC(LiCF funct, LiCL logic) {
   funct = bfunc;  logic = blogic; }

/*This method is used to generate the graph obtained
from the analysis and optimization process. The
parameters could be the components, parts or expected
changes in the assembly. */
  @Override
 public void start(Stage stage) {
if(blogic.getProcessInfo().isAnalysisProcess()){
//generate graph
}
if(blogic.getProcessInfo().isAnalysisProcess()){
//generate graph
}
if(blogic.getProcessInfo().isOptimizationProcess()){
//generate graph
}}
```

The *CxiC* class accepts LiCL object as a constructor argument. Using LiCL object, the LiCLProcess object can check the status of the effect (contextual) process. If the value is true, then it can generate the graph based on the contextual relationship between engineering objects. The outcome of Effect process is shown in Fig. 9 where contextual relation between attenuator and battery is explained by varying the battery output current with the attenuator speed. Here, XYChart class is used for generating the Line Chart graph.

```
//CxiC Class
/**
This class is written to show the Contexts Info-Chunk
object of the Behaviors layer in the MAAD structure
*@param blogic This parameter initialize the LiCL
(Layer InfoChunk in the logical layer) object */
public class CxiC extends Application {
   LiCL blogic = null;
/**
*This is the only constructor used to initialize the
CxiC (Contextual InfoChunk) object with the LiCL
object*/
  public CxiC(LiCL logic) {logic = blogic;}

/**This method is used to generate the graph obtained
from the contextual process. The parameters could be
the components, parts or expected changes in the
assembly. */
    @Override
    public void start (Stage stage) {
if(blogic.getProcessInfo().isEffectProcess()){
//generate graph}
}}
```

The *MAADStructure* class is the main method class that launches the application by calling the objects of BiC and CxiC objects.

```
//MAADStructure Class
/** This class is written to show the main application
of the InfoChunkLib API. */
public class MAADStructure {

 /** This method is used to start the main application
by calling the BiC and CxiC objects and generate the
graphs*/
    public static void main(String args[]) throws
Exception {
    Application.launch(BiC.class, args);
    new Thread(){
    Application.launch(CxiC.class, args);
    }}   }
```

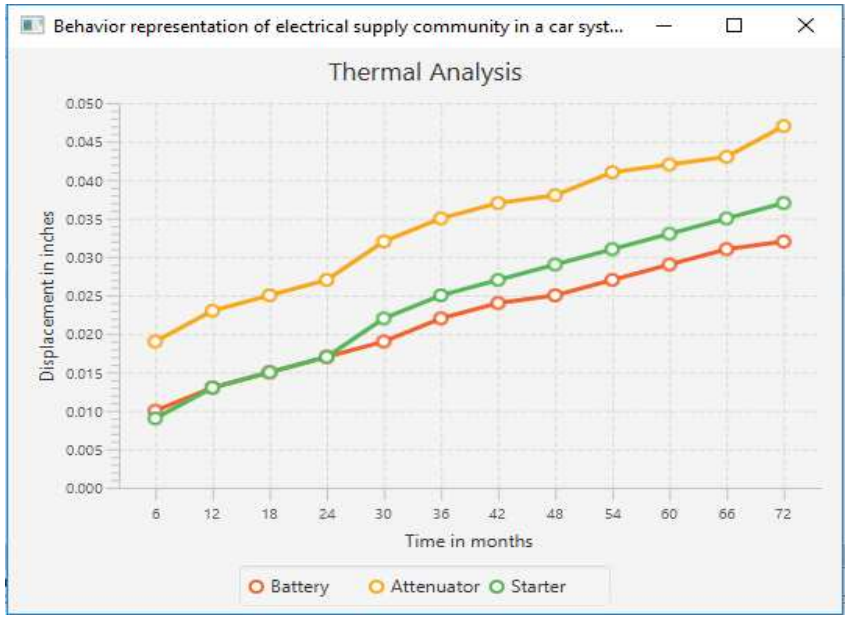


Figure 7
The graph of components after thermal analysis

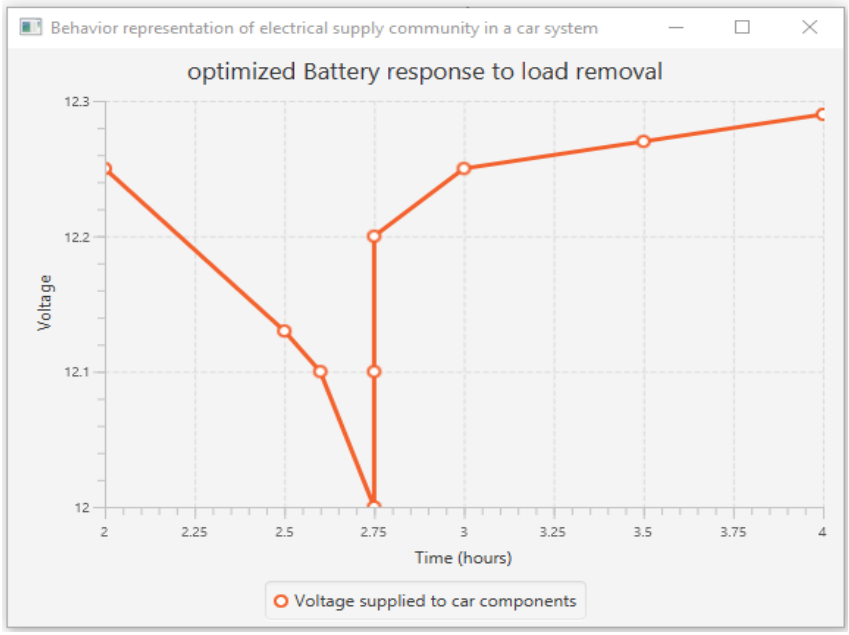


Figure 8
The graph of the battery component after the optimization process

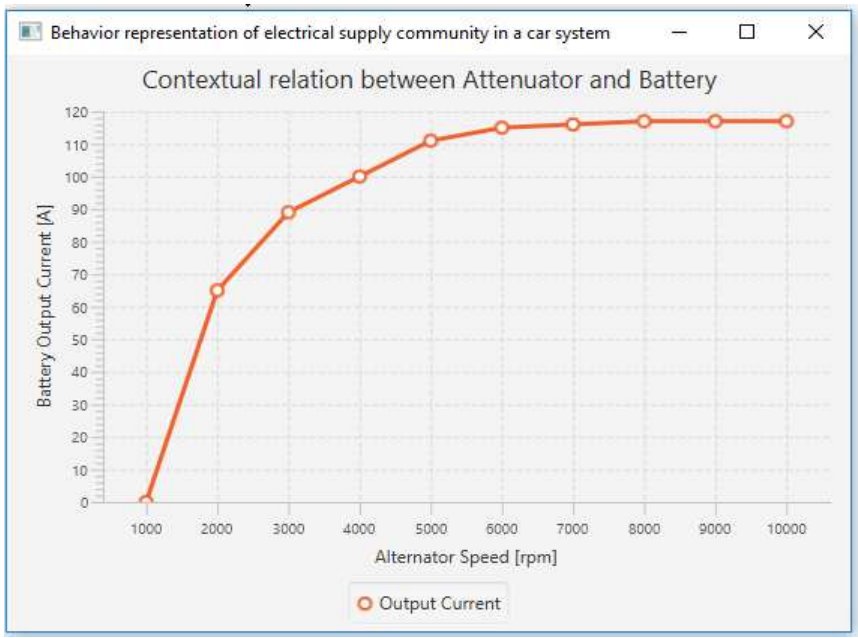


Figure 9
The graph between battery and attenuator component after effect/contextual process

6.3 Implementation of InfoChunkLib API in the Information Content

InformationContent class is the application which imports the InfoChunkLib API and handles the multidisciplinary product model. It could be a Java application or Web application. The output is stored in the database. As shown in the code below, SFiC objects and CiC objects are initialized first. Then, LiCF objects are initialized from SFiC objects and LiCL objects are initialized from CiC objects and LiCF objects. Then, BiC objects are initialized by LiCF objects and LiCL objects. CxiC objects are initialized by LiCL objects. Finally, the *MAADStructure* class is called by *InformationContent* arguments and graphs are generated for the behavior representation of multidisciplinary product model.

```
//InformationContent.java
/**This class is written to use InfoChunkLib API to
drive the multidisciplinary product model*/
import org.obuda.infochunklib.rflp.SFiC;
import org.obuda.infochunklib.rflp.SFiCLink;
import org.obuda.infochunklib.rflp.CiC;
```

```
import org.obuda.infochunklib.rflp. ConnectorCiC;
import org.obuda.infochunklib.rflp.LiCF;
import org.obuda.infochunklib.rflp. FunctionLink;
import org.obuda.infochunklib.rflp.LiCL;
import org.obuda.infochunklib.rflp.ConnectorLiC;
public class InformationContent{
public static void main(String args[]) throws
Exception{
//Extract SFiC object arguments information from the
functional layer and physical layer (.step file)
SFiC subfunc1 = new SFiC("To recharge the battery",
"Energize a field current that turns a rotor inside a
set of stators that can produce high current in
alternating directions", SubfunctionLink.DataFlow,
"Mechanical energy", "Electrical energy", "The
electrical system of a car is a closed circuit with an
independent power source the battery");

//Extract CiC object arguments information from the
logical layer and physical layer (.step file)}}
CiC comp1 = new CiC("Alternator", "Electrical supply",
"large BATT terminal connected to battery positive,
Relay Terminal connected to the connect to the dash
warning light, Sense Terminal connect the pigtail
directly to the BATT terminal", "provide power to the
car electrical system", subfunc1, "Magnet movement",
"Energy", "Battery", "Engine and Starter",
ConnectorCiC.Inner);
//Extract LiCF object arguments information from the
functional layer and physical layer (.step file)
LiCF funct = new LiCF("To power the car system", "The
battery provides juice to the starter. Then, the
alternator gives that battery the energy required to
power the car system", "Electrical Group",
FunctionLink.DataFlow, null, "Power", arrySFiC);
//Extract LiCL object arguments information from the
logical layer
LiCL logic = new LiCL("Electrical supply", "Electrical
Group", 3, "To supply power to Car system", "Mechanical
Energy", "Power", "Experienced_driver","Alternator,
Starter and Battery", assembly_info, null, true, false,
circumstnce, true, false, true, value_analysis_thermal,
null, value_optimization_global, ConnectorLiC.Extended,
"Lighting and signaling system", "Ignition electronic
system", arryCiC, func); }}
```

```
//Initialize BiC object and CxiC object from LiCL
object and LiCF object
BiC behav = new BiC(funct, logic);
CxiC context = new CxiC(logic);

//Call MAADStructure class for behavior representation
of multidisciplinary product model
String[] args = new String[0];
MAADStructure.main(args);}
```

6.4 Testing Phase of the Info-Chunk Objects

It is necessary to check the stored information in the Info-Chunk objects. In the OOP based language like Java, JUnit testing is a popular tool to check the behavior of an object. The behavior of a multidisciplinary product can be tested by varying the attributes and methods of the BiC and CxiC objects in the virtual environment. These values are compared with the values obtained from the physical environment. Further, formulas can be derived from the consistent values obtained from the virtual and physical environment.

Conclusion

This research work focuses on the behavior representation of a multidisciplinary product model by introducing Info-Chunk objects in the Information Content (IC). It started with the conversion of Info-Chunk entities into the Info-Chunk objects. Further, "InfoChunkLib" API is proposed based on the communication between the MAAD structure and RFLP structure in terms of LiC, BiC, and CxiC objects. Information Content (IC) is an application that imports InfoChunkLib API to handle and drive a multidisciplinary product model. The generated graph obtained from the IC evaluate the behavior of product components at various processes. The IC facilitate the HCI of multidisciplinary product model by initializing the parameters through the application. Info-Chunk objects provide necessary specification and knowledge representations to simulate the behavior of the complex multidisciplinary product model.

Future Work

A web server could be the next step for this research work, where a database is populated by the proposed API and a web application is used to access the IC. Dassault Systemes has implemented the RFLP structure in the CATIA V6 and 3DEXPERIENCE (3DXP) platforms for the multidisciplinary product model. Here, Dymola [16] is used to analyze the dynamic logical behavior and Modelica is used for logical and physical modeling of a product. The Java language and Modelica language are based on OOP concepts. Hence, API could be translated accordingly. The author could further update the API and database. Also, InfoChunkLib API can be extended and deployed in the IP.

Acknowledgment

The author gratefully acknowledges his supervisor, Dr. Horváth László, for guidance while writing this article.

References

[1] L. Horváth and I. J. Rudas: Towards the Information Content-driven Product Model, Proceedings of the IEEE International Conference on System of Systems Engineering, Singapore, 2-4 June 2008, pp. 1-6

[2] L. Horváth and I. J. Rudas: Systems engineering in product definition, Proceedings of the IEEE 13th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Slovakia, 22-24 Jan. 2015, pp. 181-186

[3] H. F. Krikorian: Introduction to object-oriented systems engineering.1, Journal of IT Professional, V(2), pp. 38-42, 2003

[4] L. Horváth and I. J. Rudas: Multilevel Abstraction Based Self Control Method for Industrial PLM Model, Proceedings of the IEEE International Conference on Industrial Technology, South Korea, 26 Feb.-1 March 2014, pp. 695-700

[5] L. Horváth and I. J. Rudas: Active Driving Content in RFLP Structured Product Model, Recent Advances on Mechanics, Materials, Mechanical Engineering, and Chemical Engineering, MMMCE, Barcelona, 2015, pp. 123-131

[6] Peter Fritzson: Principles of Object-Oriented Modeling and Simulation with Modelica 3.3: A Cyber-Physical Approach, Wiley-IEEE Press, John Wiley & Sons Inc, 2015, pp. 45-97

[7] Detterfelt, Jonas and Johansson, Gert: A UML Based Modeling Approach for Multi Domain System Products, Nordic Conference on Product Lifecycle Management - NordPLM, 2006, pp. 39-50

[8] John Stark: Product Lifecycle Management: 21st Century Paradigm for Product Realisation, Springer-Verlag, London, 2011, pp. 10-25

[9] Yatish Bathla: Conceptual Models of Information Content for Product Modeling, Acta Polytechnica Hungarica, XV (2), 2018, pp. 169-188

[10] Ou Y.: On Mapping Between UML and Entity-Relationship Model, The Unified Modeling Language, Schader M., Korthaus A. (eds), Springer Nature, Switzerland, 1998, pp. 45-57

[11] Bernhard Thalheim: Entity-Relationship Modeling: Foundations of Database Technology, Springer-Verlag, New York, 2000, pp. 124-145

[12] L. Horváth, J. Fodor, I. J. Rudas: Manufacturing Aspect of the IBCA Structure for Active Knowledge Content Representation in Product Model, Journal of IFAC- PapersOnLine, 48(3), 2015, pp. 1616-1621

[13] Yatish Bathla: Different types of process involved in the information content product model. In Proceedings of the IEEE 14th International Symposium on Intelligent Systems and Informatics (SISY), 2016, pp. 99-104

[14] L. Horváth and I. J. Rudas: Integrated Associative Modeling of Parts and their Machining Process by Features, Proceedings of the IEEE International Conference on Microelectronic Test Structures (ICMTS) conference, 2001, pp. 316-321

[15] Ian Sommerville: Software Engineering: 9th edition, Addison-Wesley, Pearson Education & Sons Inc, 2011, pp. 216-421

[16] Dassault Systemes AB: Dymola Dynamic Modeling Laboratory Getting started with Dymola. Dymola User Manual Volume 1, Dassault Syst emes, Lund, Sweden, 2013, pp. 23-48

[17] Yatish Bathla: Structured organization of Engineering Objects in the information content of the PLM system. In Proceedings of the IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI), 2016, pp. 473-478

[18] L. Horváth and I. J. Rudas: Behavior and Design Intent Based Product Modeling, Acta Polytechnica Hungarica, 1(2), 2004, pp. 17-34

[19] Yatish Bathla: Info-Chunk driven RFLP Structure based Product Model for Multidisciplinary Cyber Physical Systems, Proceedings of the IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), 2018, pp. 000327-000332

[20] L. Horváth and I. J. Rudas: Bringing up product model to thinking of engineer, Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2008, pp. 1355-1360

[21] László Horváth: New methods on the way to intelligent modeling in computer integrated engineering. In Proceedings of the 36th Annual Conference on IEEE Industrial Electronics Society (IECON), 2010, pp. 1359-1364

The Effect of Business Enabling Policies, Tax Treatment, Corruption and Political Connections on Business Climate

Gentjan Çera[1], Pavla Breckova[2], Edmond Çera[1], Zoltan Rozsa[3]

[1] Tomas Bata University in Zlín, Faculty of Management and Economics, Mostní 5139, 760 01 Zlín, Czech Republic. E-mail: cera@utb.cz, ecera@utb.cz

[2] University of Finance and Administration, Faculty of Economic Studies, Estonska 500, 101 00 Prague, Czech Republic. E-mail: pavla.breckova@vsfs.cz

[3] School of Economics and Management in Public Administration in Bratislava, Furdekova 16, 851 04 Bratislava, Slovakia. E-mail: zoltan.rozsa@vsemvs.sk

Abstract: The impact of the institutional environment on the business activity was a subject of several previous studies. However, the ways in which changes in institutions affect business climate have not received proper attention from scholars as of yet. The purpose of this paper is to fill this gap in the literature by examining the relationship between selected formal institutions (business enabling policies and tax treatment) and informal institutions (corruption and political connections) and business climate in the context of the developing country. To test the proposed hypotheses an ordinal regression with two link functions was applied on an original dataset of 404 firms operating in Albania. Results show that neither formal institutions, nor informal ones act as a block concerning the impact on the business climate. Tax treatment and political connections affected business climate negatively, whereas corruption seemed to have a positive impact. A positive but insignificant effect was found between business enabling policies and the business climate. Our research triggers interest of policymakers who intend to design policies to improve the business environment.

Keywords: business climate; business enabling policies; tax treatment; corruption; political connection

1 Introduction

Research problem. The state of activity (productive, unproductive or destructive) in the economy is determined by the institutional environment in which the activity is carried out [1]. This implies that the change of the institutional framework affects the entrepreneurial activity by influencing the business environment [2]–[4]. Therefore, as suggested in the literature [5], [6], certain

interconnection of institutions and the business climate is envisaged. However, ways in which changes in institutional environment affect the business climate have not received sufficient attention from scholars [7], [8]. In order to fill this gap in the literature, our research focuses on examining the relationship between formal and informal institutions and the business climate in the context of the transition economy, specifically Albania.

The contribution of the previous research, where the institutional theory was developed [9]–[11], drives us to create institutional factors affecting the business climate. An institution can be formal or informal, and it has the capability to influence entrepreneur's attitude or behavior by constraining or supporting his activity [9]. Formal institutions are considered crucial for the business activity – if they are stable and operating efficiently, they have a potential to reduce the business risk and uncertainty [12], [13]. Along with others, business enabling policies and tax administration or tax treatment are considered as formal institutions. On the other hand, informal institutions in the economy that are related to the legacy of the past, certain business practices and traditional social behavior may constrain entrepreneurial activity [14]–[16]. Corruption and unfair competition are typical informal institutions. Unfair and/or informal competition is related to political ties because entrepreneurs linked to politicians may benefit from avoiding law requirements [17], [18].

According to both official reports and previous studies, informal competition, corruption and tax administration were among the top business environment obstacles identified by Albanian entrepreneurs [19]–[21]. Entrepreneurial activity, carried out primarily by small and medium-sized enterprises (SMEs), makes an important contribution to the Albanian economic growth [22]. SMEs contribute more than 70% of value added and account for more than 80% of employment. Compared to the European Union (EU), SMEs crate an average value added of 57% and an average employment rate of 70% [23]. Given these figures, entrepreneurship support should be of a particular interest to Albanian policymakers. In addition, improving the business climate may also lead to attracting foreign direct investments and to developing a better functioning market economy, particularly in the Western Balkans [24], [25].

A reasonable level of governance leads to the strengthening of those formal institutions that enable and support entrepreneurial activity. At the same time, it leads to weakening those informal institutions that constrain it. Based on the eclectic theory of entrepreneurship, Verheul et al. [26] found that government influences the demand and supply sides of business activity. More recently, using the same theory, Thai and Turkina's study [27] concluded that the effect of governance is positive on formal institutions and negative on informal ones. The current paper aims to investigate these effects on business climate in the context of a transition and developing economy. Moreover, business climate is not affected by all formal institutions in the same way, nor by the informal ones. Another strand of literature suggests that the effect of institutions in transition and

emerging economies is in contrast to what is expected in developed countries. This aspect is more present especially in the case of informal institutions. For instance, corruption has a positive association with business growth [28] or with an innovative activity [29]. On the other hand, economic, institutional and political environments play an important role in the relationship between political connections and business performance [30].

Aim and motivation. This paper seeks to explore the relationship between business climate and selected formal and informal institutions in the context of a transition economy. For analysis purposes, business enabling policies and tax treatment are selected as formal institutions, and corruption and political connections as informal ones. In the course of conducting this study, significant evidence of this relationship being explored has not been found, in particular in the context of a transition economy. The results of this research may be of a particular interest to policymakers that intend to improve the business environment and to foster business start-ups. As Fereidouni and Masron [31] claim, from the point of view of the policymakers, it is very important to know which institutions matter the most for entrepreneurs and what their impact on business climate is.

Next part of this paper is dedicated to the literature review on the formal and informal institutions and developing the research hypotheses. Further part (no. 3) covers the issues related to the measurement of variables, the composite variables reliability test, the statistical method and the data collection technique. Analyzed results are presented in Section 4, and the Section 5 is dedicated to the discussion. At the end of the paper the concluding remarks are presented.

2 Literature Review

Structure of a country's institutional environment is made of formal and informal components [9], [11]. The institutions shape business environment and, consequently influence the business climate conditions in the economy.

Formal institutions are rules communicated through official channels, and consist of a regulatory framework and policy tools. They include the complexity and enforcement of the regulations in a country. Heavily regulated framework and unfriendly business policies may impede business start-ups and can discourage individuals from taking actions to become an entrepreneur [2], [32]. Therefore, formal institutions, such as business enabling policies and tax treatment affect business climate by stimulating or deterring entrepreneurial activity.

To encourage market entry and entrepreneurial activity, business enabling policies aimed at business environment improvement should be considered by policymakers [31], [33], [34]. Governmental interventions in the economy can lead to the improvement of business climate. Bjørnskov and Foss [35] argued that

the impact of entrepreneurship activities on productivity increases as the government becomes more active in the economy. Similar results were explored also by Fereidouni and Masron [31]. Likewise, Surfield and Reddy [36] found that business climate coincides with a lower rate of job loss. According to Blume [37], the local economic policies are associated with the business climate. Consequently, the firms' satisfaction with economic policies is associated with a set of factors related to the business environment. Nevertheless, other scholars argue that in a short-run policymakers cannot do much to change or reshape the industry profile in the country, whereas, in a long-run government involvement in public investments in education or infrastructure can affect the economy by shifting it from one set of industries to another [38]. Government can enable (designing policies) or constrain (through regulations) business start-ups and entrepreneurial activities [39]–[42]. The presence of good program aiming at assisting SMEs provided by government leads to a quality business environment [2]. Conversely, Xheneti and Bartlett [20] by performing principal component analysis and hierarchical linear regression, found that support-related impediments do not influence Albanian company growth. Similarly, Čadil et al. [43] studied the cohesion policy support for SMEs designed by European Commission in the context of the Czech Republic by applying a quasi-experimental research design and concluded by finding no impact of such policy on the value added and value added per labor cost of SMEs. Based on this discussion the following hypothesis can be proposed:

Hypothesis 1 (H1): Business climate is positively affected by business enabling policies.

Tax treatment can influence the business environment the firms operate in. Compared to high-income countries, tax administration is identified as a problem especially in middle-income countries [5], [44]. Similarly, in Central and South East European countries entrepreneurs perceive the level of taxes and, in particular tax administration, as one of the major obstacles for the business growth [28], [45], [46]. Changes in tax legislation and administration are among the most important impediments identified by Albanian entrepreneurs [20], [21]. Concerning the relation between taxation and entrepreneurship, the consensus is absent in empirical research [47]. Stallmann and Deller [48] found evidence that taxes limitations are associated with a poorer business climate and lower economic performance. Furthermore, Braunerhjelm and Eklund [49] examined the tax administration and found a negative relationship between firms' market entry and tax administrative burden. Following the Sobel's work [8], Chowdhury et al. [2] considered tax rates as a key formal institution determining entrepreneurship quality. They established a negative and significant relationship between them. Countries with cumbersome regulations have lower rates of business start-ups [50], [51] and do not stimulate a business growth [15]. For instance, complicated tax regulations might force business owners to hire external advisors to deal with

tax procedures and administration, which consequently raises their costs. Therefore, we hypothesize:

Hypothesis 2 (H2): Tax treatment has a negative effect on business climate.

Informal institutions are socially shared rules, usually not written that are communicated through unofficial channels [9]. They are deeply rooted values and norms which can influence individual behavior.

In transition economies, informal institutions are expected to be important drivers for business start-up and an entrepreneurial activity. The reason could be weak formal institutions originating from communist rule, and also inadequate institutional reforms during the transition period [17], [20]. As Belitski et al. [47] argued, a country characterized by an inadequate formal institutional environment may lead to additional pressure on informal institutions to shape organizational behavior.

Corruption is considered a classical informal institution [15] especially in transition economies [17]. It may transfer resources towards more corruptible activities because firms want to benefit from them [52]. Several researchers refer to corruption as an influential factor of business activity [31], [53], however, in the academic sources there is no consensus on the direction of its effect [29], [52], [54]. Grosanu and Bota-Avram [55] rated control of corruption as an important factor for the business environment, particularly for business start-up. Other studies have also found evidence that corruption hurts entrepreneurial activity [56], [57]. Dutta and Sobel [58] concluded that the corruption effects remain negative, but become smaller when business climate is not corruption-favorable. Nevertheless, another strand of the empirical research shows that corruption may help firm's market entry and entrepreneurial activity [16]. In the countries of South-Eastern Europe, corruption had a positive effect on business growth, whereas in the countries of Central-East Europe the opposite effect was observed [28]. Interpretation of such results could be related to deeply rooted social acceptance of corruption [59] in economies characterized by weak formal institutions. Furthermore, business owners from Western Balkan tend to justify corruption as "greasing the wheels" of business [60]. Based on this evidence in the context of a transition economy, the following hypothesis can be formulated:

Hypothesis 3 (H3): Business climate is positively affected by corruption.

Whether corruption is damaging or helping entrepreneurship can depend on the company's political connections. Political connections are other informal institutions that influence business activity. Linkages with politicians (at local or national level) can help business owners to facilitate transactions and gain benefits to improve their business [61]. In countries with weak institutions, especially in post-communist countries, entrepreneurs tend to engage in political activities [17], [18]. Such engagement leads to potential benefits that government officials may offer in the future, leading to informal competition. The informal sector

competitor's procedures are identified by enterprises as the main obstacle for doing business in some European countries and Central Asia [5]. This is in line with what researchers have documented in the Albanian business environment: business entities consider unfair competition as an obstacle [21]. In addition, due to their greater experience, social and potential political relationships, senior (older) entrepreneurs are more active in connection with government officials [20].

On the other hand, political connections are influenced by the prevailing institutional and political environment at a national level, by the business characteristics [62] and also by the economic environment [30]. Therefore, the political, institutional, and economic environments shape the relationship between political connections and business performance [30]. Companies that have ties with politicians might perform better [63], [64] and they also take a lower risk compared to the businesses with no political connections [52]. Amore and Bennedsen's [65] research results indicated that doing business with the public sector is an important channel for transferring rent to connected firms, which increases their profitability. Contrary to these empirical studies, however, some researchers have found the opposite: business performance is reduced by political connections [30]. Due to political instability and frequent changes in governmental officials, companies may be exposed to a risky and unstable political connections. As a result, the business climate is expected to be affected by political connections. Therefore, we assume the hypothesis:

Hypothesis 4 (H4): Political connections influence the business climate negatively.

3 Methods and Procedures

Unit of analysis. The unit of the analysis was a company. A face-to-face structured interview was performed with a member of each management team. As with Jolley, Lancaster and Gao [66], the owner, co-owner, financial manager, director, deputy director or manager was considered to be the appropriate person to represent the company's viewpoints.

Variable measurement. Tax treatment, corruption, political connections and business enabling policies were composed by the mean of a selected item set different per each variable. This type of variables creation has been commonly used in the literature (i.e. Batsakis [46]). The tax treatment represented by the following items: "Tax officials are competent and knowledgeable", "Tax officials are fair in their assessments and decisions", "Government is doing a good job in services offered to my business", and "There are many benefits for businesses that pay taxes". Political connections variable consisted in the item set: "Companies are involved in the local political activities", "Relationships between senior

government officials and some private sector entities include bribes or other benefits", and "Political favoritism impacts on business activity in the private sector." Corruption variable had a following item set: "In this business branch, it is common for companies to provide informal pays to get things done with regard to customs, taxes, licenses, regulation, etc.", "In this business branch, companies are familiar with the amount of informal payments to get things done", "Bribery and corruption remain an inevitable cost of doing business in my country", and "Bribery and corruption remain an inevitable cost of doing business in the Balkans." Business enabling policies variable was represented by these items: "Public investment in infrastructure has had a direct and positive impact on my firm's operations", "Public investment in the energy supply has had a direct and positive impact on my firm's operations", "Public investment in education has had a direct and positive impact on my firm's operations", and "Public investment in health services has had a direct and positive impact on my firm's operations". The items of tax treatment and corruption were formulated as a five-point Likert scale, 1 = "fully disagree" to 5 = "fully agree", whereas those of political connections and business enabling policies were in a form of four-point scale, 1 = "no, not at all" to 4 = "completely".

Business climate was measured by one question: how do you perceive current business climate / doing business conditions? Respondents were supposed to choose one of the three listed options: 1 = "unfavorable", 2 = "normal" and 3 = "favorable". This type of measurement makes business climate an ordinal variable, which limits the use of statistical methods.

Reliability test. Before computing the mean of item sets per each variable, the reliability of the scales was checked. Reliability test checks whether the measure reflects the construct that it is measuring or not. Table 1 shows the results of Cronbach's alpha, which is a test of reliability along with the mean and standard deviation per each item and composed variable. Considering DeVellis's [67] criteria, business enabling policies and corruption were respectable (between .70 and .80), whereas tax treatment and political connections were minimally acceptable (between .65 and .70).

Table 1

Cronbach's alpha per each composed variable and expected sign

Institution	Composed variables	Number of items	Mean	SD	α	Expected sign
Formal	Business enabling policies	4	1.86	.69	.750	+
	Tax treatment	4	3.37	.96	.677	–
Informal	Corruption	4	2.84	1.13	.749	+
	Political connections	3	2.55	.70	.675	–

Note: SD is standard deviation, α is Cronbach's alpha

The relation between business climate and our variables is shown in the Table 1. An improvement in business enabling policies and corruption may lead to a better

business climate, whereas, if tax treatment and political connections increase, business climate tends to become unfavorable for doing business generally.

Methods. To examine the effect of business enabling policies, tax treatment, political connections and corruption on business climate ordinal regression was employed. Ordinal regression is a statistical technique used to predict behavior of ordinal level dependent variables with a set of independent variables. As compared to multinomial logit model, it estimates one equation over all levels of the response variable. Dependent variable is the order response category variable and independent variable may be categorical or continuous. Our dependent variable was business climate, which was an ordinal variable (1 = "unfavorable", 2 = "normal" and 3 = "favorable"). There are five link functions that can be applied in an ordinal regression: logit, probit, log-log (also known as negative log-log), complementary log-log and cauchit [68, p. 362]. The link function is a transformation of the cumulative probabilities of the ordinal outcome to be used in the estimation of the model. In this paper, the logit and log-log are employed. They predict the probability of a certain level or category of dependent variable (γ) occurring with respect to the known values of the independent variables (X_i), and their equations are as follows:

Link name	Function	Inverse
Logit	$P(\gamma)=\frac{1}{1+e^{-(\beta_0+\beta_{1i}X_{1i})}}$	$\ln(\frac{\gamma}{1-\gamma})=\beta_0+\beta_{1i}X_{1i}$
Log-log	$P(\gamma)=e^{-e^{-(\beta_0+\beta_{1i}X_{1i})}}$	$-\ln(-\ln\gamma)=\beta_0+\beta_{1i}X_{1i}$

According to Norušis [69], logit should be applied if evenly distributed categories of the dependent variables are noticed, whereas log-log is recommended to be applied if lower categories are more likely. The analyses were computed by means of statistical package SPSS version 23. The SPSS Ordinal Regression procedure, known as PLUM (Polytomous Universal Model), was essential to generate the ordinal regression results. PLUM is an extension of the general linear model to ordinal categorical data [69].

Data and sample profile. In the survey, there were 404 businesses involved in Albania. Observations were conducted by IDRA Research and Consulting, a market research company based in the capital city. Data collection and quality control were completed in January 2017. For the distribution of the sample, the General Directorate of Taxation business database was used, and in order to ensure the representativeness of the results, the following criteria have been taken into account: county (12 counties), business size (number of employees) and business sector (manufacturing, service and trade). The questionnaire was a semi-adaptation of previous similar surveys by the International Labor Organization.

Table 2 introduces the final survey dataset and the results of the business climate in Albania. About 58% of the observed data was collected from the companies located in the capital, 23% from the south, 11% from the central and 8% from the northern part of the country. It corresponds with the real business distribution throughout Albania. In Tirana, the capital, there is a major part of the Albanian business located. The northern region has the lowest number of businesses compared to other regions, although it includes four different counties. That is due to the level of economic development and low population density in these areas. According to sectoral industries in the survey sample, there were about 17% manufacturing companies, 40% trade, 45% services and the rest of the examined sample represented another activity (6%).

When analyzing the data, most businesses see the business climate as "unfavorable". The highest score was recorded in the capital (63% "unfavorable"). Less than one in seven firms rated the business climate as favorable in Albania. Table 2 shows the results disaggregated by region and business activity.

Table 2

Sample profile and the distribution of business climate categories across regions and business activities

		Business climate			Total	
		Unfavorable	Normal	Favorable	*n*	%
Region	South	57%	29%	14%	92	23%
	North	50%	28%	22%	32	8%
	Central	50%	46%	4%	46	11%
	Capital city	63%	23%	14%	232	58%
Business activity	Manufacturing	61%	34%	4%	67	17%
	Trade	56%	28%	16%	162	40%
	Services	59%	25%	16%	149	37%
	Others	71%	21%	8%	24	6%
	Share	59%	28%	14%	402	100%

4 Results

The mean, standard deviation and number of observations by business climate categories for each analyzed variable are quoted in Table 3. The mean of tax treatment had a negative trend across the business climate (from 3.61 for *unfavorable* level to 2.98 for *favorable* level), whilst business enabling policies had a moderate positive trend. Based on these trends in the data, it was expected that ordinal regression would bring a positive sign between business climate and business enabling policies, and a negative sign with tax treatment. On the other hand, a negative trend was observed when business climate levels increase in cases of political connections, and the positive trend was marked in case of

corruption. Consequently, a negative association can be expected between business climate and political connections, and a positive one with corruption.

Table 3

Descriptive statistics of observations by business climate categories for each variable

Variable	Business climate											
	Unfavorable			Normal			Favorable			Total		
	Mean	SD	*n*	Mean	SD	*n*	Mean	SD	*n*	Mean	SD	*n*
Business enabling policies	1.81	.66	235	1.85	.65	111	2.12	.83	55	1.86	.69	401
Tax treatment	3.61	.98	236	3.04	.74	111	2.98	.96	55	3.37	.96	402
Corruption	2.62	1.16	230	3.06	1.06	110	3.32	.95	55	2.84	1.13	395
Political connections	2.72	.69	218	2.33	.58	101	2.21	.69	52	2.55	.70	371

Note: SD stands for standard deviation and *n* represents the number of observations

An ordinal regression analysis was performed to assess the prediction of affiliation with one of three outcome levels on the basis of four covariates. Our outcome variable was business climate (1 = "unfavorable", 2 = "normal" and 3 = "favorable") and the covariates used were tax treatment, political connections, corruption and business enabling policies, all scale measured. After deduction of 38 cases with missing values on our covariates, data of 366 companies remained suitable for analysis. Although the three levels of output were unevenly distributed, the logit link function was performed. A detailed look at frequencies of business climate categories (refer to Table 2) may lead to the selection of the negative log-log link function. Lower categories of business climate were more likely, thus the log-log link function was used.

Table 4

Model fit, goodness-of-fit and test of parallel lines for two types of ordinal regressions

Link function			-2 Log likelihood	Chi-Square	df	Sig.
Logit	Model fitting		627.903	62.837	4	.000
	Test of parallel lines		621.678	6.225	4	.183
	Goodness-of-fit	Pearson		713.948	710	.451
		Deviance		626.517	710	.989
Log-log	Model fitting		632.965	57.776	4	.000
	Test of parallel lines		627.098	5.866	4	.209
	Goodness-of-fit	Pearson		701.144	710	.586
		Deviance		631.578	710	.984

In Table 4, the summary of data for both conducted models is presented. Regarding the ordinal logistic regression (logit), the results indicate the overall model was statistically significant, $\chi^2(4, n = 366) = 62.837$, $p < .001$. Also, there was a good model fit (discrimination among levels) on the basis of our four covariates, $\chi^2(710, n = 366) = 626.517$, $p = .989$, using a deviance criterion. In addition, evidence showed no violation of the parallel lines assumptions, that state the slope coefficients in the model were the same across response categories (and

lines of the same slope were parallel), $\chi^2(4, n = 366) = 6.225$, $p = .183$ (referring the first block of Table 4). A violation of this test leads to the less restricted model usage, i.e. multinomial logit model [68]. As the ordinal logistic regression, similar results were found even in the case of log-log link function (second block of Table 4). The fitting of the model was statistically significant, $\chi^2(4, n = 366) = 57.776$, $p < .001$, and deviance criterion reported the good model fit based on our four covariates, $\chi^2(710, n = 366) = 631.578$, $p = .984$. Also, its test of parallel lines was not violated indicating that slope coefficients are the same among the dependent variable categories, $\chi^2(4, n = 366) = 5.866$, $p = .209$. Therefore, the results provided by both link functions are not misleading.

