Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science

Master`s Thesis

Probabilistic calculation of incomplete social network user data.

Nadiya Yangirova

Supervisor: **Ing. Karel Frajták, Ph.D.**

Study Program: Open Informatics
Field of Study: Software Engineering
May 24, 2019

MASTER'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Yangirova Nadiya** Personal ID number: **484746**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Computer Science and Engineering**

Study program: **Open Informatics**

Specialisation: **Software Engineering**

II. Master's thesis details

Master's thesis title in English:

Probabilistic calculation of incomplete social network user data

Master's thesis title in Czech:

Probabilistic calculation of incomplete social network user data

Guidelines:

Prediction of missing data in the user profiles of the social network "Vkontakte".
This includes collecting the necessary data and predicting the data using the Group Method of Data Handling.
Construction of a statistical model and testing this model on the real data.

Bibliography / sources:

[1] Гусарова Н.Ф. Интеллектуальные системы в управлении социальными процессами. – СПб: Университет ИТМО, 2015. 90 с.
[2] Newman M. Networks: An Introduction. Oxford University Press, 2010.
[3] Newman M., Girvan M.. Finding and evaluating community structure in networks // Phys. Rev. E 69, 026113, 2004.
[4] Easley D., Kleinberg J.. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010. – Электронный ресурс. Режим доступа: http://www.cs.cornell.edu/home/kleinber/networks-book/
[5] Метод группового учета. - А. Г. Ивахненко

Name and workplace of master's thesis supervisor:

Ing. Karel Frajták, Ph.D., Software Testing Intelligent Lab, FEE

Name and workplace of second master's thesis supervisor or consultant:

Date of master's thesis assignment: **05.04.2019** Deadline for master's thesis submission: **24.05.2019**

Assignment valid until: **19.02.2021**

Ing. Karel Frajták, Ph.D. Supervisor's signature	Head of department's signature	prof. Ing. Pavel Ripka, CSc. Dean's signature

III. Assignment receipt

The student acknowledges that the master's thesis is an individual work. The student must produce her thesis without the assistance of others, with the exception of provided consultations. Within the master's thesis, the author must state the names of consultants and include a list of references.

.	
Date of assignment receipt	Student's signature

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Ing. Karel Frajták (CVUT), Ing. Zulfira Enikeeva (KFU) and and Galim Vakhitov (KFU) for guidance and huge help on my thesis. I also want to thank my family, friends and colleagues for their support. A big thanks to Turilova Ekaterina, Enikeev Arslan, Miroslav Bureš and all Czech team for the opportunity to participate in the Double Degree program (CVUT, KFU).

Declaration

I hereby declare that I have completed this thesis independently and that I have listed all the literature and publications used.

I have no objection to usage of this work in compliance with the act §60 Zákon č. 121/2000Sb. (copyright law), and with the rights connected with the copyright act including the changes in the act.

Kazan, May, 2019

_____

Abstract

YANGIROVA, Nadiya: Probabilistic calculation of incomplete social network user data. [Master's Thesis] - Czech Technical University in Prague. Faculty of Electrical Engineering, Department of Computer Science. Supervisor: Ing. Karel Frajták, Ph.D.

Social networks are a unique source of data about private lives and interests of real people. This opens up unprecedented opportunities for solving research and business problems (many of which could not be solved effectively before due to the lack of data). In addition, this causes an increased interest in the collection and analysis of social data from companies and research centers.

In thesis, we consider the problem of predicting hidden data using Group Method of Data Handling, GMDH. To solve this problem, we find a model of the relationship of the studied data taken from the profiles of the social network VKontakte.

The Keywords are: social network, community detection, GMDH, social network analysis.

Contents

List of figures

List of tables

Introduction

Social data analysis is rapidly gaining popularity all over the world due to the importance of online social networking services in the 1990s (SixDegrees, LiveJournal, Facebook, Twitter, YouTube, and others). [1, 2] The phenomenon of personal data socialization connected with biography facts, correspondence, diaries, photo, video, audio materials, travel notes, etc. have become publicly available. Thus, social networks are a unique source of data about the personal life and interests of real people. This opens up unprecedented opportunities for solving research and business problems (many of which could not be solved effectively before due to lack of data), as well as the creation of 440 support services and applications for social network users. In addition, this causes an increased interest in the collection and analysis of social data from companies and research centers.

In 2012, the analytical agency Gartner published a report entitled “The Cycle of Rush for Developing Technologies”. According to the report, Social Analytics and Big Data technologies are currently on the so-called "the peak of high expectations". In particular, Carnegie Mellon, Stanford, Oxford, INRIA, Facebook, Google, Yahoo !, LinkedIn and many others are actively involved in social data research. Companies that own online social networking services (Facebook, Twitter) are actively investing in the development of advanced infrastructure (Cassandra, Presto, FlockDB, Thrift) and algorithmic (new search algorithms and recommendations of users, products and services) solutions for processing large amounts of user data. Commercial companies are emerging and successfully developing, providing services for accessing social data warehouses (GNIP), collecting social data according to specified scenarios (80legs), social analytics (DataSift), and expanding existing platforms using social data (FlipTop).

Thus, specialists from research centers and companies around the world use social networking data to model social, economic, political and other processes from the personal to the state level in order to develop mechanisms for influencing these processes, as well as creating innovative analytical and business applications and services.

However, when working with social data, factors such as instability in the quality of user content (spam and false accounts), problems with ensuring the privacy of users' personal data during storage and processing, as well as frequent updates of the user model and functionality should be taken into consideration. All this requires continuous improvement of algorithms for solving various analytical and business problems.

The processing of social data also requires the development of appropriate algorithmic and infrastructural solutions that take into account their dimensionality. For example, the Facebook social network database today contains more than 1 billion user accounts and more than 100 billion connections between them. Every day, users add over 200 million photos and leave more than 2 billion comments on various network objects. To this date, most of the existing algorithms that effectively solve actual problems are not able to process data of a similar dimension in a reasonable time. In this regard, there is a need for new solutions that allow for the distributed processing and storage of data without significant loss of quality results.

Social media web interfaces are real-time data sources and are designed to view and interact with social networking pages in a web browser or to use user data in specialized applications. Since the scenarios for using social network interfaces do not provide for automatic collection of data from several users to build a social graph, a number of problems arise. Here are the problems:

1. Data confidentiality
2. Poorly structured data
3. Access to restrictions and blocking

The purpose of this work is to obtain a probabilistic model for the study of incomplete social network data.

Tasks set:

1. Program the group method of data handling for analyzing social network data.

2. Get a model of data dependency social network

3. To obtain a probabilistic assessment of the quality of a probabilistic model

In my study I describe a possible mechanism for analyzing user data from social networks. In Chapter 2 I describe the processing and collection of real user data by accessing the social service web interfaces. The model of interaction of studied data is revealed. Chapter 3 analyzes the results obtained.

Chapter 1

The arguments for using GMDH in social network analysis tasks.

1. The specificity of social network data. Methods of statistical processing.

Vkontakte (international name: VK) is a social network, one of the most popular in the Russian-speaking segment of the Internet. In Russia, this network is ahead of Odnoklassniki, not to mention Facebook, and rightfully occupies the first place.[7,10]

As in other social networks, the principle of action is based on the creation by users their personal pages - the so-called profiles, and the exchange of various information, both textual, in the form of small messages, and graphical - the exchange of pictures and photos. Also, users of the network can create various communities and groups whose users are united by various interests, and organize events and meetings.

The Vkontakte network is convenient because it allows you to pass any information to a huge number of users very quickly. It is the speed of information exchange - one of the main reasons why even large companies create their profiles and groups in the network.

