CLUSTERING METHODS IN LARGE-SCALE SYSTEMS

Denis V. Gadasin,
Moscow Technical University of Communications and Informatics, Moscow, Russia
dengadiplom@mail.ru

Andrey V. Shvedov,
Moscow Technical University of Communications and Informatics, Moscow, Russia
a.v.shvedov@mtuci.ru

Alyona A. Yudina,
Moscow Technical University of Communications and Informatics, Moscow, Russia
alenka5yudina@mail.ru

DOI: 10.36724/2664-066X-2020-6-5-21-24

ABSTRACT

Interactions between people, groups, organizations, and biological cells have a relationship character that can be represented as a network. The system properties of such networks, regardless of their physical nature, but clearly determining the performance of networks, create the totality of the real world. Complex networks – are naturally existing networks (graphs) that have complex topological properties. The researchers who participate and also make discoveries in this field come from various Sciences such as mathematics, computer science, physics, sociology, and engineering. Therefore, the results of research carry both theoretical knowledge and practical applications in these Sciences. This paper discusses the definition of complex networks. The main characteristics of complex networks, such as clustering and congestion, are considered. A popular social network is considered as a complex network. The calculation of nodes and links of the considered social network is made. The main types of AI development and training are highlighted.

KEYWORDS: *complex networks, network implementation, technology, node, communication, network, artificial intelligence, algorithm, characteristics, clustering, workload, training, algorithm, analysis, solution, problem*

INTRODUCTION

Often in practice, the interaction between different objects and/or subjects is expressed in the form of established relationships or associations that show how the relationships of system elements relate or interact. Any of these interactions can be represented as a "Question-Answer" bundle, which is defined as binary relationships in which each object represents a specific point that is connected to another object using a line or arc. Another factor also comes up here – the inheritance Paradigm, which is based on distributed axioms of biosphere unity: fuzziness – clarity (1D and 2D) of continuous nature and is built on three clusters of inseparability. Two incompatible universal laws of nature are valid -Transformation and Conservation, for the closeness of the basis the third dogma is inevitable – Inheritance 3D [1].

Information about authors

Denis V. Gadasin, *Ph. D., Associate Professor, Department of Network information technologies and services, MTUCI, Moscow, Russia*
Andrey V. Shvedov, *Senior lecturer, Department of Network information technologies and services, MTUCI, Moscow, Russia*
Alyona A. Yudina, *Undergraduate of the Department of Network information technologies and services, MTUCI, Moscow, Russia*

I. CONNECTIONS AND DISTANCES BETWEEN NODES

Every day, the relationship between a person and their personal digital devices becomes stronger and stronger. Along with this, the degree of immersion of people in the virtual world – in the world of the Internet, applications and social networks is also growing. Virtuality is absorbing more and more, drawing people deeper and deeper into its networks, and making people separate nodes of this network. Gadgets are constantly expanding their capabilities and connecting people in one digital, large-scale and complex social network.

A social network is a resource, a platform that provides relationships between people. Your profile is used to communicate and create an offline connection or a connection based on common interests. People upload their photos and information about themselves to their social network profiles and thus personal information becomes public, and the intensive development of information technologies has significantly increased interest and attention to the problem of privacy and security of personal data, including during their automated processing [2]. Thus, personal data may be used by other users, including for unfavorable purposes.

Each of the network nodes has a certain number of connections, combining with other nodes of the same type. If the connection between nodes has a direction, then the network is oriented, and if the connection is symmetric for all nodes connected by it, then the network obtained by such connections is called undirected. For example, if we assume that a connection exists if two people are close friends, then the network will be undirected. If we assume that there is a connection, if one person considers himself a friend of another, then the formed network will be oriented. The number of links of a node is determined by its degree. In oriented networks, there is an outgoing and incoming node degree. The degree distribution of nodes is an important characteristic of a complex network. The main part of all complex networks is close to the power law distribution of degrees of nodes with an exponent between 2 and 3. the Distance between nodes is the minimum number of connections that must be overcome in order to get from one node to another. For all network pairs, there is an average distance between them when moving from one node to another. This distance is called the average distance between nodes – d. For most complex networks, $d \approx \log N$, where N is the number of nodes in the network [3].