Table 5 contains a summary of parameter estimates for both link functions. To differentiate the dependent variable levels, the ordinal regression has an algorithm that calculates a continuous latent variable [68]. The thresholds [Business climate = 1] and [Business climate = 2] represent the response variable in the ordinal regression. The estimated threshold for [Business climate = 1] is the cutoff value between *unfavorable* and *normal* business climate levels and the threshold estimate for [Business climate = 2] represents the cutoff value between *normal* and *favorable* business climate levels. Thus, [Business climate = 1] is the estimated cutpoint on the latent variable used to differentiate *unfavorable* business climate from *normal* and *favorable* business climate levels, when all factors and covariates are zero. Subjects that had a value of -1.70 (logit vs log-log: -0.81) or less on the underlying latent variable that caused a rise in our dependent variable would be classified as *unfavorable*. In this line, [Business climate = 2] is the estimated cutpoint on the latent variable used to differentiate *unfavorable* and *normal* categories from *favorable* category of business climate, if values of all factors and covariates are zero. Subjects (entrepreneurs) with a value of -0.05 (logit vs log-log: 0.52) or greater on the underlying latent variable that rose our dependent variable would be classified as *favorable* business climate. Subjects with a value between -1.70 and -0.05 (logit vs log-log: -0.81 and 0.52) on the underlying latent variable would be classified as *normal* business climate. According to logit link function's output, only the threshold of [Business climate = 1] proved to be statistically significant.

There was found a statistically significant effect of the tax treatment, political connections and corruption on business climate in both link functions. Referring to the case when logit was applied as a link function, if an entrepreneur was to increase his perception in tax treatment score by one point, his ordered log-odds of being in a higher business climate category would decrease by 0.47, while the other variables in the model are held constant. Alternatively, an increase by one unit in tax treatment, the odds of the *unfavorable* and *normal* categories of business climate versus to the *favorable* category of business climate was 0.62 times greater, given that the other variables in the model are held constant, $p < .01$. Because of the proportional odds assumption (referring to the test of parallel lines in Table 4), the same increase, 0.62 times, is found between *unfavorable* business

climate and the combined categories of *normal* and *favorable* business climate. Like the logit case, the model where log-log was applied as link function reported a statistically significant negative association between business climate and tax treatment. Therefore, the evidence supported our expectation related to the sign of the relationship between tax treatment and business climate leading to the acceptance of H2. Despite the positive sign, business enabling policies were insignificant in the predicting the company's affiliation with one of the three levels of business climate, which implies the rejection of H1.

Table 5

Results of two types of ordinal regressions

Variable	Logit[a]				Log-log[b]		
	Estimate	OR	95% CI Lower	Upper	Estimate	95% CI Lower	Upper
[Business climate = 1]	-1.70 (.86)*				-0.81 (.64)		
[Business climate = 2]	-0.05 (.86)				0.52 (.65)		
Business enabling policies	0.17 (.16)	1.19	0.88	1.62	0.14 (.11)	-0.08	0.37
Tax treatment	-0.47 (.12)***	0.62	0.49	0.79	-0.34 (.09)***	-0.52	-0.16
Corruption	0.25 (.11)**	1.28	1.03	1.58	0.15 (.08)*	-0.02	0.31
Political connections	-0.60 (.18)***	0.55	0.39	0.79	-0.41 (.14)***	-0.68	-0.15

Note: *p < .10, **p < .05, ***p < .01, The numbers in parentheses are standard errors, CI is confidence interval, OR is odds ratio. a. R^2 = .158 (Cox & Snell), .186 (Nagelkerke), .091 (McFadden). b. R^2 = .146 (Cox & Snell), .172 (Nagelkerke), .083 (McFadden).

In contrast to tax treatment effect, corruption displayed a positive effect on business climate. When logit was employed as a link function, an increase by one unit of business' perception in corruption, the odds of *favorable* business climate versus the combined *unfavorable* and *neutral* categories were 1.28 times greater, with a 95% confidence interval between 1.03 and 1.58, $p < .05$. The positive association was reported even when log-log had been applied as a link function, but the significance was a bit weaker compared to the case of logit, $p < .10$. Therefore, we failed to reject the H3.

Contrary to corruption variable, the increase by one unit in business' perception in political ties, the odds of business climate being *favorable* compared to *unfavorable* and *normal* categories were 0.55 times greater, with a 95% confidence interval that laid between 0.39 and 0.79, $p < .01$. The significance and a negative association between political connections and business climate was found even in case when log-log had been used as link function. Thus, this evidence supports H4.

5 Discussion

Our findings show that the impact of formal institutions on the business climate is not the same within each institution. A positive relationship was identified between business climate and business enabling policies, which goes in line with previous research [2], [33], [34]. However, business enabling policies do not seem to be an important factor in determining the business climate, which is similar to what Čadil et al. [43], Xheneti and Bartlett [20] and EBRD [70] concluded in their studies. The reason should be explored in the quality of the implemented business policies. An adoption of a similar framework for corporate / enterprise policy formulation, which Arshed et al. [71] suggested should lead to better results. Likewise, Xheneti [22] offers a conceptual framework for exploring policy formulation, linking policy formulation and the intended policy outcomes [14]. Business policies designed to improve the business environment should encourage or motivate business start-up and entrepreneurial activity. Policymakers should therefore insist on creating a friendly business environment and a well-functioning educational system [33] that would increase the supply of educated entrepreneurs [44].

Concerning tax treatment, a negative impact has been identified, which has supported the results of previous studies [2], [49], [50]. As stated in the literature review, the cumbersome regulatory framework and frequent changes in tax procedures may lead to the discouraging individuals from engaging in the business start-up process. As Jolley et al. [66] emphasized that, in order to improve the country's economy, entrepreneurs may prefer a policy that aims at reducing taxes rather than tax incentives or tax administrations that are procedure-based.

As mentioned in the official report of the European Commission [23], considerable efforts have been made in Albania to encourage individuals to engage in entrepreneurship and improve the business environment. This effort consists of establishing an action plan for cooperation between the government, industry and universities, and work on creating a friendly business environment for business start-ups. Nevertheless, further reforms encouraging entrepreneurial activity are needed to tackle deep-rooted obstacles, such as infrastructure (especially roads and electricity), property registration and contract enforcement [70].

Similar to formal institutions, even the informal ones do not have the same impact on the business climate. Our findings have shown the positive impact of corruption on the business climate supporting the "grease the wheels" theory of the business [57], [72]. This contradicts what was found in developed countries characterized by strong formal institutions, where corruption acts as an additional tax [47]. Corruption acts as the 'grease the wheels' for entrepreneurial activity in emerging and transition economies where there is an institutional weakness [17], [28], [73]. In addition, entrepreneurs cannot operate independently of corruption in these countries [74]. As Goel et al. [75] claims that entrepreneurs might also be

involved in mutual corruption to counter law requirements. This could be seen as a result of operating in the environment consisting of both weak formal institutions and a weak entrepreneurial culture which lead to business owners being willing to avoid legal requirements or the attention of tax officials, and/or engage in bribery or corruption as a way of doing business.

Contrary to the main empirical literature, we have found that political connections have had a negative impact on business climate. This is in accordance with Jackowicz et al.'s [30] results. This should be related to the political instability characteristic for transition countries. Frequent changes of government officials may cause risky and unstable business connections with local or national politicians.

Conclusion

Scholars and policymakers consider entrepreneurship to be an important factor to stimulate the economic development, so many developed and developing countries have designed and implemented policies to support entrepreneurship [2], [32]. However, due to the differences in economic, institutional and political environments, the impact on business varies from country to country. This research focuses only on the influence of selected institutions on the business climate in the context of the transition economy.

The study based on institutional theory [9]–[11] seeks to examine and explore the relationship between selected institutions and business climate. This theory proves that the role of formal and informal institutions is very important for the business climate, especially for emerging and transition economies. Such institutions include business enabling policies, tax treatment, corruption and political connections. Compared to developed countries, improving the quality of institutions has a greater impact on the quality of business in developing countries [17]. We succeeded in answering the research question that appeared in the literature, regarding the institutional impact [2] on the business environment. Our study thus proves that neither formal institutions, nor informal ones act as a block concerning the effect on business climate.

Although our study has reached its aims, there are limitations in research. First, our findings are limited to one country, which might share the same conditions in terms of regional, economic, institutional and political environments with only limited number of countries. Therefore, our findings can be generalized only for developing and transition countries. Second, it is questionable to assume that the identified relationships could continue for infinite time and affect the business climate. Our results on tax treatment and political connections showed dubious effects that requires further investigation.

The study findings are beneficial for designing policies encouraging entrepreneurship and improving the business environment. That is why our results have been of a particular interest for policymakers, as the significant relationships

between formal and informal institutions and the business climate have been identified. Consequently, this study contributes to a better understanding of the institutional theory.

Acknowledgement

The authors are thankful to the Internal Grant Agency project of Faculty of Management Economics, Tomas Bata University, "The role of institutional environment in fostering entrepreneurship", for financial support to carry out this research, and to IDRA Research and Consulting for giving us access to their data.

References

[1] R. Douhan and M. Henrekson, "Entrepreneurship and second-best institutions: going beyond Baumol's typology," *J. Evol. Econ.*, Vol. 20, No. 4, pp. 629-643, Aug. 2010

[2] F. Chowdhury, D. B. Audretsch, and M. Belitski, "Institutions and Entrepreneurship Quality," *Entrep. Theory Pract.*, pp. 1-31, Sep. 2018

[3] T. S. Manolova, R. V. Eunni, and B. S. Gyoshev, "Institutional Environments for Entrepreneurship: Evidence from Emerging Economies in Eastern Europe," *Entrep. Theory Pract.*, Vol. 32, No. 1, pp. 203-218, Dec. 2007

[4] P. Stenholm, Z. J. Acs, and R. Wuebker, "Exploring country-level institutional arrangements on the rate and type of entrepreneurial activity," *J. Bus. Ventur.*, Vol. 28, No. 1, pp. 176-193, Jan. 2013

[5] J.-J. Dethier, M. Hirn, and S. Straub, "Explaining Enterprise Performance in Developing Countries with Business Climate Survey Data," *World Bank Res. Obs.*, Vol. 26, No. 2, pp. 258-309, Aug. 2011

[6] H. Ghura, X. Li, and A. Harraf, "Moderating relationship of institutions for opportunity entrepreneurship and economic development," *World J. Entrep. Manag. Sustain. Dev.*, Vol. 13, No. 4, pp. 350-374, Oct. 2017

[7] S. Dorado and M. J. Ventresca, "Crescive entrepreneurship in complex social problems: Institutional conditions for entrepreneurial engagement," *J. Bus. Ventur.*, Vol. 28, No. 1, pp. 69-82, Jan. 2013

[8] R. S. Sobel, "Testing Baumol: Institutional quality and the productivity of entrepreneurship," *J. Bus. Ventur.*, Vol. 23, No. 6, pp. 641-655, Nov. 2008

[9] D. C. North, *Institutions, institutional change, and economic performance*. Cambridge University Press, 1990

[10] W. J. Baumol, "Entrepreneurship: Productive, Unproductive, and Destructive," *J. Polit. Econ.*, Vol. 98, No. 5, Part 1, pp. 893-921, Oct. 1990

[11] O. E. Williamson, "The New Institutional Economics: Taking Stock,

Looking Ahead," *J. Econ. Lit.*, Vol. 38, No. 3, pp. 595-613, Sep. 2000

[12] F. Welter and D. Smallbone, "Institutional Perspectives on Entrepreneurial Behavior in Challenging Environments," *J. Small Bus. Manag.*, Vol. 49, No. 1, pp. 107-125, Jan. 2011

[13] D. Smallbone and F. Welter, "Entrepreneurship and institutional change in transition economies: The Commonwealth of Independent States, Central and Eastern Europe and China compared," *Entrep. Reg. Dev.*, Vol. 24, No. 3-4, pp. 215-233, Apr. 2012

[14] M. Xheneti and J. Kitching, "From Discourse to Implementation: Enterprise Policy Development in Postcommunist Albania," *Environ. Plan. C Gov. Policy*, Vol. 29, No. 6, pp. 1018-1036, Dec. 2011

[15] S. Estrin, J. Korosteleva, and T. Mickiewicz, "Which institutions encourage entrepreneurial growth aspirations?," *J. Bus. Ventur.*, Vol. 28, No. 4, pp. 564-580, Jul. 2013

[16] S. Aparicio, D. Urbano, and D. Audretsch, "Institutional factors, opportunity entrepreneurship and economic growth: Panel data evidence," *Technol. Forecast. Soc. Change*, Vol. 102, pp. 45-61, Jan. 2016

[17] B. A. Krasniqi and S. Desai, "Institutional drivers of high-growth firms: country-level evidence from 26 transition economies," *Small Bus. Econ.*, Vol. 47, No. 4, pp. 1075-1094, Dec. 2016

[18] T. Rajwani and T. A. Liedong, "Political activity and firm performance within nonmarket research: A review and international comparative assessment," *J. World Bus.*, Vol. 50, No. 2, pp. 273-283, Apr. 2015

[19] EBRD (European Bank for Reconstruction and Development), "The business environment in the transition region," London, 2017

[20] M. Xheneti and W. Bartlett, "Institutional constraints and SME growth in post- communist Albania," *J. Small Bus. Enterp. Dev.*, Vol. 19, No. 4, pp. 607-626, Oct. 2012

[21] A. Bitzenis and E. Nito, "Obstacles to entrepreneurship in a transition business environment: the case of Albania," *J. Small Bus. Enterp. Dev.*, Vol. 12, No. 4, pp. 564-578, Dec. 2005

[22] M. Xheneti, "Contexts of enterprise policy-making – an institutional perspective," *Entrep. Reg. Dev.*, Vol. 29, No. 3-4, pp. 317-339, Mar. 2017

[23] European Commission, "European Neighbourhood Policy and Enlargement Negotiations: 2017 SBA Fact Sheet Albania," Brussel, 2017

[24] R. Osmani, "Improved Business Climate and FDI in the Western Balkans," *J. Econ. Soc. Stud.*, Vol. 6, No. 1, pp. 5-23, 2016

[25] Z. Kittova and D. Steinhauser, "The International Economic Position of Western Balkan Countries in Light of their European Integration

Ambitions," *J. Compet.*, Vol. 10, No. 3, pp. 51-68, 2018

[26] I. Verheul, S. Wennekers, D. Audretsch, and R. Thurik, "An Eclectic Theory of Entrepreneurship: Policies, Institutions and Culture," in *Entrepreneurship: Determinants and Policy in a European-US Comparison*, Boston: Kluwer Academic Publishers, 2002, pp. 11-81

[27] M. T. T. Thai and E. Turkina, "Macro-level determinants of formal entrepreneurship versus informal entrepreneurship," *J. Bus. Ventur.*, Vol. 29, No. 4, pp. 490-510, Jul. 2014

[28] I. Hashi and B. A. Krasniqi, "Entrepreneurship and SME growth: evidence from advanced and laggard transition economies," *Int. J. Entrep. Behav. Res.*, Vol. 17, No. 5, pp. 456-487, Aug. 2011

[29] M. Tomaszewski, "Corruption - A Dark Side of Entrepreneurship. Corruption and Innovations," *Prague Econ. Pap.*, Vol. 27, No. 3, pp. 251-269, 2018

[30] K. Jackowicz, Ł. Kozłowski, and P. Mielcarz, "Political connections and operational performance of non-financial firms: New evidence from Poland," *Emerg. Mark. Rev.*, Vol. 20, pp. 109-135, Sep. 2014

[31] H. G. Fereidouni and T. A. Masron, "Governance Matters and Entrepreneurial Activities," *Thunderbird Int. Bus. Rev.*, Vol. 54, No. 5, pp. 701-712, Sep. 2012

[32] J. Belás, V. Demjan, J. Habánik, M. Hudáková, and J. Sipko, "The business environment of small and medium-sized enterprises in selected regions of the Czech Republic and Slovakia," *E+M Ekon. a Manag.*, Vol. 18, No. 1, pp. 95-110, Mar. 2015

[33] Z. Brixiova and B. Égert, "Entrepreneurship, institutions and skills in low-income countries," *Econ. Model.*, Vol. 67, pp. 381-391, Dec. 2017

[34] Z. Brixiová and B. Égert, "Business environment, start-ups, and productivity during transition," *Macroecon. Dyn.*, Vol. 16, No. S2, pp. 213-231, Sep. 2012

[35] C. Bjørnskov and N. Foss, "How Strategic Entrepreneurship and The Institutional Context Drive Economic Growth," *Strateg. Entrep. J.*, Vol. 7, No. 1, pp. 50-69, Mar. 2013

[36] C. J. Surfield and C. S. Reddy, "Mass layoffs, manufacturing and state business climates: Does state policy matter?," *Contemp. Econ. Policy*, Vol. 34, No. 4, pp. 630-645, Oct. 2016

[37] L. Blume, "Local economic policies as determinants of the local business climate: Empirical results from a cross-section analysis among East German municipalities," *Reg. Stud.*, Vol. 40, No. 4, pp. 321-333, Jun. 2006

[38] J. Kolko, D. Neumark, and M. C. Mejia, "What do business climate indexes teach us about state policy and economic growth?," *J. Reg. Sci.*, Vol. 53, No. 2, pp. 220-255, May 2013

[39] M. Xheneti and D. Smallbone, "The Role of Public Policy in Entrepreneurship Development in Post-Socialist Countries: A Comparison of Albania and Estonia," *EBS Rev.*, Vol. 24, pp. 23-36, 2008

[40] D. J. Cumming, L. Grilli, and S. Murtinu, "Governmental and independent venture capital investments in Europe: A firm-level performance analysis," *J. Corp. Financ.*, Vol. 42, pp. 439-459, Feb. 2017

[41] M. Cepel, A. Stasiukynas, A. Kotaskova, and J. Dvorsky, "Business environment quality index in the SME segment," *J. Compet.*, Vol. 10, No. 2, pp. 21-40, Jun. 2018

[42] A. Kljucnikov, J. Belas, L. Kozubikova, and P. Pasekova, "The Entreprenurial Perception of SME Business Environment Quality in the Czech Republic," *J. Compet.*, Vol. 8, No. 1, pp. 66-78, Mar. 2016

[43] J. Čadil, K. Mirošník, and J. Rehák, "The lack of short-term impact of cohesion policy on the competitiveness of SMEs," *Int. Small Bus. J. Res. Entrep.*, Vol. 35, No. 8, pp. 991-1009, Dec. 2017

[44] R. La Porta and A. Shleifer, "Informality and Development," *J. Econ. Perspect.*, Vol. 28, No. 3, pp. 109-126, Aug. 2014

[45] I. Hashi and J. Mladek, "Fiscal and Regulatory Impediments to the Entry of New Firms in Five Transition Economies," *J. East-West Bus.*, Vol. 6, No. 2, pp. 59-94, Jan. 2001

[46] G. K. Batsakis, "Impediments on the way to entrepreneurship. Some new evidence from the EU's post-socialist world," *J. Small Bus. Enterp. Dev.*, Vol. 21, No. 3, pp. 385-402, Aug. 2014

[47] M. Belitski, F. Chowdhury, and S. Desai, "Taxes, corruption, and entry," *Small Bus. Econ.*, Vol. 47, No. 1, pp. 201-216, Jun. 2016

[48] J. I. Stallmann and S. Deller, "State Tax and Expenditure Limitations, Business Climate, and Economic Performance," *Public Budg. Financ.*, Vol. 31, No. 4, pp. 109-135, Dec. 2011

[49] P. Braunerhjelm and J. E. Eklund, "Taxes, tax administrative burdens and new firm formation," *Kyklos*, Vol. 67, No. 1, pp. 1-11, Feb. 2014

[50] R. Aidis, S. Estrin, and T. M. Mickiewicz, "Size matters: entrepreneurial entry and government," *Small Bus. Econ.*, Vol. 39, No. 1, pp. 119-139, Jul. 2012

[51] I. Verheul, A. Van Stel, and R. Thurik, "Explaining female and male entrepreneurship at the country level," *Entrep. Reg. Dev.*, Vol. 18, No. 2, pp. 151-183, Mar. 2006

[52] C. J. Boudreaux, B. N. Nikolaev, and R. G. Holcombe, "Corruption and destructive entrepreneurship," *Small Bus. Econ.*, Vol. 51, No. 1, pp. 181-202, Jun. 2018

[53] V. Tonoyan, R. Strohmeyer, M. Habib, and M. Perlitz, "Corruption and Entrepreneurship: How Formal and Informal Institutions Shape Small Firm Behavior in Transition and Mature Market Economies," *Entrep. Theory Pract.*, Vol. 34, No. 5, pp. 803-831, Sep. 2010

[54] M. M. Khyareh, "Institutions and entrepreneurship: the mediating role of corruption," *World J. Entrep. Manag. Sustain. Dev.*, Vol. 13, No. 3, pp. 262-282, Jul. 2017

[55] A. Grosanu and C. Bota-Avram, "The influence of country-level governance on business environment and entrepreneurship: A global perspective," *Amfiteatru Econ.*, Vol. 17, No. 38, p. 60, 2015

[56] G. Dempster and J. Isaacs, "Entrepreneurship, corruption and economic freedom," *J. Entrep. Public Policy*, Vol. 6, No. 2, pp. 181-192, Aug. 2017

[57] A. Mohamadi, J. Peltonen, and J. Wincent, "Government efficiency and corruption: A country-level study with implications for entrepreneurship," *J. Bus. Ventur. Insights*, Vol. 8, pp. 50-55, Nov. 2017

[58] N. Dutta and R. Sobel, "Does corruption ever help entrepreneurship?," *Small Bus. Econ.*, Vol. 47, No. 1, pp. 179-199, Jun. 2016

[59] D. Traikova, T. S. Manolova, J. Mollers, and G. Buchenrieder, "Corruption perception and entrepreneurial intention in a transitional context - the case of rural Bulgaria," *J. Dev. Entrep.*, Vol. 22, No. 03, p. 1750018, Sep. 2017

[60] J. Budak and E. Rajh, "Corruption as an obstacle for doing business in the Western Balkans: A business sector perspective," *Int. Small Bus. J.*, Vol. 32, No. 2, pp. 140-157, Mar. 2014

[61] H. Guo, E. Xu, and M. Jacobs, "Managerial political ties and firm performance during institutional transitions: An analysis of mediating mechanisms," *J. Bus. Res.*, Vol. 67, No. 2, pp. 116-127, Feb. 2014

[62] N. Boubakri, O. Guedhami, D. Mishra, and W. Saffar, "Political connections and the cost of equity capital," *J. Corp. Financ.*, Vol. 18, No. 3, pp. 541-559, Jun. 2012

[63] S. Dicko, "Political connections, ownership structure and quality of governance," *Int. J. Manag. Financ.*, Vol. 13, No. 4, pp. 358-377, Aug. 2017

[64] J. S. Ang, D. K. Ding, and T. Y. Thong, "Political Connection and Firm Value," *Asian Dev. Rev.*, Vol. 30, No. 2, pp. 131-166, Sep. 2013

[65] M. D. Amore and M. Bennedsen, "The value of local political connections

in a low-corruption environment," *J. financ. econ.*, Vol. 110, No. 2, pp. 387-402, Nov. 2013

[66] G. J. Jolley, M. F. Lancaster, and J. Gao, "Tax Incentives and Business Climate: Executive Perceptions From Incented and Nonincented Firms," *Econ. Dev. Q.*, Vol. 29, No. 2, pp. 180-186, May 2015

[67] R. F. DeVellis, *Scale development: theory and applications*, 4th ed. SAGE Publications, 2017

[68] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, 2nd ed. Springer, 2015

[69] M. J. Norušis, *IBM SPSS statistics 19 advanced statistical procedures companion*. Prentice Hall, 2012

[70] EBRD (European Bank for Reconstruction and Development), "Country assessments: Albania," 2018

[71] N. Arshed, S. Carter, and C. Mason, "The ineffectiveness of entrepreneurship policy: is policy formulation to blame?," *Small Bus. Econ.*, Vol. 43, No. 3, pp. 639-659, Oct. 2014

[72] P.-G. Méon and L. Weill, "Is Corruption an Efficient Grease?," *World Dev.*, Vol. 38, No. 3, pp. 244-259, Mar. 2010

[73] A. Dreher and M. Gassebner, "Greasing the wheels? The impact of regulations and corruption on firm entry," *Public Choice*, Vol. 155, No. 3-4, pp. 413-432, Jun. 2013

[74] N. Williams and T. Vorley, "Fostering productive entrepreneurship in post-conflict economies: the importance of institutional alignment," *Entrep. Reg. Dev.*, Vol. 29, No. 5-6, pp. 444-466, May 2017

[75] R. K. Goel, J. Budak, and E. Rajh, "Private sector bribery and effectiveness of anti-corruption policies," *Appl. Econ. Lett.*, Vol. 22, No. 10, pp. 759-766, Jul. 2015

A Theoretical Approach to The Implementation of Low-Voltage Smart Switch Boards

Péter Holcsik[1], Judith Pálfi[1], Zsolt Čonka[2], Mihai Avornicului[3]

[1] Óbuda University, Kandó Kálmán Faculty of Electrical Engineering, Research Group of Applied Disciplines and Technologies in Energetics, 1034 Budapest, Bécsi út 96, Hungary, peter.holcsik@elmu.hu, palfi.judith@kvk.uni-obuda.hu

[2] Technical University of Košice, Department of Electric Power Engineering, Faculty of Electrical Engineering and Informatics, 040 01 Košice, Slovak Republic, zsolt.conka@tuke.sk

[3] Babeş–Bolyai University, Cluj-Napoca, Faculty of Economics and Business Administration, TeodorMihali street, Nr. 58–60. Campus UBBFSEGA 400591, Cluj-Napoca, Romania, mihai.avornicului@econ.ubbcluj.ro

Abstract: Technological advances have made possible the fault location detection on the low-voltage distribution network using the fault location determination algorithm (FLDa). The results obtained by operating this algorithm can be implemented into a system that schedules the faults toward the electrician teams in charge of the troubleshooting. This solution, however, only addresses the processing and evaluation of signals based on remote signaling and does not provide the possibility of automatic interventions. This present paper investigates and describes the possibilities of automatic interventions on low-voltage distribution networks. This paper examines the Smart Switchboard concept developed by the Research Group of Applied Disciplines and Technologies in Energetics.

Keywords: theory; low-voltage distribution network; smart switchboard

1 Introduction

The basic task of electricity supply is to ensure safe and continuous service of the electrical networks. “The joint fulfillment of the requirements of safety, quality, and economic efficiency is a task based on compromises that represent the central issue of system management” [1]. In parallel with increasing consumer demands, power suppliers have to maintain the quality of their services on an adequate level. If not, regulatory sanctions would be applicable.

"Electricity suppliers use several indicators for measuring the quality of electricity networks. The Hungarian Energy and Public Utility Regulatory Authority are following two indicators and expects their improvement by Hungarian electricity suppliers. These two indicators are the System Average Interruption Duration Index (*SAIDI*) and the System Average Interruption Frequency Index (*SAIFI*)" [2, 3].

The System Average Interruption Frequency Index (*SAIFI*) shows the number of unscheduled outages for a consumption site in a specific interval (usually yearly), i.e., "the frequency of unplanned supply interruption per consumer" [4].

The *SAIDI* network quality indicator is given by:

$$SAIDI = \frac{\sum_{i=1}^{n}(U_i \cdot N_i)}{N_T}[sec] \tag{1}$$

where N_i is the number of customers and U_i is the annual outage time for location *i,* and N_T is the total number of customers served.

In other words,

$$SAIDI = \frac{sum\ of\ all\ customer\ interruption\ durations}{total\ number\ of\ customers\ served} \tag{2}$$

The System Average Interruption Duration Index (*SAIDI*) shows the average number of outage minutes per consumer, i.e., "the average duration of unplanned supply interruptions" [4].

The *SAIFI* network quality indicator is given by:

$$SAIFI = \frac{\sum_{i=1}^{n}(\lambda_i \cdot N_i)}{N_T}[sec] \tag{3}$$

where λ_i is the failure rate, N_i is the number of customers for location *i* and N_T is the total number of customers served. In other words,

$$\mathrm{SAIFI} = \frac{\text{total number of customer interruptions}}{\text{total number of customers served}} \tag{4}$$

One of the *SAIDI, SAIFI* determinants of the network quality indicators is the number of consumers (N_i) affected by the malfunction *i*. Energy supplier companies keep records of the number of consumers affected by the failure of a particular equipment. Due to these records, they are able to produce accurate accounts about the number of consumers left without service due to a specific network component failure [5].

Quality indicators are usually calculated per year. Thus, the number of customers (N_T) used in the calculation of indicators is a constant value (number) determined for a specific year and a specific power supplier at the beginning of the year. This customary method is necessary in order to eliminate the effects of ongoing changes during the year and for carrying out uniform and transparent calculations [6].

The calculation method of the *SAIDI* index reflects that one of the most significant factors of the network quality indicator consists in the time interval of the failure, *i* which starts from its detection (security operation or the first consumer report) and ending with the restoration of the service at the consumption point (this does not necessarily mean the restoration of the normal functioning) [7].

Hungarian legislation on electricity providers defines short-term network failures for which the time of interruption and the number of affected consumers are not included in the calculation of the *SAIDI* and *SAIFI*. The short-term network failures are the consumer interruptions of a maximum duration of 3 minutes in the following situations:

1. normal operational interruptions not exceeding the restart duration of the network automatics;
2. inefficient restarts in rigidly earthed networks not leading to consumer interruptions;
3. efficient functioning of non-UPS switch automatics and of current transformer switchback automatics. [8]

The *SAIDI* and *SAIFI* network quality indices can be influenced differently by the operational management levels of the medium (MV) and high-voltage (HV) distribution networks. These effects will be detailed in the following chapters. In the paper after presenting the possibilities of the MV and HV the authors will then present the new possibilities that are available on the LV network. The authors demonstrate the theoretical demonstration of the effectiveness of the solution developed by the Research Group of Applied Disciplines and Technologies in Energetics [2, 3, 5, 8, 9, 14].

2 Breakdown Recovery of High-Voltage Transmission Networks

"European power systems are constructed hierarchically and can be divided into three distinct parts: 750, 400, 220, and 120 kV high-voltage transmission networks (HV), 35, 20, and 11 kV medium-voltage distribution networks (MV) and 0,4 kV low-voltage distribution networks (LV)" [9, 18]. The construction and operation of the HV, MV, and LV networks is significantly different from each other influencing the troubleshooting method.

High-voltage transmission networks are looped. Their theoretical operation schematics is represented in Figure 1 [10].

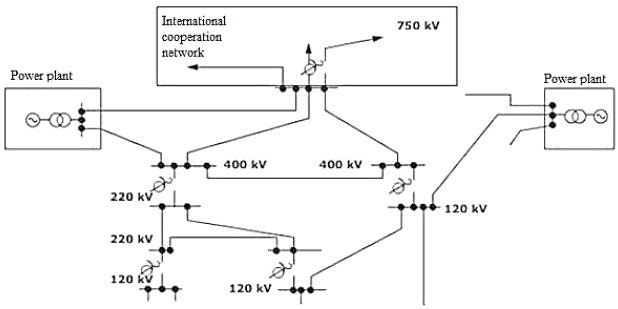


Figure 1

Theoretical schematics of a looped network [10]

The basic feature of the looped network is that there are various connections operating simultaneously in different directions between the various feeding and consumer points (Figure 1). The consumers joined to the looped network can be fed from many sides and through various routes. Hence, the looped network operates with maximum reliability. Another advantage is that multiple power routes (connectivity statuses) can be realized granting optimal power supply to the individual consumers (optimal operational parameters, minimum loss and low voltage drop) [10].

2.1 The *n-1* Principle

The looped design of the high voltage transmission network enables the realization of the *n-1* criteria. According to the *n-1* principle, the transmission system is constructed in such a way that the malfunction of element *1* of the system does not cause any loss at the consumer level (Figure 2).

Figure 2

The identified fault location of the HV operational failure not resulting in an outage [11]

Figure 2 shows the malfunction of a high-voltage network which did not result in a failure for the consumers. In areas where enhanced safety is required (for example in the vicinity of a nuclear power station) the compliance with the *n-2* criteria must be ensured [1].

2.2 Fault Recovery IT Support for the High-Voltage Transmission Network

The breakdown of high-voltage transmission networks can affect up to 100,000 individual consumers. Hence, along with the structural design, many other forms of assistance collectively named as IT support have been implemented in the operation of the HV networks.

IT support requires the constant transfer, storage and processing of large- and mostly real-time data. Online functions supporting the operational system control can be divided into two groups according to their complexity and use. These are the SCADA (Supervisory Control and Data Acquisition System) and the EMS (Energy Management System) functions.

Below are listed some of the typical functions of the SCADA system.

1. Reception of remote measurements and signals, e.g., real and reactive performance flows, busbar voltages, frequency measurements, breaker and disconnector position indicators, gear position of transformer regulators, etc.
2. Real time database creation with short refresh times, of usually a couple of seconds.
3. Representation, man-machine relation: the cyclically refreshed information usually appears on screens and on schema tables.
4. Registering and archiving.
5. Observing the limit values and gradients, recognizing endangered and dangerous states.
6. Topology analysis, inspection of the connection status and of the network continuity, registering changes, recognizing failures.
7. Issuing remote commands. The commands of the controlling personnel and the value settings calculated by the EMS and approved for dispatch are transmitted through the SCADA tele-mechanics system to the controlled objects [12].

Some of the typical EMS functions are:

1. automatic generation control (AGC),
2. load-flow or power flow,
3. real time sequence,
4. Model Update (MU),
5. State Estimation (SE)
6. Voltage Scheduler (VS), Automatic Voltage Control (AVC).
7. Operator Training Simulation (OTS). [12]

3 Breakdown Recovery of Medium-Voltage Distribution Networks

The operation of the medium-voltage distribution network is radial. However, the topology of their design is partially looped. Therefore, on the 10 kV urban cable network and on the so-called main line sections of the 20 kV and 35 kV overhead line networks, the electricity supply to consumers can be temporarily ensured through transfers without the correction of failures [9, 13]. This partially looped solution can be dubbed as a ringed or a curbed network, according to its design:

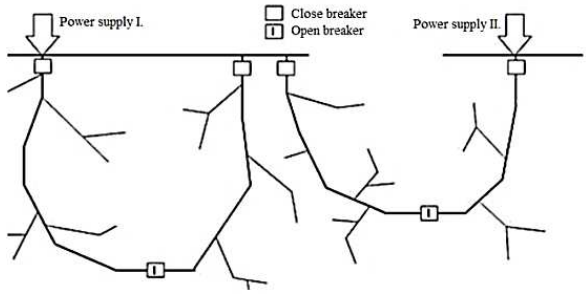


Figure 3
Ringed and curved networks [11]

Figure 3 shows a ringed- and a curved- medium voltage distribution network structure. The failure frequency on the medium-voltage distribution networks is higher than on the high-voltage networks. Figure 4 presents a failure of a medium-voltage distribution board.

Figure 4
Identified MV failure location causing an outage at a consumer connection point [11]

3.1 Remote-controlled Switches and Short-Circuit Detectors

The Hungarian power suppliers ELMŰ-ÉMÁSZ installed various remote signaling and remote controlled devices on the distribution networks in order to complete the delimitation of the failures and the speeding up of transfers, thus reducing the consumer disturbance and improving the *SAIDI* and *SAIFI* network quality index values.

Such a device is the remote controlled pole mounted disconnector (RPD) (Figure 5).

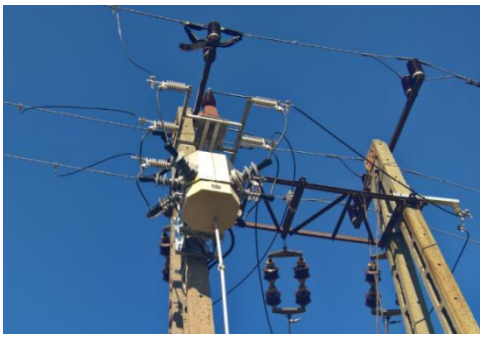


Figure 5
Remote controlled pole mounted disconnector on the ELMŰ-ÉMÁSZ network [11]

In addition to remote operation, the remote controlled pole mounted disconnector (Figure 5) provides information on the short-circuit currents and voltages flowing through it. Its application enables the automatic disengaging of shorted wires during the idle time of operation control through turning the pole mounted disconnector off.

Another device is the remote controlled switchgear on the distribution networks (RSD). This can ensure the possibility of remote operation by installing ex-post motors and current converters into the NERi, RM6 and similar devices (Figure 6).

Figure 6
RSD-ized NERi type device on the ELMŰ-ÉMÁSZ network [11]

Along with the remote signaling and operation devices, other devices with far lower investment needs are used today on medium voltage distribution networks enabling exclusively the identification of the failure location. The targeted positioning of these devices within the networks, e.g., at network junctions (Figure 7), can significantly facilitate and shorten the failure detection time consequently shortening the malfunction period.

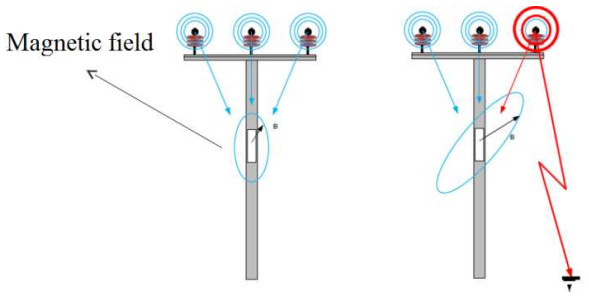


Figure 7
LineTroll R400D fault indicator (left: normal functioning, right: ground fault) [11]

According to the operational principle of the fault indicator (Figure 7), the fault perception of the device is based on the perception of the variation of the electromagnetic field below the line. This has to be installed 3 m below the middle line (Figure 7, encircled on the left-side figure).

Another solution for the same task is provided by the reinforced fault indicator pole (Figure 8):

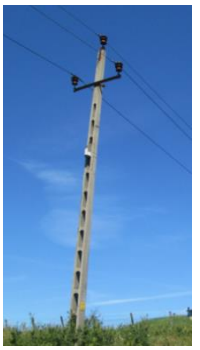


Figure 8
LineTroll R400D fault indicator installed on a reinforced concrete pole [11]

4 Developments Trends for Low-Voltage Distribution Networks

In contrast to, the partially looped topology of the medium-voltage distribution networks, the radial or tree-like topological characteristics of the low-voltage distribution networks do not enable the use of such temporary solutions as in the case of the medium-voltage distribution networks. Due to the high anticipated costs, such a development is not to be expected for the future, since, according to its definition, "the radial network consists of main lines fed from the supply point and their laterals, whose lines are not in contact either with one another or with line fed from other supply points" [8, 9]. Remote signaling and remote controlled devices are currently operating on low-voltage distribution networks only on a pilot basis. The significant number of *SAIDI* and *SAIFI* indicators stems precisely from the failures of low-voltage distribution networks, as given in Table 1.