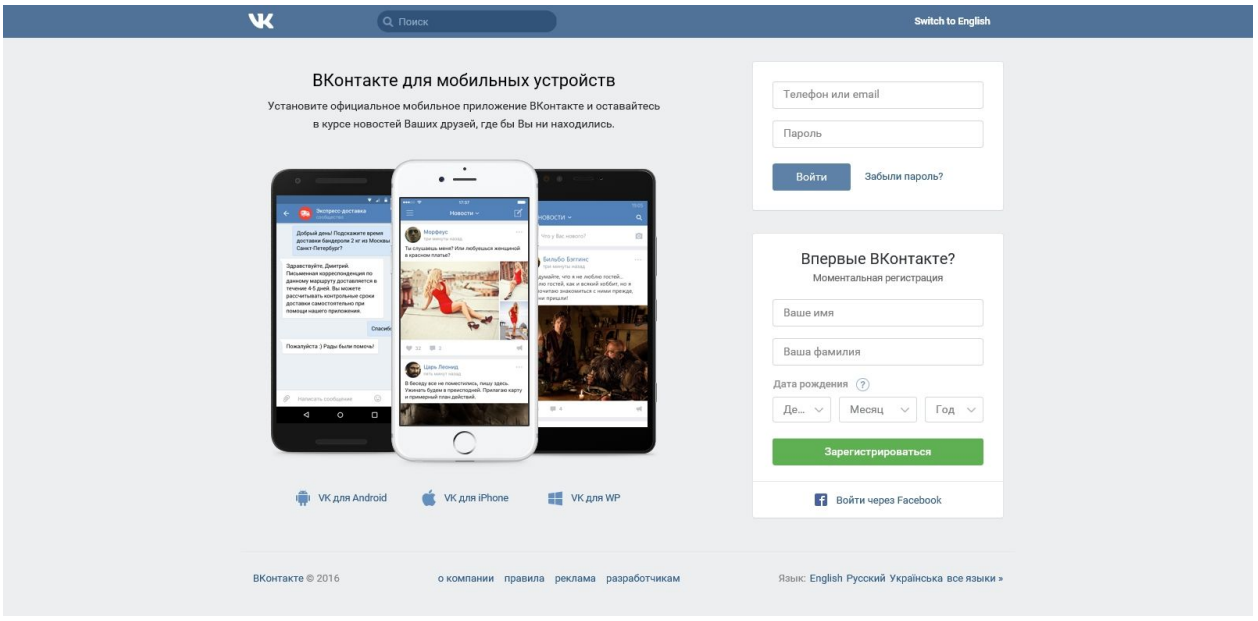

Figure 1.1 - VK Profile

The easiest way to collect data is to use the services of specialized companies that collect and constantly update data from various sources. The main advantage here is the speed of obtaining information, which is significant with large volumes of client database and the use of various social networks. The disadvantage is a paid subscription to update data.

The next way is to use software interfaces provided by almost all popular social networks. For different networks, APIs differ in the set of available data, restrictions on the number of requests and the cost to access the interfaces. For example, if you use the VKontakte network software interface, you can get complete information about the user, Facebook provides an API that returns almost “zero” information about the user.[4] The disadvantages of this method include a limit on the number of simultaneous requests and on the number of hits that an application can make per unit of time. In addition, you need to constantly monitor changes in the API and update the data collection application, and some social networks provide important data only for a specific fee. The advantages of the method are the possibility of obtaining data about a single client in a structured form (JSON or XML), as well as the ease of integrating API calls into your own application.

Another way is the manual analysis of social networking web pages, as well as the use of ready-made crawlers for data collection with subsequent analysis. In this case, there is access to all open data and there are no restrictions on how fast you want to collect the data. The disadvantages include the complexity of implementation - the web page of each social network is unique, so each time you have to develop your own passing rules. The disadvantages are also the complexity of support and the need for large computational resources. However, this process is well parallelized.

It is necessary to take into account that the corresponding characteristics in social networks are only reliable to a certain extent - they may be absent, be deliberately false, or allow different spellings. Therefore it is necessary to clean and normalize the data before processing. It's important to check the correctness of the parameters specified in the profile - for example, the city of the user can be refined based on the analysis of his subscriptions, posts, and statuses.

Some parameters can be restored by analyzing the profile of the user or his friends. For example, women very often do not indicate the year of birth, but there is a year of graduation from a university or school.

In addition to the data that clearly indicated in their profiles, a lot can be learned by analyzing posts, subscription groups, and photos. At the same time, additional facts are of interest that can be extracted from this unstructured information. For example, if most of the posts on the wall are about impressions of films, then it is clear that the user is interested in cinematography.

Automatic text analysis is impossible without linguistic technologies. In addition, statistical methods, machine learning technologies and deep-depth data analysis are also useful for solving many problems. Statistical studies and work with natural language are usually associated with some inaccuracy - in statistics we always talk about certain assumptions, heuristic assumptions that are not always fully fulfilled, and in natural language

there is always the probability of ambiguous interpretation of statements and conclusions. The correct combination of linguistic and statistical approaches improves the quality of the result and the level of its reliability. To illustrate the possible correlation of different methods for textual data enrichment, we consider several examples.

Imagine that we need to find out if the user is interested in football. Determine how often the corresponding terms are found in the texts on his wall, and upon reaching a certain level of their appearance, certain conclusions can be drawn. For such an enrichment method, it is necessary to know the terminology that can be obtained from dictionaries or thesauri in a specific subject area. In addition, you also need to be able to correctly calculate the number of uses - to understand the various forms of the same word. Thus, for this example, only linguistic means are sufficient.

The second example refers to the case when, in addition to linguistic processing, machine learning methods are needed. Let’s suppose that a user does not have a complete date of birth and is required to determine the age group on the basis of the texts that he writes. First of all, a set of texts of users whose age is known is formed. Then, for this set, using machine learning algorithms, the text features for each age group are identified and some formal model is formed that allows an arbitrary text to estimate the age of its author. Machine learning algorithms are usually designed for structured data, so before using them, the texts are replaced with sets of words found in them or with a set of topics that characterize these texts. For this purpose, linguistic algorithms are used to highlight meaningful words, normalize them, compile a lexical profile of a text, define topics, etc.

Since the scenarios for using social network interfaces do not involve the automatic collection of data from multiple users, a number of problems arise:

1. data privacy - often access to user data is allowed only for registered and authorized network participants, which requires support for user session emulation using special accounts (accounts);

2. poorly structured data - in many cases, social networking software interfaces (APIs) have limited functionality, which requires support to receive static copies of HTML pages using the web user interface, correctly processing their dynamic part (including making asynchronous requests to the social network server), extract the necessary data using an algorithm and / or a template and construct their structured presentation, convenient for further automatic processing;

3. access restrictions and blocking - in order to prevent unauthorized automatic data collection and limit the load on the social network service infrastructure, service owners often impose explicit or hidden restrictions on the allowed number of requests from one user account and / or IP address per unit of time, which requires accounting of the number of sent requests, as well as support for dynamic rotation of user accounts and IP addresses used to collect data;

4. The data dimension necessitates a parallel method of data collection, as well as methods for obtaining a representative sample of social network users (sampling).

To download publications from the VKontakte communities, a software module in the Java programming language is implemented.

To gain access to information about communities and their publications, we used API VKontakte technology, which provides methods for working with social network data. The number of calls to API methods has a limit: no more than 3 times per second.

API (Application programming interface) is a set of ready-made classes, functions, and structures provided by the service for use in external software products.