Determining the distance between nodes is closely related to clustering, which is one of the local characteristics of the network. It represents a characteristic of the level of interaction between the nearest neighbors of a node. The clustering task is to fragment objects from the Y-set into several subsets (clusters), in which the objects are more similar to each other than objects from other clusters. There are several main known clustering methods. In most cases, to choose the correct method, you need to determine the type of connection between network objects.

II. THE METHOD OF NEAREST-NEIGHBOR OR SINGLE LINK

In this method, the distance between two clusters (Figure 1) is determined by the smallest distance between two close objects in different clusters (1).

$$R^{\delta}(W,S) = \min_{w \in W, s \in S} \rho(w,s) \quad (1)$$

$$\alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$$

where W , S – clusters.

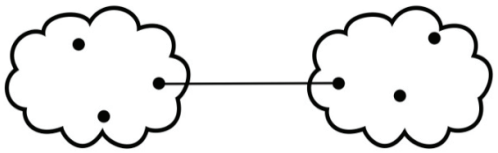

Figure 1. The method of nearest-neighbor or single link

If different parts of such clusters are connected by chains of elements that are close to each other, this method allows you to select clusters of arbitrarily complex shapes. The result of this method is clusters represented by long "chains" that are "linked together" only by individual elements that happen to be closest to each other.

III. MOST DISTANT NEIGHBOR METHOD OR FULL CONNECTION

In this method, the distance between clusters (Fig. 2) is determined by the largest distance between any two objects in different clusters (2).

$$R^{Д}(W,S) = \max_{w \in W, s \in S} \rho(w,s) \quad (2)$$

$$\alpha_U = \alpha_V = \frac{1}{2}, \beta = 0, \gamma = \frac{1}{2}$$

where W , S – clusters.

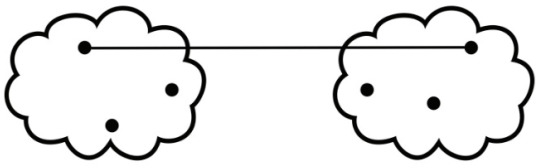

Figure 2. Most distant neighbor method or full connection

This method is good to use when objects actually come from different structures.

IV. WARD'S METHOD

One of the main clustering methods is the Ward's method, which is constructed in such a way as to optimize the minimum variance within clusters. This objective function is known as the intragroup sum of squares or sum of squared deviations (3).

$$CKO = x_j^2 - \frac{1}{n \bullet \left(-x_j\right)^2} \quad (3)$$

where x_j is the value of the attribute of the j-th object. In the first step, the sum of squared deviations is 0. This method is aimed at combining closely located clusters and tends to create small clusters.

V. UNWEIGHTED PAIR GROUP METHOD USING ARITHMETIC AVERAGES (UPGMA)

The distance between two clusters (Figure 3) is the average distance between all pairs of objects in them. The distance between these clusters is determined by (4).

$$D\big((u,\upsilon),\omega\big)=\frac{T_u D_{u,\omega}+T_\upsilon D_{\upsilon,\omega}}{T_u+T_\upsilon} \quad (4)$$

where u, ω are clusters containing T_u, T_ω objects, respectively.

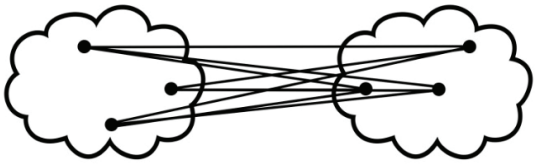

Figure 3. Unweighted pair group method using arithmetic averages (UPGMA)

This method should be used if the objects actually come from different structures. In cases where there are clusters of the "chain" type, assuming unequal cluster sizes.

VI. WEIGHTED PAIR GROUP METHOD WITH ARITHMETIC MEAN (WPGMA)

This method uses the cluster size (the number of objects contained in the cluster) as the weighting factor. The distance between clusters is determined according to the (5).

$$D\big((u,\upsilon),\omega\big)=\frac{D_{u,\omega}+D_{\upsilon,\omega}}{2} \quad (5)$$

This method should be used only if there is an assumption about clusters of different sizes.