Table 1
Unplanned consumer disturbance of ELMŰ-ÉMÁSZ Ltd. in 2017

	Number of disturbances	**Duration of outage (hours)**	**Number of affected consumers**	***SAIDI*** (minutes)	***SAIDI*** **[%]**	***SAIFI*** (minutes)	***SAIFI*** **[%]**
LV individual fault	26 500,0	65 339,6	26 500	0,0	1%	4,0	2%
LV medium fault	9 619,0	25 512,7	336 477	0,3	15%	45,0	27%
MV	1 783,0	5 238,1	1 965 311	1,8	83%	119,9	71%
HV	1,0	0,1	35 896	0,02	1%	0,1	0%
Sum:	37 903,0	96 090,4	2 364 184	2,2	100%	169,0	1,0

Due to the 15% *SAIDI* and the 27% *SAIFI* effect, the implementation of remote signaling devices on LV networks with lower investment costs is worth considering.

In the present paper we examine the automatic intervention possibilities as a further development of the LV distribution network operation. The basic idea stems from the application of this technology in Hungary since the 1980's for the fast and efficient handling of the temporary short-circuits in MV networks.

4.1 The Reclose Function

"When the reclose function is activated, the circuit breaker recloses after a previously specified time period following the defense action. If the defense continues to detect the short-circuit, the circuit breaker opens again. Following this and after another pre-set time period, the automatics will close the contacts of the circuit breaker again. The automatics seek to switch back two times (two cycles). The final release is activated if the short-circuit persists. Nowadays, reclose technology is already in use for MV networks. Its working principle and the pre-set time periods for MV networks are shown in Figure 9.

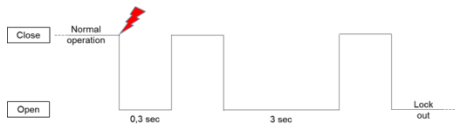


Figure 9

Working principle and pre-set time periods of the reclose function in case of MV networks

The implementation of the reclose function into the LV network will optimize the work of the electricians (they do not have to spend any more time going out to detect the faults in the network) and minimize the length of the LV power outages caused by short-circuits" [13].

4.2 Reclose Function on the LV Network

On LV networks, short-circuits are usually caused by external factors (e.g., tree branch touching the line, rain-related flashover, heavy wind, etc.) or by temporary overloads. Usually, the electrician sent to the location to resolve the outage only needs to change the fuse in phase 1, 2 or 3, depending on the number of phases affected by the event. In this case, no further mechanical or electrical interventions are required. The development of the reclose function of the Smart Switchboard for LV distribution networks enables the reduction of temporary short-circuits to failures causing at most 3 minutes of consumer outage.

The steps of the intervention are shown on the flow chart in Figure 10.

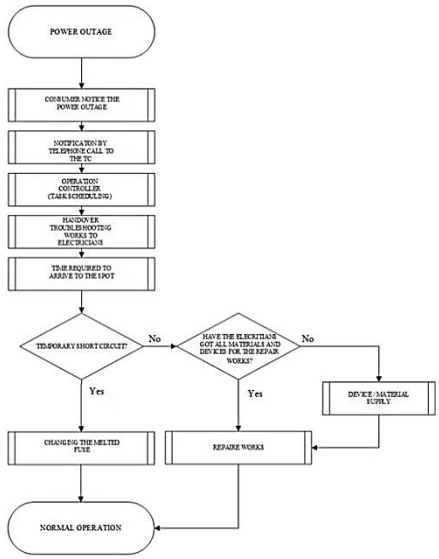


Figure 10

Flow chart of the current LV troubleshooting process [5]

Figure 10 presents the current method of troubleshooting from the occurrence of the fault until the restoring of the normal operation.

4.3 From the Smart Sensor to the SSB

Currently, the fault localization on the LV distribution networks is done manually [5]. The lack of an automated practice for fault localization on the LV distribution network is due to the lack of a failure detecting device on the LV residential consumer network. Nevertheless, for the sake of automation, researchers and suppliers have developed new plans and pilot programs [9]. Intelligent consumption meters and low-voltage distribution boards equipped with smart sensors (Smart Switch Board – SSB) are opening new perspectives and enabling modern solutions for the localization of the LV network malfunctions.

4.3.1 The Smart Sensors and the FLDa

The data of the smart sensors [9] installed on the LV distribution networks – from intelligent consumption meters to smart functions built into the distribution board – created the possibility for the development of an algorithm for the fault localization. The algorithm is capable of localizing the eventual faults. It was introduced into the scientific discourse as the fault location determination algorithm (FLDa). Running the algorithm gives a one-line fault message containing the individual ID of the defective device, its address (coordinates), and the percentage of the determination accuracy for the identified failure. The results of the algorithm can be used as input to the current fault scheduling dispatcher system (i.e. the LV fault-sheet scheduling system – the LFS) [9].

4.3.2 The Concept of the Smart Switchboard

"The Smart Switchboard (SSB) concept stands for a remote controlled LV switchboard which uses a circuit breaker for the dismantling of the short-circuit current. The detection of the short-circuit current is carried out by using detection equipment together with a corresponding measurement analysis system. It is suitable for remote switch-on (circuit breaker activation) which, if necessary, can be turned to clogging mode. It contains the possibility of visible interruption point and earthing functions as well.

The visible interruption point and the earthing functions are required for ensuring the life, health and safety protection during maintenance, reconstruction, etc. works. The remote monitoring functions could actively or passively monitor the current, the voltage and the performance of the LV system. The implementation of an automatic recloser, a so-called reclose function into the SSB is also possible" [14].

We have used the 2014 and 2015 data of ELMŰ-ÉMÁSZ Ltd. for the efficiency analysis of the SSB system, i.e., for determining the actual *SAIFI* and *SAIDI* savings. The results were published in our previous articles [5].

"By carrying out the study of the efficiency, the research group determined the number of costumers for which it is worth upgrading the existing equipment. A cumulative efficiency function has been developed. This function shows the *SAIDI* improvement which could be achieved when changing some of the fuses to SSB's.

For example, in case N=105 customers:

$$Cumulative\ utility = \sum_{N=105}^{N_{max}} (N_i \cdot U_i) \tag{5}$$

where N is the number of customers in the LV network behind the fuse, N_{max} is the maximum number of customers behind a fuse, N_i is the number of customers affected by the power outage i, U_i is the duration of the power outage i" [5].

The cumulative efficiency function is given in Figure 11.

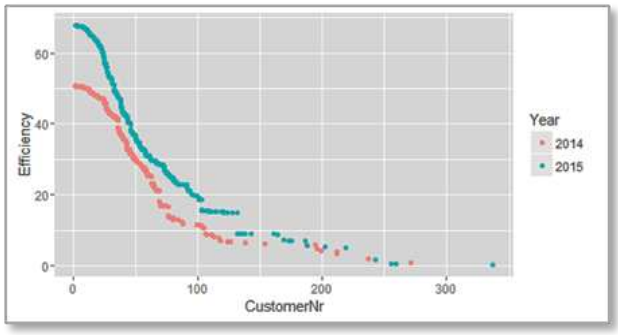


Figure 11
Efficiency in observations 2014/2015 [5]

The curve from Figure 11 regarding the measured data in year 2015 has been determined using regression analysis. To solve this problem the authors used the built-in polyfit function of the MATLAB software (fit type: poly34). The regression model was developed by using the 3rd order polynomial regression according to which:

$$CU(N) = 74.526 - 0.8611 \cdot N + 0.003317 \cdot N^2 - 0.000004095 \cdot N^3 \tag{6}$$

Hence, the present paper contains the theoretical approach for the previously published practical efficiency calculations.

4.4 The LV Fault-Sheet Scheduling System

The decisions of the dispatchers operating the low-voltage distribution network and the fault address allocations are supported by IT systems. These software are based on the LV fault-sheet scheduling system (LFS). The LFS is not an IT software but a system describing the processes of receipt, processing and allocation of the faults.

Power suppliers take notice of the outages of the residential electricity supply network from the fault reports of the consumers. Phone calls are registered at the error reporting center (TeleCentrum – TC). These registered fault addresses and the related information is received by the dispatchers through the IT system. The dispatchers allocate the reparation of faults to the trouble shooting electrician teams on the basis of the received information and taking into account the professional experience of the team members. [9]

The implementation of smart sensors with FLDa and LFS can significantly accelerate the complete lead time of the LV fault management systems [9, 19]. This solution, in this specific form, deals exclusively with the processing of signals based on the remote signaling and does not take into account the possibility of remote and automatic intervention. The expected investment costs of these devices are elevated. Hence, their system integration can be taken into account only after a significant quality improvement of the networks [20, 21, 22].

5 Implementation of the SSB in the LFS System

The LFS system supplemented with the SSB system processes is shown in Fig. 12.

According to Figure 12, the process is started by τ and ρ external excitations. The external excitation ρ might be any external factor influencing the number and the professional composition of the teams of electricians involved in the repair of the breakdowns. It can also consist in the expiry of the shifts or in the beginning of new shifts and standby periods as well as in tasks from new work management responsibilities, in capacity requirements from higher plant management levels, in human factors (e.g., sickness, electrician's vehicle breakdown), etc.

Excitation τ might be any event resulting in LV malfunctions causing more than one consumer outage, e.g. network overload, external influences, fallen trees, flashover, etc.

For new τ excitations, the fault is currently still transmitted to the work management systems by the telecentre that is expected to be replaced by the smart sensors [9]. The task of the work management system is to compile the *1 x m* sized $\overline{\overline{H}}$ matrix, extended with the k parameters on the basis of the incoming information.

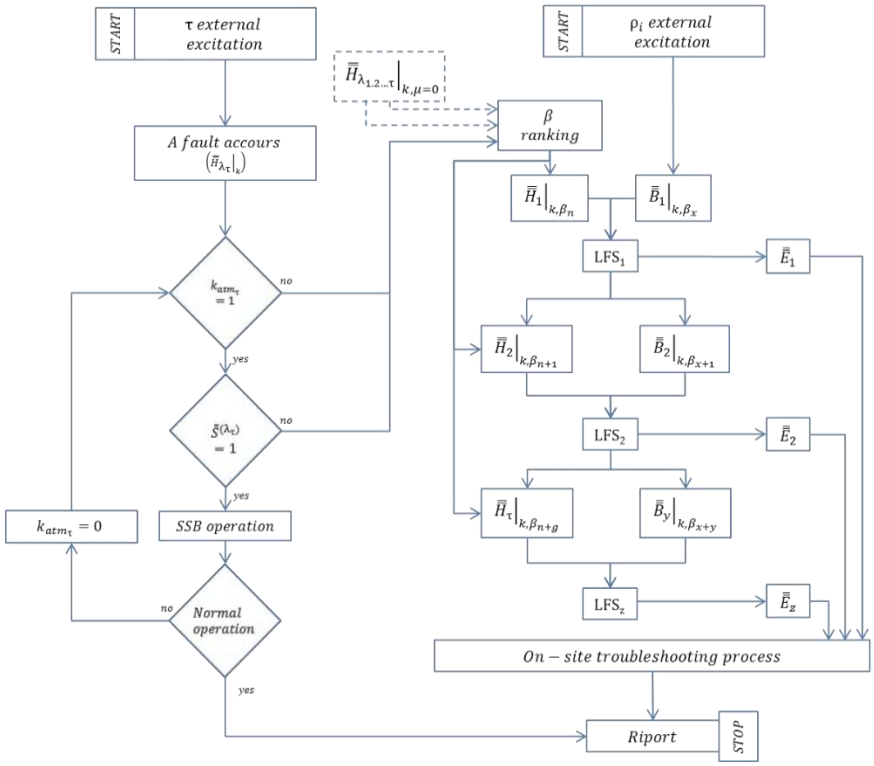


Figure 12
Implementation of the SSB in the LFS system

The *k* parameter, of which *m* are associated to the specific fault, might be the geographical location of the error (e.g., GPS coordinated), its priority or the number of affected consumers. The selection of these *k* parameters for setting up the system was based on the requirement of the faulty addresses to be allocated to the electrician trouble shooting teams at the optimal location and in possession of all the necessary tools, professional exams, etc. E.g., the $\overline{\overline{H}}$ matrix transposition of the λ^{th} incoming fault at τ location in the graph structure of the network (the transposition is for the purposes of transparent representation):

$$\overline{\overline{H}}^{T}_{\lambda_\tau} = \begin{bmatrix} Budapest \\ 1037 \\ Bécsi\ út\ 96/b \\ KDÜL \\ Gyenes \\ 42160/10 \\ bus\ consumer \\ 1\ fogyasztó \\ nem\ lakossági\ fogyasztó(k) \\ kábeles\ hálózat \\ prioritás: 3 \\ 47.5338768 \\ 19.0343905 \end{bmatrix} \tag{7}$$

$\overline{\overline{H}}_{\lambda_\tau}^T$ is the illustration of the matrix. Its content structure is given by the users, in our case the power supplier's decision. The values of the line in the $\overline{\overline{H}}_{\lambda_\tau}^T$ matrix:

$$\overline{\overline{H}}_{\lambda_\tau}^T = \begin{bmatrix} city \\ postcode \\ address \\ substation \\ name\ of\ the\ MV\ line \\ transformer\ ID \\ LV\ circuit\ ID/\ bus\ customers \\ number\ of\ the\ customers \\ type\ of\ the\ customers \\ network\ type \\ priority \\ north\ latitude\ GPS\ coordinate \\ east\ longitude\ GPS\ coordinate \end{bmatrix} \tag{8}$$

If the specific fault affecting the location τ is of a transitional character (for further details, see Chapter 4.2), i.e.,

$$k_{atm_\tau} = \begin{cases} 1,\ if\ \overline{\overline{H}}_{\lambda_\tau}\ is\ temporary\ faults \\ 0,\ if\ \overline{\overline{H}}_{\lambda_\tau}\ is\ not\ \ temporary\ faults \end{cases} \tag{9}$$

and there is an SSB at the specific τ fault location, i.e.,

$$\bar{\bar{S}}^{(\lambda_\tau)} = 1 \tag{10}$$

where

$$\bar{\bar{S}}^{(\lambda_\tau)} = \begin{cases} 1,\ if\ the\ node\ \ \lambda_\tau has\ SSB \\ 0, if\ the\ node\ \ \lambda_\tau nothas\ SSB \end{cases} \tag{11}$$

and

$$k_{atm_\tau} \cdot \bar{\bar{S}}^{(\lambda_\tau)} = \pi_\tau \tag{12}$$

where π_τ represents a variable that can take only the value 0 or 1.

In case of $\pi_\tau = 1$, the SSB becomes operational and attempts to eliminate the fault. If it succeeds in doing so, then it sends a report to the plant management. If the fault remains in spite of the operation, then $k_{atm_\tau} = 0$, where $\pi_\tau = 0$.

In case of $\pi_\tau = 0$, $\overline{\overline{H}}_{\lambda_\tau}\big|_k$ is incorporated into the β ranking process. The process will introduce the fault address into a $\tau\ x\ k$ sized $\overline{\overline{H}}_\tau\big|_{k,\,\beta_{n+g}}$ matrix on the basis of the evaluation of the k parameters. [17] The β values could stand for:

$\beta 1$: Danger of death and risk of accidents

$\beta 2$: High priority address (e.g., hospital)

$\beta 3$: Malfunction affecting a high number of customers

$\beta 4$: Malfunction affecting a low number of customers

$\beta 5$: Malfunction affecting a single customer

Hence, all faulty addresses received in the system that cannot be eliminated with the SSB, enters into a $\overline{\overline{H}}_{\tau}\big|_{k,\,\beta_{n+g}}$ matrix.

Electricians currently on shift or on call form the matrix $\overline{\overline{B}}_{y}\big|_{k,\,\beta_{x}}$. The faulty addresses of the system with the highest (numerically lowest) βranking, along with the k parameters associated with the addresses, will be the first to enter into the faulty address scheduling algorithm (LFS_1). The result of running the LFS_1 is entered into the $\overline{\overline{E}}_{1}$ matrix.

However, the $\overline{\overline{E}}_{1}$ results matrix contains not necessarily all the faults. It might happen that there will be k parameter needs that can be fulfilled only if one of the electrician teams has completed the fault correction at a given address. E.g., it is possible that there are three fault addresses where climbing the pole is needed, but there is only one electrician with the required authorization to perform this task. These addresses, for which the k parameter need cannot be immediately fulfilled, will be put into a waiting state (of course, these can also be de-scheduled on the basis of individual decisions).

In addition to the addresses forced into the waiting state the electricians who did not receive any address $\left(\overline{\overline{B}}_{1}\big|_{k,\,\beta_{x+1}}\right)$ as well as the addresses with the lowest β rankings $\left(\overline{\overline{H}}_{\tau}\big|_{k,\,\beta_{n+1}}\right)$ will remain in the system. The LFS_2 will complete its run with this input data and its result will be entered in the $\overline{\overline{E}}_{2}$ matrix.

This will go on for z cycles. z is reached when there is no more address left in the system. Then, the process will be put in standby mode until the next external excitation.

In the present case, the external excitation might be an electrician team becoming available or a new fault report. If a new external excitation occurs, the process restarts. That is to say, the LFS_z completes its run in the case of any $\overline{\overline{H}}_{\tau}\big|_{k,\,\beta_{n+g}}$ or $\overline{\overline{B}}_{1}\big|_{k,\,\beta_{x}}$ change, thus also facilitating the shortening of the entire system runtime.

The final step of the fault correction administered by the SSB and completed by the electricians is the reporting to the plant management centers.

The flow chart in Figure 13 shows with green color the processes rendered superfluous in the fault elimination by the installation of the SSB and by the activation of its reclose function at the temporary fault location.

An additional advantage to the benefit of decoupling consists in the fact that, if the function is activated on the MV network with the parameters set on an empirical basis, then the complete run-time of the system will be less than 3 minutes; i.e. they will be considered as short-term disturbances and will not be taken into account in the *SAIDI* and *SAIFI* calculations (see Chapter 1).

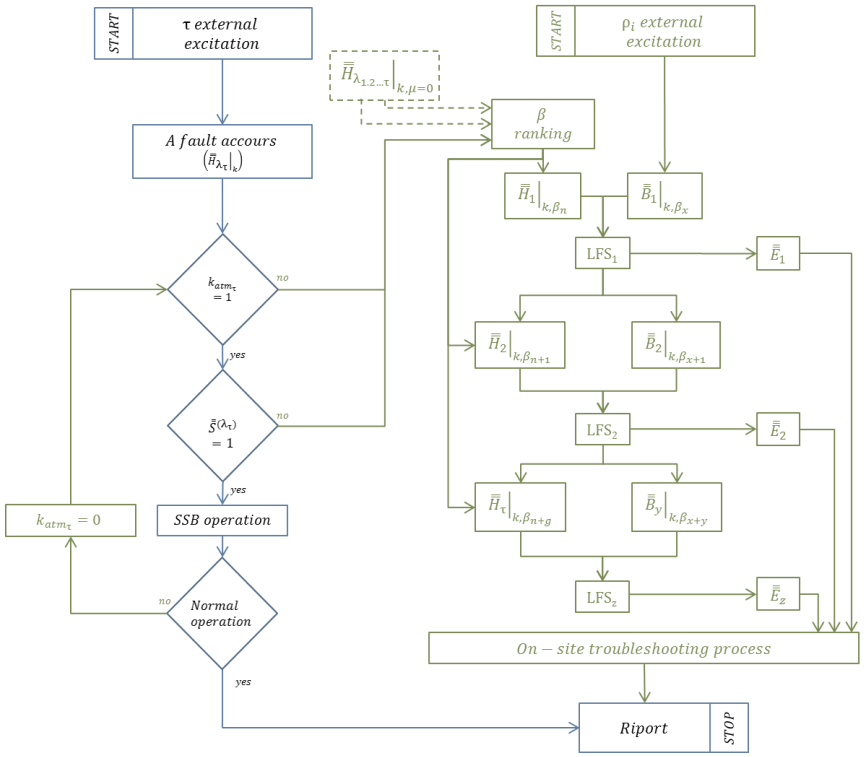


Figure 13

The effect of the SSB reclose function on the handling of temporary faults

Conclusions

This paper presented the smart switchboard devices and the theoretical advantages of their use in the LV distribution networks. This objective was derived from the existing technological developments and theoretical approaches. We have shown that the application of new results makes possible to be improve the network security on the network level through shortening the processes running time in case of faults due to temporary malfunctions.

Acknowledgements

The present paper was prepared at the Research Group of Applied Disciplines and Technologies in Energetics (AD&TE) at the University Óbuda and it was supported by the ELMŰ Network Ltd. and ÉMÁSZ Network Ltd.

EMBERI ERŐFORRÁSOK MINISZTÉRIUMA SUPPORTED BY THE ÚNKP-18-3-I-OE-88 NEW NATIONAL EXCELLENCE PROGRAM OF THE MINISTRY OF HUMAN CAPACITIES.

References

[1] F. Faludi, L. Szabó: Operation and control of the electric power system, Budapest University of Technology and Economics, BMEVIVEM265, https://vet.bme.hu/sites/default/files/tamop/vivem265/out/html/vivem265.html, 2012

[2] "1366-2012 - IEEE Guide for Electric Power Distribution Reliability Indices" Revision of IEEE Std. 1366-2003 (Revision of IEEE Std. 1366-1998), May 31 2012, DOI:10.1109/IEEESTD.2012.6209381

[3] K. Zou, W. W. Keerthipala, S. Perera: Saidi minimization of a remote distribution feeder, Australasian Universities Power Engineering Conference AUPEC, Perth, Australia, pp. 342-346, ISBN 9780646494883, 2007

[4] T Illés, G Lovics: Approximation of the Whole Pareto Optimal Set for the Vector Optimization Problem, Acta Polytechnica Hungarica, Vol. 15, No. 1, pp. 127-148, 2018

[5] J. Pálfi, M. Tompa, P. Holcsik: Analysis of the Efficiency of the Recloser Function of LV Smart Switchboards, Acta Polytechnica Hungarica, Vol. 14, No. 2, pp. 131-150, 2017

[6] G. Cordoş, M. Fülöp: Understanding audit reporting changes: introduction of Key Audit Matters, Accounting & Management Information Systems/Contabilitatesi Informatica de Gestiune, Vol. 14, No. 1

[7] P. Yashodhan Agalgaonkar, J. Donald Hammerstrom: Evaluation of Smart Grid Technologies Employed for System Reliability Improvement: Pacific Northwest Smart Grid Demonstration Experience, Power and Energy Technology Systems Journal IEEE, Vol. 4, pp. 24-31, ISSN 2332-7707, 2017

[8] E.ON Hungária Zrt.: Distributor Policy on the Rules for Cooperation on Access to the Distribution Network, 3. modification, 4. 12, Budapest, 2009

[9] J. Pálfi: Applying Big Data Methods to Power Supply, PhD thesis, Doctoral School of Applied Informatics and Applied Mathematics, Óbuda Unievrsity, Budapest, 2018

[10] Gy. Morva: Electric Power Engineering, Education College, Book of TÁMOP-4.1.2 A1 and TÁMOP-4.1.2 A2, 2012

[11] B. Czakó, G. Horváth: Operating systems and support systems, http://www.uni-miskolc.hu/~elkborzo/uzemiranyitas_Czako_Horvath.pdf,2017

[12] A. Faludi, L. Szabó: Operation and control of the electric power system, Budapest University of Technology and Economics, Book of TAMOP 4.2.5

[13] P. O. Geszti: Electric Power Systems I., Textbook Publishing Company, Budapest, ISBN:963-17-6987-9, 1983

[14] J. Pálfi, P. Holcsik: New Database and Theoretical Model for Power Distribution Networks, Proceedings of the 9th International Scientific Symposium on Electrical Power Engineering Elektroenergetika, pp. 539-544,

Technical University of Košice Faculty of Electrical Engineering and Informatics, 2017

[15] J. Pálfi., P. Holcsik, L. Pokorádi: Determination of Customer Number by Matrix Operations in Case of Network Failure, IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI 2018), Timisara, Romania, pp. 555-560, ISBN:978-1-5386-4639-7, 2018

[16] L. Pokorádi: Graph Theoretical Investigation of Network Structure System Scientific Bulletin Series C: Fascicle Mechanics, Tribology, Machine Manufacturing Technology, Vol. 2013, Issue 27, pp. 56-58, 2013

[17] M. Jocic, E. Pap., A. Szakál, D. Obradovic, Z. Konjovic: Managing Big Data Using Fuzzy Sets by Directed Graph Node Similarity, in Acta Polytechnica Hungarica, Vol. 14, No. 2, pp. 183-200, 2017

[18] T. Bagi: Measuring Current Distribution of Phase Conductor and Current Intesity Induced in Ground Wire on the 400 kV Transmission Line, Proceedings of the 9th International Scientific Symposium on Electrical Power Engineering Elektroenergetika, pp. 681-687, Technical University of Košice Faculty of Electrical Engineering and Informatics, 2017

[19] A. Dineva, A. R Várkonyi-Kóczy, V. Piuri, J. K. Tar, Point Cloud Processing with the Combination of Fuzzy Information Measure and Wavelets, Soft Computing Applications, Springer International Publishing, p. 584, 2018

[20] B. Niu, Y. Fan, H. Wang, L. Li, X. Wang, Novel bacterial foraging optimization with time-varying chemotaxis step, International Journal of Artificial Intelligence, Vol. 7, No. A11, pp. 257-273, 2011

[21] S.Vrkalovic, E. Lunca, I. Borlea, Model-free sliding mode and fuzzy controllers for reverse osmosis desalination plants, International Journal of Artificial Intelligence, Vol. 16, No. 2, pp. 208-222, 2018

[22] T. Haiddegger, L. Kovács, R-E. Precup, B. Benyó, Z. Benyó, S. Preitl, Simulation and control for telerobots in space medicine, Acta Astronautica, Vol. 181, No. 1, pp. 390-402, 2012

An Advanced Quick-Answering System Intended for the e-Government Service in the Republic of Serbia

Slobodan Nedeljković

Ministry of Interior Republic of Serbia, KnezaMiloša 101, 11000 Belgrade, Serbia, vojkan.nikolic@mup.gov.rs

Vojkan Nikolić

Ministry of Interior Republic of Serbia, KnezaMiloša 101, 11000 Belgrade, Serbia, vojkan.nikolic@mup.gov.rs &
Academy of Criminalistic and Police Studies, Cara Dušana 196, 11080 Zemun, Serbia, vojkan.nikolic@kpa.edu.rs

Milan Čabarkapa

School of Electrical Engineering, University of Belgrade, Bulevarkralja Aleksandra 73,11000 Belgrade, Serbia, ca.milan@etf.bg.ac.rs

Jelena Mišić

Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia, jelena.misic@kpa.edu.rs

Dragan Ranđelović

Academy of Criminalistic and Police Studies, Cara Dušana 196, 11080 Zemun, Serbia, dragan.randjelovic@kpa.edu.rs

Abstract: Many of the services incorporated in the e-Government of the Republic of Serbia need a quick-answer system to meet the continually increasing demands of the citizens for easy, fast and effective obtaining of the requested information. However, the public administration of the Republic of Serbia contains a significant amount of unstructured data

arranged in the documents. Thus, it is necessary to provide an automatic classification system based on the principle query-document. The question-answering (Q&A) system related to the Crime domain of the e-Government service of the Republic of Serbia represents a system for achieving the quick replies on citizens' questions. The Q&A system is based on the data mining, text mining, natural language processing, question answer, Bag of Words and N-gram analysis. A similarity measure (distance) is a significant parameter of the Q&A system due to its direct impact on searching speed and distance from wanted documents. Here, three most commonly used similarity measures are used: Cos, Jaccard and Euclid. The primary goal is to determine the similarity measure which provides the most precise results in the crime domain, and that similarity measure is used as a referent one. Due to the high importance of a similarity measure, we use the above three similarity measures, in the process of selecting the most appropriate similarity measure. The selection of the similarity measure is performed using the principles of redundancy and fault tolerance. Specifically, the principle of triple modular redundancy (TMR) with one voter is used. The proposed system is verified by the experiments with real citizen queries. The results show that the proposed system achieves good performance.

Keywords: e-Government; text mining; redundancy; TMR; unstructured documents

1 Introduction

To meet the increasing demand of citizens, for the easy, fast and effective obtaining of needed documents, it is necessary to equip the e-Government services with a quick-answer system. Public administration of the Republic of Serbia disposes of a large number of unstructured documents. These documents are mostly text type documents, so required answers are usually within these documents. To obtain an adequate reply it is necessary to provide an automatic mapping of relevant documents, i.e., an automatic classification strategy, a query – a relevant document. The mentioned strategy is incorporated by the Q&A system (Crime Domain) for e-Government services of the Republic of Serbia [3]. This system provides a quick reply on the asked question (query) achieving the faster and more effective searching and obtaining of wanted answers. The Q&A system is based on the principles of data mining, text mining, natural language processing, question answer, Bag of Words and N-gram analysis.

A similarity measure (a distance) is a crucial parameter in the Q&A system because it is directly correlated to the speed of answering and distance from wanted documents. Here, we analyze three most commonly used similarity measures: Cos, Jaccard and Euclid, with the aim to select the similarity measure which gives the most precise results in the Criminal domain, and that similarity measure is used as a reference one. Since the selection of the similarity measure is very important, our goal is to improve its selection. Namely, to increase the precision in getting a correct answer on the asked question, it is necessary to increase the similarity measure, which can be achieved if all three similarity measures are used in the selection of the most appropriate similarity measure.

Therefore, to achieve the mentioned goal, the principle of the redundancy and fault tolerance are employed. Since there are three similarity measure, i.e., three algorithms which provide the values of similarity measures as information which is further used in processing, such a redundancy represents the information redundancy. Thus, three values are obtained, but the Q&A system needs only one value, for the further operation. To determine which of the obtained similarity measures id the most appropriate one, the principle of the triple modular redundancy (TMR) with one voter is used, where the best result is determined by a voting logic.

The method which authors have proposed in this paper, is based on the example of one method given from another author [3]. It could be applied in the same way or in other methods.

The paper is organized as follows. In Section 2, the related works are presented. In Section 3, the basic theoretical principles of the redundancy and fault-tolerant systems are introduced, and a triple modular redundancy is presented. Section 4 offers the proposed Q&A system related to Crime domain and intended for the e-Government services of the Republic of Serbia is explained in detail, and its possibilities are listed. The proposed TMR system which determines the best similarity measure for a specific query (question) for the Q&A system is introduced in Section 5. The experimental results are given in Section 6. Lastly, the paper is concluded in Section 7.

2 Related Works

Considering the existing solutions for present e-Government problems, and analyzing and identifying the problems in the e-Government solutions in the Republic of Serbia, Šimić et al. [1] proposed a hybrid solution which represents a multi-layer e-document clustering based on a fuzzy concept and a usage of different measures of text similarity. The main aim was to reduce the response time of public administrations with a minimum civil clerks' involvement. To solve this issues, the authors introduced a new approach to facilitate the optimization and automation of advanced methods and techniques for information retrieving. This paper presents the ADVanced ANSwering Engine solution (ADVANSE) for wide-range e-Government services. The most important contributions of the ADVANSE project related to the e-Government services quality, quick response to the citizens' requests, and innovative use of the available content and restructured relationship between civil clerks and citizens. In particular, the accent was on response efficiency and flexibility. Namely, the authors focused on testing under different conditions and improving the ability of adaptation in the next research phases. One of the objectives of mentioned work was to find solutions for the functioning of such a system in multilingual environments and to increase the content complexity regarding the grammar and dictionaries of different languages regardless of the area of use. Consequently, different strategies were proposed.

A particular challenge was the functioning of e-Government services in different domains. Namely, different domains can use the special dictionaries, so it is necessary to use the specialized techniques to find a similarity. Besides, qualitative improvement of a given document processing is required; thus a possible solution can be the tagging of certain document parts instead of the entire document labelling.

Marovac et al. [2] analyzed texts written in different languages and according to different linguistic rules. The texts written in the Serbian language demand complex analysis because it can be written using has two alphabets, Cyrillic and Latin. Moreover, the Serbian language has very rich morphology. Therefore, the use of linguistic resources (corpus of contemporary Serbian language, morphological dictionaries, stop-words, a dictionary of abbreviations, etc.) intended for a qualitative analysis of a natural language, has become a considerable challenge.

The use of N-gram analysis achieved significant results without using the extensive lexical content or analyzing the texts written in Serbian without using the morphological vocabulary. Recently, special attention has been paid to the algorithm for keywords extraction (the N-gram) which is explained in detail in the following sections.

The Authors are of the opinion that an algorithm should be developed, such that, to cluster keywords according to their frequencies, in the text, text parts or clustering keywords (the N-gram), by separating the keywords that are frequent from the less frequent ones.

In V. Nikolić et al. [3], an approach for e-Government services intended for an automatic finding of the required answers or documents on online citizen's requests is presented. The authors described a method to overcome the problems caused by natural language processing tasks in the Serbian language and introduced the sentiment analysis as a special tool for text classification. The document pre-processing presented in this article included changing of a document format, removing the redundant and informal character, and structuring of the documents by the corresponding rules of the next step where the normalization is conducted. The used text normalization is based on the transformation of text into another form suitable for the computer processing. The results achieved by the proposed approach are very satisfactory.

The web-based framework for searching the Web content written in Serbian language, named the SEFRA, was proposed in M. Jovanović et al. [4]. The proposed SEFRA represents a hybrid solution that serves as a platform for a new search application or is used as a service for already existing applications. The SEFRA solves the indexing, searching, and displaying of search results which are adjusted to Serbian. Besides, these framework merges several web-based technologies and services for improving the e-Government citizen's services and other public-sector services. Moreover, SEFRA can be used in the administration of private companies solving the specific searching problems. Although the

SEFRA was developed for the Serbian language primarily, it can also be used for any other language containing the language morphology service. Furthermore, the SEFRA was optimized from both backend and front-end web perspective. The application of the SEFRA was validated by searching the crime law documents of Serbia, and good results were achieved.

In V. Nikolić et al. [5], the problems with large amounts of data, due to numerous implemented e-Government services, in the Serbian government was explained and a suitable solution was presented. The main problem is the specific data and information extraction from the variety of existing text documents which are usually in a format prepared for print (HTML, PDF and Microsoft Word formats). As a solution for that problem authors proposed an application that includes Lucene library, which is a specialized library for implementation of the indexing and searching over a significant amount of data. The proposed application provides a quick search within the unstructured text documents written in the Serbian language which further leads to an efficient detection and processing of criminal offenses and increases the security level of the Republic of Serbia. The proposed application is verified by searching the data and documents within the unstructured crime documents written in the Serbian language that aims to find the elements of a crime in the cyberspace. The obtained results showed that the proposed application accelerated the searching procedure significantly.

When it comes to the Serbian language in literature, except the papers of co-authors of this paper [1] [3] [4], very rare attempts are made to construct and describe the System for information retrieval i.e. Q&A system. One of these is the building of the Information Retrieval System for Serbian - Challenges and Solutions by M. Martinović et al. [6], using Natural language processing (NLP) technology as an area of computer science and artificial intelligence. In this article techniques, achieving state-of-the-art results in many natural language tasks, for example in language modelling, parsing, and many others implemented in the Serbian Information Retrieval System (SPRETS).

In article N. Milošević [7], one realized application was presented, which is, in fact, Stemmer for Serbian language. In this article is presented suffix-stripping stemmer for Serbian language, one of the highly inflectional languages. Stemming application is designed as a web application. It uses PHP script for backend and AJAX interaction with backend side.

The general effects of data redundancy have been noticed by many researchers in the field of Q&A systems, for example Lin in [8]. The redundancy-based approach has been developed as an alternative to the traditional ontology-driven knowledge-based techniques. This approach have basis in the philosophy of “data is all that matters” i.e., just give enough data and system could simply count instances and derive answers from these observations. This approach which solved problems in language processing was demonstrated in a paper written by Banko and Brill [9].

In Light et al. [10] is discussed a correlation between the number of times an answer appeared in the Text Retrieval Conference (TREC) corpus and the average performance of TREC systems on that particular question. That is, systems preferred to perform better on questions that had multiple answer instances within the corpus. Also, Clarke et al. [11] pointed an upward trend in Q&A system accuracy as a system was given even more text from which to extract answers and thereby holding everything else constant.

From standpoint of significance of using basic principles of TMR to constrict the proposed algorithm for optimization, one Q&A system and especially it's part named, voting system, it is important to notice different approaches of its implementation, like for example, using genetic algorithm [12], Bayesian techniques [13], Markov modeling [14], neural networks [15] etc.

In [16] Gruzenkin et al. Considered one compensation model of multi-attribute decision making and its application to N-version software choice.

3 Theoretical Background

With the development of IT technology, complex and sensitive processes have become automated which has a high demand for proper operation of the implemented system. This is provided by using the VLSI technology that enables practical adoption of many methods for mitigation and reduction of system faults. The fault-tolerant systems represent the high-confidential systems which even in unfavorable conditions operates in a proper way providing all functionalities due to the ability to tolerate single faults. These kinds of systems have the advantages of high reliability, availability, security, stability, manageability and serviceability [17, 18, 19, 20, 21].

In the fault-tolerant systems, one of the common methods to enhance system reliability is to use a redundancy, which represents the addition of resources, (amount of information and time) to normalize system operation. The redundancy can be related to the hardware, software, information and time.

As a basis for the considered Q & A system in [3], Text retrieved. Text retrieval is a branch of information retrieval where the information is stored primarily in the form of text. In the concrete case, it is about textual information relating to the members of the Criminal Code of the Republic of Serbia and they are written in Cyrillic and Latin scripts.

In order to optimize the existing Q&A system: Q&A system for e-Government services in the Republic of Serbia, it can be used as a basic idea, the idea TMR which uses three functionally equivalent units to provide redundant backup, regardless of type of redundancy i.e. whether it is viewed as an active hardware or corresponding to its three programming software redundancy or three different

sources of information; especially possibly application of different approaches in voting system as its obligatory part. This idea through presentation of basic content of most important titles necessary for understanding this idea authors have done in this paper, as described in the following sub-sections.

3.1 Hardware Redundancy

Hardware redundancy represents the addition of additional hardware to normalize system operation. This type of redundancy is usually employed when it is necessary to detect system failures or tolerate them to make the system robust to the failures. Hardware redundancy can be active, passive and hybrid. [22]

Passive hardware redundancy denotes the technique of fault masking with the aim to prevent a fault to cause errors.

Active hardware redundancy denotes the adoption of techniques that enhance fault toleration, by detecting the fault and repairing of faulty (inoperative) hardware. It is based on fault detection and localization within the system and its repairing.