Objects	**Usage Example**
User	users.get
Community	groups.get
Wall Post	wall.get

Wall Comment	wall.getComments
Private Message	messages.get
Chat	messages.getChat
Note	notes.get
Wiki Page	pages.get
Market Item	market.get
Market Collection	market.getAlbums
Topic	board.getTopics
Topic Comment	board.getComments
Application	apps.get
Poll	polls.getById

Table 1.1 - basic objects

Object	Usage Example
Photo	photos.get
Audio	audio.get
Video	video.get
Document	docs.get
Wall Attachments	wall.get

Message Attachments	messages.get
Attached Link	wall.get
Sticker	messages.get
Gift	messages.get

Table 1.2 - Media and Attachments

Object	**Usage Example**
Photo Sizes	photos.get
Audio Genres	audio.getById
Post Source	wall.getById
Privacy	photos.getAlbums
Push Notifications Settings	account.getPushSettings

Table 1.3 - Additional Objects and Values Sets

To download data, the software module sends requests to VKontakte API methods to perform the following tasks:

1. Getting information about users using the *users.get* API method;

2. Getting a list of popular communities for each user in question using the *groups.getCatalog* API method;

2. Overview of methods for working with incomplete data

There is often gaps in the data that you need to: to ignore, discard, or fill in the missing values. It seems important to fill in gaps. However, it's not necessary. Unsuccessful selection of the gap filling method may not only improve, but worsen the results.[11]

Excluding and ignoring lines with missing values was the default solution in some popular application packages, with the result that novice analysts may leave the impression that this solution is the right one. In addition, there are fairly simple to implement and use skip processing methods, called "ad-hoc methods": filling gaps with zeros, medians, arithmetic mean values, introducing indicator variables, and the simplicity of which may be the reason for choosing these methods.

The following 3 mechanisms for the formation of passes are distinguished: MCAR, MAR, MNAR.

MCAR (Missing Completely At Random) - the mechanism for the formation of passes, in which the probability of a pass for each record set is the same. For example, if a sociological survey was conducted in which every tenth respondent was not asked one randomly selected question, and the respondents answered all other asked questions, then the MCAR mechanism takes place. In this case, ignoring/excluding records containing missing data does not lead to distorted results.

MAR (Missing At Random) - in practice, the data are usually not passed by chance, but because of certain patterns. Gaps are referred to as MAR if the probability of a gap can be determined on the basis of other information available in the data set (gender, age, the position held, education ...) that do not contain gaps. In this case, deleting or replacing the gaps to the "Skip" value, as in the case of MCAR, will not lead to a significant distortion of the results.

MNAR (Missing Not At Random) - a mechanism for the formation of passes, in which data is not available depending on unknown factors. MNAR assumes that the probability of missing a score could be described based on other attributes, but there is no information on these attributes in the data set. As a result, the probability of missing a pass cannot be expressed based on the information contained in the data set.

Consider simple gap handling methods and related problems.

1. Remove or ignore gaps

Complete-case Analysis (also known as the Listwise Deletion Method) is a gap treatment method used in many application packages as the default method. It consists of excluding from the data set records/lines or attributes/columns containing gaps.

In the case of the first skip mechanism (MCAR), the use of this method will not lead to a significant distortion of the model parameters. However, deleting lines leads to the fact that in further calculations not all available information is used, standard deviations increase, the results become less representative. In cases where there are many gaps in the data, this becomes a tangible problem.

In addition, in the case of the second (MAR) and, especially, the third skip mechanism (MNAR), the shift of the statistical properties of the sample, the values of the parameters of the constructed models and the increase in standard deviations become even stronger.

Thus, despite the wide distribution, the use of this method for solving practical problems is limited.

Available-case analysis (also known as Pairwise Deletion) is processing methods based on ignoring gaps in calculations. These methods, like Complete-case Analysis, are also often used by default.

Statistical characteristics, such as mean values, standard deviations, can be calculated using all non-missing values for each of the attributes/columns. As in the case of

Complete-case Analysis, provided that the MCAR hypothesis is fulfilled, the application of this method will not lead to a significant distortion of the model parameters.

The advantage of this approach is that when building a model, all available information is used.

The main disadvantage of these methods is that they are applicable for the calculation of far from all indicators and, as a rule, are associated with algorithmic and computational difficulties, leading to incorrect results.

The disadvantages of the first two gap treatment methods (Complete-case Analysis and Available-case analysis) also include the fact that, by no means, the exclusion of strings is in principle acceptable. Often, post-processing procedures assume that all rows and columns are involved in calculations (for example, when there are not too many gaps in each column, but there are few rows in which there is not a single missing field).

Later in this article, we will look into methods that involve filling in gaps based on available information. Often these methods are combined into one group, called Single-imputation methods.

2. Filling the gap with an average value

Filling the gap with the mean value (Mean Substitution) (other options: filling with zero, median and etc.) - the name of the method speaks for itself.

All variants of this method have the same drawbacks.

Consider these shortcomings on the example of one of the easiest ways to fill in the gaps of continuous characteristics: filling gaps with the arithmetic mean value and mode.

Example 1. Figure 2 shows the distribution of the values of the continuous characteristic before the gap is filled with the average value and after it.

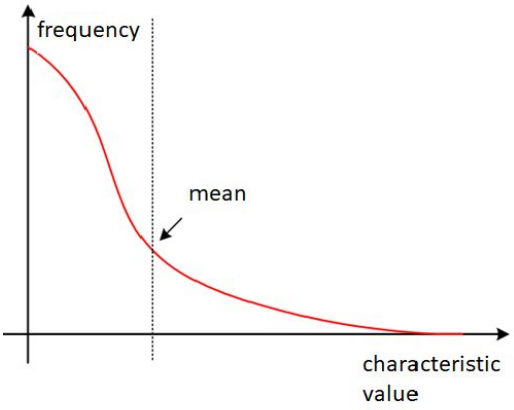

Figure 1.2a - The distribution of the values of the continuous characteristics to fill gaps

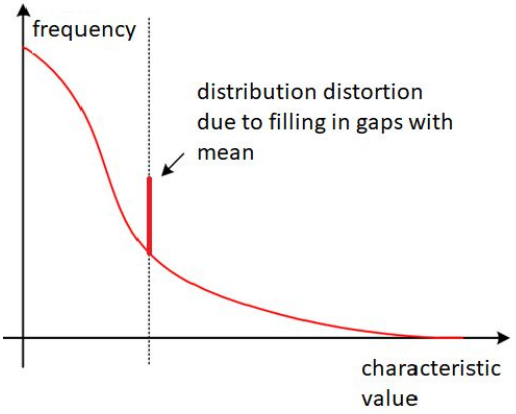

Figure 1.2b - The distribution of the values of the continuous characteristics after filling gaps

In Figure 2, it is clearly seen that the distribution after filling in the gaps looks extremely unnatural. This ultimately manifests itself in the distortion of all indicators characterizing the properties of the distribution (except for the average value), the underestimated correlation and the overestimation of standard deviations.

Thus, this method leads to a significant distortion of the distribution of the characteristic, even in the case of MCAR.

Example 2. In the case of a categorical/discrete characteristic, mode filling is most often used.

Figure 3 shows the distribution of the categorical characteristic before and after filling in the gaps.

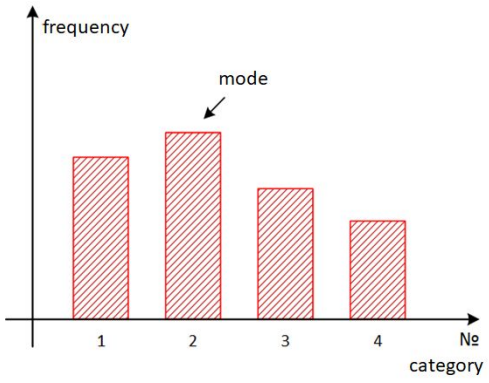

Figure 1.3a - The distribution of the discrete characteristics before filling in gaps with mode

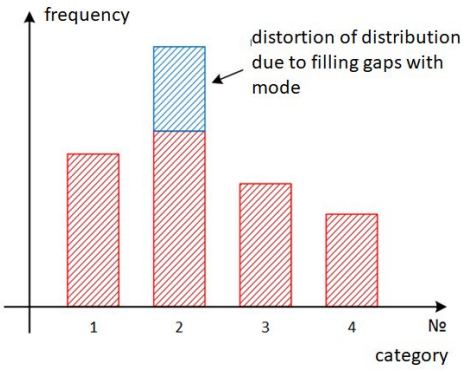

Figure 1.3b - Discrete characteristic distribution after filling in gaps with a mode

Thus, when filling gaps with categorical characteristics, the same disadvantages arise as to when filling gaps with continuous characteristics with the arithmetic average (zero, average, etc.).

