Application of Clusterization for Analysis of Virtual Community Users

Anna Synko[a], Kateryna Molodetska[b]

[a] *Department of Social Communication and Information Activities, Lviv Polytechnic National University, 12 S. Bandery str., Lviv, 79000, Ukraine*
[b] *Department of Computer Technologies and Systems Modeling, Zhytomyr National Agroecological University, 7, Staryi Blvd str., Zhytomyr, 10008, Ukraine*

Abstract
This paper deals with an artificial neural network for solving problems that would argue impossible or difficult by statistical or human standards. The article presents the analysis of virtual communities to their characteristics and components. The main goal is analysis of authorized members of virtual communities which are in the forums. It was a question of a big amount of data which have to be processed. After thorough research it was selected Kohonen neural network to deal with this issue. This computational method has self-learning capabilities that enable them to produce better results as more information becomes available to represent the operation of the algorithm. A model was built as a function modeling methodology for describing functions, decisions and activities of a system. The advantages and disadvantages of presented algorithm has been provided. Based on the research data, the results were optimized and forecasted.

Keywords 1
classification, user analysis, an artificial neural network, virtual community, clustering, forum, Kohonen network, self-organizing maps.

1. Introduction

Today, World Wide Web has become an all-encompassing phenomenon on a global scale. And it changed people life and activity forever. People communicate, do research, seek information and support for the competition of the tasks online. The big share is played by social networks that bring together people and organizations from different backgrounds. People use them to talk, exchange ideas to each other. World statistics show that the most popular sites are search engines (for example Google, Yandex, Baidu, ect.), social networks (Facebook, Instagram, Twitter, Вконтакті, Reddit, ect.), marketplaces (for example Amazon) and news sites (Yahoo, Naver) at August 2010 [1]. There has also been a tendency in recent years to the growth in the number of registered users.

When we having reliable and credible information about a qualitative assessment of members (authors) of virtual community which produce different types of materials, we can quickly find the relevant information. Owing to the large amount of data it is an important and highly urgent matter.

2. Related works

The article [2] analyzed the meaning of the concept of "social network"; the main stages of creation, development, distribution and use of social networks; highlighted their functions and features in modern period; the classification of social networks and their impact on modern society is given.

IT&AS'2021: Symposium on Information Technologies & Applied Sciences, March 5, 2021, Bratislava, Slovakia
EMAIL anna.i.synko@lpnu.ua (A. Synko); kmolodetska@gmail.com (K. Molodetska)
ORCID: 0000-0001-9864-2463 (K. Molodetska)

Management of virtual communities was studied Y. Zhang, A. Skinner, Y. Serov. Socio-demographic characteristics of users of the virtual community (S. Fedushko) [3]. Creating and managing the content of the virtual community (T. Berners-Lee, A. Croll). Systems of site indicators and methods that take into account the ratio of site indicators (A. Peleshchyshyn) [4].

One of the leading common methods of data analysis is clustering. The problem of clustering is solved using various methods, the choice of which should be based on the study of the original data set. The complexity of clustering is the need for its expert evaluation.

Theoretical aspects of the application of cluster analysis are devoted to the scientific works of many domestic and foreign scientists, in particular B. Everit, D. Cherezov, T. Harris, R. Tkachenko, L. Young, S. Schultz and others ([5], [6], [7], [8], [9]). These and other authors have formed a mathematical basis for the application of cluster analysis in various fields.

Scientists, including V. Golovko [10], B. Soilis [11] and others, pay attention to the issue of qualitative distribution of users into groups on the basis of clustering. A small amount of research in this area necessitates the development and improvement of cluster analysis techniques for qualitative characterization of authors of publications.

3. Purposes

The purposes of article are:

- to explore the role of virtual communities and their components;
- to analyze members of virtual communities by identifying the competitive advantages of registered users;
- to build a model according to the IDEF0 methodology.

4. Characteristics of virtual communities and their components

There are many different social networks on the World Wide Web that can be classified on different grounds. Any social network contains a variety of virtual communities (VC). The VC is social groups of people who communicate and interact via the Internet through computer communication [12]. Virtual communities differ from each other in subject matter, general structure, organization of content, functionality [4]. Each virtual community is unique.