Hybrid hardware redundancy denotes the combination of the previous two redundancies by via their advantages. The technique of fault masking is used to prevent errors, and when this technique does not achieve good results, one of the active techniques is used to locate and eliminate the fault.

3.1.1 Passive Hardware Redundancy

In the implementation of passive hardware redundancy, a voting mechanism such as the principle of majority voting is used to mask a fault. In passive hardware redundancy, the techniques that provide fault tolerance without a need to detect and repair the fault are used.

One of the most commonly used passive hardware redundancies is a triple modular redundancy (TMR) whose basic principle is presented in Fig. 3.1, wherein, it can be seen that quantity in the used hardware is tripled, and the principle of majority voting is employed. Namely, if one module stops working properly, the rest two modules will mask its fault mitigating the errors. [23]

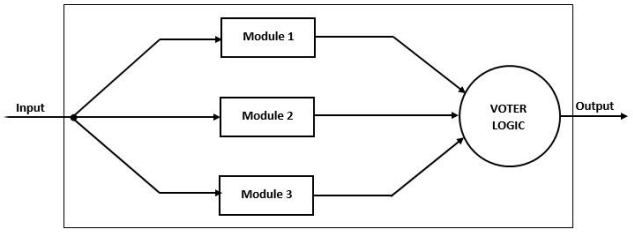


Figure 3.1
Passive hardware redundancy block diagram

The main problem in the TMR is the voter. In the case of its failure, the entire system work will be influenced. Therefore, the reliability of the simplest TMR is of the same level as the reliability of the voter. To overcome voter failure, the voter tripling technique can be used which will provide three independent outputs. [24]

In the TMR, special attention is paid to the voting techniques used by a voter. The voting process includes many problems. The first one is a decision whether to use the hardware voter or to conduct the voting process on a software level. The second problem relates to the practical realization of voting; for instance, three results of the TMR do not agree in total even when there is no any failure or fault. The processing of this disagreement can make it even more significant. Finally, the majority voter can decide that there are no two results in the TMR system that are in agreement, although the systems work correctly.

3.1.2 Active Hardware Redundancy

Active hardware redundancy tries to achieve fault-tolerance by fault detection, localization, and repair. Active hardware redundancy does not use fault masking technique, and it is used in the systems that can tolerate temporary, incorrect results under the condition that system is reconfigured and that its operation is stabilized in the satisfactory period.

3.1.2 Hybrid Hardware Redundancy

As already mentioned, the hybrid redundancy represents the combination of active and passive redundancy. The fault masking is used to prevent errors, and detection, localization, and repair od a fault are used for system reconfiguration in the case of system failure.

3.2 Q&A System for e-Government Services of the Republic of Serbia

The Q&A system (Crime domain) for e-Government services in the Republic of Serbia represents a quick-answer system which provides faster and more effective searching and obtaining of the answers on citizen's questions. [3] The Q&A system includes the following:

- Profound data analysis (Eng. Data mining - DM) which represents a tool for analysis of huge amount of data which constantly increases.
- Profound text analysis (Eng. Text mining – TM) which is used to reject the unneeded data to identify the data the citizen asked for.
- NLP which represents a set of techniques and methods for automatic generation of texts in natural language.

- N-gram analysis is used to overcome issues related to the lexical resources of Serbian language and which can provide satisfactory results. Namely, the Serbian language has a lot of language rules, and exceptions to these rules, so significant lexical resources are needed to analyze the documents properly.
- Apache Lucene represent a stable searching library which denotes the basis for the development of searching applications; also, it can provide a search within the created indexes and achieve good results for specific queries.

The working principle of the Q&A system for the e-Government services of the Republic of Serbia based on the Bag of Concept (BoC) model is presented in [3].

The Q&A system includes a framework for web services, a quick-answer model for eGovernment, a BoC based sub-system, an algorithm for classification of queries related to the criminal laws Criminal Code of the Republic of Serbia.

The main task of the Q&A system is a comparison of a short text represented as a vector with the queries made by citizens which are also represented as vectors. As a short text, we use the clauses of the Criminal Code of the Republic of Serbia in the Criminal domain (inflict of massive injuries).

The Q&A system contains the classification of questions related to the Criminal Code by using a special domain based on the BoC model as well as a comparison of performances of a classification of the based method [3].

In the Q&A system (Crime Domain) for the e-Government services of the Republic of Serbia, first, the mapping of questions from the BoC model defined by 31 terms is performed. Then, the filtering of a stop-word is conducted. Next, the stemming is performed, which denotes the removing of the common affix in words to conduct the morphological normalization and make more general characteristics of words. Therefore, a 4-gram steamer is used which is the most common stemming algorithm for the Serbian language. Finally, the similarity between a query and a BoC representation of three documents using the similarity functions. The Q&A system output is the corresponding document of the made query. The output denotes a message to the user having the following format "Please look at the clause No. n of the Criminal Code"; n will be determined by the Specific Annotator (SA) for stemming.

The pallet of words of the BoC model is defined by 31 terms in Serbian language: "teška, organ, teško, telesna, prouzrokovana, povredi, povreda, nesposobnost, naruši, telesno, povređenog, prinuda, povredi, kazniti, pretnja, trajno, zatvorom, ubistvom, meri, nehata, teškom, oštećen, napadom, telesnom, oslabljen, laka, povredom, tela, telo and posledice".

After the normalization by the 4-gram method from eight most common words, the set of seven words ("delo, učin, kazn, zatv, tele, teškipovr") is obtained by using the tf*idf criteria. Since this set is not enough for further analysis we use the

citizens' questions from the portal "PRO BONO" and daily newspapers "Blic" which contain the tag "physical injury/injuries". Consequently, the set of seven words is enlarged to the set containing 31 words.

The N-gram normalization is performed using the tf*idf criteria (term frequency and inverse document frequency) where tf denotes the number of word appearance in a specific text document, and idf denotes the importance of a given word. If a sum of all documents is divided by the sum of the documents wherein a given word is found, and the logarithmic function is applied, the idf value is obtained. By multiplying tf with idf, the TF-IDF measure is gotten, and it denotes word importance in a document or a set of documents.

In the Q&A system, the parts of the Criminal Code are presented as three different documents grouped in a whole, which shows the currently available knowledge. To complement this centralized repository, the answers by the related Experts are used, and they can be found on the website for free legal help named the PRO BONO [25]. The existing answers to citizens' questions correlated to the above-listed three law clauses of the Criminal Code which are a part of the section about the criminal law are analyzed. From a large number of questions, 45 questions are selected to help the best possible settings of the BoC model. Besides, to find the group of words correlated to the given problem, the Google search is used to find all related links to the term "serious bodily injury" which is the term - the basic lemma, obtained on the basis of an electronic dictionary for the Serbian language, using the available online-language resources "bag of words" [26]. The best results are found on the website of Serbian daily newspapers called the "Blic". From the mentioned website we use 35 text articles which correspond to the given query.

The sop words are the words which do not have any meaning for a given subject. There are certain types of words for which this is completely true (e.g. for conjuncts). In some other cases, this does not hold, so the selection of the stop words depends on the context of documents consideration. If it is needed to group the documents which contain data about current and previous events, then in the stop-word list the adverbs are not included. If the same documents are grouped regarding the meaning, then, the adverbs should be included in the list of the stop words. On the other hand, the nouns and verbs are rarely the stop words even, but that is also a possibility if the considered documents demand that. The standard list of stop words in English counts 600 words, while the SAS has a little fewer words, 330 exactly. Our Q&A system has a stop list containing 700 most commonly used words.

In the Q&A system, the N-gram analysis can be successfully used to find the words with the same root very easily. To select the value of N which will provide the valid results, we collect the extensions of words' forms in the Serbian language, only the words that are meaningful for the analysis are [27]. It is found that the majority of these extensions, about 92.70%, have up to 4 letters [2].

In addition, in [2], a 4-gram analysis was used because it achieves the best results in tasks on Serbian. Hence, we also use a 4-gram analysis for normalization of words in the document. The SA is a software agent which uses a BoC model of a specific domain from the knowledge base as a source of a map for keywords extraction. The SA assign the absence or presence of each extracted term following the relevant keywords within the BoC. The translation of the query and document form a raw data to the form needed for comparison can be done by the computer processing which represents the first obstacle in text similarity calculation. To overcome this problem, the text information is converted to the space-vector form [28].

To achieve good results by using the Q&A system, it is necessary to establish a good correlation between query classification and response type extraction. The goal of the SA is that the system "learns" how to map the corresponding type of answer on the basis of the query [29].

Parameters for supervised learning are set based on the values that have yielded good results in similar text classification processes. The proposed SA-based algorithm is based on the BoC approach with the task of automating the classification. The BoC is a list of all words ranked according to their descriptive value for three clauses of the Criminal Code of the Republic of Serbia (cluster membership). Similarity measures, in this case, precisely use vector representations of documents and questions for calculating the distance between them.

The Q&A system (Crime domain) for the e-Government services of the Republic of Serbia system was implemented using the Apache Lucene. The Apache Lucene is a scalable search library that represents the basis on which the search application is developed and which analyzes and indexes the textual content and provides the search within the created indexes and displays the search results for a particular query. The main task is to make a comparison of a short text, given as a vector, with the queries set by the citizens, which are also presented as vectors. Here, as short texts, certain articles of the Criminal Code of the Republic of Serbia relating to inflicting physical injuries are used.

To determine the threshold of similarity between the question presented in the form of short text and the Articles of the law, which is also shown as a short text, the following formula is used [30]:

$$Similarity = \frac{W(Sa) \cap W(Sb)}{\min(W(Sa), W(SB))} \quad (1)$$

where $W(S_a) \cap W(S_b)$ is an intersection set number of words in questions q_i, and the number of words in rt_j, and $min\ (W(S_a),\ W(S_b))$ is a value lower than the number of words in both documents.

3.3 Voting System Solutions in TMR and Q&A System

As we mentioned in introduction of this chapter, redundancy is a common approach to improve the reliability and availability of a system and there are various methods, techniques, and terminologies for implementing redundancy in one system.

To use one type of redundancy in one Q&A system whose optimization work through increasing its efficiency is the subject of this paper we used basic principles of TMR which refers to the approach of having multiply modules running in parallel, receive the same input information at the same time and their output values are then compared and a voter decides which output values should be used further in Q&A system.

N-version programming is one type of software redundancy and the well-known software development approach which ensures high dependability and fault tolerance of software. One algorithm have to be considered when choosing an optimal variant of N-version software, which could be and N different types of algorithms which solved same problem i.e. they received same input at the same time with the task to calculate same output but voting system as an obligatory part of this type of redundancy system has the task of selecting the best according to a particular criterion.

Voting algorithms in one TMR, as a special case of a N Modular Redundancy-NMR, play a significant role in most fault-tolerant and control systems so these algorithms are continually in develop and progress and with regard to different systems, new types of voting algorithms which could be de developing like different types of iterative algorithms, Markov modelling, neural networks etc.

For one Q&A system, to make it more efficient, it is necessary to determine which similarity measures will be used before the obligatory clustering of documents having in mind that no similarity measure is universally best for clustering of all types of documents and that this job is obligatory for each Q&A system.

We can use N but it is enough three different similarity measures in one Q&A and applying basic principles of TMR and one from mentioned voting algorithms choose the best from this three different similarity measures and in this way construct one optimized Q&A system in terms of its efficiency.

In the next chapter we will consider one such optimization using one iterative algorithm and three types of commonly used similarity measures which are considered in the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia and that: Cos, Jaccard, and Euclid, to choose the similarity measure that gives the most accurate results for the crime field.

4 Using Basic Principles of 3-TMR System to Construct One Algorithm for Q&A Systems

For the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia, it is necessary to determine similarity measures (distances) before the clustering. The similarity measure is very important due to the direct impact on the documents ranking because of its direct impact on the proximity degree or a distance from the target documents. In addition, the measurement of the similarity of documents based on characteristics that depend on the type of data that are in the context of documents and processing leads to grouping and clustering of documents within the cluster. No similarity measure is universally best for clustering of all types of documents.

Selection of an appropriate similarity measure is crucial for cluster analysis, especially for a specific type of clustering algorithms. Three types of commonly used similarity measures are analyzed in the Q&A system (Criminal domain) for the e-Government services of the Republic of Serbia: Cos, Jaccard, and Euclid, to choose the similarity measure that gives the most accurate results for the crime field.

On a specific sample query (usually 10% of the total queries used in the analysis), all three similarity measures are applied, and the one which provides the best result is chosen as the most appropriate one. In order to select the best measure of similarity, the Expert determines the correct answer for each of the queries from the query prompt. Queries are in the form of textual documents in the area of the Criminal Code of the Republic of Serbia.

The results of this analysis determine the reference measure of similarity, one of the three applied similarity measures, and it is taken as a reference measure for further algorithm calculation (Fig. 4.1).

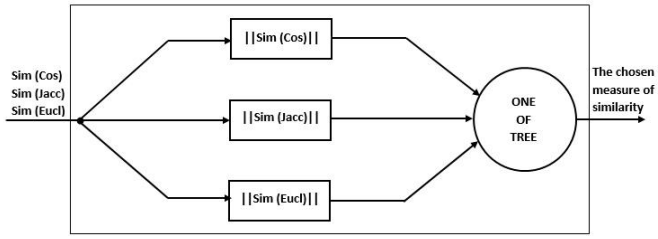


Figure 4.1

The TMR with three similarity measure (Cos, Jaccard and Euclid) and one voter (one of three)

The choice of a similarity measure for the Q&A system for the e-Government services of the Republic of Serbia is of crucial importance for determining the answers that the system delivers to citizens, businesses or employees in the public administration. Their satisfaction with the provided service depends on the quality of the answers they receive from the Q&A system.

For this reason, in this work, we put a focus on the possibilities of improving the selection of similarity measure in the existing Q&A system. The basic idea is that in the selection of a similarity measure all three similarity measures are used.

With the aim to increase the quality of the Q&A system response, we applied the principles of redundancy and fault tolerance of the system. In particular, the principles of triple modular redundancy (TMR) with one voter were applied.

The TMR is characteristic because of the hardware tripling (modules) and because a majority vote is used in determining the system output. The application of the TMR principle to obtain the best measure of similarity for the Q&A system is based on the fact that each module calculates a normalized measure of similarity, respectively: Cos, Jaccard, and Euclid. For all three normalized similarity measures, the same algorithm like module in one TMR system is used. From the point of view of the information redundancy, the TMR for obtaining the best measure of similarity for the Q&A system is an informational redundancy due to the obtaining of more than one information (three information) as needed and sufficient for the Q&A system.

For all three TMR modules, the inputs are the same: $Sim_i(Cos)$, $Sim_i(Jacc.)$ and $Sim_i\ (Eucl.)$, for the query for which the BoC Q&A system already has an answer and for which the Expert confirmed that it is correct.

In the first step, each of the algorithms calculates the normalized value of the input values according to the following formulas:

$$||Sim_i(Cos)|| = \frac{Sim_i(Cos)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{2}$$

$$||Sim_i(Jacc.)|| = \frac{Sim_i(Jacc.)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{3}$$

$$||Sim_i(Eucl.)|| = \frac{Sim_i(Eucl.)}{Sim_i(Cos)+Sim_i(Jacc.)+Sim_i(Eucl.)} \tag{4}$$

This process is repeated for the next query, i.e., for the following input values for algorithms until the following conditions are met:

$$\frac{\sum_1^i||Sim_i(Cos)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Cos)||}{i-1} = 0.000 \tag{5}$$

$$\frac{\sum_1^i||Sim_i(Jacc.)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Jacc.)||}{i-1} = 0.000 \tag{6}$$

$$\frac{\sum_1^i||Sim_i(Eucl.)||}{i} - \frac{\sum_1^{i-1}||Sim_{i-1}(Eucl.)||}{i-1} = 0.000 \tag{7}$$

Accuracy in the subtracting process is taken to three decimal places.

In order to ensure the exit from the algorithm, another criterion was introduced:

$$N < 100$$

This means that if the mean value on the third decimals is not found for up to 100 queries, then the last mean value is taken for further consideration.

This is an iterative algorithm that stops when the conditions are met. The algorithm calculates the mean values for each normalized similarity and their mean values. When the previous mean value is equal to the next one in the first three decimals, then, the algorithm stops because the condition is fulfilled.

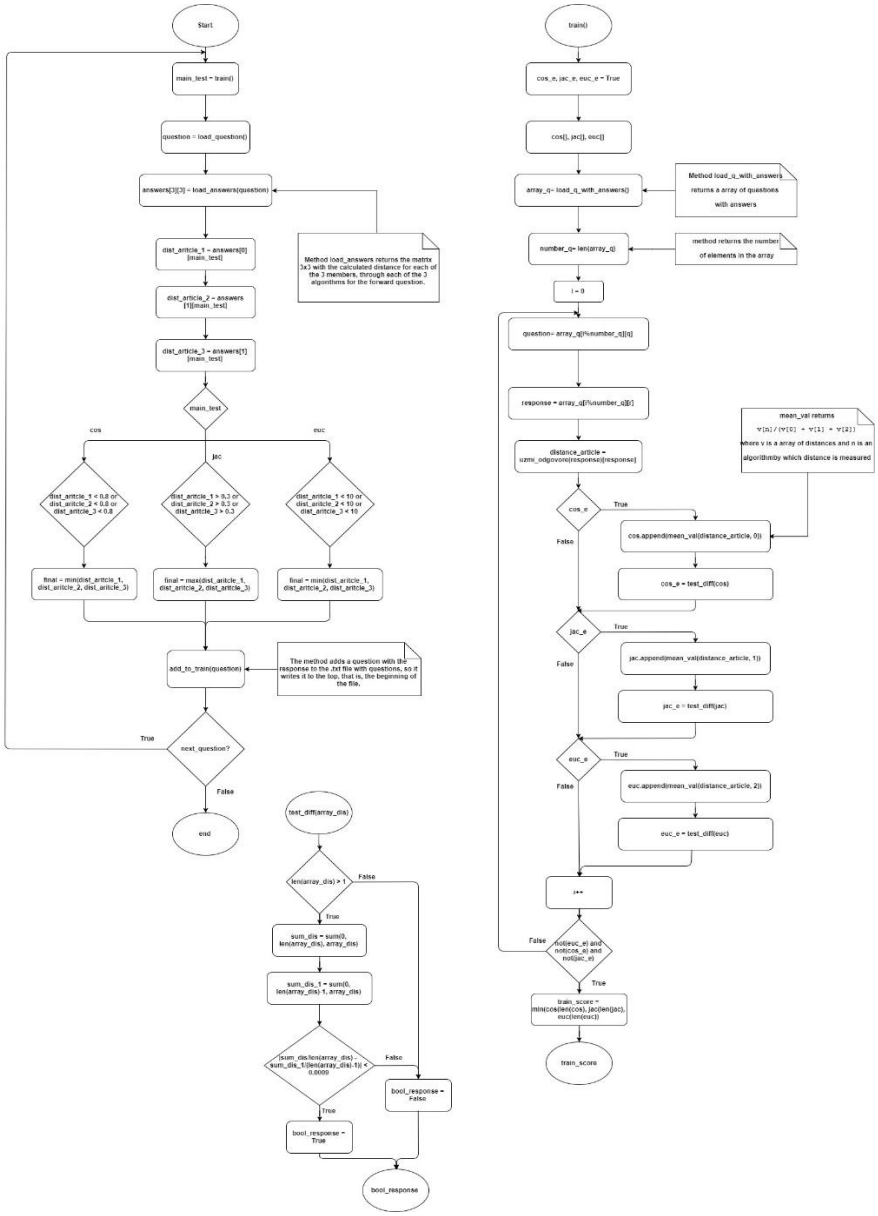


Figure 4.2
Block-diagram of proposed algorithm

To determine the mean distance in each algorithm (TMR module), the following formulas are used:

$$\frac{\sum_1^i ||Sim_i(Cos)||}{i} \quad (8)$$

$$\frac{\sum_1^i ||Sim_i(Jacc.)||}{i} \quad (9)$$

$$\frac{\sum_1^i ||Sim_i(Eucl.)||}{i} \quad (10)$$

The output of each of the algorithms (TMR modules) represents information.

These three distance values denote the three inputs of the TMR voter. The logic of the TMR voter is to determine the minimum value using the three obtained values. The smallest value represents the smallest distance, i.e. the greatest similarity between the two vectors. On the basis of the greatest similarity, the measure that will be used to searching is determined.

5 Experimental Evaluation – Case Study Q&A System for e-Government Services of the Republic of Serbia

In the Republic of Serbia, the laws related to the criminal represent a special kind of laws. One of such a law is Criminal Code of the Republic of Serbia. The current Criminal Code is presented in the "Sl. glasnik RS", br. 85/2005, 88/2005 - ispr., 107/2005 - ispr., 72/2009, 111/2009, 121/2012, 104/2013 i 108/2014) [31].

The Criminal Code of the Republic of Serbia examines the issue of guilt and the provisions of sanctions in relation to the omitted crimes in 36 segments, one of which is related to the physical injuries that are processed from many aspects. According to the possibility to inflict the physical injuries, three segments of the Criminal Code are selected:

1. Massive physical injury, Clause 121
2. Light physical injury, Clause 122
3. Coercion, Clause 135

These concepts are mapped on criminal vocabulary for three laws of Criminal Code, and they are given as vectors in the BoC:

- Clause 121 [kazn,zatv,tešk,post,javn,poku,tuži,pril,delo,tele,povr,nane, sumn,uhap]
- Clause 122 [tele,poli,napa,udar,post,nane,učin,kazn,zatv,kriv,javn,povr, lake,prij,poku,tuži]

- Clause 135 [prij,poli,kriv,post]

For specific relevant measures of similarity, 100 queries are prepared and used, wherein the Expert is determined the correct answer of each query, i.e., to which of the above three laws the query relates to.

The algorithm input consists of three measures of similarity: Sim (Cos), Sim (Jacc.) and Sim (Eucl.) which match the answer the Expert labelled as the correct one for a specific query. The measures of similarity for the first three queries are given in Table 1.

Table 1
The measures of similarities for the first three queries

R. br.	**Sim (Cos)**	**Sim (Jacc.)**	**Sim (Eucl.)**
1.	0.413528	1.000000	15.362291
2.	0.790875	0.900000	3.872983
3.	0.72175	1.000000	5.567764

The first step in any of three algorithms is to calculate the normalized values using the corresponding equations Eq. (2)-(4), as presented in Table 2.

Table 2
The normalized similarities

R. br.	$\|\|Sim\ \ (Cos)\|\|$	$\|\|Sim\ \ (Jacc.)\|\|$	$\|\|Sim\ \ (Eucll.)\|\|$
1.	0,0246502421133657	0.0596096083297036	0.9157401495569310
2.	0.1421450727175280	0.1617582619829620	0.6960966652995100
3.	0.0108698287740294	0.1371833568054060	0.7638045554202930
25.	0.0411644913582571	0.0586313221421774	0.9588354463586860
26.	0.0695500826502071	0.1332351538288670	0.7972147635209250
…	…	END	…
32.	0.0446494071297836		0.9070069585827190
33.	0.0626163530416827		0.8075977536867500
…	…		END
50.	0.0108698287740294		
51.	0.0337292888988060		
	END		

According to the values in Table 2, it can be seen that first stops the loop of Algorithm 2 (module 2) during the processing of query 26. Then, the following condition is satisfied:

$$\frac{\sum_1^{26}||Sim_i(Jacc.)||}{26} - \frac{\sum_1^{25}||Sim_{25}(Jacc.)||}{25} = 0.000 \qquad (11)$$

0.1093293107839940 - 0.1083730770622000 = **0.000**9562337217949

The next one stops the loop of Algorithm 3 (module 3) during the processing of query 33. Then, the following condition is satisfied:

$$\frac{\sum_1^{33}||Sim_i(Eucl.)||}{33} - \frac{\sum_1^{32}||Sim_{32}(Eucl.)||}{32} = 0.000 \quad (12)$$

0.8354245664315350 - 0.8362941543298090 = -**0.000**8695878982745

In the end, stops the loop of Algorithm 1 (module 1) during the processing of query 51. Then, the following condition is satisfied:

$$\frac{\sum_1^{51}||Sim_{51}(Cos)||}{51} - \frac{\sum_1^{50}||Sim_{50}(Cos)||}{50} = 0.000 \quad (13)$$

0.0604216790226415 - 0.0609555268251182 = -**0.000**5338478024767

The last step in all three algorithms is the calculation of a distance using Eq. (9)-(11). The values of *i* for all three algorithms is 26, 33 and 51, respectively.

$$\frac{\sum_1^{26}||Sim_i(Jacc.)||}{26} = 0.1093293107839940 \quad (14)$$

$$\frac{\sum_1^{33}||Sim_i(Eucl.)||}{33} = 0.8075977536867500 \quad (15)$$

$$\frac{\sum_1^{51}||Sim_i(Cos)||}{51} = 0.0604216790226415 \quad (16)$$

The logic of TMR voter is based on the determination of the smallest value in one of three line of constructed TMR because the smallest value represents the smallest distance between two vectors, i.e., the highest similarity.

In this case, the smallest value presents the normalized cosine similarity, Eq. (16).

To verify the validity of the proposed system 1830 queries related to the Criminal Code of the Republic of Serbia were collected. After the processing of collected queries, the proposed system eliminated 270 queries which related to the clauses not included in this experiment. Thus, further processing is continued with the remaining 1560 queries.

The verification results are presented in Table 3. In Table 3, Doc represents the document corresponding to the particular query. The related verification parameters were calculated using Eq. (17)-(20).

Table 3

Evaluation Metrics: Classification View

Doc / Action	**Retrieved**	**Not Retrieved**
Relevant	Relevant Retrieved	Relevant Rejected
Not relevant	Irrelevant Retrieved	Irrelevant Rejected

$$Precision = \frac{Relevant\ Retrieved}{Retrieved} = 49.67\ \% \quad (17)$$

$$Recall = \frac{Relevant\ Retrieved}{Relevant} = 49.67\ \% \quad (18)$$

$$F_{i\,(i=1,n)} = \frac{2*precision_i*recall_i}{precision_i+recall_i} \tag{19}$$

$$F_{Average} = \frac{F_1+F_2+\cdots+F_n}{N} = 49.67\ \% \tag{20}$$

The precision, recall, and F_1 parameter focused on true positives, i.e., the positive examples of the gold standard. In a monolingual alignment, the positive examples denoted the tokens that were aligned, while the negative examples denoted the tokens that were not aligned. Usually, the focus is only on whether those which should have been aligned, are indeed correctly aligned; thus the measure of $F_1$is a good *t*.

Since the correct rejection value, related to the true negatives, is important, the accuracy was computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 74.83\ \% \tag{22}$$

The accuracy weighs the true positives and the true negatives equally. The aim is to ensure that the Classifier recognizes both positive and negative examples. Specifically, in the alignment, where most tokens were not aligned, the accuracy value was likely to be very high in general, and it was difficult to determine the difference. In such a situation, only F1 on positive (aligned) examples was reported.

The result of questions based verification - With the aim to test the proposed algorithm, ten questions were used. The results obtained by the proposed algorithm are given in Table 4. Table 4 also displays the results provided by the Expert in the field of the Criminal Code of the Republic of Serbia.

Table 4
Comparison results of proposed and existing algorithm

	New algorithm	Old algorithm Cos	Old algorithm Euclid	Old algorithm Jaccard
TP	909	877	493	709
TN	921	953	1337	1121
FP	921	953	1337	1121
FN	4569	4537	4153	4369
Precision	0.496721	0.479234973	0.269398907	0.387431694
Recall	0.496721	0.479234973	0.269398907	0.387431694
$F_{Average}$	0.496721	0.479234973	0.269398907	0.387431694
Accuracy	0.748361	0.739617486	0.634699454	0.693715847

Conclusions

This work is focused on the improvement of the existing Q&A system (Crime Domain) within the e-Government services, of the Republic of Serbia, from the aspect of improving the similarity measure, which represents, a significant feature

for proper system operation. Similarity measure determines the similarity of direct influence on speed and distance from the necessary documents; the existing Q&A uses one of three similarity measures: Cos, Jaccard, and Euclid.

A new approach presented herein and validated by experiment, is based on the following principle, in the calculation of a new measure of similarity, all three similarity measures are used to increase the similarity level. Besides, the principles of the redundancy and the fault tolerant system are adopted by employing a triple modulation technique.

Using the described approach and application of the new algorithm, results are clearly better than the application of any measure of similarity individually, which proved to be true in the example of the Q&A system of the Government of the Republic of Serbia, i.e. eGovernment of the Republic of Serbia, where it is in experimental evaluation. In addition, the complete software is publicly available at the website:

https://drive.google.com/open?id=1Ny92N48JURDhhwy6tRCITwS8FH1eYh9E

For the correct operation of the Q&A system used in this experiment, it was necessary to use the Expert forgive opinion, concerning the accuracy of the results. Our future work will focus on an in-expert Q&A system, i.e. A Q&A system for which it will not be necessary to involve an Expert.

References

[1] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović, A. Presnall: A Framework for Delivering e-Government Support, Acta Polytechnica Hungarica, Vol. 11, No. 1, 2014

[2] U. Marovac, A. Pljasković, A. Crnišanin, E. Kajan: N-gram analysis of text documents in Serbian, TELFOR 2012

[3] V. Nikolić, B. Markoski, K Kuk, D. Randjelović, P. Čisar, Modelling the System of Receiving Quick Answers for e-Government Services: Study for the Crime Domain in the Republic of Serbia, Acta Polytechnica Hungarica

[4] M. Jovanović, G. Šimić, M. Čabarkapa, V. Nikolić, D. Ranđelović, SEFRA - Web-based framework customizable for Serbian language search applications, Paper is accepted in 2017. for publications in Acta Polytechnica Hungarica (aceppted)

[5] V. Nikolić, M. Ivković, S. Nedeljković, P. Djikanović, Information Retrieval for Unstructured Text Documents: Lucene Searching, AIIT 2015

[6] M. Martinović, S. Vesić, G. Rakić, Building an Information Retrieval System for Serbian - Challenges and Solutions, INTERSPEECH 2007

[7] N. Milošević, Stemmer for Serbian language, http://www.inspiratron.org (2018)

[8] Lin, J. 2007. An exploration of the principles underlying redundancy-based factoid question answering. ACM Trans. Inform. Syst. 25, 2, Article 6 (April 2007), 55 pages. http://doi.acm.org/10.1145/1229179.1229180

[9] Banko, M. Andrebill, E. 2001. Scaling to very very large corpora for natural language disambiguation. In Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001) 26-33

[10] Light, M., Mann, G. S., Riloff, E., Breck, E. 2001. Analyses for elucidating current question answering technology. Nat. Lang. Eng. 7, 4, 325-342

[11] Clarke, C., Cormack, G., Lynam, T. 2001. Exploiting redundancy in question answering. In Proc. of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001) 375-383

[12] Zahra Latifi, Abbas Karimi,"A TMR Genetic Voting Algorithm for Fault-tolerant Medical Robot," Medical and Rehabilitation Robotics and Instrumentation, V.42, pp. 301-307, 2014

[13] Eric P. Kim; Naresh R. Shanbhag, "Soft N-Modular Redundancy, "IEEE Transactions on Computers, V.61, pp. 323-336, 2012

[14] Omid NOROUZIFAR, Manouchehr KAZEMI, Mahdieh Nadi SENEJANI, Developing a New Weighted Voting Algorithm based on Markov Model, Bulletin de la Société Royale des Sciences de Liège, Vol. 86, special edition, 2017, pp. 528-534

[15] Zarafshan F, G Latif-Shabgahi, and Karimi A, "A novel weighted voting algorithm based on neural networks for fault-tolerant systems" Computer Science and Information Technology (ICCSIT) 3rd IEEE International Conference, Vol. 9, pp. 135-139, 2010

[16] Gruzenkin D. V., Grishina G. V., Durmuş M. S., Üstoğlu I., Tsarev R. Y. (2017) Compensation Model of Multi-attribute Decision Making and Its Application to N-Version Software Choice. In: Silhavy R., Silhavy P., Prokopova Z., Senkerik R., Kominkova Oplatkova Z. (eds) Software Engineering Trends and Techniques in Intelligent Systems. CSOC 2017. Advances in Intelligent Systems and Computing, Vol. 575, Springer, Cham

[17] Atkins, E. M., Abdelzaher, T. F., Shin, K. G. et al., Planning and Resource Allocation for Hard Real-time, Fault-Tolerant Plan Execution, Autonomous Agents and Multi-Agent Systems (2001)

[18] Berlizev A., Guelfi N. (2009) Fault Tolerance Requirements Analysis Using Deviations in the CORRECT Development Process. In: Butler M., Jones C., Romanovsky A., Troubitsyna E. (eds) Methods, Models and Tools for Fault Tolerance. Lecture Notes in Computer Science, Vol. 5454, Springer, Berlin, Heidelberg

[19] Cao, Z., Tian, Y., Le, TD. B. et al., Rule-based specification mining leveraging learning to rank, Automated Software Engineering, Springer (2018)

[20] Shekhar, C., Jain, M., Raina, A. A. et al., Reliability prediction of fault tolerant machining system with reboot and recovery delay, Int J Syst Assur Eng Manag (2018)

[21] Ferdinando C., Fault-Tolerant Search Algorithms, Springer-Verlag Berlin Heidelberg (2013)

[22] Nadia N., Luiza de M. M., Hardware for Soft Computing and Soft Computing for Hardware, Springer International Publishing (2014)

[23] Pan Z., Qi Z., Zhankui Z., Liman Y., The signal integrity design and simulation of triple modular redundant (TMR) computer, 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)

[24] Hossein S., Mehdi D., Mostafa S., Comparing the reliability in systems with triple and five modular redundancy, 2016 5th International Conference on Computer Science and Network Technology (ICCSNT)(2016)

[25] http://www.besplatnapravnapomoc.rs/

[26] http://hlt.rgf.bg.ac.rs/Page/Services

[27] D. Subotić, N. Forbes, "Serbo-Croatian language – Grammar", Oxford Clarendon press, str.25-31, 61-64, 101-113

[28] Magerman, Tom, et al. "Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents." 2011

[29] Metzler, D. and Croft, W.B., Analysis of Statistical Question Classification for Fact-based Questions, in Information Retrieval, 8(3), 481-504, 2005

[30] P. Djikanović, V. Nikolić, D. Sivčević, National Framework of Interoperability of the Republic of Serbia and Service-Oriented Architecture (SOA), YU INFO 2014

[31] www.paragraf.rs

The Effect of Bank Competition on the Cost of Credit: Empirical Evidence from the Visegrad Countries

Ashiqur Rahman[1], Manuela Tvaronavičienė[2], Luboš Smrčka[3], Armenia Androniceanu[4]

[1] Tomas Bata University in Zlin, Mostni 5139, 76001 Zlin, Czech Republic, E-mail: rahman@utb.cz

[2] Faculty of Economics, Vilnius Gediminas Technical University, Saulėtekio al. 11, 10223 Vilnius, Lithuania, E-mail: manuela.tvaronaviciene@vgtu.lt

[3] University of Economics, Prague, Faculty of Business Administration, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic, E-mail: smrckal@vse.cz

[4] Faculty of Administration and Public Management, The Bucharest University of Economic Studies, PiataRomana 6, 010374 Bucharest, Romania, E-mail: armenia.androniceanu@man.ase.ro

Abstract: The effects of bank competition on the cost of credit are a much-debated topic in Small and Medium enterprises financing. In this paper, we would like to examine the relationship between the cost of credit and interbank-competition in the context of Visegrad countries - the Czech Republic, Poland, Hungary, and the Slovak Republic. The dataset of this paper comes from two different sources, the firm level data provided by the latest version of the Business Environment and Enterprise Performance Survey that was conducted by the European Bank for Reconstruction and Development and the World Bank during 2012 to 2014, and the country level bank competition measures are collected from the Global Financial Database, updated in 2017 [3]. We have examined bank competition with four measures, including structural bank concentration measure and three non-structural (Lerner Index, H-Statistics, and Boone Index) measures. We find evidence that bank competition has a positive effect on the cost of credit and hence, our results are in-line with prior literature on information-based theories of bank competition. We have also assessed the firms in terms of their information opacity (micro, small, and medium), and we find that the cost of credit is higher for the information opaque firms. Thus, firm sizes have important implications for bank competition and cost of credit.

Keywords: Cost of credit; bank competition; SME; Visegrad countries

1 Introduction

The small and medium enterprises (SMEs) are the integral of many developed and developing countries, as they generate most of the employment and business activities. However, the growth of the SMEs is largely depended on the availability of external finance. The limited access to bank finance for the SMEs has been an issue that is far from settled in both advanced and emerging countries. The role of banks in facilitating the credit services to the business sectors are extremely vital for the development of private business sectors and for the economic welfare of a country. The banking market structure is considered as one of the important elements that can have a significant effect on the access to finance for firms and to reduce financial constraints.

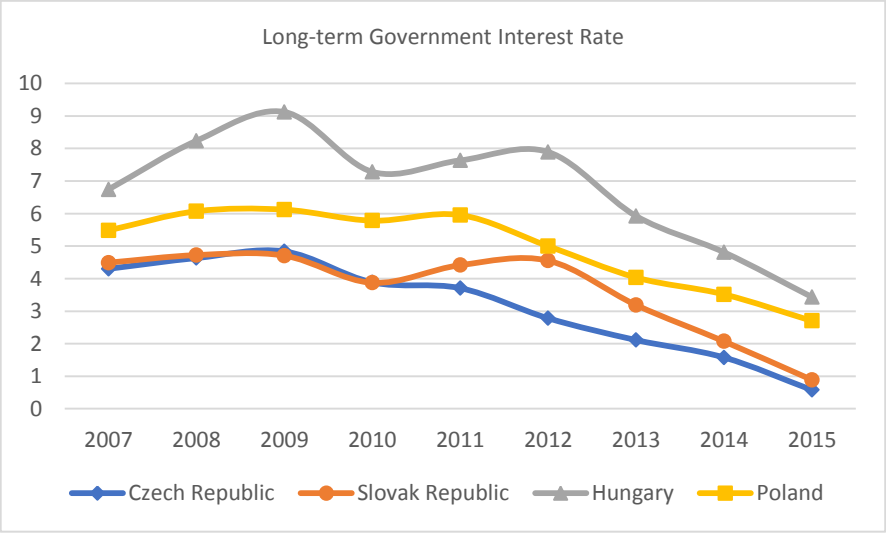
It is well documented in the prior literature that SMEs are facing problems in accessing bank loans due to information asymmetry. The reduction of information asymmetry can increase access to finance for SMEs, reduce loan interest rate, lower collateral requirements and overall facilitates the availability of finance [2] [6] [20] [27] [28]. However, the influence of information asymmetry on reducing financial constraints can be affected by the nature of bank market structure, for example, competition and concentration in the market. The literature on competition-based studies argued the effect of bank competition from two different perspectives. At one hand, the *market power hypothesis* suggests that the bank competition can increase access to finance, reduce interest rates and lower collateral requirements for SMEs [7] [26] [22]. The theory is based on the general economic assumption that higher competition can lower the cost of credit and enables better access to finance. Therefore*, the market power hypothesis* considers interbank competition is preferable for the SMEs by which financial constraints can be alleviated. On contrary, the *information hypothesis suggests* that banking competition can increase financial constraints for firms due to a high asymmetric information and agency costs. The *information hypothesis* argues that higher competition reduces bank incentive to invest in relationship lending and hence higher financial constraints due to more asymmetric information between banks and borrowers [13] [37]. A few literatures discuss that a high competition reduces bank quality of loan screening process [33], and reduces bank incentives in relationship-based lending technologies [23]. Overall, the *information hypothesis* argues that the intensive bank competition is not desirable for SMEs, as it increases financial constraints.

