Polskie Towarzystwo Informatyczne

Selected contemporary problems of information systems

Editors:
Kapczyński Adrian
Brückner Adrian

Selected contemporary problems of information systems

Editors:

Kapczyński Adrian
Brückner Adrian

Wydawnictwo PTI Oddział Górnośląski
Katowice 2009

Recenzenci:
prof. dr hab. inż. Andrzej Grzywak
prof. dr hab. inż. Jerzy Klamka

Redakcja techniczna:
dr inż. Adrian Brückner
dr inż. Adrian Kapczyński

Projekt okładki:
Jacek Uroda

ISBN: 978-83-60810-28-6

Polskie Towarzystwo Informatyczne
Oddział Górnośląski
40-040 Katowice, ul. J. Lompy 2/10
Tel. +48 (32) 253 61 09
www.pti.katowice.pl
e-mail: biuro@pti.katowice.pl

Wydrukowano w Pracowni Komputerowej Jacka Skalmierskiego
Nakład - 100 egz.

SPIS TREŚCI

Part 1. Information systems

Part 2. Social networks and knowledge engineering

Foreword

This book presents current results of scientific work related to problems of information systems.

First part of the monograph contains 5 chapters from information systems domain. The second part is related to social network and knowledge engineering problems and consists of five chapters.

Editors would like to express their gratitude to all authors who contributed in this book.

Adrian Kapczyński
Adrian Brückner

Part 1.
Information systems

Chapter 1

About the inadequacy of computer systems

Wiesław Byrski
Wyższa Szkoła Zarządzania i Bankowości
Uniwersytet Humanistyczno-Przyrodniczy
wiebyr2001@yahoo.com

Abstract

According to the dominant view the computer systems are a part of information system which realizes the functions of information system by the means of computer and informatics tools. The final objective of the informatization project is the creation of computer informatics system helping or substituting the chosen functions of information system. In initial analysis of the informatics project the influence of appearing solution to the whole is certainly considered, what i.e. in methodology SSADM has a distinct phase BSO – business system options, and in many analysis a strategic phase is distinct. Improvement of the information system activity where a predilection of new in connection to the old is made. It seems however that many of the contemporary solutions cannot keep to the pace of the changes and propose informatics systems which are surprisingly ineptly using the abilities of the contemporary informatics/ teleinformatics. The basic reason is the lack of strategic planning including the entire modified environment. The other reasons are: 1) lack of a reasonable theory describing modeled phenomena; 2) lack of informatics knowledge of the deciding body; 3) bad economies and other. Probably one of the reasons is also knowledge deficiencies of informatics engineers but the latter issue will not be discussed in the following chapter.

1. Introduction

According the popular opinion computer and informatics systems should be a part of information systems, i.e. should realize functions of the information system by the means of informatics tools . Information systems of all organizations have objectives precisely defined. The main goal and mission of the SI is supporting the operational activities of organization and helping it to achieve its strategic aims. In this article we are going to talk about the operational objectives of the organization better defined in the context of the strategy.

Usually at the moment of creating of the organization its information system is constructed according to the principles developed in the given branch. It is treated as operational, so used to the realization of well precised tasks that the organization has worked out in the course of its activities and realizes within its information system. In new and unknown situations new solutions are created commonly basing on well known analogies. New solutions are obtained not only thanks to the new technologies but also creating new organizational procedures. The invented solutions are surrendered to a fast modification raised by mutual influence if the new technology, new environment and new organizational procedures. Sometimes, by the force of inertia, new solutions are in function for many years. The famous Red Flag Act dating from the beginning of the XX c ordering an assistant with a red flag to run in front of the technological news which the car used to be was legally binding for 20 years.

Certainly, in the process of projecting the computer systems the dynamics of the processes cannot be ignored, nor the changes of the organization objectives, changes of the organization itself, its near and far environments and other less predictable changes. All methodology of developing the informatics projects should deal with changes in the project, some do it better while some worse. One of the oldest Methodology SSADM in the early phase of the projecting of the computer system considered the phase of analysis of a system business influence on the organization. It is, therefore, outside the interest of the software enginery to analyze the influence of the changes that the new informatics system brings to the existing information system of the organization. It is according to the good practice of other enginery sciences: the person preparing the projects of the bridge is only to some extend analyzing the influence of the construction in progress to the entire communication system (it is usually the task of communication engineers). The person inventing a new type of armored transporter is not involved in the changes of the war strategy, however he might expect such changes.

There is nothing strange in it as those are system changes in the environment, rather difficult to predict, caused by some unnoticed interactions and side effects.

2. Modeling of the reality and Okham's raizor

We build computer and informatics systems and their functional elements (data bases, controlling system, decisions systems, and others) assuming that they will be models of a chosen area of the reality. It seems that this conception has lead to the inflexibility of the "reality" from which engineers expect to be the simple, it is the paradigm of the science as a principle of the Okham's razor. The world of the social reality is unordered from its nature, most of the notions are not clearly defined and inaccurate and they require an individual, here-and-now interpretation. Methodology of creating the systems tolerate such reality but the rigid norms are forced whenever it is possible. It seems however that while creating a reality model the principle of phenomena model simplification, introduced by Okham, is abused. Below some examples illustrating the problem.

2.1. Example I – post address and its role in the information system

In a telecommunication enterprise the problem of issuing bills was made automatic quite early. After counting the monthly payment a bill was send to client to his/her address found it the database. The address in its semantic sense is indication of a place. There are many ways to indicate it, beginning with description as "next to an old house with red roof" and ending with geographic coordinates. It is "functioning" correctly if the message is delivered. According to traditional way of writing it could look as:

Kraków, Karmelicka 10/5
Rynek Główny 20, Kraków
34-015 Świątniki Górne 145
Pałac Kultury IX fllor, room 1245

Some small mistakes in the address, misspelling or even errors were corrected by the postman, i.e. the post information system. The system had also capacities of self correcting: if the message sent did not caused expected reaction (payment of the bill) it was a signal about a possible mistake in the address.

It seemed that the most updated addresses of the users are in the database of the payments system. But an attempt to use the data base of the phone users while compiling a phone directory showed non-usefulness of the model to this objective. An address was modeled in the data base as STRING(70) which was working enough for the purposes of the basically planned system and its special function but it was (almost) useless to the new use where an address should have a semantic structure. As a result from one screen the data were written by hand to the second data base with a correct structure.

Nowadays an address has a standard structure set by the post office and all systems are using the same model. It is clear that it is time to identify an address with the GPS indications.

2.2. Example II: when a telephone switchboard is "ready"?

Computer system of census for a big telecommunication enterprise was to stand for the reporting system used until. One of the most important tasks was generating the Yearly Report that in the traditional system lasted for over half a year. In the new a Report should be ready just while pushing a button. While projecting an integrated database a simple question raised: when a telephone switchboard is "ready" and when it could be closed in the yearly report?

a) when a real plan was made
b) when it was bought
c) when it was installed and it worked
d) when it was connected
e) when the first client was connected
f) when the trial period was successfully passed and completed?

In traditional enterprise system of reporting each subsystem had its own databases modeled and defined to support its information system. The field "switchboard readiness date" was filled with the date resulting from the application's need and it was not important that it was different for different subsystems. For planners it was an early phase of the project, for accouters- the moment of payment, for technicians, for service, for... Each of the computer systems was acting correctly supporting "its" informatics system. From the moment when the systems were integrated all old notions needed to be redefined.

2.3. Example III: evolutional changes

A group of young computer specialists from Warszawa worked out a very interesting system of application registration for telephone allotment (beginning of changes in Poland, 90s). The form taken by the engineers as a screen and database model was an essential part of the application. System was designed to operate on few departments of the local office. In effect identical forms were in the whole organization, authors knew also all the procedure, among others also the procedure of filling in the form. They have not interviewed the users in other service departments, because they have taken for granted an assumption of complete identity. The reality was revealed to be different: indeed few years beforehand all forms in all departments were identical, the same as all procedures and pragmatics, but after few years some unexpected small or bigger changes occurred, however no-one was centrally coordinating the implementation, every department was doing it according to local ways. The aim of corrections of the content of the form, its factual content, procedures of filling up etc. was a necessary updating and adjusting to the changes, but without a formal change of the form. In the paper version it was easy: some margin notes were added, in free place, overwriting of the blanks, crossing out and adding notes, briefly all changes that are permissible in the current "paper" technology. There was no necessity to coordinate those changes in all departments because local information systems (almost) never contacted each other. When all 6 departments has "seen" each other the incompatibility and divergence of the procedures were revealed, the standardization of the procedures exceeded local competencies.

2.4. Example IV: specifying definitions

In the paper ID it was possible to write any (with some general limits) sequence of letters. The system of issuing paper IDs worked in decentralized version, after changing the system to a centralized one, which was possible due to the informatics, some unexpected and unpredictable limits appeared.

The first version of the system of plastic IDs did not expected names written in another alphabet than the Polish one, so the names of our compatriots written in Czech, French alphabet or with umlauts and other will disappear. It caused protests, initially ignored by the offices – it was possible to make transcription. Only when in Lithuania the same problem occurred with Polish names, but a rebous (Świerczewski as Swierczewski, Mączyński as Maczynski, etc) the correction of the newly implemented information system was decided. Information system, not computer informatics system, as in passing this system that the informatics was to support, was being changed Informatics could easily deal with any alphabet, thus not exactly defined rules has caused the conflict.

2.5. Analysis of the requirements as the choice of the model

In all methodologies of system creation there is an initial phase of analyzing the requirements. This phase could be simply defined as the choice of the model of the fragment of the reality. Every model is either based on a theory or is searching a theory itself. Often there is a promising theory suggesting the choice of a model but the lack of technical possibilities prevents its realization. It is not working the other way round – technical possibilities cannot solve the problem which could be solved only when we choose the appropriate theory .

In practice the appropriate solution is searched through many attempts and mistakes, little by little reaching the appropriate use of - thanks to the new technologies- existing possibilities, a peculiar marriage of new technology with new theories knowing how to use the possibilities of the new technology is necessary. It is not possible to realize this conception without the informatics. Drucker, one of the creators of the modern management, considered for a long time the benefits of informatics claiming that it is not able to use its own abilities and that the observed progress could be much faster . According to Drucker at last an example of authentic progress brought by the informatics was the ABC theory (Activity Base Costing) that, in brief, consists in creating thousands of subaccounts and entering on them incomes and expenses using simple rules, that in the end enables preparing a completely new model of finances for the enterprise.

3. The role of strategic phase

Before a decision on the project realization is taken there are some other decisions considered, their consequence is, successful or not, fitting the computer information technology system to the existing information system. In this phase we can observe a big similarity to the problems solved while strategic decisions are taken. The problem is analogical to the crucial in the strategic management theory and still not solved: is possessing of the strategy giving an advantage over concurrence? It seems that having a strategy is indispensable but there are still no empirical evidences for an alternative "strategy" of a passive reacting toward occurrences [4]. While regarding the problems of information technology it means asking the question whether the presently implemented computer information system is implemented according to the assumed plan of the strategic informatization. Moreover, it often happens that introducing the information technology to the information system changes its strategic objectives, for instance enabling achieving something that used to be neglected as unattainable.

4. Conclusions from the system inadequacies

While introducing the informatics "mechanically" to the traditional information system, i.e. supporting the function of the information system with the informatics technology, more than once we observe some conventionalities of the assumptions and relativity of the procedures. We can expect serious social conflicts when our informatics systems start requiring a literal sticking to the rules assumed as justified. It could be possible even now to solve the problem of the drivers not obeying to the legal and organizational driving rules using present informatics solutions. As a matter of fact why a car is allowed to drive at a speed over 130 km/h if there is no place where you can exceed this speed limit?

The role of the strategic phase in the system designing is developing. A level of 20% of successfully realized projects is lowered by an unknown number of projects formally, i.e. according to the initial assumptions were successful, but they are not fitted to the new conditions. Huge costs of informatics technology causes situations when even if an informatics system is not a complete success it is used anyhow for a long time. Some reasonable economies could be made only when carefully planning informatization strategy, beginning form recognizing the organization objectives, correctly presenting the strategic objectives to the information structure then to the informatics and finally an information computer system can be designed.

Bibliography

1. Bartczak I.: Interview with Peter F. Drucker. Computerworld, 7 czerwca 1999r, nr. 23/387.
2. Byrski W.: Theoretical base of information systems strategy (in polish) in: in: Strategie informatyzacji, str.11-19, red. Z. Szyjewski, J.S. Nowak, J.K. Grabara, PTI Katowice 2006.
3. Kisielnicki, J., Sroka H.: Bussiness information systems (in polish), Placet, ed.III, Warszawa 2005.
4. Kreikenbaum H.: Strategic planning in organization. PWN, Warszawa 1996.

Chapter 2

The process of creating electronic documents using predefined forms

Marek Valenta*, Robert Marcjan*, Janusz Chełmiński**
** AGH University of Science and Technology in Cracow*
*** Electro Croon Poland*

valenta@agh.edu.pl, marcjan@agh.edu.pl, jchelm@croon.nl

Abstract

This document contains a presentation of achievements of a certain stage of research[1] aiming at realization of the idea of the Information Society within the scope of citizen - public administration relationship. A common element of this relationship is a repeated submission of forms completed in a written form according to legally accepted specimens. Authors, based on the example of the process of submission of documents of statements of means, present a conception and projects of realization of the processes of creating electronic documents of these statements based on structuring of these documents and multi-functional forms which describe them. For the realization of this conception there is created an infrastructure which, together with dedicated applications, creates an environment suitable for the implementation of the proposed solutions.

1. Introduction

In the era of the development of Information Society great importance is attached to the development of teleinformatic infrastructure, operating on the side of institutions and oriented to wide information transfer. However, not less

[1] Research carried out under a research project of The Ministry of Science and Higher Education, grant no. 0032/R/T00/2008/06, "Acquiring and verification if information for the needs of supporting organizational and operational actions of public safety services"

important issue is the construction of infrastructure for citizens, which would help them to provide information required by both state public and local administrative units. The need to create systems that could be the basis for realizing such functions is almost evident, especially when the transmission of such information by a citizen to an office is a statutory requirement. We have to deal with this particular situation in the case of an obligation to submit, by a large part of public citizens, statements of means [7].

The authors of this article, in realizing research within the scope of functional and non- functional assumptions of a system designed to assist the citizen obliged to submit statement of means, created a conception and determined general conditions of the technical implementation of such a system. Given the legal conditions, current solutions suggest the existence of client applications for people obliged to submit statement of means, but the way to solve the problem, allows significantly to extend this ideas with institutional systems. The project also takes into account the need to deliver to users, that is, the distribution function, appropriate documentation specimens, different for different groups of people obliged to submit statement of means. This diversity of documents forced us to start the project and invent solutions which would be, in a large extent, universal and subjecting to the process of elaboration at the stage of defining the specimens of these documents.

The content of the meta data of documents specimens is not only the definition of information structure of the created e-documents according to their original versions in written form. It is also the "parameters" allowing to realize complex procedures of data validation, and comprehensive contextual help for users. They also include detailed definitions of both editing functions of documents and formatting of ultimate documents, both in a paper and electronic form, the existence of which makes the transmission of documents outside the environment of the described system possible.

Presented solutions through their additional functionalities allow a safe storage of edited documents, basic analysis of the data included in various documents as well as support within the scope of use of the existing documents with the data to create their next updated editions as well as completely new documents. Preserving confidence of data in application designed for a group of users is guaranteed by data encryption and a user management subsystem with the authentication functions.

In the near future presented solutions could become an important part of a larger, integrated management system for statements of means. However, its implementation, at least in Poland, must be preceded by legislative changes that allow the circulation of electronic documents in this field. Such legislative changes can be the basis for using also institutional applications for a direct data collection using form techniques. And then, the presented solutions can be more widely applied in a number of dedicated systems designed for different groups of users. These systems, with a so-called web client, using widely available network environments, will be able to create a completely new quality in citizen service in his relations with public administration.

Natural applications of this type have their own enormous and widely known advantages. But there are cases in which the possibility of generating and editing electronic documents according to a well known form rule by autonomous application which functions in an environment of a private user's system has a number of very significant advantages.

2. Statements of means

On the force of legal regulations and acts a large number of citizens have been obliged to submit statements of means [7]. Primarily, representatives of various professions and professional groups fell under this obligation and started to be treated as public service officials. Relevant legislative acts oblige debt collectors, employees, police officers and customs officials, councilors, members of parliament and even the European Mps to submit statements of means. For designated entities it is not a particular problem to collect and store these statements. But legally appointed representatives of these units are also obliged to analyze the content of those statements paying special attention to the occurrence of data that could signify a possibility of the existence of activities of a corruptive nature. This process of content analysis, annually, at least hundreds of thousands of statements of means, unfortunately, is not proceeding smoothly. Therefore, the aim of researches of various teams of people was to make this process more efficient and to invent adequate solutions that would enable achieving this aim in practice [5]. In the era of ever increasing importance and capabilities of IT technologies, suggested solutions tend to substantially automate the process of analyzing statements using computer systems. And here, two basic problems arise. From the point of view of the technical realization of the idea, the way of entering data to such systems becomes a problem. Currently, statements of means are only available in a written form and in this same form they are submitted. There is no doubt that in the era of the development of Information Society and a world-wide access to computers, the process of entering data by society members should be made with the use of e-forms which would allow to create electronic forms that could be completed and send to authorities responsible for their further collection and analysis. This, to some extent, solves the problem of storing data in electronic form for the purpose of their analysis by authorized institutions. Here, however, another problem arises, not of the technical, but legal nature [2]. Current technical and legal solutions neither allow storing nor collecting statement of means in an electronic form. This problem, however, the authors of this article leave to high-level decision-making executives who have to adjust Polish law to the requirements of the XXI century in terms of public administration processes' service [4]. This can be done with a clear conscience, keeping in mind declarations and implementations of projects connected with creation of an information society in Poland specified in a document entitled „The Strategy for the Development of the Information Society in Poland until 2013" [6].

The authors believe that some solutions in the field of IT can help effectively realize many activities associated with the process of submitting statements of means. The scope of undertaken work should aim at, on the one hand, preparing society to future changes, and on the other, supporting people already obliged to submit statements of means to fulfill this not easy duty in a more simple way [8].

3. Main functional assumptions of this solution

The proposed solution should, above all, support individual obligated people in the process of creating their statements of means. Secondly, new solutions should improve, broadly understood, process of delivering this statements to the institution in accordance with this obligation. In current legal conditions supporting the process of submitting statements is limited only to the creation of their paper form. But the solution should also predict, future, after the change of legal status, submission of these documents in an electronic form and with the use of network.

The way of the realization of function of supporting the obliged individuals should also take into account the possibility of implementing similar functions by future institutional solutions, that could be implemented after the change of legal conditions. Generally, the process of creating statement of means in an electronic form should include the need to work with it in a similar way as to work with such a statement in paper form, which, for various reasons, will certainly continue to be used.

It should be noted that the act of completing the statements of means is an activity repeated at least once a year. This fact, as in other similar situations, forces us to write down again, to a new form, the same things we have written in previous years if our situation did not change of course. The number of such data may be large, especially because of the fact that the basic data we are obliged always to fill in are our personal data, which stay the same, for many years. Some may write "not applicable" here.

Moreover, you must also realize that the layout of a statement for one group of obliged people may substantially differ from a statement for another. Therefore, the process of supporting the creation of e-forms should support also citizens who will be obliged to complete different forms in subsequent years. Such assumptions have a clear relationship with a possibility, even certainty, of the change of form and content of forms which would result from the creation of their new versions introduced by the legislature.

As always in this type of solutions it is important to assume the possibility of the development of the system in the direction of, for example, the possibility of other documents service, other than the statements of means.

The authors, taking into account all the above requirements have proposed some solutions which are currently under discussion in public administration institutions by people interested in solving the problems connected with statements of means.

4. General conception of this solution

The proposed solution will primarily support the individual obliged person in the process of creating his statements of means. Secondly, it will be helpful in providing this statement to the institution in accordance with the obligation to do so. In current legal determinants, this support will be limited only to a creation of a paper version of the statement. But it is not out of the question that this solution will enable to submit statements in e-form, after the change of law in future.

The basic element of the proposed solution is to support the process of creating the statement of means by an electronic form of the statement called *document specimen*. Such specimen is created for each binding statement of means form, the filling of which is to be assisted by the system. The method of construction of the specimen provides the possibility to describe each form of the statement of means with its characteristic data structures. Specimens in the form of XML files, available to applications dedicated to them, become the basis for realizing the process of creating electronic documents of the statements of means. It is assumed that electronic documents of the statements of means will also have the form of XML files, having the features of so called partially structured document. It is necessary because of the possibility, in future, to store these documents in databases and subjecting their substantive content (the data contained therein) to legally required processes of analyses.

So understood document specimens of the statements of means become the basis not only for the processes of supporting their creation but also the processes of identification of their content. Role of document specimens of statements of means understood in such a way is presented in figure 1.

In the presented conception the task of document specimen is not limited only to define the scope and nature of the data in a future document (selecting and naming attributes and determine their types). Document specimen also defines:

- future document header information – rules for creating an individual document identifier, specimen identifier and other data describing the document which do not fall within the scope of from data,
- document formal correctness rules – all kinds of validators relating to both individual attributes and attribute groups, as well as to the entire document,

- form of the document in its various forms – rules for the presentation and formatting of data in an XML document, information which allow an unambiguous presentation of document in a visual form on a computer screen or generating "paper" version of the document or its electronic image.

These features of the document specimen allow dedicated software systems for an a single way of creating and interpreting electronic documents defined by these patterns.

The analysis of the process of creating documents of statement of means revealed the necessity of existence of additional procedures of supporting individuals obliged to this action. Therefore, additional elements of document specimen include:

- rules of generating tooltips and supportive explanations during completing the form with data,
- rules of data transformation between documents of different form, that is created based of different specimens.

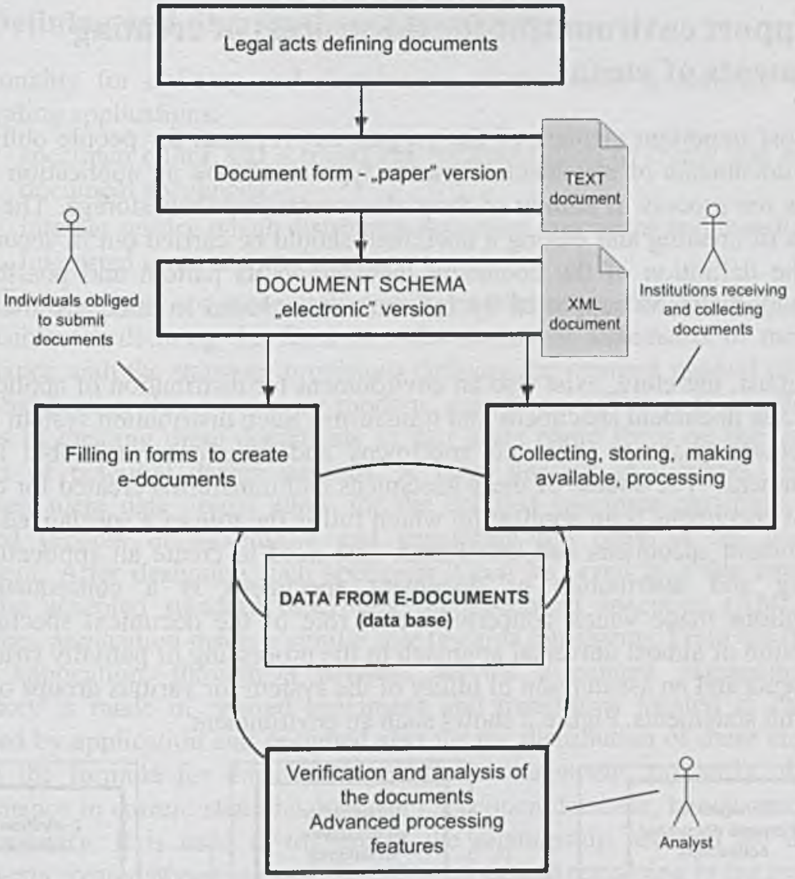

Legal acts defining documents
Document form - „paper" version
TEXT document
DOCUMENT SCHEMA „electronic" version
XML document
Individuals obliged to submit documents
Institutions receiving and collecting documents
Filling in forms to create e-documents
Collecting, storing, making available, processing
DATA FROM E-DOCUMENTS (data base)
Verification and analysis of the documents Advanced processing features
Analyst

Fig. 1. Schema of the basic role of document specimen

Realization of defining rules of data relevance in documents created according to different specimens was based on predefined set of tags attributed to individual data or groups of data. Relevance of particular data or their groups does not cover the required functionality within the scope of required and theoretically possible automation of processes of data transferring between documents of various types. Therefore, in proposed conception there was introduced an additional mechanism of moving data from one document to another (created according to different specimen) based on special files defined for pairs of specimens. These files were called *transform files* or in short *transforms* which can contain complex rules for the conversion of the data included in two documents that are created by different specimens.

5. Support environment for the process of creating statements of means

The most important element of the support environment for people obliged to create documents of statement of means type should be an application which realizes the process of edition of these documents and their storage. The entire process of creating and editing a document should be carried out in accordance with the definition of the document included in its pattern and possibly, as required, should be supported by information included in adequate transform files.

There must, therefore, exist also an environment for distribution of applications for chosen document specimens and transforms. Such distribution system is also an adequate place to store all specimens and transforms available in this environment. The source of these specimens and transforms created for chosen pairs of specimens is an application which fulfils the role of a specialized editor of document specimens and transforms. The need to create an application for defining and distribution of document specimens is a consequence of assumptions made which concern: central role of the document specimen in conception of almost universal approach to the processing of partially structured documents and an assumption of utility of the system for various groups obliged to submit statements. Figure 2 shows such an environment.