Virtual communities are divided into six main groups [5, 13-14]:

- Academic virtual communities (Academia.edu);
- Educational virtual communities (The Student Room Group, ePALS School Blog ect);
- Information virtual communities (Do-It-Yourself Community, HGTV Discussion Forums ect);
- Multimedia virtual communities (YouTube, Flickr, Periscore, Instagram ect);
- Professional virtual communities (Xing, LinkedIn, Sumry, Slack ect).
- Virtual communities for communication (Twitter, Facebook ect);

General characteristics of VC are:

- access to information content (open, closed);
- structure content (semistructured, unstructured);
- by the degree of stability (temporary (unstable), medium and stable);
- by size (large, medium, small);
- by content (socio-ethnic, socio-demographic, socio-professional, professional, etc. [15-16]);
- by types (organizational or communication);
- by methods of implementation (chat, guest book, forum, blog) [4]

The main components of all virtual communities are its members and information content [17].

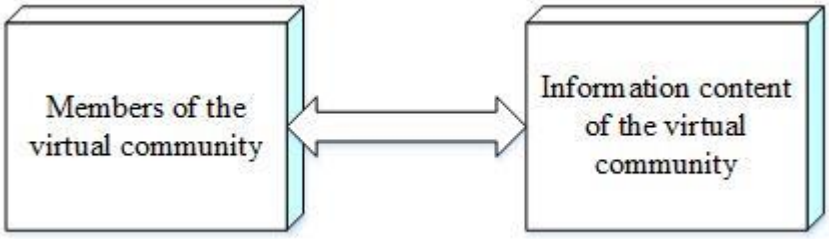

Figure1: Web community architecture

This paper will discuss the analysis of registered users of virtual communities.

5. Analysis of members of virtual communities

The main feature of the virtual community is that the user must have his own registered profile, which contains information about him. Upon receipt of this data can be used different methods and tools for analyzing users.

To achieve this goal, it was chosen to use cluster analysis, which is widely used in various fields. It is useful when it comes to classify a large amount of information [19]. Typically, clustering is the initial stage of a mathematical study of objects, followed by further steps such as optimization and forecasting [20]. One of the most important tasks of using cluster analysis in this study is to analyze the qualitative assessment of users, namely grouping users into homogeneous classes to get the most complete picture of their experience, activities and reputation (selected from other users' reviews) on the forum. As a result of cluster analysis, a map built to determine the level of quality of the user who is a member of the community in the forum, which greatly facilitates the perception of data and provides an opportunity to make new hypotheses.

Task. Today, there are many online services, such as Reddit, Stack overflow or Cyberforum, where registered users can publish their articles, scientific materials, post or direct links to certain topics, as well as leave comments and feedback on other works.

For clustering was chosen network Reddit, community – «r/C_Programming» (https://www.reddit.com/r/C_Programming/), which contains 94 thousand users.

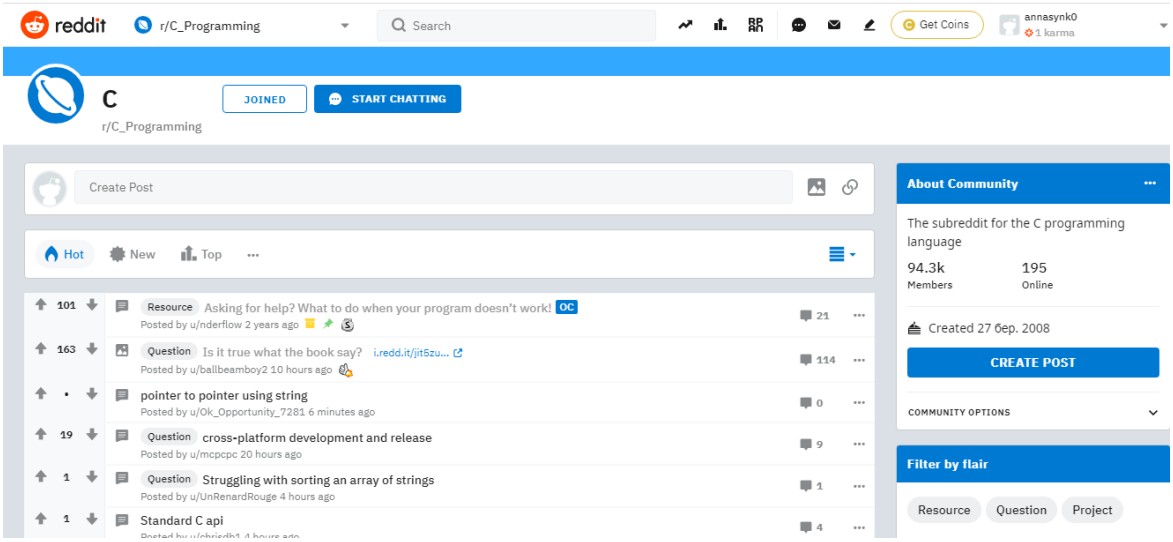

Figure 2: The community of C Programming

The components of the virtual community can be represented as a cortege:

$$VC_i = \langle \begin{matrix} Name_i, DateCreation_i, CountOfMembers_i, \\ CountOfOnlineMembers_i, Posts_i, Filters_i \end{matrix} \rangle, \quad (1)$$

where $Name_i$ is the name of the virtual community, $DateCreation_i$ – community creation date, $CountOfMembers_i$ – number of community users, $CountOfOnlineMembers_i$ – the number of users that are online, $Posts_i$ – publications, $Filters_i$ – filters for selection of posts.

Community filters are defined by a cortege:

$$Filters(VC_i) = \langle New(Posts_i), Hot(Posts_i), Top(Posts_i) \rangle, \quad (2)$$

where $New(Posts_i)$ is the selection of new posts (by date of posting), $Hot(Posts_i)$ are the most discussed publications, $Top(Posts_i)$ are the most visited posts.

This clustering is a descriptive procedure, it does not make any statistical conclusions, because the purpose of clustering is to search for existing structures, but, instead, makes it possible to conduct exploratory analysis and study the «data structure» [18].

For example, here is information about one of the users who is a member of this community:

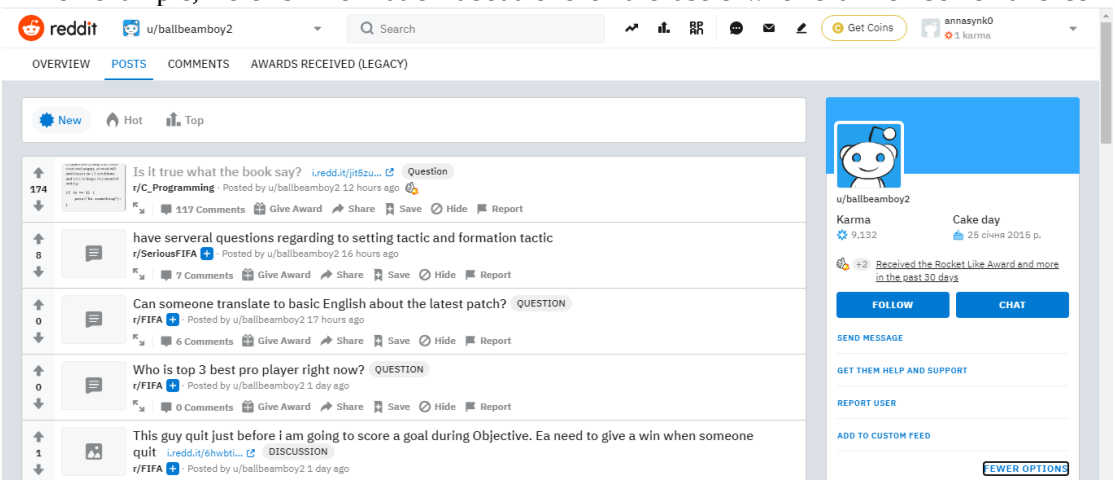

Figure 3: Information about the selected user from the community

The main nodes of the virtual community member's page are the user's personal information. User information can be represented by a tuple of such elements:

$$User(VC_i) = \langle \begin{matrix} PersonalData(User_i), Carma(User_i), Overwiev(User_i), \\ Posts(User_i), Comments(User_i), AwardsReceived(User_i) \end{matrix} \rangle, \quad (3)$$

where $User(VC_i)$ is the profile of the i-th user in the virtual community, $PersonalData(User_i)$ – personal data, $Carma(User_i)$ has a user rating, $Overwiev(User_i)$ is an overview page, $Posts(User_i)$ – publications, $Comments(User_i)$ are comments left by the i-th user, $AwardsReceived(User_i)$ are awards.

CARMA is a program that collects continuous ratings of users posts, comments and others data. CARMA gives a tool for researchers in affective computing, human-computer interaction, and the social sciences who need to capture the unfolding of subjective experience and observable behavior over time.