Regardless of the conflicting views on bank competition and financial constraints on SMEs, a great deal of empirical research tested bank market power and information hypothesis in different markets. The banking concentration has a positive effect on financial constraints for SMEs, thus supporting the market power hypothesis [2]. On the other hand, assuming bank competition is opposite to concertation, it is found that the bank concentration has a positive effect on access to finance for SMEs, hence lower credit restrictions [38].

While several studies examined the effect of bank market power in relation to access to finance [29] [41] [1] [31], in this paper, we intend to follow a different path by examining the effect of bank market power on the cost of credit for SMEs. By using the World Business Environment Survey (WBES) [4], shows that the high cost of credit is the first and foremost problems for SMEs that restrict the firms to access bank loans. Therefore, considering the importance of the cost of credit in SME financing, the objective of this current paper is to understand the association between bank market structure and its effect on pricing of SME loans.

In Figure 1, we can see the long-term government bond interest rate in the Visegrad countries. The purpose of this figure is to analyse how the interest rates have evolved over the time during the pre-and post financial crisis in the Visegrad countries. The figure shows that in the beginning of the financial crisis (2007) the interest rate in the Czech and Slovak Republic was about 4%, and the rate increased by about 1% in 2009. However, we can see that the interest rate declined for both countries in 2010. The interest rate for both Hungary and Poland is higher than the Czech and Slovak Republic. The interest rate significantly increased in Hungary during the financial crisis, from about 6.5% in 2007 to about 9% in 2010. It could be the fact that the government was providing incentivs to the investors to invest in the local bonds to collect funds to invest in the banking and other sectors. In Poland the interest rate is quite stable during the after the financial crisis, which is about 6% until 2011. However, the interest rate declined steadily after 2012 for all Visegrad countries.

Figure 1
Long-term government interest rate



Source: OECD (2015)

The current research is based on the Visegrad countries (the Czech Republic, the Slovak Republic, Poland, and Hungary). We have selected the Visegrad group on purpose because the Visegrad countries are strategically important for the European Union. On top of that, we can provide cross-country evidence from the central European countries. This paper contributes to the existing literature on the cost of credit and bank market competition in several ways. First, the competition measures of this paper are a combination of structural and non-structural measures. Second, the sample of firms are from the Visegrad countries and so far, this research is the first empirical evidence from the Visegrad group in relation to the cost of credit and bank market structure. Third, we provide a new evidence based on firm information opacity and its impact on the cost of credit.

The rest of the paper is organized as follows. Section 2 discusses the current literature on market power and its implications on access to finance and cost of credit. Section 3 presents the data set and describes the variables and empirical methodology. Section 4 discusses the descriptive and estimation results. Section 5 concludes the paper.

2 Literature Review

From a theoretical perspective, lending to SMEs requires to build-up a long-lasting relationship by which it is possible to acquire soft-information and to minimize information gap. The decrease of information asymmetry or information mismatch may positively affect access to finance for SMEs and hence minimal financial constraints. However, in competitive market banks have less incentives to provide loans based on relationship banking, because in a competitive environment a borrower can easily switch from one bank to another. Hence the minimum value added to the bank from investment in relationship banking [9]. Nevertheless, when a bank has market power it can try to develop a long-lasting relationship with the borrower by which a bank can extract exclusive private information from the borrowers [38]. Information based theory suggests that a bank can give-up immediate rent or profit margin from borrowers when they have market power, but they can take the advantage later when the bank will have superior authority over the borrowers [38]. However, the situation is opposite for banks that are operating in a competitive market. They may ask for the higher rate of interest from the borrowers' due to competitive pressure. The authors proposed that the bank market power can increase the investment in relationship banking that simplifies information asymmetry and alleviates financial constraints for SMEs. However, different authors argues that, banking competition gives opportunity for the bank to develop a more private banking relationship with the borrower and invest more in relationship lending technologies. In doing so a bank can create its superiority over other lenders by eliminating price competition [30].

The current research on banking literature examined the effect of bank competition and concentration from different perspectives such as access to finance, cost of credit, collateral requirements, financial constraints, discouraged borrowers and so on. The studies related to access to finance and bank market power provide evidence that high bank concentration can increase financial constraints on SMEs, hence, excessive bank competition is related to a greater access to finance. In this regard, [2] examined the effect of banking concentration and its effect on access to credit in developing countries and found that banking concentration is associated with higher financial constraints for SMEs. A study by [13], provided evidence from a sample of SMEs in 119 emerging countries and found that banking concentration is associated with higher financial constraints and thus, they supported the view of market power hypothesis and suggested that in emerging markets high concentration is not desirable. Similarly, [32] found that bank market power reduces access to credit for SMEs with respect to 53 developing countries. However, they argue that the negative effect of bank market power is reduced depending on the countries that are financially developed and well-structured credit market. Likewise, they find that availability of a credit information sharing system can diminish the effect of bank market power.

The empirical research on bank competition-based studies argues that the choice of competition measures can affect significantly on the outcome of results and the interpretation of results may differ by the competition indicator [11]. Therefore, it is an important issue to select the competition measure that best explains the bank market structure of the country. However, the appropriate selection of competition measure is a debatable issue in bank finance research, because different countries have different banking systems and that can affect the results of any cross-country research.

A few recent empirical studies analyzed structural and non-structural measures of competition by which they can enhance the validity and robustness of the research. [30] examined credit constraints in 69 developing countries by including both structural and non-structural measures (Concentration ration, Lerner index, Boone and H statistic) of competition and the paper finds that bank competition can alleviate credit constraints for SMEs. The results show that the banks evaluate loan applications less strictly when competition is higher. On the other hand, countries with less bank competition face higher credit rationing due to high bank concentration. [11] used Panzar – Rosse H - statistic as a proxy for bank competition measure in their analysis of 16 countries and they find that bank competition has a positive effect on the growth of firms those are largely depended on bank finance and the result is true for countries those have high competition in the market. Hence, bank competition can facilitate access to finance and growth of firms. [42] provides more evidence on bank market power and financial constraints from a sample of 20 European countries and they find that the bank competition relates to lesser credit restrictions on SMEs. To measure financial constraints, they have used the [17] investment sensitive model and the

Lerner index is used to capture the market power and it is found that in a competitive market SMEs are less sensitive to their investment policy. [7] examined the effect of bank competition and access to finance through the availability of trade credit in Spain and the authors show that in a competitive market SMEs have more access to trade finance and hence, supporting market power hypothesis. [34] examining the relationship between bank competition and the availability of finance in the Italian market and find that bank competition has a positive effect on SMEs access to external credits. Thus, they find that bank competition can minimize financial constraints on SMEs.

While most of the papers examined the issue of access to finance and bank competition, a few studies are done on how bank competition affect the pricing of loans. The preliminary research by [38] reported that bank competition has a positive effect on the cost of credit and that means that higher the competition higher is the cost of credit. [42] used a sample SMEs from 20 European countries and they have used two-structural and two non-structural measures of competition. The results reveal that the bank competition can increase the cost of credit. The authors also observed the effect of competition on the pricing of loans based on firm information opacity and they find that small and medium firms need to provide more interest rate on their borrowing than the large firms. It is argued that small firms encounter the information problems more than the large firms and thus competition has a harsher effect on the firms that are depended on relationship-based lending. Therefore, the above studies are supporting the information hypothesis of bank competition. However, [22] showed that bank competition can relax the lending terms by reducing collateral requirements and interest rates on loan contact. Hence, empirical studies on bank competition and its relationship with the cost of credit are mixed and that is why we have chosen to examine the issue in the context of Visegrad countries.

3 Data, Method and Variables

3.1 Data

This paper utilizes data from the Business Environment and Enterprise Performance Survey (BEEPS) that was conducted by the European Bank for Reconstruction and Development and the World Bank during the period of 2012 to 2014. The survey is performed to understand the overall business environment and the enterprises' performance-related factors in 30 transition and emerging countries including European, Central Asian countries and Russia. The survey covered 1,374 firms in four examined countries – 254 from the Czech Republic, 310 from Hungary, 542 from Poland, and 268 from the Slovak Republic.

According to the aim of the paper, the small and medium enterprises are defined, under the Convention of the Organization for Economic Co-Operation and Development (OECD) and the guidelines are given in the survey, as enterprises with a maximum of 250 employees. After refining the dataset by excluding the missing variables and the large firms from the sample size, the analysis involves 1,296 records about firms for descriptive statistics, and 230 firms have disclosed information about the cost of credit in the survey. Regardless, of the firm-level data from the BEEPS survey, we have collected the country level competition measures data from the [3], Global Financial Database, which is updated in 2017.

3.2 Variables

To analyze the impact of bank competition on the cost of credit for SMEs, we have collected the cost of credit information from the BEEPS survey question "Q46 - What is the annual nominal interest rate (in percent) of the most recent line of credit of loan". The Cost of credit is our main dependent variable in the context of the research. A detailed list of variables is presented in Table 1.

In this paper, we have a few firm-level control variables such as Firm size (Size), Firm age (Age), Largest Owner (Largest Own), Borrower Experience (Experience), Audit (Audit), and Innovation (Innovation). Firm size (SIZE) is counted based on the number of full-time employees the firm had during the BEEPS survey. We assume to find an inverse relationship between the size of the firms and the cost of credit because the larger firms would face lesser information opacity problem than the smaller ones and can access loans with a lower interest rate [24] [35]. We control for firm age (Age), which is measured by years the firm is in operation. We also expect to find an inverse relationship between firm size and the cost credit, because the older firms may have a better business relationship with the banks and other external lenders due to their long existence in the market and hence, they may access loans with better credit terms, such as lower interest rate [5].

In this current paper, we also control for firm ownership structure and its effect on the cost of credit. As per the agency theory, firms having concentrated ownership and those operated and controlled by the same individual have less and sometimes may have zero agency costs [16] [25]. Thus, we presume to find a negative relationship between ownership concentration and the cost of credit. Because less agency cost may reduce the credit risk of the firm, andmay induce lenders to provide loans a with lower price. Additionally, we control for borrower experience and its effect on the cost of credit. Borrower experience is counted by the number of years the top manager within the current business or related businesses. It is found that the cost of credit is lower for an experienced borrower than of the younger borrower [35]. Because, an experienced borrower can maintain the business better than an inexperienced borrower and hence, it signals lower credit

risk of the firm. Therefore, banks and other external lenders can provide loans with lower interest rates. On the other hand, [21] contend that an experienced borrower can have more bargaining power with the creditors in compared to an inexperienced borrower and which may lead to a lower cost of credit. Hence, we expect to find a negative relationship between the cost of credit and borrower experience. Afterwards, we control for firm financial reporting status and its relationship with cost of credit. We measure financial reporting status of the firm with audit (Audit) report. The Audit is a dummy variable that takes one if the business has an audited financial statement and zero otherwise. It is widely discussed in prior literature that when a firm has its financial statement audited by external auditors, it can help to minimize information asymmetry between firms and the creditors and thus can receive loans with lessercredit restrictions [34] [30] [39]. Therefore, we expect to find a negative relationship between the audit report of the firm and cost of credit. Because a third party certified financial statements may increase lenders confidence on the borrower and provide loans with a lower cost of credit. Finally, we control for firm innovation activity and its impact on the cost of credit. It is argued that innovative firms are more information opaque compared to the non-innovative firms. Thus, innovative firms face higher credit restrictions than the non-innovative ones [18] [29]. Considering the above theoretical arguments surrounding the innovative SMEs, we expect to find a positive effect of bank competition on the cost of credit.

Table 1
Definition and sources of variables

Variable	Definition	Source
Cost of credit	Annual interest rate on loan	BEEPS
Firm-level control		
Size	Size of the firm, measured as the number of full-time employees	BEEPS
Age	Age of firm, measured as the number of years that the firm has been operating	BEEPS
Largest. Own	Percentage ownership of the firm held by the largest shareholder	BEEPS
Experience	Experience of top manager measured in years	BEEPS
Audit	Equals 1 if the firm financial statement is checked by external auditors (0,1)	BEEPS
Innovation	Equals 1 if the firm has introduced any new products within the last three years	BEEPS
Competition measures		
H-stat.	A measure of the degree of competition	Beck et al. (2000)
Lerner	A measure of market power in the banking market	Beck et al. (2000)

Boone	A measure of the degree of competition based on Profit - efficiency in the banking market	Beck et al. (2000)
CR5	The asset share of the five largest banks in total banking system assets	Beck et al. (2000)

Source: This table presents variable definitions and sources of the data set. BEEPS = Business Environment and Enterprise Performance Survey.

3.2.1 Competition Measures

The goal of the current research is to inspect the relationship between bank competition and the cost of credit and thus, it is necessary to select appropriate measures of bank competition. The literature on competition-based studies classified bank competition into two segments: structural indicators and non-structural indicators. With respect to structural indicators, the theory suggests that the excessive concentration in the banking sector can be considered an opposite to bank competition and in a concentrated market a bank can ask for higher loan rates from the borrower by which it can generate more profits than in a competitive market. The commonly used structural bank competition measure is concentration ratio, which is in inverse proxy of competition and is proxied by asset share of the largest five banks in the overall banking market (Cr). We intend to use concentration ratio as a measure of structural measure of bank competition.

Apart from the concentration ratio, in this paper, we have employed three (Lerner index, H statistics, and Boone index) non-structural measures of bank competition. The Lerner index captures the market power of a bank and that is analyzed by the difference between output prices and marginal costs of inputs. The output prices are observed by total bank revenue in terms of its assets, and the marginal costs are calculated from an estimated translog cost function of three inputs (labor, physical capital and deposits; a detailed methodological explanation is cited in [31] with respect to output. The greater values of the Lerner index are associated with a lesser bank competition. That means that when a bank can set higher prices over the costs, it has more market power. Because in a competitive market, it would be difficult for a bank to charge higher prices than the marginal costs due to competition from other banks.

In this paper, we further introduced Panzar-Rosse H statistics [37], which is also a commonly used competition measure in banking literature. ThePanzar-Rosse model measures the elasticity of bank revenues to its input prices and it shows that under certain condition the prices of inputs vary conditional on the intensity of competition in the market. The H statistics value gives information about the degree of competition in a market and by which it is possible to understand the competitive nature of the banking industry in a market [37]. When a market operates under a perfect competition, the H-statistic equals 1. Whereas under a monopoly, an increase in input prices results in a rise in marginal costs, a fall in output, and a decline in revenues leading to an H-statistic less than or equal to 0.

And, H-statistic is between 0 and 1, when the banking sectors operate under monopolistic competition.

Finally, the competition measures we introduced in this paper is the Boone index. Boone [8] introduced a model grounded on the price elasticity of profits to marginal costs. To measure the elasticity, the log of profits (measured by return on assets) is regressed on the log of marginal costs. The estimated coefficient (computed from the first derivative of a trans-log cost function) is the elasticity. Hence, the more negative is the Boone indicator, the greater is the degree of competition because the effect of reallocation is stronger. The basic intuition of the model is that only the efficient banks can earn a higher level of profits in terms of their costs. Additionally, the model explains that the propensity of earnings increases with the competitive nature of the market. That means that, as the market gets more competitive the efficient banks can generate more profits than of the inefficient banks. The Boone indicator is intensively used in the banking literature because it has some advantages over other competition measures [31] [30] [12]. The Boone indicator can reflect the dynamics and non-price related factors in the market, however, there is a limitation exists in Boone index. The Boone index shows the intensity of competition for the overall economy or as a country in total, but it does not capture the regional differences within the country. Hence, the index may not be not well fitted when performing analysis on a large country. Since the regional differences in a banking environment may create differences in overall countrywide banking competition measures. To make the empirical results more understandable, in this paper we used the inverse of Boone index that means that higher the values of Boone index higher is the competition.

3.3 Methodology

The aim of this paper is to examine the relationship between bank competition and the cost of credit in the context of the Visegrad group. The cost of credit is a continuous variable and as a result, we intend to use an OLS regression model that is the best fit for our purpose. The empirical model to be examined as follows:

$$Y_{fct}\,(\text{cost of credit}) = \beta_1\,\text{Firm level controls}_{fct} + \beta_2\,\text{Competition}_{ct} + \varepsilon_{fct}$$

Where Yis the cost of credit, and fct represent firm (f), country (c), and time (t). In our baseline model, we have *Firm-level controls* (size, age, ownership, etc.); *Competition* indicates one of our competition measures, and ε_i Is the usual error term. In our model, the impact of bank competition is indicated by β. As already discussed elsewhere, the higher values of competition measures are associated with a lower level of competition (Cr and Lerner) and higher values competition measures are associated with higher levels of competition (Boone index and the H statistics). That can also be said that three of our competition measures are an inverse proxy of bank market competition. Hence, if β> 0 that means higher

concentration is associated with higher cost of credit and if we find a $\beta < 0$, that means higher concentration is associated with lower cost of credit.

4 Results

4.1 Descriptive Statistics

In Table 2, we present the descriptive statistics of our full sample. The Table shows that the average cost of credit is about 8.15% of our sample. However, the maximum cost of credit is about 70%,which is tremendously high. This preliminary result may highlight that the SMEs sometimes need to pay an extremely high price for their loans, regardless of the nature of competition in the banking sector. Considering the firm level determinants of the cost of credit, we see that an average firm employs 33 employees and hence it could be said that most of the firms in our sample are in the range of small firms (10-49). If we consider the firm age, it is possible to see that the average maturity of the SMEs is about 18.5 years. However, the sample suggests that the firm age ranges from 1 to 81 years. That may highlight that our sample covers both mature and just newly established firms. As per the ownership structure of firms, we find that SMEs are highly concentrated with 77% of concentration, hence that may reflect that SME owners are more likely to keep their control over the firm by holding a large share. In terms of borrower business experience, we can see that the mean experience of the borrower is about 21 years. Considering the borrower experience, we may find a negative association with the cost of credit because an experienced borrower may have more bargaining power in comparison to the inexperienced borrower. The descriptive statistics suggest that about 34% of SMEs in our sample have their financial statement audited. The audited financial statement can have a significant impact in determining the cost of credit since, it shows the quality of the firm's financial information also reduces information asymmetry. With respect to innovation, we can see that about 31% of the SMEs have introduced new products within the last three years. The result may reflect that the SMEs in our sample countries are not actively participating in innovation activities.

With respect to the competition measures, we find that the banking sector is highly concentrated in our sample with a five-bank concentration ratio of 68.42% (CR). On the other hand, we can see that the Lerner index was in between 0.13 to 0.40 during the survey period, and H statistics show that it ranges from 0.61 to 0.70, whereas, Boone index was 0.01 to 0.16.

Table 3 shows the cross country analysis of cost of credit and the differences in competition measures in our surveyed countries and compared with the EU average and also with OECD countries.

Table 2
Descriptive statistics (Total sample)

Variable	Obs.	Mean	Std. Dev	Min	Max
Cost of credit	230	8.15	7.27	0.00	70.00
Firm characteristics					
Size	1296	32.91	45.53	1.00	245.00
Age	1292	18.34	8.87	1.00	81.00
Largest. Own	1267	76.38	26.21	0.00	100.00
Experience	1228	20.41	9.86	1.00	57.00
Audit	1286	0.34	0.47	0.00	1.00
Innovation	1295	0.31	0.46	0.00	1.00
Competition5 measures					
CR5	1296	68.42	13.41	53.66	88.52
Lerner	1296	0.29	0.09	0.13	0.40
H-stat.	1296	0.63	0.05	0.61	0.73
Boone	1296	0.07	0.05	0.01	0.16

Note: Firm level variables are authors calculation based on the BEEPS survey and competition measures are obtained from the Beck et al. (2000) GFDD database.

The table shows that the average cost of credit in the Visegrad countries is about about 8.63 % and the lowest interest rate in the Visegrad countries is in the Czech Republic, which is about 5.7%. The interest rate is about 7.06% in the Slovak Repuplic and for both Hungary and Poland the rate is about 10.81 and 10.84%, respectively. With respect to the average of EU and the OECD countries, we can see that the interest rate is about 3.6 and 3.91% respectively. Therefore the data clearly shows that the interest rate in the Visegrad countries are significantly higher than the other EU and OECD countries. The higher interest rate in the Visegrad countries may impose significant barriers for the SMEs to borrow funds from the externa market and that can deter their business growth.

The level of bank concentration (CR5) is extremely high in the Slovak Republic, which is about 90.17%. That might reflect that the borrowers in the Slovak Republic have very limted alternative options to look for external funds where they can bargain for favourable loan terms. In the Visegrad countries, Poland has the lowest level of bank concentration that is about 54.05%. That shows the banking sector in Poland is relatively competitive in compared to the Czech, Slovak or Hungarian banking sector. We can see that the concentration in the Slovak Republic is also higher than the average of the EU and OECD countries. On the other hand, the concentration in other three countries (Czech Republic, Hungary and Poland) are lower than the EU and OECD average. Considering the H-Stat, Lerner index and Boone index, we can see that the H stat is also higher in the Slovak Republic than the other three Visegrad countries, and the result is also

higher than the EU average and OECD average. The Lerner index is comparatively higher in Poland (0.45) and in Czech Republic (0.42) than the Slovak Republic and Hungary and also in compared to the mean of Visegrad countries as well as the EU and OECD countries. The higher Lerner index may reflect that the banks in these countries are able to maintain their product prices higher than their cost of input prices. Therfore, we can assume that the banking sector in the Visegrad countries is comparatively less competitive than the other EU and OECD countries.

Table 3

Cross country analysis of average ost of credit and bank competition measures

		Mean Bank Competition Measures			
	Mean Cost of Credit (%)	H-stat.	Lerner	Boone	CR5
Czech Republic	5.79	0.61	0.42	0.04	78.84
Slovak Republic	7.06	0.74	0.30	0.00	90.17
Hungary	10.81	0.61	0.33	0.07	72.53
Poland	10.84	0.66	0.45	0.02	54.05
Mean V4 Countries	8.63	0.66	0.37	0.03	73.90
European Union (27)	3.6	0.65	0.21	0.05	81.08
OECD Countries	3.91	0.62	0.19	0.02	79.53

Source: Interest rate is based on the BEEPS survey and the all other data is collected from Beck et al.(2000) Global Financial Database.

4.2 Empirical Results

In Table 4, we present the regression results for each of the competition measures and their relationship with the cost of credit. As already discussed, Cr and Lerner index indicate that higher values of competition measures are related to lower levels of competition and conversely, H statistics and Boone index (inverse values of Boone index are used in this paper) suggest that higher values of competition measures are associated with higher levels of competition in the market. In column 1, we see that the coefficients of Cr are negative, similarly, in column 2, the coefficients of the Lerner index are also significant and negative. We find that the coefficients of H statistics are negative but not statistically significant (column 3).

Finally, the results for Boone index shows a positive significant result with the cost of credit (column 4). Hence, if we consider the results for the first two competition measures, it suggests that the higher concentration is negatively related to cost of credit and from the results we may say that the higher level of concentration can help the firms to get loans with lower interest rates. However, the cost of credit is higher when market competition is excessive. The competition

results of our paper corroborate the information hypothesis, where we argue that the competition does have a positive impact on the cost of credit, due to less benefit of banks in investing relationship lending. Hence, lack of information increases the cost of credit for borrowers. The results for Boone index suggest that the high competition in the market can increase the cost of credit and thus our results for all competition measures are in line with the information hypothesis, apart from the H statistics. Therefore, we may say that market competition is not helpful to reduce the cost of credit, rather a concentration structure of the banking system in the Visegrad group is more suitable to reduce the cost of credit for the borrowers. Our results suggest that the structural and non-structural competition measures have similar implications on the cost of credit and the selection of competition measures does not distort the interpretation of our results. The results of this paper are in line with recent literature on bank competition and the cost of credit. [19, 38] who have also found that high bank competition increases the cost of credit and which is mainly driven by the information problems associated with SMEs.

With respect to the firm level controls, we find that firm size has a negative impact on the cost of credit and the results are stable for all competition measures. Thus, the result suggests that larger firms may have easy access to finance with a lower cost of credit due to their more bargaining power than of the smaller firms or the large firms are more transparent and information asymmetry may not have a detrimental effect on the large firms' credit availability. In terms of firm age, the results show that the cost of credit is higher for the larger firms than of the smaller firms. The results are opposite to our expectation. We expected a negative association with firm age and cost of credit due to their mature business status and that might give the aged and older firms a better credit contract from the bank with a lower cost of credit. However, this result could be the fact that the banks charge more interest rates on their loans from the mature and older firms because they are able to give more interest on their borrowing than of the younger and newborn firms. We have found a positive relationship between the ownership structure of firms and cost of credit, but the results are not statistically significant across our four competition measures. However, we did not find any significant effect of borrower experience, audit, and innovation activities of firms on the cost of credit.

Table 4
Main estimation results

	Dependent variable = Cost of credit			
	CR5	Lerner	H-Stat.	Boone
Competition	-0.123***	-22.056***	-12.637	39.55***
	(-0.039)	(5.310)	(9.680)	(10.058)
Size	-0.0363***	-0.037***	-0.0341***	-0.038***
	(0.011)	(0.010)	(0.011)	(0.01)
Age	0.209***	0.245***	0.246***	0.25***
	(0.011)	(0.056)	(0.058)	(0.057)

Largest_own	0.006	0.010	0.000	0.009
	(0.018)	(0.017)	(0.018)	(0.017)
Experience	-0.031	-0.063	-0.052	-0.069
	(0.048)	(0.047)	(0.049)	(0.0477)
Audit	-0.388	-0.357	-1.091	-0.383
	(1.028)	(0.997)	(1.0156)	(1.002)
Innovation	(-0.573)	0.130	-0.559	-0.15
	(0.948)	(0.941)	(0.968)	(0.938)
Constant	15.143***	12.830***	14.714**	3.66***
	(3.33)	2.405	6.494	(2.03)
R_Squared	0.37	0.41	0.33	0.4

Source: Authors estimation. Dependent variable: cost of credit. Statistical significance at the 10%, 5% and 1% level indicated by *, ** and ***, respectively. Standard errors are in parentheses.

4.3 Empirical Results by Firm Information Opacity

The literature on information-based study suggests that the smaller firms experience the negative effect of competition more than the larger firms. The intuition is that the small firms are more information opaque and hence they need to develop a long-lasting relationship with the banks and by which it is possible to alleviate the information gap between banks and borrowers. However, when there is an intense competition in the market it reduces the bank's incentives to invest in relationship lending because in a competitive environment a borrower can easily switch from one to another bank. Hence, the switching behavior of borrowers is reducing the bank benefits of investment in relationship lending [38] [39]. Based on the above argument, we intend to examine whether market competition does affect the cost of credit of the SMEs due to their information opacity. To test the firm level information opacity, we have segmented the firms according to their sizes (micro, small and medium) and depending on the competition in the market we may expect to find a greater positive association between the cost of credit and micro firms and a lesser impact on the smaller and medium firms. The empirical results are presented in Table 6 and 6.

The results in Table 5 suggest that the effect of bank competition measures on the cost of credit differs according to the firm sizes. Our results suggest that the concentration ratio (Cr) has a positive impact on the cost of credit for micro firms, while a negative effect on the small and medium firms. It could be the fact the in a concentrated market micro firms have fewer alternative options for loans and more importantly, micro firms may not be able to get loans with lower rates not only because of information opacity but also due to their limited capacity in providing collateral or business guarantee [39, 40]. The results for Lerner index suggest that in a concentrated market micro firms provide a lower cost of credit than of the small or medium firms. Hence, the results do support the information

hypothesis that the micro and opaque firms face the negative effect of bank competition more than the medium firms.

The coefficients for the Boone index (Table 6) is positive and statistically significant for the micro and small firms and we did not find any effect of Boone index on the medium firms. These results also support the information hypothesis and we can say that micro and small firms are facing higher loan rates in a competitive market than in a concentrated market. Therefore, information asymmetry can be a significant factor in determining the cost of credit, which is also depended on the nature of the market structure in the Visegrad countries. The results for H statistics is not statistically significant for the micro and medium firms but we have found a negative effect on the medium firms. The results of the H statistics were not significant in our main estimation, but we did find a negative association with the cost of credit, that may imply that when information gap is lower it can lower the cost of credit for the SMEs and higher competition can increase the cost of credit.

Table 5
Estimation by firm opacity (1/2)

Variable	Dependent variable = Cost of credit					
	CR5			Lerner		
	Micro	Small	Medium	Micro	Small	Medium
Competition	0.038*	-0.101***	-0.283***	-31.706***	-28.816***	-9.481
	(0.094)	(0.04)	(0.115)	(12.575)	(6.297)	(16.247)
Size	0.317	-0.030	-0.066***	0.367	-0.045	-0.058***
	(0.241)	(0.050)	(0.025)	(0.588)	(0.047)	(0.028)
Age	0.195	0.041	0.170**	0.319	0.0326	0.237***
	(0.626)	(0.077)	(0.092)	(0.224)	(0.072)	(0.093)
Largest_own	-0.063	0.012	0.025	-0.014	0.018	0.026
	(0.051)	(0.019)	(0.003)	(0.049)	(0.018)	(0.048)
Experience	-0.204	-0.061	0.045	-0.189	-0.065	0.081
	(0.155)	(0.052)	(0.110)	(0.146)	(0.048)	(0.114)
Audit	4.642	-2.852***	0.107	4.550	-3.531***	0.881
	(2.786)	(0.989)	(2.935)	(2.619)	(0.910)	(3.479)
Innovation	2.32	-0.211	3.290	2.673	-0.287	-1.584
	(2.436)	(0.978)	(2.986)	(2.311)	(0.919)	(2.951)
Constant	6.82	16.596***	26.958***	13.760***	19.463***	7.366
	(8.489)	(3.741)	(10.653)	(6.777)	(3.098)	(7.444)
R_Squared	0.33	0.43	0.56	0.44	0.53	0.47

Source: Authors estimation. Dependent variable: cost of credit. Statistical significance at the 10%, 5% and 1% level indicated by *, ** and ***, respectively. Standard errors are in parentheses.

Table 6
Estimation by firm opacity (2/2)

Variable	Dependent variable = Cost of credit					
	H-Stat			Boone		
	Micro	Small	Medium	Micro	Small	Medium
Competition	14.30463	-10.538	-55.040*	39.252*	54.118***	33.73597
	(23.644)	9.713	(29.388)	(22.95)	(12.18)	(29.24)
Size	0.284	-0.019	-0.060***	0.220	-0.050	-0.064***
	(0.622)	0.051	(0.026)	(0.60)	(0.05)	(0.03)
Age	0.307	0.078	0.246***	0.306	0.037	0.243***
	(0.236)	0.077	(0.09)	(0.23)	(0.07)	(0.09)
Largest_own	-0.058	0.009	0.016	-0.034	0.021	0.029
	(0.622)	0.019	(0.05)	(0.049)	(0.018)	(0.047)
Experience	-0.200	-0.084	0.038	-0.191	-0.072	0.060
	(0.154)	0.052	(0.112)	(0.150)	(0.048)	(0.114)
Audit	4.62	-3.298***	1.439	4.821**	-3.491***	1.520
	(2.772)	0.991	(3.247)	(2.696)	(0.914)	(3.446)
Innovation	2.240	-0.316	-3.664	2.566	-0.528	-1.472
	(2.429)	1.004	(2.624)	(2.380)	(0.926)	(2.709)
Constant	-0.453	16.034***	41.624***	4.224	6.972***	2.468
	(16.880)	(6.658)	(20.497)	(7.221)	(2.225)	(6.292)
R_Squared	0.33	0.38	0.53	0.39	0.52	0.49

Source: Authors estimation. Dependent variable: cost of credit. Statistical significance at the 10%, 5% and 1% level indicated by *, ** and ***, respectively. Standard errors are in parentheses.

Conclusions

The bank competition and its implications on financial constraints have been an on-going topic in economic literature. The theory of market power hypothesis suggests that bank competition should relax financial constraints by reducing the interest rate on loans, collateral requirements and enhances access to credit to firms. In contrast, the information hypothesis predicts that bank competition can have a significant negative effect on access to credit and can increase financial constraints due to high information asymmetry between firms and banks. Because, in a competitive market, banks are reluctant to invest in relationship lending technologies and hence increases financial constraints due to a high asymmetric information.

In this paper, we examined the effect of bank competition on the cost of credit by using a sample of SMEs from Visegrad countries (Czech Republic, Slovak Republic, Hungary and Poland). To examine the information problems associated with bank competition and the cost of credit, we have used four proxies of bank

competition: concentration ratio, Lerner index, Panzar-Rosse H statistics and Boone index. The results of our paper suggest that the bank competition is associated with higher cost of credit and thus, our results are aligned with existing literature on information hypothesis of bank competition that high bank competition increases financial constraints on SMEs. Therefore, we reject the view that the bank competition can relax the lending terms and enhances access to finance.

We have also segmented the firms in terms of their sizes as per the intuition that micro and small firms may face higher credit restrictions than of the medium or large firms due to information opacity and bank competition. Our results do support that micro and small firms need to provide higher loan rates than of the medium firms. Hence, we find evidence that the effect of bank competition in increasing the cost of credit is larger on the firms those are financially opaque and need to access loans via relationship lending.

The policymakers may implement policies by which excessive bank competition can be alleviated from the market and that can lower the lending rates in the Visegrad countries. It could also be helpful to remove market barriers so that SMEs can access loans with lower restrictions. The effect of bank competition may be lessened by improving the financial literacy of the borrowers and by doing so the borrowers can prepare better loan proposals and more importantly the borrowers can keep their business accounting records more efficiently. Future research can be done to check how country-specific factors affect the cost of credit on borrowing. Additionally, whether banks are charging higher prices not only for high competition in the market but also is there any factors that force them to charge high prices needs to be investigated.

Acknowledgement

Ashiqur Rahman is grateful to the Internal Grant Agency of FaME TBU No. IGA/FaME/2019/002: The Role of Institutional Environment in Fostering Entrepreneurship" for financial support to carry out this research.