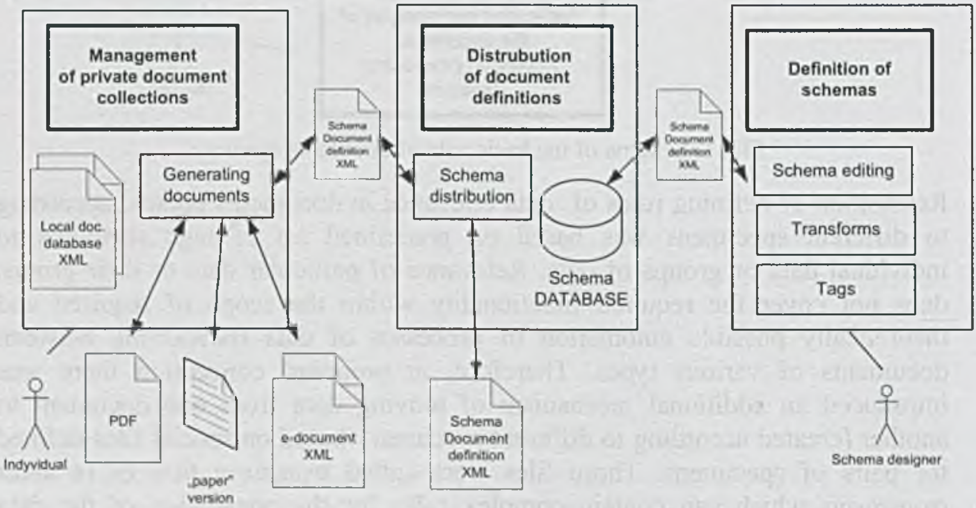


Fig. 2. Support environment for the process of creating documents

5.1. Defining and distribution of specimens

Functionality for defining and distribution of specimens is realized by two cooperating applications:

- specimen editor and a transform for application to create and manage document specimens,
- internet service which distributes document specimens and transforms to interested users and applications.

Application of the specimen editor and transforms is dedicated to people responsible for defining the form of e-documents of statements of means in accordance with the statutory provisions defining the required method of filling them. Supporting this process by modern, convenient editor is to simplify the process of creating these documents so that users could focus on the use of a number of potential design patterns. A large number of different types of specimen meta data must allow for the detailed specimen matching to the required process of its editing and generating top copy of the proposed document. After designing such specimen it can be saved as a file compliant with the accepted standard describing the document specimen (XML file). Described application meets a similar role towards transforms. From the level of editor application, through a network service a content management of repository is made of created specimens and transforms located in database operated by application and designed also for the distribution of these elements [1]. In the formula for document specimen management, primarily, there is maintenance in current state the document specimen database, transforms, and a tag dictionary. It is used to implement the relationship between the data in documents created according to different patterns, and remaining in the semantic compatibility.

Currently, the users of the document specimens service and transforms service include primarily individual users using the OMpriv application, but in future also the institutional systems which exploit databases systems of statements of means.

5.2. Application for a single OMpriv user

The main function of OMpriv application is to give the user obliged to submit a statement of means the possibility to, generally speaking, manage his own documents of these statements. In this formula of management there is, primarily, creation and completing the document, its storing and generating its top copy.

The first task is to choose the nature of this application. After considering many pros and cons, including problems of: comfort, safety, current legal status and habits of future users, it was assumed that the implementation of the applications will be dome in the form of a desktop system with access to network services provided by the document specimens service and transforms service.

To make the main tasks of the application be implemented effectively, the application was equipped with many additional functions. All of these functions are shown in Figure 3.

Part of the functionality is connected with the willingness to allow to work with this application not only individual users but also to groups of users. It's quite logical to extend the scope of tasks of the system, but results in the need to solve two additional issues. The first is the need for the existence of a single user with a privileged status, who would be responsible for the application content management within the scope of document specimens and transforms. He would be predisposed for carrying out communication function with the document specimens service and transforms service. He would also be responsible for the realization of creating backups for the entire application together with documents stored in it for all users. The second issue is to ensure the confidentiality of documents of individual users by restricting access to these documents to all other users. This implies the existence in the system user management function, including the existence of procedures for authorized access to applications and consequently only to our own documents.

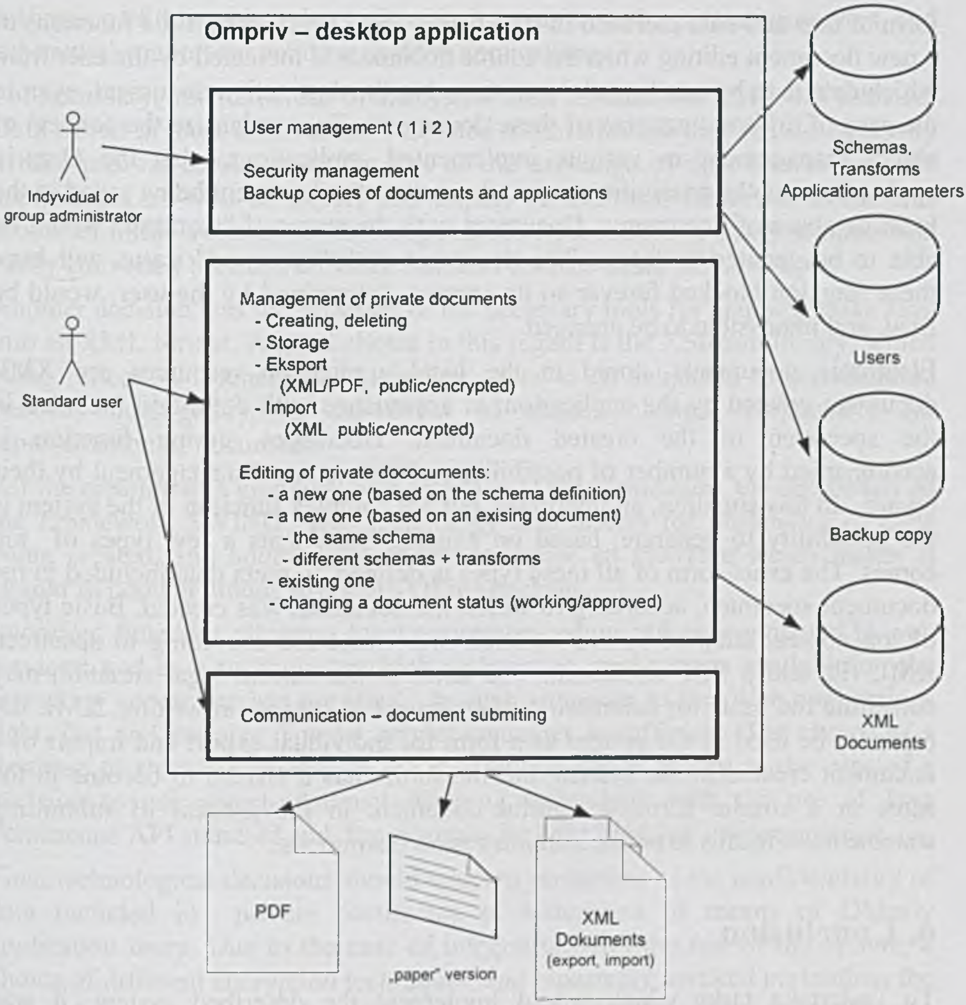

Ompriv – desktop application
User management (1 i 2)
Security management
Backup copies of documents and applications
Indyvidual or group administrator
Management of private documents
- Creating, deleting
- Storage
- Eksport
(XML/PDF, public/encrypted)
- Import
(XML public/encrypted)
Editing of private dococuments:
- a new one (based on the schema definition)
- a new one (based on an exising document)
- the same schema
- different schemas + transforms
- existing one
- changing a document status (working/approved)
Standard user
Communication – document submiting
Schemas, Transforms Application parameters
Users
Backup copy
XML Documents
PDF
„paper" version
XML Dokuments (export, import)

Fig. 3. Functional schema of an application for private document management- OMpriv

In desktop systems, however, this is not a sufficient scope of protecting confidentiality of documents stored as files. Therefore, a symmetric encryption method was introduced for these files based on a generated and stored in the system encryption key for each individual user. It also concerns the resources of created system backups. The main functionality of this application is the edition of own documents created by the users using available in the application document specimens and transforms. Procedures for the realization of all the editing functions fully use meta data of standard documents. So they not only control structuring of data in created documents, but also validation processes of values entered into documents and provide content for contextual substantive help during filling documents with data. Not to overestimate is the role which in the process of editing the documents have to fulfil meta data of schema in the form of tags and data included in transforms' files. Both support the functions of a new document editing when the source document is indicated by the user from which data is to be intelligently' transferred to the destination document, even in the case of different patterns of these documents. To standardize the concept of object management in various implemented applications, also the Ompriv application has the possibility to give the status of a document being saved in the local database of documents. Document with the status of "working" would be able to be updated in future. The document with "approved" status will have these function blocked forever so its content, determined by the user, would be final, and impossible to be changed.

Electronic documents stored in the local application resources are XML document created by the application, in accordance with definition included in the specimen of the created document. Document storing function is accompanied by a number of possibilities of their content management by their owner. In this situation, an important, but not complex function of the system is the possibility to generate, based on existing documents a few types of top copies. The exact form of all these types is defined by meta data included in the document specimen, according to which the document was created. Basic types of top copies, adopted in this solution are: structured according to specimen XML file and a PDF document. The latter in the current legal situation may constitute the basis for submission of statement of means in writing. XML file can now be used in the system as a form for individual export and import of a document created in the system, but the future has a chance to become in the same or a similar form, a rightful document in the process of submitting statements of means to public administration institutions.

6. Conclusion

To undertake tasks which would implement the described system, it was essential to take a series of choices which, despite wide diversity of implemented elements of the system will allow for their easy integration in the future for their easy development [1. 3].

The most important choice for the process of implementation of all applications was the choice of Java language for the implementation of key elements of the proposed solution. This choice was supported by a fact of large object code portability, and a good support for necessary in the system of this architecture web services and web applications. The most important choice for the process of implementation of all applications was the choice of Java language for the implementation of key elements of the proposed solution. This choice was supported by a fact of large object code portability, and a good support for necessary in the system approval of a program environment suitable for the realization of this project due to its easy integration with application servers, and substantial support for creating desktop applications.

For necessary, and numerous in the system data serialization XML was selected. Data stored in local archives are subjected to serialization the and the necessity of this process results from the fact of the exchange of data between system applications as well as export and import of the most important documents produced in the system. XML independent of the platform, easily scalable and easily converted between different standard formats using XSLT scripts.

Another decision was the selection of the necessary tools for mapping tasks Java into an XML format. A good choice in this regard is the XStream library, which using reflection mechanism for finding data to be subjected to serialization, behaves intelligently, in the event of classes changes responsible for representation of documents.

For the creation of a graphical user interface of the application, the choice fell on the Framework SWING. This technology also allows for portability of code being created, and additionally allows to create an interface which makes it similar to popular among users MS Office package.

Important functions allowing for a remote procedure call was entrusted to web services, and as a technology which makes web applications available on the repository server servlets are used. In such situation, to run Web applications light, fast and portable Tomcat servlet container is sufficient. The choice of a database to store documents on the distribution server is any in the case of a decision to use object-relational mapping technology with the use of Java Persistence API standard and, for example, its light TopLink implementation.

Final technological decisions should concern protection of the confidentiality of data included in private documents of statements of means of OMpriv application users. Due to the ease of integration with the rest of the system, a choice of different encryption techniques, and repeatedly invoked portability, the choice fell on framework Java Cryptography Architecture. It not only allows to choose 256-bit encryption algorithm (TwoFish here), but also it provides a structure to allow the secure storage of encryption keys on a disc of a local station.

Above, short characteristics of the selection of techniques and tools for implementing the system shows very real possibility of achieving its successful future implementation.

In the current phase of work, you can just talk about the possibility of achieving success. But it is worth to present such functional and realization conception of the environment for opportunities for a broader discussion on problems raised here.

REFERENCES

1. Abiteboul, S., Buneman, P., Suciu D.: Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann Publishers, 2001.
2. Fajgielski P.: Information in public administration – legal aspects of collecting, sharing and protecting (in Polish). PRESSCOM, Wroclaw, 2007.
3. Guzowski M., Sedzik K.: Systems for the support of acquisition and data validation in a heterogeneous environment (in Polish). Thesis under the supervision of M. Valenta, Department of Computer Science AGH, Cracow 2009.
4. Marcjan R., Valenta M.: Necessary legal changes in the aspect of computer systems implementation within the scope of submitting statements of means in services subordinated to Ministry of Internal Affairs and Administration (in Polish), Ministry of Internal Affairs and Administration, Warsaw, 2009.
5. Nowak M., Okrzes M.: Knowledge management in decision support systems (in Polish). Thesis under the supervision of M. Valenta, Department of Computer Science AGH, The Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, Cracow 2008.
6. Strategy for the development of information society in Poland until 201 (in Polish)3. Ministry of Internal Affairs and Administration – http://www.mswia.gov.pl/strategia, Warsaw, 2008
7. The Act of 21 August 1997 on Restrictions for State Officials on Business Activities (in Polish). Journal of Laws., 1997, No. 106 item 679.
8. Valenta M., Marcjan R.: Processing of partially structured documents. Artificial intelligence methods in activities for promoting public safety (in Polish), editor: Nawarecki E., Dobrowolski G., Kisiel-Dorohinicki M., AGH Publishing House, Cracow, 2009.

Chapter 3

Computer simulation of block-parallel algorithms for image reconstruction

Nadiya Gubareni, Mariusz Pleszczynski
Politechnika Śląska, Technical University of Częstochowa

Abstract

This chapter is about image reconstruction supported by block-parallel simulation.

1. Introduction

The problem of investigation of the internal structure of objects without destroying them arises not only in medicine but also in many scientific and technical problems. The technique of computerized tomography allows to reconstruct the internal structure of an object from projection data collected outside the object.

Let *f*(*x*,*y*) be a function which represents the spatial distribution of a physical parameter. If *L* is a line (ray) in the plane then the line integral

$$p_L = \int_L f(x,y)dL\,, \tag{1}$$

which is called a projection, is usually obtained from physical measurements.

From mathematical point of view the problem of reconstruction from projections is to find an unknown function *f*(*x*,*y*) by means of a given set of projections p_L for all *L*. Theoretically it is possible to reconstruct the function *f*(*x*,*y*) from the set p_L by means of the Radon inversion formula. However, in practice we are given only discrete set of projection data that estimate *p* for a finite number of rays. Moreover, since the projection data are obtained by physical measurements, they are given with some errors.

In many practical applications the projection data are often not available at each direction and may be very limited in number. In this case we say that we have a problem of image reconstruction with incomplete projection data. In particular, such kind of problems arises in mineral industries and engineering geophysics connected with acid drainage, the stability of mine workers, mineral exploration and others [1], [2].

In this case the use of analytical methods does not give enough good results. One of the way for solving problem of image reconstruction with incomplete data is reducing to solve the system of linear algebraic equations:

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{p} \ , \tag{2}$$

where:

$\mathbf{A} = (a_{ij}) \in \mathbf{R}^{m,n}$ is the matrix of coefficients,

$\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in \mathbf{R}^n$ is the image vector,

$\mathbf{p} = (p_1, p_2, \dots, p_m)^T \in \mathbf{R}^m$ is the measurement vector of projection data.

This system has a few characteristics: it is a rectangular as a rule and it has a very large dimension. For solving this system it is often used different kind of algebraic iterative algorithms which are based on the Kaczmarz method, the most well-known of which are the additive algorithm ART (see [3]-[7]). These algorithms are very flexible and allow to apply different *a priory* information about object before its reconstruction that is especially very important when we have an incomplete projection data. The basic idea of these algorithms is to run through all equations cyclically with modification of the present estimate $\mathbf{x}^{(k)}$ in such a way that the present equation with index *i* is fulfilled.

In practice the vector of projection data is given as a rule with some error.

Therefore instead of a system of linear equations (2) we have a system of linear inequalities:

$$\mathbf{p} - \mathbf{e} \leq \mathbf{A} \cdot \mathbf{x} \leq \mathbf{p} + \mathbf{e} \tag{3}$$

where $\mathrm{e} = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m\}$ is a non-negative vector. And we can consider that the vector e is given a priory and defines the errors of projection data.

In this paper we consider some parallel implementation of iterative algebraic algorithms for image reconstruction from incomplete projection data for some particular reconstruction schemes which arise in some problems of engineering geophysics and mineral industry. In such computing structure each elementary processor executes independently its calculations by means of the same simple algorithms connected with set of corresponding equations.

We assume that each processor executes its calculations with its own pace and we allow the communication channels to deliver messages out of order. In this case we have the chaotic character of interactions in such CPS which corresponds to some chaotic iterative algorithm. This algorithm realized on such CPS is based on the asynchronous methods [8], [9], [10].

In order to reduce the computation time and memory space of computer there were proposed another algebraic algorithms which allow their parallelization and may be realized on the fast massively parallel computing systems (MPCS) consisted of elementary processors and a central processor [11], [12], [13].

In this paper we consider some kinds of the block-parallel asynchronous algorithms for image reconstruction. These algorithms are some generalization of parallel chaotic iteration methods considered by Bru, Elsner and Neumann [14].

Numerical simulation of solving problems for image reconstruction from incomplete projection data for some modeling objects, comparing evaluations of errors and rate of convergence of these algorithms are presented. It is shown then for some choice of parameters we can obtain a good enough quality of reconstruction with these algorithms, and that these algorithms have much higher rate of convergence in comparison with corresponding synchronous algorithms.

2. Block-parallel iterative algorithms for image reconstruction

In this paper we use some of the parallel and block-iterative algorithms for solving system of linear equations (2) and system of linear inequalities (3), some of which were considered in papers [15], [4], [16].

Denote by

$$\mathbf{P}_i(\mathbf{x}) = \mathbf{x} - \frac{((\mathbf{a}^i,\mathbf{x}) - p_i - \varepsilon_i)^+ - (p_i - \varepsilon_i - (\mathbf{a}^i,\mathbf{x}))^+}{\left\|\mathbf{a}^i\right\|^2}\mathbf{a}^i, \tag{4}$$

where

$$s^+ = \begin{cases} s, & \text{if } s \geq 0; \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbf{P}_i^{\omega} = (1-\omega)\mathbf{I} + \omega\mathbf{P}_i, \tag{5}$$

where $\mathbf{a}^i$ is i-th row of a matrix $\mathbf{A}$, $0<\omega<2$ is a relaxation parameter.

Algorithm 1 (PART) [11]

1. $\mathbf{x}^{(0)} \in \mathbf{R}^n$ is an arbitrary vector;
2. The $k+1$-th iteration is calculated in accordance with such a scheme:

$$y^{k,i} = \mathbf{P}_i^{\omega_k}\mathbf{x}^{(k)} \qquad (i = 1,2,\ldots,m), \tag{6}$$

$$\mathbf{x}^{(k+1)} = \mathbf{C}\sum_{i=1}^{m}\mathbf{B}_i^k\mathbf{y}^{k,i}, \tag{7}$$

where $\mathbf{P}_i^{\omega_k}$ are operators defined by (4) and (5), $0 < \omega_k < 2$ are relaxation parameters, $\mathbf{C}$ is a constraining operator and $\mathbf{B}_i^k$ are matrices of dimension $n \times n$ with real nonnegative elements and

$$\sum_{i=1}^{m} \mathbf{B}_i^k = \mathbf{E}, \qquad \sum_{i=1}^{m} \| \mathbf{B}_i^k \| \le 1, \tag{8}$$

for all $k \in N$, where $\mathbf{E}$ is the unit matrix of dimension $n \times n$.

The parallel implementation of this algorithm may be organized as follows:

```
begin
  $\mathbf{x}^{(0)}$=initial
  for $k$=0,1,... until convergence
  do
      for $i$-th processor, $i$=1 to $m$
      do
              $y^i = \mathbf{P}_i^{\omega_k} \mathbf{x}^{(k)}$
      enddo
              $\mathbf{x}^{(k+1)} = \mathbf{C}\sum_{i=1}^{m} \mathbf{B}_i^k \mathbf{y}^i$
  enddo
end
```

Let $\mathbf{B}_i^k = (\gamma_{jj}^i)_{j=1}^n$ be a diagonal matrix with elements $0 < \gamma_{jj}^i < 1$. If $\gamma_{jj}^i = \gamma_i$ for each $j \in J$, $i \in I$, $\mathbf{C} = \mathbf{I}$, then we obtain the Cimmino algorithm [17].
The sufficient conditions of convergence of algorithm 1 are given by the following theorem:

Theorem 1 [11]. *If system* (3) *is consistent then the sequence* $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ *defined by algorithm* 1 *converges to some solution of system* (3).

These algorithms may be realized on parallel computing structure consisted of m elementary processors and one central processor. On each (k+1)-th step of iteration every i-th elementary processor computes the coordinates of vector $\mathbf{y}^{k,i}$ in accordance with formula (6) or (10) and then the central processor computes the (k+1)-th iteration of the image vector $\mathbf{x}$ in accordance with formula (7) or (9).
The main defect of parallel algorithms considered above are their practical realization on parallel computational structures because it needs a lot of local processors in such MPCS. In order to reduce the number of required local processors we proposed block-iterative additive and multiplicative algorithms.

For this purpose we decompose the matrix **A** and the projection vector **p** into M subsets in accordance with decomposition

$$\{1,2,\ldots,m\} = H_1 \cup H_2 \cup \ldots \cup H_M, \qquad (9)$$

where

$$H_t = \{m_{t-1}+1,\, m_{t-1}+2, \ldots, m_t\}, \qquad (10)$$

$0 = m_0 < m_1 < \ldots < m_M = m.$

Then in the general case we have the following block-parallel algorithm:

Algorithm 2 (BPART) [4]

1. $\mathbf{x}^{(0)}$ is an arbitrary vector;
2. The k+1-th iteration is calculated in accordance with such a scheme:

$$\mathbf{x}^{(k+1)} = \mathbf{C} \sum_{i \in H_{t(k)}} \mathbf{B}_i^k \mathbf{P}_i^{\omega_k} \mathbf{x}^k, \qquad (11)$$

where $t(k) = k(\mathrm{mod}\ M) + 1$, $\mathbf{P}_i^{\omega_k}$ are operators defined by (4) and (5), $0 < \omega_k < 2$ are relaxation parameters, **C** is a constraining operator and $\mathbf{B}_i^k$ are matrices of dimension $n \times n$ with real nonnegative elements.

The sufficient conditions of convergence of algorithm 4 are given by the following theorem:

Theorem 2 [4]. *If system* (3) *is consistent and* $\mathbf{B}_i^k$ *are matrices of dimension* $n \times n$ *with real nonnegative elements:*

$$\sum_{s \in H_{t(k)}} \mathbf{B}_i^k = \mathbf{E}, \quad \sum_{s \in H_{t(k)}} \left\| \mathbf{B}_i^k \right\| \le 1 \qquad (12)$$

for all $k \in N$, *where* **E** *is identity matrix of dimension* $n \times n$ *then the sequence* $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ *defined by algorithm* 4 *converges to some solution of system* (3).

The parallel implementation of this algorithm can be described as follows:

$$y^{k,i} = \mathbf{P}_i^{\omega_k} \mathbf{x}^{(k)} \qquad (i \in H_{t(k)}),$$

$$\mathbf{x}^{(k+1)} = \mathbf{C} \sum_{i \in H_{t(k)}} \mathbf{B}_i^k \mathbf{y}^{k,i},$$

or may be given by the following form:

```
begin
    $\mathbf{x}^{(0)}$=initial
```

```
 for $k$=0,1,... until convergence
 do
     $t(k) = k(\mathrm{mod}\ M) + 1$
     do
     for $i$-th processor, $i \in H_t$
     do
             $y^{k,i} = \mathbf{P}_i^{\omega_k} \mathbf{x}^{(k)}$
     enddo
             $\mathbf{x}^{(k+1)} := \mathbf{C} \sum_{i \in H_t} \mathbf{B}_i^k \mathbf{y}^i$
 enddo
end
```

Consider a particular case of algorithm 2. Let $\mathbf{W} = (w_{ij})$, $\mathbf{U} = (u_{ij})$ be matrices of dimension $M \times n$ with elements

$$w_{ij} = \sum_{s \in H_i} a_{sj}, \quad u_{ij} = \frac{w_{ij}}{\sum\limits_{k=1}^{M} w_{kj}}$$

where $i = 1,2,\ldots, M$; $j = 1,2,\ldots, n$.

Consider matrices $\mathbf{B}_i = (b^{i}_{sk})$ of dimension $n \times n$ with elements

$$b_{sk}^{i} = \begin{cases} u_{ik}, & \text{for} \quad s = k \\ 0, & \text{otherwise} \end{cases}$$

where $i = 1,2, \ldots, M$; $s, k = 1,2,\ldots, n$.

Algorithm 3 (SZB-3).

1. $\mathbf{x}^{(0)} \in \mathbf{R}^n$ is an arbitrary vector;

2. The k+1-th iteration is calculated in accordance with the following scheme:

$$\mathbf{x}^{(k+1)} = \mathbf{C} \sum_{i \in H_{t(k)}} \mathbf{B}_i^k \mathbf{P}_i^{\varpi_k} \mathbf{x}^k, \tag{13}$$

where $t(k) = k(\mathrm{mod}\ M) + 1$, $\mathbf{P}_i^{\omega_k}$ are operators defined by (4) and (5), $0 < \omega_k < 2$ are relaxation parameters, $\mathbf{C}$ is a constraining operator and $\mathbf{B}_i^k$ are matrices of dimension $n \times n$ with real nonnegative elements and

$$\mathbf{B}_i^k = diag\{b_1^{k,i}, b_2^{k,i}, \ldots, b_n^{k,i}\}, \tag{14}$$

where

$$b_p^{k,i} = \frac{y_p^{k,i}}{\sum_{i \in H_{t(k)}} y_p^{k,i}} \tag{15}$$

for $p = 1,2,...,n.$

The parallel implementation of this algorithm can be described as follows:

$$y^{k,i} = \mathbf{P}_i^{\omega_k} \mathbf{x}^{(k)} \quad (i \in H_{t(k)}),$$

$$\mathbf{x}^{(k+1)} = \mathbf{C} \sum_{i \in H_{t(k)}} y_p^{k,i}.$$

We also consider the following block-parallel algorithm.