3. Repetition of the result of the last observation

LOCF (last observation postponed) - a repetition of the result of the last observation.

This method is applied, as a rule, when filling in gaps in time series, when the subsequent values are closely related to the previous ones.

However, the method can also lead to significant distortions of the statistical properties even in the case of MCAR. So, it is possible that the use of LOCF will lead to duplication of the emission (filling the gaps with an anomalous value). In addition, if there are a lot of consistently missing values in the data, then the hypothesis of small changes is no longer fulfilled and, as a result, the use of LOCF leads to incorrect results.

4. Indicator Method

Indicator Method is a method that assumes the replacement of missing values with zeros and the addition of a special attribute that takes zero values for records where the data did not initially contain gaps and non-zero values where there were gaps previously.

Also, when filling gaps with nonzero values, the interaction of the flag-field and the source field is often added.

The advantages of this method include:

1) use of the entire data set (representativeness of the sample does not suffer),

2) explicit use of information about missing values.

Despite these advantages, even with the MCAR hypothesis and a small number of missing values, this method can lead to a significant distortion of the results.

5. Recovery of gaps based on regression models

This method is that all values are populated using linear regression models based on known data set values.

Figure 4 shows examples of the results of missing values of characteristics 1 based on known characteristics 2.

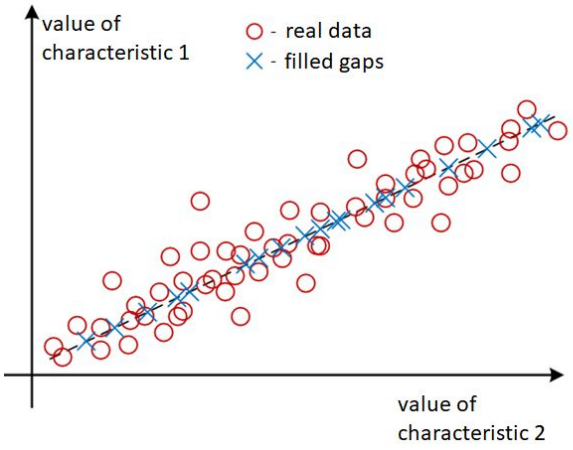

Figure 1.4 - Filling gaps based on linear regression

The linear regression method allows you to get plausibly filled data. However, there is some variation in the real data, which is absent when filling in gaps based on linear regression. As a result, the variation in the characteristic values becomes smaller, and the correlation between characteristic 2 and characteristic 1 is artificially enhanced. As a result, this method of filling gaps becomes worse, the higher the variation of the characteristic values, gaps in which we fill, and the higher the percentage of missing lines.

It is worth noting that there is a method that solves this problem: the stochastic linear regression method, illustrated in Figure 5 (similar to Figure 4).

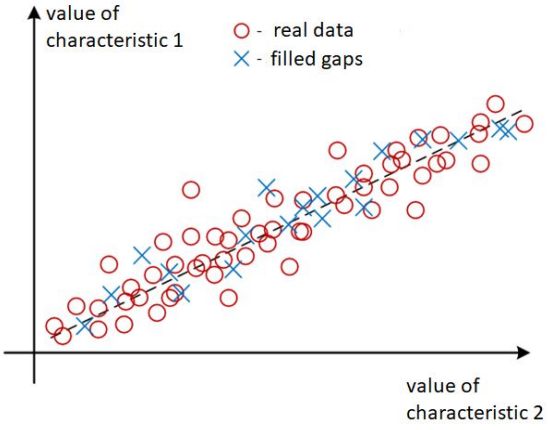

Figure 5 - filling in gaps based on stochastic linear regression

The stochastic linear regression model reflects not only a linear relationship between characteristics but also deviations from this linear relationship. This method has a positive gap filling properties based on linear regression and, moreover, does not distort the correlation coefficients so much.

Of all the methods that we considered in this part of the article, filling in the gaps using stochastic linear regression in general leads to the smallest distortion of the statistical properties of the sample. However, for more correct data recovery, more subtle methods are needed. Therefore, we use GMDH.

3. Mathematical foundations of the Group Method of Data Handling.

The author of the group method of data handling (GMDH) is A. G. Ivakhnenko. GMDH is used in various fields using structural, parametric identification and forecasting. The method is based on the recursive selective selection of models on the basis of which more complex models are built.[3,8]

The Group Method of Data Handling consists of several algorithms for solving various problems. It includes both parametric algorithms and non-parametric algorithms for clustering, combining analogs, rebnarization and probabilistic algorithms. This approach of self-organization is based on the selection of gradually complicating models and the choice of the best solution according to the minimum of the external criterion. Not only polynomials are used as basic models, but also nonlinear, probabilistic functions or clustering.

The latest developments of GMDH led to the creation of expert systems based on normative forecasting systems (under the if-then scenario) and optimization of control using simplified linear programming algorithms and neural networks with active neurons. In such neural networks, separate modeling algorithms are used as neurons in a multi-row neural

network. This makes it possible to improve the accuracy of the prediction, approximation or pattern recognition above the boundaries that are achieved by conventional neural networks with simple neurons or conventional statistical methods.

Most GMDH algorithms use a polynomial basic function. The general connection between input and output variables can be expressed in the form of a Volterra functional series, the discrete analog of which is the Kolmogorov-Gabor polynomial:

$$Y = a_0 + \sum_{i=1}^{k} a_i X_i + \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} X_i X_j + \sum_{i=1}^{k} \sum_{j=1}^{k} \sum_{l=1}^{k} a_{ijl} X_i X_i X_l, \quad (1.1)$$

where $X(x_1, x_2, ..., x_k)$ - input vector of variables;

$A(a_0, a_1, ..., a_k)$ - vector of coefficients or weights.

The components of the input vector X can be independent variables, functional forms or finite difference terms. Other non-linear basis functions, such as the differential, logistic, probabilistic, or harmonic, can also be used to build a model. The method allows you to simultaneously obtain the optimal structure of the model and the dependence of the output parameters on the selected most significant input parameters of the system.

GMDH has advantages when

1. there is no or almost no a priori information about the structure of the model and the distribution of its parameters

2. observational data is extremely small, to the extent that the model parameters are greater than the number of observations

In our case, the obtained data set fully meets the conditions.

We describe the stages of the implementation of this method.

1. Consider various component subsets of the base function (1.1) called partial models. The most commonly used dependencies are:

$$Y = a_0 + a_1 x_i x_j, (1.2)$$

$$Y = a_0 + a_1 x_i + a_2 x_j, (1.3)$$

$$Y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j, (1.4)$$

$$Y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2 (1.5).$$

2. Various models are built for some or all of the arguments. For example, polynomials with one variable are constructed, polynomials with all possible pairs of

variables, polynomials with all possible triples of variables, and so on, a polynomial with all variables. For each model, its coefficients are determined by regression analysis methods.

3. Among all models, several (from 2 to 10) are selected. The quality of models is determined by the coefficient of determination, or the standard deviation of the error, or other external criteria.

4. If a sufficiently “good” model is found or the maximum permissible complexity of the models is reached, the algorithm ends.

5. Otherwise, the models found at the 3rd step are used as arguments for the support functions of the next iteration stage (go to the 2nd item). That is, already found models are involved in the formation of more complex relationships.

Usually, the degree of the polynomial of the support function is chosen not higher than N-1, where N is the number of sample points. It is often enough to use polynomials of the second degree as a support function. In this case, at each iteration step, the degree of the resulting polynomial is doubled.

Instead of the Kolmogorov-Gabor polynomial, Fourier series can be used. It makes sense to apply them if periodicity is observed in the source data. The model obtained, in this case, will be polyharmonic.

Often, the initial sample is divided into two subsamples A and B. Sample A is used to determine the model coefficients, and subsample B is used to determine the quality (coefficient of determination or standard deviation). The ratio of the amount of data in both samples depends on theoretical assumptions.

