Using Very Large Volume Data Sets for Collaborative Systems Study

Ion IVAN, Cristian CIUREA
Economic Informatics Department,
Academy of Economic Studies, Bucharest, Romania
ionivan@ase.ro, cristian.ciurea@ie.ase.ro

This article presents the study requests for collaborative systems, the structure and volume of data necessary for collaborative systems analysis. The paper defines procedures for collecting and validating data. This article identifies algorithms to construct homogeneous collectivities. Calculations are carried out with very large data sets and the results are interpreted.
Keywords: *collaborative systems, data sets, metric.*

1 Dynamics of collaborative systems

The collaborative systems are an interdisciplinary field at the intersection of economy, informatics, management, and sociology. Collaboration involves organizations that have a common mission and join together to form a new structure [1]. A collaborative informatics system is also a distribution company whose goal is to sell increasingly quantities of his products.

The collaborative informatics systems represent, from the implementation viewpoint, software entities that are developed during a life cycle process that starts with the problem analysis and ends with the implementation of a fully functional software system.

The systems consist of components and interactions between them. When collaborative systems are used voluntarily, one of the key drivers to success is how users feel that their experience with the system: if they like, if the system offers them what to expect from him, if they are able to communicate freely and naturally with other participants and whether to recommend it to others [4].

Collaborative systems are classified according to the following criteria:

a) *level of complexity*, and by this criterion are identified:

- *collaborative systems with low complexity level*, have few components and the number of relationships is limited;
- *collaborative systems with medium complexity level*, have small number of components, but do not have large number of streams or *systems with large number of flows* and which have large number of components;
- *collaborative systems with large or highly complexity level*;
- *collaborative systems extremely complex*, have many components and many streams: banks, police, internal chain of hotels, airline transport; the banking system is among collaborative systems with very high level of complexity, because it consists of many components and is characterized by a large variety of links between them.

b) *type of application*, criterion which groups systems in:

- *collaborative systems in education*;
- *collaborative systems of defense*;
- *productive collaborative systems*.

c) *method of organization*, criterion which divide systems into:

- *linear systems*, in which subsystems interact with each other in both directions;

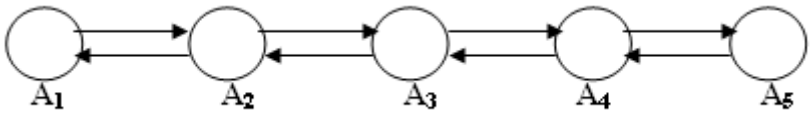

Fig. 1. Linear collaborative system

Between activities $\mathbf{A_1}$ and $\mathbf{A_2}$ is changed the message $\mathbf{M_1}$, between $\mathbf{A_2}$ and $\mathbf{A_3}$ is changed the message $\mathbf{M_2}$, between $\mathbf{A_3}$ and $\mathbf{A_4}$ is changed the message $\mathbf{M_3}$, and between $\mathbf{A_4}$ and

$\mathbf{A_5}$ is changed the message $\mathbf{M_4}$.
These types of collaborative systems are encountered in the field of education, each subsystem representing a graduate school.
- *tree systems*, organized by levels, as in figure 2:

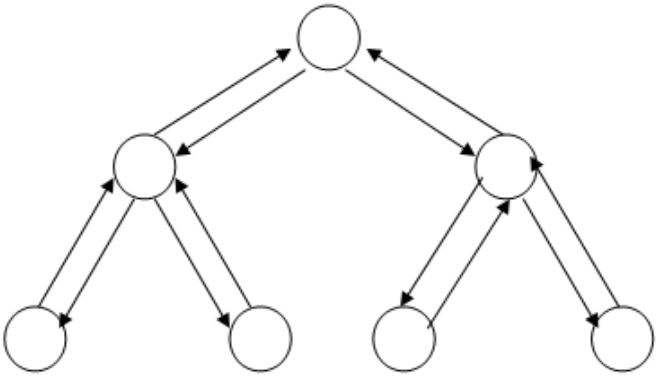

Fig. 2. Tree collaborative system

In a tree system, messages are moving between activities in a hierarchical manner, a message from the second level will reach the level zero only if he move and at level one, and a message of basic activity, represented by the tree root, will be propagated only to activities on the immediately below level. From this level, the message will be forwarded to the activities represented by child nodes of the nodes from level one;
Considering the collaborative system as a tree structure, there are taking into consideration:
- the degree of vertical collaboration as the number of links between components from level *k* to the ones on level *k+1*;
- the degree of horizontal collaboration as the number of links between components on same level.
Systems of this kind meet in organizational management and public administration.
- *network systems*, the components communicate with each other regardless of the level that is;
In the case of a collaborative system, network type, subsystems are all interconnected, that all transfers are interrelated. In such a system, messages circulate between all components without any restriction. Network type collaborative systems meet in the field of production and banking.
The business collaborative system works under the black box principle set out by Zadeh, the entries being given by raw materials and information and the outputs being materialized in finished products, services and other information which turns into costs for that business.