Algorithm 4 (RB-3).

1. $\mathbf{x}^{(0)} \in \mathbf{R}^n$ is an arbitrary vector;
2. The k+1-th iteration is calculated in accordance with such a scheme:

$$\mathbf{x}^{(k+1)} = \sum_{i=1}^{M} \mathbf{B}_i \mathbf{y}^{k+1,i}, \tag{17}$$

where

$$\mathbf{y}^{k+1,i} = \mathbf{Q}_i \mathbf{x}^{(k)},$$

$\mathbf{Q}_i = \mathbf{P}^{\varpi}_{m_i} \mathbf{P}^{\varpi}_{m_i-1} ... \mathbf{P}^{\varpi}_{m_{i-1}+1}$, $\mathbf{P}_i^{\omega}$ are operators defined by (4) and (5), $0 < \omega < 2$ are relaxation parameters and $\mathbf{B}_i$ are matrices of dimension $n \times n$ with real nonnegative elements and

$$\mathbf{B}_i = diag\{b_1^i, b_2^i, ..., b_n^i\}, \tag{18}$$

where

$$b_p^i = \frac{\sum_{s \in I_i} a_{s,p}}{\sum_{i=1}^{M} a_{s,p}} \tag{19}$$

for i=1,2,…,M and $p = 1,2,...,n.$

3. Block-parallel asynchronous algorithms for computer tomography

In this section we apply the generalized model of asynchronous iterations, considered in [19], for implementations of block-parallel algorithms on non-synchronous computer structure. For this aim we recall some main notions of the theory of asynchronous iterations which were introduced by Chazan and Miranker (1969) in [10] and Baudet (1978) in [8].

The important notion in the theory of asynchronous iterations is the sequence of chaotic sets.

Definition 1. A sequence of nonempty subsets $I = \{I_k\}_{k=0}^{\infty}$ of the set {1,2..., m} **is a sequence of chaotic sets** if

$$\limsup_{j\to\infty} I_j = \{1,2,...,m\} \tag{20}$$

(another words, if each integer $j \in \{1,2,...,m\}$ appears in this sequence infinite number of times).

Definition 2. If each subset I_k of a sequence of chaotic sets $I = \{I_k\}_{k=0}^{\infty}$ consists of only one element, then such sequence is called **acceptable**.

Definition 3. A sequence $J = \{\sigma(k)\}_{k=1}^{\infty}$ of m-dimensional vectors $\sigma(k) = (\sigma_1(k), \sigma_2(k),...,\sigma_m(k))$ with integer coordinates, satisfying the following conditions:

$$1)\ 0 \le \sigma_i(k) \le k-1;$$

$$2)\ \lim_{k\to\infty} \sigma_i(k) = \infty, \tag{21}$$

for each i = 1, 2..., m and $k \in \mathbf{N}$, is called a **sequence of delays**.

Suppose that PCS (Parallel Computing System) consists of m of processors working locally independent. In this case the notion of the sequence of chaotic sets has a simple interpretation: it sets the time diagram of a work of each processor during non-synchronous work of PCS. So the subset I_k is the set of the numbers of those processors which access the central processor at the same time.

Let $T = \{T_i\}_{i=1}^{m}$ be a set of nonlinear operators acting on the Euclidean space $\mathbf{R}^n$ and S be an algorithmic operator. Consider the following iterative process:

$$y^{k,i} = T_i x^{(k-1)}, \tag{22}$$

$$x^k = S(x^{(k-1)}, \{y^{k,i}\}_{i=1}^{m}), \tag{23}$$

where **x** is an n-dimensional vector of the space $\mathbf{R}^n$, $i \in \{1,2,...,m\}$ for every $k = 0,1,2,...$.

We shall consider the parallel asynchronous implementation of such iterative process on a parallel multiprocessor structure consisting of m independent elementary processors and some central processor. Each i-th elementary processor executes its calculations with its own pace in accordance with formula correspondent to operator $\mathbf{T}_i$. It has its own local memory and connects only with the central processor. We assume that each elementary processor can have access to the central processor at any time. After each cycle of calculations of vector $\mathbf{y}^{k,i}$ the i-th processor sends this value to the central processor and loads from it the new value $\mathbf{x}^k$ as its new initial data. And the sequence of delays determines the numbers of using iterations by each fixed processor, and the number L shows a depth of used iterations and actually reflects possibilities of the concrete computing system. For synchronous implementation of the iterative process the difference

k-$\sigma_i(k)$ is equal to 0 for all $i = 1,2..., m$ and $k \in$N.

Recall the definition of generalized model of asynchronous computational process (see [15]):

Definition 4. Let $\mathbf{T}_i: \mathbf{R}^n \to \mathbf{R}^n$, $i \in \{1,2,...,m\}$ be a set of nonlinear operators and let $\mathbf{x}^0 \in \mathbf{R}^n$ be an initial value of a vector $\mathbf{x}$. A **generalized model of the asynchronous iterations with limited delays** for the set of operators $\mathbf{T}_i$, i=1,2,...,m is a method of building the sequence of vectors $\left\{\mathbf{x}^k\right\}_{k=0}^{\infty}$, which is given recursively by the following scheme:

$$\mathbf{y}^{k,i} = \begin{cases} \mathbf{T}_i \mathbf{x}^{\left(\sigma^i(k)\right)}, & \text{if } i \in I_k \\ \mathbf{y}^{k-1,i}, & \text{otherwise} \end{cases} \tag{24}$$

$$\mathbf{x}^{(k)} = \mathbf{S}\left(\mathbf{x}^{(k-1)}, \left\{\mathbf{y}^{k,i}\right\}_{i \in I_k}\right)$$

whe$\left\{\sigma^i(k)\right\}_k^{\infty} I_k \}_{k=1}^{\infty}$ is a sequence of chaotic sets such that $I_k \subset \{1,2,...,m\}$ and $J_i = \left\{\sigma^i(k)\right\}_{k=1}^{\infty}$ are sequences of limited delays (i=1,2,...,m).

Now we apply the generalized model of asynchronous iterations for implementation of algorithm BPART on non-synchronous computer structure. In this case we obtain the following algorithm, where the numbers of operators are chosen by the chaotic way:

Algorithm 5

1. $\mathbf{x}^{(0)} \in \mathbf{R}^n$ is an arbitrary vector;
2. The k+1-th iteration is calculated in accordance with such a scheme:

$$\mathbf{x}^{k+1,i} = \mathbf{C} \sum_{i \in H_{\iota(k)}} \mathbf{B}_i^k \mathbf{P}_i^{\omega_k} \mathbf{x}^{(\sigma^i(k))}, \tag{25}$$

where $\mathbf{P}_i^{\omega_k}$ are operators defined by (4) and (5), $0 < \omega_k < 2$ are relaxation parameters, $\mathbf{C}$ is a constraining operator, $\iota(k) = I_k$, $I = \left\{I_k\right\}_{k=0}^{\infty}$ is a sequence of chaotic sets such that $I_k \subset \{1,2,..., M\}$ and $\mathbf{B}_i^k$ are matrices of dimension $n \times n$

with real nonnegative elements which satisfy conditions (16), $J_i = \left\{\sigma^i(k)\right\}_{k=1}^{\infty}$ are sequences of delays.
The convergence of this algorithm is given by the following theorem:

Theorem 3. *Let system* (2) *be consistent,* $I = \left\{I_k\right\}_{k=0}^{\infty}$ *be a regular sequence of chaotic sets* $I_k \subset \{1,2,\dots,M\}$ *with a number of regularity T,* $J_i = \left\{\sigma^i(k)\right\}_{k=1}^{\infty}$ *be sequences with limited delays and* $\sigma^l_j(k) = \sigma_i(k)$, *and let the number of delay be equal to T. Then for every point* $\mathbf{x}^{(0)} \in \mathbf{R}^n$ *the sequence* $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ *defined by algorithm 6 converges to some point* $\mathbf{x}^* \in H$, *which is a fixed point of orthogonal projection operators* $\mathbf{P}_i$ *(i* = 1,2,..., *M).*

The full proof of this theorem one can find in [4]. We now consider the particular case of algorithm 5 when we have no delays and the sequence of chaotic sets is acceptable.

We decompose the matrix **A** and the projection vector **p** into *M* subsets in accordance with decomposition (9) and (10). We consider $s_t = \left|H_t\right| = m_t - m_{t-1}$ cardinality H_t.

Algorithm 6 (CHRB-3)

1. $\mathbf{x}^{(0)} \in \mathbf{R}^n$ is an arbitrary vector;
2. The *k*+1-th iteration is calculated in accordance with such a scheme:

$$\mathbf{x}^{k+1} = \mathbf{C}\sum_{i=1}^{M} \mathbf{B}_i^k y^{(k+1),i}, \tag{26}$$

where

$$y^{(k+1),i} = \mathbf{Q}_i \mathbf{x}^k,$$

$$\mathbf{Q}_i = \mathbf{P}_{i,s_i} \mathbf{P}_{i,s_i-1} \dots \mathbf{P}_{i,1},$$

$$\mathbf{P}_{i,j} = \mathbf{P}_j^{\varpi} \sum_{i=1}^{M} \mathbf{B}_i^k, \quad j \in I_{i(j)},$$

$\mathbf{P}_i^{\omega}$ are operators defined by (4) and (5), $0 < \omega < 2$ are relaxation parameters, **C** is a constraining operator, $I = \left\{I_{i(k)}\right\}_{k=1}^{\infty}$ is a sequence of chaotic sets such that $I_{i(k)} \subset$ {m_{i-1}+1, m_{i-1}+2,..., m_i }=H_i and $\mathbf{B}_i^k$ are matrices of dimension $n \times n$ with real nonnegative elements which satisfy conditions (18) and (19) for each $k \in N$.
In this paper we consider that **C**= C_1 C_2 where

$$\left(C_1[\mathbf{x}]\right)_i = \begin{cases} a, & \text{if } x_i < a; \\ x_i, & \text{if } a \le x_i \le b; \\ b, & \text{if } x_i > b; \end{cases} \tag{27}$$

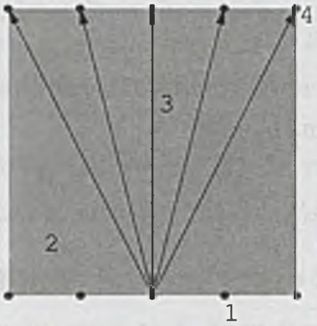
$$\left(C_2[\mathbf{x}]\right)_j = \begin{cases} 0, & \text{if } p_i = 0 \text{ and } a_{ij} \ne 0; \\ x_j, \text{otherwise} \end{cases} \tag{28}$$

4. Computer simulation and experimental results

In dependence on the obtaining system of projections there are many image reconstruction schemes, the main of them are parallel and beam schemes in the two-dimensional space. In some practical problems, in engineering for example, it is impossible to get projections from all directions because of the existing some important reasons (such as situation, size or impossibility of an access to a research object). This situation arises, for example, in the coal bed working. In such a coal bed during the preparing process for working in dependence on the scheme the access to longwalls may be very difficult or impossible at all. Sometimes it is impossible to access to one or two sides of longwalls, and sometimes it is impossible only to access to the basis but all the longwalls are accessible. Each this situation has its own scheme of obtaining information.

In this paper we present result for image reconstructions only for two different schemes, which are described below.

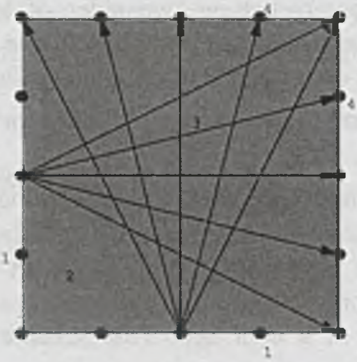
In the first scheme of obtaining projection data, which we shall call as the system (1 × 1), we have an access to a research object from only two opposite sides. This situation often arise in engineering geophysics. In this case the sources of rays are situated only on one side and the detectors are situated on the opposite side of the research part of a coal bed. This scheme of obtaining information is shown in Fig.1.



1 – Sources of rays
2 – Research object
3 – Rays
4 – Detectors

Fig. 1. The system (1×1)

And the second scheme of obtaining projection data, which we shall call (1 × 1, 1 × 1), is shown in Fig.2. In this situation we can have an access to all four sides of an object. Therefore the sources can been situated onto two neighboring sides, and the detectors can been situated on the opposite sides. So the projections can be obtained from two pair of the opposite sides.



1 – Sources of rays
2 –Research object
3 – Rays
4 – Detectors

Fig. 2. The system (1×1, 1×1)

In order to evaluate the goodness of the compute reconstruction of a high-construct image from a limited number of projections and incomplete data we tested different kind of geometric figures and reconstruction schemes.

An important factor in the simulation process of image reconstruction is the choice of modeling objects which describe the density distribution of research objects. In a coal bed, where we search the reservoirs of compressed gas or interlayers of a barren rock, the density distribution may be considered discrete and the density difference of these three environments (coal, compressed gas and barren rock) is significant. Therefore for illustration of the implementation of the algorithms working with incomplete data we chose the discrete function with high contrast which is given by the following form:

$$f(x,y)=\begin{cases}1, & (x,y)\in D_1\subset E\subset \mathbf{R}^2,\\ 2, & (x,y)\in D_2\subset E\subset \mathbf{R}^2,\\ 3, & (x,y)\in D_3\subset E\subset \mathbf{R}^2,\\ 4, & (x,y)\in D_4\subset E\subset \mathbf{R}^2,\\ 0, & \text{otherwise}\end{cases} \tag{29}$$

where E is a square $E=\{(x,y):-1\le x,y\le 1\}$, and D_i are subsets of E of the following form:

$$D_1 = [-0.7,-0.4] \times [-0.5,0.2],\ D_2= [-0.2,0.2] \times [-0.1,0.1],$$

$$D_3 = [-0.2,0.2] \times [0.3,0.5],\ D_4 = [0.4,0.7] \times [0.4,0.7].$$

The plot of this function is given in Fig.3.

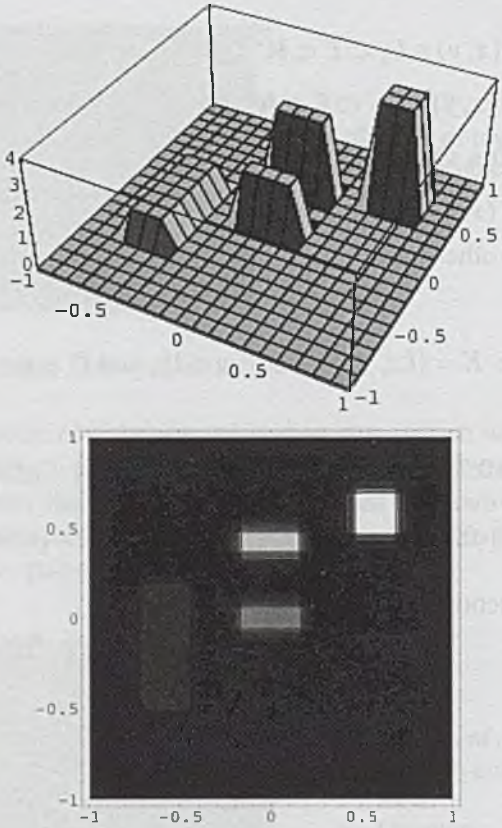


Fig. 3. The original function *f(x,y)*

As was shown earlier (see, for example [3], [4], [5]), the image reconstruction of such objects from complete data gives a good enough results after 6-7 full iterations.

In this paper we present the numerical results of image reconstruction of this function with algebraic iterative algorithms ART-3, block-parallel algorithm SZB-3, RB-3 and chaotic block-parallel algorithm CHRB-3. We compare the results of reconstructions, and we investigate the influence of various parameters of these algorithms such as a pixel initialization, relaxation parameters, number of iterations and noise in the projection data on reconstruction quality. The convergence of these algorithms was studied in dependence on different these parameters. The convergence characteristic plots are given in view of plots for the mean absolute error

$$\delta_1 = \frac{1}{n}\sum_i \left| f_i - \tilde{f}_i \right|$$

and the maximal relative error

$$\delta_2 = \frac{\max\limits_i \left| f_i - \widetilde{f}_i \right|}{\max\limits_i \left| f_i \right|} \cdot 100\% ,$$

where f_i is the value of a given modeling function in the center of the i-th pixel and $\widetilde{f}_i$ is the value of the reconstructed function in the i-th pixel.

In the results of computer simulation we assume, that
n - is the number of pixels, i.e. the number of variables,
m - is the number of rays, i.e. the number of equations,
M - is the number of blocks,
iter - is the number of full iterations.

In our simulation we also assume that *M* is equal to the number of detectors.

The results of image reconstructions for function *f(x,y)* with block-parallel algorithm SZB-3 and RB-3 in the system (1 × 1,1 × 1) for the same parameters are given in Fig.4 - Fig.5.

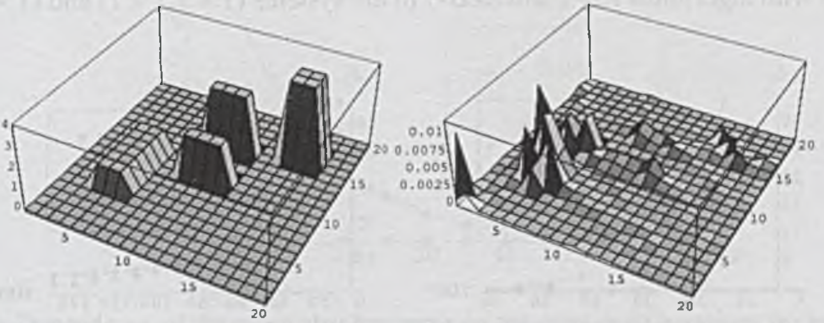


Fig. 4. The image reconstruction and the mean absolute error for *f(x,y)* obtained with algorithm
SZB-3 for *n* =20 × 20, *m*=644, *M*=36, *iter*=25 in the system (1 × 1, 1 × 1)

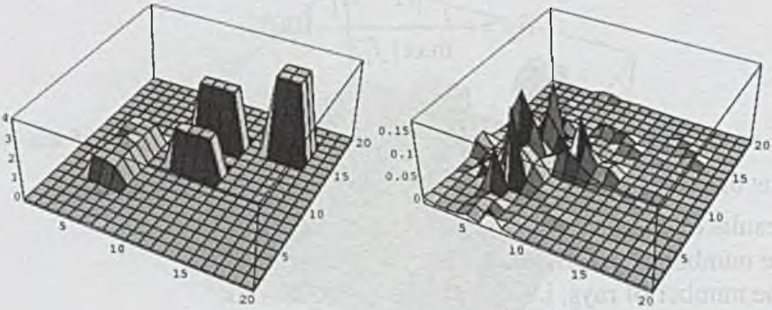


Fig. 5. The image reconstruction and the mean absolute error for *f(x,y)* obtained with algorithm

RB-3 for *n* =20 × 20, *m*=644, *M*=36, *iter*=25 in the system (1 × 1, 1 × 1)

The plots, which are presented in Fig.6, illustrate the dependence of the mean absolute error on the number of iterations of image reconstruction of function *f(x,y)* with algorithms RB-3 and SZB-3 in the systems (1 × 1,1 × 1) and (1 × 1):

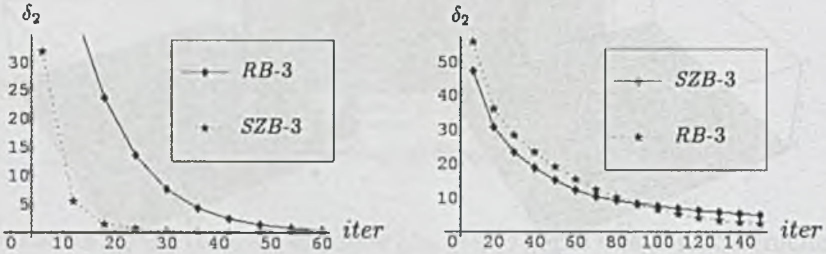


Fig. 6. Dependence of the mean absolute error on the number of iterations for image reconstruction of *f(x,y)* with algorithm RB-3 and SZB-3 in the system (1 × 1, 1 × 1) for *n*= 20×20, *m*=644, *M*=36 (on the left side) and in the system (1 × 1) for *n* = 20 × 20, *m*=788, *M*=28 (on the right side)

The results of image reconstructions for function *f(x,y)* with chaotic block-parallel algorithm CHRB-3 in the system (1 × 1,1 × 1) for the same parameters are given in Fig.7.

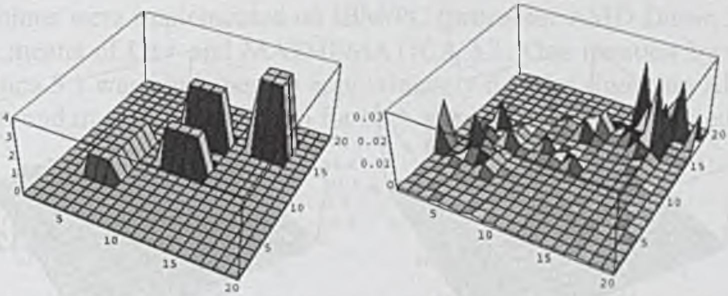


Fig. 7. The image reconstruction and the mean absolute error for *f(x,y)* obtained with algorithm
CHRB-3 for n=20 × 20, m=644, M=36, *iter*=25 in the system (1 × 1, 1 × 1)

The plots presented in Fig.8 illustrate the dependence of the mean absolute error on the number of iterations of image reconstruction of *f(x,y)* with algorithms RB-3 and CHRB-3 in the system (1 × 1, 1 × 1):

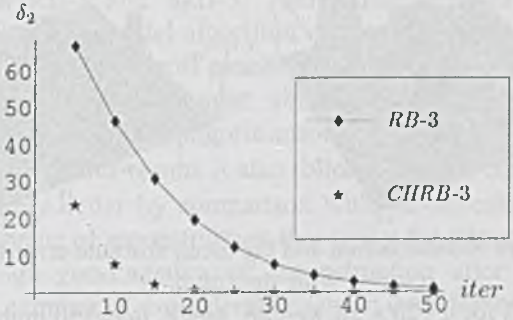


Fig. 8. Dependence of the mean absolute error on the number of iterations for image reconstruction of *f(x,y)* with algorithm RB-3 and CHRB-3 in the system (1 × 1, 1 × 1)

The results of reconstruction of the function *f(x,y)* with block-parallel algorithm SZ-3 and chaotic block-parallel algorithm CHRB in the system (1 × 1) is shown in Fig.9 and Fig.10.

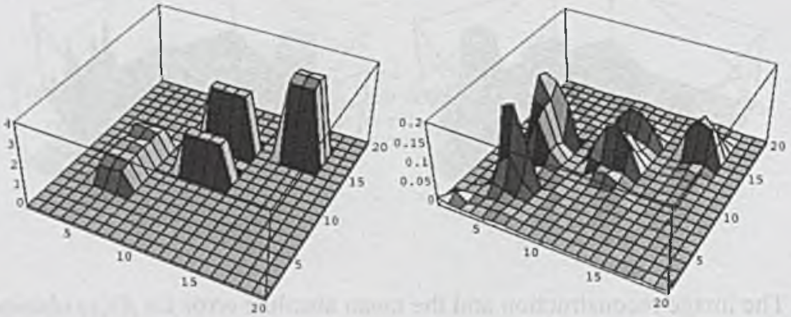


Fig. 9. The image reconstruction and the mean absolute error for *f(x,y)* obtained with algorithm
SZB-3 for *n* =20 × 20, *m*=788, *M*=28, *iter*=150 in the system (1 × 1)

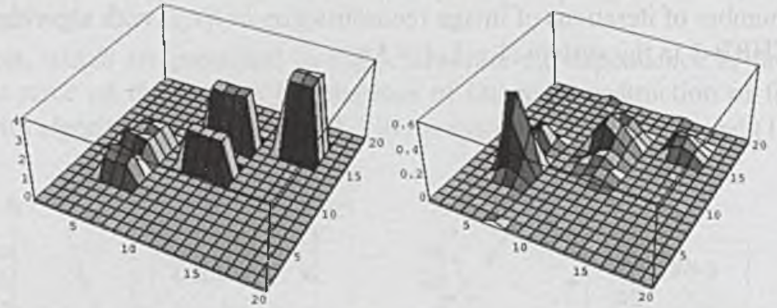


Fig. 10. The image reconstruction and the mean absolute error for *f(x,y)* obtained with algorithm
CHRB-3 for *n* =20 × 20, *m*=788, *M*=28, *iter*=150 in the system (1 × 1)

The plots, which are presented in Fig.11, illustrate the dependence of the mean absolute error on the number of iterations of image reconstruction of function *f(x,y)* with algorithms RB-3 and SZB-3 in the systems (1 × 1,1 × 1) and (1 × 1):

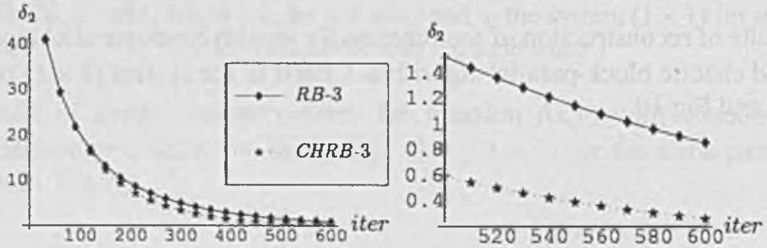


Fig. 11. Dependence of the mean absolute error on the number of iterations for image reconstruction of *f(x,y)* with algorithm RB-3 and CHRB-3 in the system (1 × 1) for *n*=20×20, *m*=788, *M*=28 and *iter* form 1– 600 (on the left side) and 500 – 600 (on the right side)

All algorithms were implemented on IBM/PC (processor AMD Duron XP, 1600 MHz) by means of C++ and MATHEMATICA 5.1. One iteration by means of Mathematica 5.1 was implemented approximately 0.5s for algorithm ART-3 and BPART-3, and in C++ one iteration for both algorithms is implemented in a real time.