Statistics show that with each iteration step, the standard deviation decreases. But after reaching a certain level of complexity (depending on the nature and amount of data, as well as the general type of model), the standard deviation starts to grow.

The group method of data handling finds knowledge about an object directly from data sampling. It is an inductive selection method that has advantages for rather complex objects that do not have a specific theory, in particular for objects with fuzzy characteristics. The GMDH algorithms find the only optimal model for each sample using a complete enumeration of all possible candidate models and evaluate it by an external exact or balanced criterion on an independent data subselection.

There are 4 basic algorithms GMDH and many of their modifications

· COMBI - combinatorial algorithm

· MULTI - combinatorial selection algorithm

· MIA - multi-row iterative algorithm

· RIA - relaxation iterative algorithm

Consider the basic algorithms on the example of a linear model of many variables

$(x_1, x_2, ..., x_{20})$

All algorithms are multi-row. Each row is associated with one level of model complexity.

Obviously, there can be several models in each row. Algorithms differ in terms of forming and selecting variables when moving from one row to another. Algorithms have settings, including the number of the best models of each row. Parameters are set by the user.

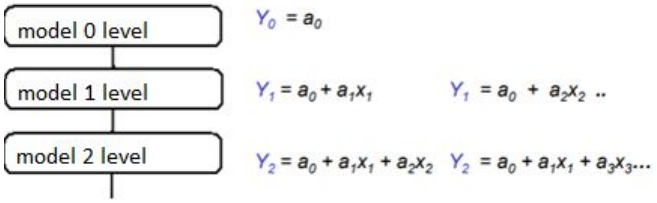

Figure 1.5 - Multi-row in the algorithm

COMBI

The COMBI idea is a combinatorial algorithm. COMBI is the simplest of the basic algorithms of the GMDH. The idea of the algorithm: do not miss any of the various models. Therefore, at each level of complexity:

- all models are considered
- selection of the best combinations of variables is not carried out

If the number of variables of the model is N, then the number of all combinations is

M = 2^N

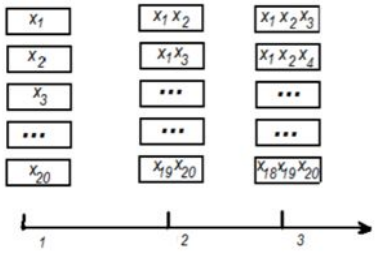

Figure 1.6 - COMBI Algorithm

Brute force rule

- Calculates the quality of all models of a given series. Among these models, choose the best model, called a record

- If the current record is worse than the record of the previous row, then the process of complicating the model stops

MULTI

The idea of MULTI is a combinatorial selection algorithm. MULTI is a development COMBI. His main idea is to reduce the number of models considered on each row, and at the same time if possible not to lose the best combination of variables. Therefore, at each level of complexity:

- A fixed number of the best combinations of model variables is selected.

- These best combinations are combined with all other variables (one by one) when moving to the next level.

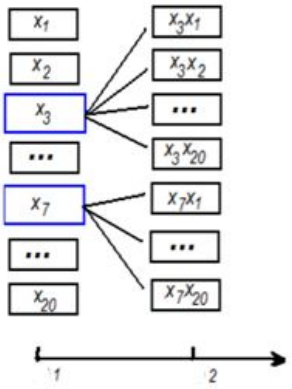

Figure 1.7 - MULTI Algorithm

MIA

The MIA idea is a multi-row iterative algorithm. MIA is historically the first of the MGUA algorithms. Key MIA ideas:

- reduce the number of models considered on each row

- reduce the number of rows, and thereby accelerate the achievement of the optimum level of complexity.

Therefore, on each row:

- a fixed number of the best models is selected, (each model is considered as a variable!)

- each pair of the best variables generates a new variable when moving to the next level

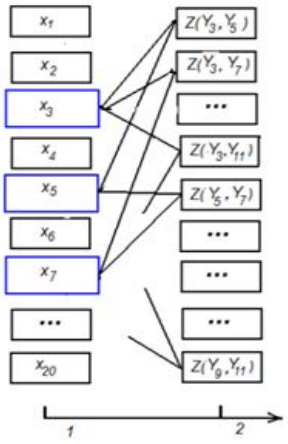

Figure 1.8 - MIA Algorithm

In my work, I use two algorithms. COMBI allows find the best model for each row, and the generation of a new variable gives the best result.

4. Criteria for the selection of models used in GMDH.

In order to reduce number of options, i'm using the method of selection. When choosing the optimal structure of the equation, the arguments are searched in groups (or in pairs), and the coefficients are determined by the least squares method on the training sample of the source data, and the resulting version of the model is evaluated using the specified selection criterion on the test sample of data.[9]

As internal criteria for the selection of the prognostic model, the criteria for the accuracy of the point forecast and the criteria derived from them are used.

The criterion for selecting a model can be called external if it is obtained using additional information that is not contained in the data that was used to calculate the model parameters. For example, such information is contained in an additional, test sample.

The GMDH algorithm uses both internal and external criteria. The internal criterion is used to adjust the parameters of the model, the external criterion is used to select the model of the optimal structure. You can select models by several external criterias.

Coefficient of determination. (R ^ 2 - R-square)

The proportion of the variance of the dependent variable, explained by the dependency model under consideration, that is, the explanatory variables.

The true coefficient of determination of the model of the dependence of the random variable y on the factors x is determined as follows:

$$R^2 = 1 - \frac{V(y|x)}{V(y)} = 1 - \frac{\sigma^2}{\sigma_y^2}, \quad (1.6)$$

where $V(y/x) = \sigma^2$ - conditional (in terms of factors x) variance of the dependent variable (variance of the random error of the model).

In our case

$$R^2 = 1 - \frac{\sum_{i=1}^{k} (y_i - \widehat{y_i})^2}{\sum_{i=1}^{k} (y_i - \overline{y_i})^2}, \quad (1.7)$$

$\sum_{i=1}^{k} (y_i - \widehat{y_i})^2$ - the sum of the squares of the regression residuals,

$\sum_{i=1}^{k} (y_i - \overline{y_i})^2$ - common dispersion

$\overline{y_i} = \frac{1}{k} \sum_{i=1}^{k} y_i$ - sample mean

The coefficient of determination for the model takes values from 0 to 1 and is a total measure of the quality of the resulting equation used for the selection of models. If the value

of the coefficient of determination is close to 0, then the model constructed is not qualitative, not applicable in practice. If the value of the coefficient of determination is close to 1, then the quality of the constructed model is subject to furthermore thorough investigation. In our work, we analyzed the law of distribution of random deviations of the model, counting the number of incorrect answers. The threshold value of the coefficient of determination, after which the model is selected for further research, varies depending on the nature of the problem being solved. If we are talking about the possibility of catastrophes, the value of the coefficient of determination is considered satisfactory if R ^ 2> 0.99. If we are talking about some managerial decisions, the researcher can call R ^ 2 = 0.2 a satisfactory value. The nature of the data we studied allowed us to choose the threshold R ^ 2 = 0.7.

The criterion of regularity.

The criterion of regularity $\Delta^2(C)$ includes the root-mean-square error on the training subsample *C* obtained with the model parameters configured on the test subsample *l*.

$$\Delta^2(C) = |y_C - A_C \widehat{w}_\ell|^2 = (y_C - A_C \widehat{w}_\ell)^T (y_C - A_C \widehat{w}_\ell), \quad (1.8)$$

$$\eta^2_{bs} = |A_W \widehat{w}_\ell - A_W \widehat{w}_C|^2 = (\widehat{w}_\ell - \widehat{w}_C)^T A_W^T A_W (\widehat{w}_\ell - \widehat{w}_C). \quad (1.9)$$

where $\widehat{w}_\ell = (A_\ell^T A_\ell)^{-1} (A_\ell^T y_\ell)$ and $\widehat{y}_C(\ell) = A_C \widehat{w}_\ell$.