The personal data of the user will be defined by the following cortege:

$$PersonalData(User_i) = \langle \begin{matrix} Name(User_i), DateOfBirth(User_i), DateOfRegistration(User_i), \\ Photo(User_i), WorkExperience(User_i), Other(User_i) \end{matrix} \rangle, \quad (4)$$

where $Name(User_i)$ – is a name of i-th user, $DateOfBirth(User_i)$ is a date of birth, $DateOfRegistration(User_i)$ is the date of i-th user registration on the forum, $Photo(User_i)$ – user photo, $WorkExperience(User_i)$ – work experience, $Other(User_i)$ – other information.

The sets of posts, comments and awards posted by the user are defined as follows:

$$Post(User_i) = \{Post_j(User_i)\}_{j=1}^{N_i^{UPost}}, \quad (5)$$

$$PComments(User_i) = \{Comments_j(User_i)\}_{j=1}^{N_i^{UComments}}, \quad (6)$$

$$AwardsReceived(User_i) = \{AwardsReceived_j(User_i)\}_{j=1}^{N_i^{UAwardsReceived}}, \quad (7)$$

where N_i^{UPost}, $N_i^{UComments}$, $N_i^{UAwardsReceived}$ is a number of publications, comments and awards of the *i*-th user.

The overview page contains the following sets:

$$Overwiev(User_i) = \left\{ \begin{array}{c} \{Post(User_i)\}_{j=1}^{N_i^{UPost}}, \\ \{Comments(User_i)\}_{j=1}^{N_i^{UComments}}, \\ \{AwardsReceived(User_i)\}_{j=1}^{N_i^{UAwardsReceived}} \end{array} \right\}, \quad (8)$$

Based on these data, clustering can be made. Since the clustering itself does not give specific analysis results, it is necessary to perform a meaningful interpretation of each cluster to obtain the effect. So, based on the data about the presence of users on the forum, we will divide them into three groups by element DateOfRegistration($User_i$):

1) those who are less than a year on the online service (junior);
2) those who are from one to five years on the online service (middle);
3) those who are from five years and older on the online service (senior).

In Figure 2 we see that the user has been on the forum since February 2015, so we can refer it to the third group (senior).

Clustering of the data of the task will be implemented in the following stages:

a) Selection of characteristics (selection of properties that characterize the selected objects. The obtained data must be normalized. Then all objects are represented as characteristic vectors (Figure 4), which makes it possible to further identify the object with its characteristic vector);
b) Definition of metrics (choice of metrics, which determines the proximity of objects. Metrics are selected depending on: the space in which the objects are located; implicit characteristics of clusters. Usually use the classical Euclidean metric - formula 7);

$$d^2(x_i, x_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2\right)^{1/2} = \|x_i - x_j\|_2, \quad (7)$$

c) Presentation of results in a convenient form for further evaluation of the quality of clustering (cluster representation by a set of characteristic points was used).

The characteristics of users to identify internal relationships, dependencies, patterns that exist between objects are:

- number of posts (messages) - activity. Evaluate from 1 to 5;
- experience in the field ($WorkExperience(User_i)$). Evaluate from 1 to 10;
- reviews from other users ($Carma(User_i)$). Evaluate from 1 to 6 (score 1– 1-2 points, score 2 - 3-5 points, etc.).

Based on the above, we choose a neural network that is implemented by the method of training without a supervisor [20].

Learning without a teacher is one of the methods of machine learning, in solving which the test system spontaneously learns to perform the task, without interference from the experimenter. As a rule, this is suitable only for problems in which the description of a set of objects is known (training sample), and it is necessary to identify the internal relationships, dependencies, patterns that exist between objects. Methods for solving such problems are graph clustering algorithms, k-means clustering, deep network of persuasion, Kohonen neural network. To solve this problem, the Kohonen neural network was chosen, which has its advantages.

Software implementation. Software that allows you to work with Kohonen maps is now represented by many tools. These can be both tools that include only the implementation of the method

of self-organizing maps, and a neural package with a set of neural network structures, including Kohonen maps. Also, this method is implemented in some universal data analysis tools.

The tools that include the implementation of the Kohonen map method include MATLAB Neural Network Toolbox, Statistica, SoMine, Deductor, NeuroShell, NeuroScalp and many others. To solve this problem, the Python programming language was selected and its built-in functions and commands were applied.

