PROBABILISTIC CALCULATION OF MISSING DATA VALUES OF THE SOCIAL NETWORK USERS: PROBABILISTIC ESTIMATE OF THE VALUES OF THE MISSING DATA OF THE SOCIAL NETWORK USERS

Nadiya Yangirova[1]
Zulfira Enikeeva[1]
Galim Vakhitov[1]

1 Kazan Federal University
E-mail: GZVahitov@kpfu.ru

ABSTRACT

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 effectively solved earlier due to lack of data). In addition, this causes increased interest in the collection and analysis of social data from companies and research centers. However, many of them are hidden or not always correct. Therefore, before proceeding with the analysis of the data, it is necessary to carry out their adjustment, normalization, and propose a probabilistic estimate of the values of the missing data. This article explores the problem of predicting incomplete data using the method of group accounting of arguments (MGUA). To solve this problem, a model of the relationship between the studied data taken from the profiles of the social network VKontakte was found. The resulting model determines the relationship of a user's subscription to one interesting page of the network and user subscriptions to a group of other interesting pages, allows to predict the user's interest in a certain topic underlying the selected interesting page, depending on which interesting pages are already subscribed to at a certain moment

Keywords: social network, analysis of the social network, identification modeling, the method of group accounting of the argument, virtual behavior, forecasting, method.

1.INTRODUCTION

Social networks, such as Facebook, Twitter, YouTube, allow you to chat, share materials with people around the world. Thanks to such socialization, data became available that were previously inaccessible or for the collection of which a tremendous amount of time and money was spent. For example, there are opportunities to timely identify psychological problems in people who have accounts on social networks [1]. In this regard, data analysis has become very popular, new opportunities have opened up for research tasks [2], [3].

High interest in the topic of data analysis was confirmed by the global company Gartner. She analyzes current research methods, such as in-depth patent research, industry best practices, trend analysis, and quantitative modeling. In his 2017 Hype Cycle for Emerging Technologies article [4], Gartner singled out Social Network Analysis

and Big Data technologies and revealed that they are now at Peak of inflated expectations. Also, the research of data collected from social networks is carried out by many prestigious universities such as Stanford, Oxford, INRIA, and large corporations find this relevant and useful. Also, the owners of social networks themselves are investing in the development of algorithms for processing huge amounts of data. This is done for more accurate recommendations to users, for example, topics of interest to them, a possible friend, audio and recommendations, etc. Also important is the problem of identifying factors that positively affect the well-being of a person [5]. In particular, a well-tuned model of profile interests can easily identify the target audience for the sale and advertising of certain content. There are more and more companies involved in the collection and processing of the information submitted to users, as well as its storage. Researchers at many companies model various processes using social network data.

However, it is necessary to consider the possibility of low-quality data (false data), as well as problems with the storage of personal information. To work with data obtained by collecting it from a social network, one must also take into account their frequent updates. All this cannot do without continuous improvement of algorithms [6,11,12].

Social media web interfaces are real-time data sources and are intended for viewing and interacting with social network pages in a web browser or for using user data by specialized applications. Since the scenarios for using social network interfaces do not imply the automatic collection of data from many users in order to build a social graph, a number of problems arise:

1. data privacy;
2. poor data structure;
3. access restrictions and blocking.

The purpose of this study is to obtain a probabilistic model for investigating incomplete social network data. This work describes a possible mechanism for analyzing user data of social networks, namely finding dependencies between network objects.

The study was conducted in 2018-2019. Undergraduate students of the Institute of computational mathematics and information technologies of the Kazan Federal University (Kazan, Russian Federation, www.kpfu.ru). The social network VKontakte (www.vk.com) was used. As the source data, the URLs of their personal pages on the social network were used.

2.METHODS

VKontakte is a social network; it occupies the first place in terms of popularity among the Russian-language segment. Each network user has his own personal page - his profile. He also has the ability to share various information: send messages, create post-articles, share the photo and video materials. One of the main opportunities of a social network is the appearance of a group and a community where users can find the information they need on various topics and just chat with friends of interest and even organize meetings. With such images, you can quickly convey information to a large number of users, without limiting yourself to any framework. It is precisely the speed and optimality that attract entrepreneurs and business-less projects.