5. Conclusion

In this paper we have presented a general model of asynchronous iterations and new chaotic iterative algorithms for reconstruction of high-contrast objects from incomplete projection data. These algorithms can be realized on a parallel computing structure consisting of elementary processors and some central processor, all of which are connected with shared memory. We study the quality and convergence of these algorithms by computing simulation on sequential computer. The experimental results show that convergent characteristics of block-parallel chaotic algorithm CHRP-3 are better by comparison with block-parallel algorithms RB-3 and SZB-3. Taking into account that the time of implementation of block-parallel algorithm on parallel is approximately less in *M* times (where *M* is the number of processors) with comparison with sequential computer, from results of computer simulation it follows that the time characteristics of block-parallel algorithms are better with comparison with sequential ART-3. From our results it also follows that the configuration (1 × 1, 1 × 1) is considerably better by comparison with the scheme (1 × 1). And for each considered scheme of reconstruction there exist the parameters which allow to obtain an enough good quality of reconstruction after some number of iterations but this number is considerably larger than for reconstruction with complete projection data. The number of iterations for achieving the stable reconstruction is approximately two times more for the second scheme by comparison with the first one. And this number is approximately 10 times more for the scheme (1 × 1, 1 × 1) by comparison with the case of the complete data.

REFERENCES

1. Patella D.: Introduction to ground surface self-potential tomography. Geophysical Prospecting, vol.45, 1997, p.653-681.
2. Williams R.A., Atkinson K., Luke S.P., Barlow R.K., Dyer B.C., Smith J., Manning M.: Applications for Tomographic Technology in Mining, Minerals and Food Engineering. Particle and Particle Systems Characterization, vol.12, N.2, 2004, 105-111.
3. Eggermont, P.P.B., Herman, G.T., Lent, A.: Iterative algorithms for large partitioned linear systems with applications to image reconstruction. Linear Algebra and Its Appl. v. 40, 1981, 37-67.

4. Gubareni N., Computed Methods and Algorithms for Computer Tomography with limited number of projection data, Naukova Dumka, Kiev, 1997, 328p. (in Russian)
5. Herman, G.T.: Image Reconstruction from Projections. Academic Press, New York, 1980.
6. Herman, G.T.: A relaxation method for reconstructing objects from noisy x-rays. Math. Programming v.8, 1975, p.1-19.
7. Herman, G.T., Lent, A., Rowland, S.: ART: Mathematics and application (a report on the mathematical foundations and on the applicability to real data of the Algebraic Reconstruction Techniques). Journ. of Theoretica Biology v. **43**, 1973, 1-32.
8. G.M.Baudet, Asynchronous iterative methods for multiprocessors. J.Assoc. Comput. Mach. v. **25**, 1978, p.226-244.
9. D. P. Bertsekas, J. N. Tsitsiklis, Parallel and Distributed Computation: Numerical Methods. Prentice-hall, Englewood Cliffs, NJ, 1989.
10. D.Chazan, W.Miranker, Chaotic relaxation. Linear Alg. its Appl., v.**2**, 1969, p.199-222.
11. Y. Censor, Parallel application of block-iterative methods in medical imaging and radiation therapy, Math. Programming, vol. **42**, pp. 307-325, 1988.
12. A.R. De Pierro, A.N. Iusem, A simultaneous projections method for linear inequalities. Linear Algebra and its Appl., vol. **64**, pp. 243-253, 1985.
13. A.R. De Pierro, A. N. Iusem, A parallel projection method of finding a common point of a family of convex sets. Pesquisa Oper., vol. **5**, pp.1-20, January 1985.
14. R.Bru, L.Elsner, M.Neumann, Models of Parallel Chaotic Iteration Methods. Linear Alg. its Appl. v.**103**, 1988, p.175-192.
15. Gubareni N., Generalized Model of Asynchronous Iterations for Image Reconstruction, Proc. of the third Int. Conf. on PPAM, Kazimierz Dolny, Poland, 1999, p.266-275.
16. N. Gubareni, A. Katkov, Simulation of parallel algorithms for computer tomography, in Proceedings of the 12-th European Simulation Multiconference, Manchester, United Kingdom, June 16-19, 1998, pp. 324-328.
17. Censor, Y. Parallel application of block-iterative methods in medical imaging and radiation therapy. Math. Programming, 1974, v.**42**, p.307-325.
18. De Pierro, A.R. 1990. Multiplicative iterative methods in computed tomography. Lecture Notes in Mathematics, 1990, v.**1497**, p.133-140.
19. N.Gubareni, A.Katkov, J.Szopa, Parallel Asynchronous Team Algorithm for Image Reconstruction, Proc. of the 15-th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics, Berlin, 1997, v.1, Computational Mathematics (ed.A,Sydow), p.553-558.

Chapter 4

Reverse engineering for different version of instant messaging log's such as Gadu-Gadu

Marek Piotr Stolarski
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Rafał Orlik
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Insitute of Physics, Wroclaw University of Technology

Borys Łącki
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Abstract

Analyzing of user's chats in Gadu-Gadu network gives completely new tool in computer forensic research. Due to false anonymity while using Internet networks one can easily violate the law. It often occurs while using instant messaging programs such as Gadu-Gadu, which is the most widely used network in Poland. Due to its proprietary type using reverse engineering methods is needed to allow others to read Gadu-Gadu archive. This can be used in prosecutions while specialized software can easily make several analysis automated.

6. Introduction

As the Internet infrastructure becomes much accessible, the more people star using it to communicate among them. Such a program, called instant messaging program, as Gadu-Gadu, Tlen, WP Kontakt and others are much more popular. In contrary to either the public telephone system or cell phones this type of communications gives its users the false feeling of the anonymity. It can make that people, who use this sort of programs, more easily break the law. Therefore, the analysis of the evidence secured by the Police or prosecutor, is the key aspect during process.

In this article we will focus on how to analyze Gadu-Gadu's logs. This communicator was chosen due to its popularity in Poland. Two formats of archives will be presented, namely, archives.dat and archibe.db files. These files corresponds to different versions of Gadu-Gadu, up to 7.0 (*.dat) and 8.0 (*.db).

7. File archives.dat

File archive.dat, which was used in the older version of Gadu-Gadu, is placed in user's home dictionary. All information such as either incoming and outgoing messages or short messages services (SMS) are stored inside it. Its internal structure is, unfortunately, quite complicated. Namely, this file contains a header, an index, the blocks and, finally, the messages. Their description is shown in the following part of this Section.

The main logical unit, which exists in the archives.dat, is called header, which contains such an information as user's ID number (UIN), control sum (based on CRC-32 algorithm), and the addresses of the following logical bloks. Format of the header is presented below:

- 0x00 (4 bytes) – string 'RC03' which identifies Gadu-Gadu archive,
- 0x08 (4 bytes) – offset (from the file's beginning) of the next logical unit, namely, index,
- 0x0C (4 bytes) – index's size in bytes,
- 0x14 (4 bytes) – offset (from the file's beginning) to the point where the data are placed,
- 0x24 (4 bytes) – user's identification number (UIN) in its 'hidden' format, namely, UIN^0xFFFFFD66,
- 0x28 (4 bytes) – control sum based on CRC-32 algorithm calculated for the first N bytes of archives.dat, where N is stored in 0x14 field.

If the calculated control sum for the first N bytes, assuming that in the 0x28 is stored 0x000000, is the same as the one taken from 0x28, one can assume that archive file is not corrupted. Then one can start analysing the next logical unit such as index. This unit stores where next logical units are placed inside archives.dat. The index is described as follows:

- 0x00 (4 bytes) – section's number,
- 0x04 (4 bytes) – number of blocks, which belongs to each section,
- 0x08 (4 bytes) – offset to the first block (the offset is calculated from the beginning of the data section in the archives.dat, see 0x14 field in the header),
- 0x0C (4 bytes) – offset to the last block.

As one can see in the previous description, the index is nothing more than a structure in which the offsets to the next logical units (such as blocks) are stored. These blocks have the following structure:

- 0x00 (4 bytes) – CRC-32 control sum,
- 0x04 (4 bytes) – section's number for which the block belong.
- 0x08 (4 bytes) – block's length,
- 0x0C (4 bytes) – offset to the next block,
- 0x10 (4 bytes) – number of bytes, which are stored just after each block; there are the messages stored.

At this moment, during the analysis of the Gadu-Gadu archive, one almost reaches the most important part of the archive, namely, the messages. These messages are described using two different structures, depending on its type – either incoming or outgoing. Each messages is described inside block as the following structure:

- 0x00 (4 bytes) – flags describing messages' state, it is either zero (0) and one (1) for deleted and undeleted messages, respectively,
- 0x04 (4 bytes) – offset to each message,
- 0x08 (4 bytes) – messages' size,
- 0x0C (4 bytes) – offset to the block for which the message belongs.

The last step, which is needed to read the messages, is to analyze the date which are stored in each blocks. The structure of it differs due to the message's type (either incoming or outgoing). For the incoming message it reads:

- 0x00 (4 bytes) – date when the message was sent (number of seconds from od January 1st 1970),
- 0x04 (4 bytes) – sender's UIN,
- 0x08 (4 bytes) – always zero (0),
- 0x0C (4 bytes) – date when the message was received (approximately),
- 0x10 (4 bytes) – message's length in bytes,
- 0x14 (0x10 bajtów) – encoded message.

The messages, both incoming and outgoing, are encoded using very simple algorithm. Namely, each byte of the decoded message (o_i) is defined as follows: o_i=ei ^ e_{i-1}, for o_i it is assumed that o_0=0xFF.

A slighty modified structure describes the outgoing message:

- 0x00 (4 bytes) – date when the message was sent,
- 0x04 (4 bytes) – sender's UIN,
- 0x08 (4 bytes) – number of receivers (for incoming message this field reads zero),
- 0x0C ([0x08]*4 bytes) – receiver's UIN's,
- 0x0C + [0x08]*4 (4 bytes) – date when the message was received,

- 0x10 + [0x08]*4 (4 bytes) – message's length,
- 0x14 + [0x08]*4 – encoded message.

In case when the outgoing messages describes Short Message Service (SMS) its structure is as follows:

- 0x00 (4 bytes) – date when message was sent,
- 0x04 (4 bytes) – receiver's name, a null-terminated string (contains N characters),
- 0x04 + N (4 bytes) – message's size (K bytes),
- 0x08 + N (K bytes) – encoded message.

As one can see the internal format of the archives.dat cannot be called user friendly. It is described using several logical units such as header, index, blocks and messages. From the user's point of view, the messages stored inside the archive, are not protected (see i.e. An algorithm used for their encoding). Fortunately, this was changed in the next version of the Gadu-Gadu archives, namely, archive.db.

8. File archive.db

Gadu-Gadu uses a new file to store user's messages since version 8.0. Namely, all the messages are stored using SQLite database. The archives.dat is no longer used. This approach makes that analyzing logs become much more simpler.

Using SQLite databse makes that all data are stored/receiver using SQL language. The archive.db file is, in fact, a databse described as follows:

```
CREATE TABLE chats (
chat_id                    INTEGER PRIMARY KEY NOT NULL,
interlocutor_id     NUMERIC NOT NULL,
is_initialized_by_user NUMERIC NOT NULL,
start_date                 TEXT NOT NULL,
first_communication_item_id   INTEGER DEFAULT 0
);
```

Table 'chats' is, from the log's analysis, nothing else than place where all user's chats are placed taking into account such conditions as (i) interlocutors, and (ii) send/received time. Most important fields are 'chat_id' and 'interlocutor_id' which describe (i) chat's identification number, and (ii) interlocutor's id number (primary key in 'interlocutors' table), respectively. These fields such as 'is_initialized_by_user' denotes message's type (0 – incomming, 1 – outcomming), 'start_date' – date and time when the message started (using 'yyyy-MM-ddThh:mm:ss' format), 'first_communication_item_id' – primary key in 'communications_items' table.

As it was mentioned before, the message is stored in 'communication_items' table, which is defined as follows:

```
CREATE TABLE communication_items (
communication_item_id INTEGER PRIMARY KEY AUTOINCREMENT
NOTNULL,
chat_id                          NUMERIC NOT NULL,
is_sent_by_user     NUMERIC NOT NULL,
start_date          TEXT NOT NULL,
content_type        NOT NULL,
content            TEXT,
plain_text_content TEXT
);
```

Field 'chat_id' in this table corresponds to 'chat_id' in 'chats' table. Due to this conntecion one can know who was the interlocutor, see 'interlocutor_id' from 'chats' table. The new fileds called 'content_type', 'content' and 'plain_text_content' appear. These fields are used to store such an information as message's type (0 – chats, 1 – SMS, 2 – file transfer), communication's string (HTML for chats, XML for either SMS or file transfer), string used for archive searching (could be an empty string).

The last table, which is defined in archive.db, is called 'interlocutors' and is defined as follows:

```
CREATE TABLE interlocutors(
interlocutor_id     INTEGER PRIMARY KEY NOT NULL,
identification_type    NUMERIC NOT NULL,
identification      TEXT NOT NULL
);
```

This table contains either UIN (for chats) or phone number (for SMS).

One can see that the new format of Gadu-Gadu archive is much more consistent. Namely, all the questions are described using SQL language such as

```
'SELECT * FROM 'chats' WHERE interlocutor_id = 123;'
```

This allows us to defined our user defined questions, which are handled by SQLite database. It is much more efficient that the methods used in the older one archives' format.

What is more important, from the user's point of view, its privacy is much better handled while using archive.db. Namely, this archive can be encrypted using AES256 algorithm instead of simple exclusive of method in archives.dat. Of course, for encrypted archive.db file one has to know its password using either (i) brute-force approach or (ii) dictionary method, or others.

9. Summary

We have shown how to reads messages which are stored using two completely different archive's format, namely, archives.dat and archive.db. Using a specialized programs to the forensics analysis one can quite simply finds these messages which fulfill specific conditions, such as date of sent, to whom it was send etc. On the other side, using two different internal formats requires that the forensic program should be carefully planned. It seems to convert both formats to some intermediate format, which is used by our program. In this case, when a new format of Gadu-Gadu archives will appear we need to write a rather simpler converter, the rest of the program will be the same.

REFERENCES

Authors have not provided any references.

Chapter 5

Effective methods for similarity detection in source texts

Marek Piotr Stolarski
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Rafał Orlik
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Insitute of Physics, Wroclaw University of Technology
Mateusz Kocielski
iConsulting Marek Piotr Stolarski
iconsulting@iconsulting.pl

Abstract

Finding similarity in source texts as a tool for intellectual property control. Methods of creating effective, both speed and accuracy. Typical methods to cheat detection system and ways to avoid it. Normalizing source texts such as converting from PDF, HTML, ODT, DOC into TXT, synonyms etc. Different languages in one document. How to handle with new words which do not exist in the dictionary?

1. Introduction

In today's world the information is the most important things. Who control it can control others. This makes that the information is very vulnerable, i.e. can be easily stolen from the personal WWW page. In this article we present our system for similarity detections in source text. This system which initially was created for academic centers such as universities and technical universities. With small modification it can be used by press agencies, WWW portals etc. In the next Section we present the methodology of development of such a system and typical problems which must be solved to make it bullet-proof. This all system uses Open Source programs such as Python (server side application), PostgreSQL (database engine for storing texts), Apache (user side).

2. Methodology of development effective algorithms for similarity detection

The simplest way for similarity detections is to use one of the string search algorithm, i.e. naïve, Rabin-Karp, Knuth-Morris-Pratt. This approach works well for identical texts. This method, although correct, will be very slow and easy to avoid. One can speed-up it by split source text into smaller segments, for example 10 words long, and calculate it's hash function, i.e. MD5. Then by simple asking database for these texts which contains the same hash signature, one can obtain these documents which the same fragments of text. Due to very small chances that different texts would have the same MD5 signature, one can be sure that it detects similarity.

The approaches mentioned earlier can both be easily deceived when one who steals some fragments of other's text, changes it a little, i.e. by either changing the construction of tense, or rewriting it from direct speech into reported speech. Eventually, those algorithms do not work well for smart enough thief.

So, we have to develop another method which will be as fast as text's signature approach, but will not be so easy to cheat as plain text search.

3. How to normalize source texts?

Due to several different document formats, such as *.pdf, *.ps, *.odt, *.doc, *.txt etc., a perfect system for similarity detections has to do some text normalization. The simplest way is to convert documents into a text format. Taking into account that there exists alphabets with some natiolan-specific character, one has to decide whether to use either ISO-8859-1 or i.e. UTF-8/UTF-16 encoding. Using Unicode encoding seems to be the best choice, because one can distinguish between all (?) available characters. Unfortunately, this approach is not the best. In fact using Unicode makes text analysis harder due to the fact that the same character could be coded into several different ways. This makes that Unicode is not the best choice from the practical way of view.

There is also another issue which should be taken into account. Namely, let us suppose that the source document was OCR'ed (Optical Character Recognition). During this process some part of texts could be wrongly recognized. For example, Polish character “ą” could be changed into “a”. For a human this error is easy to detect and avoid. The same process for a computer program it is extremely difficult. So, there is very important task to handle with: how to avoid this kind of situation? In our system we decided to treat all diacritic sign as there was none of them. Namely, all these character “z”, “ż” and “ź” are converted into “z”. Moreover, during the normalization process all characters are changed into their lower case version. At this point it is clearly seen that the ISO-8859-1 encoding system is used.

At this point we have normalized the source text into its lower-case ISO-8859-1 representation. Of course, in the original text there exists characters which are relevant in the written text, but, when spoken, are hardly to distinguish. So, characters such as “,” (colon), “.” (period), “;” (semicolon) have to be treated separately. We decided that we just simply remove them from the source text. Eventually, using our approach text like this “Litwo! Ojczyzno moja! ty jesteś jak zdrowie. Ile cię trzeba cenić, ten tylko się dowie, Kto cię stracił. Dziś piękność twą w całej ozdobie Widzę i opisuję, bo tęsknię po tobie” would be normalized into “litwo ojczyzno moja ty jestes jak zdrowie ile cie trzeba cenic ten tylko sie dowie kto cie stracil dzis pieknosc twa w calej ozdobie widze i opoisuje bo tesknie po tobie”.

Such a method would be sufficient for a simple similarity detection system. But, how to handle with methods for detection avoiding? This will be presented in the next Section.

4. Typical methods for avoiding similarity detection and how to handle them

In the previous Section we presented how to normalize the source documents. But what happed if someone changes the source text a little? Namely, it take the original document and changes all first person into the third one? For example, instead of write “I go to school” it writes “Tom goes to school”. As long as the “I” from the first sentence means exactly the same as “Tom” from the second one, both sentences have the same meaning. So, how to handle with such a smart person? There are more ways to deceive detection systems, i.e. combine two simple sentences into one complex sentence, i.e. rather than “Tom went to school. The weather was foggy.“ one writes “The weather was foggy when Tom went to school.” Another way of avoiding detection is using synonyms. This changes the original text and makes, usually, detection systems helpless. Let us compare two sentences: “The battery was discharged” versus “The battery was empty”. Both means the same, but different word appears. One again simple similarity detection systems might fail.

All these mentioned examples shows that designing smart similar detection system is not so easy as it looks. To prevent our system against such a problems we decided to use the following techniques: all words are transformed into their core version (i.e. makes is treated as make), and check whether there exists synonyms in the source texts. If it is true we change the synonyms into its one chosen representation of the synonym's group.

All these techniques together make that if someone want to cheat our system it will be forced to make a lot of changes in the source text and, therefore, it become an author instead of a deceiver.

5. Dictionary and works with words which do not exist in it?

At this moment our system for similarity detection seems to work perfectly. As an input it gets the source document, then converts it into a text format. After some techniques that makes cheating much more difficult the source text changes into words' cores representation. Due to the fact that all computers works using only numbers we make the last normalization. Namely, all the words' cores are changed into their numerical representation, i.e. each word, for which we know its core, has well defined integer number.

But, one open question arise: what should we do when a not known word appear? One possible solution is to calculate its new identification number. Of course, each word could have several forms due to the flexion. But, on the other hand, the core of the word can be easily (in most cases) determined. Usually, all the forms of the core has the form prefix-core-suffix, where prefix mostly have form like un-, non-, etc. So, by removing prefix one handle only with the core-suffix version of the new word. Moreover, only the first few letters belong to the core. At this moment we have almost our task done: let us assign for each letter some numerical value. By summing value with different weights (the most important are letters at the beginning, the last ones can be neglected, we can calculate the new word's identification number.

The algorithm, mentioned above, has still some open questions, i.e. what is the form of the weight function, how to assign numerical values for each letters, and more. On the other hand, it is a good point to start with. Moreover, for each words, which have similar core it gives either the same or close identification number. So, one can treat these new words, for which the identification number are close to each other, as a different versions of the same core word. Eventually, all the words, which appear in the source text, should be converted into their numerical id. This representation of the source text is placed in the database and its further analyzed.

6. Two and more languages inside source text

The similarity detection system, which was presented step-by-step in the previous Sections, works perfectly for a one-language document. But, what happen when two or more languages appear? Usually, the foreign languages denote that there are some quotation in the original text. So, the main document's language can be easily determined by counting the fractions of each languages.

To make this process simpler we could required that these languages are preselected by the document's owner.

The easiest way is as follows: as long as the words can be identified as a part of a one selected language, one can add some new some constant value for each word identification number. When all possible languages have different constant and for each words these identification number do not overlap, the system should work without any significant changes.

7. Similarity detection algorithm

The normalization process is finished. The source text was transformed from its original format, i.e. *.pdf, into its numerical representation. During this some additional techniques were used such as synonyms detections, converting words into its cores. One can start checking whether there are some similarity among source text and the others, which are stored in the database. But one question arised: how to detect such similarities? We have used very simple but useful algorithm. Namely, the normalized text is divided into frames, each N-words length. Then, for each frame S we ask the database if there exist another frames S_i (from different documents) that the cardinality of the intersection of these sets ($S \cap S_i$) is greater than i.e. M<N. So, one might suspect that these two frames (S from the original document, and S_i from the other one) have M/N common words. Hence, there is similarity between them. Of course, the system could only tell that there is some nonzero probability of such an event. The ration r=M/N gives us an information about the similarity level. For completely different frames r is equal zero (0), whereas for the same frames – one (1). When the similarity occurs this information should be stored in the database.

This part of similarity detection is the most crucial part of the system. One has to chose the values of both N and M correctly. These numbers cannot be either too small or too large. For small value of N the system sees only the fragment of the sentence, whereas for N too large too many sentences are taken into account. The same rule applies for M. The thumb-rule says that N should be equal to the average sentence's length, and M approximately 75-80% of N. As it was mentioned before, this is the most important part of the system. Moreover, it is also the slowest one. The experiments shows that the normalization process takes about few minutes to be accomplished. The next stage, detection, could span over several hours, depending on the database's size. This is needed to know which frame is similar to each one, and then, knowing where each frame starts and ends, to show these fragments to user.

8. Summary

We have shown a step-by-step approach for development the similarity detection system. We have led Reader from the source document emerging, than by its normalization and, finally, its conversion for storing in the database system. Typical problems, such as changing sentence type, using synonyms, and their solutions have been presented. There has been also shown how to deal with new words, which do not exists in the dictionary.

REFERENCES

Authors have not provided any references.

Part 2.
Social networks and Knowledge engineering

Chapter 6

Heterogeneous connections in social networks

Anna Zygmunt, Jarosław Koźlak, Marek Kałużny
Department of Computer Science,
AGH University of Science and Technology
{azygmunt, kozlak}@agh.edu.pl

Abstract

In this chapter, social networks constructed on the basis of two kinds of interactions between users - phone calls and exchanged e-mails – were analyzed. We present the algorithms for assigning roles and the results obtained when the network built from phone call data was expanded to include data from e-mail communication.

1. Introduction

Interactions between people in the current global climate are becoming more and more diverse and together with the development of computer technologies – easier for us to observe and analyze. These interactions may manifest themselves in the form of participating in meetings, phone calls, exchange of e-mails, connections with the use of Internet communicators as well as connections between people who send money via bank transfers to each other. These dependences may be described using a network of connections, where the role of the nodes, often called actors, are played by persons or groups of persons and the edges represent interactions among these persons.

The domain which focuses on research of these dependences manifesting in these relationships is known as a Social Network Analysis (SNA). It focuses on the analysis of dependencies between people, groups of people, organizations and markets. Analyzing the basic measures of the network (such as different kinds of centralities) one can study roles played by individuals in the network. Usually the analysed networks have a homogenous structure, where the dependencies between individuals are expressed using only one means of communication, for example, a network built on the basis of gathered data about phone calls.

In this paper, we will present possible approaches to the problem of adding new kinds of connections to such a network. Our method is presented taking as an example data about phone calls and the data originating from the exchange of e-mails.