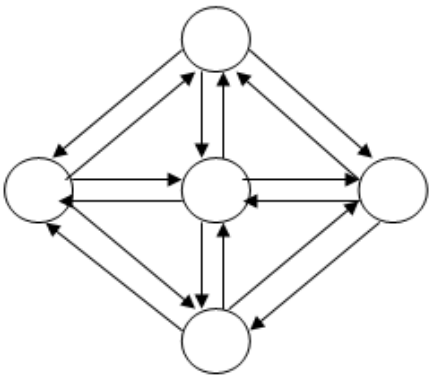

Fig. 3. Network collaborative system

Dynamics of collaborative systems concern changes regarding the quality, structure, functions, size, their procedures and standards. Dynamics of collaborative systems are studied using mathematical analysis, providing long-term behavior of each major systems, winning a look inside the system design: which parameters determine the group behavior and how the system characteristics are affected. Is developed a class of mathematical models which describe the collective dynamics of the collaborative system and which illustrates the approach by applying to several case studies, including both software agents and robots. For each system, is trans-

formed a set of equations that describe how the system is changing in time and analyze their solutions. Finally, is shown what say these solutions about the collaborative system behavior.

2. Decisions in collaborative systems

The decision system is very closely related to the information system. The link between the decision and the information systems is as follows:

- first of all, at the entry, the information system provides to the decision system the data which it needs and which it process giving them the appropriate form to be used in the decision process;
- secondly, the data, once processed and taken the decision, this is captured by the information system and directed to compartments where it must reach to be put into practice;
- thirdly, the decision system requires even inside it a interim circuit information under conditions in which the collaborative system is led through participatory management.

Decisions relating to collaborative systems are divided into:

- *current decisions*, that relate to daily decisions, contribute to achieving individual objectives and prevails at medium and lower management. Inputs and outputs of such decisions concern the daily problems taking place in a collaborative system. The costs regarding a wrong current decision are small and risks are also reduced;
- *short-term decisions*, which concern the decisions for periods of maximum several months and contribute to achieving the group objectives and prevails at medium and lower management. The costs incurred in making such decisions are higher than for current decisions;
- *medium-term decisions*, which concern the decisions for periods ranging from several months to a year, contributing to achieving the objectives of the group and prevailing at medium and high management. Costs and risks incurred when a medium-term decision is wrong are quite high;
- *long-term decisions*, that relate to decisions for periods ranging between 1 and 3 years, contributing to achieving the group objectives and overbear at high management level. A wrong decision of this kind involves very large risks and costs and have catastrophic consequences on the collaborative system involved;
- *strategic decisions*, that relate to a period between 3 and 5 years and contributing directly to achieving the fundamental or derivative objectives and aims the overall system activities or its main components.

Economic decision is the action line consciously chosen in the driving the system process, from a certain number of possibilities in order to achieve some objectives in an efficient maximum. The decision put the resources at work, establishes and achieves the system objectives. Requires a training and preparation in which participate a large number of individuals and departments within the system.

Environmental decision consist in all heterogeneous and exogenous elements of a collaborative system, which make up the decision situation characterized by the expression of direct and indirect influences on the content and significant results of the decision. In environmental decision is an evolving contradictory: on the one hand are a number of changes likely to provide the best prerequisites for an effective decision process, and on the other, environmental decision tends to become increasingly more complex due deepening social division, reducing the life cycle of products and accelerating the pace of moral attrition. In decision making, these elements are expressed in a number of variables and limit conditions and in the involvement of interdependencies between these.

3. Procedures for data collection

Since the first computers become fully operational in 1950, people have realized their power to collect, manipulate, classify, store and retrieve data in a much faster, more flexible and more effective manner than human power. The volume and complexity of information exceeds the processing capacity of one person. That is why the information systems allow to store data for later analysis and

provides methods for processing and filtering to extract only relevant data. By using mathematical and statistical algorithms that implement known methods for understanding the data is obtained information with new meanings.