Today, there are several ways to collect data:

1. specialized companies that collect information and constantly maintain its relevance. This method is very good for the speed of data acquisition, but has a paid basis;
2. use of API - software interface. In the social network, we are considering, you can get complete data about the user by JSON requests; however, it has its own limitations (the number of calls in a certain period of time). However, the simplicity and ability to combine the API with the application adds an advantage to this method. However, this feature is not good for all social networks. For example, if you use the API provided by Facebook, then at the output the requested one will receive almost "zero" information;
3. independent collection of information from profiles. In this case, the speed of information collection will greatly lose. Each profile is unique and for this method, there will not be a specific analysis rule. To support this method, a large amount of computing resources is required. However, this process is well parallelized.

It must be borne in mind that trusting the information provided by social networks is not always necessary. It can be deliberately false or absent altogether. Therefore, before processing, it is necessary to normalize and verify the correctness of the data. Missing data must be restored based on the data we received: photos, posts, statuses, analysis of the profiles of friends, the community in which he is a member, etc.

Since the social network is not designed for automatic data collection, a number of problems arise:

1. poor data structure; Due to the limited functionality of the API, pulling out the necessary data is carried out using separate methods that need further transformation into the view used for analysis;
2. closed, private profiles and locks. Access to data is provided only to registered users of the network;
3. Access Restrictions Since online servers do not provide for automatic data collection, owners for data security limit the number of requests per unit time from one IP address, which protects the server from various DoS attacks;
4. The huge amount of data. Due to a large amount of data, it is necessary to use parallel data collection, as well as subsequently converting them to a form convenient for analysis, using various sampling methods.

To access information about users and the communities in which they are members, VKontakte API was used, which provides methods for working with social network data. The number of calls to API methods has a limit of no more than 3 times per second [7,13].

To collect the data we study, the program module sends requests to the VKontakte API methods:

1. Retrieving user information using the users.get method;
2. Getting a list of popular communities for each user studied using groups.getCatalog;

Often there are gaps in the data that you need to work with, resulting in a choice: ignore, discard, or fill in the missing values. Filling in the gaps often, and quite justifiably, seems to be the preferred solution. However, this is not always the case. An unsuccessful

choice of the method of filling in the gaps may not only not improve, but also greatly worsen the results [8,14].

Excluding and ignoring strings with missing values has become the default solution in some popular application packages, as a result of which novice analysts may have the idea that this solution is correct. In addition, there are methods for processing gaps that are quite simple to implement and use, called “ad-hoc methods”: filling gaps with zeros, a median, the arithmetic mean value, introducing indicator variables, and the like, the simplicity of which can be the reason for choosing these methods [9].

There are the following 3 mechanisms for creating passes: MCAR, MAR, MNAR. Briefly reveal their contents.

MCAR (Missing Completely At Random) - a mechanism for creating gaps, in which the probability of skipping for each record in the set is the same. For example, if a sociological survey was conducted in which one out of ten respondents did not ask one randomly selected question, and all the other questions asked were answered by respondents, then the MCAR mechanism takes place. In this case, ignoring/excluding records containing missing data does not distort the results.

MAR (Missing At Random) - in practice, data is usually not skipped accidentally, but because of some patterns. Gaps are classified as MAR if the probability of a skip can be determined on the basis of other information available in the data set (gender, age, position, education, etc.) that does not contain gaps. In this case, deleting or replacing the blanks with 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 creating gaps in which data is not available depending on unknown factors. MNAR suggests that the probability of skipping could be described based on other attributes, but there is no information on these attributes in the dataset. As a result, the probability of omission cannot be expressed based on the information contained in the data set.

Filling gaps using stochastic linear regression in the general case leads to the least distortion of the statistical properties of the sample. However, for more correct data recovery, more subtle methods are needed, therefore we use the method of group accounting of the argument. It is based on the recursive selective selection of models, on the basis of which more complex models are built [10].

This method is based on enumerating gradually more complicated models and choosing the best solution according to an external criterion. As basic models, not only polynomials can be used.

Most MSUA algorithms use a polynomial basis function. The general relationship between input and output variables can be expressed as a functional series of Volterra:

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

where X ( ) is the input vector of variables;

A ( ) is a vector of coefficients or weights.

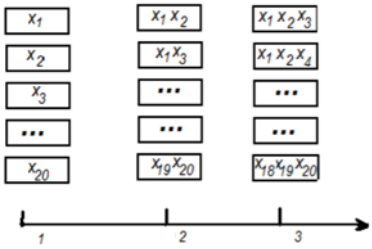
The components of the input vector X can be independent variables, functional forms, or finite difference terms.