VII. UNWEIGHTED PAIR GROUP METHOD WITH CENTROID AVERAGE (UPGMC)

In this method, the distance between two clusters (Figure 4) is taken as the distance between their centers of gravity. The distance between the centers of gravity is determined by (6).

$$R^{U}(W,S)=\rho^2\left(\sum_{w\in W}\frac{w}{|W|},\sum_{s\in S}\frac{s}{|S|}\right) \quad (6)$$

$$\alpha_U=\frac{|U|}{|W|},\alpha_V=\frac{|V|}{|W|},\beta=-\alpha_U\alpha_V,\gamma=0$$

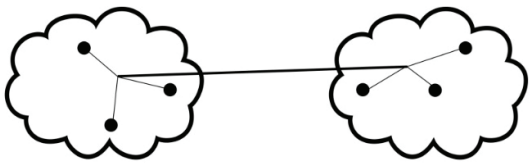

Figure 4. Unweighted Pair Group Method with Centroid average (UPGMC)

VIII. WEIGHTED PAIR-GROUP METHOD USING THE CENTROID AVERAGE (WPGMC)

This method is similar to the UPGMC method, the only difference is that weights are used to account for the difference between cluster sizes (the number of objects in them). This method is preferable if there are assumptions about significant differences in cluster sizes.

The clustering coefficient of a node is the probability that the two nearest neighbors of that node are themselves nearest neighbors. For example, as we are friends with our friend, so we are friends with his friends. The concept of friends of friends can also be found in social networks. Connection through friends of friends or friends at ambiguous numbers may be submitted to the original node. That is, going through a huge amount of links, we can return to the starting point at the source node, as all nodes in the network are interconnected. Each node must be processed and analyzed, which is very time-consuming work.

Using the example of one of the most popular social networks, such as Vk.com, which has been included in the top 10 most popular sites in Russia for several years, let's look at the task that the network has to face. The monthly audience of this site is 71 million people, this is the number of nodes involved in the social network. There are many more connections between these nodes. Let's assume that the text information about one participant corresponds to the volume that fits on a standard A4 sheet.
On average, a sheet contains 25 lines, each line containing 75 characters of 8 bits each, i.e. the size of a single page is approximately 1.5 KB, and the amount of information for active users is about 105 GB. Based on the fact that this information is presented in the form of typewritten text and contains certain items (full name, address, place of study, etc.), it is structured and the easiest to analyze. In addition to text information, there is also information in various graphic and multimedia formats (drawings, photos, videos, audio), which is more difficult to structure and is not considered in this work. We can assume that if we consider a social network as an information network consisting of nodes and communication channels, it contains 71 million nodes. Communication channels determine the interaction of users, i.e. their familiarity, and it is more likely that all users of the network are familiar with each other through their friends. Thus, we need to determine such connections, that is, find a route from one node to another, and then conduct an analysis. It is known that the number of routes will be equal to $n!$, where n is the number of nodes, that is, in our case, the number of routes is equal to 71000000!. Based on the Stirling's approximation:

$$n!\approx\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \quad (7)$$

After making calculations, the value of the number of routes is $10^{284000000}$. Therefore, the total amount of information analysis is 1.5 KB * $10^{284000000}$ which is much more than the original 105 GB.

This volume must be analyzed every time a change occurs. The problem is difficult to understand and only AI can solve it. The more times It is used to solve the same type of problem, the faster it learns, and the more difficult this task is, the higher the level of intelligence. Training uses a large amount of data, which allows the AI to choose its own method and approach to training, and for complex network technologies, it is the basis for technological development.

The most popular task for segmentation data mining requires that a given group be fragmented into internal homogeneous clusters in order to better identify different groups of people who share a common set of characteristics.

Classical approaches model this problem on relational data. Each individual (data point) is described by a structured list of attributes. Indeed, in several scenarios, this modeling choice is an excellent proxy for dealing with context-sensitive issues. However, such methodologies alone cannot answer the natural but non-trivial question: What does it mean to segment a population for which the social structure is known in advance? The first way to solve this problem can be defined in a complex network analog of the data mining clustering problem-Community Discovery.

So far, many algorithms have been proposed for efficiently and efficiently splitting graphs into connected clusters, often maximizing specially adapted quality functions. One of the reasons why this task is considered one of the most difficult is its failure: there is no single, generally accepted definition of what a community should look like.

With today's huge amounts of data generated by next-generation networks, and the evolution of computing involving coordination between edge and core platforms, the need for data center evolution will be paramount. The next-generation mobile capabilities in sensing, visualization, and location will generate huge amounts of data that must be managed on behalf of network owners, service providers, and data owners.

When developing and researching AI, several different systems and methods are distinguished: state-space search; natural language processing; knowledge representation; expert systems and decision support systems; machine learning and artificial neural networks; genetic algorithms; multi-agent systems [4].