References

[1] R. Alvarez and M. J. Berten, "Banking competition and firm-level financial constraints in Latin America", Emerging Markets Review, Vol. 28, pp. 89-104, 2016

[2] T. Beck, A. Demirguc-Kunt, and V. Maksomovic, V, "Bank competition and access to finance as a growth constraint", Journal of Banking and Finance, Vol. 30, No. (11), pp. 2931-2943, 2004

[3] T. Beck, A. Demirguc-Kunt, and Levine, R, "A New Database on Financial Development and Structure", World Bank Economic Review, Vol. 14, pp. 597-605, 2000

[4] T. Beck, A. Demirguc–Kunt, L. Laeven, and V. Maksimovic, "The

determinants of financing obstacles" Journal of International Money and Finance, Vol. 25, pp. 932-952, 2006

[5] A. Belluchi, A. Borisov, and A. Zazzaro, "Does gender matter in bank–firm relationships? Evidence from small business lending", Journal of Banking and Finance, Vol. 34, pp. 2968-2984, 2010

[6] A. N. Berger, and G. F. Udell, "Small business credit availability and relationship lending: the importance of bank organisational structure", The Economic Journal, Vol. 112, No. 477, pp. 32-53, 2002

[7] D. Besanko, and A.V. Thakor, "Banking deregulation: allocational consequences of relaxing entry barriers", Journal of Banking and Finance, Vol. 16, No. 5, pp. 909-932, 1990

[8] J. Boone, "A new way to measure competition". The Economic Journal, Vol. 118, No. 531, pp. 1245-1261, 2008

[9] A. Boot, and A. Thakor, "Can relationship banking survive competition" The Journal of Finance, Vol. 55, No. 2, pp. 679-713, 2000

[10] S. Carbo-Velverde, D. Humphery, J. Maudos, and P. Molyneus, "Cross-country comparisons of competition and pricing power in banking." Journal of International Money and Finance, Vol. 28, No. 1, 115-134, 2009

[11] S. Claessens, and L. Laeven, "Financial dependence, banking sector competition, and economic growth", Journal of the European Economic Association, Vol. 3, No. 1, 179-207, 2005

[12] M. D. Delis, "Bank competition financial reform and institutions: the importance of being developed" Journal of Development Economics, Vol. 97, No. 2, pp. 450-465, 2012

[13] G. Dell' Ariccia, and R. Marquez, "Lending booms and lending standards" Journal of Finance, Vol. 61, No. 5, pp. 2511-2546, 2013

[14] Dong, Y. and Men, C.: SME financing in emerging markets: firm characteristics, banking structure and institutions. Emerging Markets Finance and Trade, 50 (2014) No. 1, pp. 120-149

[15] F. D. Duarte, A. P. M. Gama, and J. S. Esparanca, "The role of collateral in the credit acquisition process: evidence from SME lending" Journal of Business Finance and Accounting, Vol. 43, No. 5, 2016

[16] E. F. Fama, and M. C. Jensen, "Separation of ownership and control". Journal of Law and Economics, Vol. 26, No. 2, pp. 301-325, 1983

[17] S. M. Fazzari, R. G. Hubbard, and B. C. Petersen, "Financing constraints and corporate investment", NBER Working Papers, 2387, National Bureau of Economic Research. 1988

[18] M. S. Freel, "Are small innovators credit rationed?" Small Business Economics, Vol. 28, pp. 23-35, 2007

[19] Z. Fungacova, A. Shamshur, and L. Weill, "Does bank competition reduce cost of credit? Cross-country evidence from Europe", Journal of Banking and Finance, Vol. 83, pp. 104-220, 2017

[20] C. J. Godlewski, and L. Weill, "Does collateral help mitigate adverse selection? A cross–country analysis" Journal of Financial Services Research, Vol. 40, pp. 49-78, 2011

[21] L. Grunert,and L. Norden, "Bargaining power and information in SME lending"*Small Business Economics*, Vol. 39, pp. 401-417, 2010

[22] C. Hainz, L. Weill, and C. J. Godlewski, "Bank competition and collateral: Theory and evidence"*Journal of Financial Services Research*, Vol. 44 pp. 131-148, 2013

[23] R. Hauswald,and R. Marquez, "Competition and strategic information acquisition in credit markets"*Review of Financial Studies*, Vol. 19, No. 3, pp. 967-1000, 2006

[24] Y. Z. Hsiaoand N. T. Chou, N. T. "Owner characteristics and the cost of Bank loan: Evidence from Small Business", http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2562981, 2015

[25] M. C. Jensen, and W. Meckling, "Theory of the firm, managerial behaviour, agency costs and ownership structure", *Journal of Financial Economics,*Vol. 5, pp. 305-306, 1980

[26] G. Jimenez, V. Salas, and J. Saurina, "*Determinants of collateral, Journal of Financial Economics*" Vol. 81, pp. 255-281, 2006

[27] A. Kljucnikov, and B. Poposko, "Export and its financing in the SME segment, Case study from Slovakia",*Jouranal of Competitiveness*, Vol. 9, No. 1, pp. 20-35, 2017

[28] A. Kljucnikov, J. Belas, L. Kozubikova, and M. Pasekova, "The entrepreneurial perception of SME business environment quality in the Czech Republic"*Journal of Competitiveness*, Vol. 8, No. 1, pp. 66-78, 2016

[29] N. Lee, H. Sameen,and M. Cowling, "Access to finance for innovative SMEs since the financial crisis"*Research Policy,*Vol. 44, pp. 370-380, 2013

[30] F. Leon, "Does bank competition alleviate credit constraints in developing countries?"*Journal of Banking and Finance*, Vol. 57, pp. 130-142, 2015

[31] I. Love, and M. S. Peria, "How bank competition affects firms access to finance", World Bank Economic Review, Vol. 29, No. (3), pp. 413-448, 2015

[32] I. Malafronte, S. Monferra, C. Porzio, and G. Sampagnaro, "Competition, specialization, and bank-firm interaction, what happens in credit crunch periods", Applied Financial Economics, Vol. 24, No. 8, pp. 557-571, 2014

[33] R. Marquez, "Competition, adverse selection, and information dispersion in

the banking industry", Review of Financial Studies, Vol. 1, No. 3, pp. 901-926, 2002

[34] A. Moro, M. Fink, and D. Maresch, "Reduction in information asymmetry and credit access for small and medium sized enterprises", The Journal of Financial Research, Vol. XXXVIII, No. 1, pp. 121-143, 2015

[35] D. Neuberger, and S. Rathke–Doppner, "The role of demographics in small business loan pricing" Small Business Economics, Vol. 44, pp. 411-424, 2015

[36] S. Nguyen, and S. Wolfee, "Determinants of successful access to bank loans by Vietnamese SMEs: new evidence from the red river delta", Journal of Internet Banking and Commerce, Vol. 21, No. 1, pp. 1-23, 2016

[37] J. Panzar, and J. Rosse, "Testing for monopoly equilibrium". The Journal of Industrial Economics, Vol. XXXV, No. 4, pp. 443-453, 1987

[38] M. A. Petersen, and R.G. Rajan, "The effect of competition on lending relationship" The Quarterly Journal of Economics, Vol. 110, No. 2, pp. 407-443, 1995

[39] A. Rahman, J. Belas, T. Kliestik, and L. Tyll, "Collateral requirements for SME loans: empirical evidence from the Visegrad countries". Journal of Business Economics and Management, Vol. 18, No. 4, pp. 650-675, 2017

[40] A. Rahman, Z. Rozsa, and M. Cepel, Trade credit and bank finance-evidence from the Visegrad group. Journal of competitiveness, Vol. 10, No. 3, pp. 132-148, 2018

[41] A. Rahman, Z. Rozsa, L. Kozubikova, and M. Cepel, "Determinants of loan maturity in small business lending" Journal of International Studies, Vol. 10, No. 2, pp. 104-118, 2017

[42] R. M. Ryan, C. M. O'Toole, and F. McCann, "Does bank market power affect SME financing constraints?" Journal of Banking and Finance, Vol. 49, pp. 495-505. 2014

Comparison of the Level of Robotisation in Poland and Selected Countries, including Social and Economic Factors

Miroslaw Smieszek[1], Pawel Dobrzanski[2], Magdalena Dobrzanska[1]

[1]Department of Quantitative Methods, Rzeszow University of Technology, Poland, msmieszk@prz.edu.pl, md@prz.edu.pl

[2]Department of Computer Engineering in Management, Rzeszow University of Technology, Poland, pd@prz.edu.pl

Abstract: Two opposing tendencies are observable in the field of industrial production in economically developed countries. From the perspective of the management staff, for whom economic effects are paramount, the most important are increased work efficiency and reduced cost of production. Such actions are intended to provide an appropriate competitive position for the company. From the standpoint of the employees, aware of their growing position on the labour market and their value, the most important are suitable working conditions and adequate compensation. These opposing tendencies can be reconciled by the widespread automation and robotisation of the production processes. This requires substantial investment and, considering the growing costs of labour, can provide an increase in efficiency and a reduction in the costs of production. Growth in robotisation and automation are also necessary due to the shrinking labour resources. This paper is intended to analyse the level of robotisation in selected countries, and to investigate the relations between labour costs and the degree of robot utilisation. The final part of the paper characterises the condition of Polish industry from the perspective of robotisation, based on a more in-depth analysis of selected factors. On this basis, the directions are outlined for the necessary changes to achieve further growth and approach the levels observed in European Union countries.

Keywords: robotisation; industrial production; efficiency; labour costs; labour market

1 Introduction

Modern manufacturing companies strive to effectively manage their resources [17, 20]. Within its operations, a company has to reconcile many opposing tendencies. Employees expect growing compensation and improving working conditions, while the employer desires lower production costs and improved efficiency. Robotisation is an element that can contribute to satisfying the needs of both

employees and employers. It provides businesses with new opportunities [1, 11, 24], including increased production capacity [22], reduced downtime, and improved efficiency, quality and work safety [12, 7].

Industrial robots and manipulators [3] are some of the ways to automate the production and support processes [10], by performing specific actions without human participation. Other than robots and manipulators, this category also includes automatic production lines, treatment centres, numerically controlled machine tools, and automatic welding, paint shop and assembly machines.

The progress of science and robotics-related technologies means that robots are:
- able to replace humans in hazardous or inconvenient working conditions. The work can be cyclical or irregular;
- programmable, where after completing one task the robot can be reprogrammed and retooled to perform a completely different task;
- controllable by computers and connected to other computer systems to achieve computer-integrated production.

The working environment is one of the characteristics that must be considered [9, 6] when selecting an appropriate robot. In production-related fields, robots are most commonly used for transporting and handling the materials processed [14], production operations, assembly [15] and control. The use of robots in industry must be technically and economically viable for the industry [23]. Transporting materials involves the robot collecting parts from one location and moving them to another. An example of this type of application is 'pick and place', while another, more complex one is palleting [14]. In the latter case, robots have to take parts or other objects from one location and place them on a pallet or in another multi-location container.

For processing-related operations, industrial robots may be employed in heat treatment, welding, spray painting, drilling, marking out, laser cutting, riveting, grinding, brushing, and water jet cutting operations [13].

The final type of industrial robot application is assembly and control [15]. This is a combination of the two previous types, and involves both material handling and transport. Typical assembly and control applications include operations related to both materials and tools. Assembly and control are traditionally work-intensive, tedious and highly repetitive operations. For this reason, they are increasingly performed using robots.

The sectors that most commonly employ [8; 23] robotised solutions are the automotive, electronic, electromechanical, precision, machine construction, and rubber and plastic product industries. Companies belonging to these sectors must be aware that the failure to use robots will in their case entail the loss of competitiveness and effectiveness, and consequently loss of market share.

This paper aims to analyse the level of robotisation in Poland and around the world. Factors affecting the robotisation of production processes are discussed,

and the directions and requirements to be met by the process of robotisation are outlined.

2 Object and Methodology of the Study

The issues addressed in this paper are related to numerous fields of scientific research. The issue of robotisation development can be analysed from the perspective of economics, work organisation and safety, technical sciences, and ergonomics.

There are numerous factors responsible for the development of robotisation. Among the most important are the technical, economic and organisational factors. The technical factors include:

- development of precision mechanics and control systems;
- development of new material technologies;
- necessity to ensure high and uniform product quality;
- development of compact robots;
- growing demand for processes that require high precision under harmful and hazardous conditions, or on high-weight objects of complex shapes.

The economic factors are related to the need to:

- reduce operating costs;
- improve competitiveness;
- reduce energy costs;
- improve efficiency;
- reduce costs caused by short production cycles.

Within the organisation factors, the following can be defined:

- lack of workers for simple and burdensome physical jobs;
- shrinking working age population due to society ageing;
- increased work safety standards;
- developing consumer markets requiring frequent production changes and quick expansion of production capacity.

This paper first analyses the level of robotisation in selected countries. Due to Poland's location, particular attention is paid to European Union (EU) countries. As was mentioned earlier, the level and development of robotisation depend on numerous factors, with the economic factors, including labour profit and costs, being among the more notable. It was decided to use the available statistical data to study the relation between labour costs and robot saturation levels. Studying this relation should enable the determination of the position of robotisation in Poland relative to selected EU countries and outlining the potential development perspectives. The final factor taken into account, whose importance in economically developed countries grows with each year, is society ageing.

This process leads to manpower shortages and therefore places greater pressure on the development of robotisation. This phenomenon has already been observed in many developed countries, and a similar tendency is also visible in Poland.

3 Analysis of the Level of Robotisation in Selected Countries

3.1 Level of Robotisation around the World

According to International Federation of Robotics data [8], 1.8 million industrial robots were in use worldwide by the end of 2016. In 2016, the global population of active robots increased by 12%. Most of these robots were in Asian countries and Australia, where the rate of robot sales is the highest. Companies in this part of the world purchased as many as 160 600 robots in 2016, which means an increase of 19% over the previous year. In terms of sales dynamics, Europe as a whole remained not only behind Asia, but also the entire American continent. For Europe, the yearly growth in sales was 12%, with most robot sales in 2016 occurring in Germany and Italy. Robot numbers are also growing in Central and Eastern Europe, in particular in the Czech Republic, more than in Poland while having less than a third of the population. Figure 1 shows [8] the robot purchase data of the 15 largest global consumers in 2016.

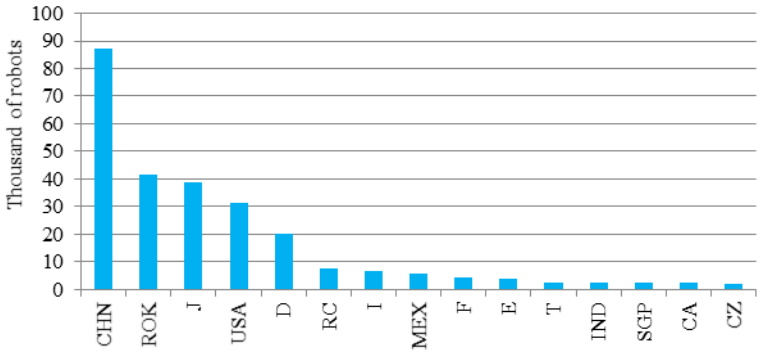


Figure 1
Robot purchases by the largest consumers in 2016 [8]

It is assumed that in 2020, 3 million robots will be in use in industry. Figure 2 shows the numbers of robots used by year, and a prediction of their use up to 2020.

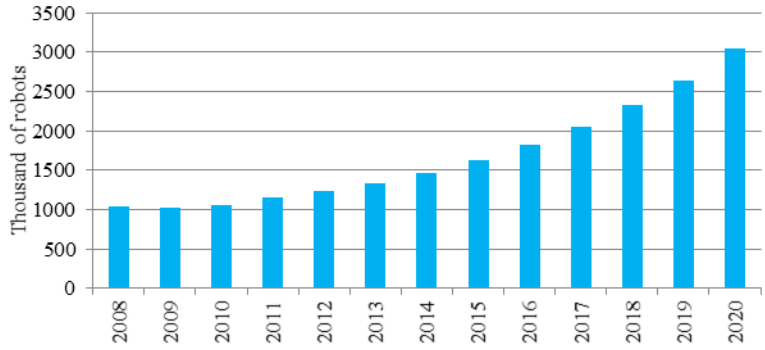


Figure 2
Numbers of robots in the economy [8]

The predicted growth in the number of robots requires the ensuring of an adequate supply. The global robot industry is prepared for this challenge, and has expanded its production capacity. The supply of robots for industrial applications, with a prediction for the years 2018-2020, is shown in Figure 3.

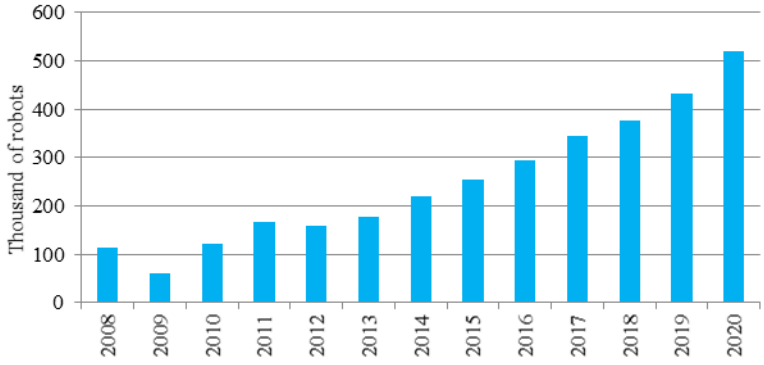


Figure 3
Supply of robots for industrial applications [8]

Due to the varying sizes of world economies, it is not useful to directly compare the numbers of installed robots. Figure 4 shows the index of robotisation density in selected countries around the world, with particular attention on Europe, calculated for the year 2016. The index is a very important parameter used to compare the degree of national robotisation. Unlike the number of installed industrial robots, the density index is a relative value and takes into account the differences in the sizes of economies. It shows the number of active industrial robots per 10 thousand people employed in industry.

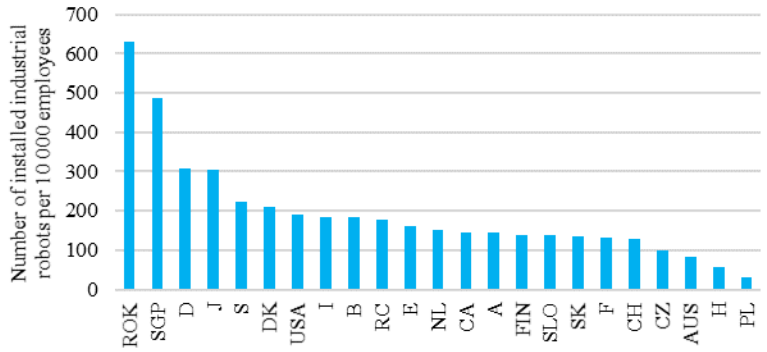


Figure 4

Robotisation density in different economies in 2016 [8]

For Poland, the value of this index was approximately 32, less than a third of the value for EU as a whole (99) and less than half the global index (74). An analysis of the data in Figure 4 indicates that Poland is at the bottom end of the European league. It is behind not only highly developed countries, but also countries at a similar level of development, such as Slovakia (135) and Hungary (57). While the robotisation density index showed a growing trend for the years 2003-2015, its growth was low compared to all of Europe or Central and Eastern European countries. The level of robotisation is also related to efficiency per worker. Figure 5 shows the labour productivity index, measured using the value of national product per worker in EU countries [16], taking into account the purchasing power standard. This chart does not differ significantly from the chart shown in Figure 4, considering only European countries.

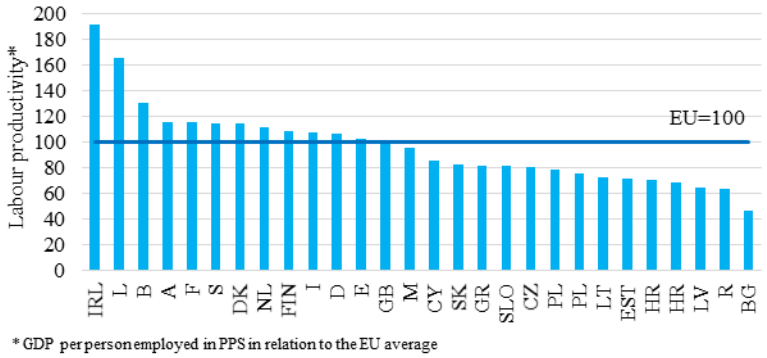


Figure 5

Labour productivity in European countries in 2016 relative to the mean [16]

Another important factor in evaluating the level of robotisation are the labour costs. Among EU countries, Poland is characterised by one of the lower labour cost values, at approx. $8/h. A summary of these costs for EU countries is shown in Figure 6.

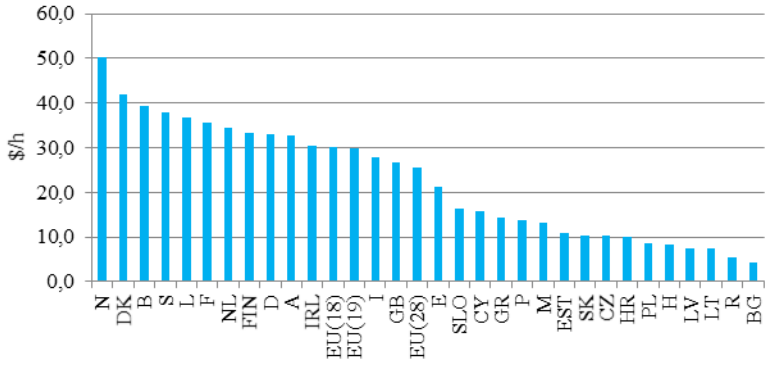


Figure 6
Costs of labour in individual EU countries in 2016 [4]

Both charts shown in Figs. 5 and 6 are similar and indicate that the GDP index and labour costs are highly related. When the costs of labour are compared with robotisation density in EU countries, an interesting relation emerges, as shown in Figure 7.

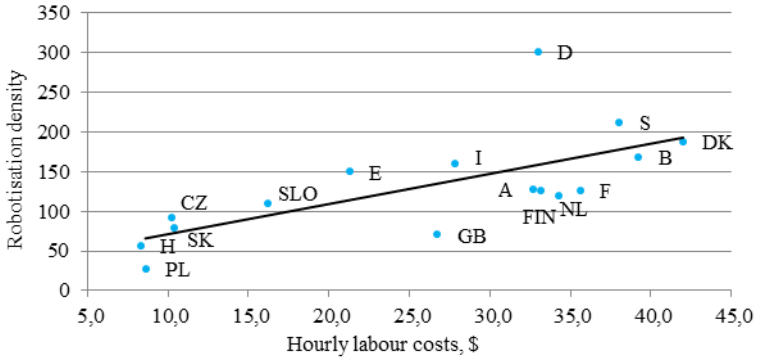


Figure 7
Robotisation density vs hourly labour cost

When analysing the above chart, it can be assumed that other than three countries - Poland, Great Britain and Germany - the relation between robotisation density and hourly labour cost is nearly linear.

3.2 Level of Robotisation in Poland

In 2014, robots and manipulators constituted 11.5% of the means of production process automation in industrial companies. The most robotised companies in Poland were those that manufactured means of transport [18]. They possessed more than 40% of the robots and manipulators in use in the Polish economy.

Another group of companies, with 37.2% of the robots and manipulators, involved companies related to the manufacture of metal products, electronic and optical products, electric devices, as well machines and devices. The number of industrial robots in Polish industry is growing systematically (Fig. 8). In the years 2010-2015, the number of industrial robots and manipulators increased by more than 50%.

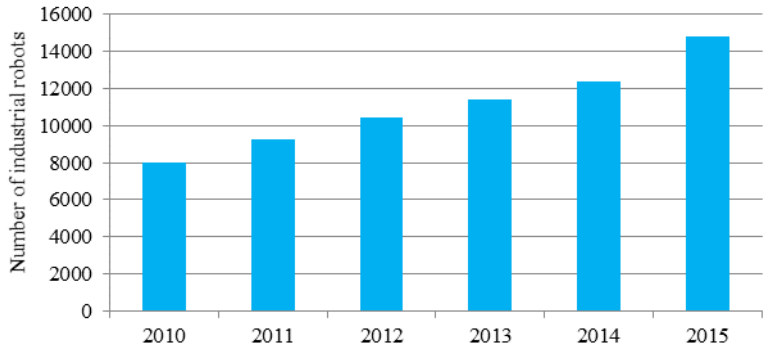


Figure 8
Number of industrial robots and manipulators in Poland (data concerning entities where the employment level exceeded 49 people) [18]

According to the International Federation of Robotics data [8], 692 robots were delivered to the Polish market in 2013, 1267 robots in 2014, and a record of 1795 in 2015. This is an increase of more than 259% over the whole period. In 2015, the number of robots and manipulators in industry was 14 847.

Considering the number of people employed in industry, most robots and manipulators were used by companies having more than 250 employees (Tab. 1). In 2014 a total of 9396 industrial robots and manipulators were used by 480 companies. In companies with 50-249 employees, the number was 2957, used in 600 companies. In companies with 10-49 employees, the number in 2014 was 699, used in 284 companies.

Table 1
Robotisation in industry relative to the number of company employees in 2014 [18]

	Company size (number of employees)		
Number of employees	10-49	50-249	250 or more
Number of robots and manipulators in industry	699	2 957	9 396
Number of companies that own robots or manipulators	284	600	480

The distribution of robots and manipulators in individual voivodeships of Poland is not uniform. The disproportions in distribution are shown in Figure 9.

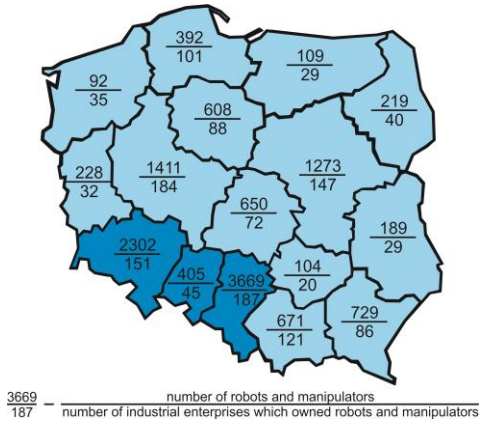


Figure 9
Distribution of robots and manipulators in individual voivodeships of Poland [18]

The descriptions for individual voivodeships concern the numbers of robots and manipulators (data above the horizontal line) and numbers of companies using these resources (below the line). Individual voivodeships differ not only in the numbers of installed robots, but also the degree of their utilisation in individual companies, as well as revenue. Data on the per capita GDP for individual voivodeships are shown in Figure 10 [5].

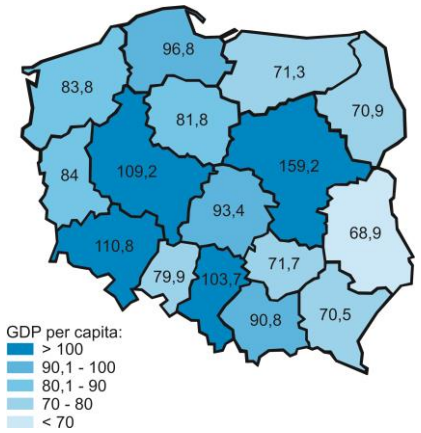


Figure 10
Data on per capita GDP for individual voivodeships [5]

The data indicate a relation between labour costs and thus income, and robotisation density in selected EU countries. In the case of Poland, at the

voivodeship level, a similar relation between the numbers of installed robots and income can be observed by comparing Figures 9 and 10.

The four voivodeships with the highest income exceeding national average, also have the highest numbers of installed robots and means of production automation. Each voivodeship in Poland has a different population, however. Given the above, the most beneficial would be to compare the number of robots per million inhabitants with the income achieved in the voivodeship. This comparison is shown in Figure 11.

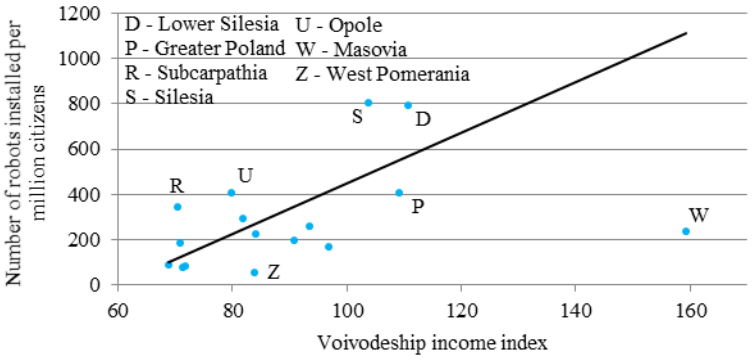


Figure 11
Relation between the number of robots installed per million inhabitants and the voivodeship income index

The relation shown in Figure 11 with a continuous line applies to 15 voivodeships. When compiling this chart, the wealthiest voivodeship Masovia, was not taken into account. The voivodeships include Warsaw, the capital of Poland, with its population of almost two million. The high income achieved by this voivodeship is to a great degree generated by governmental and financial institutions and banks with a national and international reach.

4 Potential Factors Forcing the Growth of Robotisation in Poland

4.1 Economic Development of Poland

Poland's per capita domestic product is significantly lower than the EU mean, and does not exceed 70%. Improving this index requires intense economic growth, exceeding the rate of economic growth of wealthy EU countries. The changes of income in Poland, defined as per capita GDP, is shown in Figure 12 [4].

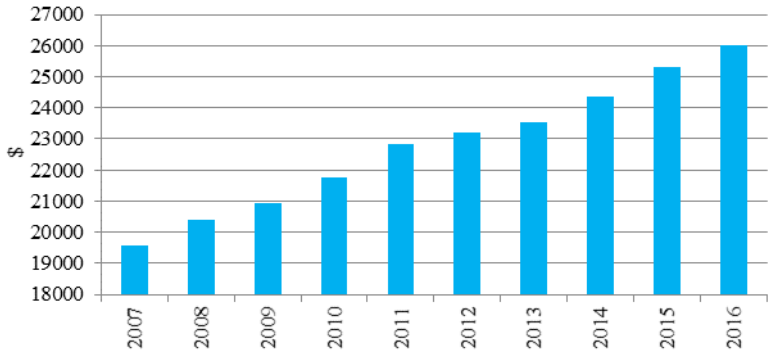


Figure 12
Poland - GDP per capita PPP [4]

One driver for this growth is industrial production. Figure 13 shows industrial production for selected years against gross national product.

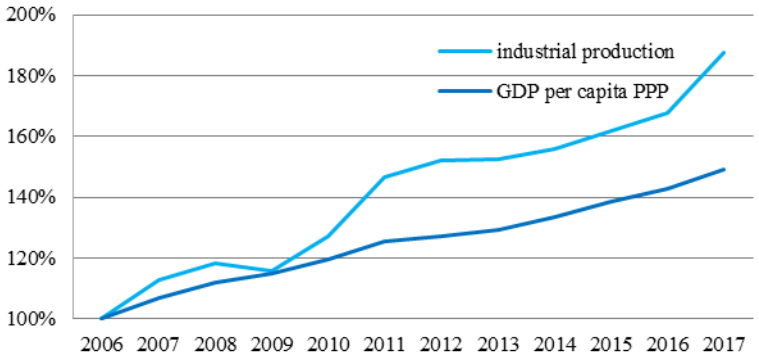


Figure 13
GDP and industrial production costs in relation to the base year [5]

An increase in GDP leads to an increase in wages. The wages for industry are shown in Figure 14. In order to ensure industry competitiveness, the increase in wages must not increase the unit labour costs. Keeping this cost at a constant level or even reducing it is possible by increasing labour productivity. See Figure 15 for productivity in selected years.

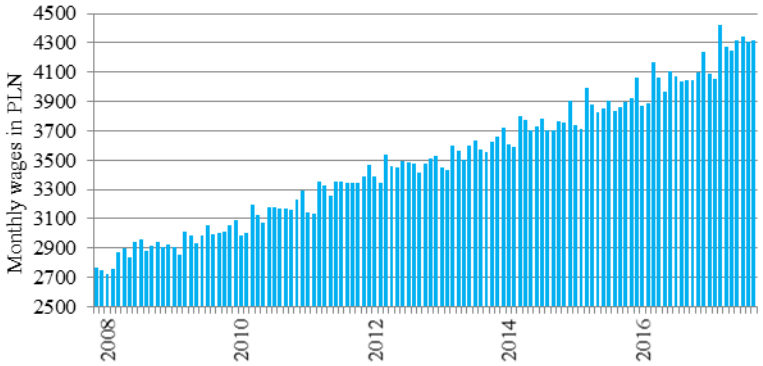


Figure 14
Poland - wages in industry [4]

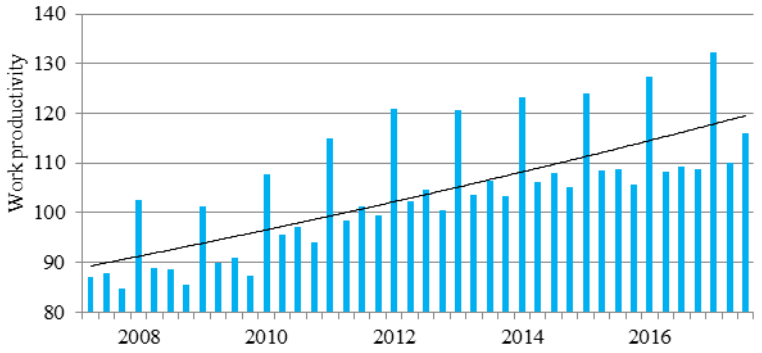


Figure 15
Poland - work productivity [4]

The unit labour costs taking into account wages and labour efficiency are shown in Figure 16. Slowing down the increase in these costs while maintaining the growth of wages was largely possible thanks to increasing the efficiency, resulting from better work organisation and improved means of production, such as industrial robots and other devices enabling greater automation of production processes.

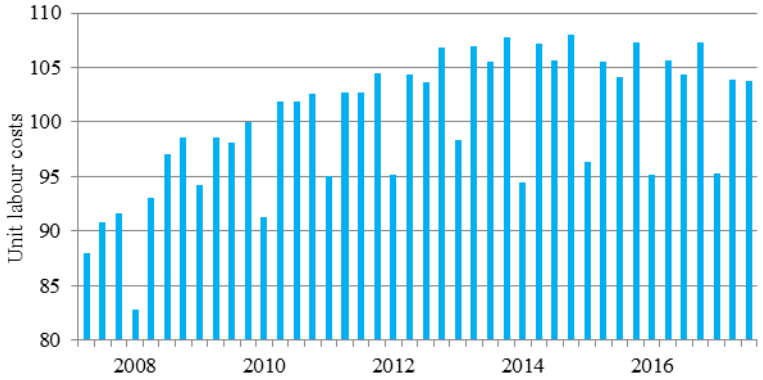


Figure 16
Poland - unit labour costs [4]

The tendency observed in Poland is consistent with global tendencies. In Poland, the dynamics of increasing robot numbers are significantly higher than the wage growth dynamics. This trend is shown in Figure 17. In the years 2010-2015, a significant increase was achieved in the use of robots and means of automation at a much lower increase in wages.

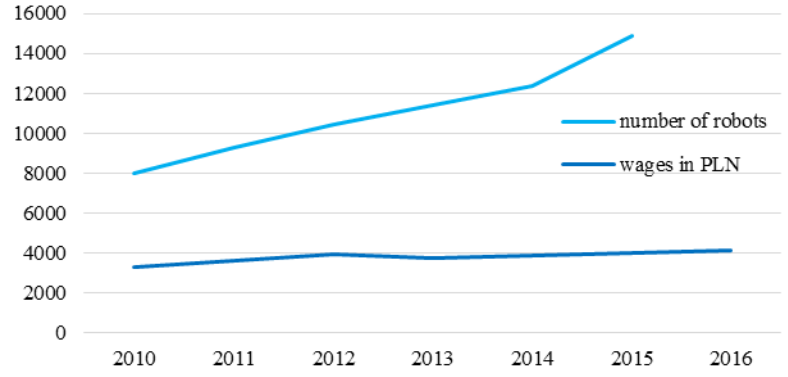


Figure 17
Growth of robot number and wage in years

According to Boston Consulting Group [19], the implementation of automated solutions will result in a global increase in productivity by approx. 10-30% by 2025. This will also result in lower costs of labour. The greatest reduction will be observed in South Korea (33%), Japan (25%) and Canada (24%), while in Poland it will only be 13%.

4.1 Effects of Society Ageing on the Labour Market

Until now it was profitable for Polish companies to employ new workers to achieve growth. Currently, such growth is encountering two fairly significant barriers. The first is labour costs, the second is difficulties in obtaining new employees. Robotisation is, therefore, an answer to both these issues. For several years an inexpensive, easily available and well-educated workforce was used to create an attractive business environment in Poland. However, in the coming decade, these conditions may change. The predictions [5] for 2050 show a great reduction in the size of the working age population. This prediction is shown in Figure 18.

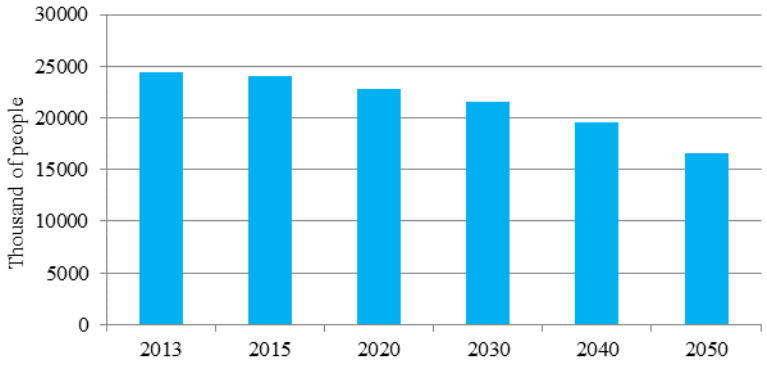


Figure 18
Numbers of working age population [5]

It is estimated that in the immediate future, until 2025, the size of the working age population will dwindle by as much as 2.6 million people. Some of this loss can be replaced by immigrants and some by transfers between individual industries. However, these effects are unlikely to fully satisfy the needs of the labour market. To ensure an adequate level of production and growth, a sufficient number of robots will have to be installed. This is an important factor indicating a need for companies to invest in the purchase of robots. With the aging of society, the ratio of professionally inactive people to working age people also changes. In Poland, the ratio of these two groups is approximately 0.75 today, and will increase to 1 within 15 years, and over the following fifteen years will increase to 1.4. All simple, burdensome and tedious jobs must therefore become automated. An aging society will not be interested in performing simple, burdensome and tedious jobs.

Conclusions

The development of robotisation processes is a global trend. The presence of this tendency is also visible in Poland. In many cases, robotisation contributes to lowering the costs of production, increasing the flexibility of production lines, and

enabling humans to be replaced by robots in repetitive and frequently hazardous jobs. However, this process frequently requires fairly significant financial investments, and especially a change in how a company is managed. Thanks to the progress in technology, the capabilities of robots improve with every passing year, and thus forms one of the cornerstones of economic growth. Compared to selected countries around the world and the EU, the level of robotisation in Poland is low. Poland is exceeded in this field by its immediate neighbours in the EU. In the group of countries with similar labour costs, such as Hungary, Slovakia and Czech Republic, Poland is characterised by the lowest level of robotisation. Due to limited manpower, Poland's further development without investments in robotisation becomes problematic, and therefore investments intended to develop robotisation and automation become a necessity. Despite the high dynamics of growth in industrial robot purchase and installation, the situation for the next few years is unlikely to improve significantly, thus reaching the average EU level will take many years. The process is inevitable, however, as without it the competitiveness level of Polish companies will degrade and, in many cases, manpower shortages caused by society ageing will result in production reduction.