2. Domain overview

Work on the analysis of social networks is nowadays very popular among researchers. Usually however, this kind of research focuses on networks which have only one kind of node and one kind of edge and such networks are called unimodal networks. Meanwhile, it is very often more useful to use more complex networks when analyzing situations and relations from real life. This complexity consists of different kinds of nodes (which represent people, locations and organizations), different kinds of arcs (which describe different kinds of relations which are as a result of the kinds of acquaintance - family members, friends, colleagues etc. or of the information exchange average: direct talk, phone calls, SMS or e-mail exchanger, a message written on a blog etc.) or having elements with different types of attributes.

Heterogeneous networks with different kinds of nodes are called multimodal networks, with different kinds of arcs – multi-relational networks, and these with different kinds of attributes associated with the elements – multi-featured networks [4]. Because of the additional complexity of the problems that appear, research on heterogeneous networks are carried out on a much lower level than that on unimodal networks. However, in our work presented in this paper, we decided to focus on these kinds of networks and especially on the multi-relational networks. The problem which appears during multi-relational network analysis is how to take into consideration the given relations. To treat them in the same way deprives the analysis of the information provided by the kind of relations, however to treat each relation in a special way, it is necessary to know how these differences should be considered, especially that a given relation may be more important for the description of the behavior of some nodes and less important for others: For example, there are people who are intensively using e-mail communication and there are other people who rarely or never use the Internet but only rely on phone calls.

In [1] an effort is made to determine the degree of importance for the different kinds of relations for a given query to a system. This approach is based on the learning of the optimal linear combination of weights of these relations which suits the needs of the user to the highest degree. The user may set different constraints (for example, a preference or requirement of a link or the path between given nodes to exist or the contrary – the preference or requirement of these links or paths not to exist) and on the basis of this, the weights for each kind of link is calculated. The main idea of this approach is based on finding such weights for given relations, that the connections between nodes which should be in the same sub-organization, are the closest possible and between these which should be separated into different organizations, should be as weak as possible.

In [3] a general approach to the transformation of algorithms for the analysis of the unimodal networks is proposed that takes on a form making it possible to analyze multi-relational networks. It is performed by an appropriate multi-relational network mapping into a uni-relational one, which then may be analyzed using the known methods. This operation is performed with the use of multi-relational path algebra, described in the paper .

3. Description of the system

The developed system is designed to analyze data which may be represented as a social network . We especially focused on the analysis of data about phone calls. In the obtained network, the nodes represent interlocutors and calls or SMS-es exchanged are represented as edges. Our system [2, 5] analyses such social networks and calculates the measures used in SNA, which are independent from the problem domain as well as domain-dependent measures.

The SNA measures describe the importance of the node in the network considering its different aspects like for example: average shortest (geodesic) distances between it and other nodes in the network (*Closeness/Bary Center*), location on the shortest paths between all pairs of nodes which may be selected in the network (*Betweenness*), numbers of incoming connections (*Degree In Centrality*) and outgoing connections (*Degree Out Centrality*), *Hubness* (which represents connections with important nodes in the network), *Authoritativeness* (represents connections of the nodes with nodes which have a lot of connections with other nodes), *Page Rank* (connections with nodes linked to important nodes, algorithms are similar to the ones used by Google search engine for building a ranking of matching sites) and *Markov Centrality* (probability that in random wandering in the graph, the token arrives to a given node, which means that a node and its neighbors have numerous connections with other nodes). A more detailed description of the measures might be found, for example, in [Jung].

The second group of measures are domain dependent measures: Mobility, Spatial range of incoming and outgoing calls, Length of calls, Average number of incoming/outgoing calls for one day, Average number of incoming/outgoing SMSs, Number of different incoming/outgoing interlocutors, Calls/SMSs ratio, Time period of activity in the network.

In our first approach [2] we assumed that there is a default set of roles which may appear in each criminal organization. This default set embraced the following roles: *Organiser, Isolator, Communicator, Watchman, Extender, Monitor, Liaison, Soldier, Recruit, Outsider, Accidental*. We then added a functionality of creating a set of roles for the given case and kind of criminal organization. The roles have assigned characteristics expressed by ranges of SNA measures. During the process of building the social network, the values of all measures taken into consideration are calculated. To make a comparison possible, they are transformed into the normalized brackets [0, 1].

The whole normalized range/period is divided into sub periods [0; 0.2], (0.2; 0.4], (0.4;0.6], (0.6; 0.8], (0.8; 1.0], the objective of this is to distinguish five states of the node for a given measure:

- the value of the given measure is significantly lower than an average for the whole network, which corresponds to the range [0; 0.2],
- the value of the measure is slightly lower than an average for the whole network. i (range/period (0.2; 0.4]),
- the value of the measure is similar to the average for the whole network, it corresponds to the middle range (0.4;0.6],
- the value of measure is slightly higher than the average for the whole network (range (0.6; 0.8]),
- the value of the given parameter is significantly higher than the average for the network and the node belongs to the set of nodes with the highest value of this measure (0.8; 1.0].

Then, the values of measures of nodes are compared with the patterns defined for roles. The closer the state, to which the value of the measure for the node is assigned, is to the state expected for the given role, the more points these nodes obtain with regard to this role. This operation is performed for all measures, roles and nodes. Finally, nodes have assigned roles in regard to which obtained the highest scores.

This solution was performed for the network with one kind of link. To apply it in the network with different kinds of links, the algorithm has to be changed.

4. Different kinds of edges

Our system enables us to analyze a social network which is constructed based on information about phone calls between different people. Analyzing the structure of connections between individual persons, values of parameters are determined. These values characterize the roles of the users in the network. Based mainly on the SNA parameters, the system attempts to determine the character of a given person and assigns them the most appropriate and fitting role, describing their behavior against other members of the network.

Taking into account only one way of communication – in our case the phone calls – simplifies the analysis, but it does not model reality completely. Generally, people belonging to the same organization communicate in different ways: e-mails, internet communicators etc. Additionally, communications between persons can be understood as establishing some relations and therefore analysis of who transfers money to who can bring in important information about the characteristics of relationships between people.

The manner of communication can also depend on the role a person plays in the organization. For example a boss can only communicate with directors of individual departments using phones, then middle management between themselves by sending e-mails. Therefore, such information indicates not only the fact that communication between persons exists, but also the nature of such communication.

Each kind of possible way of communication is unique. In this chapter the possibilities of integrating two ways of communications - phones and e-mails - will be presented. The prototype of the system has been developed in a way that enables us to expand this easily with additional ways of communications.

4.1. Methods of taking into consideration different kinds of links when assigning roles

Entering several ways of communications causes problems of assigning the roles in the form of a graph with aggregated locations of communications. One should consider what to do in cases when several persons using different kinds of communications interrelate with other people using only one kind of communication. When, for example, there is a communication in two different ways (phone, e-mail), between three persons, and in only one way in other cases. Evaluating roles in such a graph turn out to be fairly complex, and there are several solutions.

First, the simplest way to assign SNA parameters to roles is to treat each way of communicating equally.

Thus, in calculating the SNA parameters phase, we do not take the type of connections between people into account and we treat every occurrence of contact (edge) similarly, irrespective of their type. Such an approach has however disadvantages in that every person using more varied forms of communication receives extra bonuses due to only using various kinds of communication, in spite of their quality (more incoming and outgoing edges)

The second approach to the network with different kinds of communications takes into account the number of connections of a relevant kind. One should calculate which part of the entire connections constitute connections of a relevant type and acknowledge this ratio when calculating the aggregation of points for every role.

The third method is a manual definition of factors for each kind of communication. Then the scores from the individual kinds of communication is multiplied by that factor when calculating aggregated scores for every role. Such an approach could be used when it is known that some kind of communication (for example phone) is more important for the given analysis than others (for example e-mail). One could then give a higher factor for phone communication.

4.2. Modification of the algorithm determining the roles

Current algorithm determining the roles should be presented in several steps:

1. Designing the graph of interlocutors and the calculation of SNA parameters using the algorithms implemented in the JUNG library.
2. Saving computed values into the database.
3. Calculating the standardized values of SNA parameters.
4. Assigning the interlocutors to every role based on the scores they have received.

The range in the score matrix is defined for every role. Based on this range and the value of parameters, the scores for every role are assigned (the closer the defined range is to the standardized value of the parameter, the more scores are assigned to this role)

Example:

scores =

scores_of_BaryCenter
+scores_of_BetweennessCentrality
+scores_of_DegreeDistIn
+etc...

The sum of assigned scores for each parameter gives the overall amount of scores the interlocutor obtains for a given role.

5. Choosing the role for which interlocutor acquires the most amount of scores.

Entire analysis

The system will treat every kind of communication in the same way. It means constructing only one aggregated graph, in which the occurrence of whatever kind of communication (edge) is equally important. Next, based on this graph, the values of SNA parameters will be calculated. Thus, the algorithm of assigning roles will remain unchanged. Only the method of graph construction, which is used to calculate SNA parameters, should be modified.

Separate analysis

The system will treat every kind of communication independently, that is it should build the series of separate graphs; one for each kind of communication. For each of these graphs, the SNA parameters will be calculated for each site of communication according to the algorithm described above. Next, while the scores are being assigned (4. point of algorithm), the values of SNA parameters for every kind of communication will be summarized. It could also be multiplied earlier by the factor reflecting the importance of a given kind of communication defined earlier by the user.

Example 1:

```
scores =
     1.0*scores_of_BaryCenter[phone]
     +0.6*scores_of_BaryCenter[e-mail]
     +1.0*scores_of_BetweennessCentrality[phone]
     +0.6*scores_of_BetweennessCentrality[e-mail]
     +1.0*scores_of_DegreeDistIn[phone]
     +0.6*scores_of_DegreeDistIn[e-mail]
     +etc...
```

We assume that the user specifies the importance of the phone communication as 1.0, and the e-mail communication as 0.6.

In order to have a full view of relations between people in a network, it is necessary to join the knowledge obtained from the separate analysis of these two networks.

A logical approach combines the networks in such a way, that no node roles indicated as important during the analysis of one kind of communication could be reduced when applying another analysis.

The easiest way of joining two networks could be by ignoring the kind of communication and treating each fact of communication between two people uniquely (irrespective of the kind of communication). But there is a risk in such a joining: The roles of the nodes indicated as important during analysis of one kind of communication only, could be weakened. If we have, for example, much more data about e-mails than phone calls, then the phone calls would seem to be in decline.

It is worth noticing, that such joining only makes sense if we are able to join nodes for two sides of the different kinds of communication into one node (person). People, who do not communicate by more than one kind of communication, should be assigned to the role which was calculated in the analysis of the network regarding the given way of communication.

However, in the case of people who communicate using different kinds of communication methods, the joining should increase their position in the network, and at best, never decrease it.

In the simplest case, one could assume that the person who communicates with others in more than one way, obtains the highest role depending on the kind of communication.

In our work, we have developed more complex procedures of joining the locations of communication to a person. Every kind of communication could have influence on the final role. The roles, which the system assigns to each person, are determined based on the amount of scores obtained by that given person. For each role, the scores are determined for each type of SNA parameters, according to how far from the defined range for the given role, the normalized value of SNA parameters is located.

One can assume that, according to the character of role, the given type of SNA parameter is more or less important for any given roles. For example, for one role, the most important parameter could be PageRank, but for another – BaryCenter.

It is worth taking into consideration the method of reflecting these dependences between roles and kinds of parameters when combining the knowledge from different networks. When assigning the roles for single kinds of communications, each role declares the range of normalized values for each type of SNA parameters. If the normalized parameter values for each location of communication fit in this range, then the communication location receives the maximum defined amount of scores for that role. If the value is near that range, then it receives fewer defined amounts of scores.

In the original algorithm, the same amount of assignments of scores for fitting in the range for each parameter is defined. The role matrix can be extended by a store factor indicating the significance of the given type of parameter for a given role. Then it could be possible to indicate that, for example, for a given role, the parameter PageRank is more important than BetweennessCentrality.

Example 2:

Assigned number of scores, when the normalized value fit into the range: 10

Assigned number of scores, when the normalized value is near the range: 5

Importance factor for the parameter PageRank: 1.2

Importance factor for the parameter BetweennessCentrality: 1.0

If the values of parameters PageRank and BetweennessCentrality for a given site of communication fit in the defined range for that role, then the site of communication receives:

1.2*10+1.0*10=22 scores for that role.

In the original algorithm this site of communication would receive 20 scores.

This way, by properly defining factors, we can specify the importance of each parameter for each role.

In order to indicate the roles for each node combining different kinds of communications, it will be sufficient to add additional dimensions to the role matrix which describes a new kind of communication. Then describing the importance of each parameter separately for each kind of communication will be possible.

Such an approach seems to be sufficiently flexible. It is possible to easily specify which kind of communication is primary, by the relevant defining of the factors of importance.

Algorithm:

```
Foreach Person in AllPerson
  Foreach StronaKomunikacji in PersonSiteCommunication
    Foreach ParameterType
      var
RoleSet=AllRoles[SiteCommunication.KindOfCommunication AND
ParameterType]
      Foreach Role in RoleSet
        If SiteCommunication.ParameterValue[ParameterType] >=
Role.Left
            && SiteCommunication. ParameterValue
[ParameterType] <= Role.Right

   Person.ScoresForRole[Role]+=Role.ImportanceFactor*ScoresFor
Fit
```

5. Use case: A network of illegal drug traders

In the framework of the analysis we carried out, we focused on the case concerning the trade of illegal drugs. The performed analysis was based on the classical model of roles in a criminal organization, without taking into consideration the special features of the case. We also had information about the roles played by the most important members of the organization and their phone numbers. This information was used to evaluate the quality of the results. The obtained results consisted of the values of measures and role scores for the given users.

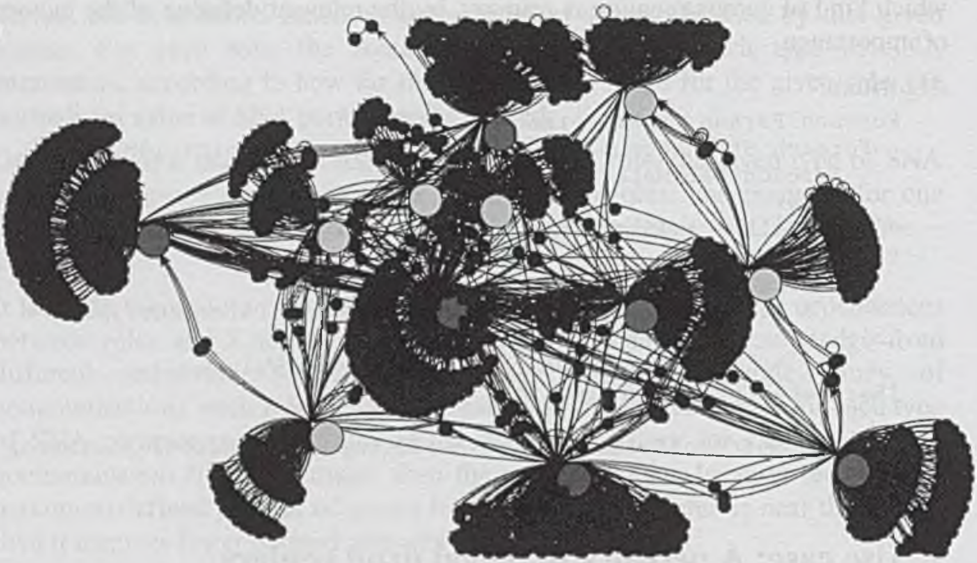


Fig. 1. A network built on the basis of phone calls

In this figure, the obtained social network for the phone calls data is presented. The users are represented by circles. If a user is represented as a larger circle, it means that the complete information about calls were gathered for it, in contrary to the small ones, for which we may have only parts of call data. The colour of the larger circle represents its number in the network: red – Organiser, violet – Liaison, blue – Monitor, pink – Recruit, green – Solder, grey – Outsider.

The next step was to create a network with two different kinds of links. To achieve this objective of three selected nodes (U2, U3, U6) and their interlocutors, generated information concerning e-mail communication was added. These modified nodes are emphasized in the fig. 2 by gold circumferences.

The goal of these changes was to represent a U2 node as the one which exchanges a lot of e-mail information, which in turn has a positive impact on its importance in the network. In the process of calculating roles in the multi-relational networks, we used the second approach from the previous section, with appropriate settings of the importance of weights for roles and networks.

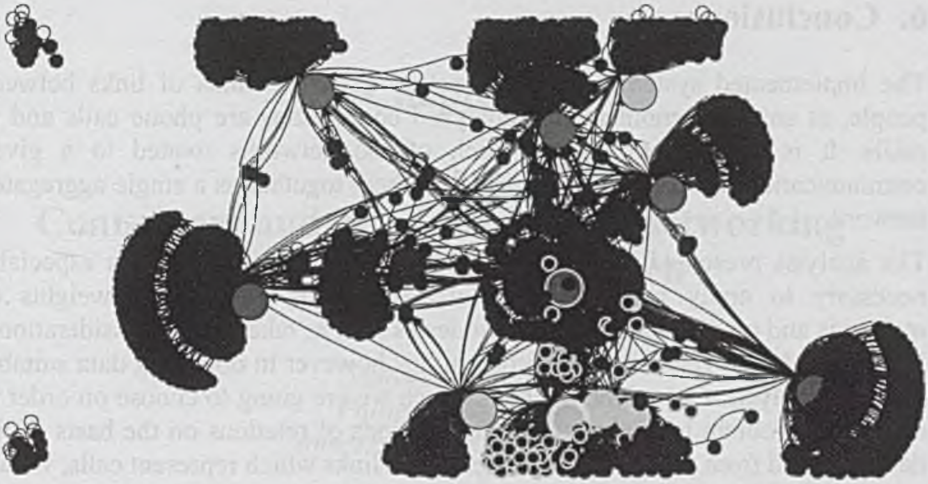


Fig. 2. The network after adding edges with information about e-mail interactions

Tab. 1. Case of illegal drugs trade, a selection of the most important numbers for the network built from phone calls. First values in the cells are scores assigned with regards to the given roles, the second number is the change of scores after adding three nodes (U2, U3, U6) and their neighbor nodes with the links representing e-mail communication

User	Liaison	Monitor	Organiser	Outsider	Recruit	Extender	Watchman
U_1	60, +5	30, +1	0, 0	0, 0	5, 0	15, 0	5, -5
U_2	**35, +20**	**45, +50**	0, 0	**15, +20**	**60, +55**	10, 0	**50, +10**
U_3	**40, +30**	**50, +50**	0, 0	**5, +5**	**30, +20**	10, 0	**25, +75**
U_4	50, 0	50, +1	0, 0	10, 0	40, 0	10, 0	**35, -35**
U_5	45, 0	**55, +0,5**	0, 0	0, 0	25, 0	10, 0	20, 0
U_6	10, 0	0, 0	65, 0	0, 0	0, 0	20, 0	15, 0
U_7	10, 0	15, +5	45, 0	0, 0	0, 0	35, 0	20, 0
U_8	50, 0	50, -9	50, 0	50, 0	**50, -45**	**50, -15**	50, 0
U_9	70, 0	20, +1	0, 0	5, 0	20, 0	30, 0	10, 0
U_{10}	30, +7,5	50, 0	5, 0	45, 0	15, 0	10, 0	50, 0
U_{11}	40, 0	45, 0	0, 0	10, 0	50, 0	10, 0	35, 0

The modifications of the network performed showed an important increase of some measures for the two modified nodes (U2, U3). Especially, the increase of the scores for the roles of Monitor, Liaison and Watchman were high. This is in accordance with our expectations and predictions, as the character of these roles is associated with the dissemination of information which may be provided by a node being a bridge between the sub-networks or having many connections to other nodes. The increase of the score for the recruit role (especially for U2) we would have to explain as a random side effect.

6. Conclusions

The implemented system is able to analyze different kinds of links between people, at any one moment; the analyzed connections are phone calls and e-mails. It is possible to analyze each of the networks related to a given communication method separately or all of them together as a single aggregated network.

The analysis presented here in this paper needs continuing as it is especially necessary to apply more complex algorithms for determining weights of measures and relations for the given nodes as well as take in into consideration a higher number of relations. The problem lies however in obtaining data suitable for such analysis. One of the methods which we are going to choose on order to avoid this problem is to generate different kinds of relations on the basis of the data obtained from phone calls. Alongside the links which represent calls, we are going to introduce links between the nodes which represent their participation in frequent sequences of calls and also a separate representation for exchanged SMSs.

REFERENCES

1. Cai, D., Shao, Z., He X., Yan, X., Han J., Mining Hidden Community in Heterogeneous Social Networks, International Conference on Knowledge Discovery and Data Mining, Proceedings of the 3rd international workshop on Link discovery, Chicago, Illinois, pages: 58 – 65, 2005.
2. Piekaj, W., Skorek, G., Zygmunt, A., Koźlak, J.: Środowisko do identyfikowania wzorców zachowań w oparciu o podejście sieci społecznych. w: Technologie Przetwarzania Danych : II Krajowa Konferencja Naukowa, Poznań, 2007.
3. Rodriguez M. A., Shinavier J., Exposing multi-relational networks to single-relational network analysis algorithms, The Computing Research Repository (CoRR), June , 2008, *Journal of Informetrics* (2009).
4. Singh, L., Beard, M., Getoor, L., Blake, M. B., Visual Mining of Multi-Modal Social Networks at Different Abstraction Levels. in: Information Visualization, 11th International Conference, 4-6 July 2007, pages 672-679.
5. Zygmunt, A., Koźlak, J.,: Zastosowanie podejścia sieci społecznych do wspomagania prowadzenia analizy kryminalnej dotyczącej danych billingowych. rozdział w: Praktyczne elementy zwalczania przestępczości zorganizowanej i terroryzmu : nowoczesne technologie i praca operacyjna, , Wolters Kluwer, Warszawa, 2009.

Chapter 7

Connectors and mavens in social networking - an agent-based approach

Tomasz Owczarek
Politechnika Śląska
tomasz.owczarek@polsl.pl

Abstract

The article presents an agent-based model in which artificial users choose a service they use at each iteration. Its aim was to study the role of mavens and connectors in social networking. The received results show that this role is nontrivial and ambiguous.

1. Introduction

Social networking services are becoming an integral part of our lives. With the use of web 2.0 technologies [13] and relying on network effect [18] they are examples of new models of business activities. However, although services like Twitter or Facebook exist for a few years, during last several months their popularity have grown much faster than earlier [19, 22].

One explanation of these phenomena is given by Malcolm Gladwell [5]. He explains that the ability of an idea or product to "tip" can depend on a very small group of people with some special abilities (he calls them connectors, mavens and salesmen). He provides some very convincing examples such as rapid popularity growth of Hush Puppies shoes in the middle of 1990s or Paul Revere's ride at the beginning of the American Revolution. The idea seems especially intriguing when comparing dates of the information about famous sportsmen [20] or other celebrities [21] starting using Twitter with graph of Twitter's popularity.

In this article an agent-based approach to study the role of mavens and connectors in social networking is presented. Model was constructed in which artificial users choose a service they use at the moment. Motivation for their choices is the number of their neighbors (in the social network graph) using the service, but there is also a lot of place for chance and randomness, representing all other unexpected factors. The aim is to test the two hypothesis. First, that mavens and connectors can speed up the moment when one service gains advantage over others. Second, that the circumstances when mavens and connectors all use the same service at some moment results in its more popularity in the long term. The basic assumption made in the model is that services are substitute goods and user can only use one of them at the same time. And following the network effect, the more friends choose the service, the more incentives a user has to choose the same one.

The article is organized as follow. Sections 2 and 3 provide theoretical background. Roles in social networks are discussed and conception of agent-based modeling and simulation is presented. Sections 4 and 5 contain model description and the results. In section 6 some ideas for further research are proposed.

2. Roles in social networks

When social networks are considered the deliberations almost always start with Stanley Milgram and his experiment [10] which initialized debates about so called "small world phenomena" and the concept of "six degrees of separation". Mathematical properties of this concept were studied by Watts and Strogatz [15]. They gave an algorithm for creating small world graphs (Fig. 1) which can be treated as a representation of real-life social networks. Graph theory is widely used in social networks researches (see e.g. [7]). Vertices (nodes) of a graph represent people and connections between them are edges (if there is an edge between two nodes it means that the people know each other).

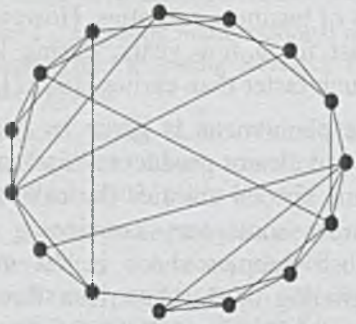


Fig. 1. Small world graph

In a small world graph there can be distinguished regular edges (linking nearest nodes), but there are also some irregular, "shortcuts", which connect distant nodes. These shortcuts lowers the number of intermediaries between any two nodes.

In Gladwell's "law of the few" concept [5] *connectors* are people with many contacts. In social network graph nodes which represent them have degree (i.e. number of connections to other nodes) above the average. But it is not only the number of people they know that matters. Connectors work as intermediaries between different groups, they link different circles of interest and are the channels for message and opinions passing between "different worlds".

Gladwell's *mavens* are people who know a lot and they are eager to share their knowledge. They are often obsessed with looking for occasions (e.g. finding a bakery where cheaper bread can be bought), but their knowledge is useful and they are often asked for advices. Others know that a maven person usually spends much more time before choosing a product or a service and their decisions are taken after long deliberations. So they must be accurate.

Gladwell mentions also about *salesmen*. They are "persuaders" (which mavens are not), they are very convincing and can get you to act like they want.