State-space search is when one of the search options is always in the focus of other artificial intelligence technologies. This method is a method in which there is a need and need to know which of the methods exist in order to perform a search and understand why everything is based on searching in the state space of an artificial intelligence system, as well as how and how artificial intelligence systems can use various types of heuristics. Natural language processing-using this method, AI systems are able to communicate with users in a language they understand, not only by entering commands, but also by voice.

Knowledge-based artificial intelligence systems use various formalizations that represent knowledge. There are still several such methods that are absolutely universal and to some extent reflect the ability of people to describe their knowledge. In addition, since the creation of the first knowledge-based systems, many mathematical methods have been developed to address the so-called factors of ignorance-completeness, unreliability, uncertainty, fuzziness, attenuation, and many others, which makes artificial intelligence systems able to work and make decisions in conditions of uncertainty, as a person does.

Thus, the most important class of knowledge-based systems are expert systems, which in turn often form the core of various decision support systems. Expert systems include an extended knowledge base of any problem area that collects and integrates expert knowledge (dynamic systems are constantly updated) and allows to make decisions based on them.

Artificial intelligence is still far from being able to generate algorithms on its own, and it is still being helped by humans. To improve the quality of work, the algorithm needs to be constantly expanded, fixing existing methods. For example, for social networks, the method of expert systems is used, which allows you to process a large amount of information, and as a result gives a conclusion. Functional work requires a fast response of the analysis and processing system, which will help to form data clusters in a timely manner to simplify problem solving in order to implement the tasks of social networks.

One of the main tasks is to form social and target groups. These groups and clusters are formed in stages and take milliseconds of time. Also popular in social networks is AI, equipped with the following functions: video editing and photo editor, telephone operator, designer and spam defender, consultant, marketing psychologist, diagnostic assistant.

An effective structure will help organizations gain confidence in their AI technology. This approach should go deep into AI in the enterprise and in the individual model to help ensure that key trust imperatives are integrated and controlled throughout. It must continually evaluate and maintain control over complex, evolving algorithms, establishing methods, controls, and tools that provide anchors of trust throughout the lifecycle, from strategy to evolution. It should also provide clear guidance to organizations and stakeholders in the various management and oversight functions [5].

IX. Conclusion

Performing tasks that are usually handled by humans is the main property of AI. These problems lead to the fact that our life is greatly simplified. Each appearance of new devices is reduced to the fact that their dimensions become minimal, and performance increases significantly. Many of these devices are equipped with special sensors that allow you to "communicate" with the network, with a person and with other devices, transmitting information through communication channels along various routes. In this case, you need to perform a large data analysis, which is capable of AI. After all, the task of statistics, collecting information, as well as storing information about a selection of objects, and ordering it is not an easy task even for AI, but it is still possible to solve. To help, the AI uses clustering, the choice of which method depends on the goals and objectives set. AI intelligence, in turn, can provide reliability, speed, and the ability to independently make decisions on a number of tasks. However, as the level of integration of new technological architectures increases, the risks of security and privacy violations also increase.

REFERENCES

1. D.V. Gadasin, A.V. Shvedov and Y.S. Litvin, "Paradigm of Inheritance in Large-Scale Systems,"*2019 Systems of Signals Generating and Processing in the Field of on Board Communications*, Moscow, Russia, 2019, pp. 1-5, doi: 10.1109/SOSG.2019.8706804.

2. V.A. Dokuchaev, V.V. Maklachkova, D.V. Makarova and L.V. Volkova, "Analysis of Data Risk Management Methods for Personal Data Information Systems,"*2020 Systems of Signals Generating and Processing in the Field of on Board Communications,* Moscow, Russia, 2020, pp. 1-5, doi: 10.1109/IEEECONF48371.2020.9078547.

3. A.T. Terekhin, E.V. Budilova, M.P. Karpenko, L.M. Kachalova, E.V. Chmyhova. Lyapunov function as a tool for studying cognitive and regulatory processes of the body. *Computer research and modeling*-2009 Vol. 1 No. 4. P. 449-456.

4. R.V. Dushkin, D.A. Movchan. Artificial intelligence. Moscow: DMK-Press publishing House, 2019. 160 p.

5. F.V. Grechnikov, V.R. Kargin. Fundamentals of scientific research: textbook. Samara: SSAU Publishing house, 2015. 111 p.