Computers have become indispensable for the collection and handling of information, when information is made in high volumes.

With the specification of a collaborative system logical project, are designed procedures for collection, validation, transmission, storage and processing of data entry. Entries in a collaborative system are achieved by retrieving the data:

- from primary documents, introduced by operators;
- transmission by all the modalities, including by satellite;
- by scanning;
- in the form of pulses from machinery of technological process.

Document processing involves extracting information contained in the data fields and converts them to series-bit format that allows storing them in a database. Forms processing is considered complete when all information in the documents have been produced, verified and saved in a database. Collection of data on forms is made either involving an appreciable number of people who collect them and bring them in a program using the computer keyboard, or by entering data automatically in the system.

Among the benefits resulting from the use of a solution for automatically data collection from the forms printed on paper, identifies:

- increase efficiency and accuracy in taking and processing the information;
- reduce the number of personnel used in such an activity and implicitly, minimize human errors;
- increased accessibility of information collected and easily in their use;
- the storage of documents image on magnetic disks lead to substantial reduction of space allocated to the classic records, and related costs of administration;
- shortening the processing time of information as much you want to achieve the STP (Straight Through Processing);
- improving the quality of services tendered by beneficiaries to their customers.

A brief description of how is working an application of automatically data collection is as follows:

- the set of forms is scanned using a speed scanner;
- most data is recognized automatically;
- a small number of characters that the program is not sure he is recognized them make that the image of scanned document to be passed to a human operator for verification and correction;
- the verified data are saved in a database.

Documents archives require development of algorithms for flexible retrieval of components or groups of components to meet requirements of collaborative decision-reliance.

Confidentiality is ensured by the fact that the information is delivered to authorized person only and only at a permitted location as access control and authentication are very important to conduct processes on the transfer of messages. Access to the database is done only by using an access procedure. The access procedure is structured on hierarchical levels and on parameterized access passwords. This procedure is available to the user for customization.

The quality of information provided by a collaborative system to the decision factors and users depends on the quality of input data in a collaborative system depends. It is desirable to prevent, if possible, plugging in the system incorrect data. The validation data procedures are grouped as follows:

- traditional procedures, which consist in checking by eye data from primary documents;
- automatic control and data validation procedures, in which validation is achieved through validation programs developed by programmers and procedures for automatic control and automatic data correction that identifies and corrects errors automatically, without human factor intervention.

Procedures are built for validating data: alphabetic, numeric and correlations between

fields. For each application are used standard procedures and the error messages should be much nuanced so as to help troubleshooting a breeze. These error messages must show to the user where and what is wrong.
The presentation manner of the results of validation is as follows:
- displays a confusing message which specifies that are errors in data;
- on a different form than they have completed the data shows a list of errors stating the field and the nature of the error;
- the form in which data is completed, at the right of the fields wrong filled appears the error messages, colored in red.
There are more options for the data validation, namely:
- *option 1:* rigid validation, meaning that the data entered are validated field with the field and the user can not move forward until it completes correctly previous fields. This validation is accompanied by a reminder regarding the nature of the error in the form, like a text message or a sound signal;
- *option 2:* flexible validation on the form, displaying messages about erroneous data;
- *option 3:* the form fields erroneous become red;
- *option 4:* on the form, after sending the data, a list of errors appears [3].
Should apply any of the four validation options, the application must solve a single problem: the information recorded in the collaborative system must be accurate and complete.

4. Dynamics indicators of collaborative systems

These indicators characterize homogeneous subsets of data and are divided into the following categories:
- *indicators operating with time moments*; in a collaborative system represented by a bank which has many branches and agencies located in many countries, it is interesting to discover the hourly intervals during the day or the days from a week with most of the cash withdrawals; to find the schedule, consisting of two hours, during most of the cash withdrawals in a day, is measured every two hours, the cash withdrawals. During the day there are twelve measurements. The schedule with the numerous withdrawals of cash is determined as follows:

$$IHM = max\ \{IH_1, ..., IH_{12}\}, \text{ where:}$$

IHM - the schedule of a day, with the most numerous cash withdrawals;
IH_i - the volume of cash withdrawals from the interval of time i; to determine the day with the most numerous withdrawals in a week shall be similar, with the difference that the maximum is calculated from seven values, representing the daily withdrawals;
- *indicators operating with levels of a variable*, in the bank collaborative system, indicators operating with levels of variables are the volume of deposits and the volume of daily withdrawals;
- *indicators operating with time moments and levels of a variable*.
These indicators make a metric of collaborative systems.
If are considered the collaborative systems $S_1, S_2, ..., S_n$, we can build and other indicators for the implementation of quality metric of collaborative systems. For each system S_i are collected the data $d_{i1}, d_{i2}, ..., d_{im}$ regarding its dynamics. Through the intersection of $d_{i1}, d_{i2}, ..., d_{im}$ values are obtained some data, which is common to all collaborative systems. These information are necessary to create new indicators $I_1, I_2, ..., I_h$. It selects from these indicators some of them which must be sensitive, stable, representative. With the new indicators we evaluate what is unique in collaborative systems.
Also, the indicators for quantification of characteristic levels for maintainability, reliability, portability, complexity has a variety of analytical expressions, from homogeneous expressions to reports of homogeneous expressions, leading to constructions in which logarithmic and exponential function appear.
The analytical forms of the indicators must be built such as the indicators simultaneously assure the following conditions. They must be:
- *sensitive*, that is at small variations of the influence factors the result variable has small variations; at big variations of the influence

factors the result variable has big variations;
- *non-compensatory*, that is at different variation sets of the factors, small values of the result variable are not obtained;
- *non-catastrophic*, that is at small variations of the factors, big variations of the result variable have not to obtain;
- *representative,* it represents the quality to be accepted by users in analysis making assuring the significance of the results.

The economic process evolution is represented by the continuous dynamic models: by differential equations or by systems of differential equations, as outlined by a single main indicator or a set of indicators related with the model equations, both among themselves and with the factorial variables which makes the process. Continuous linear dynamic models of cybernetic systems are frequently encountered in researching the dynamics of economic processes and are represented by linear differential equations.

The quality of a collaborative system is defined as all features and characteristics, bearing ability to meet the needs specified or implied. To measure the quality of a collaborative system and assess its performance is used the indicator:

$$I_{calit} = \frac{\min(A,B)}{\max(A,B)} * p + \frac{\min(X,Y)}{\max(X,Y)} * q$$

, where:

A – the amount planned;
B – the amount realized;
X – the quality planned;
Y – the quality achieved;
p – represents the share of the quantitative characteristics (generally amount 0.4);
q – represents the share of the qualitative characteristics (generally amount 0.6) [5].

In [6] are presented algorithms implementations for classifying data, using the top-down inductive method for decision tree construction. They are built on the testing of each node of the tree, beginning with the root node, for each entry. Each node represents the name of an attribute. The instance is inserted in an existing class, on the basis of common features, evaluating the appropriate attribute of the node reached. Depending on the value, the instance crosses a branch. When no more nodes are evaluated, the instance is classified. If a particular class do not differ in an obviously manner from another, following the introduction of more and more records, the two classes are merged, a process recognized as the *pruning* process.

The development of collaborative systems is accelerated, along with the wireless networks and, the quality characteristics become strictly related to the security characteristics. The extensions to metrics should include, in the future, indicators of collaborative systems security.

5. Software for very large datasets management

The storage of huge volumes of data in a company and retrieving such data, in 1990 led to the development of new technologies, building:
- Data warehouses;
- Data concentrations;
- Data extractions.

Artificial intelligence has a special role to develop and implement technologies for creating and manipulating extremely high data sets and with very high level of complexity [2].

Were properly developed several software products for managing large volumes of data, namely:
- Oracle Datawarehouse, used for data organizing in data warehouses following the multidimensional model;
- Oracle OLAP (**O**n **L**ine **A**nalytical **P**rocessing), for retrieving data from data warehouses, using OLAP technology;
- Oracle Discoverer, to obtain output situations from data warehouses, including obtaining dynamic reports;
- Oracle Miner, used for advanced statistics, for the discovery, knowledge and information extraction from data.

The WEKA product, the Waikato Environment for Knowledge Analysis, free available from specialized department of the University of Waikato, Hamilton, New Zealand, allow automatically learning techniques to practical problems and integrates various automatically learning tools that are used in a

typical work environment, characterized by a uniform interface. Users use the wide range of automatically learning techniques for extraction of useful information from the very large database. It should be noted that WEKA is used in any area of interest, thus having a major advantage over other applications of data mining, especially on those commercials, which are intended for a single area. WEKA contains tools for data preprocessing, and for data classification are used decision trees, regression, clusterization, rules for association and visualization. The application is developed in Java and the source code is open, released under the license GNU General Public License. This is a big advantage of the WEKA system unlike other applications, because it allows changing the system by users in how they need it, possibly with the development of new automatically learning techniques and implementing their own algorithms. Also equally important is that the system is used on multiple platforms: Unix, Linux and Microsoft Windows.