In the article only connectors and mavens are dealt with, but it should be remarked that the latter are considered as having also the features of salesmen. Connectors are people with many contacts. Mavens in this perspective are simply people whose opinions are more important.

3. Agent-based modeling and simulation

A simplest definition of agent-based simulation can be found in [12] where it is defined as *a simulation made up of agents, objects or entities that behave autonomously*. The agent's definition varies and different features are emphasized depending on authors [11, 17]. But they all agree that an agent is situated in some environment and able to make autonomous decisions [8]. If there are more than one agent then we deal with multi-agent system and some kind of communication between agents is also required [17].

As it is claimed by North and Macal, *agent-based modeling and simulation (ABMS) is a new modeling paradigm* [8]. It became used in social science with the publication of SugarScape model [4]. Axtell and Epstein called their approach "generative social science" as its aim was to generate artificial society of agents in which some macro-level similarities to the real-world situations could be observed. In this context ABMS can be treated as a computer research laboratory in which assumptions about real-world phenomena can be tested and explored [9]. And they gain more and more attention, especially when it comes to model networks of social interactions [1, 14].

There are many ABMS tools and environments [9]. For the purpose of this article the simulations were created and performed with NetLogo 4.04 [16].

4. Model description

Simulation starts with initializing users. Each of n_u users is randomly placed in a 2-d world with a torus topology (see Fig. 3). In the beginning users have equal probabilities of choosing one of n_s services. In the initial phase connections between users are also made (see subsection 4.1).

During each iteration users check which services were chosen by their neighbors and update their own preferences (it is described in subsection 4.2). After that they make their choices.

Simulation ends when all users use the same service. General overview of a simulation is presented below. The simulation is available at www.roz6.polsl.pl/pl/strona/zmi/ owczarek/sym/sym-users.html.

```
//initial phase
randomly place $n_c$ users
make connections
for each user
    choose service
//iteration
repeat
    for each user
            check neighbors' choice
    for each user
            update preferences
    for each user
            choose service
//stop condition
until all users choose the same service
```

Fig. 2 presents sample services' ratings during one simulation. There were two services and simulation lasted 530 iterations.

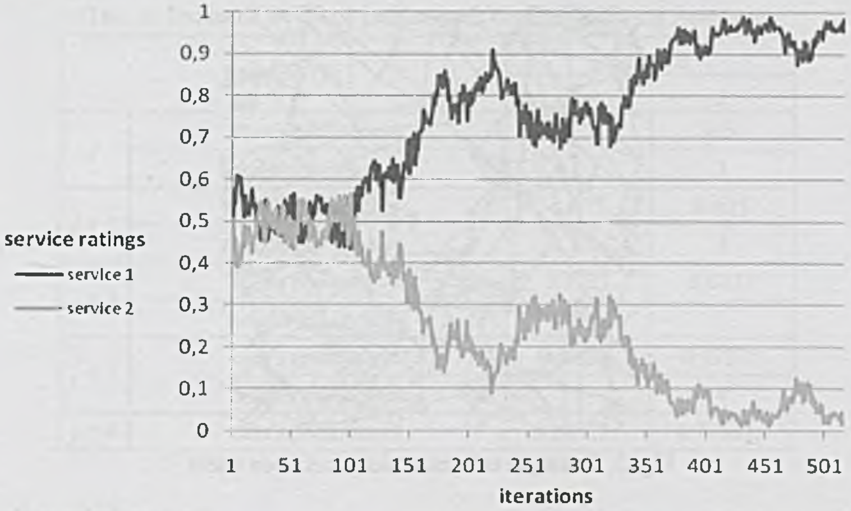


Fig. 2. Example of services' ratings during simulation

4.1. Connections making

The algorithm which adds connections between neighbors is a modified version of model proposed by Watts and Strogatz [15]. The modifications were made because of agents' random arrangement on a 2-dimensional plane. The algorithm consists of two main steps.

In the first step each agent makes connections to n_l (global variable which represents the basic connections number for each user) nearest users. If there are more than n_l users in the same distance then links are added to all of them. Notice that some users may have more connections than others.

In the second step r (global variable) fraction of all links are 'rewired', i.e. one end of link is exchanged for a randomly chosen node from the other users. There is a condition that this changed node cannot be a connector. Although there is no guarantee that the network graph will be connected, the probability that it is disconnected is very low when parameters n_u, n_l and r are chosen carefully.

A network example is shown in Fig. 3. Notice that the world used in simulation has a torus topology, i.e. its opposite edges are connected.

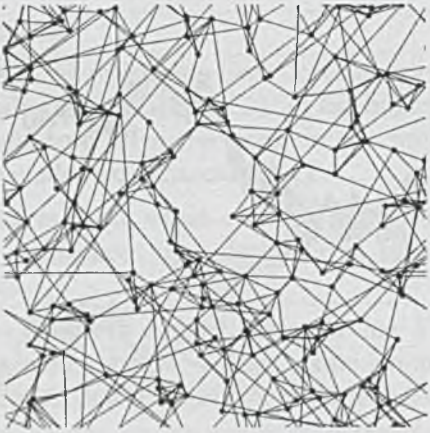


Fig. 3. Example of connections between users

4.2. Service choosing

The process of service choosing is based on a roulette wheel known from genetic algorithms reproduction mechanism [6]. Each user has preferences represented by a table with n_s elements. Values of this table sum up to one and the i-th element of the table represents probability that user will choose i-th service. Preferences of each user are updated during each iteration and they are combination of users past preferences and choices of its neighbors. The more friends using some services, the better chances that a user will choose this service during next iteration.

Let p_j^i be the probability that user chooses i-th service in j-th iteration, let n be the number of users neighbors and let n_j^i be the number of user's neighbors choosing i-th service in j-th iteration. Then the probability p^i_{j+1} is calculated in the following way:

$$p^i_{j+1} = \frac{1}{2}\left(p^i_j + \frac{n^i_j}{n}\right) \qquad (1$$

Tab. 1 presents the way in which user's preferences can change in time under the influence of other users. There are two available services and values in rows with "neighbors' choices" labels represent number of user's neighbors which decided to choose the service.

Tab. 1. Example of user's preferences changing during iterations

Iteration No.		Services	
		1	2
j	user's preferences	0.5	0.5
	neighbors' choices	1	3
$j+1$	user's preferences	0.375	0.625
	neighbors' choices	1	3
$j+2$	user's preferences	0.3125	0.6875
	neighbors' choices	2	2
$j+3$	user's preferences	0.40625	0.59375
	neighbors' choices	0	4
$j+4$	user's preferences	0.203125	0.796875

4.3. Special users

The probabilities that user is connector and (or) maven are determined by global variables r_c (describing connectors rate) and r_m (mavens rate). These probabilities are independent, so any user can be a connector, a maven, or both.

Connectors have more neighbors in network (which is determined by a global variable r_c). Mavens opinion are more important – in the model it means that their choices count as they were two or even more users (described by global variable m responsible for the strength of mavens influence).

5. Simulation results

There were two kinds of simulations performed. Each of them consisted of a few simulation series, differed in some parameters. Each variant of simulation was repeated 500 times and then some aggregate numbers were calculated. Statistical significance was tested according to [2], as a significance level α accepted 5%.

Tab. 2. Notation used to distinguish simulation runs with different parameters

	description
c0	connectors rate $r_c = 0$ (no connectors)
m0	mavens rate $r_m = 0$ (no mavens)
c10	minimal neighbors number for connector is 10
c14	minimal neighbors number for connector is 14
m2	$m = 2$ (maven's choice counts twice)
m3	$m = 3$ (maven's choice counts three times)

All simulations were performed with following parameters: $n_u = 200$, $n_s = 2$, $n_c = 4$, $r = 0.15$. Tab. 2 presents symbols denoting different simulation variants used in the article.

5.1. Simulation duration

First variants of simulations were to test if the presence of special users can shorten the time when one service gains maximum popularity. There were seven series of simulation runs, conducted with different parameters. In series where mavens and (or) connectors were present parameters r_c and r_m were equal 0.1 (10% chance that a user was maven and/or connector). Fig. 4 presents average numbers of iterations until simulation stopped in each series (averaged from 500 runs).

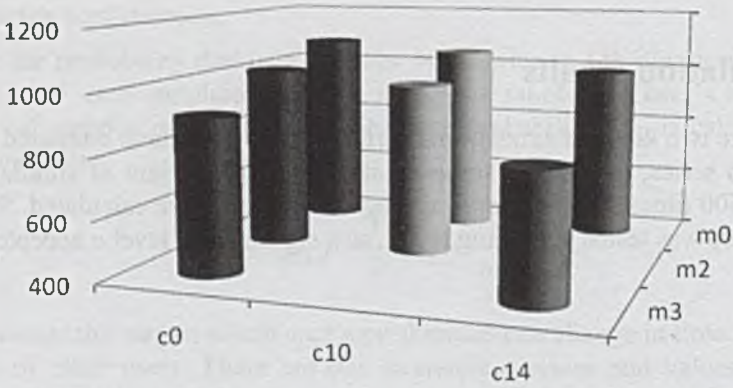


Fig. 4. Average simulations' durations

The figure clearly shows that increasing the "power" of special users' features (i.e. neighbors number in case of connectors and strength of influence in case of mavens) results in shortening the number of iterations. The accurate numbers are presented in Tab. 3. The average number of iterations until simulation stops when there are no mavens and no connectors (c0m0) is 1089.5. The presence of special users lowers the average number of iterations. In three cases (c0m2, c10m2, c14m3) the difference is statistically significant.

Tab. 3. Simulations' durations – means and standard deviations

		m0	m2	m3
c0	mean	1083.498	1016.052	917.608
	stdev	761.641	801.994	675.948
c10	mean	1051.996	979.898	-
	stdev	777.917	675.716	-
c14	mean	993.752	-	817.216
	stdev	762.581	-	616.1006

It is worth noting that the winnings' proportion (by "winnings" it is meant that one service has 100% ratings) of each of the two services was very close to 50% in every series. It proves that the implementation is not biased towards any of the services.

5.2. Winnings' proportion

This time six simulation series were performed. Their aim was to check if the proportion of winnings will be significantly different from 50% if one service will be the first choice (i.e. will be chosen in the initial phase) by all special users. There was a 20% chance that a user was maven and (or) connector, but only users which at the beginning chose service 1 were considered. This way, like in the previous case, about 10% of all users were mavens and about 10% were connectors. Service 1's winnings ratios in each of the six series are presented in Fig. 5.

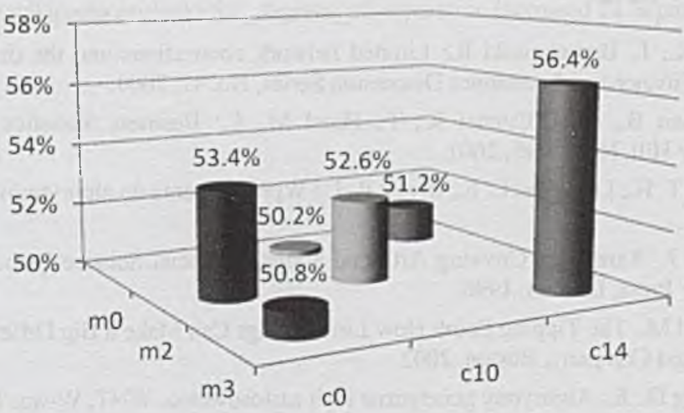


Fig. 5. Service 1's winnings ratios

It can be read from the figure that every time simulations ended more often with the success of service 1, but the differences in the winnings number were very small. Only in one series (c14m3) the proportion was significantly different from 50%.

5.3. Conclusions

The results clearly show that role of mavens and connectors in social networking is nontrivial. Even when the rules of behavior are very simple, they can contribute in rapid popularity growth of one of the services. However their presence is not unambiguous. It is not enough to ensure that all or at least most of the users identified as connectors or mavens use the same service at the same moment, hoping that network effect will guarantee its success.

6. Summary and further research

The article presents studies over the role of connectors and mavens in social networking. An agent-based simulation was built and series of simulation runs were performed. One potential extension to the model could be implementation of the mechanism of agents learning about services available to them. In the current model users have full information – results could be different if they recognized new options through the social interactions. Another future work is development of a more sophisticated agents' behavior. Users could make their decisions not only relying on their friends' choice, but there could also be included some kind of preferences or even external factors (e.g. commercials).

LITERATURE

1. Arrow K., J., Borzekowski R.: Limited network connections and the distribution of wages, Finance and Economics Discussion Series, No. 41, 2004.
2. Bowerman B., L., O'Connel R., T., Hand M., L.: Business Statistics in Practics. McGraw-Hill, New York, 2001.
3. Cormen T. H., Leiserson C. E., Rivest R. L.: Wprowadzenie do algorytmów. WNT, W-wa 2001.
4. Epstein, J., Axtell, R.: Growing Artificial Societies: Social Science From the Bottom Up. MIT Press, London, 1996.
5. Gladwell M.: The Tipping Point: How Little Things Can Make a Big Difference. Little, Brown and Company, Boston, 2002.
6. Goldberg D., E.: Algorytmy genetyczne i ich zastosowania. WNT, W-wa. 1995.
7. Huberman B., Romero D., Wu F.: Social networks that matter: Twitter under the microscope. First Monday [Online], Vol. 14, No 1 (20 December 2008), at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063, accessed 21 September 2009.

8. Macal, C., North, M.: Tutorial on Agent-Based Modeling and Simulation Part 2: How to Model with Agents. [in:] Proceedings of the 2006 Winter Simulation Conference, Monterey, California, 2006.

9. Macal, C., North, M.: Managing Business Complexity. Discovering Strategic Solutions with Agent-Based Modeling and Simulation. Oxford University Press, New York 2007.

10. Milgram S.: The Small-World Problem. Psychology Today(1), 1967, p. 60-67.

11. Russell S., Norvig P.: Artificial Intelligence: Modern Approach. Prentice Hall, 2002.

12. Sanchez S., Lucas T.: Exploring the world of agent-based simulations: simple models, complex analyses. [in:] Proceedings of the 2002 Winter Simulation Conference, San Diego, 2002, p. 116-126.

13. Shuen A.: Web 2.0: A strategy guide: Business thinking and strategies behind successful Web 2.0 implementations. O'Reilly Media, 2008.

14. Tassier T., Menczer F.: Emerging Small-World Referral Networks in Evolutionary Labor Markets. IEEE Transactions On Evolutionary Computation, Vol. 5, No. 5, 2001.

15. Watts D., J., Strogatz S., H.: Collective dynamics of 'small-world' networks. Nature, Vol. 393, 1998.

16. Wilensky, U.: NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL, 1999.

17. Wooldridge M.: An Introduction to MultiAgent Systems. John Wiley & Sons, 2002.

18. http://en.wikipedia.org/wiki/Network_effect (accessed 21 September 2009).

19. http://www.alexa.com/siteinfo/twitter.com (accessed 21 September 2009).

20. http://mashable.com/2008/11/21/shaq-twitter/ (accessed 21 September 2009).

21. http://jezebel.com/5217693/the-celebrity-twitter-backlash (accessed 21 September 2009).

22. http://www.alexa.com/siteinfo/facebook.com (accessed 21 September 2009).

23. http://en.wikipedia.org/wiki/Six_degrees_of_separation (accessed 21 September 2009).

Chapter 8

Analysis of social data dynamics

Jarosław Koźlak, Anna Zygmunt, Tomasz Grabiec, Anna Pietraszko
Institution Department of Computer Science,
AGH University of Science and Technology
{kozlak, azygmunt}@agh.edu.pl, {aniapek, tgrabiec}@gmail.com

Abstract

The paper will present work on the analysis of the dynamics of social networks built on the basis of data originating from telephone calls. Different approaches will be used to analyze these data - social network analysis metrics and their variability over time, prediction of future values of these measures and the detection of anomalies causing significant changes in the behavior of the network.

1. Introduction

The behaviour of social organisations, dependences between their members and roles performed by them are a subject of research of the domain called Social Network Analysis (SNA) [4, 9]. This kind of analysis may determine the importance of the given nodes on the basis of measures which describe their location in the graph, their relations with other nodes (such as the average shortest path to other nodes in the graph, existing on the shortest paths between pairs of nodes, numbers of connections with other nodes and the importance of neighbours – the nodes with which a node has direct connections) and a recognition of strongly connected substructures in the organisation.

Actual organizations, described by a social network, are subject to change over time. There are changes of roles in the organisation associated to individual members, the members may also leave or join the network. Similarly, an internal structure of the organisation, existing sub-structures and cliques may vary in time. As a result of these changing interactions between members, values of the SNA measures to corresponding nodes are changed, and the nodes are assigned other roles.

This paper deals with the problem of the values of measures of nodes changing related to the flow of time and how these changes affect the roles assigned. We also predict future values of these measures and compared results originating from different prediction techniques.

2. Domain overview

The analysis of the social network dynamic is a subject of a lot of research being carried out today. Different measures of the description of changes of the analysed society are proposed in [12], which was a criminal organisation. These measures are: centrality of nodes, density, cohesion and stability of groups. The analysis described in [7, 12] concerned the illegal trafficking of drugs and a terrorists networks (Al-Qaeda).

A classification of approaches to the analysis of the social network dynamic is presented in [7]. The authors distinguish between descriptive methods, statistical methods and simulation methods. The objective of the descriptive methods is to find changes of the network structure, analysis of the conformity of the analysed empirical data and verified sociological assumptions. The statistic methods describe changes of the network and explain the reasons of these changes, the changes come from the stochastic processes, such as reciprocity, transitivity and balance, determining the network behaviour. In [11] a wider overview of the statistic methods is given. In particular, the models of the network can be considered as continuous-time Markov chain models. The simulation methods are based on the application of the Multi-agent approaches to the analysis of the network dynamics.

Important elements in analysis are the identifying of substructures and cliques as well as the evolution of these structures in time. In [12], apart from the classical measures which describe nodes in SNA, the group level measures, such as link density, group cohesion and cluster stability are introduced. The link density describes the completeness of connections between group members and is expressed as a quotient of the sum of connections existing between the members of the group to the maximum possible number of connections between them (which means that each member of the group has connections with every other member of the group). Group cohesion informs us if the relations of members of the group one with another are stronger than their relations with nodes outside the group. The measure is calculated as a quotient of the average number of connections of the group members with other group members to the average number of its connections with the nodes not belonging to the group. Finally, the third measure, the group stability, describes how the group is calculated in time. The stability of the group from the time point t_1 to the time point t_2 is defined as a quotient of the quantity of the set being the common part of the two sets which contain elements of this group in times t_1 and t_2 to the quantity of set being the sum of these sets.

The [3] uses the Hidden Markov Models for identifying subgroups with suspicious behaviour in a social network. The goal is to analyse criminal activities.

An idea of a multi-agent system used for the analysis of social networks is presented in [6]. The approach is based on the assumption that each user in the network tries to optimise its individual utility. This utility is dependent on the existence of the connections with other users with a given configuration. The existence of each connection on the one hand increases the utility of the user but on the other, the user has to incur costs which lead to a decrease in utility. Such networks evolve up to the moment of achieving a maximum global utility.

Various systems for analysing static and dynamic social networks with the use of different approaches have been developed. The interesting attempt of constructing an environment which integrates features of different systems of these kinds, either cooperating with them and using their analysis or exchanging data with different formats with them is described in [2].

In our analysis, in addition to values of measures, the information about the roles attributed to the neighbouring nodes was taken into consideration. Alongside this, tests were carried out to detect the time of occurrence of important events affecting the organization, and for this we used the control cards method (cards preservation process) [10], and in particular CUSUM card counting based on the cumulative sum of parameters describing the process [5]. The concept of applying the CUSUM approach for analysing the dynamics of social networks was proposed in [7].

3. Analysis of network dynamics

In this section the concept of our network dynamics analysis, with its objective to describe the behaviour of nodes and their relations with other nodes, is presented.

In its current stage, our approach is based on the calculation for each node in the network, which represent phone users, values of various measures, either classical measures used in the Social Network Analysis, or special measures, dependent on the problem domain, introduced by us.

The SNA measures taken into consideration are: *Bary Center*, *Betweenness Centrality*, *Degree In Centrality*, *Degree Out Centrality*, *Hubness*, *Authoritativeness*, *Page Rank*, *Markov Centrality* [8, 13].

The second group of measures, the domain dependent measures, are: mobility, spatial range of incoming and outgoing calls, length of calls, average number of incoming/outgoing calls for one day, average number of incoming/outgoing SMSs, number of different incoming interlocutors, number of different outgoing interlocutors, calls/SMSs ratio, time period of activity in the network

In the analysis carried out, the schemes of roles in a criminal organisation were considered. These schemes were either the default - domain independent - or domain dependent, depending on a given kind of criminal activity. Domain-independent schemes embrace the following roles (majority of them were proposed in [1]): Organiser (constitutes the core of an organisation, managing the functioning of the organisation), Isolator (individuals or groups, isolate the core of the organisation), Communicator (individuals responsible for the transfer of information between the parts of the network), Watchman (take care on the security of the organisation and minimise its susceptibility from external attacks and infiltration), Extender (extends the organisation, recruits new members), Monitor (takes care of the efficiency of the organisation, reports weak points), Liaison (individuals belonging to the criminal organisation but at the same time are active in other networks, legal governmental, politics or financial institutions), Soldier (does not play an important function, executes orders of other members with higher position), Recruit (new members of the organisation), Outsider (person who does not belong to the criminal organisation, for example, a family member of the criminal), and Accidental (a person who has some occasional interactions with the members of the criminal organisation, for example a pizza vendor).

After we carried out a general analysis of the network structure and network, and its behaviour, we then completed a detailed analysis, including an analysis of the roles and assignments to the neighbourhood of roles, variability of measurement and dynamics models.

General analysis on the network. For the analysis of social networks, particularly the analysis of the dynamics, it is useful to carry out a general network analysis at the beginning. This analysis is intended to detect the quantities of callers on the network and the calls carried at subsequent intervals. This allows us to define the time period of analysis and possible periods for which the amount of information collected is significantly lower, which may be due to limited data collected only for certain callers. For a comprehensive analysis, the application of CUSUM cards for detecting clear disturbances in the network and the times in which they occurred can also be included.

Detailed analysis. Further detailed analysis may be carried out using various methods to select the interval for analysis (analysis of the neighbourhood of roles and its dynamic as well as metrics, their average values and variations) and use different algorithms for the classification of roles (which affects the analysis of roles and assignments to the neighbourhood of roles). We decided to choose the following modes for defining the ranges of intervals to be analysed:

- using a simple division into disjointed segments, such as the period of a week, and then carry out separate analysis for each of the sections,

- using cumulative distribution – analysis includes calls from the previous interval, then the testing interval is extended by one interval unit (lists of calls were successively analysed with the first week, the first two weeks, etc. up to and including the last week,)
- using the sliding window – a fixed-size analysis interval (as in the simple division) is selected for analysing calls taking place in a time window. At each successive analysis interval, the window is shifted forward by one day. The result is that conversations at the earliest times of the previous analysis were not taken into account in the next analysis.

Classification algorithms are used for assigning different roles taken from collections of roles which also have patterns defined for these roles. In the analysis carried out, either a core/default set of defining roles or the domain-dependent sets of roles adjusted to given kinds of cases were taken into consideration,. For further analysis, presented in the scope of car theft, the following roles were used in addition to a standard set of roles: Organizer, Receiver, Soldier, and Outsider.

4. Performed analysis of the network

During this analysis, the numbers of active speakers in the coming weeks of storing the data are given. The use of the CUSUM analysis cards, verifies whether distributions of values for each measurement in the network can be treated as normal distributions (this assumption is important because it is used in further analysis), and whether and when a disturbance of the network takes place. We were analysing classical measures used in a SNA, calculated using Jung library, such as authority, concentration, centrality, center of gravity, betweenness centrality , input and output degrees of vertex, Markov centrality and Page rank. Analysis using CUSUM cards is based on the detection of deviations in measurements from a given distribution. When it exceeds an established threshold, an alarm is triggered. In addition to identifying when the alarm is set off, this approach enables the user to identify the beginning time of the trend that has caused the alarm Alarms were observed for the bary center measure (interval 26), betweeness centrality (interval from 13 to 26), output degree (intervals from 16 to 19 and from 25 to 26). Analysis of the distribution of Markov centrality showed that the value of observation does not have a normal distribution due to the disturbances taking place in intervals 3, 16 and 18, making it difficult to analyse this measure.
In fig. 1 CUSUM analysis is presented for the betweeness centrality.

The analysis is based on the calculation of the calculated sum. Also, the alarm values are marked. The value of the cumulated sum is calculated using the equation (1).

$$C_i = \sum_{j=1}^{i} (xj - \mu) \tag{1}$$

where:

C_i – value of the cumulated sum in the i-th bracket,

x_j –average value of the SNA measure (which is in this case the Betweeness Centrality) in the j-th bracket

μ – estimated average value of the measure, which serves as a reference value, calculated as an average of several first x_i-values in the sequence.

This analysis is based on the differences between mean values, collected at regular intervals and the baseline. It allows us to detect the following changes to the average value of the process. The marked point is the sum of deviations of the measured values of the value target, in all previous measurements.

As a level indicating the change, we selected 3 sigma distance, equals to three standard deviations, which are indicated by a dotted line

The normality chart, not enclosed in these papers, proves that the analysed values have normal distribution, which allows a correct application of the approach. There is an alarm signal in the bracket number 26, it indicates a decrease of the average value of the Betweeness Centrality measure, which started in bracket 23. Another alarm takes place in bracket 13 and represents an increase of the average value of the measure starting in the 10-th bracket.