References

[1] Day C.-P.: Robotics in Industry—Their Role in Intelligent Manufacturing, Engineering 4 (2018), pp. 440-445

[2] Geren N, Redford A.: Cost and performance analysis of a robotic rework cel, Int. J. Production Economics 58 (1999), pp. 159-172

[3] Global Industrial Robotics Market Overview Infographic https://www.robotics.org/blog-article.cfm/

[4] https://tradingeconomics.com

[5] http://swaid.stat.gov.pl

[6] Iglesias I., Sebastián M. A., Ares J. E.: Overview of the state of robotic machining: Current situation and future potential, Procedia Engineering 132 (2015), pp. 911-917

[7] Industries and Economies Leading the Robotics Revolution. The Boston Consulting Group https://www.bcg.com/publications/2015/lean-manufacturing-innovation-industries-economies-leading-robotics-revolution.aspx

[8] International Federation of Robotics https://ifr.org/

[9] Kamble S. S., Gunasekaran A., Sharma R.: Analysis of the driving and dependence power of barriers to adopt industry 4.0 in Indian manufacturing industry, Computers in Industry 101 (2018), pp. 10-119

[10] Khana Z. H., Khalidb A., Iqbalc J.: Towards realizing robotic potential in future intelligent food manufacturing systems, Innovative Food Science and Emerging Technologies 48 (2018), pp. 11-24

[11] Kosea T., Sakatab I.: Identifying technology convergence in the field of robotics research, Technological Forecasting & Social Change, Article in Press

[12] Landscheidta S., Kansb M.: Method for Assessing the Total Cost of Ownership of Industrial Robots, Procedia CIRP, Volume 57, 2016, pp. 746-751, DOI: doi.org/10.1016/j.procir.2016.11.129

[13] Landscheidta S., Kansb M., Winrothc M. Westerd H.: The future of industrial robot business: Product or performance based? Procedia Manufacturing, Volume 25, 2018, pp. 495-502, DOI: doi.org/10.1016/j.promfg.2018.06.125

[14] Mahalik N. P.: Processing and packaging automation systems: a review, Sens. & Instrumen. Food Qual. (2009) 3, pp. 12-25, DOI: 10.1007/s11694-009-9076-2

[15] Michels A. S., Lopes T. C., Stall Sikora C. G., Magatão L.: The Robotic Assembly Line Design (RALD) problem: Model and case studies with practical extensions, Computers & Industrial Engineering 120 (2018), pp. 320-333

[16] Poland in the European Union. Warsaw 23.07.2018. http://stat.gov.pl/en/topics/other-studies/other-aggregated-studies/poland-in-the-european-union-2018,10,12.html

[17] Povolná, L., Švarcová, J. (2017) The Macroeconomic Context of Investments in the Field of Machine Tools in the Czech Republic. Journal of Competitiveness, Vol. 9, Issue 2, pp. 110-122

[18] Science and technology in 2014. Warsaw 2015. http://stat.gov.pl/obszary-tematyczne/nauka-i-technika-spoleczenstwo-informacyjne/

[19] Takeoff in Robotics Will Power the Next Productivity Surge in Manufacturing. The Boston Consulting Group https://www.bcg.com/d/press/10feb2015-robotics-power-productivity-surge-manufacturing-838

[20] Tuček, D. (2016) Process Segmentation Typology in Czech Companies. Journal of Competitiveness, Vol. 8, Issue 1, pp. 79-94

[21] Ungerman, O., Dedkova, J., Gurinova, K.: The Impact of Marketing Innovation on the Competitiveness of Enterprises in the Context of Industry 4.0. Journal of Competitiveness, Vol. 10, Issue 2, pp. 132-148, June 2018

[22] Zhang J. Fang X., Qi L.: Sensitivity-analysis based method in single-robot cells cost-effective design and optimization, Roboticsand Computer-Integrated Manufacturing 38 (2016) pp. 9-15

[23] Zwicker C., Hammerstingl V., Possin C, Reinhart G.: Life Cycle Cost Estimation Of Robot Systems in an Early Production Planning Phase, Procedia CIRP 44 (2016) pp. 322-327

A Hybrid Time Series Forecasting Model for Disturbance Storm Time Index using a Competitive Brain Emotional Neural Network and Neo-Fuzzy Neurons

Umar Farooq[1,3], Jason Gu[1], Valentina E. Balas[2], Ghulam Abbas[4], Muhammad Usman Asad[1], Marius M. Balas[2]

[1]Department of Electrical and Computer Engineering Dalhousie University, Halifax, N.S. B3H 4R2, Canada

[2]Department of Automatics and Applied Software, University of "Aurel Vlaicu" Arad, Romania

[3]Department of Electrical Engineering, University of The Punjab, Quaid-e-Azam Campus, Lahore, 54590 Pakistan

[4]Department of Electrical Engineering, The University of Lahore, Pakistan

umar.farooq@dal.ca, jason.gu@dal.ca, valentina.balas@uav.ro, ghulam.abbas@ee.uol.edu.pk, mh549096@dal.ca, marius.balas@uav.ro

Abstract: The Disturbance storm time (Dst) index is an important indicator of the occurrence of geomagnetic storms, which can damage communication and power systems, as well as, affect Astronauts performance. Such potential consequences of this fatal event has challenged researchers to develop Dst predictors, with some success. This paper presents the design of a computationally fast, neuro-fuzzy network to forecast Dst activity. The proposed network combines a class of emotional neural networks with neo-fuzzy neurons and is named, Neo-fuzzy integrated Competitive Brain Emotional Learning (NFCBEL) network. Equipped with five competing units, the hybrid model accepts only the past two samples of Dst time series, to predict future values. The model has been tested in the MATLAB programming environment and has been found to offer superior performance, as compared to other state-of-the-art Dst predictors.

Keywords: geomagnetic storms; Dst time series; emotional neural networks; neo-fuzzy neurons; MATLAB

1 Introduction

Geomagnetic storms are the result of interactions between the solar winds and the earth's magnetic field. During this interaction, energy is transferred from the magnetic field carried by solar winds to the Earth's magnetosphere which gives rise to increased electric currents inside the magnetosphere and ionosphere. This enhanced electrical activity further results in the modification of the magnetosphere's magnetic field and can lead to geomagnetic storms. These storms can cause disruptions in electrical power systems, radio communication systems, satellites and navigation systems [1].

The presence of the geomagnetic storms is dictated by the disturbance storm time index which is an estimate of the variation in the horizontal component of Earth's magnetic field and is measured with the help of magnetometers placed at four different stations near the Earth's equator. No geomagnetic storm is reported if these measurements fall between +20 to -20 nT while the storm is classified as moderate, intense and super if these measurements lie in the range of -50 to -100 nT, -100 to -250 nT and lower than -250 nT respectively [2]. Two intense geomagnetic storms have already hit the earth in 1859 and 1989 besides other intense and moderate storms. The Dst index was measured to be roughly -1760 nT during an 1859 storm, named 'Carrington Event', which caused the disruption of telegraph services across the United States and Europe [3], while it was estimated to be -589 nT during an 1989 storm, which resulted in the collapse of Hydro-Quebec power grid, leaving six million people without power for nine hours [4]. To prevent such disaster events, it is important to have a good prediction model of the Dst index.

Several studies have reported the one step ahead prediction model for the Dst index, based on the differential equations and intelligent networks [5]. The first mathematical model appeared in [6], which describes the time variation of Dst index having a constant decay rate through a first order differential equation which is driven by a linear function of the interplanetary electric field's dawn-dusk component. This earlier model was modified [7], by adding the solar wind pressure to the source term and reducing the decay rate, to predict high geomagnetic activities. The constant decay rate was made variable, in another study [8], to predict the low and high geomagnetic activities.

Amongst the intelligent networks, recurrent type neural networks have been widely studied to predict the Dst index because of their capability of implicit implementation of time dynamics [9]. In [10], a recurrent type neural network is presented to predict the Dst index using the interplanetary magnetic field (IMF) and the plasma parameters of the solar wind. However, when the plasma parameters of the solar wind are either not available or inaccurately measured by the relevant instruments, the performance of the prediction model is degraded.

To overcome this drawback, another recurrent neural network is proposed in [11] which can provide better predictions of the Dst index using only the IMF data.

More recently, brain emotional learning (BEL) networks are explored, to forecast the geomagnetic activity indices [12]-[16]. These neural network models are based on the mechanism employed by the limbic system of the mammalian brain in processing the stimuli and differ only in the generation of reinforcement signal during learning. The emotional networks in [12], [13] used a specially designed reward signal to predict the geomagnetic activity index. Although the network predicted the peak points well, the performance of the network degraded for predicting the valley points which is the case for Dst index, where the valley points are an indicator of the strength of geomagnetic storms. A modification to this network is presented in [14], [15] where the reinforcement signal is set to be the target value and a decay rate is introduced in the learning rules, thereby, enabling the network to learn through the input-target samples in a supervised fashion. The modified network is named as ADBEL and has shown superior performance in predicting the valley points in Dst profile. A fuzzy model of ADBEL network is presented afterwards [16] where the weights of the network are kept as fuzzy numbers and the predicted value is generated through the defuzzification process.

Neuro-fuzzy networks form another class of intelligent algorithms which have been studied for the prediction of Dst index. A locally linear neuro-fuzzy model combined with a recursive locally linear model tree algorithm is proposed in [17] to predict the Dst index along with other space weather indices. The recursive modification allows the online adjustment of neuro-fuzzy parameters so as to cope with the time varying nature of geomagnetic activity indices. In another study, authors have presented a novel neuro-fuzzy model named NFADBEL network which has shown superior performance as compared to BEL network in forecasting the chaotic Dst index along with some other benchmark time series [18]. The network works in an online fashion, to predict the next value based on the past four occurrences of the time series data.

The present work also deals with the design of a neuro-fuzzy network for the prediction of Dst index by combining the recently proposed competitive brain emotional learning (CBEL) network [19] with neo-fuzzy neurons [20]. To the best of the authors' knowledge, such a hybrid model is proposed for the first time. The proposed network is named as NFCBEL and uses only past two samples of the Dst time series to predict its future value. It employs five competitive units each comprising of a BEL network fused with neo-fuzzy (NF) neurons. The fusion of neo-fuzzy neurons takes place in the orbitofrontal cortex which is the knowledgeable part of the BEL network. The proposed network is trained on the Dst dataset acquired from [15] and the test results reveal its superior performance as compared to some of the current Dst predictors, in terms of the normalized mean square error (NMSE) criterion.

This paper is structured as follows: Brain emotional learning and neo-fuzzy networks are briefly reviewed in Sections 2 and 3 respectively, the proposed network is described in Section 4 and results are presented in Section 5 followed by conclusions.

2 Brain Emotional Learning Network

First proposed by C. Lucas et. al [21], brain emotional learning network is the computational model of the emotional processing in the mammalian brain based on the work of Moren and Balkenius. The model generates a response to the stimulus based on the interaction of two parts of the brain namely orbitofrontal cortex and amygdala. Amygdala quickly responds to the stimulus owing to its close proximity to thalamus and sensory cortex which are the carriers for stimulus. The response generated by amygdala is then inhibited by orbitofrontal cortex based on the context. During this interaction, reward signals are generated and the weights of the network are adjusted. The generation of these reward signals has been the point of discussion in the literature. Further, the computational model developed by Lucas cannot be adjusted by pattern-target samples. To address this limitation, a supervised version of brain emotional learning network is proposed by E. Lotfi et. al which has been shown to perform well for time series prediction and pattern classification tasks [14] [15]. A competitive version of the network is also proposed in which a particular block of the network is triggered to produce output response based on the proximity of stimulus to that block [19]. One such block of the competitive brain emotional learning network is shown in Figure 1.

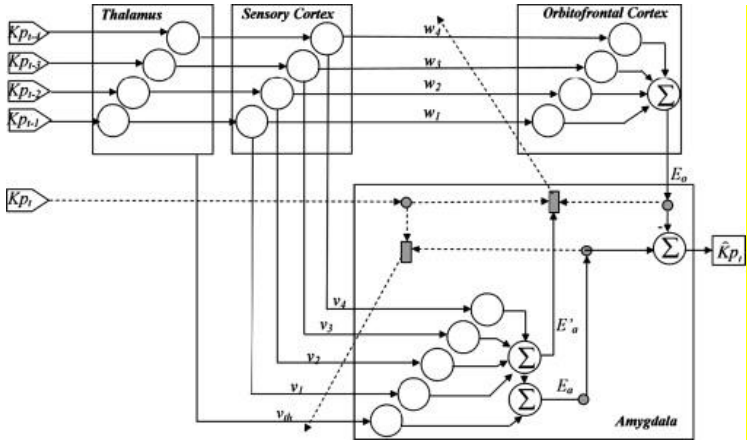


Figure 1
Brain emotional learning network for DST prediction [15]

3 Neo-Fuzzy Network

Neo-fuzzy neurons are characterized by their transparent structure, simplicity and effectiveness in time series prediction, classification and control tasks. Constructed from triangular membership functions (μ_{ij}), a neo-fuzzy neuron has a nonlinear synapse and can map the input-output data by adjusting its weights through gradient descent technique. The network constructed from neo-fuzzy neurons has also been shown to possess generalization ability. One such network is shown in Figure 2, where past samples of the time series are used to perform one step ahead prediction task.

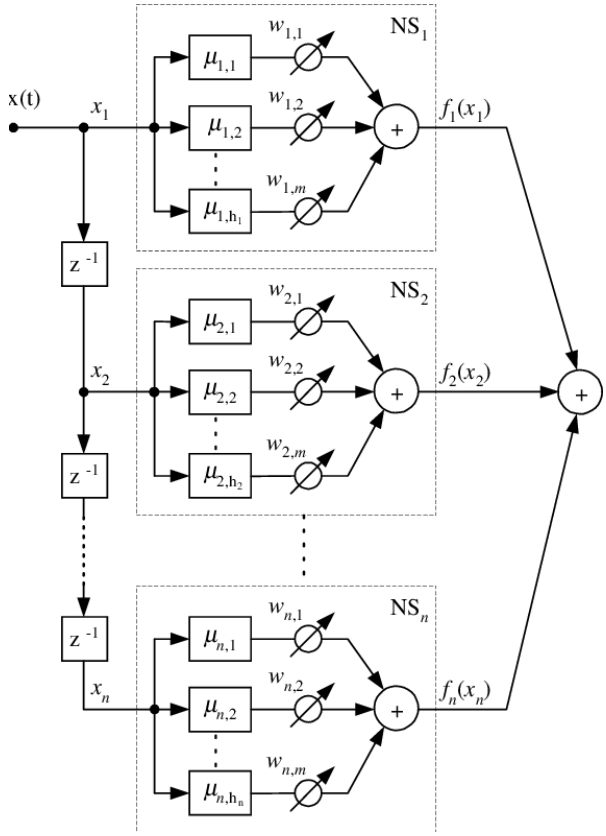


Figure 2
Neo-fuzzy network [20]

4 Proposed Hybrid Dst Predictor

The proposed Dst predictor is constructed from competitive brain emotional learning neural network and neo-fuzzy neurons. It is a single layer network with two inputs and one output as shown in Figure 3. The inputs are the past two samples in Dst time series while output is the one hour ahead predicted Dst value. The network has five competing units and only one of them is active at any time. The activation of a particular unit is based on the Euclidean distance between the input sample and the weights associated with that competing unit. The unit offering the least distance to the input sample is selected:

$$i* = \arg\min_{i} \left(\left\| c_i - y_t \right\| \right), i = 1,2,3,4,5 \tag{1}$$

Where $c_i = \begin{pmatrix} c_{1i} & c_{2i} \end{pmatrix}^T$ is the weight vector associated with i^{th} unit, $y_t = \begin{pmatrix} y_{t-1} & y_{t-2} \end{pmatrix}^T$ is the input sample containing past two Dst values and $i*$ is the winner unit. The output from the winner unit is the predicted value of Dst which can be given as:

$$y(t) = f\left(v_i^* y_{te} - w_i^* h_i^* \right) \tag{2}$$

Where f is the log-sigmoid function and other winner-unit's entries in (2) are given as:

$$\begin{aligned} v_i^* &= \begin{pmatrix} v_{1i}^* & v_{2i}^* & v_{3i}^* \end{pmatrix} \\ w_i^* &= \begin{pmatrix} w_{1i}^* & w_{2i}^* & w_{3i}^* & w_{4i}^* & w_{5i}^* & w_{6i}^* \end{pmatrix} \\ y_{te} &= \begin{pmatrix} y_{t-1} & y_{t-2} & \max\left(y_{t-1} & y_{t-2} \right) \end{pmatrix}^T \\ h_i^* &= \begin{pmatrix} h_{1i}^* & h_{2i}^* & h_{3i}^* & h_{4i}^* & h_{5i}^* & h_{6i}^* \end{pmatrix}^T \end{aligned} \tag{3}$$

Where v_i^* and w_i^* are the weights associated with the amygdala and orbitofrontal cortex sections of the winner unit respectively while y_{te} and h_i^* are the expanded inputs to the amygdala and orbitofrontal cortex sections respectively. The first three entries of h_i^* are computed as:

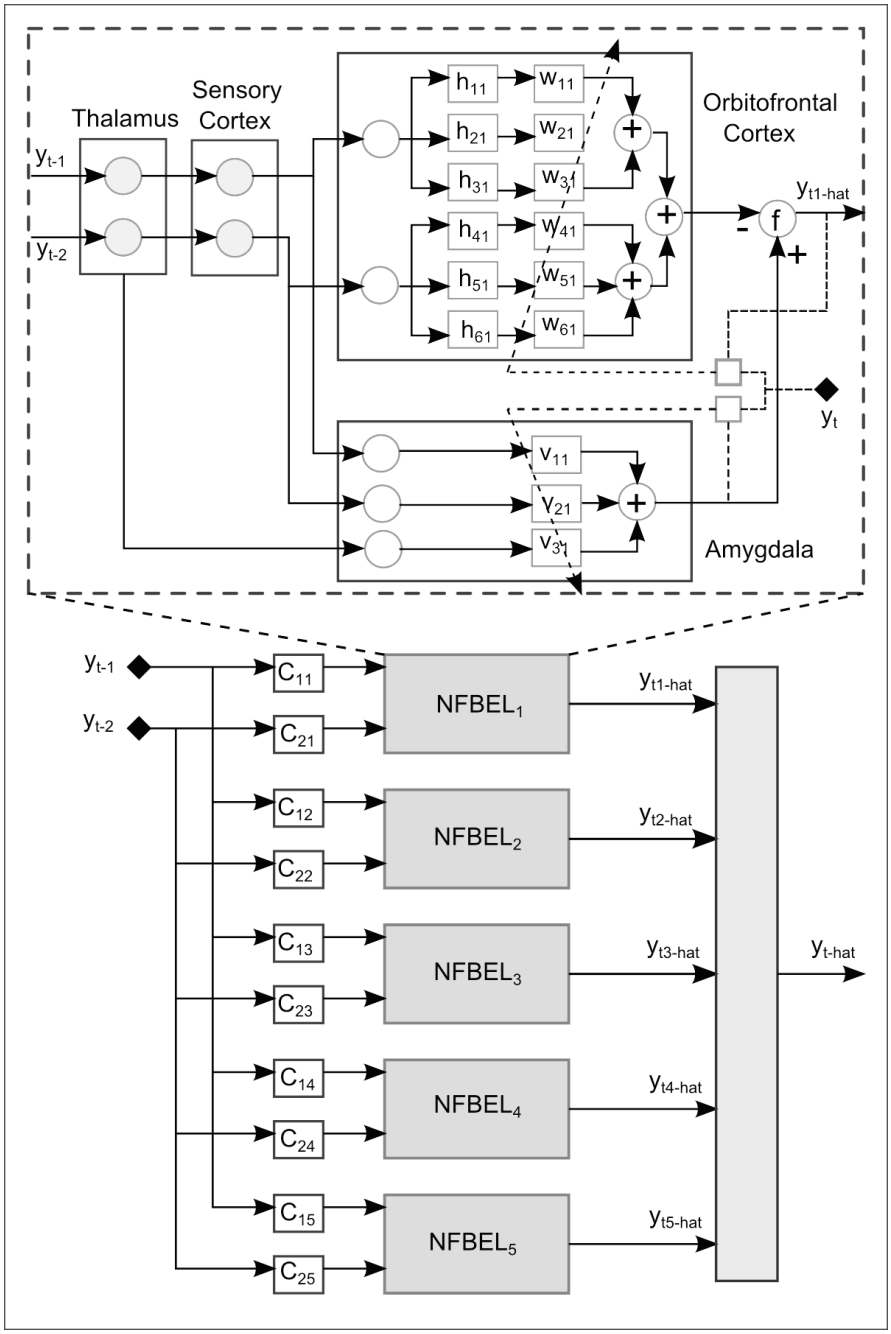


Figure 3
Proposed Dst index predictor

$$h_{1i}^{*}=\begin{cases}-2y_{t-1}+1, & 0\le y_{t-1}<0.5\\ 0, & y_{t-1}\ge 0.5\end{cases}$$
$$h_{2i}^{*}=\begin{cases}-2y_{t-1}, & 0\le y_{t-1}<0.5\\ -2y_{t-1}+2, & 0.5\le y_{t-1}\le 1\end{cases} \quad (4)$$
$$h_{3i}^{*}=\begin{cases}2y_{t-1}-1, & 0.5< y_{t-1}\le 1\\ 0, & y_{t-1}\le 0.5\end{cases}$$

The last three entries of (3) can be computed by replacing y_{t-1} with y_{t-2} in (4). We now compute the prediction error as:

$$e(t)=y(t)-y(t) \quad (5)$$

With the knowledge of (5) and the expanded inputs of (3), the weights of the winner unit of the proposed Dst predictor are adjusted in the following way:

$$w_i^{*}(t+1)=w_i^{*}(t)-\beta e(t)h_i^{*T} \quad (6)$$

$$v_i^{*}(t+1)=v_i^{*}(t)-\gamma v_i^{*}(t)+\alpha\max\left(y(t)-v_i^{*}(t)y_{te},0\right)y_{te}^{T} \quad (7)$$

Where α and β are constants representing the learning rates of amygdala and orbitofrontal cortex respectively while γ is the decay rate. The complete algorithm is shown in Figure 4.

Remark 1: In the proposed hybrid model, neo-fuzzy neurons are only utilized in the orbitofrontal cortex sections of competitive emotional neural network. This is done purposefully as orbitofrontal cortex is believed to have more knowledge of the underlying process. Thus, more degrees of freedom are available in the proposed neuro-fuzzy hybrid model.

Remark 2: The integration of neo-fuzzy network in the amygdala section can also be considered but it will increase the computational complexity of the resulting hybrid model.

Remark 3: In the proposed hybrid model, learning laws for the amygdala sections are the same as in case of competitive emotional neural networks. However, learning laws for orbitofrontal cortex sections are changed to incorporate fuzzified stimuli.

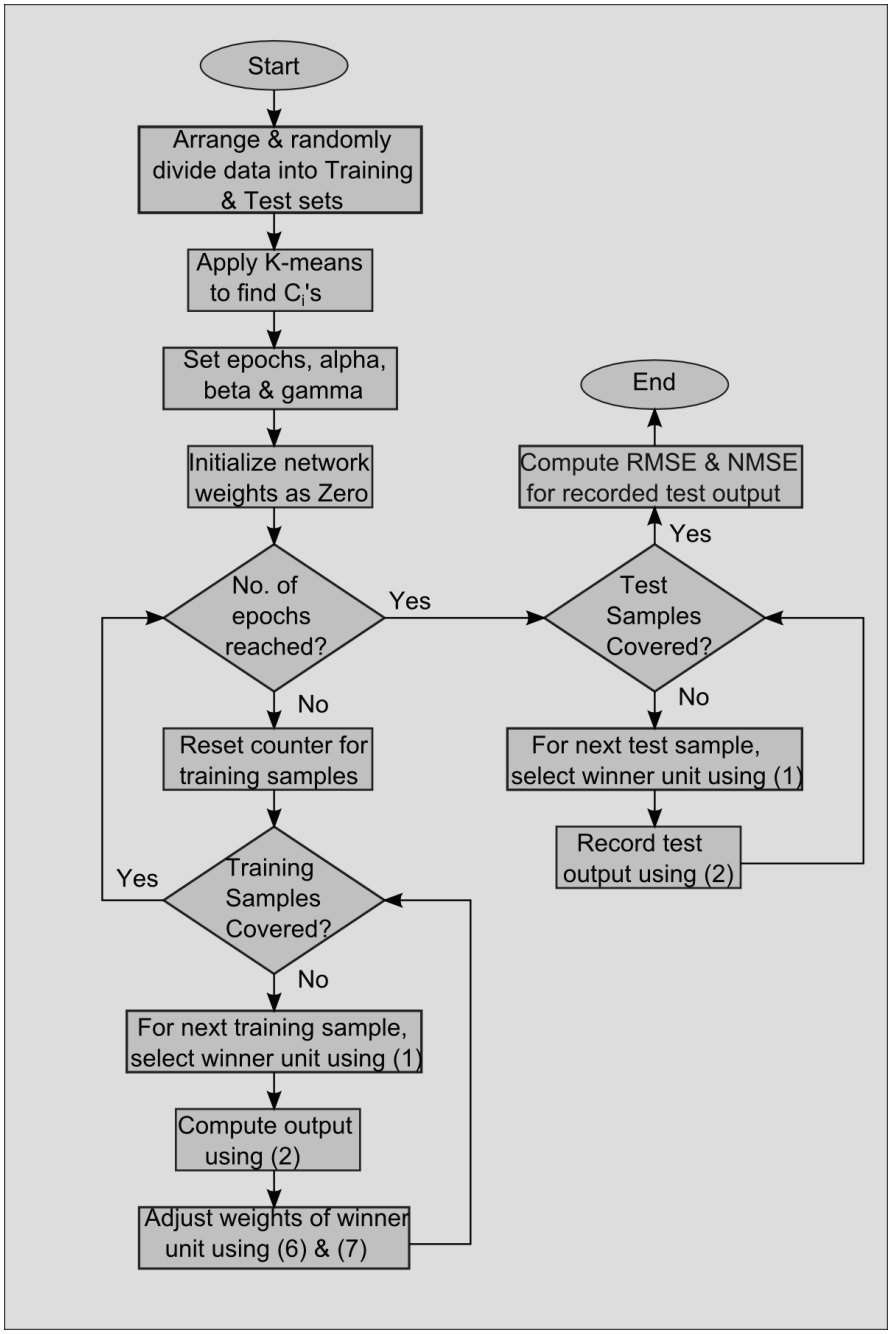


Figure 4
Flow chart of proposed Dst predictor

5 Results & Discussion

The proposed Dst predictor is implemented in the MATLAB programming environment and its performance is evaluated on the dataset obtained from [15] which contains the hourly Dst measurements between the years 2000 and 2008. At first, these 78912 Dst samples are arranged as input-output pairs where each input pair $\left(Dst_{t-1}, Dst_{t-2}\right)$ contains the past two Dst values while output is the current value Dst_t. The resulting 78910 patterns are scaled between 0 and 1 and the scaled dataset is then randomly divided in the ratio 70:30 where 70% of the dataset (55237 patterns) is used for training the proposed model while 30% (23673 patterns) is used for accessing its performance in terms of Root Mean Square Error (RMSE) and Normalized Mean Square Error (NMSE) as defined below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y - y\right)^2} \tag{8}$$

$$NMSE = \frac{\sum_{i=1}^{n}\left(y - y\right)^2}{\sum_{i=1}^{n}\left(y - \bar{y}\right)^2} \tag{9}$$

Where n is the number of test samples and $\bar{y}$ is average of the recorded output Dst values in the test sequence. By setting the parameters as given in Table 1, the proposed predictor is first trained under varying epochs. It is found that the network performance is improved when the number of epochs are increased. The trained network is then deployed to forecast the hourly Dst values and its performance is recorded on the test dataset as shown in Table 2. The predicted values for the first 200 hours in Dst test dataset are plotted in Figure 5. It can be observed that regions of low Dst activity are well-recognized by the proposed model which play vital role in the prediction of geomagnetic storms. Regression analysis of the developed predictor is also performed on the test dataset as depicted in Figure 6 which shows a good amount of correlation between the target and predicted values.

Table 1
Parameters of NFCBEL Dst predictor

Parameters	Description	Values
n_{nf}	Number of neo-fuzzy neurons	2
n_{mfs}	Number of membership functions for one neuron	3
n_{cu}	Number of competing units	5
α	Amydala learning rate	0.1
β	Orbitofrontal cortex learning rate	0.3
γ	Decay rate associated with Amygdala	0.0001

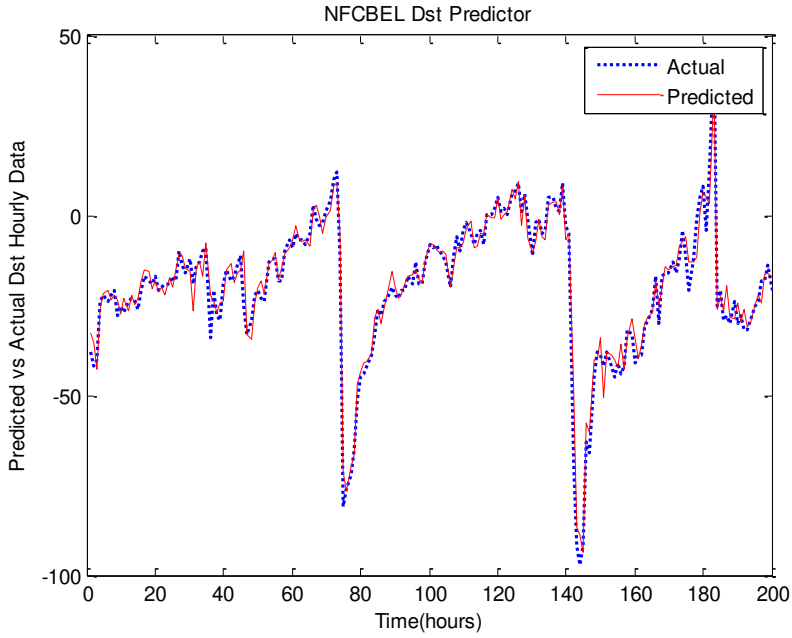


Figure 5
NFCBEL Dst predictor on test dataset

Table 2
Performance of NFCBEL Dst predictor

Epochs	RMSE	NMSE	COR
10	4.9578	0.0436	0.97932
25	4.8040	0.0409	0.98032
50	4.7694	0.0404	0.98047

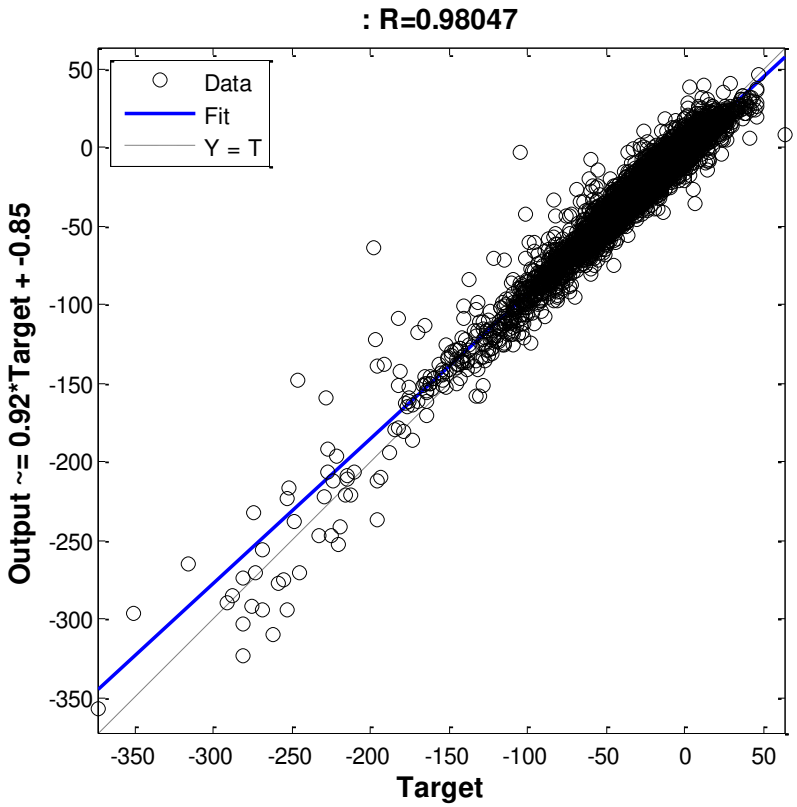


Figure 6
Regression analysis of proposed Dst predictor

The proposed hybrid NFCBEL model is also compared with some state of the art Dst predictors. Following the lines of [15], [18], NMSE is chosen as the basis of comparison for the Dst data between the years 2000 and 2006 which are approximately 61392 patterns. The network with the previously learned weights is deployed to forecast a total of 61392 target Dst values and NMSE is recorded. It is found that the proposed model offers the lowest NMSE as shown in Table 3 which shows its superior performance as compared to other Dst predictors. Further, the predicted Dst values are also plotted for some critical hours when considerable geomagnetic activity is observed. These results are shown in Figures 7 through 9. It can be observed that predicted values are in close agreement with their actual values which validates the good performance of the proposed Dst predictor.

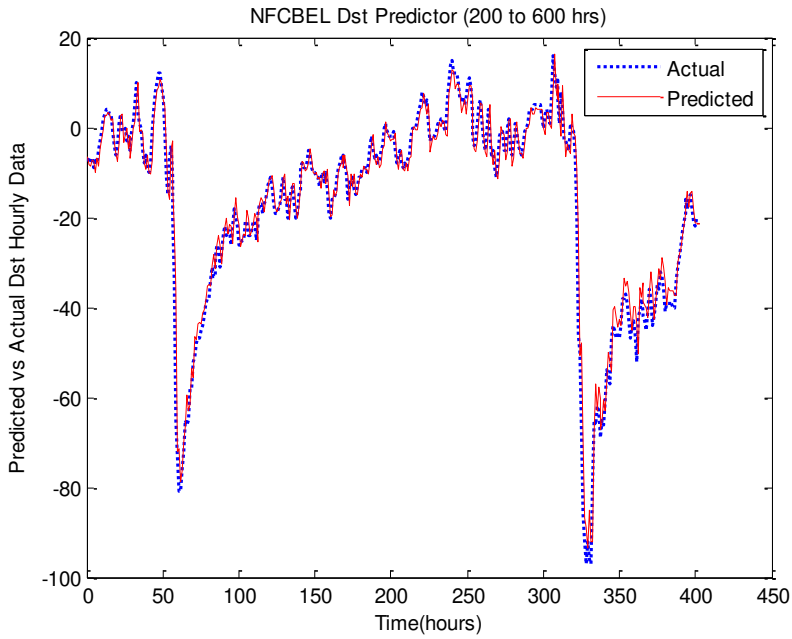


Figure 7
Result of the proposed Dst predictor for the hours 200 to 600 during the years 2000-2006

Table 3
Comparison of Dst predictors based on NMSE between 2000 & 2006

Algorithm	Learning	NMSE
LLNF	LoLiMoT	0.5348
Adaptive LLNF	RLoLiMoT	0.0968
ADBEL	Emotional Decaying	0.1123
Proposed NFCBEL	Emotional Decaying	0.0400

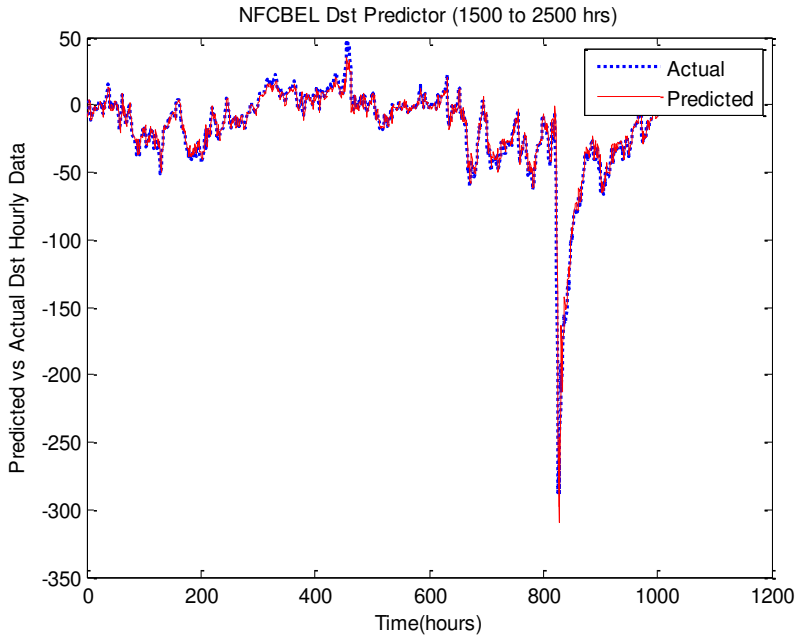


Figure 8
Result of the proposed Dst predictor for the hours 1500 to 2500 during the years 2000-2006

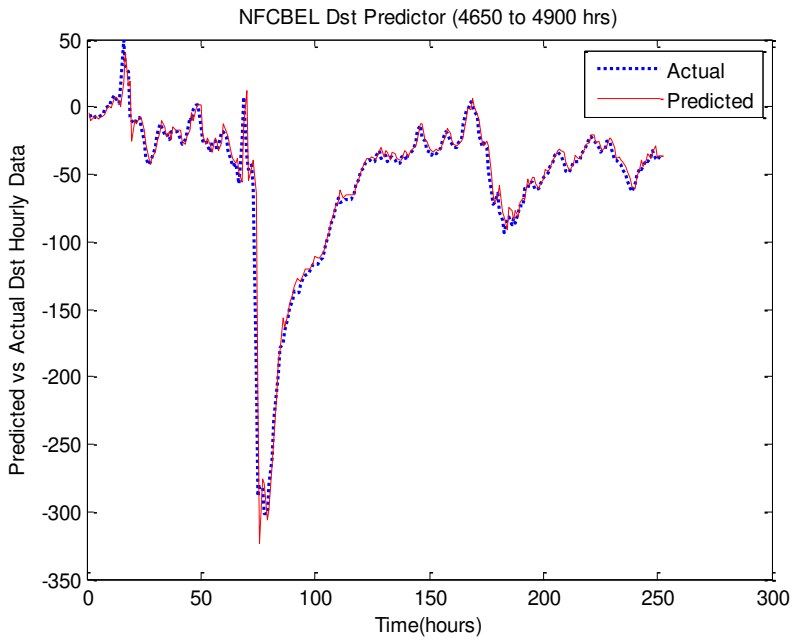


Figure 9
Result of the proposed Dst predictor for the hours 4650 to 4900 during the years 2000-2006

Conclusions

This paper presents the design of a novel hybrid model for the hourly forecast of the Disturbance Storm Index, which is an important parameter for predicting geomagnetic storms. The model combines competitive emotional neural networks with neo-fuzzy neurons to yield an effective Dst predictor which offers features such as, low computational complexity and fast learning. Low complexity is the result of fewer inputs, neo-fuzzy neurons and competing units, while fast learning is the result of employing emotion processing mechanism of the mammalian brain. The proposed model is trained offline and then deployed for the hourly prediction of Dst activity. The performance of the model is also evaluated in terms of RMSE, NMSE and COR. The comparison of the proposed model with some state-of-the-art predictors, reveals its superior performance, as the model offers the lowest NMSE. Future work involves determining model parameters, like optimal number of neo-fuzzy neurons, competing units and weighting through metaheuristic algorithms.

Acknowledgement

This work was supported by Natural Sciences & Engineering Research Council of Canada (NSERC).