Matrix A_W - a set of column vectors a_i. Other modifications of the criterion of regularity:

$$\Delta^2(C) = \frac{|y_C - A_C \widehat{w}_\ell|^2}{|y_C|^2} \quad (1.10)$$

and

$$\Delta^2(C) = \frac{|y_C - A_C \widehat{w}_\ell|^2}{|y_C - \bar{y}_C|^2}, \quad (1.11)$$

where $\overline{y}$ - mean vector y.

The criterion $\Delta^2(C)$ is also denoted $\Delta^2(C \setminus \ell)$, i.e an error in the subsample C with the parameters obtained in the subsample ℓ.

Chapter 2

Building dependency models of social network data.

1. Collection of information

For the practical implementation and verification of the theoretical representation of the model, information from the profiles of KFU's students social network was reviewed.

Uploading data from the social network was carried out using the social networking API Vkontakte, which allows you to get the data that the user independently opened for only public access.

Lists of groups and pages to which the user is subscribed and the number of subscriptions was extracted.

After uploading the date and additional data analysis, it is necessary to prepare the data and convert it into a form that can be used with the selected analysis tools. Namely, to find out where the user will belong in the list of studied groups.

The list of groups was formed on the basis of the base of all subscriptions of the studied users. From the final data set obtained, variables with a low number of students who entered this group were removed.

1168 student profiles were reviewed. 63 of these were closed accounts, i.e. 5 %.

Information was obtained on the 24,129 groups and the number of students who were in this group.

Identified the most popular groups:

Id group	Number of participants	Group name
38959783	389	“Казань | Казань. Куда пойти? | Афиша”
92943238	332	“Подслушано КФУ”
54530371	235	“Библиотека программиста”
40876092	233	“Student's life”
157299408	216	“2ойка”
31480508	198	“Пикабу”
72034968	193	“Цитаты преподавателей КФУ”
50983956	193	“Include”

36887378	182	“ЭЛЕКТРОННАЯ БИБЛИОТЕКА ИВМиИТ 1-4 КУРС”
57867786	165	“ВКазани Поймут | Главное сообщество Казани”

Table 2.1 - 10 most popular groups

2. The use of GMDH to obtain a model of dependence.

During the analysis of groups, 306 of the most popular groups remained in the data set. In each of their remaining groups, the number of students in this group is at least 21 years old.

The use of GMDH to obtain a model of dependence of belonging to some specific group on belonging to other groups.

Here is an example. Let x be a network user. We will consider x as an array of 0 and 1, which shows that they belong to groups selected during data analysis, i.e.

$$x_i = \begin{cases} 1, \text{if } x \in G_i \\ 0, \text{if } x \notin G_i \end{cases} \quad (2.1)$$

where G - the set of groups that we are considering, i= (1,306).

Let Y be the output value. Y is 0 or 1 :

$$Y = \begin{cases} 1, \text{if } x \in G_0 \\ 0, \text{if } x \notin G_0 \end{cases} \quad (2.2)$$

where G_0 - where is the group to which we are a member?

Partial model in the first interaction:

$$Y = c_0 + c_1x_1 + c_2x_2 + \ldots + c_kx_k \quad (2.3)$$

We divide our sample into a training 45%, validation 45% and test 10%. We use the training sample to build a model, validation for its evaluation.

We use the COMBI algorithm to find the optimal number of variables for the further construction of the model.

Plot a graph based on the data collected

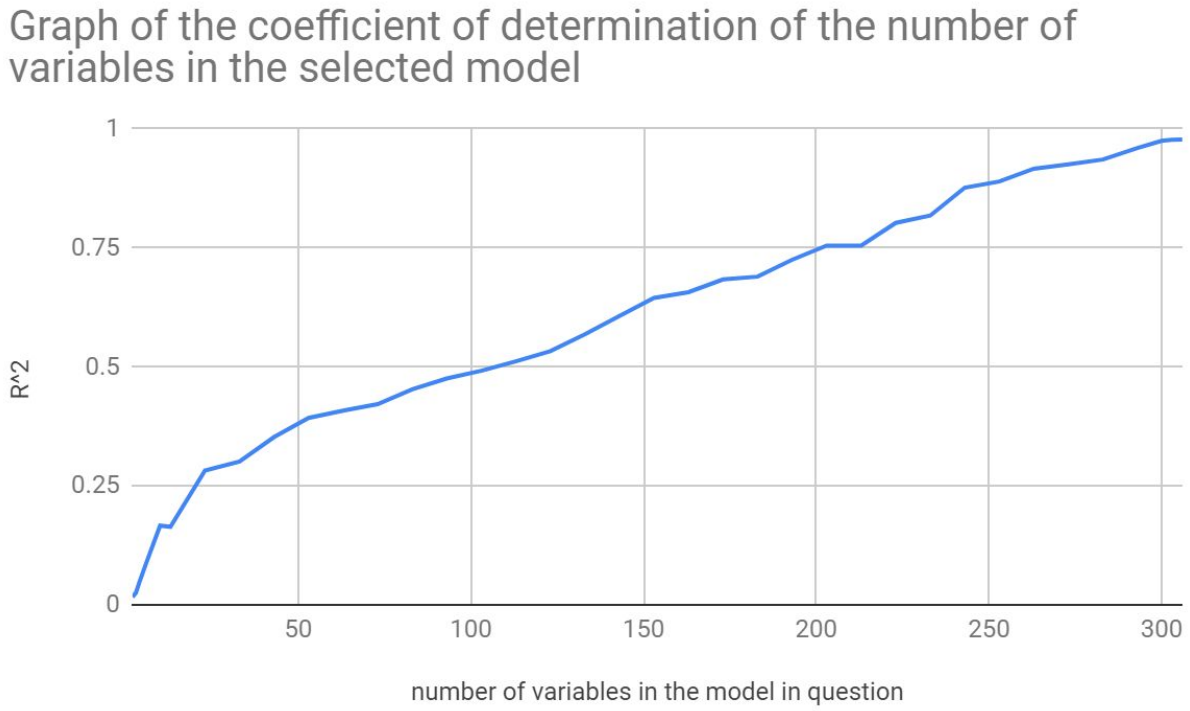

Figure 2.1 - Graph of the coefficient of determination of the number of variables in the selected model

As you can see in the figure 2.1, the optimal number of variables in the first iteration is 200.