The latest version available to users is WEKA 3.4.3 and is installed on the Windows platform and other platforms such as Linux or UNIX. It should be noted that for MacOS X is not available at this time than WEKA 3.4.2 version. Java 1.4 virtual machine must be installed on the system in order to run WEKA. An earlier version of WEKA is WEKA 3.0, which is a command line application. When launching WEKA, a GUI Chooser window appears that allows users to elect to work in CLI command line or for opening the work in Explorer graphical interface. WEKA Explorer provides in graphical interface the system packages, namely:

- *preprocessing,* in which datasets are opened both as ARFF files and from a specific database; also the system allows a unattended data filtering with one of the available filters;
- *classify,* allowing the choice and run of any classification algorithm from the defined six categories of algorithms;
- *cluster,* in which is chosen and is running the method of data clusterization;
- *associate,* allowing the setting and application of a rule for data combination;
- *select attributes* is another WEKA package and allows the configuration and implementation of any combination of attributes from these that define the data set for detect which are the most relevant attributes from the data set;
- *visualize* allows viewing the current dataset in one or two dimensions, and if the attributes have continuous values is used a spectrum of shades of the same color to represent the values, while for discrete attribute each value is represented by a different color. Additionally to these tools packages for working with datasets, WEKA contains a classifier based on decision trees WEKA CLASSIFIERS TREES USERCLASSIFIER and a graphical interface for building neural networks WEKA CLASSIFIERS FUNCTIONS NEURAL NEURALNETWORK.

The data set used in the WEKA program must be in ARFF format in order to be processed. The data come mostly from an Excel table or from a database and must be converted into ARFF format, the most widespread for databases in text files. Using this format in parallel with the direct support for databases is another advantage of WEKA. In addition to these positive elements that characterize the WEKA system, there are some disadvantages, namely that requires the learning use of interface, understanding algorithms and how are interpreted the numerical and graphical results. In addition, WEKA uses statistical terms instead of using data entry corresponding terms (e.g., in economic applications), as do other software products, business specialized and more intuitive for a manager or an economist.

To evaluate collaborative systems is implemented a software product that allows you to get levels for associated metric.

Data quality is ensured in the situation of analysis and processing of a large volume of data. On these data are made operations such as sorting, regrouping, concatenations or elimination of aberrant values. As with a dataset reduced like volume, in the case of a very large data volume certain indicators are calculated and highlighted some features of this,

based on the following ranges of values, empirically obtained:
- *[0; 0,78]* is the range that express the bad character of the calculated indicator;
- *(0,78; 0,92]* express the good character of the indicator;
- *(0,92; 1]* express the very good character of the indicator.

For a dataset of 300 values are calculated the following indicators, whose obtained values are shown in Table 1:

Table 1. A dataset analysis

Indicator 1	*Indicator 2*	*Indicator 3*	*Indicator 4*	*Indicator 5*	*Indicator 6*	*Indicator 7*	*Indicator 8*
C_1	G_1	H_1	P_1	A_1	F_1	S_1	B_1
C_2	G_2	H_2	P_2	A_2	F_2	S_2	B_2
C_3	G_3	H_3	P_3	A_3	F_3	S_3	B_3
...	...	...	...	...	...	...	...
C_{300}	G_{300}	H_{300}	P_{300}	A_{300}	F_{300}	S_{300}	B_{300}

Where the indicators in Table 1 do not fall within the range [0; 1] will give the normalization of indicators, according to the calculation relationship:

$$VN_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \in [0; 1], \text{ where:}$$

VN_i - the normalized value of the indicator i;
V_i - the value of the indicator i;
V_{max} - the maximum value of the indicator i;
V_{min} - the minimum value of the indicator i.

Normalized values of the indicators are sorted descending. Is chosen a distance h and is determined an item I_{min}. Are calculated the levels $I_{min} + h$, $I_{min} + 2h$, $I_{min} + 3h$, $I_{min} + 4h = I_{max}$ and are builded datasets that belong to the intervals:
$[I_{min}; I_{min} + h)$,
$[I_{min} + h; I_{min} + 2h)$,
$[I_{min} + 2h; I_{min} + 3h)$,
$[I_{min} + 3h; I_{min} + 4h)$.