For the experiment was selected 500 members. Due to the fact that there are many objects (the first group has 22 users, the second group has 240, the third group has 238), all their data were entered in a separate file (Figure 4). So we have three user groups (500 objects in total), each of which has with three characteristics. 1000 iterations were selected for training. The map has a size of 7×7 (Figure 5).

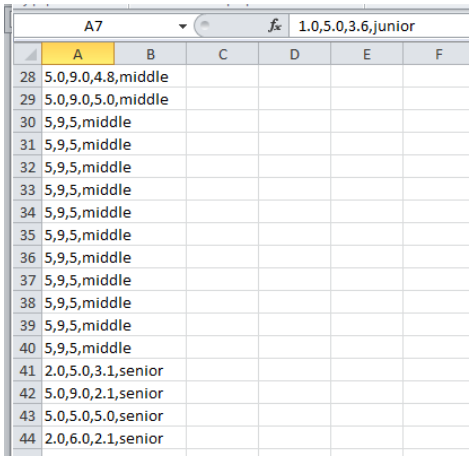

Figure 4: The data is arranged according to user groups

The process of learning the Kohonen map (Self Organizing map) takes place in two stages: the stage of ordering the vectors of weights in the feature space and the stage of adjustment.

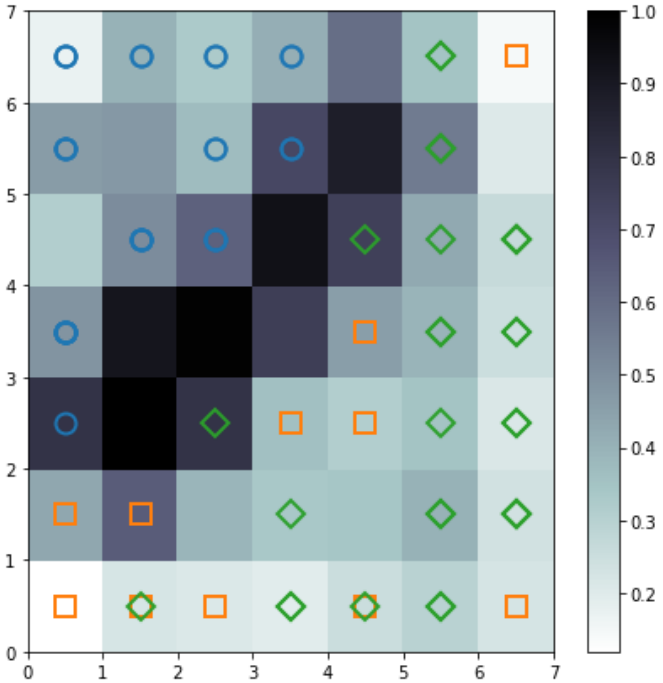

Figure 5: Solve the problem using a Kohonen map that has been trained without a supervisor

As can be seen in Figure 5, each group of users has its own color: junior - blue; middle - green; senior - brown. A scale was also constructed to show the distances between objects (the darker the paint, the greater the distance).

However, it should be emphasized that this experiment is not an end in itself because the ultimate goal of clustering is to obtain meaningful information about the structure of the studied data. The

obtained results require further interpretation, research and study of the properties and characteristics of objects to be able to accurately describe the formed clusters [19].

To solve the problem of identifying the competitive advantages of users who are registered at the forum, using self-organizing maps (SOM), give the algorithm:

1) data collection from the selected community;
2) distribution by user groups according to certain characteristics;
3) training without a supervisor using SOM.

To solve the problem, the notation of the description of business processes IDEF0 was chosen. This is where the interaction of processes, mechanisms and control signals is reflected. Therefore, a context diagram was constructed, which is the most general description of the system and its interaction with the external environment.