The method allows to obtain the dependence of the output parameters on the selected input, as well as to find the optimal structure of the model.

There are 4 basic algorithms of MGUA: COMBI, MULTI, MIA, RIA [8].

COMBI is combinatorial the algorithm, also is not skip nor is one of all kinds of models. Therefore, at each complexity level, all models are considered and the best combinations of variables are not selected

If the number of model variables N, then the number of all combinations M = 2^N



COMBI Algorithm

MIA is a multi-row iterative algorithm. The main ideas of MIA:

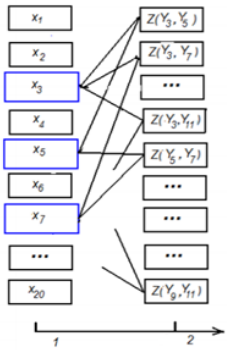
- reduce the number of models considered on each row

- reduce the number of rows, and thereby speed up reaching the optimal level of complexity.

Therefore, on each row:

- a fixed number of the best models are selected

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



MIA Algorithm

In this study, a combination of two algorithms was used. COMBI allows you to find the best model for each row, and spawning a new variable gives the best result.

3.RESULTS AND DISCUSSION

For the practical implementation and verification of the theoretical presentation of the model, information from the profiles of the VKontakte social network of studying KFU students was considered. Data was unloaded from the social network using the Vkontakte social network API, which allows you to get only the data that the user independently opened for public access. Lists of interest groups and pages to which the user is subscribed and the number of subscriptions were extracted.

After unloading and additional analysis of the data, it is necessary to prepare the data and transform it to the form that is possible for use with the selected analysis tools. Namely, finding the user's membership in the list of groups studied.

The list of groups was formed on the basis of the base of all subscriptions of the studied users. Groups with a low number of students in this group were removed from the final data set received.

1168 student profiles were reviewed. Of these, 63 closed accounts, i.e. 5). Information was obtained on 24,129 groups and their number of students in this group.

The most popular groups are identified and presented in table 1.

Table 1. Most Popular Groups

Group ID	Number of participants	Group name
38959783	389	“Kazan | Kazan. Where to go? | Poster ”
92943238	332	“Overheard KFU”
54530371	235	“Programmer's Library”
40876092	233	“Student's life”
157299408	216	“2oyka”
31480508	198	Picabu
72034968	193	“Quotes from teachers of KFU”
50983956	193	"include"
36887378	182	“ELECTRONIC LIBRARY IWMIiIT 1-4 COURSE”
57867786	165	“Come to Understand | The main community of Kazan ”

During the analysis of the 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 not less than 21.

Let's make a selection as follows. Let x be a network user. We will consider x as an array of 0 and 1, which shows the belonging to the groups selected during the data analysis i.e.

$$x_i = \begin{cases} 1, \text{если } x \in G_i \\ 0, \text{если } x \notin G_i \end{cases}$$

Where G is the set of groups we are considering,

Let Y be the output value. Y takes the value 0 or 1 if

$$Y = \begin{cases} 1, \text{если } x \in G_0 \\ 0, \text{если } x \notin G_0 \end{cases}$$

Where G_0 is the group to which membership is predicted.

We select the general view of the models being searched.

Assign the base model. Support function in the first iteration. has the form:

$$Y = f(x_1, x_2, x_3, x_4, \ldots, x_n)$$

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

We will use the COMBI algorithm and find the optimal number of variables for the further construction of models.

Based on the data obtained, we construct graph 1.