The sets are more homogeneous and allow the application of defined indicators to the metric, obtaining representative levels.

The real problem is to apply the metric and most important to validate it. This will give the confidence that the values are real and the results are reflecting the actual image of the problem. Once the model is defined, it must be implemented in real development or maintenance cases and it must be tested.

6. Conclusions

The collaborative system is developed based on a set of specifications that were defined in the analysis stage in order to define objectives for the development process. The system must behave and must give the results the users want and that they have stated at the start.

Increased volume of information and improvement of operating software products have led to a new quality of data use, as a analysis that shows to the organization management the information difficult or even impossible to obtain in other ways. It is obtained such information on customers' preferences, their profile or distribution. Is provided to the management the data, such as what region of the country is better selling a product, which are the preferences for a specific market segment.

It is obvious that such information is obtained only by using certain processes, such as multidimensional analysis, some statistical methods of forecasting and other mathematical methods applied to a very large volume of data. These mathematical methods require the use of a extremely complex specialized software.

From economic perspective, the trade globalization, a dramatic sharpening of competition, the spectacular shortening of the products life-cycle due to the dynamics of technology, the imposition of extremely high quality requirements, and other such developments, has shown even more the strategic value of information. Intelligence information manipulation requires new management models, more flexible and more functional. The need to respond in optimum time to market requirements has led to decentralization and to reduce the number of decisional

levels, and enshrines the so-called flat hierarchies, which are based on delegation of operational decision power to the second management echelon. Basically, classical officer is being replaced by workers with information. Both at strategic management and at operational management level, the need for pure information, correct and significant has become vital for collaborative systems.

The knowledge-based society evolves only through the high quality of citizen-oriented collaborative systems.

The current research was financed from European Structural Funds, project no. 7832, "PhD and doctoral in the triangle Education-Research-Innovation, DOC-ECI" (cercetare finanțată din Fondurile Structurale Europene, proiect nr. 7832, "Doctorat şi doctoranzi în triunghiul Educație-Cercetare-Inovare, DOC-ECI").

References

[1] R. Arba, "Collaborative Electronic Marketplace," in *International Workshop „Collaborative Support Systems in Business and Education"*, Cluj-Napoca, pg. 11, October 2005.

[2] C. Bodea, *Baze de date: tendinţe în evoluţia tehnologiilor şi aplicaţiilor*, Editura Inforec, Bucureşti, 2001.

[3] I. Ivan and C. Ciurea, "Entry data validation in citizen oriented applications", paper of *The "4th International Conference on Applied Statistics"*, November 20-22, 2008, NIS Publishing House, (Bucharest, Romania), ISBN 1018-046x.

[4] I. Ivan, C. Boja and C. Ciurea, *Metrici ale sistemelor colaborative*, Editura ASE, Bucuresti, 2007.

[5] I. Ivan and C. Ciurea,"Quality characteristics of collaborative systems," in *Proc. The Second International Conferences on Advances in Computer-Human Interactions, ACHI 2009*, vol. I, (Cancun, Mexico), pp. 164-168, February 2009.

[6] R. Quinlan, "Data Mining from an AI Perspective," *Proceedings of the 15th International Conference on Data Engineering*, Sydney, Australia, March 1999.

Ion IVAN has graduated the Faculty of Economic Computation and Economic Cybernetics in 1970, he holds a PhD diploma in Economics from 1978 and he had gone through all didactic positions since 1970 when he joined the staff of the Bucharest Academy of Economic Studies, teaching assistant in 1970, senior lecturer in 1978, assistant professor in 1991 and full professor in 1993. Currently he is full Professor of Economic Informatics within the Department of Economic Informatics at Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies. He is the author of more than 25 books and over 75 journal articles in the field of software quality management, software metrics and informatics audit. He is distinguished member of the scientific board for the magazines and journals like: Informatica Economică, Economic Computation and Economic Cybernetics Studies and Research, Romanian Journal of Statistics.

Cristian CIUREA has a background in computer science and is interested in collaborative systems related issues. He has graduated the Faculty of Economic Cybernetics, Statistics and Informatics from the Bucharest Academy of Economic Studies in 2007. He is currently conducting doctoral research in Economic Informatics at the Academy of Economic Studies. Other fields of interest include software metrics, data structures, object oriented programming in C++ and Windows applications programming in C#.