References

[1] Y. Cerrato, E. Saiz, C. Cid, M. A. Hidalgo: Geomagnetic storms: Their Sources and a Model to Forecast the Dst Index, Lecture Notes and Essays in Astrophysics, Vol. 1, 2004, pp. 165-176

[2] L. R. Cander, S. J. Mihajolvic: Forecasting Ionospheric Structure During the Great Geomagnetic Storms, Journal of Geophysical Research, Vol. 103, No. A1, 1998, pp. 1998

[3] B. T. Tsurutani, W. D. Gonzalez, G. S. Lakhina, S. Alex: The Extreme Magnetic Storm of 1-2 September 1859, Vol. 108, No. A7, 2003, pp. 1-8

[4] P. Stauning: High Voltage Power Grid Disturbances During Geomagnetic Storms, Proc. Solar Cycle and Space Weather Euroconference, 2001, pp. 521-524

[5] H. Lundstedt, H. Gleisner, P. Wintofit: Operational Forecasts of the Geomagnetic Dst Index, Geophysical Research Letters, Vol. 29, No. 24, 2002, pp. 1-4

[6] R. K. Burton, R. L. McPherron, C. T. Russell: An Empirical Relationship between Interplanetary Conditions and Dst, Journal of Geophysical Research, Vol. 80, No. 31, 1975, pp. 4204-4214

[7] F. R. Fenrich, J. G. Luhmann: Geomagnetic Response to Magnetic Clouds of Different Polarity, Geophysical Research Letters, Vol. 25, No. 15, 1998, pp. 2999-3002

[8] T. P. O'Brien, R. L. McPherron: Forecasting the Ring Current Index Dst in Real Time, Journal of Atmospheric and Solar-Terrestrial Physics, Vol. 62, No. 14, 2000, pp. 1295-1299

[9] G. Pallocchia, E. Amata, G. Consolini, M. F. Marcucci, I. Bertello: ANN Prediction of the Dst Index, Memorie della Societa Astronomica Italiana Supplemnt, Vol. 9, 2006

[10] J-G. Wu, H. Lundstedt: Geomagnetic Storm Predictions from Solar Wind Data with the use of Dynamic Neural Networks, Journal of Geophysical Research, Vol. 102, No. A7, 1997, pp. 14255-14268

[11] G. Pallocchia, E. Amata, G. Consolini, M. F. Marcucci, I. Bertello: Geomagnetic Dst Index Forecast Based on IMF Data only, Annales Geophysicae, Vol. 24, 2006, pp. 989-999

[12] T. Babaie, R.Karimizandi, C. Lucas: Learning Based Brain Emotional Intelligence as a New Aspect of an Alarm Aystem, Soft Computing, Vol. 12, No. 9, 2008, pp. 857-873

[13] A. Gholipour, C. Lucas, D. Shahmizadi: Predicting Geomagnetic Activity Index by Brain Emotional Learning, WSEAS Transactions on Systems, 2004, pp. 296-299

[14] E. Lotfi, M. –R. Akbarzadeh-T: Supervised Brain Emotional Learning, Proc. IEEE World Congress on Computational Intelligence, 2012, pp. 208-213

[15] E. Lotfi, M. –R. Akbarzadeh-T: Adaptive Brain Emotional Decayed Learning for Online Prediction of Geomagnetic Activity Indices, Neurocomputing, Vol. 126, 2014, pp. 188-196

[16] E. Lotfi, M. –R. Akbarzadeh-T: Emotional Brain Inspired Adaptive Fuzzy Decayed Learning for Online Prediction Problems, Proc. IEEE International Conference on Fuzzy Systems, 2013, pp. 1-7

[17] M. Mirmomeni, C. Lucas, B. Moshiri, B. N. Araabi: Introducing Adaptive Neurofuzzy Modeling with Online Learning Method for Prediction of Time-Varying Solar and Geomagnetic Activity Indices, Expert Systems with Applications, Vol. 37, 2010, pp. 8267-8277

[18] H. Milad, U. Farooq, M. El-Hawary, M. Usman Asad: Neo-fuzzy Integrated Adaptive Decayed Brain Emotional Learning Network for Online Time Series Prediction, IEEE Access, Vol. 5, 2017, pp. 1037-1049

[19] E. Lotfi, O. Khazaei, F. Khazaei: Competitive Brain Emotional Learning, Neural Processing Letters, 2017

[20] T. Yamakawa, E. Uchino, T. Miki, H. Kusanagi: A Neo Fuzzy Neuron and its Applications to System Identification and Prediction of the System Behavior, Proc. International Conference on Fuzzy Logic and Neural Networks, 1992, pp. 477-483

[21] C. Lucas, D. Shahmizadi, N. Sheikholeslami: Introducing BELBIC: Brain Emotional Learning Based Intelligent Controller, Intelligent Automation and Soft Computing, Vol. 10, No. 1, 2004, pp. 11-21

Optimal Boolean Programming with Graphs

Benedek Nagy

Eastern Mediterranean University, Department of Mathematics, Famagusta, North Cyprus, via Mersin-10, Turkey

Abstract: In this paper, we solve some special types of linear and non-linear Boolean programming problems. We present a method for transforming the used linear 0-1 *inequalities into a weighted directed graph. Allowing equalities our conditions are non-linear, but the transformation to weighted directed graphs works also in these cases. In graph representations, the "critical edges" are used to represent the non-linear conditions. Basic, modified and extended Boolean programming problems are investigated. Linear goal-functions are used in optimization. The presented algorithm, similarly as algorithms for knapsack-problems, gives a relatively good solution, moreover, the algorithm extended by the backtrack graph-search strategy, guarantees optimal solutions for the considered problems.*

Keywords: Boolean-programming; 0-1 *inequalities; optimization; knapsack problem; greedy algorithm; backtracking; graphs*

1 Introduction

There are some special integer-valued programming problems in which the values of variables must be in the set {0,1}. These Boolean programming (BP) problems are well known in the literature. We refer to [7] as a classical textbook on the topic related to optimization and Boolean programming. BP problems are ubiquitous in Artificial Intelligence. Planning with resource constraints, satisfiability testing and winner determination in combinatorial auctions are all belonging to this type of problems [18, 21]. The knapsack problems [9, 19, 22] are also in this class. In many logical problems, e.g., in some logic puzzles, the same or nearly similar conditions occur [10-17, 20]. In [3] special redundancy criteria were used to produce minimal number of extended clauses for transforming problems to equivalent ones.

In this paper, we continue the work that has been started in [11-14]. The conditions of a basic BP problem have a translation to a set of linear-programming conditions such that in each condition two variables occur. This is a key issue for representing such problems with graphs. The conditions of a modified BP problem can also be translated to linear conditions, however these conditions, in general,

cannot be written as a set of conditions such that each condition contains only two variables. Despite this, we show how these conditions can still be represented by graphs. In the extended BP problems, the conditions are more general than in the basic or modified problems. Graph-representation [2] of the conditions will be used, where the so-called critical-edges refer to the conditions where some of them are considered together in the linear programming model in cases of modified and extended BP problems. There is a goal-function to do optimization. We search the possible solution(s) under the criteria that the goal-function has maximal (or minimal) value. Based on our graph representation we use local steps (comparing variables) to infer more knowledge about their relations. The algorithm is extended with a greedy and with a backtracking part to ensure to obtain the solution(s). We use a greedy algorithm to find a reasonable solution and, if it is not enough, backtracking to find a better, or even, the best solution.

Our algorithm is based on the algorithms of [12-14]. The algorithm uses the graph representation. It can modify the graph of conditions in such a way that the solutions for the new graph are exactly the same as for the original problem. Using the fact that the value of every variable must be in {0,1} and using some other observations, we can conclude a unique possible value for some variables. Consequently, the steps of the algorithm are the graph modifications and assigning values to variables. We solve the mentioned types of BP problems by our algorithm.

2 Conditions of BP

In this research, we solve problems where each variable can have a value of the set {0, 1}. Now we give formal definitions of various types of BP problems we are working with. Capital letters (sometimes with indices) are used to denote the variables of the problem.

First, the definitions of basic and modified BP problems are given. Then, we define the general case which are called extended BP problems and, in fact, they are generalizations of the modified Boolean programming problems.

In this paper, we will solve the generalized problem, and we will show that the basic and modified problems are its special subcases.

Definition 1. (Basic-, Modified- and Extended BP problem) Let B_i denote the variables of the problem for $i \leq n$ where n denotes the number of variables. If every condition has the form

(1) $B_i \leq B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

then our conditions are of basic BP type.

If every condition is written in one of the following forms

(1) $B_i \le B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

(2) $B_i = B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

and every B_i appears on the left-hand side of at most one condition of type (2), then the conditions of the problem are of modified BP type.
If each condition has one of the following four forms:

(1) $B_i \le B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

(2) $B_i = B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

(3a) $B_i < B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

(3b) $1 - B_i < B_{r1} \cdot B_{r2} \cdot ... \cdot B_{rm} \cdot (1-B_{p1}) \cdot (1-B_{p2}) \cdot ... \cdot (1-B_{pk})$

and every B_i appears on the left-hand side of at most one condition of type (2), then, these conditions are of extended Boolean programming type.
Let our goal-function be in the form

(4) $Z = a_1 \cdot B_1 + a_2 \cdot B_2 + ... + a_n \cdot B_n$, where the values a_i are fixed real numbers.

If our goal is to minimize or maximize the function Z and our conditions are of type basic, modified or extended Boolean programming, then our problem is basic, modified or extended Boolean problem (BP), respectively. ■

We have no restriction on the occurrences of the variables in basic BP and we have only one restriction in case of modified and extended BP. While the main difference between the basic and the modified BP is the possibility of equations, the difference between the modified and the extended BP is the possibility of strict inequalities. Sometimes it happens that our conditions are not in the form as we have defined above, but we can use graph representation and we can still use our method. It may occur in the case when there are more than one conditions of type (2) with the same variable on the left hand side. If we have special type (2) conditions such that only one variable appears in both sides, then we may interchange the variables on the left and on the right hand side, it may help to transform our conditions to the defined form. If some of our conditions look like (3a) but in the left hand side there is not only one variable, but the sum of more than one variables, we can separate the condition to more than one conditions of type (3a) in which the left hand side contains only one variable and the right hand side is the same as in the original condition. If some of our conditions look like (3b) but in the left hand side there is not only one variable, but the number of the variables minus the sum of these variables, then we can write several conditions of type (3a) such that in each of them the left hand side contains exactly one variable and the right hand side is the same as in the original condition.

All basic BP problems have possible solutions, for example, each $B_i = 0$ is a trivial possible solution. Opposite to this, there are some modified and extended BP

conditions which have no solution. A simple example is as follows. Let the only one variable be B. Let our unique condition be $B = 1-B$. It is easy to see that there is no possible solution in the set $\{0,1\}$.

Now, we will define the basis of our graph theoretic approach. We will use directed graphs to represent the conditions of a BP problem, to do this, we need the conditions to be written as a set of conditions containing at most two variables.

Lemma 1. We can write the type (1) conditions of a BP-problem in the form:

(5a) $B_i \le B_{rj}$
(5b) $B_i \le 1-B_{pl}$

Proof. If $B_i = 0$, then the conditions trivially hold in both in the original and in the stated cases. If $B_i = 1$, then at each condition the equality holds. ■

The conditions in any of the forms (5a) and (5b) are called *atomic conditions*. Observe that they are, in fact, linear. Now we show that the conditions of not only the basic type, but the modified and extended BP problems can also be written in linear programming form. The corresponding type (5a) and (5b) conditions are also satisfied for a type (2) condition, but using these new forms, the new conditions are not equivalent to the original condition in form (2).

Lemma 2. The conditions (2), (3a) and (3b) can be written in linear form. A condition of type (2) can be written as the set of corresponding conditions in the form (5a) and (5b) and an additional condition

(2*) $1-B_i \le (1-B_{r1}) + (1-B_{r2}) + ... + (1-B_{rm}) + B_{p1} + B_{p2} + ... + B_{pk}$

Moreover, conditions of the form (3a) can be written as a set of conditions

(3a*) $B_i \le 0$ $\qquad 1 \le B_{rj}$ $\qquad 1 \le 1-B_{pl}$

while conditions of the form (3b) are equivalent to a set of conditions

(3b*) $1 \le B_i$ $\qquad 1 \le B_{rj}$ $\qquad 1 \le 1-B_{pl}$

Proof. Let a type (2) condition be given. Then, in case $B_i = 1$, all corresponding type (5a) and (5b) conditions are equalities and (2*) does not mean any further restrictions. In case $B_i = 0$, however, (2*) takes care about the equality of (2) by forcing either at least one B_{rj} to be 0 or at least one B_{pl} to have value 1. Considering type (3a) or (3b) conditions, the strict inequality hold only if the left hand side has value 0 and the right hand side contains a product of only 1's. Moreover, all the formulae of (3a*) and (3b*) are representing equalities. ■

In the graph we represent the atomic conditions. We will use weighted arrows as edges in the graph according to the types of the conditions. We use abbreviations LHS and RHS for left hand side and right hand side, respectively. We call a relation critical if it is based on an original type (2) relation. It is easy to show that if the variable of the LHS of a type (2) condition has value 1, then the conditions

in forms (5) are equivalent to the original condition. However, in case the value of the variable on the LHS is 0, the conditions are not the same, we must pay attention that at least one element of the product in the RHS must be 0. The critical edges are used to represent these kinds of possible equalities.

Definition 2. (Critical condition) If an atomic condition may represent an equality of type 0 = 0, which comes directly from a type (2) condition of the BP problem, then this atomic condition is critical. ■

We use various weights (labels) to represent the possible conditions between any two variables:
If the value is not a multiplier of 3, then it means that the condition (5a) is true for these variables. If the value is not less than 3, then it means that the condition (5b) is true. The possible weights with their meanings are shown in Table 1.

Table 1
Edge weights and their meaning

weight	relation between the variables		
0	we have not any information (not yet)		
1	critical condition (5a)		
2	not critical condition (5a)		
3			critical condition (5b)
4	critical condition (5a)	and	critical condition (5b)
5	not critical condition (5a)	and	critical condition (5b)
6			not critical condition (5b)
7	critical condition (5a)	and	not critical condition (5b)
8	not critical condition (5a)	and	not critical condition (5b)

Definition 3. (Associated graph of conditions) Let a basic or a modified or an extended BP programming problem be given. Let the number of the vertices of graph G be *n*, the same as the number of variables in the BP problem. Assign the variables of the BP problem to the vertices of the graph G. If variable A is on the left side of a condition of type (1) or type (2), then we will use arrows from A to the nodes representing variables on right hand side of that condition. The weights of these arrows depend on the condition. First we draw all edges that are needed. Then we will calculate their weights in the following way. Let all weights be 0 initially (we can draw all possible arrows in the graph with weight 0). After this, we read each condition one by one, and modify the weights according to the next steps:

- If variable A is on the LHS of a type (2) condition and variable B appears as a member of the product on the RHS and the weight of the corresponding arrow is not 1 mod 3, then the weight is increased or decreased by 1 such that it will be 1 mod 3.

- If variable A is on the LHS of a type (2) condition and variable B appears in the product written as (1 – B) on the RHS and the weight of the corresponding arrow is not between 3 and 5 (inclusively), then the weight is increased or decreased by 3 such that it becomes between 3 and 5.
- If variable A is on the LHS of a type (1) condition and variable B appears as a member of the product on the RHS and the weight of the corresponding arrow is divisible by 3 then its value is increased by 2.
- If variable A is on the LHS of a type (1) condition and variable B appears in the product on the RHS written in the form (1 – B) and the weight of the corresponding arrow is less than 3 then 6 is added to its value.

And now, using the type (3) conditions, we can assign values to some nodes from the set {0,1}. If variable A appears on the LHS of a condition of type (3a), then we assign 0 to it. If it occurs on the LHS of a condition of type (3b), then we assign the value 1 to it. If a variable is a member of the product of a type (3) condition (either 3a or 3b) on the right hand side, then we assign 1 to it, if A appears as (1–A) in the product on the RHS of any condition of type (3), then we assign 0 to it.
Then G is the (initial) graph of the BP problem.

We say that an assignment of the values {0,1} to the variables is a solution of the graph if it is compatible with all the conditions represented, i.e., all conditions are satisfied. ■

By Lemmas 1 and 2, Definition 2 and Table 1, one can see that all information about the conditions of the problem is encoded in the graph, and thus, the solution(s) of the graph and the BP problem coincide.

3 Manipulation of the Graph

In this section we give and explain the graph modifying steps. They are defined in such a way that the possible solutions do not change, hence, first, we define the equivalence among graphs.

Definition 4. (Graph-Equivalence) We say that two graphs are equivalent if the sets of the possible solutions of them (i.e. the possible solutions of the BP problems which are represented by these graphs) are the same. ■

In this part some possible steps for modifying the graph are shown. We will use some kinds of changing steps as node-evaluating and arrow-adding or changing.

There are two types of node-evaluating steps, let us see them first. We start with those that are used in some cases with an already known value of a node. They are called valuable arrows and listed in the following list.

Let A and B be two distinct vertices.

a) If A has value 1 and the arrow from A to B has a weight that is not a multiplier of 3, then let the value of B be 1.

b) If A has value 1 and the arrow from A to B has a weight that is larger than 2, then assign 0 to B.

c) If A has value 1 and the arrow starting at B and ending at A has a weight that is not less than 3, then assign 0 to B.

d) If A has value 0 and each arrow starting at A has a weight different from 4 and there is exactly one edge starting at A with odd weight, and its weight is either 1 or 7, and it goes to B, then let the value of B be 0.

e) If B has value 1 and each edge from A has a weight different from 4 and there is a unique edge starting at A with odd weight, and its weight is either 1 or 7 and it goes to B, then let 1 be assigned to A.

f) If B has value 0 and the arrow starting at A and ending at B has a weight that is not a multiplier of 3, then let the value of A be 0.

g) If B has value 0 and each arrow from A has a weight different from 4 and there is a unique arrow from A with odd weight, and its weight is either 3 or 5 and it goes to B, then assign 1 to A.

h) If B has value 0 and each arrow from B has a weight different from 4 and there is a unique edge starting at B with odd weight, and its weight is either 3 or 5 and this arrow ends at A, then let A have a value of 1.

The other types of node-evaluating steps are the so-called basic schemes. They work without any known values.

α) If a loop arrow occurs at vertex A such that its weight is larger than 2, then let 0 be assigned to A.

Actually, by viewing its structure, the following scheme is between the valuable arrows and the basic schemes.

β) If vertex A has a starting arrow with weight 4, 5, 7 or 8, then let A have the value 0.

γ) Let A, B and C be three vertices. If there are edges both from A and from B to C with weights non-divisible by 3 and there is an arrow from A to B with weight 3 or 5 such that there is no other arrow from A with weight 4 or with an odd weight, then let 1 be assigned to C.

We continue with arrow adding and arrow changing steps.

Considering two (A, B) or three vertices (A, B and C) in the graph such that they have the connections satisfying any of the following conditions, we can modify the weight of an arrow as it is specified below:

1) If the weight of the edge from A to B has a value more than 2 then
if the weight of the edge from B to A is at most 2, then it is increased by 6.

2) If each edge from A has a weight different from 4 and there is a unique edge from A with odd weight, moreover it is either 1 or 7, and it ends at B, then
if the weight of the edge from B to A is divisible by 3, then let it be 2 more than it was.

3) If none of the values of the weights of the edges A to B and B to C are divisible by 3, then,
if the weight of A to C is a multiplier of 3, then let this weight be increased by 2.

4) If the weight of the edge from A to B is not a multiplier of 3 and the weight of the edge from B to C is greater than 2, then

   - If the weight of the edge from A to C is at most 2, then it is increased by 6
   - If the weight of the edge from C to A is at most 2, then it is increased by 6

5) If the weight of the arrow from A to B is not a multiplier of 3 and the weight of the edge from C to B is not less than 3, then

   - If the weight of the edge from A to C is at most 2, then it is increased by 6
   - If the weight of the arrow from C to A is at most 2, then it is increased by 6

6) If the weight of the arrow from A to B is not less than 3 and each arrow from B has weight different from 4 and there is a unique edge from B with an odd weight, and it is either 1 or 7, and this edge ends at C, then

   - If the weight of the arrow from A to C is less than 3, then it is increased by 6
   - If the weight of the edge from C to A is less than 3, then it is increased by 6

7) If each edge from B has a weight different from 4 and there is a unique edge from B with odd weight and this weight is either 1 or 7, and this edge ends at C, and the weight of the arrow from B to A is not less than 3, then

   - if the weight of the arrow from A to C is less than 3, then it is increased by 6
   - if the weight of the edge from C to A is less than 3, then it is increased by 6

8) If the weight of the edge A to B is at least 3 and each edge from B has a weight different from 4 and there is a unique edge from B with odd weight and this weight is either 3 or 5, and this edge ends at C, then
if the weight of the arrow from A to C is a multiplier of 3 then it is increased by 2.

9) If the weight of the arrow from A to B is not less than 3 and each edge from C has a weight different from 4 and there is a unique edge from C with odd weight and this weight is either 3 or 5 and this edge ends at B, then
if the weight of the edge from A to C is a multiplier of 3, then it is increased by 2.

10) If each edge from B has a weight different from 4 and there is a unique edge from B with odd weight and this weight is either 3 or 5 and this edge ends at C, and the weight of the arrow from B to A is not less than 6, then
if the weight of the arrow from A to C is a multiplier of 3 then it is increased by 2.

11) If the weight of the arrow from B to A is not less than 3 and each edge from C has a weight different from 4 and there is a unique edge from C with odd weight and this weight is either 3 or 5, and this edge ends at B, then
if the weight of the edge from A to C is a multiplier of 3, then it is increased by 2.

We are continuing with 3 arrow-changing steps, however, these steps use already known values at some vertices.

12) If 1 is assigned to B and the weight of the edge from A to B is 1 mod 3 and there is another edge starting at A with either an odd weight or with weight 4, then
the weight of the edge from A to B is incremented by 1.

13) If 0 is assigned to B and the weight of the edge from A to B is either 3 or 5 and there is another edge from A either with weight 4 or with an odd weight, then
the weight of the edge from A to B is increased by 3.

14) If A has the value 1 and

- If the weight of an edge from A is 1, then let the weight of this edge be 2
- If the weight of an edge from A is 3, then let the weight of this edge be 6

Finally, we have some arrow-changing steps for subgraphs containing three vertices. Let A, B and C be three vertices.

15) If the value of the edge from A to B and the value of the edge from A to C are both 1 mod 3 and the weight of the edge from B to C is not a multiplier of 3, then
let the weight of the edge from A to C be increased by 1.

16) If the value of the edge from A to B and the value of the edge from A to C are both between 3 and 5 (inclusively) and the weight of the edge from B to C is not a multiplier of 3, then ,
the weight of the edge from A to B is increased by 3.

Now let us discuss these steps briefly. The next lemma shows how we can use the critical edges to gain information.

Lemma 3. If there is a unique critical condition for a vertex (exactly one of the weights of the edges from there is odd, i.e., has weight 1, 3, 5 or 7), then this condition must represent an equality.

Proof. It goes by indirect method. ■

For example, we use the previous lemma at arrow changing steps 6) or at node evaluating step d). Using this property and the meaning of atomic conditions (type (5a) and (5b) inequalities) represented by the edges of the graph, the justification of valuable arrow steps a) to h) are based on inferences to find the unique value for a variable represented at a node where the value of one of the involved variables are already known.

It is easy to check that at the basic scheme steps α), β) and γ), there is only one possibility for the specified variable to satisfy the conditions and it is assigned for the corresponding variable.

Condition (5a) is transitive and (5b) is symmetric. We use these properties in some steps, for example in 1), 2) etc. In those arrow-adding cases we increase the weight by 2 or 6 according to the new condition (5a) or (5b), and these new conditions are not critical, because they are not from an original type (2) condition. If a condition cannot be critical (see its definition) or our graph is equivalent to the original in the case of change a critical condition to non-critical one, then we use those steps which modify the given condition to non-critical one: we increase the respective value by 1 in case of (5a) or by 3 in case of (5b). Hence, steps 1) thru 16) are correct.

Remark 1. The arrows which are critical in the solution are also critical in the original graph.

4 Algorithm to Solve BP Problems

In this part, we present an algorithm which works in two variations. The first variation provides a relatively good solution in a short time by a greedy approach, if there is a possible solution of the problem. The second variation works more slowly: it includes the case when we do not stop after the first solution is found. This variation of the algorithm will find the optimal solution, because it works after the first solution is found till it is guaranteed that no better solution can be found than the latest one that has already been found.

Without loss of generality, for simplicity, we may assume that the variables indexed in non-decreasing order by the absolute values of their coefficients (multipliers) in the goal function, i.e. $|a_i| \geq |a_j|$ if and only if $i < j$ for all i,j between 1 and n.

Algorithm 1 searches in almost the whole tree of the state space. It is a mixture of the known knapsack and backtrack [4, 9, 19, 21] algorithms extended with our graph-modifying method.

Algorithm 1

Input: a BP problem.
Output: a solution/the best solution (or "Contradiction" if no solution exists).
0. If the goal is to minimize Z then let Z' = – Z be the new goal function, and maximize it. (Else let Z' = Z.)
1. Draw the graph representation of the BP problem.
2. Apply the possible graph-changing steps.
3. If there is a vertex with both values (1 and 0), then the problem is not solvable. (Contradiction.) STOP
4. If there is a unique value at each vertex, then this assignment could be the solution. Check it (because using graph-changing steps we may get a contradiction; see previous step)). If it satisfies all the conditions, it is the solution; otherwise the problem is not solvable (Contradiction). STOP
5. If there are no more usable graph-changing steps, then choose the variable which has the smallest index among the variables which do not have any assigned value yet. If the multiplier is positive than choose and assign value 1 to this vertex, if the multiplier is negative, then assign value 0 for this variable.
6. Use the possible graph-changing steps.
7. If there is a vertex with both values (1 and 0), then this is contradiction. (BACKTRACK)
If there is a unique value at each vertex, then this may be a solution. (Check it! and if it is fairly good, then STOP, else memorize this solution (if this is better than the previously found solution) and BACKTRACK.)
8. If BACKTRACK then go back to the previous value assigning in step 5, and find the last chosen variable for which we have not tried both values, and assign the other value to it. Cancel all the graph-changing steps done after the step we have assigned the previous value to this variable. If such a variable does not exist, then we have already tried all possibilities and finished the search, if there is a memorized solution, then the last one is the best, else there is no possible solution. STOP
9. Go to step 5.

We can use cuts to speed up the algorithm. If we already have a solution, then we can check the possible maximal value of Z' for the remaining part of the state space:

Let our goal be, to maximize the function $Z' = a_1 \cdot B_1 + a_2 \cdot B_2 + ... + a_n \cdot B_n$

We call the values B_i fixed values if they cannot change in the remaining part of the search. We use the concept critical node of the search tree, for which we gave the value of a variable by step 5 of the algorithm, and it is the variable with the smallest index such that we have not tried with both values. The fixed values are the variables which have got values before we used the value assigning at the critical node. (The fixed values are the variables which have smaller indices than the variable at the critical node and the variables which got their values by using graph-changing steps using only nodes with other fixed values.)

The possible best solution in the remaining state tree is the following: if the fixed values have their actual values and all other variables has the best value (i.e. 1, if the multiplier is positive and 0 if the multiplier is negative.) Therefore, if we remember the fixed values we can easily calculate that value:

$$\sum_{\substack{\text{for } i, \text{ where} \\ B_i \text{ fixed}}} a_i B_i + \sum_{\substack{\text{for i, where} \\ B_i \text{ not fixed,} \\ \text{and } a_i > 0}} a_i$$

If this amount is not greater than the best found (memorized) solution, then we can finish the search. We can calculate this amount after we make a BACKTRACK. Or we can evaluate this sum after all node-value assigning (at step 5 or via graph-changing steps) and we can use BACKTRACK earlier if this value is already less than at the best found solution. If a variable with negative multiplier get the value 1 or a variable with positive multiplier get the value 0, then the sum will decrease by the absolute value of the multiplier.

As a consequence of the description given in Section 3, the graph-changing part of our algorithm works properly. Step 5 of the algorithm represents the knapsack problem and we use greedy algorithm to solve it. Since this step cannot guarantee the (best) solution we expand our algorithm by the backtrack method. We use a kind of branch-cost backtrack, in which we can calculate the maximal value of the goal function of the possible solutions of the remaining search space and we can use this value to decide whether we continue the search.

Remark 2 (On the difference between various types of BP) By analyzing the original graphs of various types of BP-problems one can see that in basic problems only even numbers are used as weights in the graph (2, 6 and 8). In modified BP-problems we use more weights, and in extended problems we have knowledge about values of some nodes originally.

5 Examples

In this part we will show some examples.

Example 1 (BP problem without possible solution)

Let A, B and C be the variables. The conditions:
$A < B$
$1 - B < C$
$B \leq A \cdot C$
$Z = 5A + 3B + C$ and the task is to minimize Z.

Solution. It is an extended BP problem. We want to maximize the function $Z' = -5A - 3B - C$. First we draw the graph of the example (Fig. 1). The first condition gives the values of A and B (0 and 1, respectively). The second

inequality gives the value B and C (both of them are 1). And we have two weighted edges from the third condition.

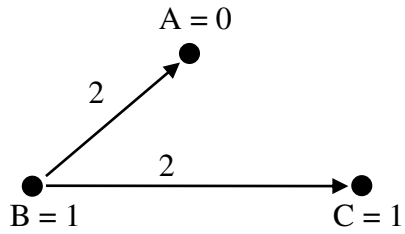


Figure 1
Graph representation of Example 1

We have a value at each node, but it is contradictory. (Between B and A, the weight of the arrow means that it is impossible that B is 1 and A is 0, it is described in point f) with interchanged role of A and B.)

Example 2

$A \leq D$
$A = 1 - B$
$B = D \cdot (1 - E)$
$C \leq (1 - B) \cdot D$
$E \leq B \cdot F$
$Z = 3A - 2.21B + 3D - 22E - 3.25F$, the task is to maximize value of Z.

Solution. This is a modified BP problem. Let us draw its graph (Fig. 2).

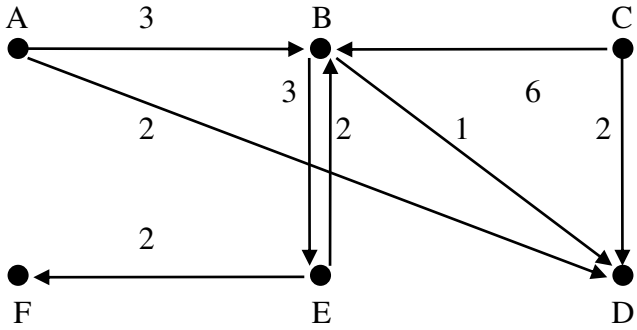


Figure 2
Graph of Example 2

In this graph, one can use some local graph changing steps: There is a basic scheme γ) with vertices A, B and D, consequently, D = 1. Then one can use step 12) for the edge between B and D, thus its weight becomes 2. One can use step 1)

between B and E, therefore the weight of the edge EB goes to 8. But according to step β) value 0 is obtained at E. Now one can use step g) for the arrow BE getting value 1 at vertex B. Then, step c) for A and B and for B and C are used to get values at A and C: they both receive value 0. The values of five vertices are already known. There is no step to get the value of F. We are at point 5 of the algorithm. The goal is to maximize Z, in which F has negative coefficient. Thus, let F = 0. This is the best possible solution: B = D = 1 and A = C = E = F = 0, yielding Z = 0.79.

Example 3

$A \le (1 - E)$
$B = (1 - C) \cdot (1 - D)$
$C \le A \cdot D$
$E < A \cdot (1 - E)$
$E = B \cdot C \cdot (1 - D)$
$Z = -5A + 9B + 3C + 7D + 0.5E$

The goal is to maximize Z.

Solution. We will use the next ordering on the variables: B, D, A, C and E, and function Z' = Z. It is an extended BP problem. The graph of these conditions is shown in Fig. 3.

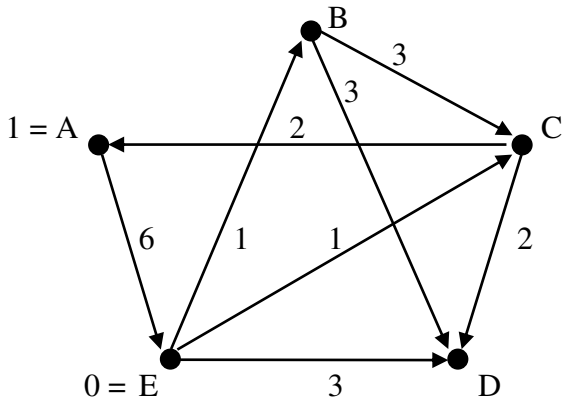


Figure 3
The graph of conditions in Example 3

From the fourth condition we know the values of two variables. There are some arrow-changing steps 3) and 4) for arrows starting from E and from B to C we can use step 16) etc., but we have no information about the values of the nodes B, C and D. See Fig. 4.

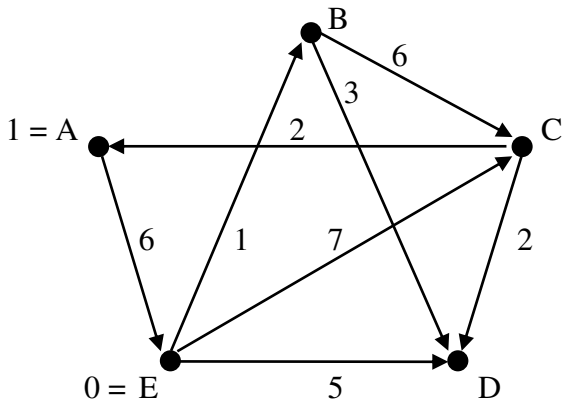


Figure 4
The graph of example 3 after changing arrows

Now we are at point 5 of the algorithm. A and E are fixed values. Let us choose 1 for the value of B. We can use step a) and we get C = 0 and D = 0. One can check it; it is a solution with Z = 4. Now we want the best solution, thus we make a BACKTRACK and let B = 0. Now we check the value Z of the remaining possible best solution: A and E were fixed before B got the value and we are trying the second value for B, so they are fixed. (It is easy to check that there are no more fixed values.) Therefore, the amount is -5 + (3+7) = 5, which is greater than the value of our memorized solution, thus we continue the search: We have no graph-changing steps to determine the values of C and D, so we are at point 5 of the algorithm and let D = 1. We cannot get the value of C by graph-changing, so we are at point 5 again, let C = 1. We can check, it is also a solution, with the value Z = 5. We memorize it, and we finish the search because we know from the previous calculation that it is impossible to get a better solution.

Conclusions

In this paper, various types of Boolean Programming problems have been considered. Using graph-theoretical approach we have solved these special 0-1 integer programming problems. In practice, we can approach similar problems, if we have conditions, for switches or Boolean circuits. We can solve such problems in which each variable is binary, i.e., it has a value of a characteristic function. In some logical exercises and puzzles the conditions are similar to the conditions investigated here [13, 15-17], or we can write them in the form of the conditions of BP problems and we can use graphs to solve them [11, 14]. We used linear goal-function in the optimization. The conditions are strong (i.e. strict) and weak inequalities and equations. In the graph representation, the critical edges, are used to represent the non-linear conditions. Our algorithm is based on local information: we can modify the graph by changing the weight values of the edges and by assigning values to nodes. Interesting properties of graphs are noted, such

as the arrows with weights non-divisible by 3 are specifying a transitive relation among nodes, while the arrows with weights at least 3 are representing a symmetric relation. In the future, we expect some more interesting phenomena by a more detailed analysis of our theory. Our method, similarly to knapsack algorithms, can give a relatively good solution in a short time in many cases. The algorithm uses backtracking graph-search strategy to find also the optimal solution of these problems.

We can use our algorithm for the case of arbitrary goal-functions, in step 5, we need to choose the most 'important' variable to be 1 (or the variable which is the least 'important' to be 0), where the importance property is specified based on the goal function. Our method can easily be implemented by using matrices of the graphs.

When we allow more than one type (2) condition for a variable in the LHS, we must use and-or graphs (i.e., hypergraphs) to represent the conditions. We hope that a variation of the presented method will work for other integer-valued programming problems for which we allow more values than 2 for the variables.

Finally, we note that the technique used here is related to methods to solve logical puzzles [14]. On the other hand, SAT solvers [1, 6, 8] provide valuations of the variables such that the given formula evaluates to true, if it is possible. The SAT problem is one of the most known NP-complete problems. SAT solvers can also be used to solve certain puzzles. Moreover, the problem of finding a satisfying valuation of a logical formula that optimizes a linear function of variables is called MinCostSAT [5]. Thus, we can see that Boolean Programming is an important and challenging topic with various new approaches.

Acknowledgements. The author is very thankful for the anonymous reviewers for their constructive comments helping to improve the paper.

References

[1] Bengt Aspvall, Michael F. Plass, Robert Endre Tarjan: A linear-time algorithm for testing the truth of certain quantified Boolean formulas, Information Processing Letters 8 (1979) 121-123

[2] Béla Bollobás: Graph Theory: An Introductory Course. Springer-Verlag, New York, 1979

[3] Peter Barth: Linear 0-1 Inequalities and Extended Clauses, Operations Research '93 (Springer) pp. 24-27

[4] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein: Introduction to Algorithms. MIT Press and McGraw-Hill, 1990 (3rd edition, 2009)

[5] Zhaohui Fu, Sharad Malik: Solving the minimum-cost satisfiability problem using SAT based branch-and-bound search. In Proceedings of the

2006 IEEE/ACM International Conference on Computer-Aided Design (ICCAD '06) 2006, San Jose, CA, 2006, pp. 852-859

[6] Weiwei Gong, Xu Zhou: A survey of SAT solver. AIP Conference Proceedings 1836, 020059 (2017) https://doi.org/10.1063/1.4981999

[7] Peter L. Ivanescu, Sergiu Rudeanu: Boolean methods in operations research and related areas. Econometrics and Operations Research, Vol. VII, Springer-Verlag New York, Inc., New York 1968

[8] Gábor Kusper, Csaba Biró: Solving SAT by an Iterative Version of the Inclusion-Exclusion Principle. 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2015) Timisoara, 2015, pp. 189-190

[9] Ken McAloon, Carol Tretkoff: Optimization and Computational Logic, Wiley-Interscience series in discrete mathematics and optimization, 1996

[10] Benedek Nagy, Márk Kósa: Logical puzzles (Truth-tellers and liars), ICAI'01, Eger, 2001, pp. 105-112

[11] Benedek Nagy: Boole programozás gráfok segítségével (in Hungarian, Boolean programming and related graphs) Szigma 23 (2002) 115-130

[12] Benedek Nagy: Truth-teller-liar puzzles and their graphs, Central European Journal of Operations Research 11 (2003) 57-72

[13] Benedek Nagy: SW-type puzzles and their graphs, Acta Cybernetica 16 (2003) 67-82

[14] Benedek Nagy: Boolean programming, truth-teller-liar puzzles and related graphs, ITI 2003: 25th International Conference on Information Technology Interfaces, Cavtat, Croatia (2003) 663-668

[15] Benedek Nagy: Duality of logical puzzles of type SW and WS - their solution using graphs, Pure Mathematics and Applications 15 (2005) 235-252

[16] Benedek Nagy: SS-típusú igazmondó-hazug fejtörők gráfelméleti megközelítésben (in Hungarian, SS-type truthteller-liar puzzles and their graphs), Alkalmazott Matematikai Lapok 23 (2006) 59-72

[17] Benedek Nagy, Gerard Allwein: Diagrams and Non-monotonicity in Puzzles, Diagrams'2004, LNCS, LNAI 2980 (2004) Cambridge, England, 82-96

[18] András Prékopa: Studies on mathematical programming, Mathematical Methods of Operations Research, Vol. 1, Budapest, Akadémiai Kiadó 1980

[19] Steven S. Skiena: The Algorithm Design Manual, Springer-Verlag, New York, 1997

[20] Raymond Smullyan: Forever Undecided, Alfred A. Knopf, New York, 1987

[21] Stuart Russell, Peter Norvig: Artificial Intelligence – A Modern Approach. Prentice Hall, 1995 (3rd edition, 2009)

[22] Béla Vizvári: On the optimality of the greedy solutions of the general knapsack problems, Optimization 23/2 (1992) 125-138