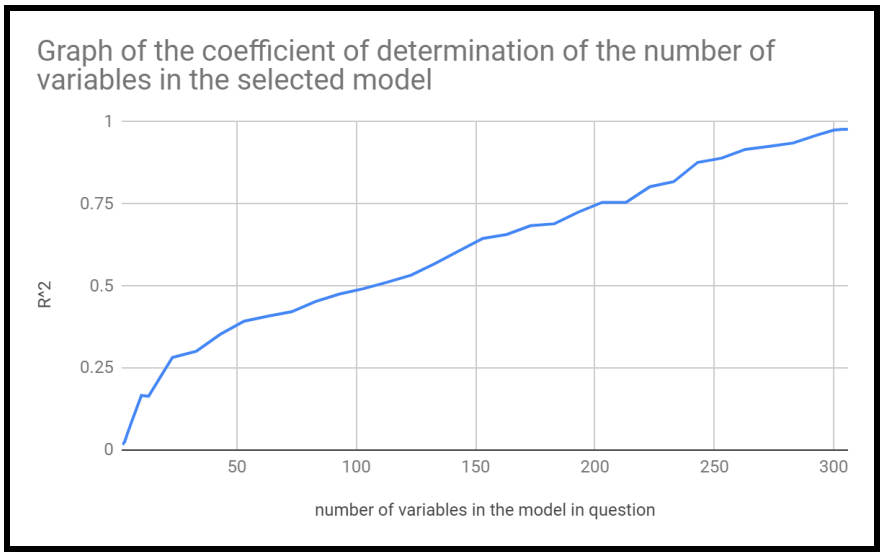


Figure 1 - Graph of the coefficient of determination of the number of variables in the selected model

As you can see in graph 1, the optimal number of variables in the first iteration is 200.

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

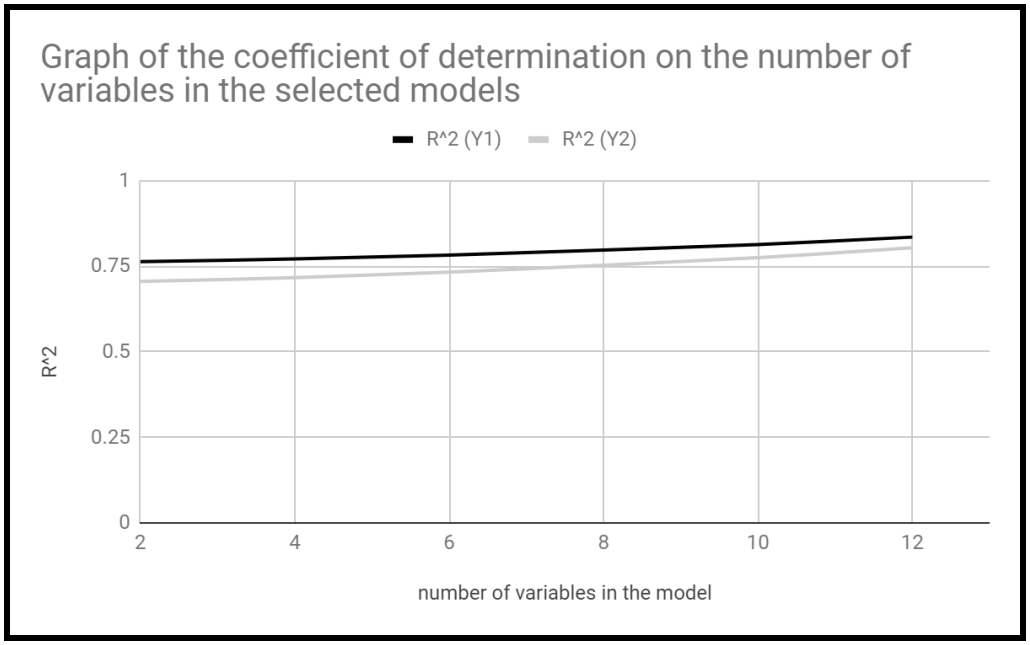


Figure 1 - Graph of the coefficient of determination of the number of variables in the selected model

As can be seen in graph 2, the model

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

for our data shows a better result.

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

We estimate the optimal number of variables for the third iteration and present the data in table 2.

Table 2. The optimal number of variables

Number of variables	R^2	The number of correct answers
2	0.87996	341
3	0.88199	341
4	0.88442	341
5	0.88960	341
6	0.89910	341
7	0.90150	341

All experiments showed a good grade. According to the result, the last experiment showed the best estimate.

Thus, we obtained a model, with G_0 - “Programmer's Library” with id = 54530371:

4.SUMMARY

The test sample was 10% of the number of individuals for whom data were collected, i.e. 74 students. 56 correct answers were received, which amounted to 75% of.

The study obtained a model of the dependence of the user's membership in one of the groups on the user's membership in several other identified groups. The general view of the resulting model can be represented as follows:

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

wherein ε - random quantity characterizing a difference between the theoretical values and statistical models, regression residue.