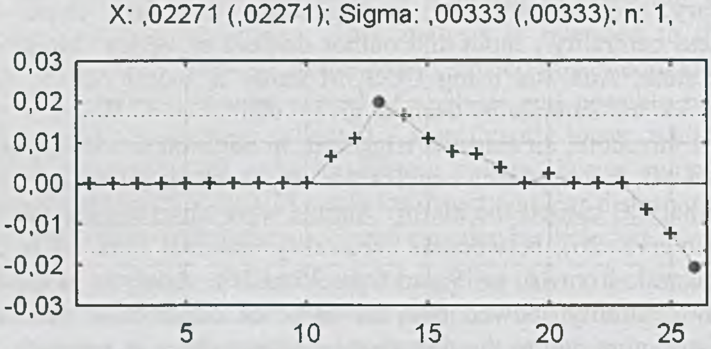


Fig. 1. Betweeness centrality. The deviations from the average value in normal distribution.

Analysis of the roles of the neighbourhood. In order to verify the correctness of the results and to observe how the organization functions, it is useful to observe the neighbourhoods between the roles.

We performed numerous studies of such neighbourhoods both for the static model and for the different versions of a dynamic model.

The experiment case concerns car theft. The analysis for the two kinds of time brackets: simple period (7 periods, 30 days length each) and cumulated period (each period has the same starting point, but each subsequent is 30 days longer). The experiments and analysis was performed for the two sets of roles: the default and domain dependent. The domain dependent set of roles embrases: Organiser, Receiver, Soldier and Outsider.

In the case of the domain dependant set of roles for both methods of defining time periods, one can see smaller variability in graphs in the subsequent time periods. The set of roles in the subsequent periods is stable, which is different from the graphs for the standard set of roles where the level of variation is high.

For the default set of roles and cumulative method of defining time periods, the results obtained in subsequent periods are more and more similar. It was to be expected as the influence of the old historical results is more and more significant. In the simple division method, the subsequent images of the network are different from one another to a degree independent from the number of the period, but they are only consequences of the phone calling activity in a given time. For the cumulative division roles, migrations are easy to recognize, it is especially visible for the roles of Organiser and Receiver. The migrations end only in the last, very wide, periods when the network becomes more stable.

In the cumulative division the new periods contain the previous periods in themselves, so the decrease of the values describing the communication between roles may be explained only by the migration of users between nodes. For the simple division, the observation of migration is more difficult and to analyse the degree of the phenomenon, a special, separate analysis is necessary.

Sample results are in fig. 2. The thickness of the arrows in the figure represents the relative number of conversations between roles, the numbers in the nodes - the number of instances of the given roles. The three numbers on the branches are: the number of different interlocutors initiating connections between the roles, the number of different receivers of these calls and the number of calls made between the roles.

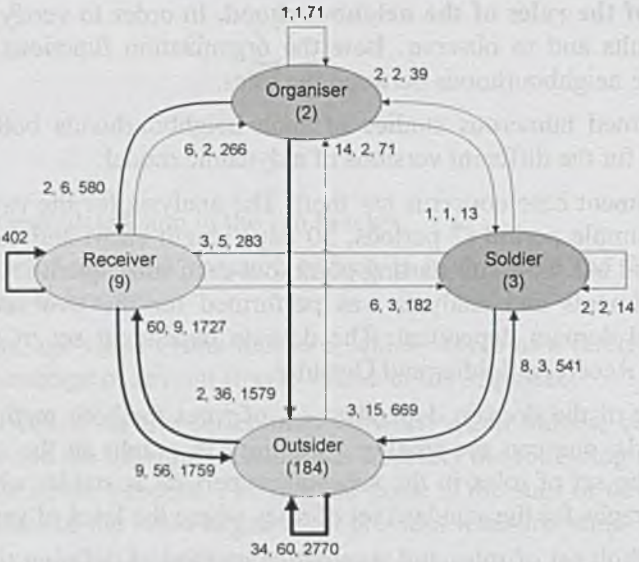


Fig. 2. Relationships between roles. The thickness of the arrows represents the relative number of conversations between roles, the numbers in the nodes - the number of instances of the given roles, the three numbers on the branches are: the number of different interlocutors initiating connections between the roles, the number of different receivers of these calls and the number of calls made between the roles

You can see significantly strong ties between the roles of Organizer and Receiver and Receiver and Soldier, which is consistent with the model of the organization, where Organiser is a leader, the Soldiers carry out the theft, and Receiver buys stolen goods. You can also note the relatively strong relationship between Receiver and Organizer or a Soldier and Outsider, those relations are not the result of the organizational model, and this may require additional studies to clarify the reasons of their existence.

Analysis of variability of measures. Analysis of variation of measures for each node has as its goal to show how the behaviour of the user in the organisation changed. In addition, it enables us to predict future behaviour of the user. Based on these predicted values, we can assign a predicted future role of the user. A number of experiments were performed, concerning the prediction of measures for different nodes (particularly the significant ones). Here we will present the results obtained for the case of car theft. We were using different kinds of approximation: polynomial, trigonometric and exponential. Prediction errors obtained were significant, but it can be concluded that the prediction correctly showed existing trends of an increase, decrease or the stability of the measure. The highest accuracy could be obtained using the exponential prediction; the results presented below will apply this kind of prediction. In the following figures we present the selected results of the prediction for the measurement of hubness (both in a simple - fig. 3 - and cumulative versions – fig. 4)) for the node, which was later identified as the leader - the Organizer of the criminal group. All charts have two vertical axes, the one on the left - the main – is for visualising the values (as specified by the name of the measure and is denoted by a continuous line) and predicted values (determined by the serial alignment L). The one on the right is the secondary axis scale with greater accuracy, which presents the rest/remainder which is the difference between the obtained prediction and the actual value. Analysis of tests for measure of hubness shows that the results for a simple division (fig. 3) are affected by a very rapid change and therefore a very large prediction error. This is understandable, since each interval is treated completely independently and any derogation periodic behaviour of nodes is very clearly visible. Concerning the value of the hubnesss, analysed in the cumulative form (fig. 4), the errors are clearly smaller, since successive values were also calculated on the basis of previous periods.

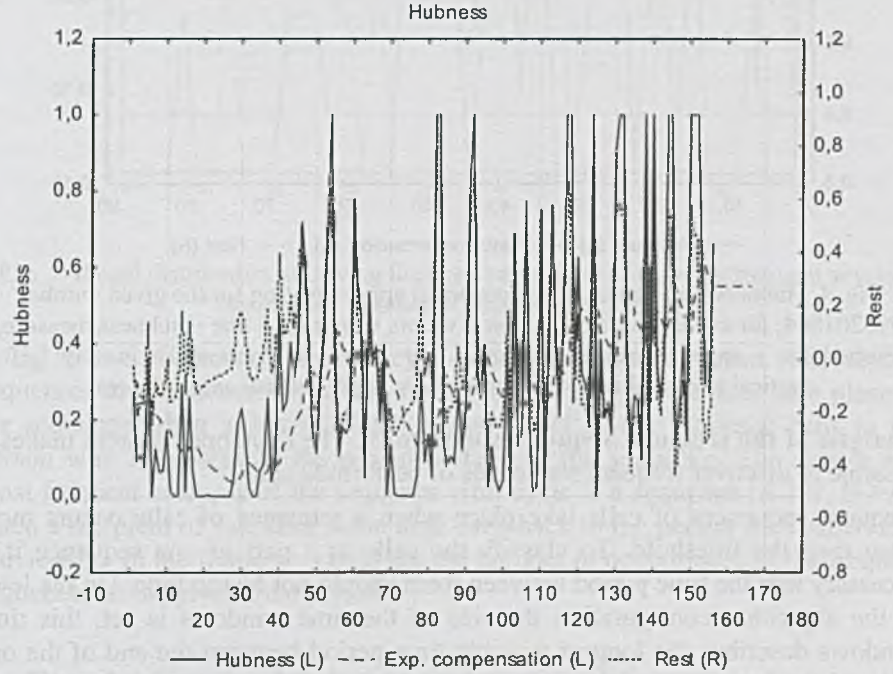


Fig. 3. Hubness measure and its exponential approximation for the given number 998201864, for periods of simple division. Continuous line – hubness measure, dashed line – approximated value, dotted line – the rest. Horizontal axis –day: Left vertical axis –hubness or its prediction: Right vertical axis – the rest

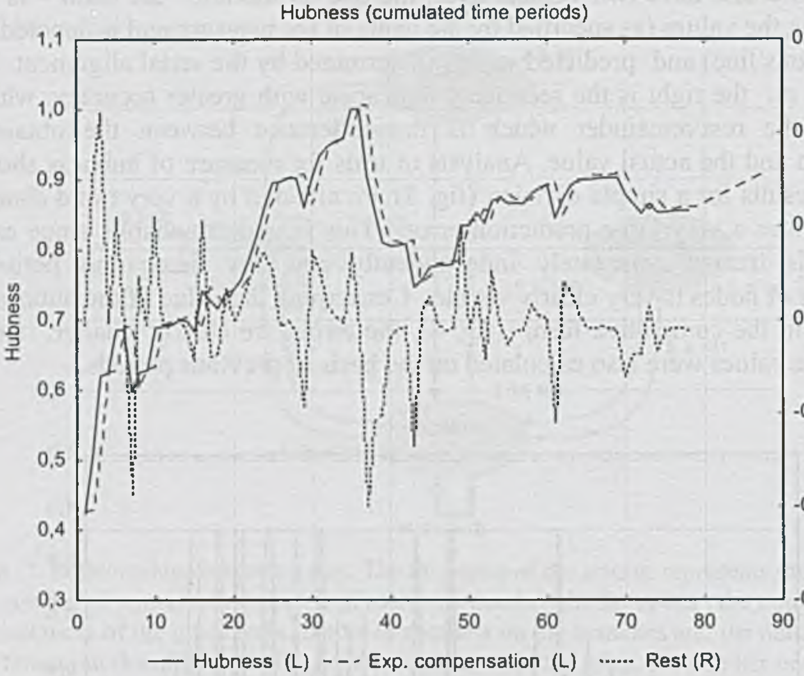


Fig. 4. Hubness measure and its exponential approximation for the given number 998201864, for cumulative distribution division. Continuous line – hubness measure, dashed line – approximated value, dotted line – the rest. Horizontal axis –day: Left vertical axis –hubness or its prediction: Right vertical axis – the rest

Analysis of the frequent sequences dynamics. The developed system makes it possible to discover frequent sequences of performed calls.

Frequent sequences of calls take place when a sequence of calls occurs more often then the threshold. To classify the calls as a part of one sequence it is necessary that the time period between them should not be too long. On the level of the algorithm configuration the size of the time windows is set, this time windows describes the longest possible time period between the end of the one call and the beginning of the second one which allows to acknowledge them as an elements of the one sequence.

Frequent sequences of calls are recognised using the Prefix Span algorithm. After identifying the existence of a given kind of sequence, we would like to verify the regularity of its occurrence in time. Different conclusions may be drawn if a given sequence appears several times during the day and different when it occurs regularly in the given day of the week at the given time. Identifying users who initiate phone calls with users which are under observation or with users with important roles may suggest that these users are associated to a high degree with the organisation being analysed.

Analysis was performed for several test cases such as a car theft organisation and an organisation which sold and distributed, in an illegal way, the possibilities of the successful passing of exams driving licence exams.

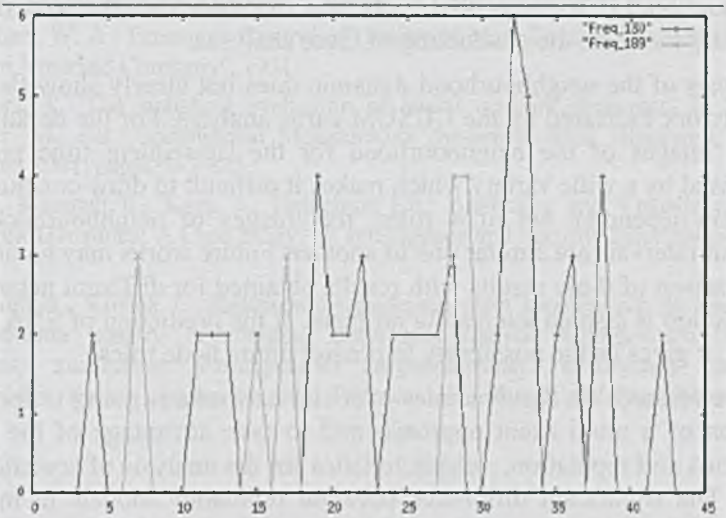


Fig. 5. Illegal distribution of driving licence exams. Two of the most frequent acyclic sequences, number of occurrences in time

In fig. 5 two the most frequent acyclic sequences are presented. The acyclic sequences are sequences which do not contain cycles. The cycles take place in the sequence when a person which appears later in the sequence calls to the person who is present in the previous stage of the sequence. The simple and most frequent example of the sequence with cycle is a sequence (A->B, B->A), when a recipient of call after some time calls back to the person who called him previously. In the figure we could see the number of occurrences of the frequent sequences in the given time periods.

5. Conclusions and future works

The implemented environment makes it possible to analyse the dynamics of the network, based on tracing the changes of the measure values, but also from the point of view of the role neighbourhood and the dynamics of the occurrence of frequent sequences. As a results, we succeeded in developing a more complete view of behaviours of a social network than the one obtained when applying only the statistic approach.

The analysis using the CUSUM cards confirms its usefulness as methods that can be applied, for quickly detecting changes in the social network. To better evaluate the results and get answers for several questions, more tests are necessary. Some questions are as follows: How the alarms change for different measures and different methods of dividing time periods or how to adjust the size of elements used while calculating the reference value – estimated average value (equation 1)? Implemented tools and extensions of the systems should significantly facilitate the conducting of these analyses.

The analysis of the neighbourhood dynamic does not clearly show the changes of the network indicated by the CUSUM cards analysis. For the default set of 9 roles the images of the neighbourhood for the subsequent time periods are characterised by a wide variety which makes it difficult to draw conclusions. For the domain-dependent set of 4 roles, the images of neighbourhoods in the subsequent intervals are similar one to another. Future works may be oriented on the comparison of these results with results obtained for different networks. The role prediction is also an interesting problem. If the prediction of SNA measures is correct, it gives us the possibility to predict future node roles.

Our future research on the dynamics of social networks is going to focus on the application of a multi-agent approach and to take advantage of the measures such as trust and reputation, characteristics for the analysis of societies in such systems. The significant difference between reputation models in multi-agent systems and the measures used in social network analysis is that agent measures has a local and individual character, each agent may have different values of measure evaluations of other agents, which depend on its local connections with them and the history of common interactions.

REFERENCES

1. Arquilla, J., Ronfeldt, D., Networks and Netwars : The Future of Terror, Crime, and Militancy (Consumer One-Off), RAND Corporation, 2002.
2. Carley K.M., Diesner, J., Reminga, J., Tsvetovat, M.,: Towards and interoperable dynamic network analysis toolkit. Decision Support Systems 43, 2007, pages 1324-1347.
3. Coffman T., Marcus S., "Dynamic Classification of Groups Using Social Network Analysis and HMMs," Proc. IEEE Aerospace Conf., March 6-13, 2004, Big Sky, MT.
4. Carrington P., Scott J., Wasserman S.: Models and Methods in Social Networks Analysis, Cambridge University Press, 2005.
5. Grigg, O.A., Farewell, V. T., Spiegelhalter, D. J. : The Use of Risk-Adjusted CUSUM and RSPRT Charts for Monitoring in Medical Contexts. Statistical Methods in Medical Research , vol. 12, number 2, 2003.
6. Hummon, N.P.: Utility and dynamic social networks. Social Networks, Volume 22, Number 3, July 2000 , pp. 221-249(29), Elsevier.
7. McCulloh, I., Carley, K.M. : Social Network Change Detection. Technical Report CMU-ISR-08-116, Institute for Software Research, Carnegie Mellon University, School of Computer Science, Center for the Computional Analysis of Social and Organizational Systems (CASOS), March 17, 2008.

8. Piekaj W., Skorek G., Zygmunt A., Koźlak J.: Środowisko do identyfikowania wzorców zachowań w oparciu o podejście sieci społecznych. Technologie Przetwarzania Danych: II Krajowa Konferencja Naukowa, Poznań, 2007.
9. Scott J.: Social Network Analysis: A Handbook. Sage Publication, 2009.
10. Shewhart, W. A.: Economic Control of Quality of Manufactured Product". New York: D. Van Nostrand Company", 1931.
11. Snijders, T.: The statistical evaluation of social network dynamics. In Sobel, M., Becker, M., eds.: Sociological Methodology dynamics. Basil Blackwell, Boston & London, 2001, pages 361-395.
12. Xu J., Marshall B., Kaza, S., Hsinchum Ch.: Analyzing and Visualizing Criminal Network Dynamics: A Case Study. in: Intelligence and Security Informatics, Springer, 2004.
13. Zygmunt A., Koźlak J.: Zastosowanie podejścia sieci społecznych do wspomagania prowadzenia analizy kryminalnej dotyczącej danych billingowych. Praktyczne elementy zwalczania przestępczości zorganizowanej i terroryzmu : nowoczesne technologie i praca operacyjna. red. Paprzycki L., Rau Z., Wolters Kluwer Polska. 2009.

Chapter 9

Theory of possibility applied in option's pricing

Arkadiusz Banasik
Politechnika Śląska

arkadiusz.banasik@polsl.pl

Abstract

This chapter presents approach and experiment involving experts' knowledge in investment databases. Presented approach indicates a way of use of AI in investment field of interest.

1. Introduction

It is a fact that application of Computational Intelligence is world-wide used. During last years the application areas expanded rapidly and the one of most developing application area is financial world [1].

The modeling and trading of financial is a great challenge because of great number of factors involved in financial processes. These factors are among others: interest rates, exchange rates, the rate of economic growth, liquidity [2]. It occurs that 8.5 billion of shares are being traded daily in the United States. That great number of shares is worth 120 billion dollars [2]. That value shows that this point of interest has a financial aspect indeed.

It is hard to estimate the influence of factors having impact on the prices of financial assets because the effects can be non-linear, time-lagged and non-stationary [2]. All important features of economy are based on information and knowledge. Obtaining and representation of knowledge is an important research area in AI [1]. A clue for investors is to gather information and aggregate it into suitable form. An important feature which helps people to gather data is use of database.

Experts' knowledge is widely used in field of economy and their clues are often expressed in natural language. That shows the possibility of using AI methods in knowledge aggregation.

2. Concept

Pieces of information gathered in database are concerned on indexes of stocks and values of stocks. Historical data are input of the decision support system. An input to the system are also expressions of experts' opinions in natural language. Natural language is a way of naming different situation, its feature is imprecision. A way of combining imprecise data into countable way is Fuzzy Logic [4]. A part of fuzzy logic concerned on possibility – Theory of possibility – is a way of transformation of those statements into countable version for computer [5-9].

Fuzzy set is an object which is characterized by its membership function. That function is assigned to every object in the set and it is ranging between zero and one. The membership (characteristic) function is the grade of membership of that object in the mentioned set [4].

That definition allows us to declare more adequate if the object is within a range of the set or not and to be more precise the degree of being in range. That is a useful feature for expressions in natural language, e.g. the price is around thirty dollars, etc. It is obvious that if we consider sets (not fuzzy sets) it is very hard to declare objects and their membership function.

It is necessary to indicate the meaning of membership function [4]. The first possibility is to indicate similarity between object and the standard.

Another is to indicate level of preferences. In that case the membership function is concerned as level of acceptance of an object in order to declared preferences.

And the last but not least possibility is to consider it as a level of uncertainty. In that case membership function is concerned as a level of validity that variable X will be equal to value x.

The third approach is concerned with Theory of possibility [3, 6-9]. This approach is an alternative towards theory of probability. It is assumed that theory of possibility is related to fuzzy sets [3]. It is managed by defining the possibility distribution as a fuzzy restriction which behaves as an elastic constraint on the values (possible to be assigned to a variable) [3]. This approach is definitely more useful in order to imprecise statement in natural language. It is hard to use the theory of probability in such a imprecise statement as e.g. the interest rate is very high. It is rather level of similarity than probability of high interest rate. That approach is an important postulate, which is the ground for analyzing natural language information and imprecise (fuzzy) statements.

Membership function is a base of gathering objects in fuzzy sets. It is also a base for further building of rules and contributions, etc. That is the main objective in fuzzy approach.

As we can see all methods are based on fuzzy approach and are predefined as suitable tools for operating with information and data [6-9]. That is the key point of interest in presented paper, which contains an approach towards application of mentioned tools in knowledge base for investor.

3. Experiment and implementation metod

The implementation of system is based on SQL language with fuzzy rules based on centroids. The defuzzyfication process is made towards forming of group of sets of stocks which gives specific values of income. Those expressions are implemented in PROLOG, which gives the possibility of creating database.

The statements expressing experts' knowledge, e.g.:

```
Table 1(A1, 25.00,15.00, around(3), 3, around(15))
Table 1(A2, unknown, unknown, 3.5, 3, 20)
```

or

```
Table 2(A1, 25.00, 15.00, 3.5, Low, High)
Table 2(A2, 30.00, 10.00, 2, Medium, Low)
```

That approach provides possibility of gathering information for better understanding Stock Exchange reality. Combination of experts' knowledge and historical data is a well developed system for better understanding of reality. That gives the opportunity to use experts' statements and their verification of risk levels in Black-Scholes Model (BSM).

The output of presented system gives information of forecast for stocks based on both experts' opinions and historical data. It is a part of information system for investor. For purpose of this article the implementation is based on ten different stocks from Giełda Papierów Wartościowych w Warszawie (Polish Stock Exchange). The historical data are gathered from researches of that scientific field.

The experiment shows the possibility of using natural language in values of data for BSM. To be precise to gather imprecise data and imprecise expressions. That approach combines historical data and experts' opinions.

Tab. 1. Experiment data[1]

Name	Asset Price [$]	Option Strike Price [$]	Maturity [years]	Risk-free interest rate [%]	Volatility [%]
A1	25,00	15,00	3,5	Low	High
A2	30,00	10,00	2	Medium	Low
A3	40,00	50,00	1	High	High
A4	100,00	20,00	5	Low	High

As you can see in Table 1 there are expressions in natural language instead of numerical values. That is the approach connected with use of natural language. That approach shows possibility of using natural language in input and output data for databases. Dealing with that kind of data is a key field for use of fuzzy logic and theory of possibility.

The result for investor is the description of stocks' environment, which should improve decision making process during investment process. The results of implemented rules are compared with real values from that period of time, which shows more adequate results than only statistical approach.

Gathering that kind of information is a way of transforming reality into countable form for investors, who are involving their money into that part of their activity.

4. Conclusions

Combining of information and money is commonly seen in today's economy. That combination is also effective and dangerous. Investing is a part of life of all people, but most important for people earning money on that kind of transactions. Information becomes as important as other production goods. That approach forces a way of obtaining, gathering, organizing and use of information. This aspect of peoples' behavior becomes an important scientific field of interest.

Researchers were using different methods of helping people with that problems. A division of them created mechanisms based on artificial intelligence tools to provide the best solutions. Fuzzy expressions and imprecise data are common now days.

Experts' opinions used in that field help people to gain better positions on markets and ability to compute their opinions is commonly needed. Application of AI in decision support systems is the only way and Theory of possibility is a key point in that field.

[1] Authors own

Databases using imprecise statements are a way of solving problems with uncountable data. The apparatus based on presented approach is a good option for investors, who are dealing with those kind of information everyday.

REFERENCES

1. Algos 3.0: Developments in Algorithmic Trading. Traders Magazine 2007. Special Report. SourceMedia's Custom Publishing Group.
2. Brabazon A., O'Neill M., Dempsey I.: An Introduction to Evolutionary Computation in Finance. IEEE Computational Intelligence, Vol. 3 No. 4 (November 2008), pp. 42-55.
3. Zadeh L.: Knowledge Representation in Fuzzy Logic. IEEE Transactions on Knowledge and Data Reasoning, Vol. 1 No. 1 (March 1989), pp 89-100.
4. Zadeh L.: Fuzzy Sets. Information and Control (8) (1965), pp. 338-353.
5. Zadeh L.: Fuzy Sets as a Basis for Theory of Possibility. Fuzzy Sets and Systems 1 (1978), pp. 3-28.
6. Łęski J.: Systemy Neuronowo-rozmyte. WNT. Warszawa, 2008. p. 689.
7. Chojcan J., Łęski J. [Eds.]: Zbiory Rozmyte i Ich Zastosowania. Wydawnictwo Politechniki Śląskiej. Gliwice, 2001. p. 479.
8. Dubois D., Prade H.: Fuzzy Sets in Approximate Reasoning, Part 1: Inference with Possibility Distributions. Fuzzy Sets and Systems 40 (1991), pp. 143-202.
9. Dubois D., Prade H.: The Three Semantics of Fuzzy Sets. Fuzzy Sets and Systems 90 (1997), pp. 141-150.

Chapter 10

Acquiring knowledge and advanced analysis procedures in supporting the process of examining the credibility of statements of means

Marek Valenta, Marcin Nowak, Marcin Okrzes
Department of Computer Science,
AGH University of Science and Technology in Cracow
valenta@agh.edu.pl, mar-nowak@wp.pl, marcin.okrzes@gmail.com

Abstract

Presented researches[1] *were carried out to identify possibilities of supporting procedures in the scope of realization of tasks concerning analyses of statements of means submitted by people who were obliged to do so by many Acts legally binding since 1997. Results of the researches showed potential and technical possibilities of increasing effectiveness of actions which result from fulfillment of the essence of these Acts. Authors indicate a method and some technical means possible to use in this field primarily based on data gathering systems (databases and data warehouses), integrated access to external data sources and a technology of data exploration for a substantial increase of the effectiveness of control actions resulting from Acts and government orders concerning statements of means.*

1. Introduction

For a long time it is observed a constant dynamic increase in using technologically advanced data processing systems and data infrastructure in realization of tasks of the state administration and local government

[1] Researches carried out under the project of The Ministry of Science and Higher Education: R & D project No. R00/001/02 and research project No. 0032/R/T00/2008/06.

administration. Computer science technologies are more common in helping to fulfill the majority of administration duties. The significant development of these systems helps not only to fulfill tasks of the public administration units, but also it enables to implement a new standard of quality of realization of those tasks. However, the observed increase in the areas of public administration is not equal, and a lot of units are being left without any support of information technology.