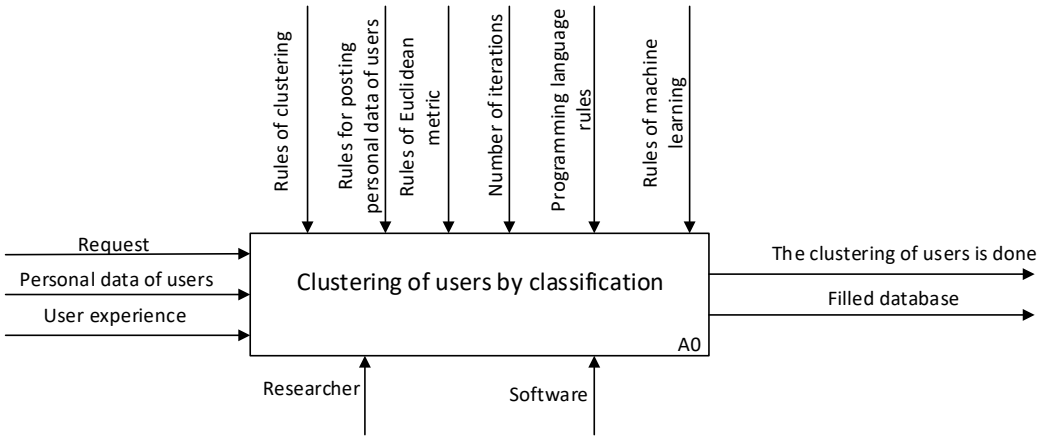

Figure 6: Contextual functional diagram for classification

After the description of the system, its detailing is performed in the form of a functional diagram of the 1st level (decomposed context diagram):

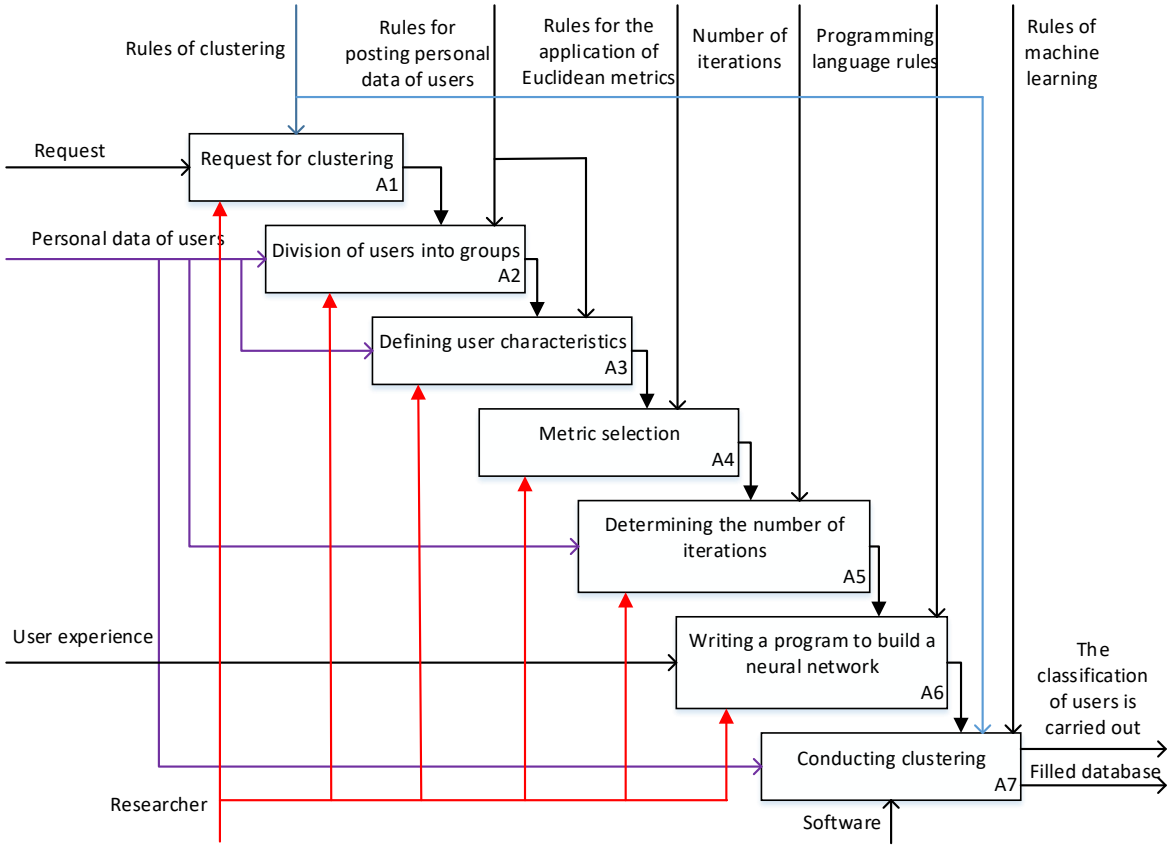

Figure 7: Functional diagram of the 1st level

The advantages of using this method to solve this problem are:

- resistance to data noise;
- unmanaged learning;
- possibility of visualization (built map);
- fast learning;
- the possibility of simplifying the multidimensional structure