The random variable ε has a Bernoulli distribution.

$$\varepsilon = \begin{cases} 1, \text{ с вероятностью } p \\ 0, \text{ с вероятностью } 1-p \end{cases}$$

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

On the data of users belonging to the same academic group, the relative frequency of random deviations was calculated, which were interpreted as selective p values. To obtain a set of sample p values, 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 random variable in question was used to estimate the probability value p: $E\hat{p} \approx 0{,}2$. Thus, Y takes the correct values with probability $1 - p \approx 0{,}8$.

5.CONCLUSIONS

Based on the materials studied and the analysis of the social network, a model for the interaction of the studied data using the method of group accounting of arguments was revealed. The best base models for such data were found. An optimal algorithm for enumerating variables was also identified. An analysis was also carried out of the interests of the student communities of the Institute of Computational Mathematics and Information Technologies of Kazan Federal University, and the dependence on the professional community "Programmer's Library" was revealed.

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

ACKNOWLEDGMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University. The study (all theoretical and empirical tasks of the research presented in this paper, except for the payment of publishing services) was supported by a grant from the Russian Science Foundation (Project No. 19-18-00253, "Neural network psychometric model of cognitive-behavioral predictors of life activity of a person on the basis of social networks").

REFERENCES

1. L.M. Popov, P.N. Ustin. (2016) Psychological alienation problem in moral and ethical psychology of personality. Mathematics Education. -Vol. 11. -№ 4. pp. 787-797 https://www.iejme.com/article/psychological-alienation-problem-in-moral-and-ethical-psychology-of-personality
2. D. Tumakov, C. Godovykh, A. Valeeva. Mathematical model of socio-psychological adaptation through a person's interaction with the environment // Herald National Academy of Managerial Staff of Culture and Arts, 2018, No. 3, pp. 310-318.
3. J. Leskovec, C. Faloutsos. Sampling from large graphs. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. - ACM, 2006. - pp. 631-636.
4. Key Trends to Watch in Gartner 2017 Emerging Technologies Hype Cycle. https://www.gartner.com/smarterwithgartner/gartners-top-10-technology-trends-2017/
5. Khakimzyanov Ruslan N., Yunusova Diana A., PSYCHOBIOLOGICAL DETERMINANTS OF PSYCHOLOGICAL WELL-BEING OF AN INDIVIDUAL // QUID-INVESTIGACION CIENCIA Y TECNOLOGIA. - 2017. - Vol., Is. 28. - pp. 1433-1437.

6. M. Najork, J.L. Wiener. Breadth-first crawling yields high-quality pages. Proceedings of the 10th international conference on the World Wide Web. - ACM, 2001. - pp. 114-118.
7. S.N. Martyshenko. Methods for recovering gaps in the data presented in various measuring scales // Territory of new opportunities. 2013. No. 4 (22).
8. 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.
9. A.A. Popov, A.A. Holdonov. Structural optimization of fuzzy regression models with minimization of forecast errors in the test sample // Vestnik NGIIE. 2018. {2-81}
10. V.M. Ponyatsky, S.I. Veleshki, A.V. Zhirnova. Using the method of group accounting of arguments to select the structure of a model of a dynamic object // Izvestiya TulGU. Engineering sciences. 2013. No 2.
11. Rezaei, M., & Nemati, K. (2017). The Impact of Purchase Intent, Word of Mouth Advertising and Skill Domain of Seller on Quality of Customer Relationship to Sale Life and Savings Insurance Policies (Case Study: Dana Insurance Co., Bushehr Province). Dutch Journal of Finance and Management, 1(2), 43. https://doi.org/10.29333/djfm/5819
12. Cota, M. P., Rodríguez, M. D., González-Castro, M. R., & Gonçalves, R. M. M. (2017). Analysis of Current Visualization Techniques and Main Challenges for the Future. Journal of Information Systems Engineering & Management, 2(3), 19. https://doi.org/10.20897/jisem.201719
13. Mendoza, D. J., & Mendoza, D. I. (2018). Information and Communication Technologies as a Didactic Tool for the Construction of Meaningful Learning in the Area of Mathematics. International Electronic Journal of Mathematics Education, 13(3), 261-271. https://doi.org/10.12973/iejme/3907
14. Feizuldayeva, S., Ybyraimzhanov, K., Mailybaeva, G., Ishanov, P., Beisenbaeva, A., & Feizuldayeva, S. (2018). Vocational training of future elementary school teacher by means of realization of inter-subject continuity. Opción, 34(85-2), 479-516.