We have to deal with such a situation in the areas of accomplishing tasks which result from acts and regulations concerning statements of means. Problem has appeared with act on limiting the freedom of conducting business activity by people who perform public functions [10] which introduced a term of *the statements of means.*

Succeeding acts and regulations were definitely describing the obligation to make the statement of means by a wider range of occupations. They also defined the form of such statements, responsibilities of employees obliged to submit them, and what is most important - the way and scope of content analysis of submitted statements. The process of property statement submission and a scope of content analysis is fundamental to all acts regulating that obligation.

All of that activities aim at prevention and countering of corruption in the environment of public officers. The scope of anti-corruption acts refers to hundreds of thousands of citizens obliged to submit a statements of means. Therefore, people must be aware that at least[1] the same number of statements of means have to be annually analyzed paying special attention to occurrence of data that could signify a possibility of the existence of corruption situations. Problems arise here because the administration specialized in gathering data based on standard paper sheets is not prepared to carry out a purposeful data analysis of not only current statements of means but also of their previous versions.

It must be stressed that currently the process of the analysis of the statements of means made by supervisors is the weakest point of accomplishing tasks related to fulfilling duties resulting from acts and regulations concerning statements of means, and quite often these actions have nothing to do with the essence of anti-corruption regulations [2].

2. Assumptions and stages of the process of analysis

Natural assumption made by authors was to assume that data which constituted the content of statements of means must be available for their analysis in the form of attributes in database dedicated for these documents. In this part of the

[1] There are factors other than the annual mandatory obligation to make statements

research, the process of data gathering and low regulations of data gathering was not important. Similarly, to achieve the purpose of this research, vital legal conditions related to process of analysis [1, 3] were omitted.

In next part of this report the process of analysis will be related to statutory obligations, that were imposed on services which come under the authority of the Ministry of Internal Affairs and Administration, particularly Police [12] what allows a thorough analysis of the details of this problem. Proposed solution have, however, a universal nature and can be related to almost any other group of obliged people.

The scope of property statement content analysis was reduced only to such data that could signify unjustified enrichment or impoverishment of the person who submitted a statement. Notification of such a fact should legally result in the necessity of additional analysis of this statement of means, authorized by The Central Anticorruption Bureau[11].

While analyzing data of current and previous statements of means of an officer, his supervisor is legally obliged to uncover a possible event of unjustified enrichment or impoverishment of his subordinate. In practice, it is practically unfeasible because the scope of real analysis, according to authors, should in this case include:

- Checking the reliability of data included in this statement,
- Determining the value of assets listed in the current and previous statement,
- Determining an approximate amount of legally acquired income,
- Estimating the amount of necessary annual cost of living of people who are kept by the officer
- Determining a possible event of unjustified enrichment or impoverishment of the person submitting the statement.

Taking this steps without access to additional data sources and performing laborious decisions adjacent to operational activities, is almost impossible. In this situation, it seems to be a reasonable to entrust these activities to a computer system equipped with appropriate tools. Making a credible analysis by a computer system enables to get results without people's participation in the whole process.

The crucial idea of such system might be a use of large amounts of data collected to acquire a valuable knowledge. In the first case, the knowledge necessary to estimate amount of annual income and expenses for different groups of people.

In the more general case, this knowledge allows to define an a'priori classification and share them on groups that require and do not require analysis according to accepted analytical procedures. It may turn out that in the next phase of system implementation, the vast majority of statements of means would not have to be subjected to a complex process of analyzing the content of these statements at all.

3. Main aspects of analytic functions

On the grounds of the clarity of the analysis of the statements of means, the entire process was divided into many stages. Logical schema of this process of analyzing statements with the visualization of its main steps was illustrated in Figure 1.

The basic scope of data included in statements of means, which should be subjected to the process of analysis refer to basic data of amount of personal assets or joint assets of a person filling in the statement.

Basic, typical components of these assets in the light of currently binding regulations include:

- real-estates in the form of plots, houses and flats,
- possession of value, mostly motor vehicles, cars and motorcycles,
- financial assets in various currencies,
- securities,
- credit liabilities.

All of the elements in a statement of means are characterized in a descriptive manner by the selection of their features, however, without mentioning their current values (with the exception of the financial assets). Each analysis of statements made on the basis of data included in them, in the first place should concern Verification of the credibility of these data.

This credibility can only be verified by comparison of data included in the statement of means with data included in other reliable sources. Such reliable sources may be registers of bodies of public administration. In this context, the most important external data sources would be:

- in the field of real estates - Central Land and Mortgage Register
- in the field of possession of value - Central Register of Vehicles and Drivers
- in the field of credit liabilities– The Credit Information Bureau
- in the field of income verification – systems of databases of Tax Offices

Basing on these records, we are able to check the reliability and completeness of the data given in the statement. Case studies of enrichment or impoverishment of the person completing the declaration, must be based on estimates of income shown in the previous and the current statement. Therefore, the next stage of general analysis should be *Valuation* of assets presented in statements.

The process of Valuation of assets we need not only data included directly in the statement but also some characteristic features of these assets that can be obtained from external data sources during the stage of Verification.

This new set of data may constitute a starting point for the processes of assets valuation and determining the amount of legal incomes. Valuation of assets should be made based on reliable sources revised according to marketing trends. Services of such kind are provided by a universally accessible web portals such as : www.gratka.pl and www.bankier.pl .

Next stages of analysis should be able to compare gained incomes, financial statuses and an approximate cost of living declared in subsequent statements. Easiest to determine in the next stage of Advanced Analysis 1, are situations in which the identified incomes are equal or lower than the increase in the wealth of the person submitting a statement. Finding such values may raise some suspicion of corruption. Therefore, it requires additional analysis of the statements made by the authorized people and institutions.

The above described situation is, however, only one case in which the financial statement, the difference in possessed properties and the declared cost of living substantially diverges from zero. To make the system detect such situations it must have the data of the approximate amount of costs of living. We can try to determine them, but only for certain groups of people who submit statements. Determining potential approximate costs of living for particular groups is connected with acquiring this knowledge from appropriately juxtaposed data which would cover large population of people who represent the environment of people who submit statements.

Furthermore, these data must be data obtained from statements of means, but this time, extended with available to supervisors data gathered in personnel systems (next external source of data).

Generally, these additional data should define social, professional and family status of person obliged to make a statement. Getting knowledge of potential approximate costs of living in particular groups of people obliged to make statements of means allows in the next stage of *Advanced Analysis* 2 for assigning people to adequate groups and selecting these statements in which the financial statement is much greater that assumed threshold heuristically defined.

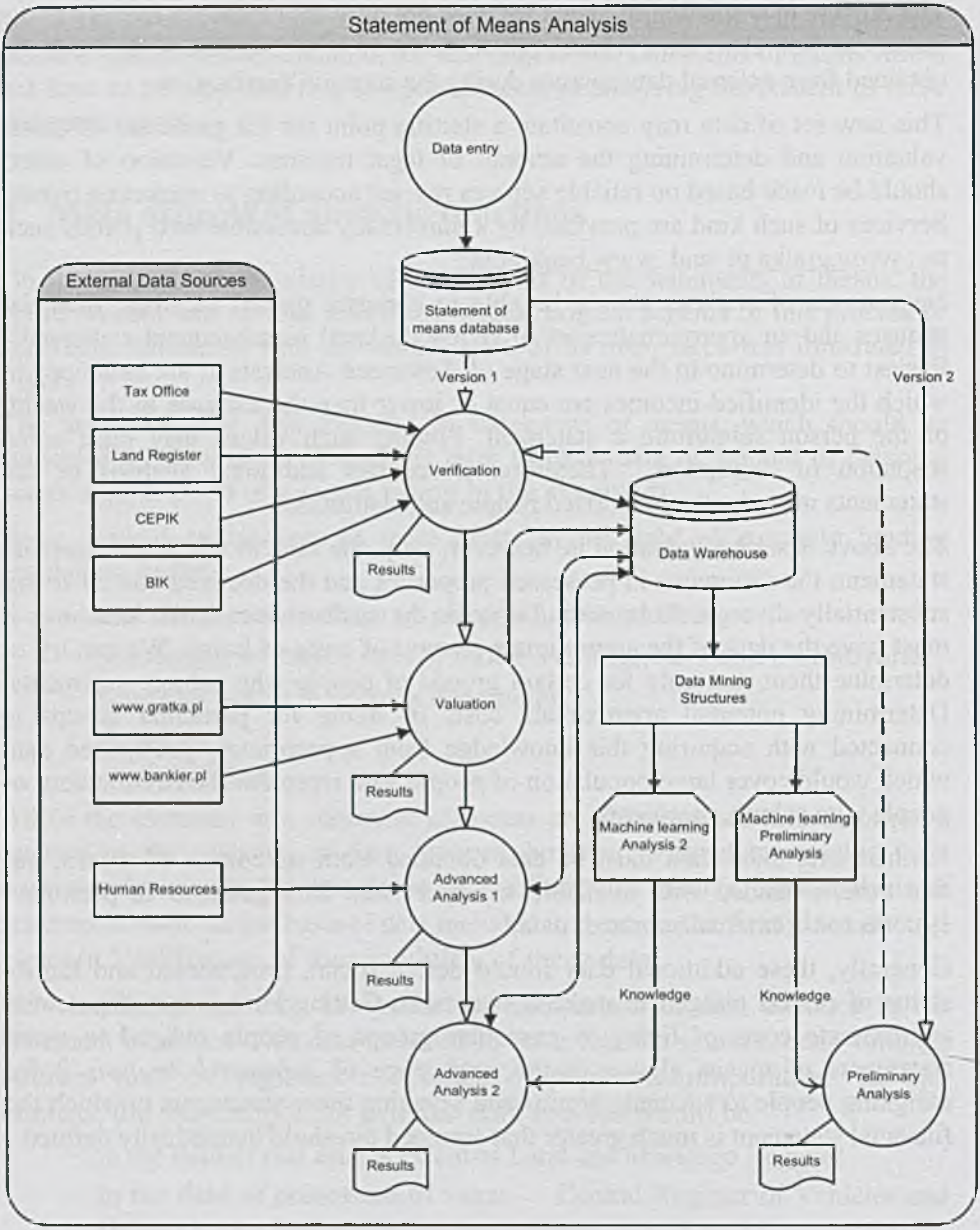


Fig. 6. Logical schema of the process of analyzing statements of means

Having stored a large quantities of data in the system related to statements of means, (and some experience in methods of acquiring knowledge from their verification) we may try to create one more analytic function. Its responsibility would be to qualify, only on the basis of the data included in the statement of means, with what degree of certainty, this statement does not require additional procedures of Verification, Valuation and Advanced Analysis 1 and 2. Such situation would only allow for Preliminary Analysis of all submitted statements, after which next stages of analysis would be made, but only for a selected, definitely smaller group of statements.

4. Acquisition and using knowledge in processes of analysis

Logical patterns of though described in chapter 3 was designed and implemented as a MaxiSOMP. (Maxi System of Statements of Means of Police). The implementation has a character of a prototype that demonstrates the possibility of realization concepts presented in the schema above. Apart from legal and technical aspects associated with obtaining access to the external data sources one of the most important elements which provide a new quality into the process of analysis is the automatic acquisition of knowledge and its use in the analysis of statements of means.

Data gathered by the system may be used by authorized users for the various types of analyses that are part of a process of the statutory statements of means' verification. This may be simplified by the appropriate data structure contained in the Data Warehouse. But despite this, the task of "manual" analyzing large number of statements of means is almost impossible to achieve. Therefore, this procedure should be supported by an automatic statements selection, which will limit the number of statements that need to be subjected to such analysis.

The Advanced Analysis in version 1 and 2 should serve that purpose, and after collecting a great amount of data also the Preliminary Analysis that would drastically narrow the number of statements designed for processing even in the processes of Verification and Valuation.

For accomplishment of task specified in such a way, data gathered in the system after previous preparation, become a learning sets for the procedures of acquiring knowledge during the process of machine learning.

5. The concept of technical implementation of complex analytic functions

Advanced analysis process and preliminary analysis process work independently on a subset of statements without any interaction with users. Processes of analyses take place on previously acquired data stored in a dedicated data warehouse. Algorithms of these operations are using knowledge, obtained earlier in the process of data mining. Due to the different nature of the tasks of advanced analyses and preliminary analysis and, processes are carried out by a different algorithms that require different types of machine learning.

Schema of realization of main processes of knowledge acquisition and analyses : preliminary advanced 1 and advanced 2 is presented in Figure 2.

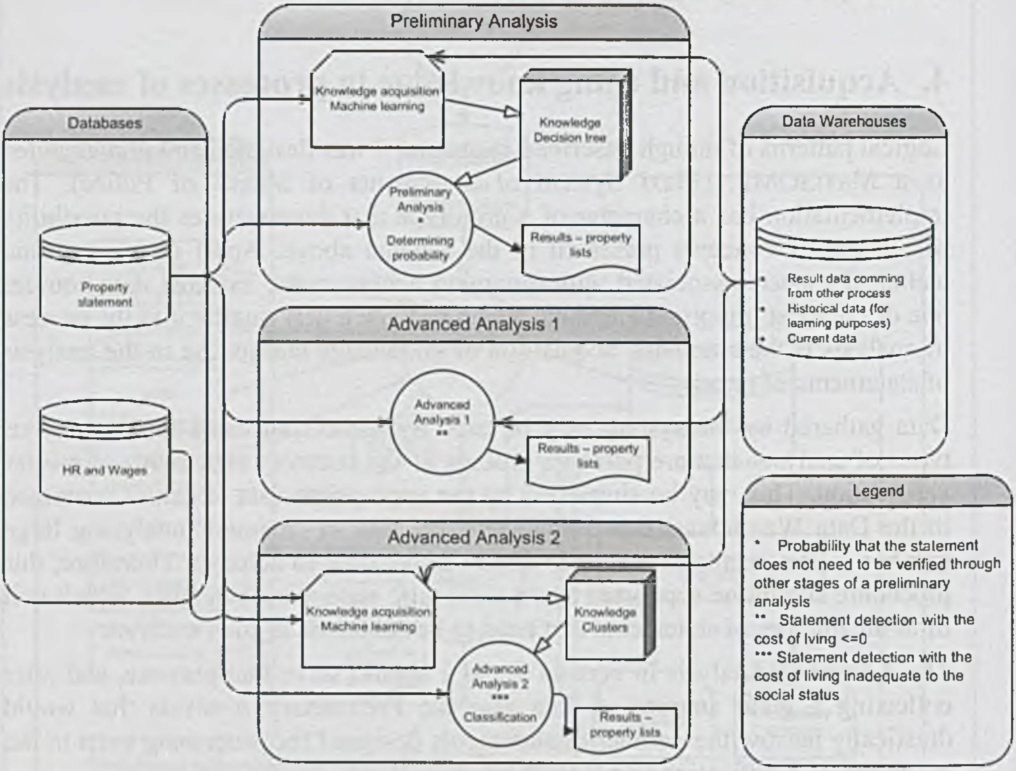

Preliminary Analysis
Databases
Knowledge acquisition Machine learning
Knowledge Decision tree
Preliminary Analysis * Determining probability
Results – property lists
Property statement
HR and Wages
Data Warehouses
Result data comming from other process
Historical data (for learning purposes)
Current data
Advanced Analysis 1
Advanced Analysis 1 **
Results – property lists
Advanced Analysis 2
Knowledge acquisition Machine learning
Knowledge Clusters
Advanced Analysis 2 *** Classification
Results – property lists
Legend
* Probability that the statement does not need to be verified through other stages of a preliminary analysis
** Statement detection with the cost of lving <=0
*** Statement detection with the cost of living inadequate to the social status

Fig. 7. Schema of realization of main processes of knowledge acquisition and analyses : preliminary advanced 1 and advanced 2

5.1. Preliminary Analysis

Preliminary Analysis is based on the Decision Trees algorithm, which for each statement of means specifies the probability of a random variable which determines a hypothetical result of the process of verification of a statement of means based on confrontation with external sources of data. The algorithm uses knowledge in the form of a decision tree. This tree is built in the process of machine learning. and uses a number of criteria such as characteristic features of particular elements of statements of means. Results of the algorithm which uses a decision tree, are also results of a preliminary analysis process have a form of a logical value: true and false with values of probability of their occurrence assigned to them.

5.1.1. Acquiring knowledge

In order to create a learning set of Preliminary Analysis data mining attributes of Data Mining algorithm are calculated for each statement. In addition, the statements included in the learning sets undergo a preliminary verification which checks their compatibility with external systems eg. (Tax Office, Central Land and Mortgage Register, Central Register of Vehicles and Drivers, The Credit Information Bureau). The results of verification process, are stored in a data warehouse, and constitute an integral part of the learning set of preliminary analysis. Thanks to availability of previous verifications' results, the process of acquiring knowledge for the preliminary analysis algorithm is possible, without the necessity to use expert knowledge which is a manual determining results of the algorithm for statements which are part of learning set. Here, it should be stressed that with sufficiently large learning set, knowledge acquired in such a way has fully objective character.

In the project for the process of acquiring knowledge for the Preliminary Analysis there were used data mining tools Microsoft Decision Trees [x] and learning data from ultimate sets coming from the earlier processes of verification of statements of means. During the process of learning the Microsoft Decision Trees algorithms build the structure of a decision tree. It consists of a set of nodes, edges, and elements that correspond to the hypothetical result of verification and probability of this result and a condition concerning attributes of algorithm (data from statements of means).

5.1.2. Analysis

Preliminary analysis in the project was based on the Microsoft Decision Trees algorithm. For each statement of means, subjected to the process of preliminary analysis firstly there are determined designated attributes of the algorithm. Then, based on the values of these attributes, for each statement, it is determined, based on the knowledge in the form of a decision tree, a sought value of probability. Analysis of each statement is made starting from the root node of the tree by the successive set of nodes determined by a values of previously determined attributes. Verification result as it was mentioned earlier are logical values true/ false with probabilities assigned to them. Depending on the settings of the system parameters, the final result of a preliminary analysis is generated in the form of statements lists that passed and which failed this stage of analysis. Statements, which have successfully passed this stage, no longer need to be subjected to other control functions such as: verification, valuation and subsequent advanced analyses.

5.2. Advanced Analysis 1 – expenses on living

The first part of the advanced analysis is to point out those statements which content is unreal. The range of unreality which is possible to find, is the existence of an information that cost of living declared by a person who made a statement was negative. This means that the increase in value of this person's property is equal, or greater than the declared income in that period. Such situation may raise the suspicion of the existence of undisclosed sources of income which may come from activities of a corruptive nature.

Knowledge of the system needed to implement the process of calculating amount of money annually spent on living comes from the statement of means itself and external data sources. 'Expenses on living' create a special kind of a measure. That measure is calculated as the difference between the concise incomes (from all types of activities, liabilities and sold properties), and all expenses (paid liabilities and obtained property). Processes that may assist in obtaining the necessary data are also processes of valuation of assets. The result of this phase of analysis is a list of statements, which should be subjected to additional, automatic analysis made by authorized institutions.

5.3. Advanced Analysis 2 - extended analysis of the cost of living

Advanced Analysis 2 is a stage which realizes the procedure of extended analysis of credibility of data which concern amounts spend on living in a period between submitted statements. It detects situations in which the calculated 'expenses on maintenance' are positive, but too small or too large when taking into account the characteristic features of the person. To determine the expenses on living recognized as standard for various groups of people who submit statements and obtaining characteristic features of these groups it is necessary to get adequate knowledge. If this knowledge comes directly from the data of the description of real situations then it is plausible. Acquiring this knowledge is possible in the machine learning which uses clustering method. Successively, the analysis that uses this knowledge, makes the classification of people, and compares the amount of the standard maintenance costs assigned to a specific group, with the cost of maintenance obtained at the stage of Advanced Analysis 1.

Authors suggest that the Microsoft Clustering algorithm is a preferred tool in this stage of analysis. It is a segmentation algorithm included in a packet of the Microsoft Analysis Services. The algorithm uses iterative techniques to determine and cluster data from the data set. Clustering algorithm teaches a model basing on data relations coming from a learning set, and creates clusters identified by the algorithm.

5.3.1. Acquiring knowledge

In order to develop a learning set of Advanced Analysis 2, for each statement the attributes of this data mining algorithm of this analysis are calculated. Thanks to the use of valuation and the learning process in a first part of the Advanced Analysis, it is possible to teach an algorithm in the second part of the Advanced Analysis, without the need to use expert knowledge and just set of learning data.

Microsoft Clustering Algorithm, during the teaching process of advanced analysis 2 creates a statement clusters, which divide them into groups according to the criteria. Each cluster is related to the range of expenses on living, which is typical for each group. It is determined on the basis of calculated for the learning set the real value of expenses on living.

5.3.2. Analysis

Advanced Analysis 2 is based on the Microsoft Clustering Data Mining algorithm, which aim is to set the expected value of "expenses on living" for a given statement based on personal data of the person submitting it. Based on the values of these attributes, each statement is classified and assigned to one of the previously formed clusters. That cluster has a fixed at the stage of knowledge acquisition value. The expected value measures "expenses on living". That measure is compared with the real value of the 'expenses on living' being calculated in the first part of the advanced analysis. For each statement, a relative deviation based on the real value spent on living is calculated and is compared with expected value obtained during the classification process. Depending on system settings defined heuristically, finally the process of the qualifies each statement to adequate ultimate subsets containing: respectively: statements of means which from the point of view of the algorithm of the Advanced Analysis 2 do not raise objections and those which must be subjected to additional analysis, not automatic one, but analysis made by people and institutions authorized to do so.

6. Summary

Realization of any system should include carrying out a number of tests. Unfortunately, the authors of this research were unable to carry out tests that would check the correctness of proposed solutions primarily in the field of realization of procedures of knowledge acquisition. This is related to the fact of carrying out tests of the system on the basis on the data publicly available (in the scope of data covered by statements of means) and data automatically generated. In that situation, tests of meritorical correctness of the machine learning action were practically without a chance of success.

As it was mentioned, the prototype of the system [6] was to play a role of a demonstration system, designed for a small group of decision makers connected with the development of computer science and implementation of its achievements in public administration, taking especially into account the probl4ms of the statements of means. And it seems that such a role was fulfilled. The system, the scope of its functionality and methods of achieving its implementation aroused not only a great interest in this subject [3,4] but also made the decision makers aware of potential possibilities of technical solutions available in this field which enable realization of tasks of this kind. Naturally, the most urgent debates concern social and legal conditions for its implementation. However, almost all interested people in this subject agree with the fact that data from the statements of means should be available for their analysis in digital form. And this is a prerequisite for taking appropriate legislative initiatives allowing for the existence of statement of means forms in the form of e-documents which could be corrected online by people who are obliged to fill them in and submit. Such trends can be noted in important for these problems document, which is the strategy of information society development in Poland [8].

REFERENCES

1. Fajgielski P.: Information in public administration – legal aspects of collecting, sharing and protecting (in Polish). PRESSCOM, Wroclaw, 2007.
2. Information on the results of control of statements of means submission (in Polish) – evidence number: 2/2006/P/05/004/P/05/028/KAP/KBF. NIK, Warsaw, 2006, http://bip.nik.gov.pl/pl/bip/wyniki_kontroli_wstep/inform2006/2006002.
3. Marcjan R., Valenta M.: Necessary legal changes in the aspect of computer systems implementation within the scope of submitting statements of means in services subordinated to Ministry of Internal Affairs and Administration (in Polish), Ministry of Internal Affairs and Administration, Warsaw, 2009.
4. Marcjan R., Valenta M.: Police statements of means – not only (in Polish), MIAA - KGP, Warsaw, 21.10.2008.
5. Mundy J., Thornthwaite W., Kimball R.: The Microsoft Data Warehouse Toolkit: With SQL Server 2005 and the Microsoft Business Intelligence Toolset. John Wiley & Sons, 2006.
6. Nowak M., Okrzes M.: Knowledge management in decision support systems (in Polish). Thesis under the supervision of M. Valenta, Department of Computer Science AGH, The Faculty of Electrical Engineering, Automatics, Computer Science and Electronics, Cracow 2008.
7. Directive of the Ministry of Internal Affairs and Administration of 7 December 2001 on particular rules and procedures for giving police officers the right to have a gainful employment, submit statement of means and duties of their supervisors (in Polish). Journal of Laws, 2001, No. 148 item 1659.

8. Strategy for the development of information society in Poland until 2013 (in Polish). Ministry of Internal Affairs and Administration – http://www.mswia.gov.pl/strategia, Warsaw, 2008.
9. Tang Z., MacLennan J.: Data Mining with SQL Server 2005. John Wiley & Sons, 2005.
10. The Act of 21 August 1997 on Restrictions for State Officials on Business Activities (in Polish). Journal of Laws, 1997, No. 106 item 679.
11. The Central Anticorruption Bureau Bill of 9 June 2006 (in Polish). Journal of Laws, 2006, No. 104 item 708.
12. Police Act of 6 April 1990 (in Polish), Journal of Laws 1990, No. 30 item 179.

PTI

Polskie Towarzystwo Informatyczne
Oddział Górnośląski

ul. J. Lompy 2/10, 40-040 Katowice
www.pti.katowice.pl
Katowice@pti.org.pl

ISBN: 978-83-60810-29-3