As any method can have its drawbacks, the chosen system has its:

- the choice of learning factor (affects both the speed of learning and the stability of the decision);
- randomization of weights (randomization of weights of the Kohonen stratum can cause serious learning problems, as this operation distributes weight vectors evenly over the surface of the hypersphere. As a rule, the input vectors are unevenly distributed and grouped on a relatively small part of the surface of the hypersphere. Therefore, most weight vectors will be so distant from any input vector that they are not activated and become useless. Moreover, the remaining activated neurons may be too small to split the nearest input vectors into clusters);
- selection of initial values weights of vectors and neurons (if the initial values are chosen unsuccessfully, i.e., for example, are located far from the proposed input vectors, the neuron will not be the winner for any input signals, and therefore will not learn);
- selection of the distance parameter (If the initially selected parameter is small or decreases very quickly, then far apart neurons will not be able to influence each other. Although the two parts on such a map are located correctly, the overall map will have a topological defect (Fig. 8)).

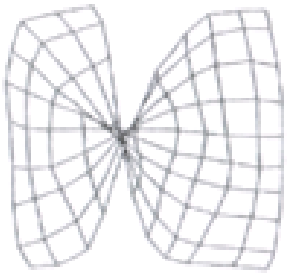

Figure 8: Topological defect of the map

So, after conducting research, taking into account all the disadvantages and advantages of the chosen method, we obtained the following result. Firstly, each user in the community can be useful (post relevant and reliable materials, publications regardless of his age, position, experience work, etc.), because the junior user group, those who have been in the community for less than a year, may be young people who could not register on the site ten years ago, have a fairly high score according to the characteristics of the study.

Second, in order to better select data on users who publishing materials, developers need to encourage users to enter as much information as possible about themselves, or to make these fields mandatory fields when registering. Of course, there may be an error, because not enough characteristics are selected for the study of this industry (the more data the less error). What makes developers think about this issue.

Third, this method of user selection should be of interest to analysts, developers to create additional functions for user selection. So that looking for the necessary information on the forum or in the community you can choose the "most useful" author of materials in the selected field

6. Conclusion

So, looking at the solution of the problem, we can conclude that no matter how long the user is on any online resource (and, of course, is registered on it), he can be no less useful to society, even regardless of his age, position, work experience, etc. Of course, these are important factors that cannot

be ignored, but there is another factor in how quickly a person finds, perceives and masters new information. For example, if it used to take years to study a particular discipline, now it has become more accessible thanks to new technologies and previous discoveries.

Also, if it used to take years to be a programming specialist, now it takes months, because everyone has access to a computer and the Internet, where they can find theory and practice. Of course, the Internet has its drawbacks, because there is a lot of information that is not reliable or even false. To do this, a classification of users by their characteristics was created to select the highest quality, reliable information.

Because clustering is one of the leading methods of data analysis, one of its approaches was chosen the Kohonen neural network. For a clear idea of the sequence of steps that the system has gone through to achieve the task, the scheme was given Figure 6 and Figure 7. The advantages and disadvantages of using the chosen approach, which should not be neglected in other similar studies, have also been described.

The chosen direction of research is relevant and needs further study, because the problem of redundant information, and not always reliable, is current and important for everyone who uses the Internet to search for any data. The chosen field of research is relevant and needs further study, because the problem of redundant information, and not always reliable, is relevant and important for everyone who uses the Internet to search for any data. Therefore, the issue of selecting high-quality, reliable information published by users is not fully disclosed, as there are other factors that can be used to analyze the data (for example, to make it mandatory for users to link to literature sources where they got this information. These may be the following characteristics: knowledge of foreign languages, the position held by the user, etc.).

To use all of these features for further analysis, developers of such communities and forums need to encourage users to enter as much information as possible about themselves, or to make these items mandatory fields when registering.

References

[1] Top Websites Ranking. Top sites ranking for all categories in the world. SimilarWeb, 2020. URL: https://www.similarweb.com/top-websites/

[2] O. Hotko, O. Chaikovska, N. Nalyvayko, Social internet networks and virtualization of public life, 2016, Vol.2, p. 94-98. URL: http://nbuv.gov.ua/UJRN/Mir_2016_2_23.