Brute-force calculation of the coefficient of determination and weight. Sample file received:

```
0.7445868749743929 0.039649083074011 98 1.1096845687184809 1.0247493299023032 0.775530606977008 -0.25128073345267243 0.7108240587466591 -0.06540332900396359
0.7440846045954603 0.03217916606612611 1.107570631648369 1.0204955884278322 0.7784648631642408 -0.26340198496447076 0.704788407092021 -0.10925229241238504 0
0.7437749939084204 0.036216880418894946 1.1050174979863117 1.0081719637995803 0.7772384997688542 -0.2393856441260639 0.7213077298277599 -0.11087596418107296
0.7435912879362341 0.035808246438274446 1.1036267910888482 1.0209235314546012 0.7741626357154303 -0.24917550677579525 0.7174836121405213 -0.1093316039584211
0.7481306483424217 0.042808164380881 5 1.0806001437163246 1.0182805763629887 0.8122890779628027 -0.30436907650959255 0.7122538848476385 -0.03605785256236993
0.752520285387982 0.04455838585679818 1.156747712461203 0.9994496687184371 0.8150118198474161 -0.18621617192792939 0.6862022308868703 -0.1278356899502169 0.
0.743600490847629 0.036001149911680484 1.110147953984638 1.0113174925668524 0.776296109041031 -0.24643875920074876 0.716134529668899 -0.11443407385381185 0.
0.7440239543691194 0.03395374600276413 1.1072585167064197 1.0165050096140251 0.7821130827788385 -0.24540469959824704 0.7061834427222573 -0.11102190346229666
0.7437094695708641 0.03545703512482306 1.1043156993060603 1.0232876039554888 0.765905233001727 -0.2566624072915793 0.7213653460675238 -0.10886662408271776 0
0.7439065654430803 0.03151705312821538 1.1111401041176705 1.0276282654648452 0.7651690252562651 -0.2580053132269544 0.722476678599351 -0.1299504331530003 0.
0.7461382267108863 0.03749129557122194 1.1182127902637926 1.0245963323001888 0.7803034495934811 -0.1868694328435032 0.703225453308775 -0.1398380983895 4023 0
0.7478818807692007 0.02597148840016 3143 1.075800346849475 1.0095934374451283 0.7756954153134737 -0.30953003397108214 0.6761988432375456 -0.09594921452225849
0.7508971255827354 0.02591872305935 4148 1.029964680687745 1.07951983302332 0.7815990133658428 -0.2201417311345595 0.7701500866968692 -0.11135058526754647 0.
0.7467718126568792 0.03596462639520903 1.1390318283193281 0.9995225419329115 0.7721589246515245 -0.2533379609294947 0.7139503270044976 -0.12339650073282282
0.7451523363706143 0.03317547484441659 1.1111518014732025 1.0133870538214043 0.782720795419212 6 -0.2764413655267177 0.6985365826115949 -0.09308068107802667
0.7449928857315116 0.03410998800868863 1.1221599128735436 1.0372205279295381 0.7840023338619614 -0.26448481317883143 0.6905969956237296 -0.11873357601132828
0.7435296810553778 0.03550718677733022 4 1.1100631798800031 1.017307183103506 0.7773009069029848 -0.25365784967141997 0.7160233592005639 -0.11304093022349955
0.7458399542086134 0.03706299265528234 1.122929589445942 1.0082619803142008 0.8183767350213524 -0.2024732040768822 0.7437668283440932 -0.07963720860875603 0
0.7467465656939214 0.03443331812272503 1.124274542573089 1.0489236193184994 0.7789337020034902 -0.23097273362298545 0.7257492279962814 -0.10220300815255788
0.7501682086317216 0.03857513261494033 1.1236939642130475 1.0054078547419991 0.7890880467136021 -0.2200253113473412 0.7416620527133195 -0.08953038612636911
0.7443805411479905 0.03435177460976903 1.11273930290775 1.0169543998692352 0.7785563545930877 -0.2564607774385246 0.73663801906079 -0.1009430049840441 0.839
0.7448492593409843 0.03496087907301324 1.1225344488388194 1.026071006782991 0.7637226824759483 -0.24312202471948766 0.7402753330114504 -0.1416133613074001 0
0.7466318631672674 0.02618171941638 3538 1.1065799147006339 1.01815689406686 0.7446195640459041 -0.2134767669361839 0.7653225267751315 -0.10139794846806185 0
0.7580287007412962 0.03084550566286 9102 1.193911009395754 0.9450174415982787 0.8159572448969661 -0.26333618209892395 0.7308977812336357 -0.12613821948185483
0.7437358709388897 0.03447510218982403 1.1141044817210892 1.0250070262324755 0.7752583945472918 -0.25057344046918917 0.7273246008894803 -0.10855213170823112
0.7460617583530416 0.02768165988340 45 1.1453904926690164 0.9948482200726282 0.7565680365260391 -0.25971217349827014 0.7060587662372891 -0.10806705825506127
0.7435542522189075 0.03515845087277018 1.1131392432596054 1.0223118778826257 0.7727851622491846 -0.25349741058051223 0.717594965518167 -0.11248394323094207
```

Figure 2.2 - Example of received file.

Investigate the main functions of the second and third iterations of the form:

$$1.Y = c_0 + \sum_{i=1}^{k} c_i Y_i + \sum_{i=1}^{k} \sum_{j=i+1}^{k} c_{ij} Y_i Y_j$$

$$2.Y = c_0 + \sum_{i=1}^{k} c_i Y_i^2 + \sum_{i=1}^{k} \sum_{j=i+1}^{k} c_{ij} Y_i Y_j$$

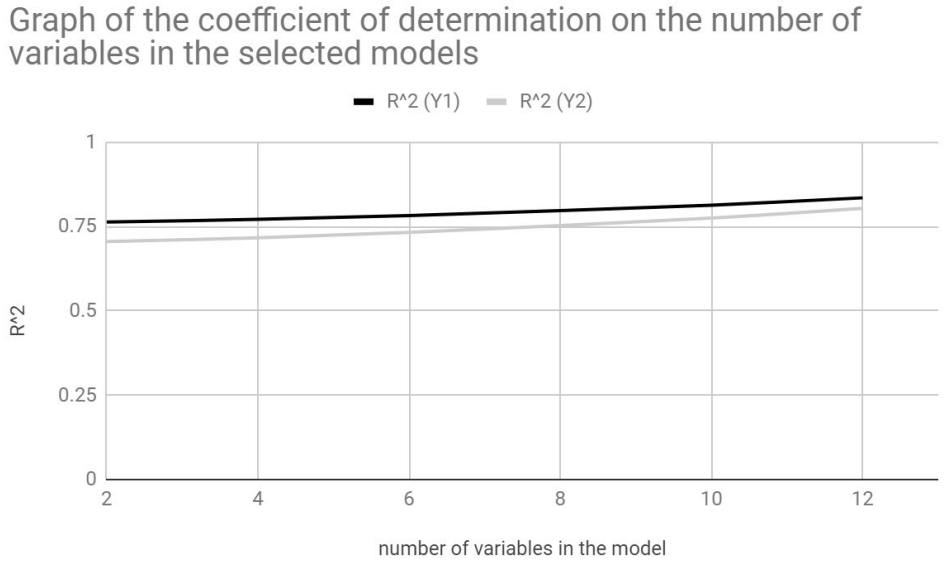

Figure 2.3 - Graph of the coefficient of determination of the number of variables in the selected models

As seen on figure 10, the model in second and third iteration

$$Y = c_0 + \sum_{i=1}^{k} c_i Y_i + \sum_{i=1}^{k} \sum_{j=i+1}^{k} c_{ij} Y_i Y_j \quad (2.1)$$

for our data shows the better result.

The number of variables at the second iteration giving the best result is 12.

Lets estimate the optimal number of variables for the third iteration

The number of variables	R^2	The number of correct answers
2	0.8799632614171315	341
3	0.8819918535378112	341
4	0.8844280354824443	341
5	0.8896014237326587	341
6	0.8991089662961465	341
7	0.9015088163336399	341

Table 2.2 - Result in 3rd iteration

All experiments showed a good rating. According to the obtained result, the last experiment showed the best rating. Thus, we construct a model of 7 variables on the third interaction.

Thus, we have obtained a model, where G_0 - “Библиотека программиста” with id = 54530371:

$$
\begin{aligned}
Y = c_{03} + \sum_{i=1}^{7} c_{i3} \Bigg(& c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,\ 1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \Bigg) \\
& + \sum_{i=1}^{7} \sum_{j=i+1}^{7} c_{ij3} \\
& * \Bigg(c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \Bigg) \\
& * \Bigg(c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \ldots + c_{200,1}x_{200}\right) \Bigg)
\end{aligned}
\tag{2.2}
$$

where the coefficients c_{ij} are obtained in the resulting file:

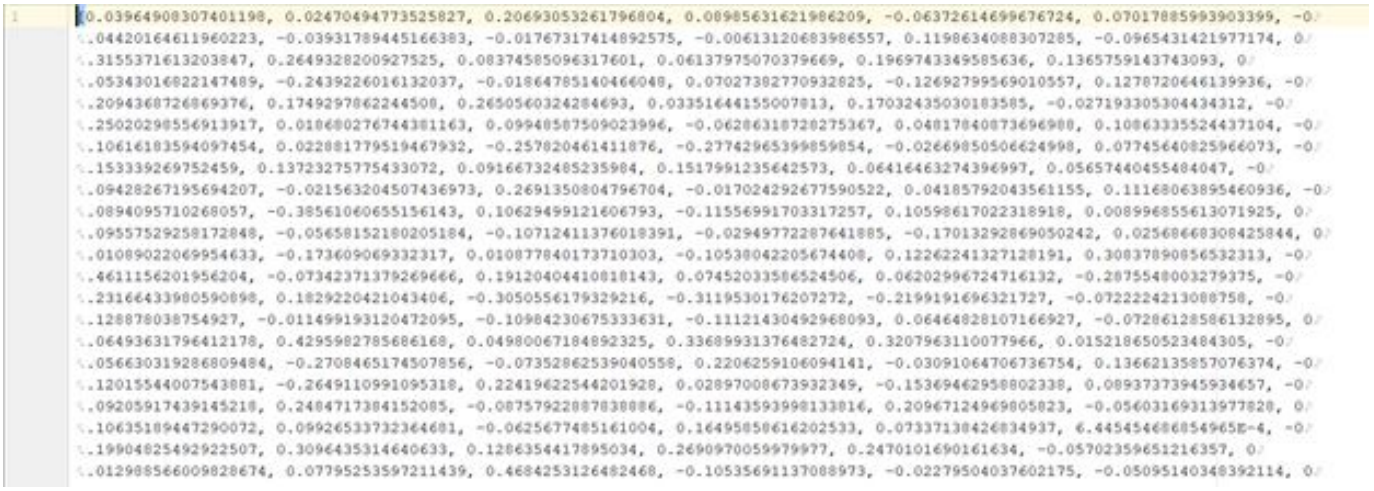

Figure 2.4 - Example of the result file

Chapter 3

Probabilistic assessment of the number of correct answers obtained using the constructed model.

1. Testing the model.

A test sample is 10% of the data collected, total 74 students.

Testing the resulting model

$$
\begin{aligned}
Ytest = c_{03} + \sum_{i=1}^{7} c_{i3} \Bigg(& c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,\ 1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \Bigg) \\
& + \sum_{i=1}^{7} \sum_{j=i+1}^{7} c_{ij3} \\
& * \Bigg(c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \Bigg) \\
& * \Bigg(c_{02} + \sum_{i=1}^{12} c_{i2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \\
& + \sum_{i=1}^{12} \sum_{j=i+1}^{12} c_{ij2}\left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \\
& * \left(c_{01} + c_{11}x_1 + c_{21}x_2 + \dots + c_{200,1}x_{200}\right) \Bigg)
\end{aligned}
$$

Consider the resulting function as

$$Y = \begin{cases} 1, \text{if } |Ytest - 1| < |Ytest| \\ 0, \text{else} \end{cases}$$

As a result, we received 56 correct answers, i.e. 75%

2. Probabilistic assessment of the number of correct answers obtained using the constructed model.

In previous studies, a model of dependence of the user's belonging to one of the groups on the user's belonging to several identified other groups was obtained. The general view of the model obtained can be represented as follows:

$$Y = f(X_1, X_2, \ldots, X_n) + \varepsilon,$$

where **ε** is a random variable, the characteristic difference between the statistical and theoretical values of the models, the remaining regression.

The random variable **ε** has the size of Bernoulli.

$$\varepsilon = \begin{cases} 1, with\ probability\ p \\ 0, with\ probability\ 1 - p \end{cases}$$

To estimate a numerical value, the probability p was considered as a random variable.

The data of users belonging to the same academic group was used to calculate the relative frequency of random deviation values, which were interpreted as selective values of p. To obtain a set of sample values of p, this procedure was performed for other academic groups. The calculations were used to construct an empirical distribution function for which a theoretical approximation was obtained.

The mathematical expectation of the considered random variable was used to estimate the probability value p: $E\,\hat{p} \approx 0,2$ .Thus, Y take the correct values with probability $1-p \approx 0,8$.

Conclusion

The main results of this thesis are as follows:

1) Real data of real user data were processed and collected by accessing social service web interfaces

2) Program group method of data handling for analyzing social network data.

3) A model of a social network data dependency model has been identified.

4) Probable estimate of the quality of a probabilistic model is obtained

The main goal of the work was to create probabilistic model for researching incomplete social network data.Thus, the main aspect of the mechanism for analyzing incomplete user data from social networks were investigated and analyzed.

As for further research, there are several possible directions. The first is to build a methodology for filling in the missing data of the social network, taking into account the specifics of the data being processed. The second is the identification of the dependence of the student's progress on his profile on the social network. The third is the construction of a system of equations in which the interdependence of many factors is reflected. Since the equations in the system will have more complex forms of dependencies, to obtain them, it will be necessary to construct a joint law for the multidimensional distribution of probabilities of the values of the quantities being studied.

Bibliography

1 Najork M., Wiener J. L. Breadth-first crawling yields high-quality pages. Proceedings of the 10th international conference on World Wide Web. – ACM, 2001. – pp 114-118.

2 Leskovec J., Faloutsos C. Sampling from large graphs. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2006. – pp 631-636

3 Giyasov, B.I., Antonov, A.I., Matveeva, I.V. Energy method for calculating noise penetrating flat rooms through walls / B.I. Giyasov, A.I. Antonov, I.V. Matveyev // Vestnik MGSU. - 2014. - Vol. 9. - pp22-31 (in Russian)

4 Anton Korshunov, Ivan Beloborodov, Nazar Buzun, Valery Avanesov, Roman Shepherds, Kirill Chikhradze, Ilya Kozlov, Andrey Gomzin, Ivan Andrianov, Andrey Sysoev, Stepan Ipatov, Ilya Filonenko, Christina Chuprina, Denis Turdakov, Kuznetsovтатья Analysis of social networks: methods and applications - 2013 - pp.439-441

5 Data analysis of social networks. Available at [https://www.osp.ru/os/2015/03/13046896/] (in Russian)

6 Key Trends to Watch in Gartner 2012 Emerging Technologies Hype Cycle. Available at [http://www.forbes.com/sites/gartnergroup/2012/09/18/key-trends-to-watch-in-gartner2012-emerging-technologies-hype-cycle-2/]

7 VK. Available at [https://vk.com/]

8 Group Method of Data Handling. Available at [http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%93%D0%A3%D0%90]

9 Vasiliev, A.A. (2012) Criteria for selection of forecast models (review). Bulletin of Tver State University. Series: Economics and Management (13). Pp. 133-148.(in Russian)

10 What is that VK? Available at [https://netrocket.com.ua/blog/chto-takoe-vkontakte/]

11 Handling gaps in data. Available at [ttps://basegroup.ru/community/articles/missing]

12 Burton A., Altman D. G. Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. British Journal of Cancer, 2004, 91(1):4–8.

13 Horton N.J., Kleinman K.P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am. Stat. 2007; 61: pp 79–90.

14 Karahalios A., Baglietto L., Carlin J.B., English D.R., Simpson J.A. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Med Res Methodology, 2012;12:96.

15 Knol, M. J., Janssen, K. J. M., Donders, A. R. T., Egberts, A. C. G., Heerdink, E. R., Grobbee, D. E., Moons, K. G. M., and Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. Journal of Clinical Epidemiology, 63: pp 728–736.

16 Miettinen, O. S. Theoretical Epidemiology: Principles of Occurrence Research in Medicine. John Wiley & Sons, New York. 1985, p. 232.

17 Molenberghs, G. and Kenward, M. G. Missing Data in Clinical Studies. John Wiley & Sons, Chichester, UK. 2007 - pp. 47-50.

18 Panteha Hayati Rezvan, Katherine J Lee, Julie A Simpson -The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Medical Research Methodology, 15(30), pp 1–14.

19 Vach, W. and Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. American Journal of Epidemiology, 134(8), pp 895–907.

20 Van Buuren S. Flexible Imputation of Missing Data. Chapman and Hall/CRC; 1 ed., 2012 - 342 p.