[3] S. Fedushko, Development of a software for computer-linguistic verification of socio-demographic profile of web-community member. Webology, Volume 11, Number 2, Article 126. URL: http://www.webology.org/2014/v11n2/a126.pdf

[4] I. Korobiichuk, Y. Syerov, S. Fedushko, The Method of Semantic Structuring of Virtual Community Content. Mechatronics 2019: Recent Advances Towards Industry 4.0. Advances in Intelligent Systems and Computing, vol 1044. Springer, Cham, 2020. pp 11-18. https://doi.org/10.1007/978-3-030-29993-4_2

[5] B. Everitt, S. Landau, M. Leese, D. Stahl, Cluster Analysis. Wiley, 2010. 346 p.

[6] O. Anisimova, H. Lukash, Y. Syerov, Formation of the portrait of the specialist in social networks. CEUR Workshop Proceedings, 2020, 2616, pp. 39–52. http://ceur-ws.org/Vol-2616/paper4.pdf

[7] R. Tkachenko, I. Izonin, Model and Principles for the Implementation of Neural-Like Structures based on Geometric Data Transformations. Advances in Computer Science for Engineering and Education. ICCSEEA2018. Advances in Intelligent Systems and Computing. Springer, Cham, vol.754, pp.578-587, 2019. DOI: 10.1007/978-3-319-91008-6_58

[8] A. Barsegyan, M. Kupriyanov, V. Stepanenko, I. Holod, Data analysis technologies. Data Mining, Visual Mining, Text Mining, OLAP. BHV-Petersburg Publisher, 2007. 384 p.

[9] J. H. Wang, J. D. Rau, W. J. Liu, Two-stage clustering via neural networks, IEEE Transactions on Neural Networks, 2003, Vol. 14, p. 606-615.

[10] I. F. Yasinskiy, Neural network self-organization method for processes prediction with penalty for complexity and optional structure, 2013. http://vestnik.ispu.ru/sites/vestnik.ispu.ru/files/publications/str._61-63.pdf

[11] B. Solis. The role of modern social networks in society and political technologies. Direct Media, 2012, 8 p.
[12] K. Shakhovska, N. Shakhovska, P. Veselý, The sentiment analysis model of services providers' feedback. Electronics (Switzerland), 9 (11), art. no. 1922, 2020, pp. 1-15. DOI: 10.3390/electronics9111922
[13] A.V. Turchin, Classification of social networks. 2016. PhD Thesis. KNTU
[14] Ntarlas, Gerasimos, Athina Ntavari, and Despina A. Karayanni, The Strategic Use of Social Media in the Business-to-Business Context. Two Social Media Users' Clusters. Strategic Innovative Marketing and Tourism. Springer, Cham, 2020. 825-833.
[15] V.V. Verbetsi, O.A. Subot, T.A. Khrystyuk, Sociology. Kyiv: Kondor Publishing, 2009, 550p.
[16] O.O. Nestulya, Sociology, Kyiv: Center for Educational Literature, 2009. 272 p.
[17] Di, Mu, Structural equation modelling for influencing virtual community networks. International Journal of Web Based Communities 16.3, 2020, 249-261.
[18] Ahamed. M. Mithun and Z. A. Bakar, "Empowering Information Retrieval in Semantic Web," IJCNIS, vol. 12, no. 2, pp. 41–48, Apr. 2020, doi: 10.5815/ijcnis.2020.02.05.
[19] E. E. Haji, "Proposal of a Digital Ecosystem Based on Big Data and Artificial Intelligence to Support Educational and Vocational Guidance," p. 11, 2020.
[20] O. Ebenezer, "Influencing Children: Limitations of the Computer-Human-Interactive Persuasive Systems in Developing Societies," IJMECS, vol. 12, no. 5, pp. 1–15, Oct. 2020, doi: 10.5815/ijmecs.2020.05.01.
[21] N.B. Paklin, V.I. Oreshkov, Business Intelligence: From Data to Knowledge. 2013, 704 p.
[22] Musci, Mirto, et al. A scalable multi-signal approach for the parallelization of self-organizing neural networks. Neural Networks 123, 2020, 108-117.
[23] T. Kohonen, Self-organizing cards. M.: Binom Publisher, 2008. 656 p.
[24] R. Hryshchuk, K. Molodetska, Y. Syerov, Method of Improving the Information Security of Virtual Communities in Social Networking Services. CEUR Workshop Proceedings. 2019. Vol 2392: Proceedings of the 1st International Workshop on Control, Optimisation and Analytical Processing of Social Networks, COAPSN-2019. p. 23–41. http://ceur-ws.org/Vol-2392/paper3.pdf